The MetaLex Document Server

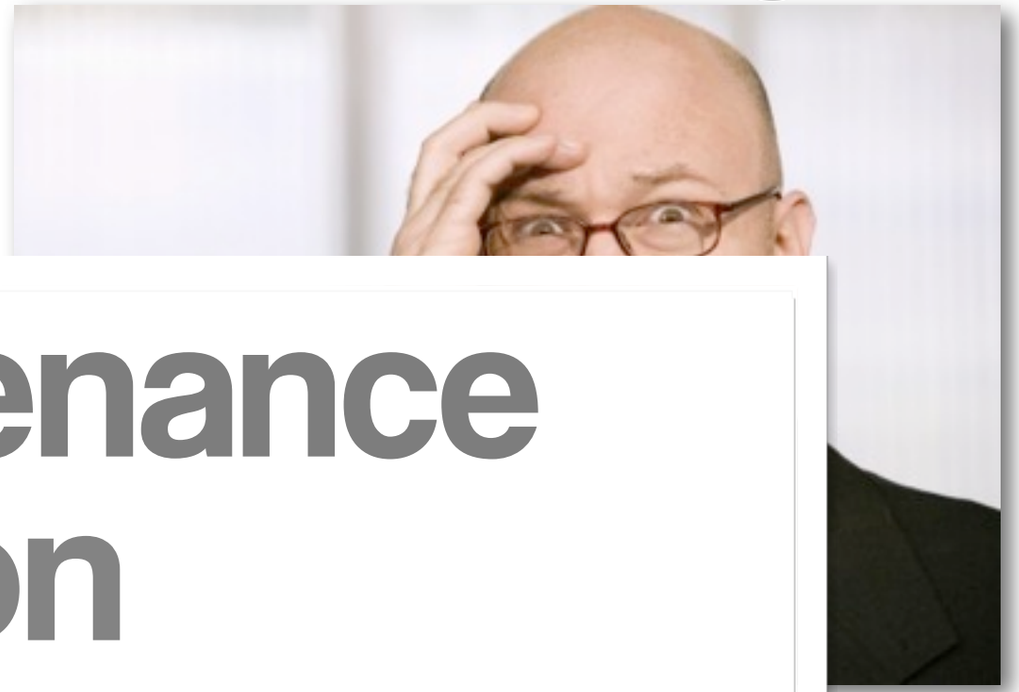
Legal Documents as Versioned Linked Data

Rinke Hoekstra
Universiteit van Amsterdam & VU University Amsterdam

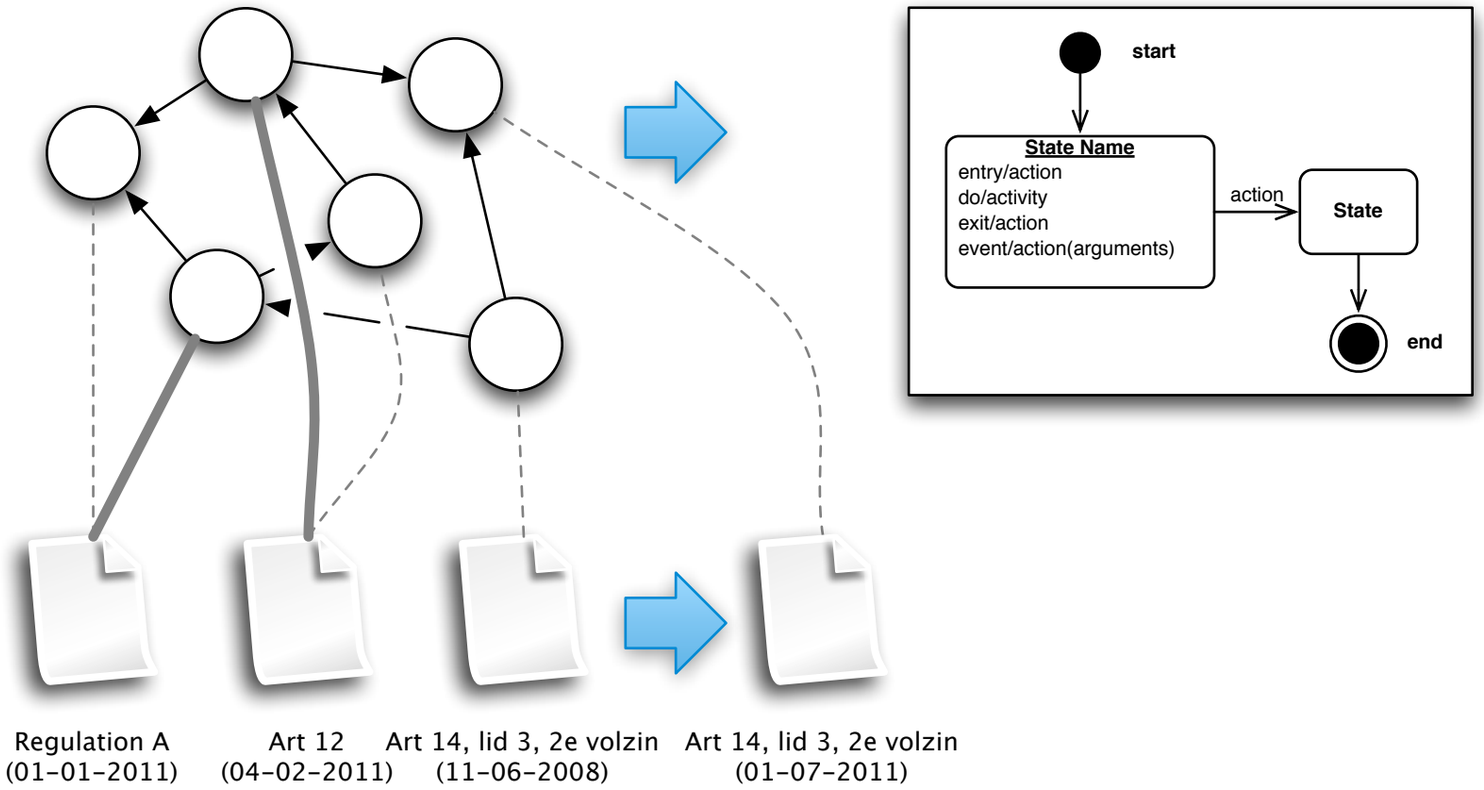
Um... why is this interesting?

