Bringing parliamentary debates to the Semantic Web

Damir Juric[1,3], Laura Hollink[2], Geert-Jan Houben[1]

[1] Delft University of Technology, [2] VU University Amsterdam
[3] FER University of Zagreb

Abstract. An analysis of parliamentary debates and media resources that cover them can provide insight into the political climate of a country. Although debates are now regularly published on official government portals, their analysis remains a cumbersome and challenging task for historians and political scientists. One of the main tasks of the PoliMedia project is to allow easy cross-media comparisons and give better insight into choices that different types of media outlets make when covering parliamentary debates. As a first step of that task, Dutch parliamentary debate data available in XML files is being translated into Semantic Web standards, which will allow users to easily query the data. In this paper we discuss design choices we made to build a semantic model that will represent events and topics from the Dutch parliamentary debates.

Keywords: XML, RDF, SEM, Semantic Web, parliamentary debates

1 Introduction

In this paper we discuss ongoing work on the representation of political events on the Semantic Web. We present the design choices of a model in which we capture parliamentary debates, including how they are covered by various media.

Analyzing media coverage across several types of media outlets is a challenging task, especially for people who need deep understanding of the data and its implications, like media historians. Previous research has focused mainly on newspaper articles, because they are generally available in digital, computer-readable format. To make cross-media comparisons between different types of media outlets, links between datasets would need to be produced. For example, to support researchers that want to know how political debates are represented in the media and how the representation of topics and people change over time. We aim to facilitate this kind of analysis by providing links between datasets of political debate events and media data.

The media-historic research questions that guide the project are: "*What choices do different media make in the coverage of people and topics while reporting on debates in the Dutch parliament since the first televised evening news in 1956 until 1995? Does the representation of topics and people change over time and how do the various media types differ?*" These questions specify a number of things that needs to be expressed in the model, like people, topics, time and media types.

To answer these questions we first created a semantic model that is expressive enough to allow us to represent all important information about events from the Dutch parliament, that are recorded in the form of debate transcripts (and later in XML files). After this step, an RDF repository is created in which we instantiate the model with instances of debate events, that allows various interesting information to be extracted from this dataset using SPARQL queries.

This paper is organized as follows: first we describe the PoliMedia project in which this work is carried out. In Section 2 we give a description of our datasets of debate events and media items. In Section 3 we discus the semantic model, and in Section 4 we describe our future work.

1.1 Background: the PoliMedia project

The PoliMedia project[1] is driven by research questions from historians with respect to media coverage across several types of media outlets. Cross-media comparisons will be conducted over a longer period of time, on different topics. The project will focus on the coverage of the debates in the Dutch parliament and give insight on the different choices that different media make while reporting on those debates. Also, when research can be performed with time and media type in mind, another question can be answered: Does the representation of topics and people change over time and how do the various media types differ?

The project will be carried out in three phases: (1) a modeling phase: creating a semantic model (that is the phase described in this paper), (2) a data production phase: creating links between debates and associated media sources, and (3) an application phase: searching and navigating linked datasets.

1.2 Related work

Related work for this project comes from three domains: other projects using parliamentary debate data, event modeling and relatedness discovery.

In [1], the author describes the structure of parliamentary proceedings and sketches a widely applicable DTD. He also describes how proceedings in PDF format can be transformed into deeply nested XML files. The work described is done as part of a project called War In Parliament [2]. In the work described in this paper, we use structured XML files from War In Parliament as a basis for our model. This dataset can be searched on the Political Mashup portal [3]. [4] presents an approach that extends existing metadata enrichment processes with a method to discover historical events. The events are structured in a historical event thesaurus to enrich object metadata. As such, the event thesaurus is used as a bridge between objects in different collections. The results of the approach allows for topic-based and event-centered browsing, searching and navigating in integrated collections. In [5], the authors put events as the central elements in the representation of data from domains such as history, cultural heritage, multimedia and geography. The Simple Event Model (SEM)

