POSTDATA:
Towards publishing European Poetry as Linked Open Data

Mariana Curado Malta (mariana@iscap.ipp.pt; mariana.malta@linhd.uned.es), ISCAP, IPP | LINHD – UNED

Paloma Centenera (paloma.centenera@linhd.uned.es), LINHD - UNED

Elena González-Blanco (egonzalezblanco@flog.uned.es), LINHD - UNED
Outline

• The Context
• The Problem
• The Approach
• Where are we now?
• Future Work
• Conclusions
Context

Project POSTDATA
Poetry Standardization and Linked Open Data

ERC Grant
PI: Elena Gonzalez-Blanco
A digital humanities center is an entity where new media and technologies are used for humanities-based research, teaching, and intellectual engagement and experimentation. The goals of the center are to further humanities scholarship, create new forms of knowledge, and explore technology’s impact on humanities based disciplines."

Context – Where?

• LINHD: a bridge between different fields of knowledge

• LINHD has:
  ➔ Philologists
  ➔ Software Developers
  ➔ Natural Language Processing Experts
  ➔ Ontologists & LOD technologists
DH in the World
Context - What

• The metrics on European Poetry

Estuans intrinsecus
in amaritudine
factus de materia
folio sum similis,

Cum sit enim proprium
supra petram ponere
stultus ego comparor
sub eodem aere

Feror ego veluti
ut per vias aeris
non me tenent vincula,

Carmina Burana, 191

pravis
The problem

• At least 21 repertoires on Poetry metrics & other information (in the Web of Documents)

• This community wants to share all the data among repertoires

• ....to enhance its research

• And more...
The problem

• First issue: standardize poetic features
  ➔ Different languages
  ➔ Different cultures/traditions

Philologists take care of this issue!
Philological barriers: different ways of conceptualization

Estuans intrinsecus
in amaritudine
factus de materia
folio sum similis,

Cum sit enim proprium
supra petram ponere
stultus ego comparor
sub eodem aere

Feror ego veluti
ut per vias aeros
non me tenent vincula,
quero mei similes

Carmina Burana, 191

Alexandrines

Goliardic

4x(7pp+7p)
(Romance)

4x(7pp+7p)
(Classic Latin)
Philological standardization: starting point

<table>
<thead>
<tr>
<th>Author</th>
<th>Title</th>
<th>Incipit</th>
<th>Manuscript</th>
<th>Post quem</th>
<th>Ante quem</th>
<th>Language</th>
<th>Topics</th>
<th>Edition</th>
<th>Online edition</th>
<th>Work</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Isometrism</th>
<th>Isostrophism</th>
<th>Metrical scheme</th>
<th>Rhyme scheme</th>
<th>Rhyme</th>
<th>Musical notation</th>
<th>Number of stanzas</th>
<th>Number of lines</th>
<th>Poetic form</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The problem

Second issue: repertoires locked in their silos of information:

- Different paradigms: Local and Web of Docs
- Different technologies: XML, Excel, Access, MySQL, SQL, Data stored in Perl Objects (so far)
- Different data models
The problem

• How to overcome these differences?
• LOD technology
• Development of a Metadata **Application Profile** for the European Poetry community
The Approach

- Method for the development of Metadata Application Profiles (Me4MAP)
  - Me4MAP establishes a well defined process for the development of a MAP:
    - defines activities
    - when those activities should take place
    - how they interconnect
    - and their resulting deliverables
The Approach

A1 - Development of the Glossary

A2 - Development of the DCAP Documentation

A3 - Environmental Scan

S1 - Developing the Functional Requirements

S2 - Developing the Domain Model

S3 - Developing the Description Set
  S3.1 - Vocabulary Alignment
  S3.2 - Constraints Matrix
  S3.3 - Constraints Matrix Test
  S3.4 - DSP Encoding

S4 - Developing the Syntax Guidelines

S5 - Developing the Usage Guidelines

A4 - Validation in Production

Postdata
Poetry Standardization and Linked Open Data

LiNHD
Laboratorio de Innovación en Humanidades Digitales
Were are we now?
Where are we now?

- **S1**: Defining the Functional Requirements
  - Analysing the Websites’ functionalities and the Logical Models of the databases (when possible)

- **S2**: Defining the Domain Model
Where are we now?

