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

Context – Where?

POSTDATA
Poetry Standardization
and Linked Open Data

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

LiNHD
LABORATORiO DE iNNOVACiÓN
EN HUMANiDADES DiGiTALES

> *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".*

Diane M. Zorich, A Survey of Digital Humanities Centers in the United States, 2008

UNED

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

POSTDATA
Poetry Standardization and Linked Open Data

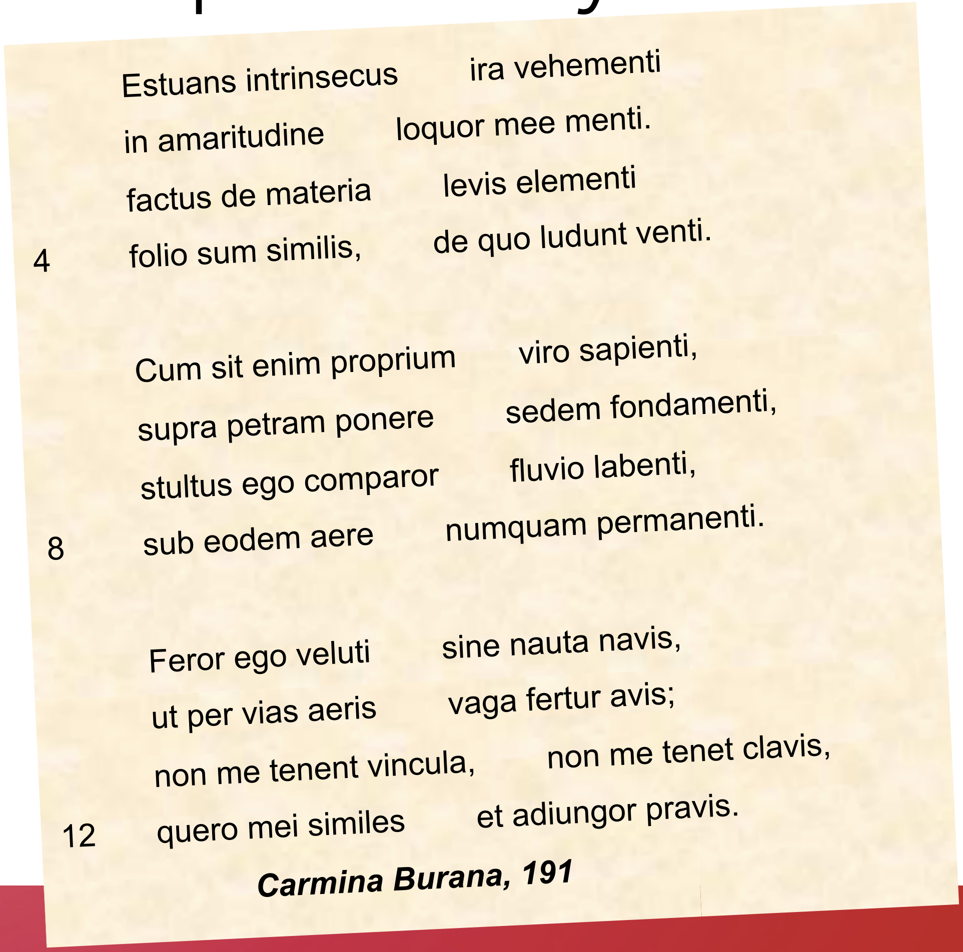
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

CulturePlex
Cultural Complexity and Digital Humanities
mlab (at)h
hyperstudio
Digital Humanities at MIT
HUMlab
Umeå University
digilab
ricerca e servizi
KING'S College LONDON
Digital Hu
Центар за дигиталне хуманистичке
ПОЧЕТНА
НОВОСТИ
ПРОЈЕКТИ
ДИГИТАЛНА ХУМАНИСТИКА
DIGHUMLAB
Digital Humanities Lab Denmark
GCDH
GÖTTINGEN CENTRE FOR DIGITAL HUMANITIES
BROWN LIBRARY
Center for Digital Scholarship
Studio@Butler
A collaboratory for educators, scholars, and librarians
CENTRE FOR E-RESEARCH
DIGITAL HUMANITIES CENTER FOR JAPANESE ARTS AND CULTURES
UNIVERSITY of VIRGINIA LIBRARY
SCHOLARS' LAB
MITH
MARYLAND INSTITUTE FOR TECHNOLOGY IN THE HUMANITIES
NULAB for texts, maps & networks
Northeastern University
Huma-Num
la TGIR des humanités numériques
臺灣大學 數位人文研究中心
Research Center for Digital Humanities
DH DIGITAL HUMANITIES initiative
THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL
Digital Innovation Lab
DARIAH-EU
CLARIN-D
centerNet
huni
Humanities Networked Infrastructure
North Pacific Ocean
United States
Russia
Ukraine
China
South Korea
Japan
Afghanistan
Iran
Iraq
Pakistan
Egypt
Chad
Nigeria
Ethiopia
Venezuela
Peru
Bolivia
Chile
Argentina
South Atlantic Ocean
Botswana
Madagascar
South Africa
Indian Ocean
Indonesia
Australia