Um... why is this interesting?

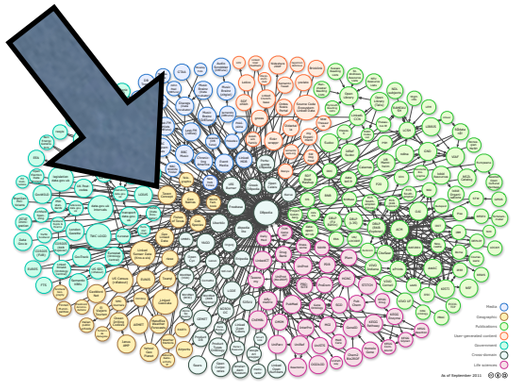
Um... why is this interesting?



provenance
annotation
import interpretation
deemingprovision scope
integration
versioning
compliance importance



- **Open Data**: current public service falls short
- **Hidden agenda**:
  - Large scale **validation** of CEN MetaLex
  - Linked Open Government Data



CEN MetaLex

"Open XML Interchange Format for Legal and Legislative Resources"

- CEN Workshop Agreement
- Interchange format
- Highly generic XML elements
  (hcontainer, block, inline)
- "Content models" signal content
  (e.g. chapter, article, sentence)
- Schema extension
- Metadata as RDFa
- Naming convention

http://www.metalex.eu

Current Situation

Public content services hosted at wetten.nl

Wetten.nl XML Service

http://wetten.overheid.nl/xml.php?regelingID=...

- Only available format is **BWB XML**
- Only **current** version
- Content at **document level**
- Identification at **document level**
- Identifiers are **not dereferencable**
- Hardly any **metadata** (e.g. version date)
- Only available **context** is position in text

BWBId Web Service

http://wetten.overheid.nl/BWBIdService/BWBIdList.xml.zip

```
    <__NS1:InwerkingtredingsDatum>2003-04-01</__NS1:InwerkingtredingsDatum>
</__NS1:Citeertitel>
</__NS1:CiteertitelLijst>
    <__NS1:NietOfficieleTitelLijst></__NS1:NietOfficieleTitelLijst>
    <__NS1:AfkortingLijst>      <__NS1:Afkorting>HRWN</__NS1:Afkorting>
</__NS1:AfkortingLijst>
      <__NS1:RegelingSoort>circulaire</__NS1:RegelingSoort>
    </__NS1:RegelingInfo>
</__NS1:RegelingInfoLijst>
</__NS1:BWBIdServiceResultaat><?xml version="1.0" encoding="UTF-8"?>
<__NS1:BWBIdServiceResultaat xmlns:__NS1="http://schemas.overheid.nl/bwbidservice">
  <__NS1:GegenereerdOp>2011-05-31</__NS1:GegenereerdOp>
  <__NS1:RegelingInfoLijst>
    <__NS1:RegelingInfo>
      <__NS1:BWBId>BWBR0001821</__NS1:BWBId>
    <__NS1:DatumLaatsteWijziging>2009-01-08</__NS1:DatumLaatsteWijziging>
    <__NS1:VervalDatum>2003-01-01</__NS1:VervalDatum>
      <__NS1:OfficieleTitel>Wet van 21 april 1810, Bulletin des Lois 285</__NS1:OfficieleTitel>
    <__NS1:CiteertitelLijst>    <__NS1:Citeertitel>      <__NS1:titel>Loi concernant les Mines, l
      <__NS1:status>officieel</__NS1:status>
```

NB: The problem with the XML processing instruction was reported and fixed, but returned some weeks later

Identifiers & Juriconnect

1.0:c:BWBR0005416&artikel=6

vs

http://wetten.overheid.nl/cgi-bin/deeplink/law1/bwbid=BWBR0005416/*article*=6/date=2005-01-14

vs

http://wetten.overheid.nl/BWBR0005416/TitelII*698946*/HoofdstukII/Artikel16/geldigheidsdatum_14-01-2005

- Existing identification standard: **Juriconnect**
- URN-like... but no naming server
  cf. Document Object Identifiers
- Named elements do not **carry** identifier
- No **explicit** version information, only **contextual**

Step 1

Requirements

Goals

- "**Deserialize**" regulation content
  (e.g. topic-based browsing)
- **Extract** and **reconstruct** implicit information
  (identifiers, metadata)
- **Annotate** regulations
  (reconstructed metadata, third-party metadata)
- Annotate **using** regulations
  (knowledge based systems, services, business processes ...)
- **Accessible** and **reusable** for any other party
  (shared vocabularies, standard access)

Requirements

- Unique, persistent identification
  (URL-like URIs)
- Generic XML structure of documents
  (CEN MetaLex XML documents)
- Extensible metadata framework
  (Linked Open Data)
- Flexible web services
  (Transparent REST services, Cool URIs)

Available Sources...

