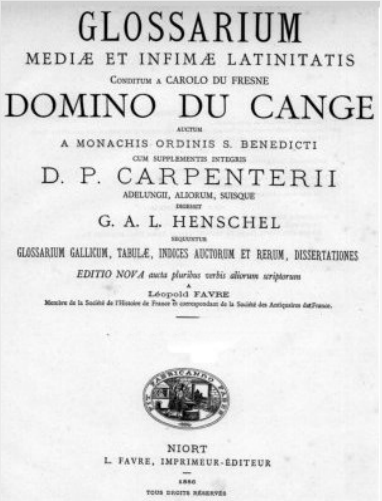
XVI EURALEX International Congress, Bolzano/Bozen, 15-19 July 2014

The Problem of Scale
Handling (More) Data with the *WikiLexicographica*

Bruno Bon
Institut de Recherche et d'Histoire des Textes, CNRS
Paris, France

Krzysztof Nowak
Institute of Polish Language
Polish Academy of Sciences
Kraków, Poland

(Pre)History of the Project (1)


GLOSSARIUM
MEDIÆ ET INFIMÆ LATINITATIS
CONDITUM A CAROLO DU FRESNE
DOMINO DU CANGE
AUCTUM
A MONACHIS ORDINIS S. BENEDICTI
CUM SUPPLEMENTIS INTEGRIS
D. P. CARPENTERII
ADELUNGII, ALIORUM, SUISQUE
DIGESSIT
G. A. L. HENSCHEL
SEQUUNTUR
GLOSSARIUM GALLICUM, TABULÆ, INDICES AUCTORUM ET RERUM, DISSERTATIONES
EDITIO NOVA auctis pluribus verbis aliorum scriptorum
A
Léopold FAVRE
Membre de la Société de l'Histoire de France & correspondant de la Société des Antiquaires de France.
NIORT
L. FAVRE, IMPRIMEUR-ÉDITEUR
1886
TOUS DROITS RÉSERVÉS


NOVUM GLOSSARIUM
MEDIAE LATINITATIS
AB ANNO DCCC USQUE AD ANNUM MCC
EDENDUM CURAVIT
CONSILIUM ACADEMIARUM CONSOCIATARUM
JACQUES MONFRIN
HAFNIAE

ca. 1920

Union Académique Internationale launches an intitiative aiming at replacing Du Cange's *Glossarium* with a new scholarly dictionary – *Novum Glossarium Mediae Latinitatis*

soon after

several national editorial committees decide to create separate national Medieval Latin dictionaries

21st century

several dictionaries varying in their

- geographical and chronological scope
- work progress *etc.*

(Pre)History of the Project (2)

2011-2015

Cost Action *Medioevo Europeo*

- main goal

  VCMS = „Virtual Center of Medieval Studies”

- WG 3 goal

  unified access to the corpora and dictionaries

2013 -
Prototype of the *WikiLexicographica*, a tool of integrated access to the dispersed lexicographic content

Outline

- *WikiLexicographica* – prototype and data import process
- macrostructural organisation
- issues with data quantity & quality
- gardening the *WikiLexicographica*
  - templates
  - redirections
  - concept pages

WikiLexicographica. Software

+

this word's headword is
mandragora

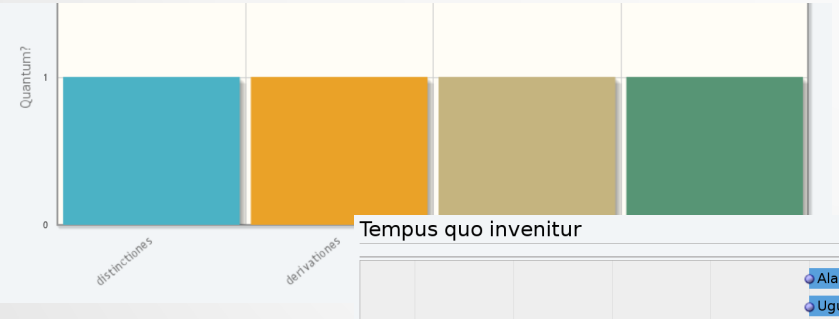
↓

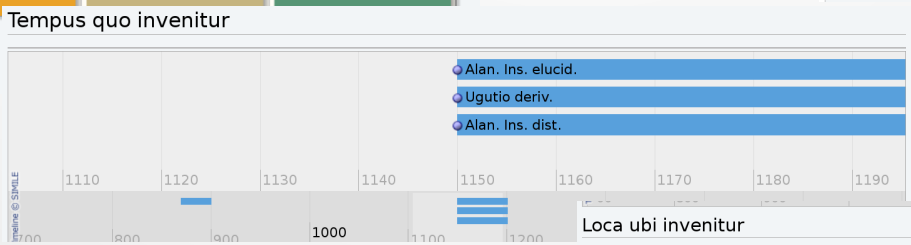
this word's headword is
[[headword::mandragora]]