UNED

Context - What

- The metrics on European Poetry

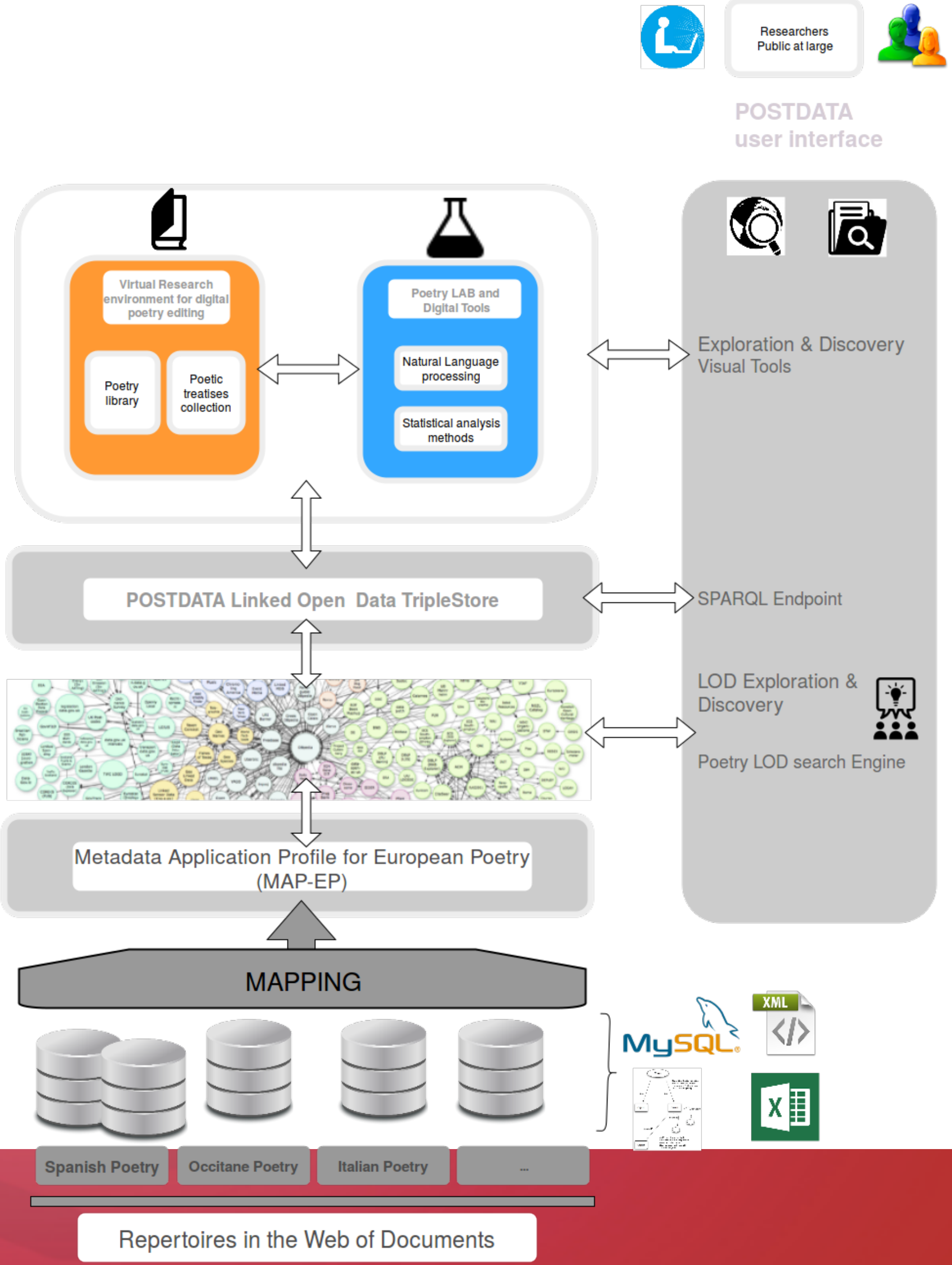


Estuans intrinsecus ira vehementi
in amaritudine loquor mee menti.
factus de materia levis elementi
folio sum similis, de quo ludunt venti.

Cum sit enim proprium viro sapienti,
supra petram ponere sedem fondamenti,
stultus ego comparor fluvio labenti,
sub eodem aere numquam permanenti.

Feror ego veluti sine nauta navis,
ut per vias aeris vaga fertur avis;
non me tenent vincula, non me tenet clavis,
quero mei similes et adiungor pravis.

***Carmina Burana*, 191**



Researchers
Public at large

POSTDATA
user interface

Virtual Research environment for digital poetry editing

Poetry library

Poetic treatises collection

Poetry LAB and Digital Tools

Natural Language processing

Statistical analysis methods

Exploration & Discovery Visual Tools

POSTDATA Linked Open Data TripleStore

SPARQL Endpoint

LOD Exploration & Discovery

Poetry LOD search Engine

Metadata Application Profile for European Poetry (MAP-EP)

MAPPING

MySQL

XML

Spanish Poetry

Occitane Poetry

Italian Poetry

...

Repertoires in the Web of Documents

The problem

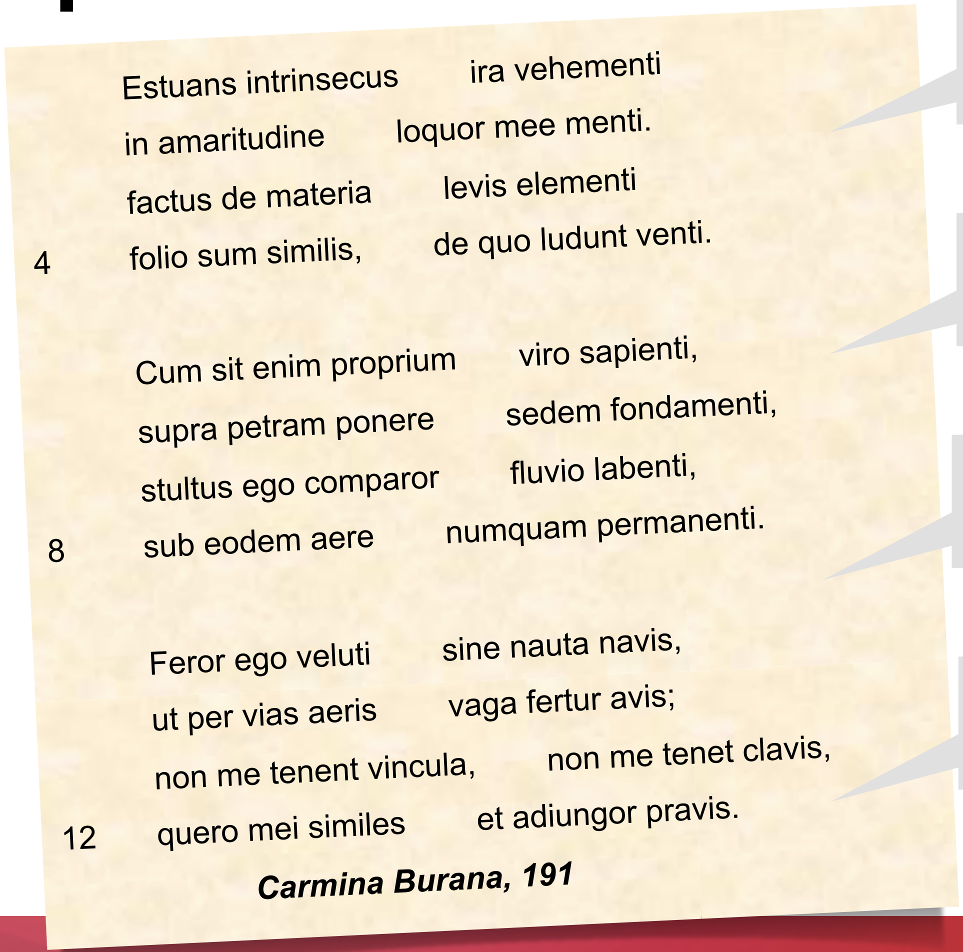
- At least 21 repertoires on Poetry metrics & other information (in the Web of Documents)
- This community wants to share all the data among repertoires
- ….to enhance its research
- And more…