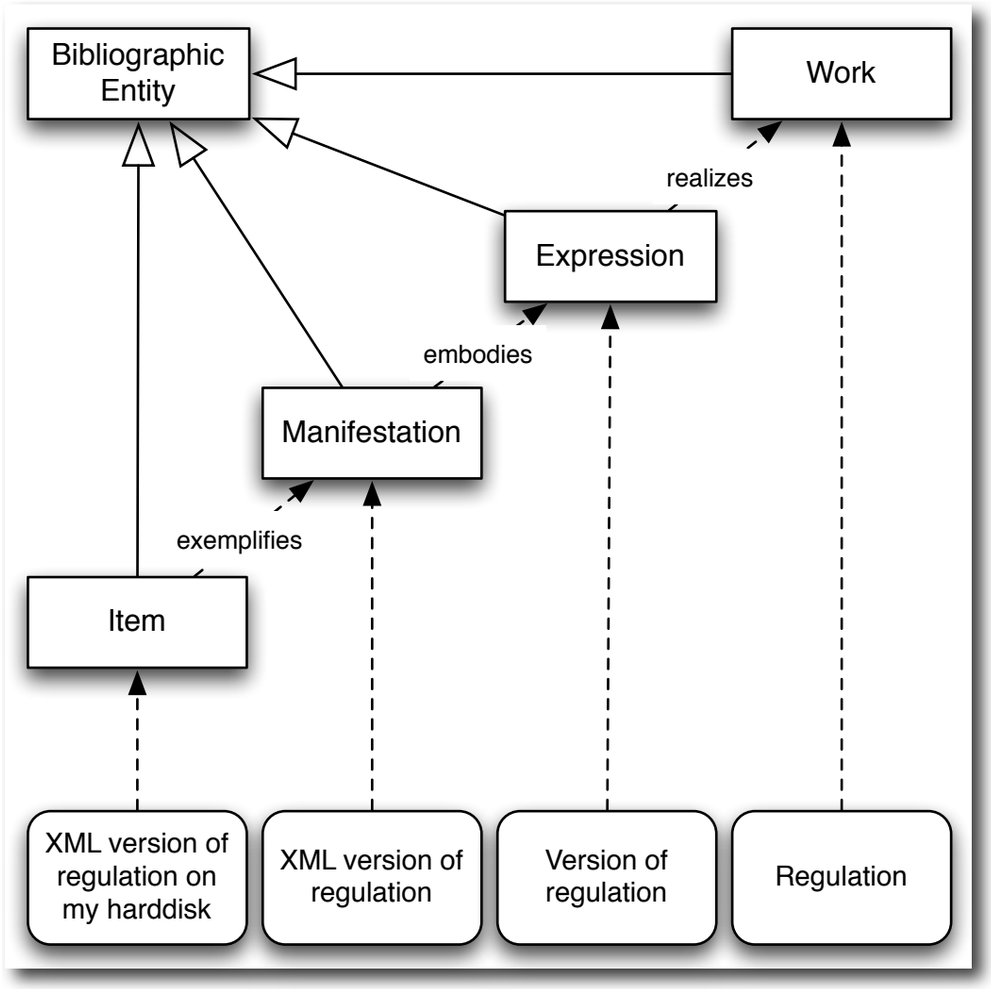
- List of all regulations in “XML”
- Wetten.nl XML Service
- Metadata in HTML table on wetten.nl
  (the “info page”)
- ... so let’s get started already

Step 2

Come up with persistent identifiers at element level and a solid versioning scheme

Levels of Identification

- IFLA FRBR levels
- Work
- Expression
- Manifestation



Transparent Identifiers

- **Hierarchical** information (work)

  http://doc.metalex.eu/id/BWBR0011823/hoofdstuk/1/artikel/1

  http://doc.metalex.eu/id/BWBR0011823/artikel/1

- **Version** and **language** (expression)

  http://doc.metalex.eu/id/BWBR0011823/hoofdstuk/1/artikel/1/**nl/2010-09-01**

- **Format** information (manifestation)

  http://doc.metalex.eu/**doc**/BWBR0011823/hoofdstuk/1/artikel/1/nl/2010-09-01/**data.xml**

Problem

- URIs don't carry semantics...
- Detect changes:
  - which *element* versions are **the same**
  - ... and which versions are **different**?

Indien de opvolger rechtspersoon is, gaat de kennisgeving gepaard met de statuten van deze rechtspersoon, een uittreksel uit diens inschrijving in het handelsregister en een bewijs van aanstelling van de natuurlijke persoon die is aangesteld als vertegenwoordiger van de vertegenwoordiger die rechtspersoon is.

Art. 44, lid 4
(2011-03-26)

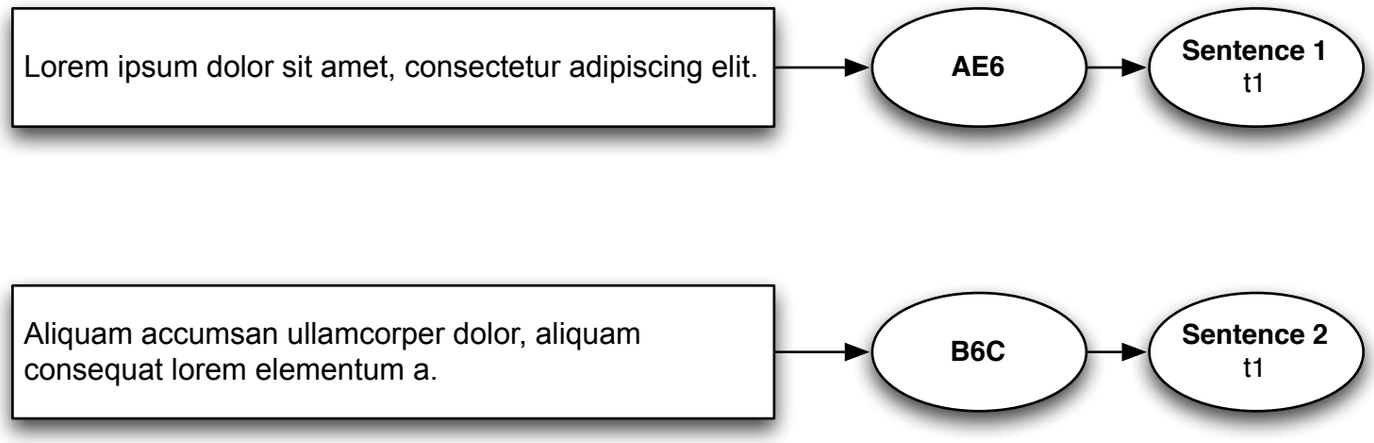
Art. 44, lid 4
(2011-04-05)

Indien de opvolger rechtspersoon is, gaat de kennisgeving gepaard met de statuten van deze rechtspersoon, een opgave van het nummer van inschrijving in het handelsregister en een bewijs van aanstelling van de natuurlijke persoon die is aangesteld als vertegenwoordiger van de vertegenwoordiger die rechtspersoon is.

