



A Text Analysis Framework for Automatic Semantic Knowledge Representation of Legal Documents

Ákos SZŐKE, Krisztián MÁCSÁR, György STRAUSZ
Budapest University of Technology and Economics
Multilogic Ltd, Budapest

ICAIL 2013 – NAIL Workshop
14 June 2013
Rome, Italy

Motivation

- Assumption:
 - The electronic representation of legal texts exist
- Problems:
 - How can we
 - exploit data discovery to answer specific legal questions?
 - assist information access by discovering latent topics / trends?
 - assist information organization to discover hidden structures / outliers?



Applying text mining approach:
Semi-automated knowledge extraction from
unstructured data sources to build knowledge base

Our Goals

- Extracting information and build knowledge base from unstructured regulations using
 - Metadata (such as title, authority)
 - Concepts based on controlled vocabularies (such as EUROVOC: <http://eurovoc.europa.eu/>)
 - Citations
 - Versions (such as validity, language versions)
- To provide
 - Semantic search possible
 - Understanding the law more deeply

Our overall goal

Legal text

1990. évi XCIII. törvény az illetékekről

§ 16.

(1) Mentés az öröklési illeték alól:

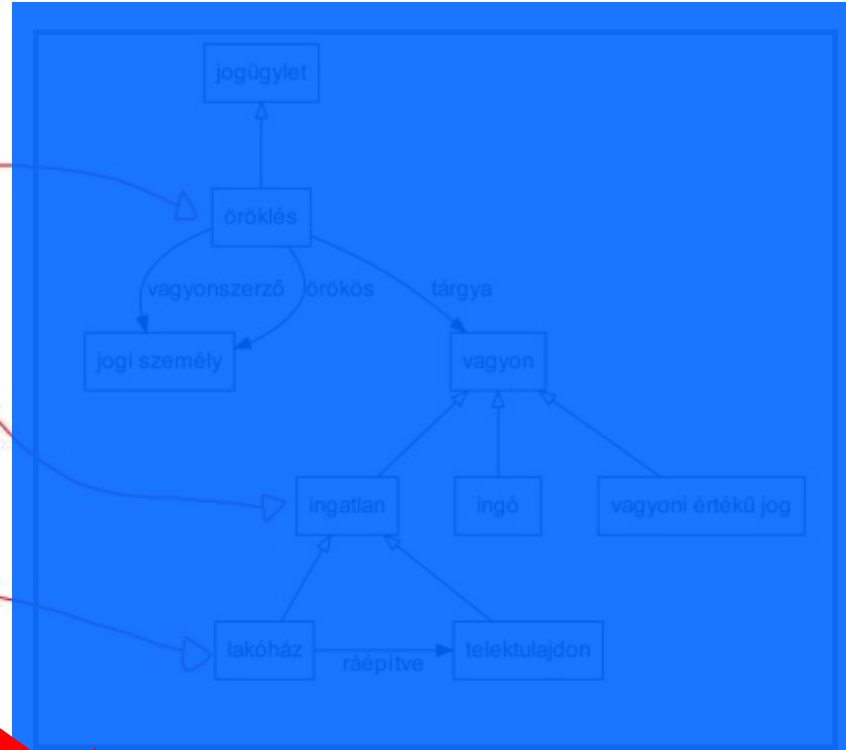
[...]

(g) a lakóház építésére alkalmas telektulajdonnak (tulajdoni hasznadnak), valamint az ilyen ingatlanra vonatkozó vagyoni értékű jognak az öröklése, ha az örökös az örökölő ingatlanon a hagyaték jogerős átadásától számított 4 éven belül lakóházat épít, és a felépített lakóházban a lakást(ök) hasznos alapterülete eléri a településrendezési tervben meghatározott maximális beépíthetőség legalább 10%-át. [...]

(2) Az (1) bekezdés g) pontjában említett lakóház felépítésének igazolása érdekében az ott meghatározott 4 éves határidő elteltét követően, azonban belül az állami adóhatóság megkeresi az ingatlan fekvésére illetékes hatóságban az építési hatóság igazolása szerint a lakóházra a vagyonszerző nevére a használatbavételi engedély kiadása megtörtént, az állami adóhatóság a megállapított, de a megfizetés tekintetében felfüggesztett illetéket törli. Törli az állami adóhatóság az illetéket akkor is, ha a 4 éves határidőn belül a vagyonszerző a nevére szóló jogerős használatbavételi engedéllyel igazolja a lakóház felépítését. [...]