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   ira vehementi
in amaritudine   loquor mee menti.
factus de materia   levis elementi
folio sum similis,   de quo ludunt venti.

Cum sit enim proprium   viro sapienti,
supra petram ponere   sedem fondamenti,
stultus ego comparor   fluvio labenti,
sub eodem aere   numquam permanenti.

Feror ego veluti   sine nauta navis,
ut per vias aeris   vaga fertur avis;
non me tenent vincula,   non me tenet clavis,
quero mei similes   et adiungor pravis.

***Carmina Burana*, 191**

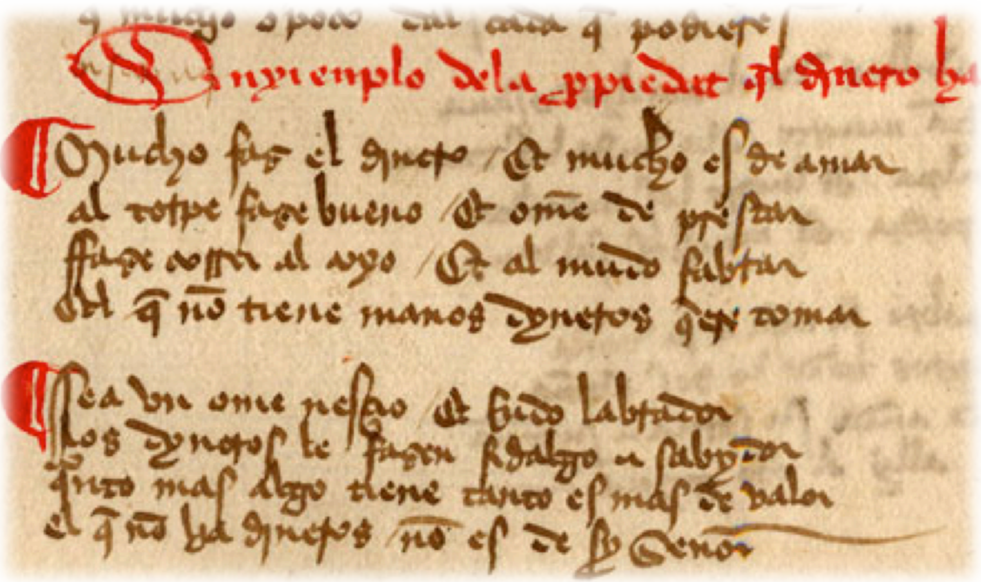
Alexandrines

Goliardic

12A12A12A12A
(Romance)

4x(7pp+7p)
(Classic Latin)

Philological standardization: starting point



Enxienplo dela ppiedat q el dinero ha

Mucho faz el dinero e mucho es de amar
al torpe faze bueno e ome de prestar
ffaze correr al coxo e al mudo fablar
el q no tiene manos dyneros q'r tomar

Sea vn ome nesçio e rudo labrador
los dyneros le fazen fidalgo e sabydor
quanto mas algo tiene tanto es mas de valor
el q no ha dyneros no es de sy señor

Author
Title
Incipit
Manuscript
Post quem
Ante quem
Language
Topics
Edition
Online edition
Work

Isometrism
Isostrophism
Metrical scheme
Rhyme scheme
Rhyme
Musical notation
Number of stanzas
Number of lines
Poetic form

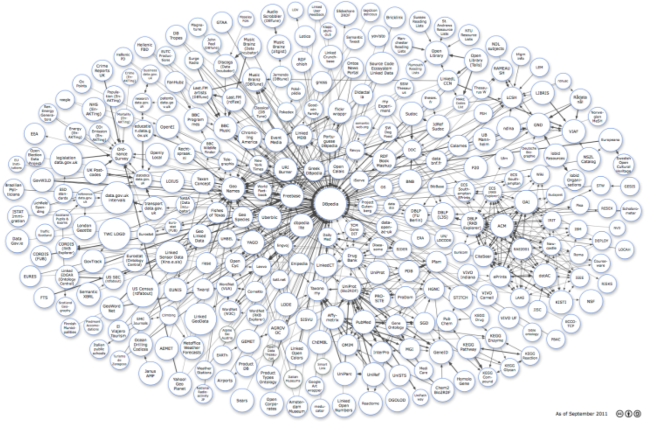
The problem

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

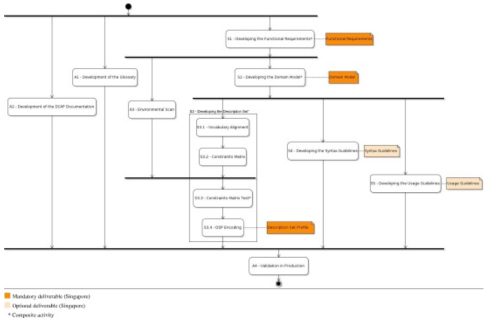
The problem

- How to overcome these diferences?
- LOD technology
- Development of a Metadata **Application Profile** for the European Poetry community