from: *Besluit prudentiële regels Wft, BWBR0020420*

Opaque Identifiers

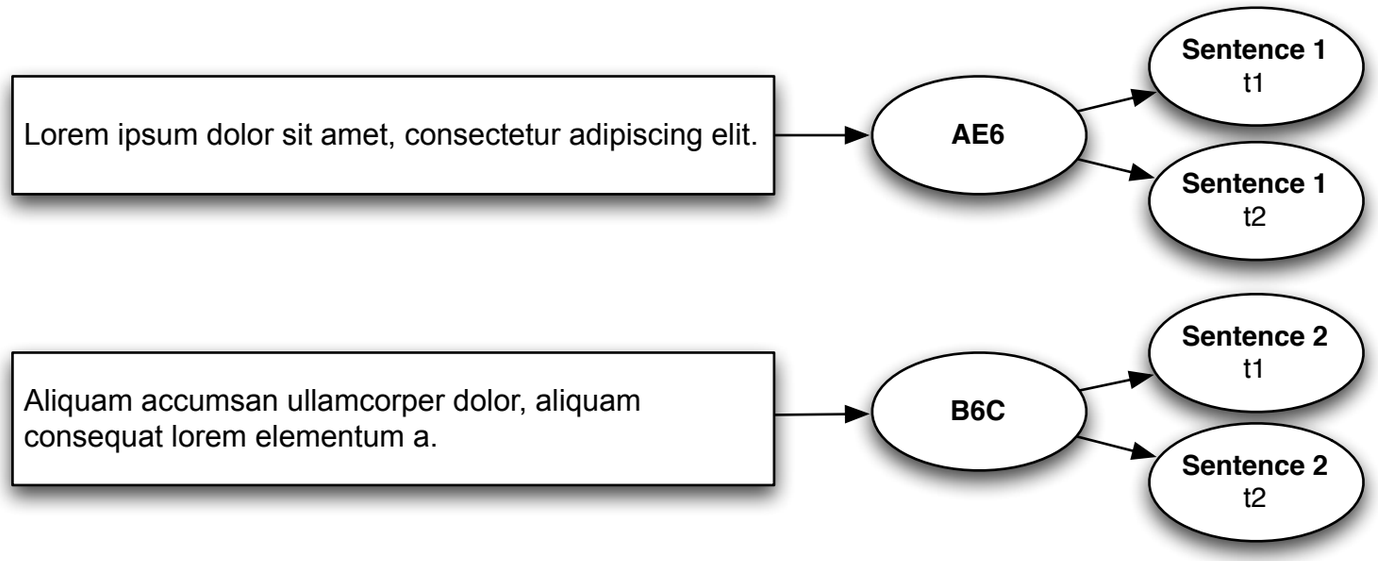
http://doc.metalex.eu/BWBR0011823/hoofdstuk/1/artikel/34b0cee26ee5138c74aa2c62caf2c117d3c616e9



- **Content** information
- Unique SHA1 Hash of **text**

Opaque Identifiers

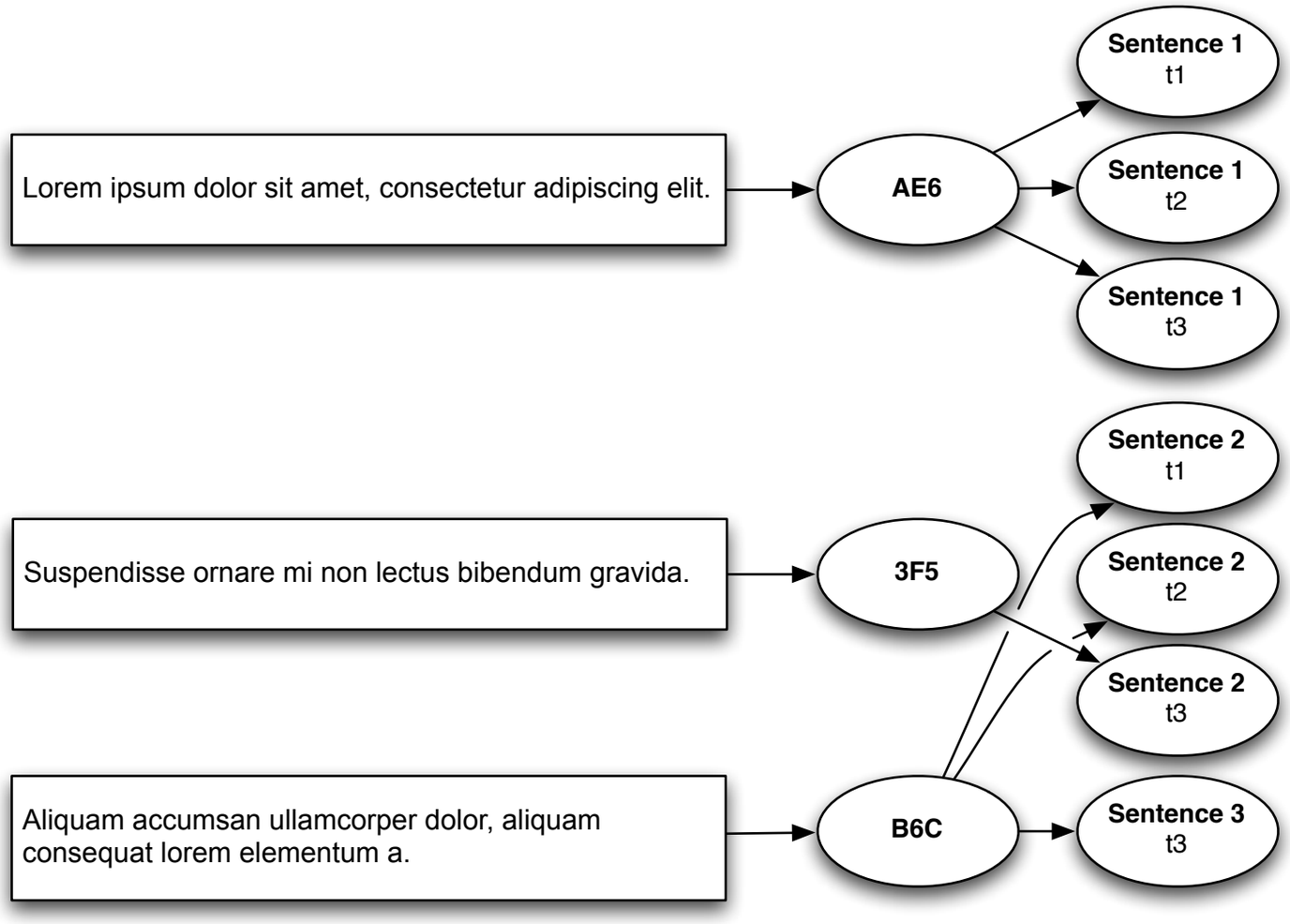
http://doc.metalex.eu/BWBR0011823/hoofdstuk/1/artikel/34b0cee26ee5138c74aa2c62caf2c117d3c616e9



