Retos, perspectivas y próximos pasos dentro de POSTDATA

Pablo Ruiz Fabo & Helena Bermúdez Sabel
(LINHD-UNED)

Málaga, Octubre 2017

UNED

LINHD
LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

erc

POSTDATA
Poetry Standardization and Linked Open Data

Contenidos

Inicio: modelación conceptual

¿Cómo se hace?

Inicio: modelación conceptual

¿Cómo se hace?

Entidad Identificar las "cosas" del dominio

Inicio: modelación conceptual

¿Cómo se hace?

Entidad Identificar las "cosas" del dominio

Atributos Propiedades de las cosas

Inicio: modelación conceptual

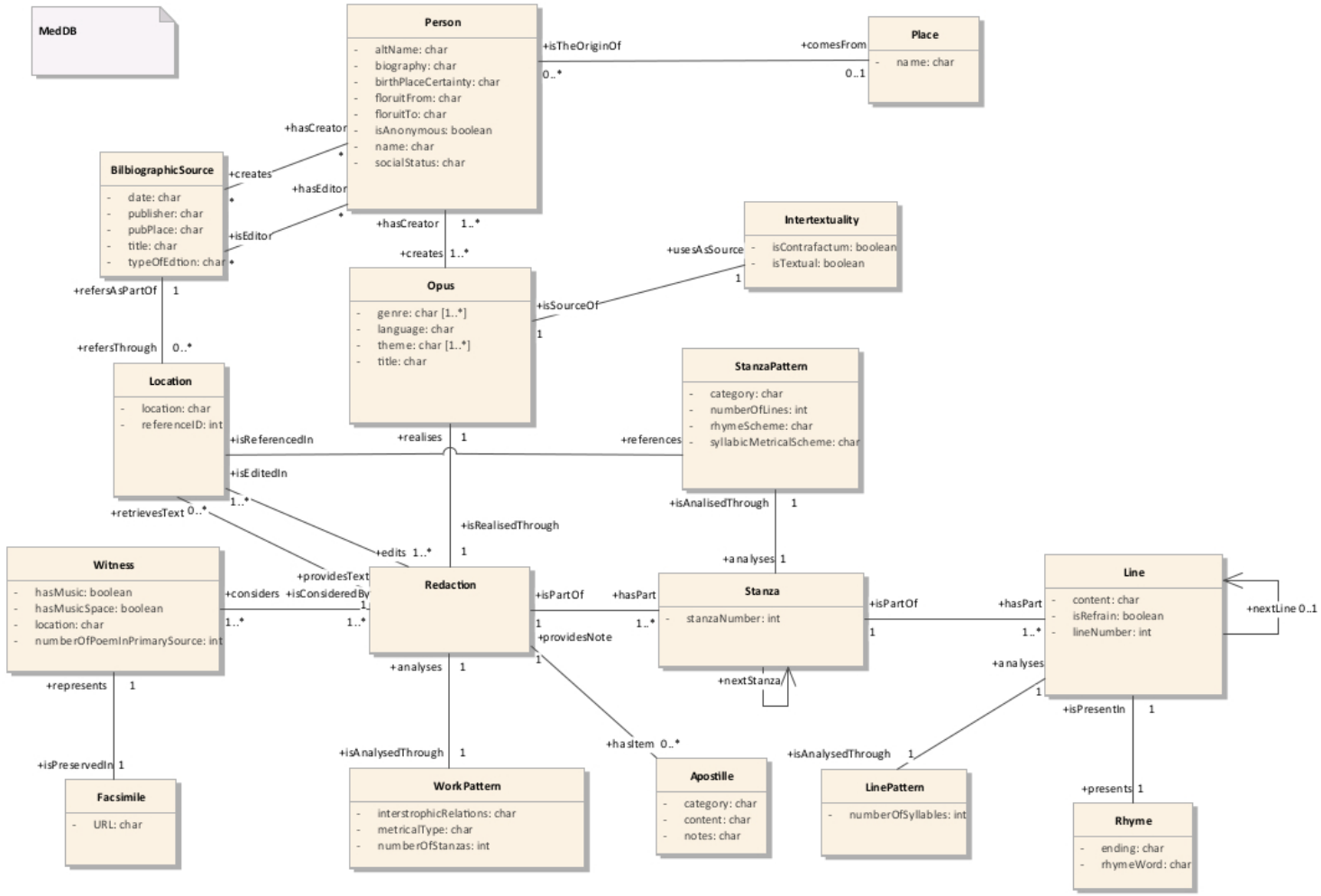
¿Cómo se hace?