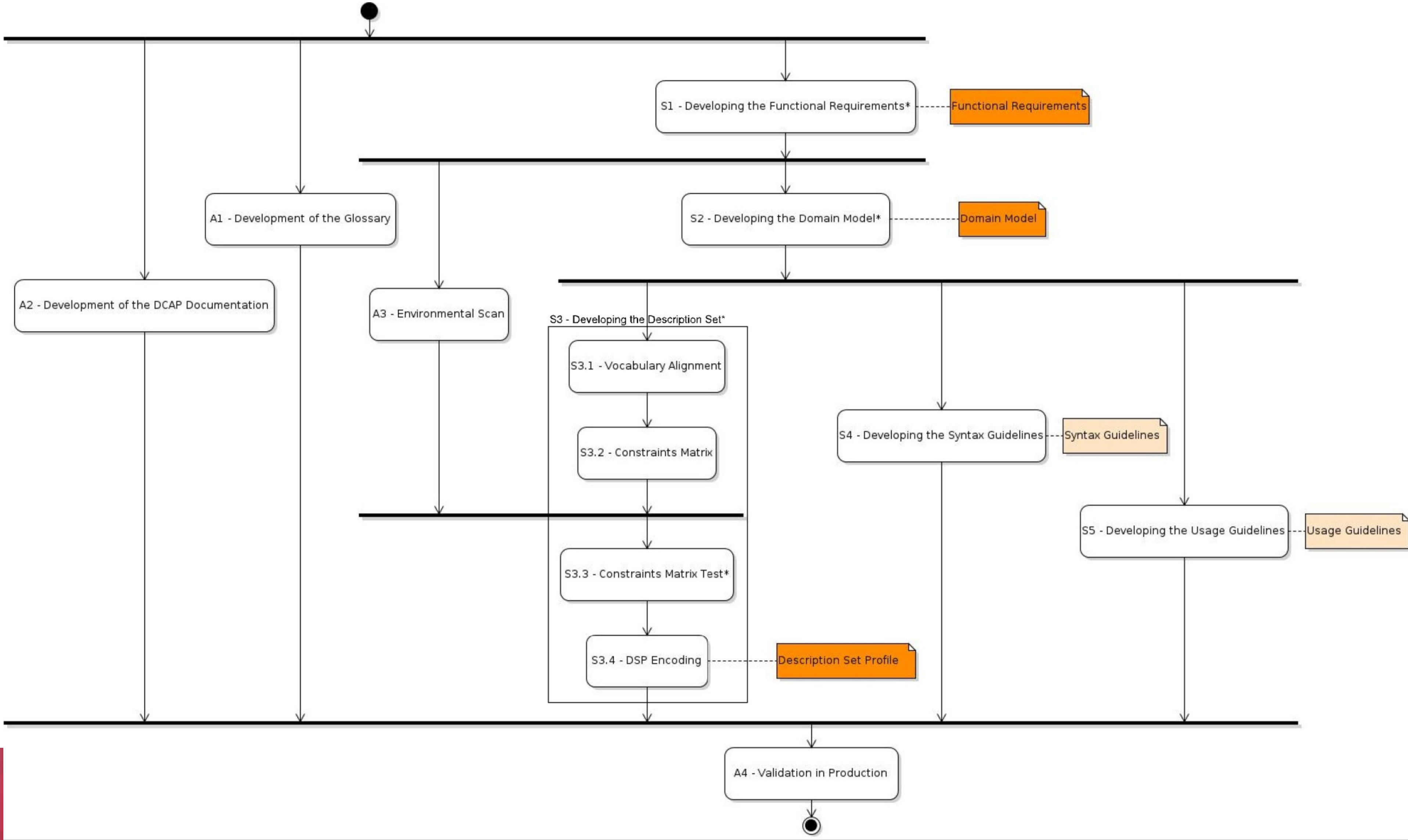
The Approach

- 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



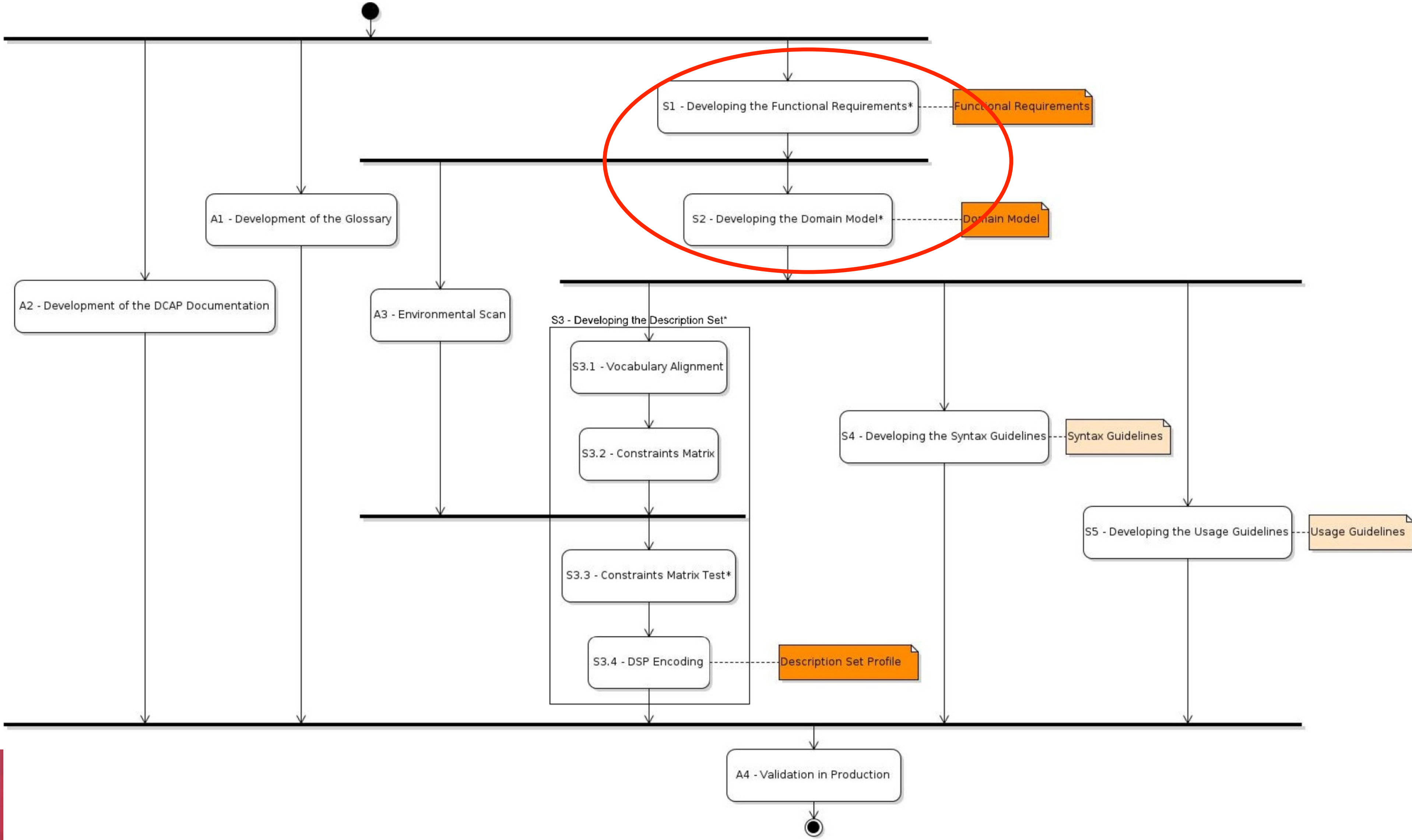
S1 - Developing the Functional Requirements*
Functional Requirements
A1 - Development of the Glossary
S2 - Developing the Domain Model*
Domain Model
A2 - Development of the DCAP Documentation
A3 - Environmental Scan
S3 - Developing the Description Set*
S3.1 - Vocabulary Alignment
S3.2 - Constraints Matrix
S4 - Developing the Syntax Guidelines
Syntax Guidelines
S5 - Developing the Usage Guidelines
Usage Guidelines
S3.3 - Constraints Matrix Test*
S3.4 - DSP Encoding
Description Set Profile
A4 - Validation in Production
🟧 Mandatory deliverable (Singapore)
▨ Optional deliverable (Singapore)
* Composite activity

The Approach



S1 - Developing the Functional Requirements*
Functional Requirements
A1 - Development of the Glossary
S2 - Developing the Domain Model*
Domain Model
A2 - Development of the DCAP Documentation
A3 - Environmental Scan
S3 - Developing the Description Set*
S3.1 - Vocabulary Alignment
S3.2 - Constraints Matrix
S4 - Developing the Syntax Guidelines
Syntax Guidelines
S5 - Developing the Usage Guidelines