- **Content** information
- Unique SHA1 Hash of **text**

Opaque Identifiers

http://doc.metalex.eu/BWBR0011823/hoofdstuk/1/artikel/34b0cee26ee5138c74aa2c62caf2c117d3c616e9



- **Content** information
- Unique SHA1 Hash of **text**

Step 3

Generic conversion of BWB XML to a generic XML format (CEN MetaLex) and appropriate metadata

Procedure

For **each** BWB XML file listed,
if **update** has occurred since latest run,
download latest version,
scrape metadata, and
produce:

Persistent URIs

CEN MetaLex + Citations

Inline RDFa *(optional)* or RDF graph *(optional)*,
Pajek ".net" files *(optional)*

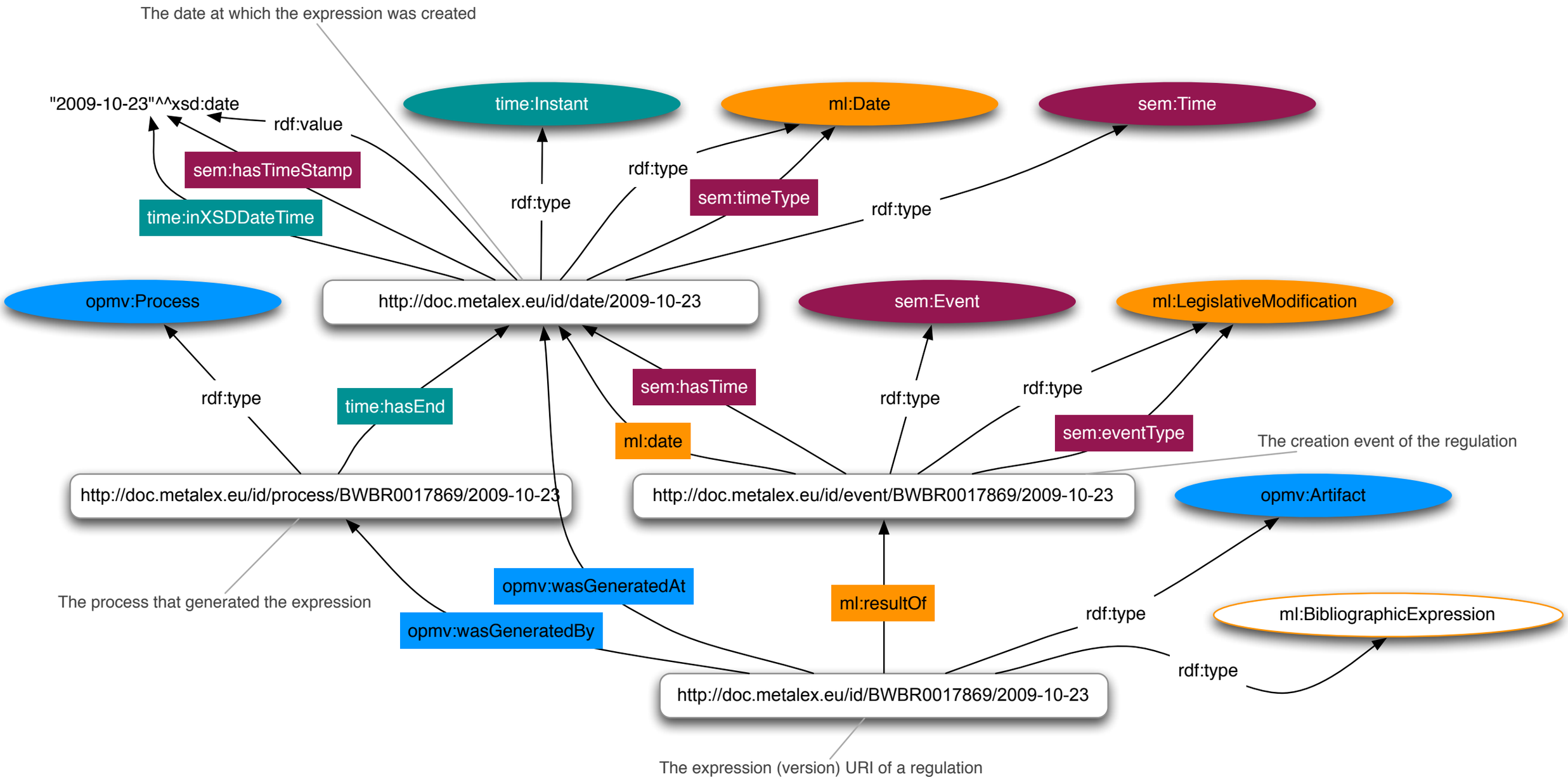
BWB to CEN MetaLex?

- Straightforward 1:1 mapping
  - ... some minor fixes

Table 1. Conversion performance for 300 randomly selected regulations.

	Number	%		Number	%
Substitutions[42]			**Corrections**		
container	22312	29 %	artikel	2525	72 %
hcontainer	3730	5 %	divisie	519	15 %
htitle	3730	5 %	colspec	289	8 %
block	34325	44 %	illustratie	54	2 %
inline	13527	17 %	*others*	99	3 %
Total	77624		*Total*	3486	
			Total no. of regulations	300	
			Revoked regulations	109	30 %
			Correction %		4 %