Entidad Identificar las "cosas" del dominio

Atributos Propiedades de las cosas

Relaciones La manera en que las cosas se relacionan

Modelación conceptual: ejemplo con técnica UML



Modelación lógica

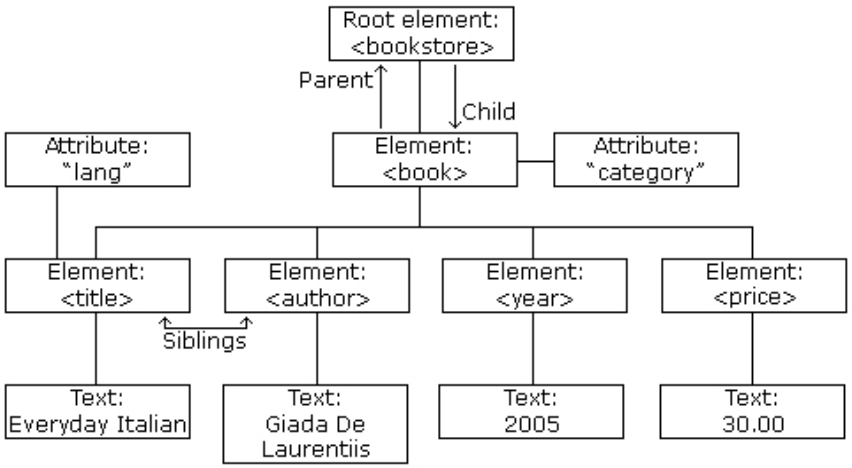
El paso siguiente, directamente relacionado con la implementación y limitaciones impuestas por algún tipo de modelación (Relacional, XML ...)

Paradigmas

- Modelación Jerárquica
- Modelación Relacional
- Modelación Orientada a Objetos
- Modelación Resource Description Framework (RDF)

Modelación Jerárquica: XML

Modelo formal basado en una jerarquía ordenada (árbol)
Lenguajes de schema: definición de elementos y atributos, y de sus valores