Usage Guidelines
S3.3 - Constraints Matrix Test*
S3.4 - DSP Encoding
Description Set Profile
A4 - Validation in Production

Were are we now?


S1 - Developing the Functional Requirements*
Functional Requirements
A1 - Development of the Glossary
S2 - Developing the Domain Model*
Domain Model
A2 - Development of the DCAP Documentation
A3 - Environmental Scan
S3 - Developing the Description Set*
S3.1 - Vocabulary Alignment
S3.2 - Constraints Matrix
S4 - Developing the Syntax Guidelines
Syntax Guidelines
S5 - Developing the Usage Guidelines
Usage Guidelines
S3.3 - Constraints Matrix Test*
S3.4 - DSP Encoding
Description Set Profile
A4 - Validation in Production

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

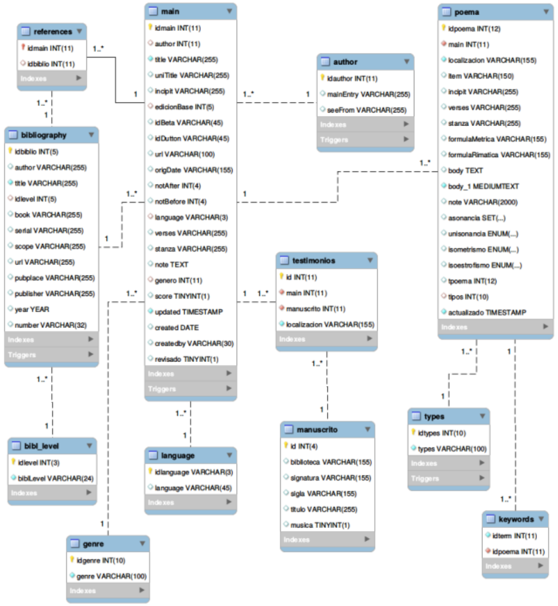
Where are we now ?

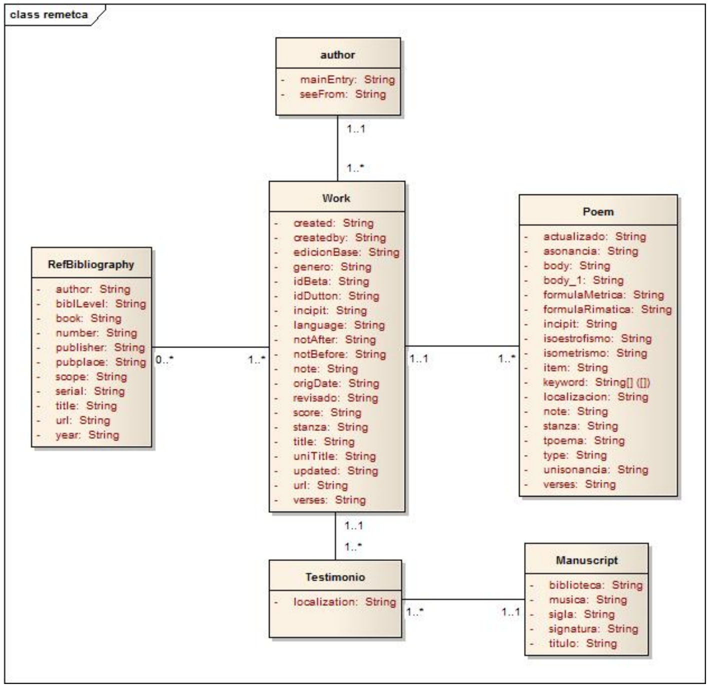
- Reverse engineering process eliminates all the details that have to do with the implementation/ representation
- We have followed the process:
  - ID keys deleted
  - Separate different concepts that are represented in the same table
  - Tables that enumerate terms deleted → become properties that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