Events & Provenance



http://doc.metalex.eu/id/BWBR0020486/2009-01-01

http://www.w3.org/1999/02/22-rdf-syntax-ns#type	• http://purl.org/net/opmv/ns#Artifact • http://www.metalex.eu/schema/1.0#BibliographicExpression
page	• http://doc.metalex.eu/doc/BWBR0020486/2009-01-01/data.xml
http://purl.org/dc/terms/title	• Besluit toezicht financiële verslaggeving
http://doc.metalex.eu/bwb/ontology/bwb-id	• BWBR0020486
http://xmlns.com/foaf/0.1/homePage	• http://doc.metalex.eu/doc/BWBR0020486/2009-01-01/data.html
isDefinedBy	• http://doc.metalex.eu/doc/BWBR0020486/2009-01-01/data.rdf
http://doc.metalex.eu/bwb/ontology/dtdversie	• 2.0
http://doc.metalex.eu/bwb/ontology/id	• 363109
http://purl.org/net/opmv/ns#wasGeneratedBy	• http://doc.metalex.eu/id/process/BWBR0020486/2009-01-01
http://www.metalex.eu/schema/1.0#resultOf	• http://doc.metalex.eu/id/event/BWBR0020486/2009-01-01
http://purl.org/net/opmv/ns#wasGeneratedAt	• http://doc.metalex.eu/id/date/2009-01-01
http://doc.metalex.eu/bwb/ontology/soort	• AMvB
http://www.metalex.eu/schema/1.0#realizes	• http://doc.metalex.eu/id/BWBR0020486
http://purl.org/dc/terms/valid	• 2009-01-01
http://purl.org/dc/terms/source	• http://wetten.overheid.nl/xml.php?regelingID=BWBR0020486
http://purl.org/dc/terms/alternative	• Btfv
is http://www.metalex.eu/schema/1.0#partOf of	• http://doc.metalex.eu/id/BWBR0020486/wet-besluit/1/2009-01-01 • http://doc.metalex.eu/id/BWBR0020486/intitule/1/2009-01-01
is http://www.metalex.eu/schema/1.0#result of	• http://doc.metalex.eu/id/event/BWBR0020486/2009-01-01

Dublin Core
OPMV
SEM
W3C Time
MetaLex
FOAF

Step 4

Publish: The MetaLex Document Server (MDS)

The MetaLex Document Server Data Search

119,307,040 triples

This service hosts (almost) all Dutch national regulations in CEN MetaLex XML and as RDF Linked Data.
Current coverage is **29,120** document versions, covering practically all regulations available through http://wetten.overheid.nl

Search »

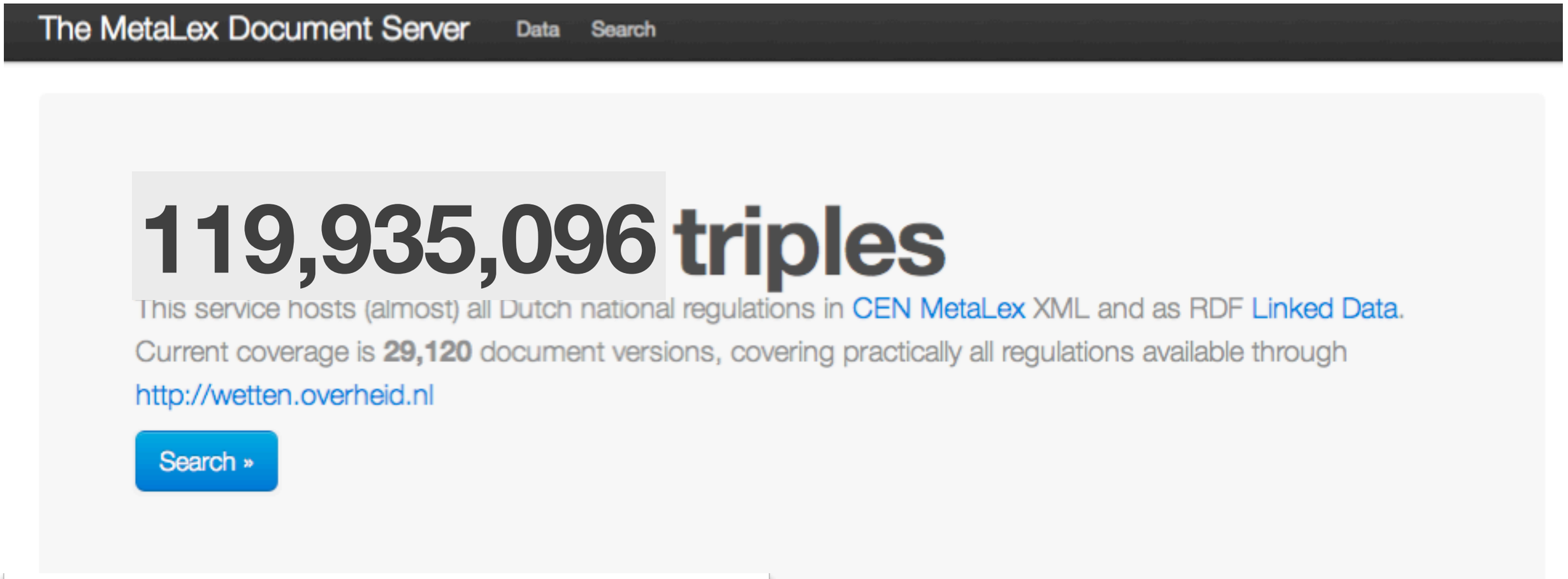
- RESTful API
  - Cool URIs
    (Dereference to XML, RDF, .net)
  - Shorthands (‘/latest’)
  - SPARQL endpoint
  - Citation graphs
- Whoosh-based search
- CSS Stylesheet for CEN MetaLex

MetaLex and Linked Data

CEN MetaLex is a standard for how sources of law and references to sources of law are to be represented in XML. It is an interchange format, a lowest common denominator for other standards, intended not to replace jurisdiction-specific standards and vendor-specific formats in the publications process but to impose a standardized view on legal documents for the purposes of information exchange and interoperability in the context of software development.

Linked Data is a W3C sanctioned approach to publishing metadata on the Web using a set of standard languages and metadata vocabularies.

MetaLex » Linked Data »