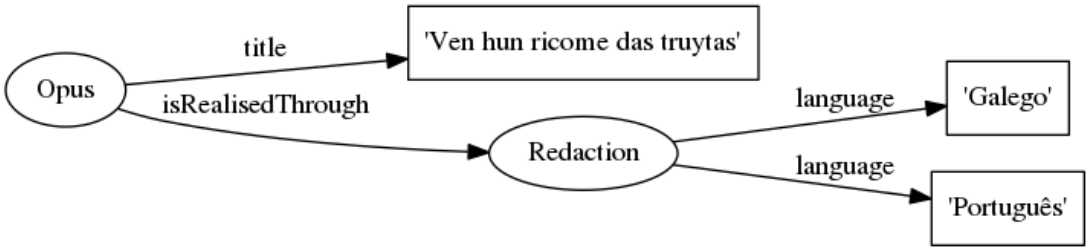



```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
  </book>
  <book category="children">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="web">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

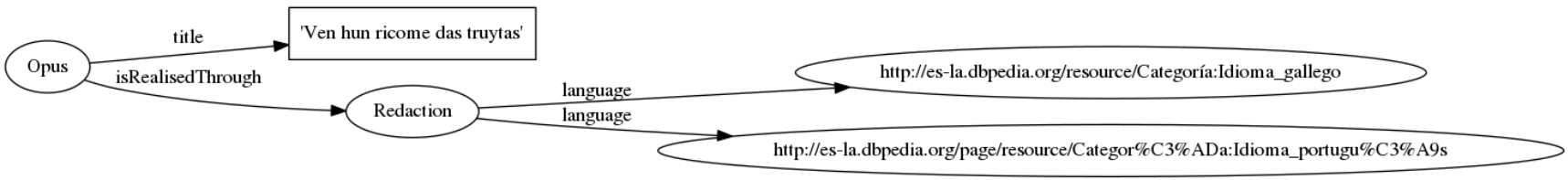
Modelación RDF: triple

Modelación: ejemplo tripleta

Modelación RDF: ejemplo tripleta

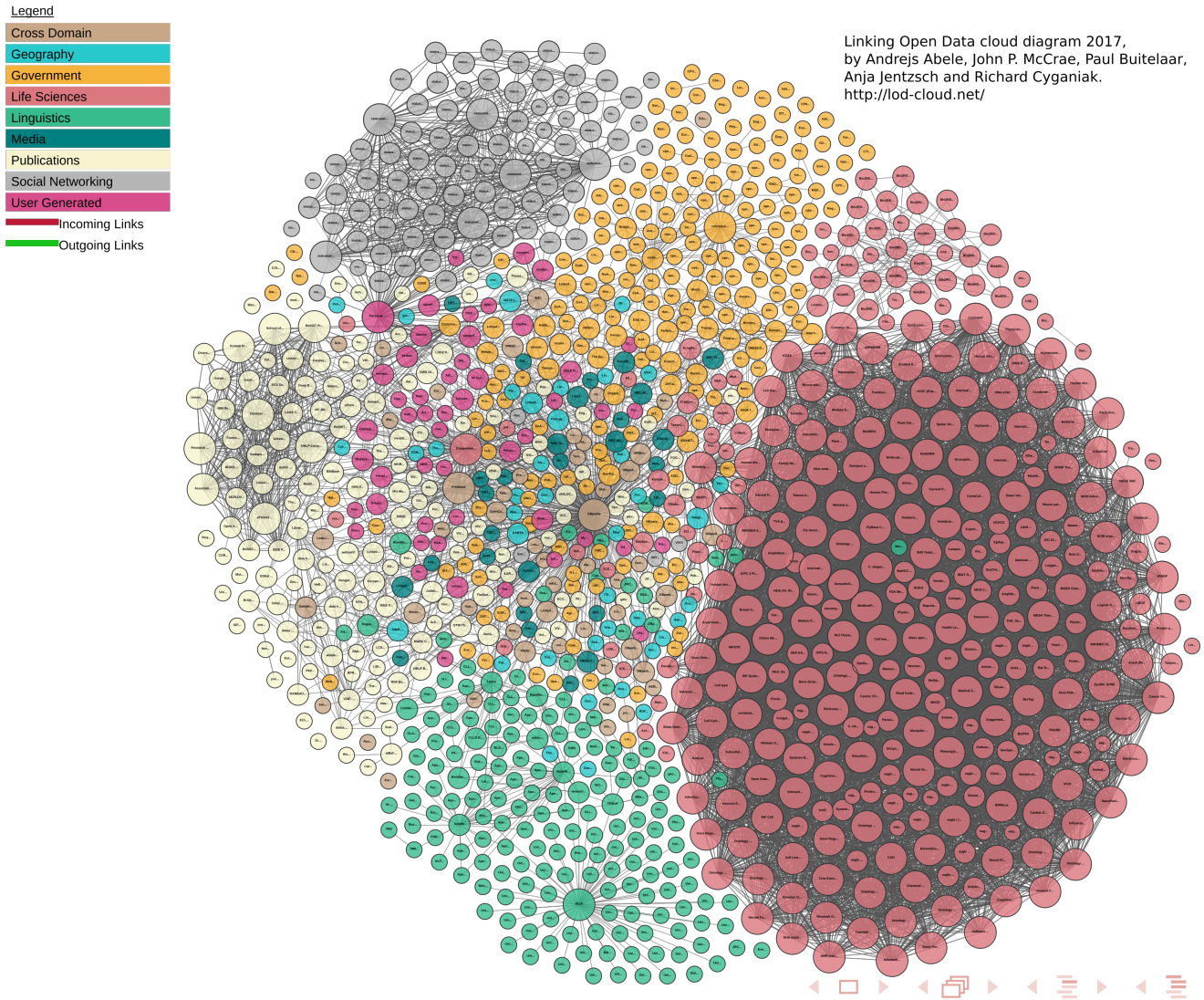


Modelación RDF: ejemplo tripleta



Del modelo jerárquico (XML) al modelo en red (RDF)

Linked Open Data: LOD



Propuesta

Nuestro artefacto: Perfil de Aplicación de Metadatos (MAP)

¿Qué es?

Un modelo semántico

¿Qué es?

Un modelo semántico

- Modelo de datos con equivalencias a términos de vocabularios RDF,
- incluyendo la definición de restricciones.
- Un MAP es un *esquema* que organiza conceptos, y define estructura y semántica.
- El esquema MAP presenta las propiedades, en el modelo de datos RDF, utilizando términos de vocabularios RDF.

Objetivo: costumización TEI

Punto de partida: esquema MAP (y no el modelo conceptual)

- el esquema XML tiene una estructura más similar al modelo de tripletas RDF que al modelo conceptual

Objetivo: costumización TEI

Punto de partida: esquema MAP (y no el modelo conceptual)

- el esquema XML tiene una estructura más similar al modelo de tripletas RDF que al modelo conceptual
- el modelo conceptual tiene semántica pero la del modelo RDF es más exacta

Objetivo: costumización TEI

Punto de partida: esquema MAP (y no el modelo conceptual)

- el esquema XML tiene una estructura más similar al modelo de tripletas RDF que al modelo conceptual
- el modelo conceptual tiene semántica pero la del modelo RDF es más exacta