Analysing Data Models

Relational Database

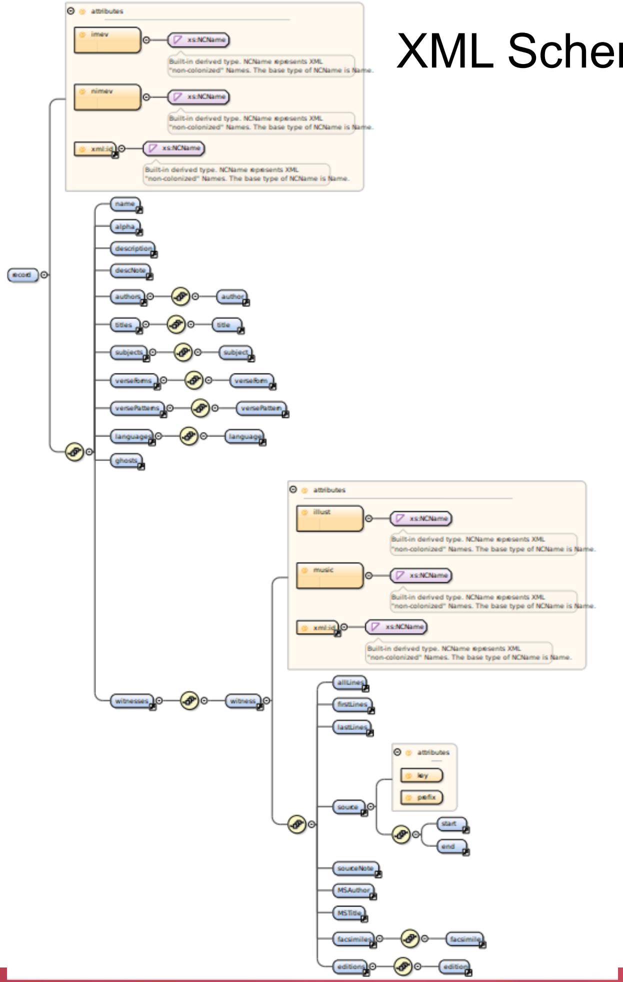
Conceptual Model



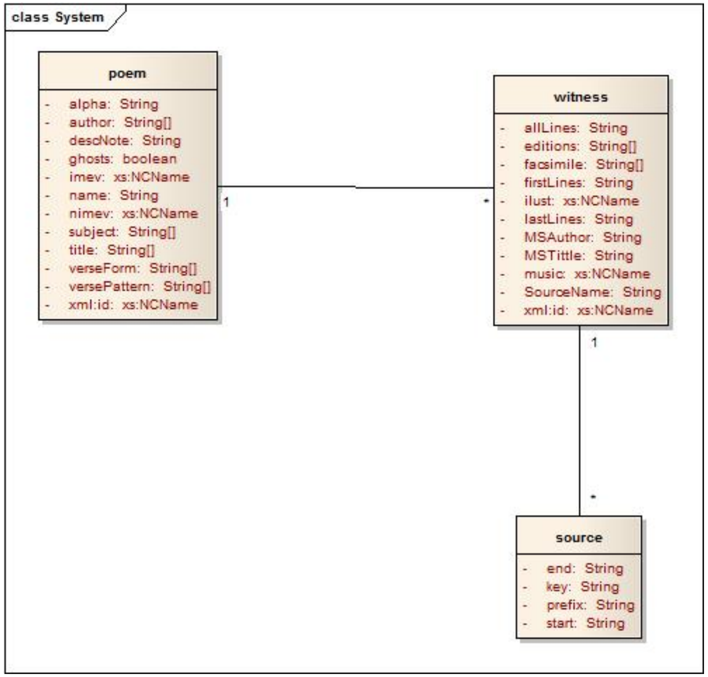


Analysing Data Models

XML Schema Model


attributes
imev
xs:NCName
Built-in derived type. NCName epresents XML "non-colonized" Names. The base type of NCName is Name.
nimev
xs:NCName
Built-in derived type. NCName epresents XML "non-colonized" Names. The base type of NCName is Name.
xml:id
xs:NCName
Built-in derived type. NCName epresents XML "non-colonized" Names. The base type of NCName is Name.
record
name
alpha
description
descNote
authors
author
titles
title
subjects
subject
verseforms
verseform
versePatterns
versePattern
languages
language
ghosts
attributes
illust
xs:NCName
Built-in derived type. NCName epresents XML "non-colonized" Names. The base type of NCName is Name.
music
xs:NCName
Built-in derived type. NCName epresents XML "non-colonized" Names. The base type of NCName is Name.
xml:id
xs:NCName
Built-in derived type. NCName epresents XML "non-colonized" Names. The base type of NCName is Name.
witnesses
witness
allLines
firstLines
lastLines
source
attributes
key
prefix
start
end
sourceNote
MSAuthor
MSTitle
facsimiles
facsimile
editions
edition