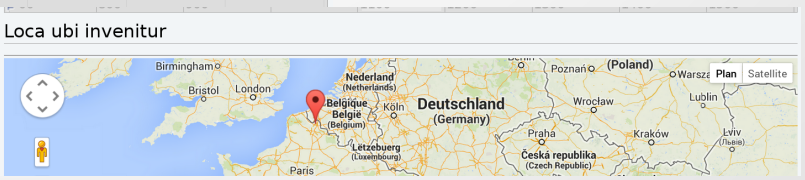
WikiLexicographica. Advantages (1)

- separation of basic and advanced view

- lexicographic data visualisation

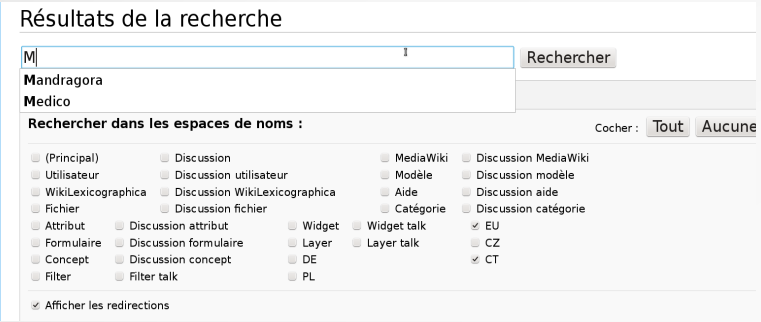






WikiLexicographica. Advantages (2)

- efficient and varied data access



simple search

Semantic Drilldown

Factbox browsing

advanced search

- collaboration support

Importing (More) Data

imported data

NGML

NGML sources

LMILP

LMILP sources

Latin Wiktionary

DuCange

import tools

Import pages

Upload XML data

Please export the file from the source wiki using the export utility. Save it to your computer and upload it here.

Filename: Przeglądaj... Nie wybrano pliku.

Comment:

Destination root page (optional):

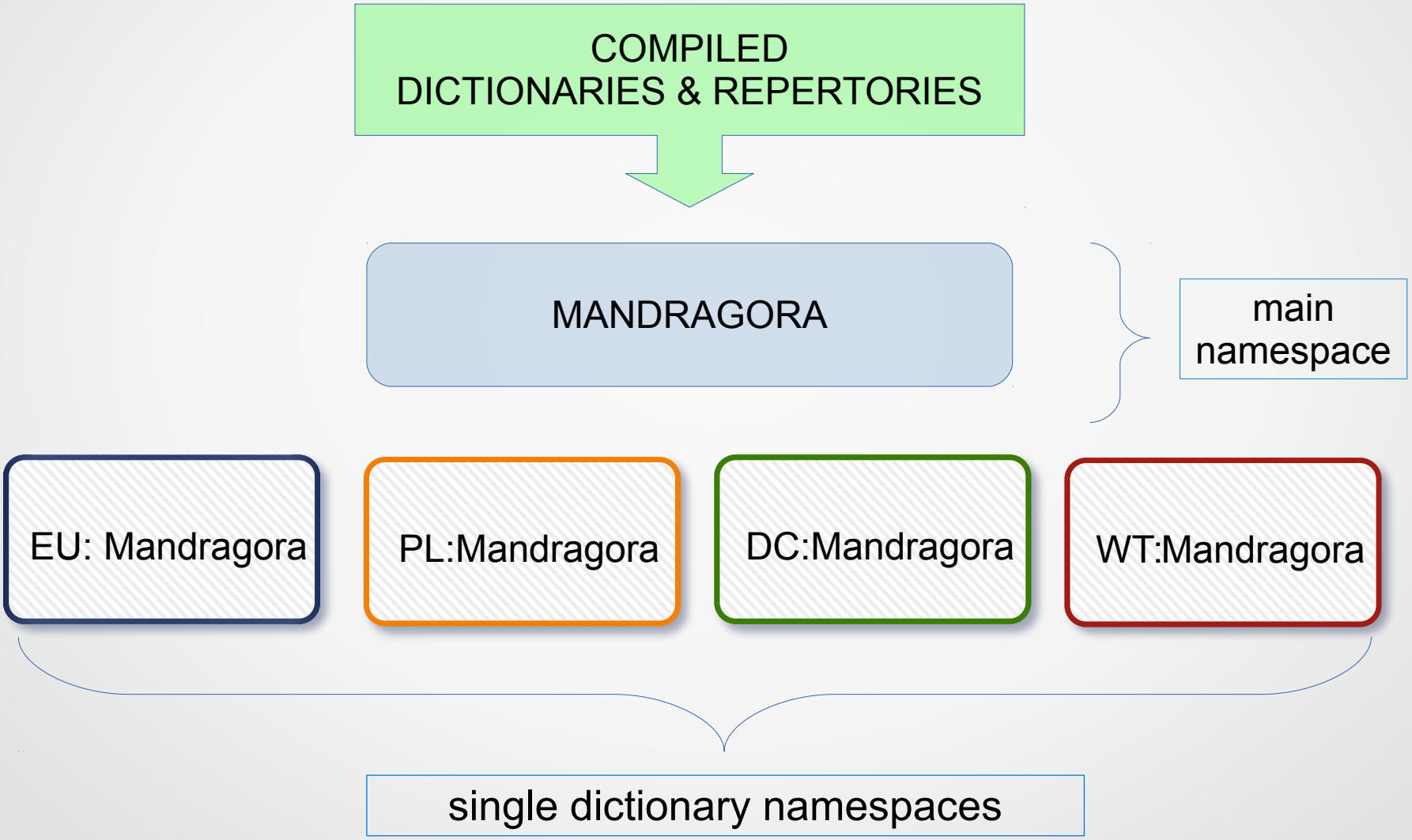
Upload file

```
1775  php importDump.php --conf ../LocalSettings.php < ../../work/WikiLexicographica/export/A-Z_export.xml
1779  php importDump.php --conf ../LocalSettings.php < ../../work/WikiLexicographica/export/A-Z/A-Z_export.xml
1785  php importDump.php --conf ../LocalSettings.php < ../../work/WikiLexicographica/export/A-Z/A-Z_main_export.xml
1786  php importDump.php --conf ../LocalSettings.php < ../../work/WikiLexicographica/export/A-Z/A-Z_main_export.xml -v
```