Regulation

Relations



Metadata

- Title
- Validity
- Authority
- Language
- Etc.

Rules



Concepts

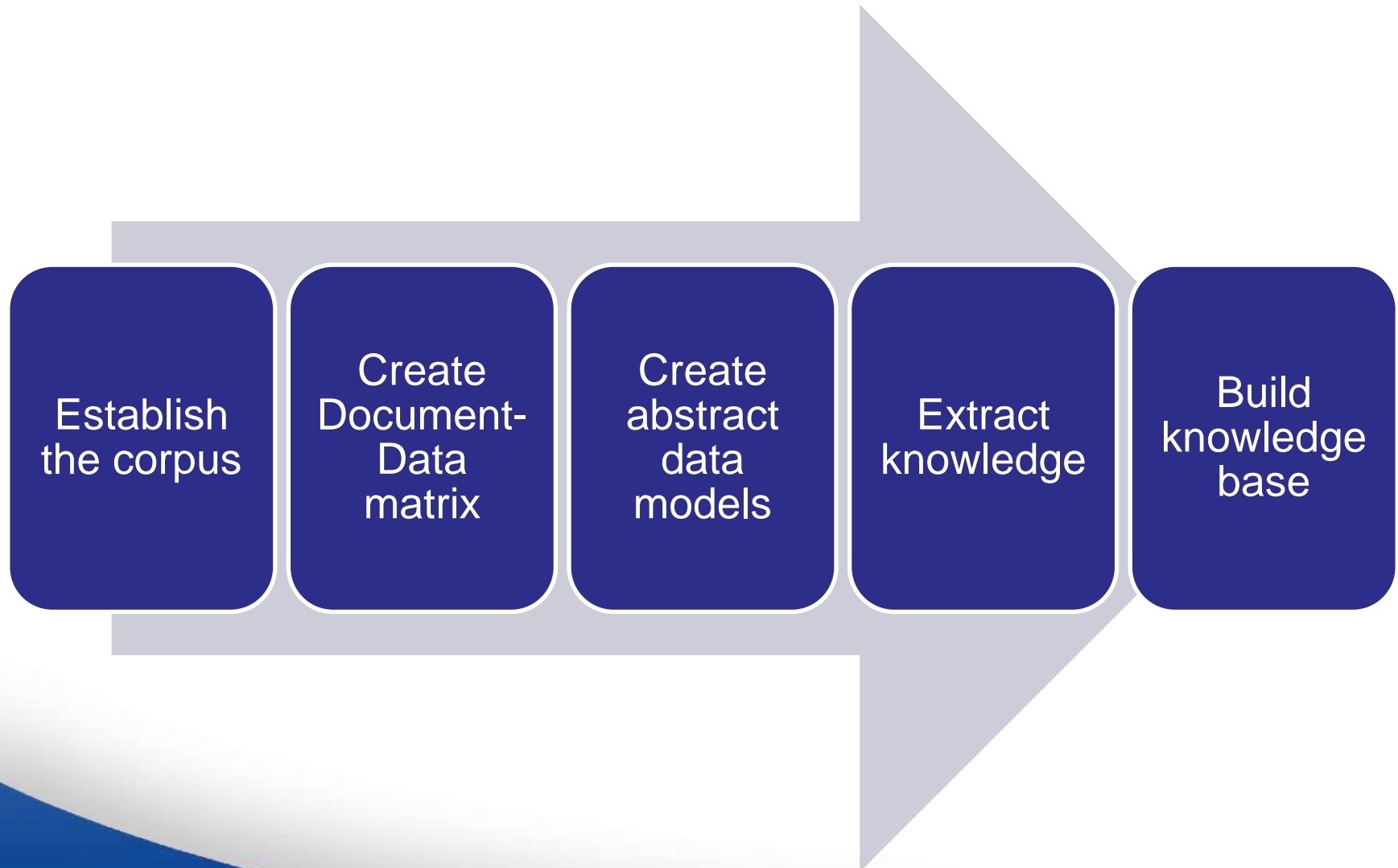
Formal model

Agenda

- Our Methodology
- Architecture of the Framework
- Evaluation and Future Work

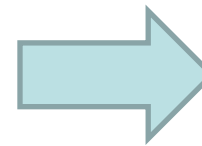
Our Process of Extraction

The Process of Extraction

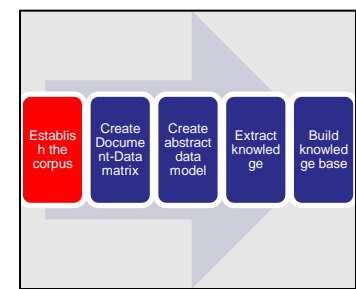