[1] PoliMedia project: http://www.polimedia.nl/

is created to model events in these various domains, without making assumptions about the domain-specific vocabularies used. The researchers designed SEM with a minimum of semantic commitment to guarantee maximal interoperability. In [6] the authors describe real life problem using SEM. Some properties of SEM are used in the semantic model described in this paper. We used SEM model as a starting point on which we build our own model that conforms to the events in the parliament. The problem of link discovery is tackled in [7]: there a validation approach is presented of detected alignment links between dialog transcript and discussed documents, in the context of a multimodal document alignment framework of multimedia events (meetings and lectures). The validation approach consists of an entailment process of the detected alignment links. This entailment process exploits several features, from the structural level of aligned documents to the linguistic level of their tokens. In [8] the authors present a function that discovers relatedness between news articles across four aspects: relevance, novelty, connection clarity, and transition smoothness.

2 Description of datasets

The PoliMedia project is aimed at cross-linking four different datasets, each from different media outlets. All datasets, which are textual and audiovisual, are available via the CLARIN infrastructure.

Primary dataset for this project is a collection of Dutch parliamentary debates, the so-called *Handelingen der Staten-Generaal* or the *Dutch Hansard*. Parliamentary debates used in this project, are actual transcripts of speeches that politicians had in the parliament. At the time of writing this article, three sources of Dutch parliamentary debates are available online. On the Officiële Bekendmakingen portal, which is an official source for parliamentary debates from the Dutch parliament, debates can be found in an XML format, using XML schema and permanent identifiers. Existing identifiers point only to the debate as a whole, not specifically to parts of the debate. Also, only debates from 1995 till present are available at this source.

A second source for Dutch parliamentary debates can be found online, on the Staten-Generaal Digitaal portal[2], which contains debates from the parliament from before the year 1995. Data can be accessed publically using the SRU (Search and Retrieval via URL) [9] or OAI (Open Archives Initiative) [10] protocols. Contrary to the previous source, debates from this source posses no further structure (data is provided in txt or pdf formats).

A third source for political debates from the Dutch parliament can be found on Political Mashup [3]. This data is created by the CLARIN project War in Parliament (WIP). The project is still ongoing, and the way debates are published is continuously improving. At the time of writing this article, all debates until the year 1995 are published as XML documents (OCR with satisfactory quality is being used). This data shows a fine-grained structure.

[2] Staten-Generaal Digitaal: http://www.statengeneraaldigitaal.nl/

Secondary datasets contains different media types: newspaper articles, radio bulletins, and newscasts. The newspaper and radio bulletins dataset is available from the National Library of the Netherlands, which allows users to analyze the text of the articles and see in which way they are layered. Metadata of the articles and bulletins are available from the metadata store of the Koninklijke Bibliotheek (KB), the KB-MDO or Koninklijke Bibliotheek metadataopslag [11] as DIDL (Digital Item Declaration Language – an XML dialect [12]). The newscast dataset contains evening news and current affairs programs. Audiovisual content include program level metadata in Dublin Core and CDMI format, enriched with thesaurus terms from the Gemeenschappelijke Thesaurus Audiovisuele Archieven (GTAA). Data can be accessed using the OAI-PMH protocol.

3 Semantic model

The semantic model for the PoliMedia project is built to satisfy the requirements of the project, i.e. the research questions from the users. The model is based on the Simple Event Model [5] developed in the NWO CATCH project Agora. SEM is a model to represent events on the Web, and to explicate complicated semantic relations between people, places, actions and objects: not only who did what, when and where, but also the roles each actor played, the time during which this role is valid and the authority according to whom this role is assigned. Because the PoliMedia project deals with a specific domain, our semantic model is adapted to it so it can express important information associated with the events and actors in political debates.

3.1 Requirements for the semantic model

The goal of the project is to publish the links on the Web, so using open Web formats and standards, a Web query language, and unique identifiers (URI's) is compulsory.

The semantic model of the PoliMedia project is to be expressive in a way that it allows important information regarding parliamentary debates to be easily accessed. Important information for every parliamentary debate is:

- The **time** on which the debate is held
- **What** is being said in the debate (**topics**)
- **Who** is giving the speeches in the debate and in which **role** (**persons**)
- Links to **additional information** about actors involved in the event (names of the politicians, their party, age, etc.)
- **Subparts of the debate** have their own identifiers (part of the debate where only one speaker can be identified as actor)
- Important information about subparts is their **chronological order** (the order in which the subparts where occurring inside the parliament debate,
- **Named entities** apart from politicians (persons, locations, etc.)