import format

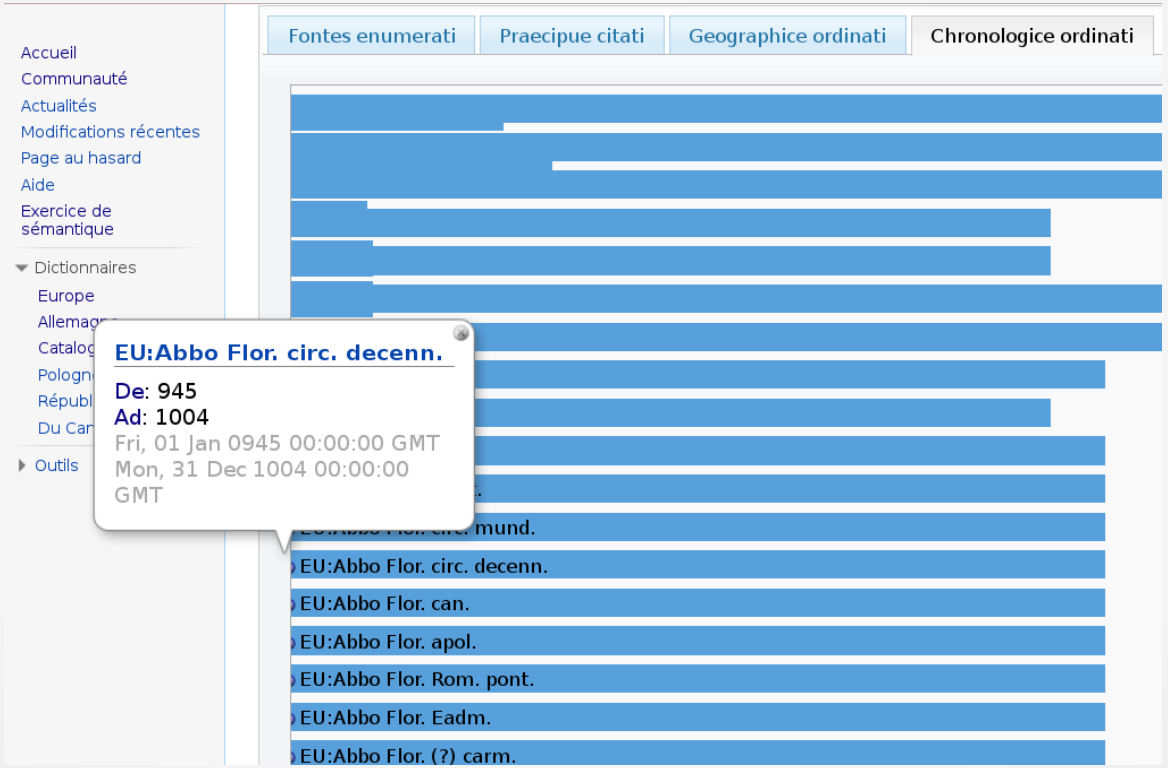
```
1  <?xml version="1.0" encoding="UTF-8"?>
2  <mediawiki version="0.8" xml:lang="">
3     <page>
4        <title>a</title>
5        <ns>0</ns>
6        <revision>
7           <text>{{Art|{{Art lemma|a|silent}}{{Art form|a|silent}}{{Art form|ab|silent}}{{Art
            form|abs|silent}}{{Art inv|a|silent}}{{Art inv|ba|silent}}{{Art inv|sba|silent}}{{Art
            iType|-|silent}}{{Art iType|-|silent}}{{Art iType|-|silent}}{{Art pos|PRE|silent}}{{Art def|(+ abl.) à
            partir de, loin de, du côté de, par suite de, par|silent}}|silent}}</text>
8        </revision>
9     </page>
10    <page>
11       <title>abacinus</title>
12       <ns>0</ns>
```

