



**C.I.R.S.F.I.D**

**Alma Mater Studiorum Università di Bologna**

**Research Centre of History of Law, Philosophy and Sociology of Law,  
Computer Science and Law**

# Automated Extraction of Normative References in Legal Texts

---

## Legal informatics in Italy

**Bologna, Italy  
18-19 June 2003**

**Dr. Matteo Massini  
mmassini@cirfid.unibo.it**



# Project goal

---

automatic recognition of normative references:

- localising of every word sequence that identifies a reference
- interpretation of the contained reference data in order to:
  - find out the reference destination
  - extract the reference properties



# APPROACH AND METHOD

---

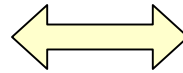
- Programming Language:
  - PERL (Practical Extraction and Report Language)
- References Template Language:
  - RegEx (Regular Expression)
- Dictionaries
  - Dictionaries retrieved from legistic analysis of legal texts (CIRSFID background experience)
  - Code libraries (ex: roman number interpretation)



# RegEx Examples: finding section citations

---

- ([part] \s\* [numbers])
- ([ord\_it] \s+ [part])
- ([part] \s+ [ord\_it])
- ([ord\_it] \s+)? ([part] \s\* [numbers])



- art.15  
letter i)
- first section
- second title
- next article 17

\s: space, tab character or any kind of separator  
[ord\_it]: first, second, third.... previous, next  
[part]: book, part, section, title, article ...or abbreviations  
[numbers]: arabic or roman numerals



# RegEx Examples : locating the candidate references

---

- `[element] ( [charSep]+ ( [element] | [wordSep] ) )*`

- `[article 6], [first section], of [act] [1 april 1981], [n. 121]`
- `[article 371-bis], [section 3], of [civil code]`
- `[directive] [83/189/EC]`
- `[article 3] [section 1]`

`[element]`: act types or their abbreviations, known compilations or their abbreviations (ex.: cc instead of civil code), section references, dates (in several formats), act IDs (ex.: 83/189/CEE)

`[charSep]`: separator, comma, quotes, minus....

`[wordSep]`: articles, prepositions and meaningless words



# minimum-content criteria

---

In order to classify the located text sequence as link, it must contain at least one of the following:

- Act type + accompanying number (or ID)
  - act n. 675
  - directive 83/189/EC
  - dlgs. 3 February 1993, n. 29
- Partition name + number (arabic or roman)
  - article 5
  - letter i)
- Known compilation
  - Civil Code
  - Maastricht Treaty (EU)



# Link Interpretation

---

The extraction of the data contained in the found references occurs using:

- Conversion tables attached to the dictionaries
  - ex: Civil Code → <DTID=7><ID1=262><DD=16/03/1942>
- Conversion rules
  - RegEx for dates interpretation:
    - 1st march 1986
    - 17/01/1975
  - Interpretation algorithms for:
    - Roman numbers
    - Lists numbered with letters (ex: “a), b), c), ...”)
    - Ordinal numbers (ex: first, second, ...)
    - Latin adverb (ex: bis, ter, ...)
    - Their compositions (ex: 4-ter, 1bis.1, ...)



# Document Base

---

- The document base used to verify the results is a consolidated data base about law of IT (1640 historical texts)
- The documents has been analysed and marked with *manual* tools by the CIRSIFID staff
- The documents are marked in NML, an *XML like* Mark-Up language, that defines structures and information extracted from the plain text; in NML the normative references are marked with a special tag named N\_LINK



# Match rules

---

The comparison between found references and N\_LINKs occurs in five levels:

- **WELL-FORMED-REFERENCE MATCH**
- **NOT WELL-FORMED-REFERENCE MATCH**
- **TEXT-ONLY MATCH**
- **PARTIAL TEXT-ONLY MATCH**
- **NULL MATCH**
- References not marked with N\_LINK tag and located by the parser are classified as **LINK PLUS**



# Match Rules (2)

---

## **WELL-FORMED-REFERENCE MATCH:**

- the normative reference detected by the parser has properties identical to those of the corresponding n-Link. These properties are such that they enable us to identify unambiguously the specific act and text partition referred to
  - ex: *article 1 of act n. 675, 31 December 1996*

## **NOT WELL-FORMED-REFERENCE MATCH:**

- the properties match completely, but they are such that we cannot identify unambiguously the specific act referred to
  - ex: *article 1, section 2*