Step 1: Establish the corpus

- Collect all relevant unstructured data
 - Web pages (HTML documents)
- Standardize the collection
 - XML format → CEN MetaLex compliant
- Place the collection in a common place



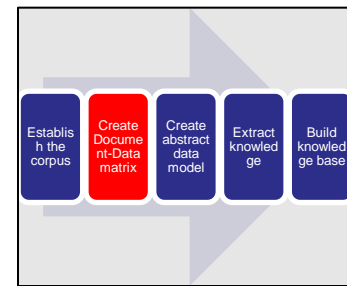
Reflects URI



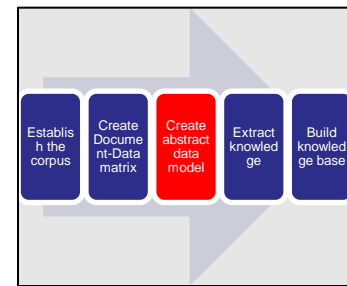
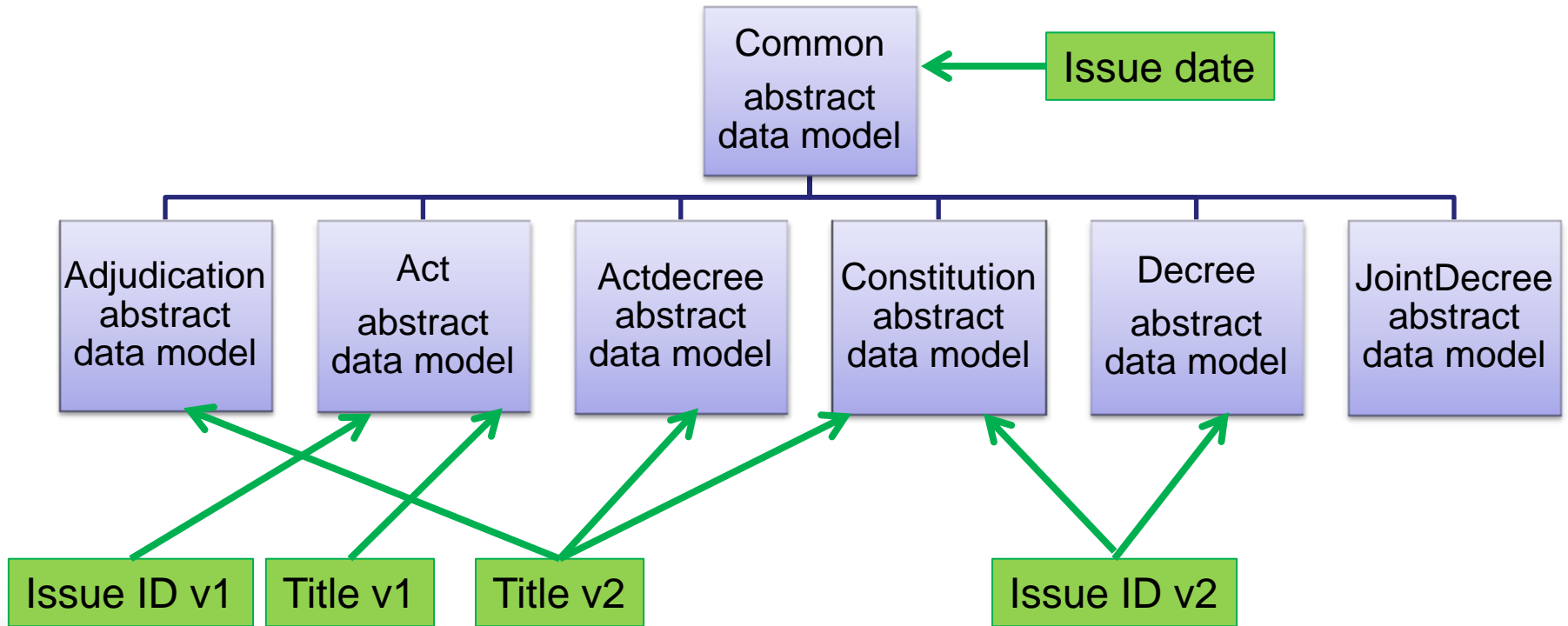
Step 2: Create Document-Data matrix