- las restricciones de sintaxis y valores son definidos únicamente en el esquema MAP

Contenidos

Ejes de POSTDATA

Infraestructura de web semántica

LOD: Datos abiertos enlazados

Entorno Virtual de Investigación

Edición crítica digital

Poetry Lab

Procesamiento del Lenguaje Natural

¿Qué es el Poetry Lab?

Espacio con herramientas para automatizar análisis de poesía

¿Por qué? ¿Cómo?

- Existen técnicas de Procesamiento del lenguaje natural (PLN) que permiten anotar automáticamente características lingüísticas relevantes para el análisis literario
- Ejemplos de características:
  - Fonemas
  - Categorías gramaticales
  - Funciones sintácticas

¿Por qué? ¿Cómo?

- Existen técnicas de Procesamiento del lenguaje natural (PLN) que permiten anotar automáticamente características lingüísticas relevantes para el análisis literario
- Ejemplos de características:
  - Fonemas: Patrones de sonidos
  - Categorías gramaticales
  - Funciones sintácticas

¿Por qué? ¿Cómo?

- Existen técnicas de Procesamiento del lenguaje natural (PLN) que permiten anotar automáticamente características lingüísticas relevantes para el análisis literario
- Ejemplos de características:
  - Fonemas: Patrones de sonidos
  - Categorías gramaticales: Predominancia de verbos vs. sustantivos
  - Funciones sintácticas

¿Por qué? ¿Cómo?

- Existen técnicas de Procesamiento del lenguaje natural (PLN) que permiten anotar automáticamente características lingüísticas relevantes para el análisis literario
- Ejemplos de características:
  - Fonemas: Patrones de sonidos
  - Categorías gramaticales: Predominancia de verbos vs. sustantivos
  - Funciones sintácticas: Sujetos que actúan en un texto, objetos afectados

Áreas de trabajo

Tarea

- Patrones fonológicos