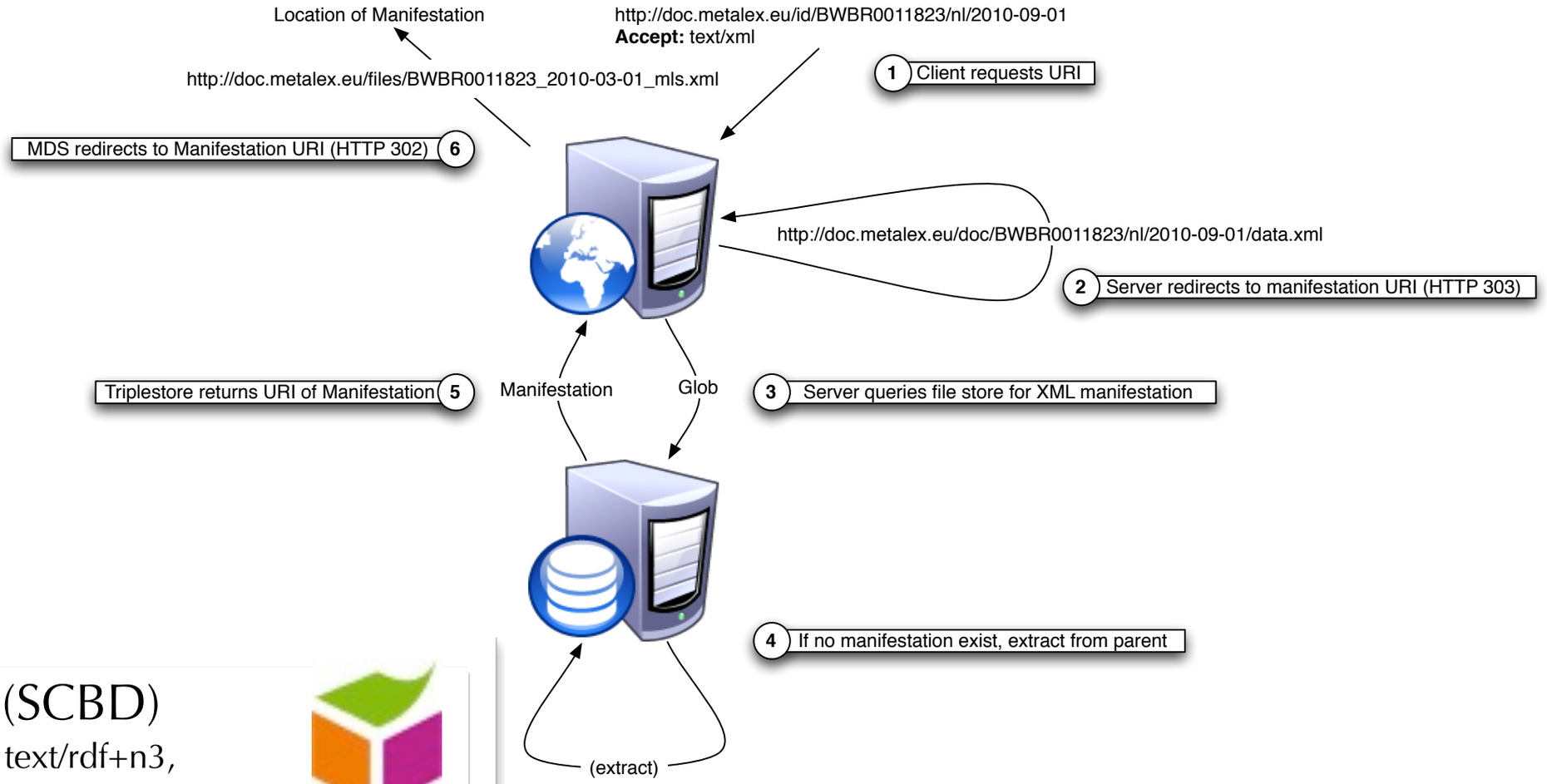
- RESTful API
  - Cool URIs
    (Dereference to XML, RDF, .net)
  - Shorthands (‘/latest’)
  - SPARQL endpoint
  - Citation graphs
- Whoosh-based search
- CSS Stylesheet for CEN MetaLex

MetaLex and Linked Data

CEN MetaLex is a standard for how sources of law and references to sources of law are to be represented in XML. It is an interchange format, a lowest common denominator for other standards, intended not to replace jurisdiction-specific standards and vendor-specific formats in the publications process but to impose a standardized view on legal documents for the purposes of information exchange and interoperability in the context of software development.

Linked Data is a W3C sanctioned approach to publishing metadata on the Web using a set of standard languages and metadata vocabularies.

MetaLex » Linked Data »



- RDF syntaxes (SCBD)
  application/rdf+xml, text/rdf+n3, application/x-turtle
- XML documents
  text/xml
- HTML clients
  application/xml, application/xhtml+xml, text/html
- Pajek clients
  text/plain
  - Download .net file
  - View using Gephi Toolkit

http://gephi.org

Technical Details

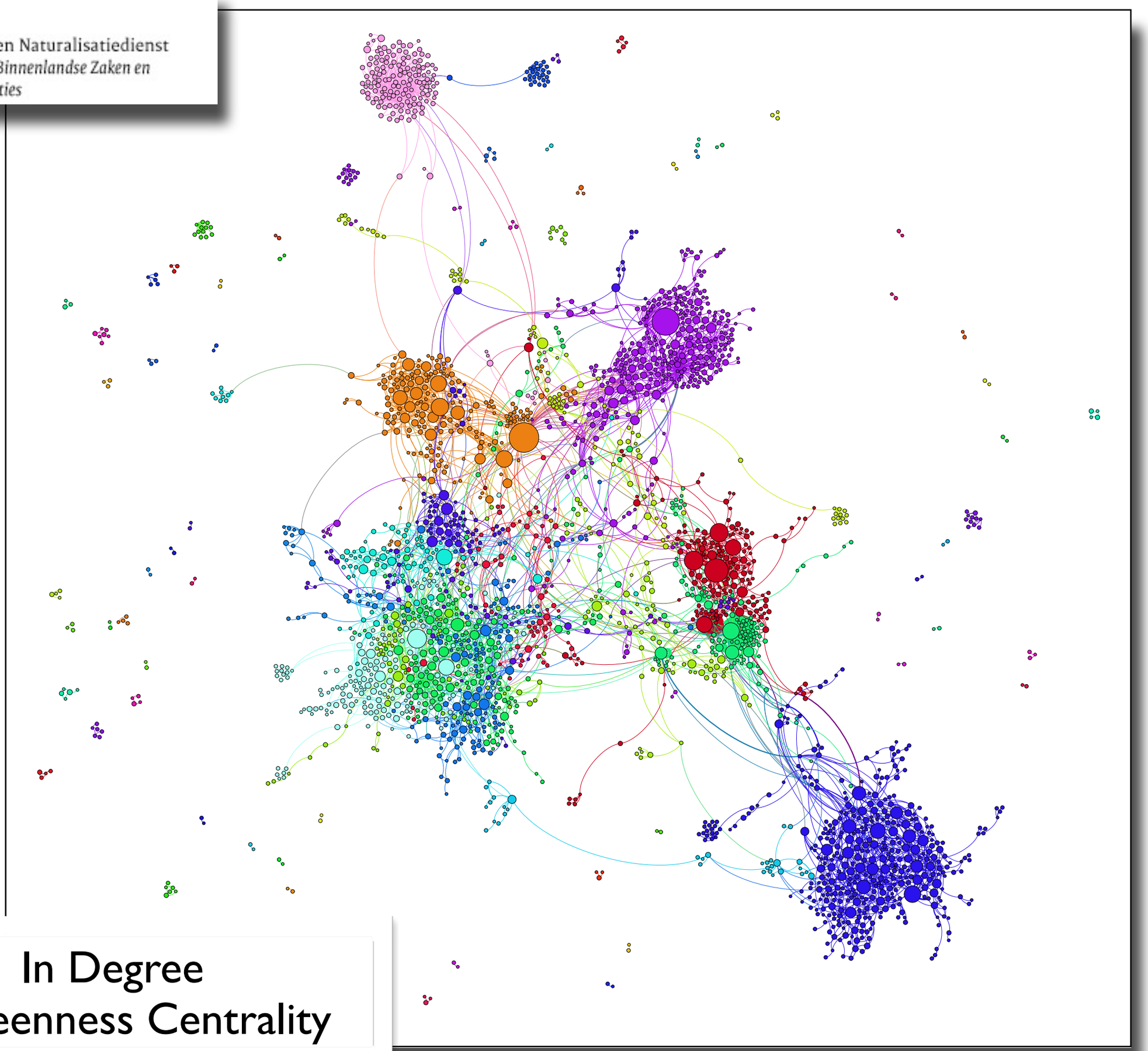
- Current situation
  - +/- 29 **thousand incremental** regulation versions
  - 119,3 **million** triples (legislation.gov.uk: 1.9 billion)
  - Updated **daily**
- Technical details
  - Dell PowerEdge II T110, 32GB RAM
  - Garlik 4Store triplestore (http://4store.org)
  - Python Django web applications
  - Tomcat servlet + Gephi Toolkit API