Important information for parliamentary debates that are specific to PoliMedia project:

- **Links** between subparts of the debate and news articles, radio bulletins and television newscasts
- Various information about media items linked to the debate

Data from the parliamentary debates is available online, so unique identifiers are created for:

- Debates (as a document as a whole) and for the parts of the debates
- Individual news articles, radio bulletins, and television newscasts
- All political parties of the speakers in the debates as well as the speakers them self

All important information about debates listed here are represented in the semantic model.

3.2 URIs as identifiers

On the Semantic Web, all entities are identified by a URI. In our case, all source datasets already contain URIs. Our preference is to use these existing URIs directly instead of creating our own URIs. For example, we link to the newspapers of the *Koninklijke Bibliotheek* with statements like:

<http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19590000048.1.9>
<http://purl.org/linkedpolitics/nl/polivoc:coveredIn>
"http://resolver.kb.nl/resolve?urn=ddd:**010688440**:mpeg21:a0001:ocr" ;

We have made a different choice for the debate events, as these are the core of our dataset. Also for debates, URIs do already exist: the government website *officielebekendmakingen.nl* provides persistent URIs to debates after 1995, and the project War In Parliament provides URIs for debates as well as parts of debates. Nevertheless, we create our own URIs for each debate and parts of debates, for example:

<http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19590000048.1.9>
a <polivoc:Speech> ;

The reason for this is that we want the URIs to be dereferenceable, i.e. we want to serve informative and descriptive RDF when the event URI is requested. Neither *officielebekendmakingen.nl* nor *War in Parliament* does this. We use so-called PURLs (**P**ersistent **U**niform **R**esource **L**ocators), Web addresses that act as permanent identifiers.

3.3 Provenance

We build on existing data and tools. It is important to preserve this provenance information, both to give credit where credit is due and to provide information about how much the data can be trusted. For every debate in our model we add information about the original source of the debate. For example, the next statement uses the dc:source property to state that the original debate came from Political Mashup:

<http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19590000048>
<http://purl.org/dc/elements/1.1/source>
"http://resolver.politicalmashup.nl/nl.proc.sgd.d.19590000048" ;

Named entities were extracted in the War In Parliament project using the NER tool Folia[14]. We use the *dc:provenance* property to state the source of extracted entity.

3.4 Description of the semantic model

The semantic model, as well as the links between datasets, is expressed in the RDF format, W3C Standard for Semantic Web. Also, the data is made compatible with the ISOCAT standard[3], Dublin Core[4] and SKOS[5].

We created this semantic model to conform to the rules and regulations of the Dutch parliament, although the model can be easily adapted to follow different rules (of parliaments in other countries), because in its core all parliamentary debates consists of the same most important elements like the topics and the speeches.

All debates conform to the same rule, where speakers give speeches in the parliament in some chronological order. First speaker is always the "voorzitter" (the person who is in charge of the actual debate and can be called chairman). The chairman gives usually an introduction to the topic and after his speech he gives the floor to some member of the parliament.

Every debate has three main structural elements:

- The topics – the themes or agenda of the meeting
- The speeches – every word by every speaker is transcribed including the names of the speakers and their affiliation
- Actions – descriptions, lists, etc.

Every transcript contains metadata with important information about the debate as a whole, like the date when the debate actually happened in the parliament, the title of the debate etc. In the PoliMedia semantic model, as can be seen in Fig1., a debate is represented as a resource with its own unique identifier (for example: http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002). This resource serves as a domain for Dublin Core properties like *dc:date*, *dc:title*, *dc:identifier*, *dc:publisher*, *dc:source* and *dc:language*, which points to the literals that contain information about the date when the debate happened, its official title, unique identifier and original source, the publisher and the language on which the debate was published (an RDF example is given in Fig. 2).

The PoliMedia specific property *hasPart* is attached to the resource containing the debate URI and points to the range of possible parts of the debate that the debate as a whole can contain (this element is shown in Fig.3). One specific part of a debate always contains elements called *DebateContext* and *Speech*. Element *DebateContext* contains text that is read by the chairman (*voorzitter*) of the debate and that text represents the short description of subjects that will be addressed in the forthcoming speech.