- Determine the extractable data
- Determine their relations with document types

		Data					
		Title v1	Title v2	Issue ID v1	Issue ID v2	Issue date	...
Document types	Adjudication		X		X	X	
	Act	X		X		X	
	Actdecree		X		X	X	
	Constitution		X			X	
	Decree		X		X	X	
	JointDecree		X		X	X	



Step 3: Create abstract data models



Step 4: Extract knowledge

- Determine the patterns of extraction
 - Clustering the information
 - Improve search recall
 - Improve search precision

Example: Clustering the information

Different forms of one citation style:
(IssueYear, IssueID, WorkType, Anum, Pnum, Cnum)

Issue year

Issue ID

Paragraph number

Clause number

1995.	évi	LIII.	törvény	56.	§	(4)	a)	pontja
1995.	évi	LIII.	törvény	56.	§	(4)	a)	pontjában
1995.	évi	LIII.	törvény	56.	§	(4)	bekezdés	a) pontja
1995.	évi	LIII.	törvény	56.	§	(4)	bekezdésének	a) pontja
1995.	évi	LIII.	törvény	56.	§-ának	(4)	bekezdésének	a) pontjában

multilogic

Work type

Article number

Establish the corpus

Create Document-Data matrix

Create abstract data model

Extract knowledge

Build knowledge base

Step 3: Extract knowledge - Example



■ Applied semantic vocabularies:

- MetaLex OWL
- Legal Knowledge Interchange Format
- Dublin Core, Dublin Core Terms
- Open Provenance Model Vocabulary
- W3C Time ontology
- Friend of a Friend

Step 4: Build knowledge base

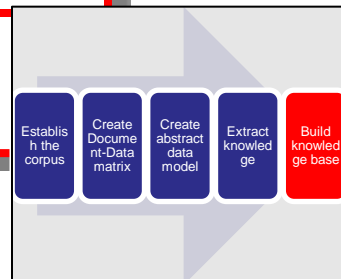
■ Data model excerpt of regulations

```
[WorkURI]
  rdf:type metalex-owl:BibliographicWork
  metalex-owl:resultOf [WorkCreationEventURI]
  rdf:type metalex-owl-ext#<work-category> # e.g. "metalex-owl-ext#act"
  dct:terms:title "<Title of the legal document>"^^<xsd:string>
  metalex-owl:Author [WorkAuthorURI] # e.g. some ministry
  metalex-owl:countryCode "<jurisdiction-code>"^^<xsd:string> -- e.g. "hu"
  metalex-owl:issueID "<document-id>"^^<xs:integer> -- e.g. "11"
  metalex-owl-ext:issueDate "<date>"^^<xs:integer> -- e.g. "1960"
```