# Match Rules (3)

---

## **TEXT-ONLY MATCH:**

- The text segment (normative reference) detected by the parser is identical to the text tagged by n-Link, but the parser was unable to extract every useful added knowledge from that segment concerning the reference
  - *ex: articles 5, 8, 20 sections 9, 10 and 11*

## **PARTIAL TEXT-ONLY MATCH:**

- The text segment detected by the parser matches only in part the text segment tagged by N\_LINK



# Examples (1)

- Act n. 675: WELL-FORMED-REFERENCE MATCH

```
*LINK*: <DTID=2><ID1=400><DD=23/08/1988><PID=n_0_0_0_0_0_17_1_[a]_0_>  
<TEXT=article 17, section 1, letter a), act n. 400, 23 august 1988>  
N-LINK: <DTID=2><ID1=400><DD=23/08/1988><PID=n_0_0_0_0_0_17_1_1_0_>  
<TEXT=article 17, section 1, letter a), act n. 400, 23 august 1988>
```

- Act n. 675: NOT WELL-FORMED-REFERENCE MATCH

- The document structure has not been used
- Many references are internal or in abbreviated form

```
*LINK*: <PID=n_0_0_0_0_0_7_1_0_0_><TEXT=first section of article 7>  
N-LINK: <DTID=2><ID1=121><DD=01/04/1981><PID=n_0_0_0_0_0_7_1_0_0_>  
<TEXT= first section of article 7>
```

**\*LINK\***: text retrieved by the parser and its extracted properties

**N-LINK**: information contained in the N\_LINK tag



# Examples (2)

---

- TEXT-ONLY MATCH

**\*LINK\*:** <TEXT= articles 5, 8, 20, sections 9, 10 e 11, 22, 25, sections 1 e 3, 27, sections 2, 30, 32, 40, 41, 42, 43, 44, 45, section 2, 53, section 2, 57, 62, 72, section 2 e 3, of dlgs. N. 29, 3 February 1993>  
**N-LINK:** <TEXT= articles 5, 8, 20, sections 9, 10 e 11, 22, 25, sections 1 e 3, 27, sections 2, 30, 32, 40, 41, 42, 43, 44, 45, section 2, 53, section 2, 57, 62, 72, section 2 e 3, of dlgs. N. 29, 3 February 1993 >

- PARTIAL TEXT-ONLY MATCH

**\*LINK\*:** <ID1=143><DD=24/06/1975><TEXT=n. 143 of 24 June 1975>  
**N-LINK:** <DTID=74><ID1=158><DD=10/04/1981>  
<TEXT=OIL agreement n. 143, 24 June 1975, ratified with act n. 158, 10 April 1981>

**\*LINK\*:** text retrieved by the parser and its extracted properties  
**N-LINK:** information contained in the N\_LINK tag



# Examples (3)

---

- NULL MATCH

```
<DTID=68><ID1=495><DD=16/12/1992><TEXT=regulations>  
<DTID=56_A><ID1=93/15/CEE><TEXT=attachments II and III of the directive>  
<DTID=2><ID1=1058><DD=07/10/1947><PID=n_0_0_0_0_0_47_1_0_0_><TEXT=the words>  
<DTID=8><ID1=305><DD=10/07/1991><PID=n_0_0_0_0_0_11_7_0_0_>  
<TEXT=second sentence>
```

- LINK PLUS

```
<PID=n_0_0_0_0_0_[THIS]_0_0_0_><TEXT=this article>  
<DTID=7><ID1=929><DD=21/06/1942><TEXT=regio decreto 21 giugno 1942, n. 929>  
<PID=n_0_0_0_0_0_-1]_[LAST-1]_0_0_><TEXT=next to last section of previous article>
```



# Results

- Processing of the entire document base (1641 docs)

TOTAL NUMBER N-LINK	=	<b>23530</b>	
%WELL-FORMED-REFERENCE MATCH	=	<b>35.6 %</b>	
%NOT-WELL-FORMED-REFERENCE MATCH	=	<b>49.4 %</b>	<b>93.6%</b>
%TEXT-ONLY MATCH	=	<b>4.3 %</b>	
%PARTIAL TEXT-ONLY MATCH	=	<b>4.3 %</b>	
%NULL MATCH	=	<b>6.4 %</b>	
TOTAL	=	<b>100 %</b>	
TOTAL NUMBER OF PLUS-LINKS	=	<b>3281</b>	



# Prototype limits and future...

---

- Complex lists analysis
- Definition of separation methods for connected links
- The use of the document structure information to improve the properties precision
- Suggestion for incomplete links



---

*thank you for your attention*