Conceptual Model



class System

poem

- alpha: String
- author: String[]
- descNote: String
- ghosts: boolean
- imev: xs:NCName
- name: String
- nimev: xs:NCName
- subject: String[]
- title: String[]
- verseForm: String[]
- versePattern: String[]
- xml:id: xs:NCName

witness

- allLines: String
- editions: String[]
- facsimile: String[]
- firstLines: String
- ilust: xs:NCName
- lastLines: String
- MSAuthor: String
- MSTittle: String
- music: xs:NCName
- SourceName: String
- xml:id: xs:NCName

source

- end: String
- key: String
- prefix: String
- start: String

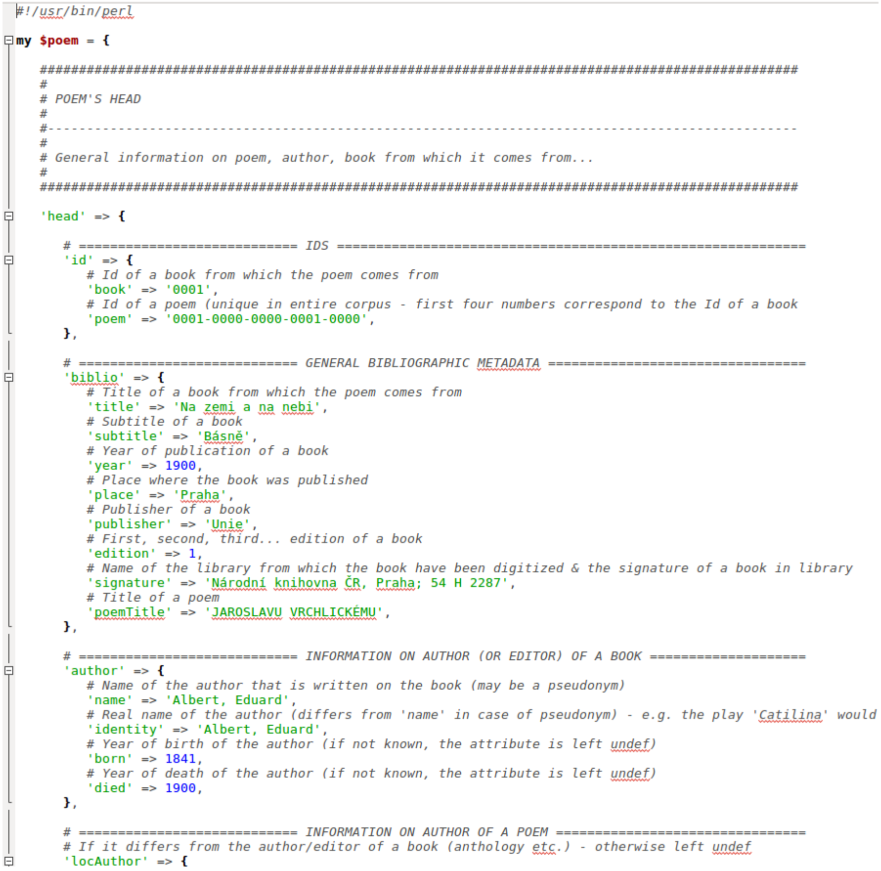
Analysing Data Models

POSTDATA
Poetry Standardization and Linked Open Data

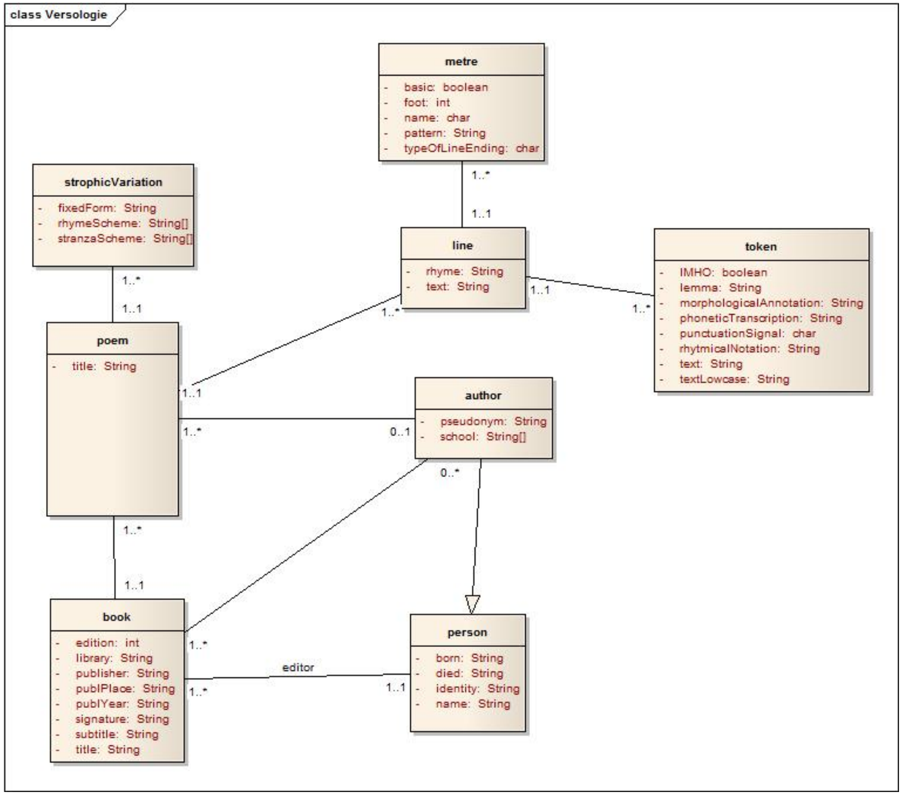
LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Perl Scritp structure