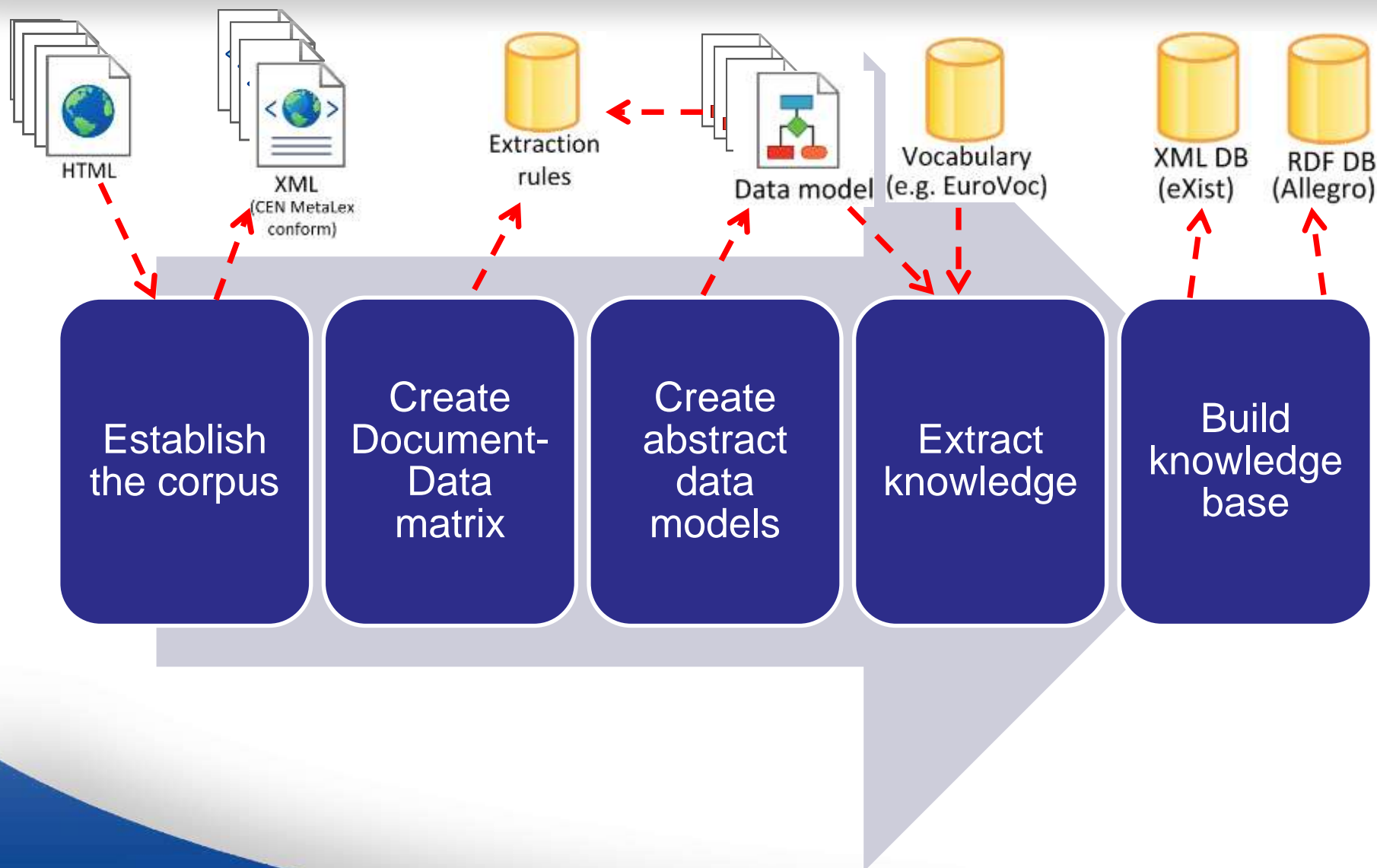
```
[ExpressionURI]
  rdf:type metalex-owl:BibliographicExpression
  metalex-owl:resultOf [ExpressionCreationEventURI]
  time-modification:in_force [ExpressionTimeModificationURI]
  metalex-owl:languageCode "<language code>"^^<xsd:language> -- e.g. "hun"
  metalex-owl#realizes [WorkURI]
  metalex-owl:cites [WorkURI]
```

```
[WorkCreationEventURI]
  rdf:type metalex-owl:LegislativeDelivery
  metalex-owl:participant [ParticipantURI] -- e.g. "Parliament"
```