- Escansión silábica

Áreas de trabajo

Tarea	Tecnologías
• Patrones fonológicos • Escansión silábica	• Transcripción automática • Normalización ortográfica • Etiquetado gramatical

Áreas de trabajo

Tarea	Tecnologías
• Patrones fonológicos • Escansión silábica	• Transcripción automática • Normalización ortográfica • Etiquetado gramatical
• Análisis temático	

Áreas de trabajo

Tarea	Tecnologías
• Patrones fonológicos • Escansión silábica	• Transcripción automática • Normalización ortográfica • Etiquetado gramatical
• Análisis temático	• Semántica distribucional • Extracción de entidades nombradas

Áreas de trabajo

Tarea	Tecnologías
• Patrones fonológicos • Escansión silábica	• Transcripción automática • Normalización ortográfica • Etiquetado gramatical
• Análisis temático	• Semántica distribucional • Extracción de entidades nombradas
• Otros rasgos de estilo	

Áreas de trabajo

Tarea	Tecnologías
• Patrones fonológicos • Escansión silábica	• Transcripción automática • Normalización ortográfica • Etiquetado gramatical
• Análisis temático	• Semántica distribucional • Extracción de entidades nombradas
• Otros rasgos de estilo	• Etiquetado gramatical • Análisis sintáctico automático

Resultados

Herramientas desarrolladas para automatización de análisis:

- Detección de entidades nombradas en español medieval (Análisis temático)
- Detección de encabalgamiento (Análisis de estilo)

Detección de entidades en español medieval

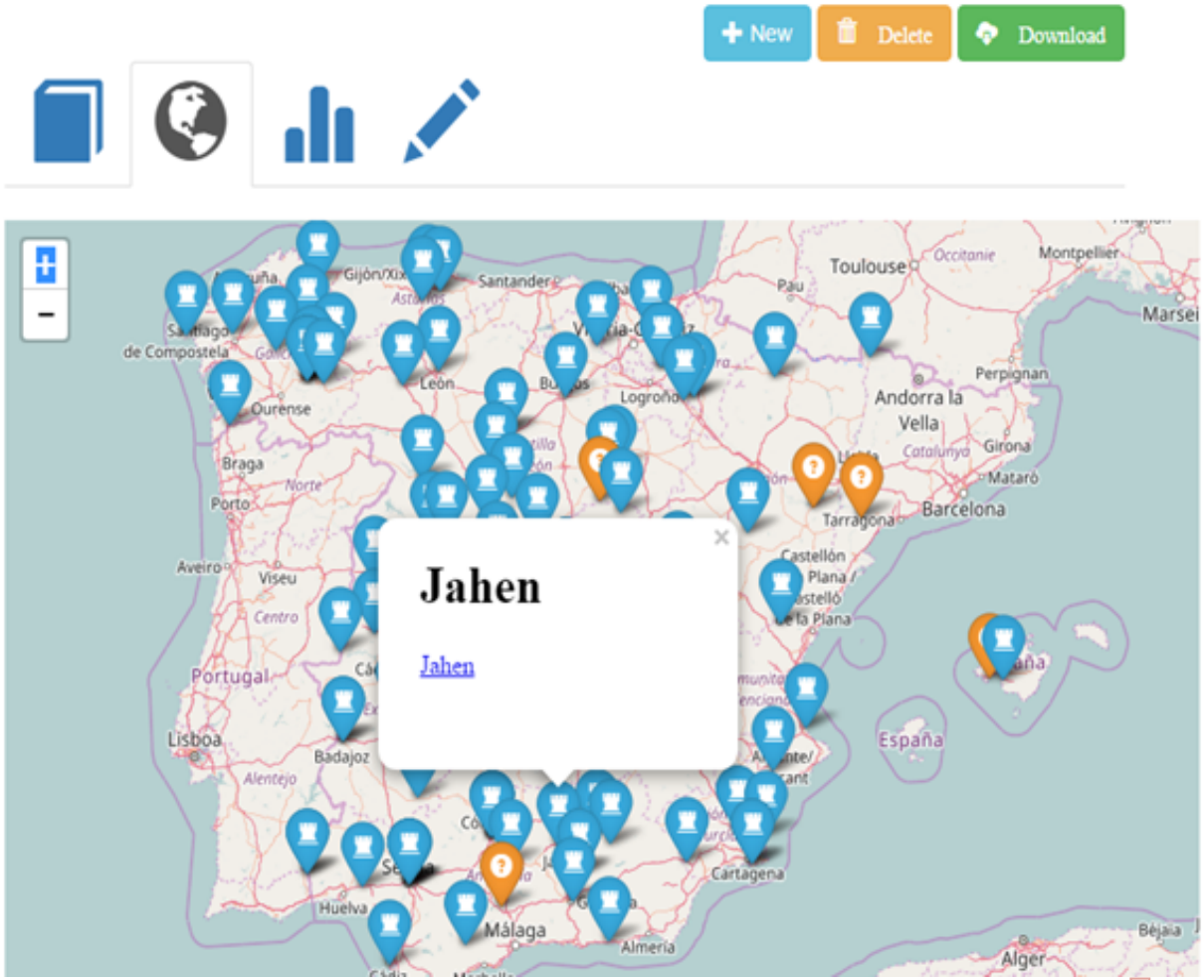
HisMeTag: Hispanic **Me**dieval **Tag**ger (Díez Platas et al., 2017)

Demo **ahora** en Polo Digital **Sala 21**

- Detección de nombres de persona, lugares, organizaciones y "roles" en español medieval