- See http://doc.metalex.eu

Step 5

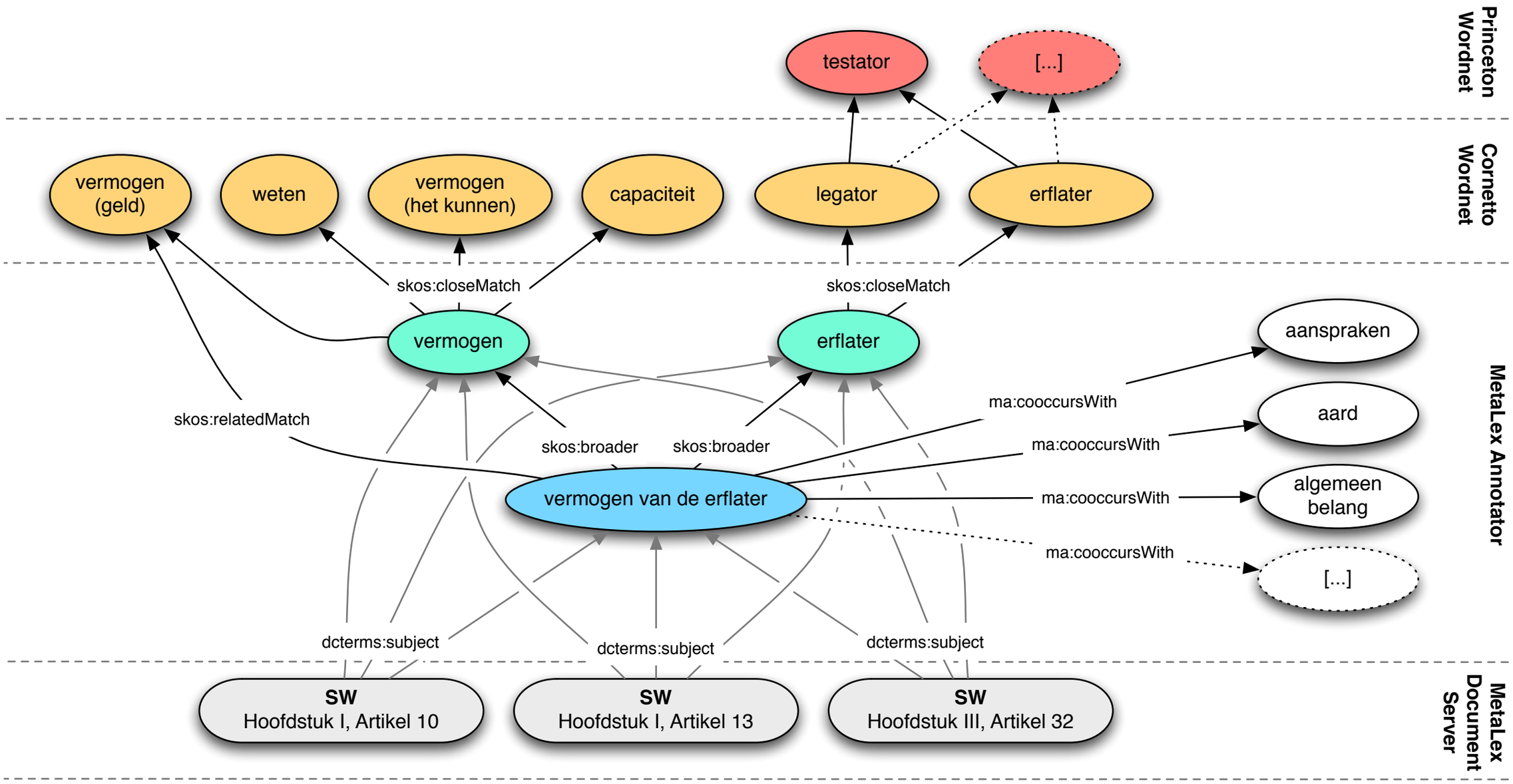
Use it

Immigratie- en Naturalisatiedienst
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties



In Degree Betweenness Centrality


Belastingdienst
Leuker kunnen we 't niet maken. Wel makkelijker.



>500k triples for Inheritance Tax law

Discussion

- Linked Open Data is not just for citizens
- Content service of Dutch gov. falls short
- Successful transformation to CEN MetaLex
- Successful transformation to RDF
- Dual versioning at element level

- Extend to other (international) XML formats
- Empirical study of network analysis

Fin

http://doc.metalex.eu

http://github.com/RinkeHoekstra