[3] http://www.isocat.org/
[4] http://dublincore.org/
[5] http://www.w3.org/2004/02/skos/

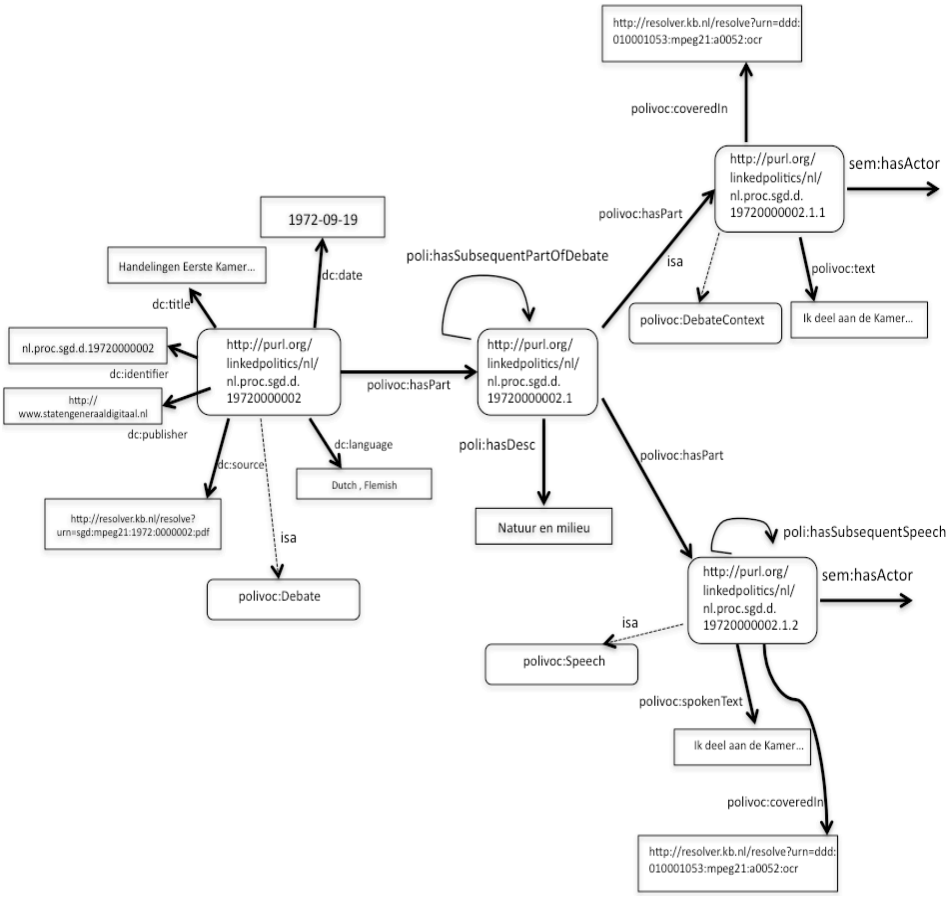

Fig. 1. Part of semantic model representation of the debates dataset (second part on Fig. 4.)

```
<http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002>
      a       <polivoc:Debate> ;
      <http://purl.org/dc/elements/1.1/date>
              "1972-09-19" ;
      <http://purl.org/dc/elements/1.1/identifier>
              "nl.proc.sgd.d.19720000002" ;
      <http://purl.org/dc/elements/1.1/language>
              "Dutch; Flemish" ;
      <http://purl.org/dc/elements/1.1/publisher>
              "http://www.statengeneraaldigitaal.nl" ;
      <http://purl.org/dc/elements/1.1/source>
              "http://resolver.politicalmashup.nl/nl.proc.sgd.d.19720000002" ;
      <http://purl.org/dc/elements/1.1/title>
              "Handelingen Eerste Kamer 1972 19 september 1972, Paginas 3-10." ;
      <http://purl.org/linkedpolitics/nl/polivoc:hasPart>
              <http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.2> ,
              <http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.3> ,
              <http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.4> ,
              <http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.1> .
```