Macrostructure. Namespaces



Data Quantity Issues. Timelines

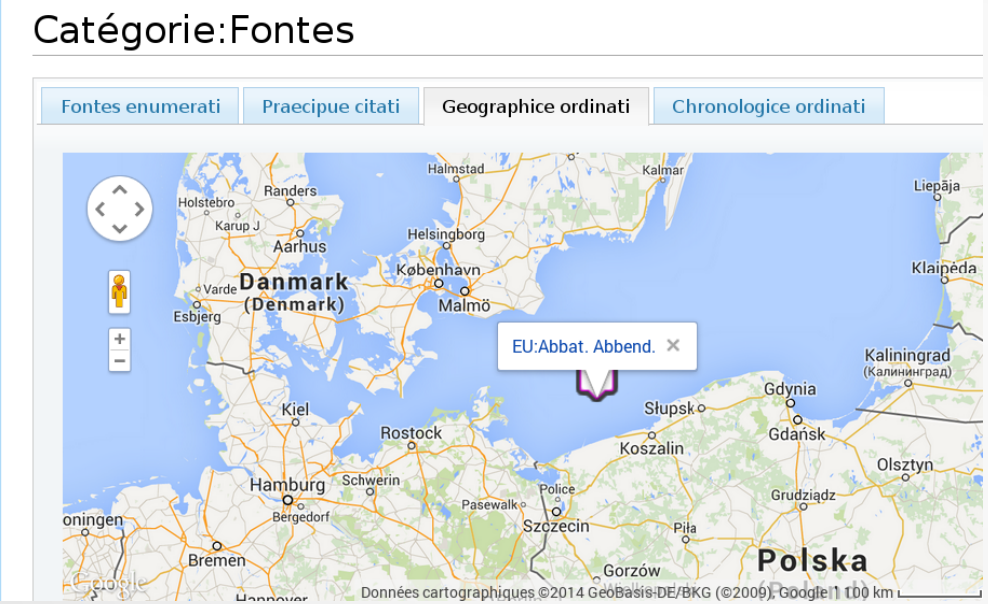
too much data



limiting data display

Data Quality Issues. Maps (1)

low granularity of the geographical data



eliminating uncertain data?