Conceptual Model



```
#!/usr/bin/perl

my $poem = {

    ##########################################################################################
    #
    # POEM'S HEAD
    #
    #-----------------------------------------------------------------------------------------
    #
    # General information on poem, author, book from which it comes from...
    #
    ##########################################################################################

    'head' => {

       # ============================ IDS ==========================================================
       'id' => {
          # Id of a book from which the poem comes from
          'book' => '0001',
          # Id of a poem (unique in entire corpus - first four numbers correspond to the Id of a book
          'poem' => '0001-0000-0000-0001-0000',
       },

       # ============================ GENERAL BIBLIOGRAPHIC METADATA ================================
       'biblio' => {
          # Title of a book from which the poem comes from
          'title' => 'Na zemi a na nebi',
          # Subtitle of a book
          'subtitle' => 'Básně',
          # Year of publication of a book
          'year' => 1900,
          # Place where the book was published
          'place' => 'Praha',
          # Publisher of a book
          'publisher' => 'Unie',
          # 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 ČR, Praha; 54 H 2287',
          # Title of a poem
          'poemTitle' => 'JAROSLAVU VRCHLICKÉMU',
       },

       # ============================ 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 pseudonym) - e.g. the play 'Catilina' 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' => 1900,
       },

       # ============================ INFORMATION ON AUTHOR OF A POEM ===================================
       # If it differs from the author/editor of a book (anthology etc.) - otherwise left undef
       'locAuthor' => {
```



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 responsibles did not provide database definition →analyse the functionalities of the Website

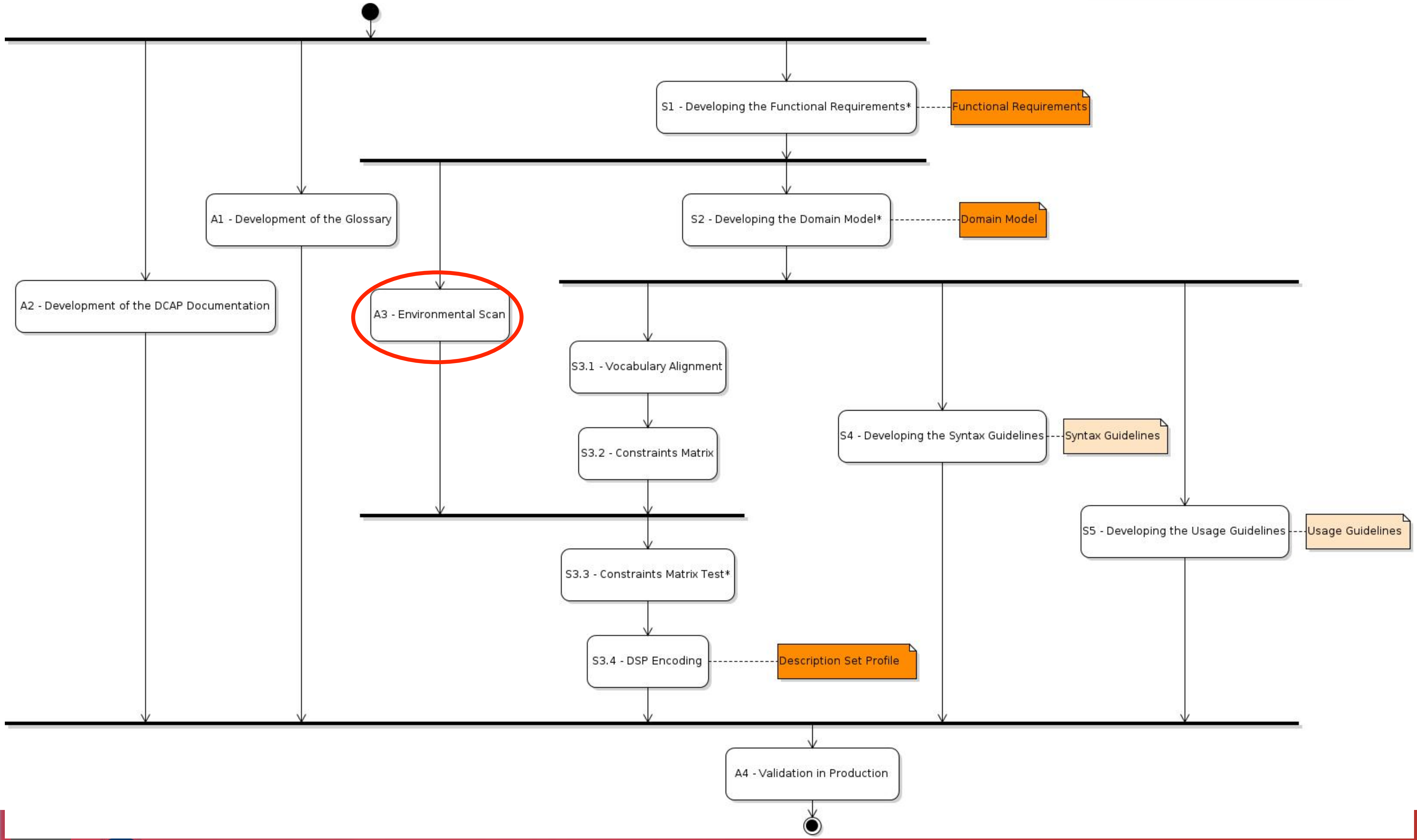
- Study the controlled vocabularies and standardize them: 1) ask for them, 2) collect them in the Websites

Where are we now ?

- Study other communities/projects for interoperability, ex:
  - ➢ Biblioteca Nacional de España - http://www.bne.es
  - ➢ Biblissima - http://www.biblissima-condorcet.fr
  - ➢ 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



S1 - Developing the Functional Requirements*

Functional Requirements

A1 - Development of the Glossary

S2 - Developing the Domain Model*

Domain Model

A2 - Development of the DCAP Documentation

A3 - Environmental Scan

S3.1 - Vocabulary Alignment

S4 - Developing the Syntax Guidelines

Syntax Guidelines

S3.2 - Constraints Matrix

S5 - Developing the Usage Guidelines

Usage Guidelines

S3.3 - Constraints Matrix Test*

S3.4 - DSP Encoding

Description Set Profile

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?

POSTDATA
Poetry Standardization
and Linked Open Data

LiNHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Mariana Curado Malta
mariana.malta@linhd.uned.es
Paloma Centenera
paloma.centendera@linhd.uned.es
Elena González-Blanco
gonzalezblanco@flog.uned.es

European Research Council
Established by the European Commission
Supporting top researchers
from anywhere in the world

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

UNED