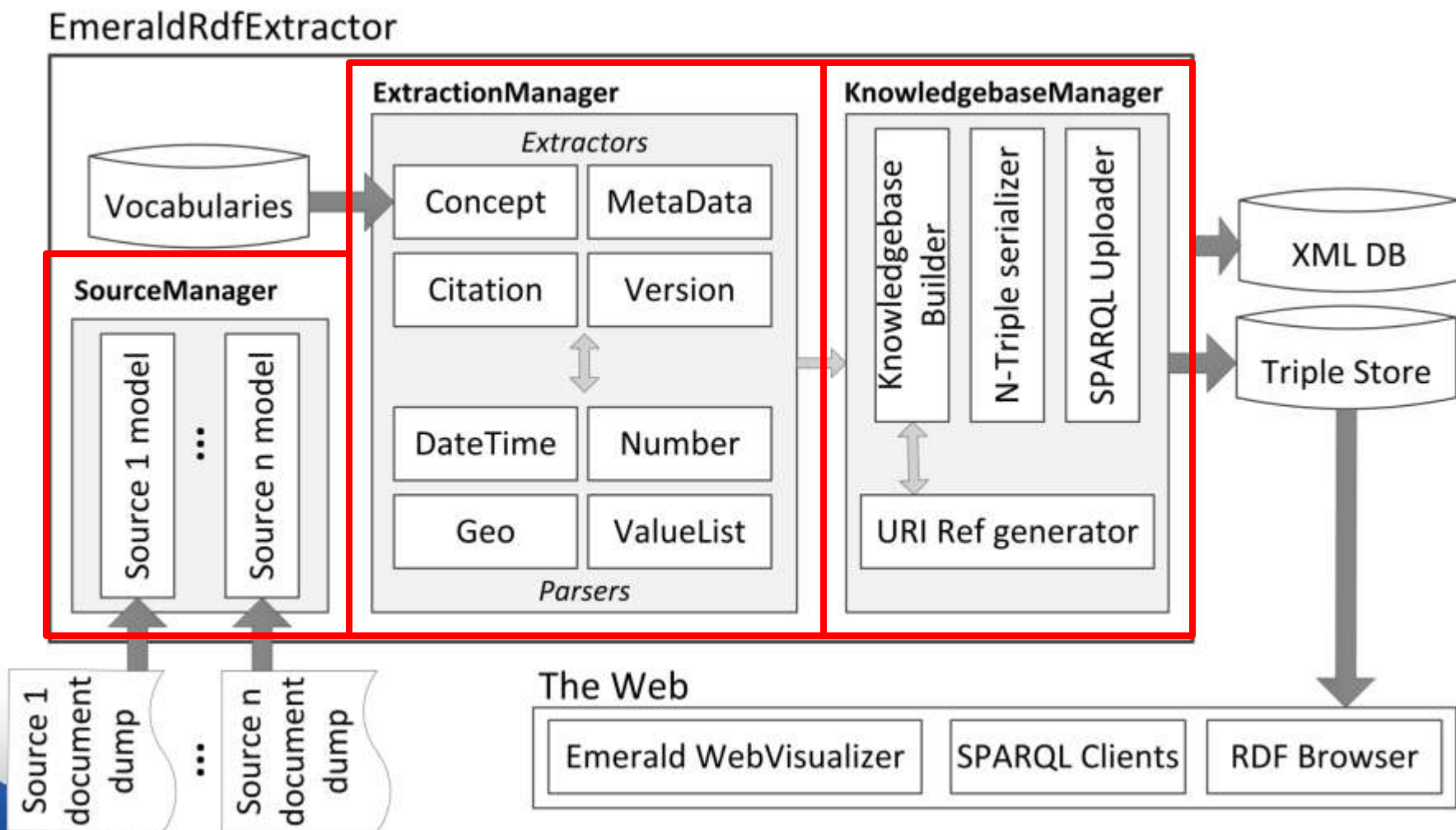
```
[ExpressionCreationEventURI]
  rdf:type metalex-owl:LegislativeModification
  metalex-owl:participant [ParticipantURI] -- e.g. "Parliament"
```



Summary: The Process of Extraction



Architecture of the Framework



Evaluation and Future Work

Evaluation with a case study

■ Purpose

- To evaluate our proposed method on a real-life situation

■ Background

- Hungarian legislative documents (www.njt.hu)
- 256 most popular Hungarian regulations (work)
- 3502 version of documents (expressions)
- HTML format
- Different types (6):