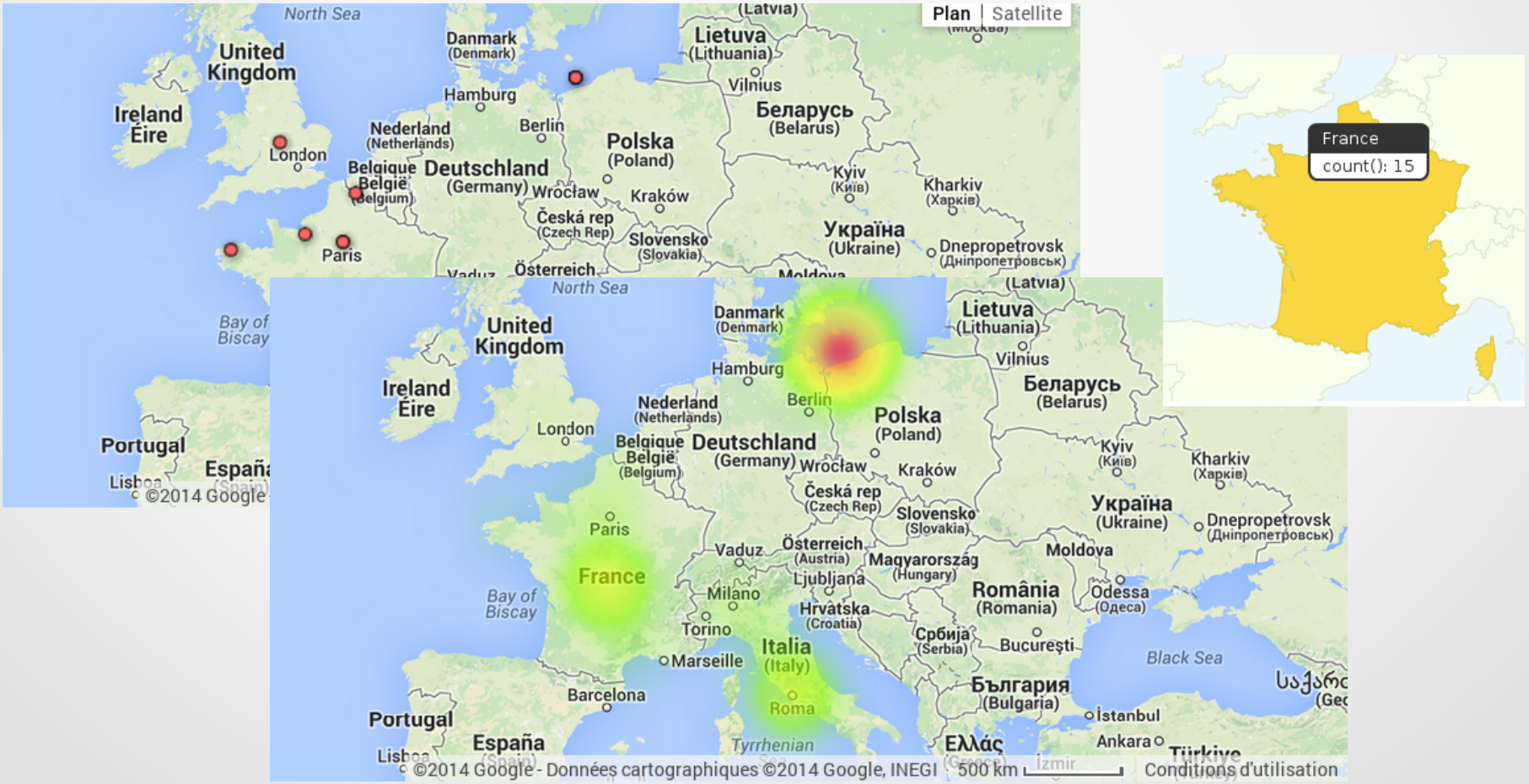
Kraków, Poland, Europe

Poland, Europe

Europe

Data Quality Issues. Maps (2)

implementing static choropleths with Google Fusion Tables



Microstructural Problems

poor formatting

difficult to reflect nested structures (senses, notes *etc.*)

data extraction approach

different lexicographic conventions, differing encoding

copyright fears

Gardening the WikiLexicographica. Templates

automatisation of repetitive tasks

```
{{#set:Lemma=  {{#regex:
{{{1}}}|/([^0-9.]+\s*)+/}}
}}
```

1. pinna
↓
[[Lemma::Pinna]]

cuts homonym number from the headword's form and sets [[Lemma]] property

adding redundant annotation

```
{{#set: POS= {{#if: … [[iType::~1*]] … VBE … }} }}
```

pinno ... 1. → [[POS::VBE]]

checks if word's inflectional type begins with number;
if this is the case, sets [[POS]] property to VBE

Gardening the WikiLexicographica. Redirection Pages

dealing with reference entries

pinguesso *v.* **pinguesco**

EU:Pinguesso

```
#REDIRECTION
[[EU:pinguesco]]
```

dealing with varying labels

Lemma
Lemme

Lemme

```
#REDIRECTION
[[Lemma]]
```

Gardening the WikiLexicographica. Concept Pages

adding hierarchy, exposing data to the users

<u>1st Conjugation Markers</u>

Lemma ending with **-o(r)**

&

EU → conjugation number || PL → inflectional endings

pinguesco 1. … **-ar(e|i)**

<u>1st Conjugation Concept Page</u>

`Concept:Coniugatio_I`

`[[Lemma::~*o||~*or]] [[iType::~1*||~-ar*]]`

More?

E-mail

- Bruno Bon bruno.bon@irht.cnrs.fr
- Krzysztof Nowak krzysztofn@ijp-pan.krakow.pl

WWW

http://scriptores.pl/wiki