Fig. 2. Debate represented in RD

```
<http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.2>
      a       <polivoc:PartOfDebate> ;
      <http://purl.org/dc/elements/1.1/source>
              "http://resolver.politicalmashup.nl/nl.proc.sgd.d.19720000002.2" ;
      <http://purl.org/linkedpolitics/nl/poli:hasDescription>
              "behandeling van het wetsontwerp Gemeentelijke herindeling van het Land van Heusden en Altena ( 11 284 )." ;
      <http://purl.org/linkedpolitics/nl/polivoc:hasPart>
              "http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.2.4" ,
              "http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.2.3" ,
              "http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.2.5" ,
              "http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.2.6" ,
              "http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.2.2" ,
              "http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.2.1" ;
      <http://purl.org/linkedpolitics/nl/polivoc:hasSubsequentPartOfDebate>
              "http://purl.org/linkedpolitics/nl/nl.proc.sgd.d.19720000002.3" .
```

Fig. 3. Part of the debate represented in RDF

The most important element of the PoliMedia semantic model is the element *Speech* that represents the actual speech that a certain member of Parliament has spoken while addressing the issues of the debate topic (Fig.4 and Fig.5). The content of the speech is saved as a *Literal*. Every speech has its speaker and those two resources are connected with the *sem:hasActor* property described in the Simple Event Model[5]. Property *hasActor* points to the blank node with three other properties leaving from the node. Objects of those properties are URIs that lead to the pages of the politician giving the speech, to the party the mentioned politician is member of, and SEM properties denoting the role of the *hasActor* property.

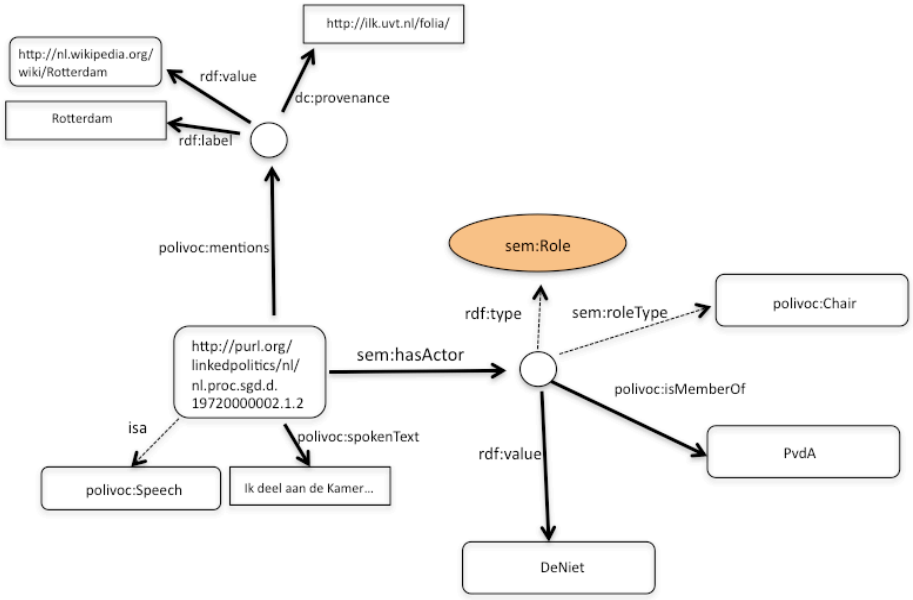

Fig. 4. Semantic model representation of the debates dataset

Fig. 5. Example of one speech in RDF

By nature, speeches in the parliament usually contain a great number of named entities, such as names of politicians or business people, names of different organizations, and geographical locations. Named entities were recognized in parliamentary debates in the project War in Parliament. Names of persons, organizations, locations, and miscellaneous entities were extracted from transcripts using a tool for Linguistic Annotation[13]. Named entities are connected with four different properties where each one points to different objects of the triple (either person, location, organization or miscellaneous entity). A literal is created for every named entity found in the speech together with a URI that leads to the Wikipedia page of the entity, in case that page exists.