- Reverse engineering process eliminates all the details that have to do with the implementation/representation.
- We have followed the process:
  - ID keys deleted.
  - Separate different concepts that are represented in the same table.
  - Tables that enumerate terms deleted → become properties that can be repeated.
  - When models have conceptual problems → fix problems.
Analysing Data Models

Relational Database

Conceptual Model
Analysing Data Models

XML Schema Model

Conceptual Model
Analysing Data Models

Perl Script Structure

```
#!/usr/bin/perl

my $poem = {
    # General information on poem, author, book from which it comes...
    'head' => {
        # Id of a book from which the poem comes
        'id' => 4,
        # Id of a poem unique in entire corpus - first four numbers correspond to the Id of a book
        'poem' => '0001-0002-0003-0004',
    },
    # GENERAL BIBLIOGRAPHIC METADATA
    'bibliography' => {
        # Title of a book from which the poem comes from
        'title' => 'Na poetu a na nebi',
        # Subtitle of a book
        'subtitle' => 'Hsing',
        # Year of publication of a book
        'year' => 1900,
        # Place where the book was published
        'place' => 'Praha',
        # Publisher of a book
        'publisher' => 'Unic',
        # First, second, third... edition of a book
        'edition' => 1,
        # Name of the library from which the book have been digitized & the signature of a book in library
        'signature' => 'Národní knihovna CR, Praha; 54 M 2208',
        # Title of a poem
        'poemtitle' => 'STROPHELAVI VODNI CHLOP',
    },
    # INFORMATION ON AUTHOR (OR EDITOR) OF A BOOK
    'author' => {
        # Name of the author that is written on the book (may be a pseudonym)
        'name' => 'Albert, Eduard',
        # Real name of the author (differs from 'name' in case of pseudonyms) - e.g. the play 'Catinell' would
        'identity' => 'Albert, Eduard',
        # Year of birth of the author (if not known, the attribute is left undef)
        'born' => 1841,
        # Year of death of the author (if not known, the attribute is left undef)
        'died' => 1909,
    },
    # INFORMATION ON AUTHOR OF A POEM
    'locAuthor' => {
        # If it differs from the author/editor of a book (anthology etc.) - otherwise left undef
        'locname' => undef,
    }
};
```

Conceptual Model
Analysing Data Models

- During the process of reverse engineering we standardize, i.e.
  - Call the same concepts by the same name (working together with the philologist)
  - Try to call the same names to tables or properties as classes or terms that already exist in A3: Environmental Scan
Where are we now?

• If the repertoire’s responsibilities did not provide database definition → analyse the functionalities of the Website

• Study the controlled vocabularies and standardize them: 1) ask for them, 2) collect them in the Websites
Where are we now?

• Study other communities/projects for interoperability, ex:
  ➢ Biblioteca Nacional de España - http://www.bne.es
  ➢ Pelagios - http://pelagios.org/
  ➢ Claros - http://www.clarosnet.org/
Where are we now?

• A3: Environmental Scan:
  • A report
  • contains a review schemas available in any serialization of the Semantic Web
  • that may serve the needs of the Domain Model
A1 - Development of the Glossary

A2 - Development of the DCAP Documentation

A3 - Environmental Scan

S1 - Developing the Functional Requirements

S2 - Developing the Domain Model

S3.1 - Vocabulary Alignment

S3.2 - Constraints Matrix

S3.3 - Constraints Matrix Test

S3.4 - DSP Encoding

S4 - Developing the Syntax Guidelines

S5 - Developing the Usage Guidelines

A4 - Validation in Production
Future Work

• At the end of S1 & S2 processes we will have Functional requirements & Domain Model defined
• We will validate the Domain Model in two meetings (Jan/Feb 2017) with:
  1. The repertoire’s responsibles (circa 20)
  2. Semantic Modelers experts (3)
Next step?

- Defined the Domain Model & the Environmental Scan → develop the S2.1: Vocabulary Alignment
  
  ➢ To match the terms of the metadata schemas identified in the Environmental Scan (A3) with the needs of the Domain Model.
Conclusions

• POSTDATA aims to put poetry metrics data in LOD

• There are at least 21 repertoires on the Web of Documents

• To achieve that we need to: 1) standardize the way poetry metrics is defined, 2) create a Metadata Application Profile (MAP) for the European Poetry community

• We are following Me4MAP to create this MAP
Europeana Poetry?
Mariana Curado Malta
mariana.malta@linhd.uned.es
Paloma Centenera
paloma.centendera@linhd.uned.es
Elena González-Blanco
gonzalezblanco@flog.uned.es

linhd.uned.es
postdata.linhd.es
@linhduned