Main types	Subtypes
Adjudication	67 + 5
Act	1
Actdecree	1
Constitution	1
Decree	67
JointDecree	1

Results - 1

- We created
 - 6 document data models
 - 67 extraction rules of 32 kinds of citations
- We extracted
 - 524k triples (cca. 150 triples per document)
 - Cca. 73% of the triples were citations
 - 1672 documents were cited from the 256 documents → 7350 links

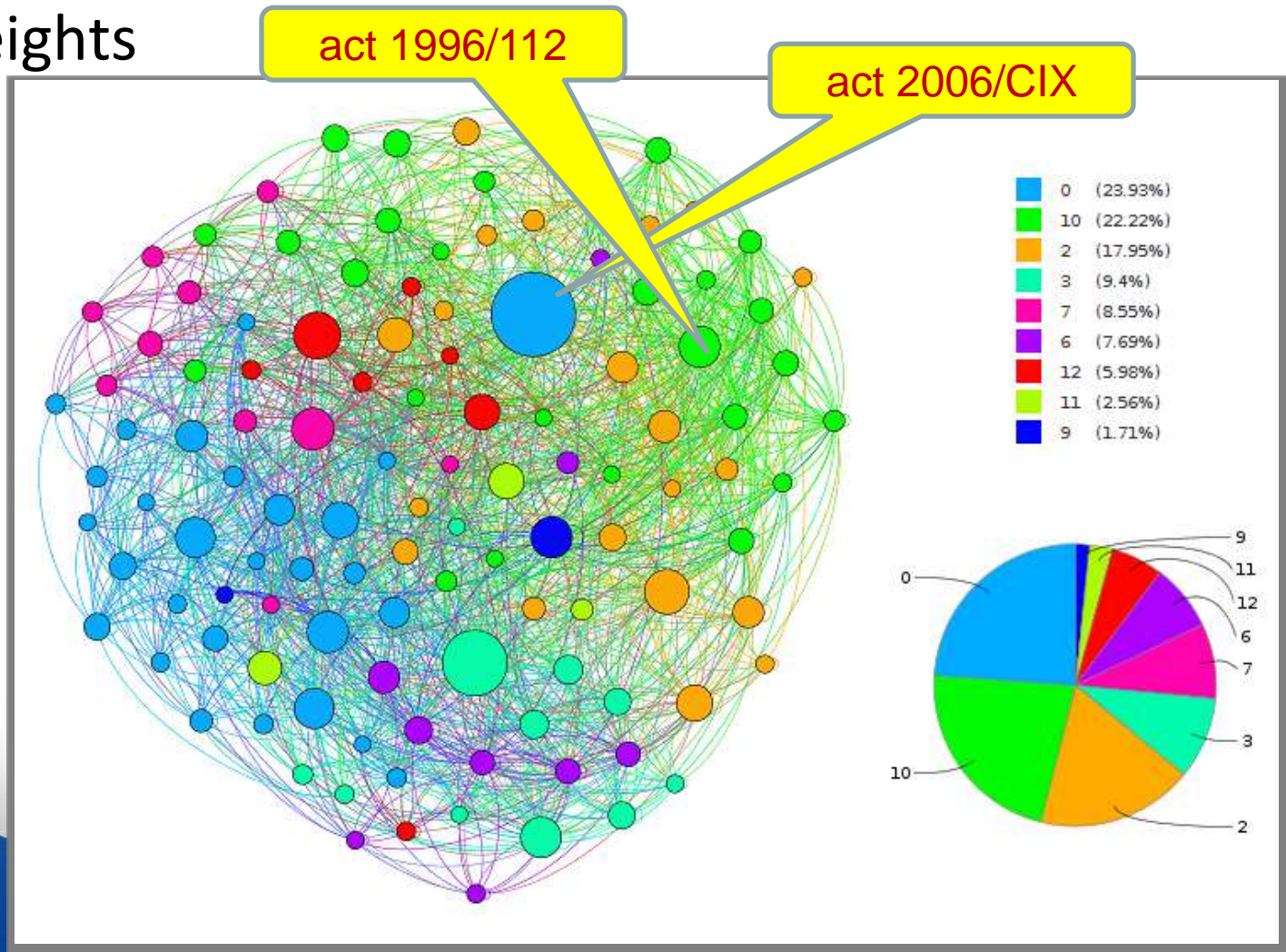
Results - 2

■ Citation network analysis:

Metrics	Definition	Value
In-degree	the number citations from a document	[1, 129]
Out-degree	the number citations to a document	[1, 121]
Average degree	the number citations from/to a document	4.4
Network diameter	the maximum citation-distance between all pairs of documents	9
Average path length	the average number steps along the shortest path for all documents	3.4
Modularity	the strength of division of the network into communities	0.39
Number of communities	detected communities using resolution = 1:0	15

Results - 3

- Filtered citation network (nodes less than 35 degree) and its communities with their relative weights



Implementation: Emerald

EMERALD Webmegjelenítő Perspektíva: Dokumentum

Dokumentum leíró adatok

Kiemelt adatok Minden adat

Cím:	EVA		
Típus:	law	Hatályosság kezdete:	2011-01-01
Azonosító:	doc.law.EVA.2011.01	Hatályosság vége:	
Belső azonosító:	EVA_2002_XLIII_20110509	Publikáció dátuma:	2010-08-25
Alkalmazás:	Hungary	Kihirdetés dátuma:	
Nyelv:	hu	Visszavonás dátuma:	

Forrásdokumentum

2002. évi XLIII. törvény az egyszerűsített vállalkozói adóról

Az Országgyűlés az egyszerűsített adómegállapítási és beszedési szabályok alkalmazása érdekében azon kisvállalkozások számára, amelyeknél az általános szabályok szerint történő adóztatás tevékenységük jellege miatt nehézségekbe ütközne, valamint az [Európai Unióhoz](#) való csatlakozásból eredő szempontok érvényesítése céljából a következő törvényt alkotja:

Fejezet I.

A TÖRVÉNY HATÁLYA ^(p1)

§ 1.
Az egyszerűsített vállalkozói adó ^(p1a1)

(1) A Magyar Köztársaságban egyes személyek vállalkozási (gazdasági) tevékenységből származó bevételét az e törvényben meghatározott egyszerűsített vállalkozói adó (a továbbiakban: eva) terheli.

(2) Az evából származó bevétel a központi költségvetést illeti meg.

(3) Az adózó az evával összefüggő adókötelezettségeit e törvény, valamint az adózás rendjéről szóló törvény rendelkezései szerint teljesíti.

Fogalom találatok: Előző << 1 / 1 >> Következő

Terminológia Megjegyzések

Fogalmak Nyelv ▾

- EGT Finanszírozási mechanizmus [def]
- elektronikus úton kibocsátott számla [def]
- ellenérték [def]
- Európai Közösség [def]
- Európai Unió [def]
- gazdasági tevékenység [def]
- gyűjteménydarab [def]

Fogalomdefíció

Az Európai Unió (EU) egy túlnyomórészt Európában található, 27 tagállamból álló gazdasági és politikai unió. A regionális integráció iránt elkötelezett szervezetet 1993. november 1-i hatállyal hozta létre az 1992. február 7-én aláírt Maastrichti szerződés az Európai Gazdasági Közösség alapjain.

Fogalom gráf

```
graph TD; Thing --> TargetArea[Tárgyterületi fogalom]; Risk[Kockázatvállalás] --> TargetArea; Retail[Retail fogalom] --> TargetArea; Limit --> TargetArea; Actor[Szereplő] --> TargetArea; EU[Európai Unió] --> TargetArea; Credit[Hitelintézet] --> TargetArea; Debt[Kiteltség] --> TargetArea;
```

Conclusions

The presented framework aims at:

- Standardizing data collections (CEN MetaLex Standard)
- Building knowledge models based on
 - document metadata (MetaLex, LKIF, ...)
 - controlled vocabularies (EuroVoc and other custom vocabularies)
 - citation informations
 - version information
- Providing joint management of legal texts (XML) and their formal representations (formal model): metadata and concepts