The semantic model for secondary datasets is straightforward. Both SRU and OAI-PMH protocols allow the client to submit a search and retrieve request for matching records from the secondary datasets. A response on a query containing the matched keywords contains Dublin core properties such as *dc:identifier*, *dc:type*, *dc:publisher*, *dc:date*, *dc:source* and *dc:title* which are used in our PoliMedia semantic model in case of newspaper articles. The model will contain the instance of a newspaper article with a URI that uses a resolver for accessing the OCR text or pdf document at the National Library. Both radio bulletins and newscast datasets have very similar models. The newscast dataset contains very rich metadata about its resources, so except information about the date, type and publisher, this metadata contains spatial information and names of subject that appears in the videos.

As a final result of the first phase of our project, we created an RDF repository that contains around 38,8 million triples, that came from 10,924 XML files containing information about debates in Dutch parliament. Important elements from XML files were extracted using Java libraries (SAX) and RDF triples were created (JENA). The semantic repository is created using OWLIM[6], a software component for storing and manipulating huge quantities of RDF data. OWLIM is packaged as a Storage and Inference Layer (SAIL) for the Sesame OpenRDF framework.

[6] OWLIM – Semantic repository: http://owlim.ontotext.com/display/OWLIMv51/Home

4 Summary and next steps

In this paper we described the process of creating the semantic model for the purpose of building a semantic repository for the PoliMedia project. The semantic repository is filled with triples that describe events and topics that happened in the Dutch parliament and allows us to use queries to fetch interesting information that was not as easily available before (for example, how many times a particular politician spoke of a particular person in the parliament).
As previously stated, the PoliMedia project will be carried out in three phases. Phases that will be carried in the future are phases (2) and (3), with an automatic detection of the semantic links between primary and secondary datasets and the creation of a demonstrator application.
For the creation of links Named Entities (that appears in primary and secondary datasets) will be used to decide whether the media resource is on some way connected to the events discussed in the debates. Important entities are persons but also locations and time. As debate events consist of smaller sub-events, namely speeches of consecutive speakers (as it is expressed in the semantic model described in this paper), we will search for possible links between those sub-events and media items that cover that particular part of the debate. A virtual research environment will be built that allows the exploration of the debate events and media coverage thereof via search and browsing. Next to the use of standard information retrieval libraries (Lucene), navigation options will be implemented that will allow users to browse through the linked datasets of debates and media.

References

1. Maarten Marx: Advanced Information Access to Parliamentary Debates. J. Digit. Inf. 10(6): (2009)
2. War In Parliament project: http://www.clarin.nl/page/about/projects/162#WIP
3. Political Mashup project: http://politicalmashup.nl/
4. van Erp, Marieke et al, :Automatic Heritage Metadata Enrichment with Historic Events. In J. Trant and D. Bearman (eds). *Museums and the Web 2011: Proceedings*. Toronto: Archives & Museum Informatics. Published March 31, 2011. Consulted March 5, (2012).
5. W. van Hage, V. Malaisé, R. Segers, L. Hollink, and G. Schreiber: Design and use of the Simple Event Model (SEM). J. Web Semantics, 2011.
6. Hage, W.R. van, V. Malaisé, G. de Vries, G. Schreiber and M. van Someren: "Combining Ship Trajectories and Semantics with the Simple Event Model (SEM)". In: Proceedings of the 1 stACM International Workshop on Events in Multimedia 73-80, (2009)
7. Dalila Mekhaldi, Denis Lalanne: Multimodal Document Alignment: Feature-based Validation to Strengthen Thematic Links. JMPT 1(1): 30-46, (2010)
8. Y. Lv, T. Moon, P. Kolari, Z. Zheng, X. Wang, and Y. Chang: Learning to model relatedness for news recommendation. In *WWW*, (2011).
9. SRU: Search/Retrieval via URL: http://www.loc.gov/standards/sru/
10. OAI protocol: http://www.openarchives.org/OAI/openarchivesprotocol.html
11. Koninklijke Bibliotheek metadataopslag:http://research.kb.nl/documenten.html
12. Digital Item Declaration Language: http://xml.coverpages.org/mpeg21-didl.html
13. FoLiA: Format for Linguistic Annotation: http://ilk.uvt.nl/folia/