- Normalización de variedades históricas
- Salida en TEI

Hispanic Medieval Tagger

Geolocalización permite ver lugares mencionados en una obra



Detección de encabalgamiento

ANJA: **A**utomatic e**NJ**ambment **A**nalyzer
(Ruiz Fabo, Martínez Cantón et al., 2017)

Ponencia **16:30** en Polo Digital **Sala 21**

- Encabalgamiento: Definido según categorías gramaticales y constituyentes + funciones sintácticas (Quilis, 1964, Spang, 1983)
- Definición se presta a su detección basándose en Procesamiento del lenguaje natural

ANJA Online

http://prf1.org/anja/index/

AUTOMATIC ENJAMBMENT ANALYSIS IN SPANISH

Text

1	Vino primera frívola -yo niño con ojeras-
2	y nos puso en los dedos un sueño de esperanza
3	o alguna perversión: sus velos y su danza
4	le ceñían las sílabas, los ritmos, las caderas.
5	Mas quisimos su cuerpo sobre las escombreras
6	porque también manchase su ropa en la tardanza
7	de luz y libertad: esa tierna venganza
8	de llevarla por calles y lunas prisioneras.

Standoff | Inline | PosTags | FullNLP

legend

Start	End	Enjambment Type
3	4	ex_subj_verb
6	7	pb_noun_prep
7	8	pb_noun_prep

LINHD LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES

Retos

- Procesamiento del Lenguaje Natural

- Humanidades Digitales

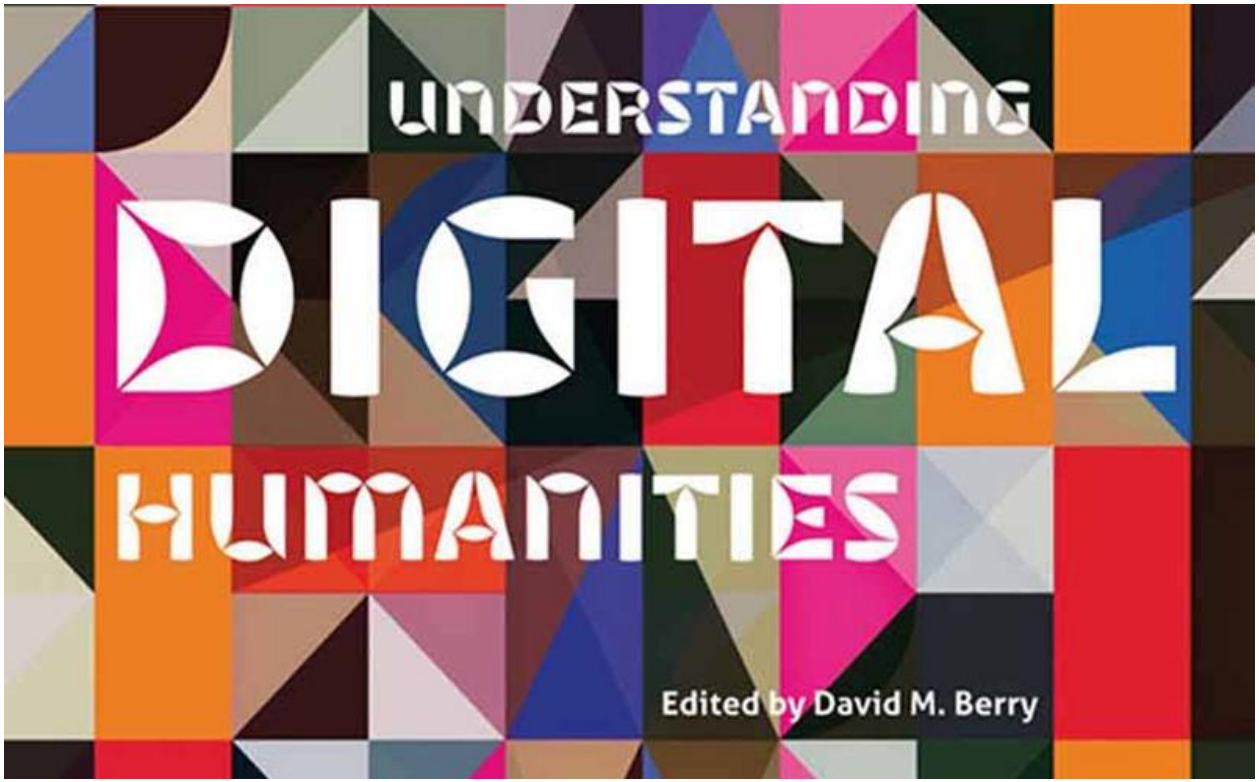
Retos

- Procesamiento del Lenguaje Natural
  - Complejidad del lenguaje poético
  - Variedades históricas: Escasez de recursos lingüísticos
- Humanidades Digitales

Retos

- Procesamiento del Lenguaje Natural
  - Complejidad del lenguaje poético
  - Variedades históricas: Escasez de recursos lingüísticos
- Humanidades Digitales
  - Interpretación de resultados cuantitativos
  - Obtenidos con medios automáticos

Interpretación



Interpretación

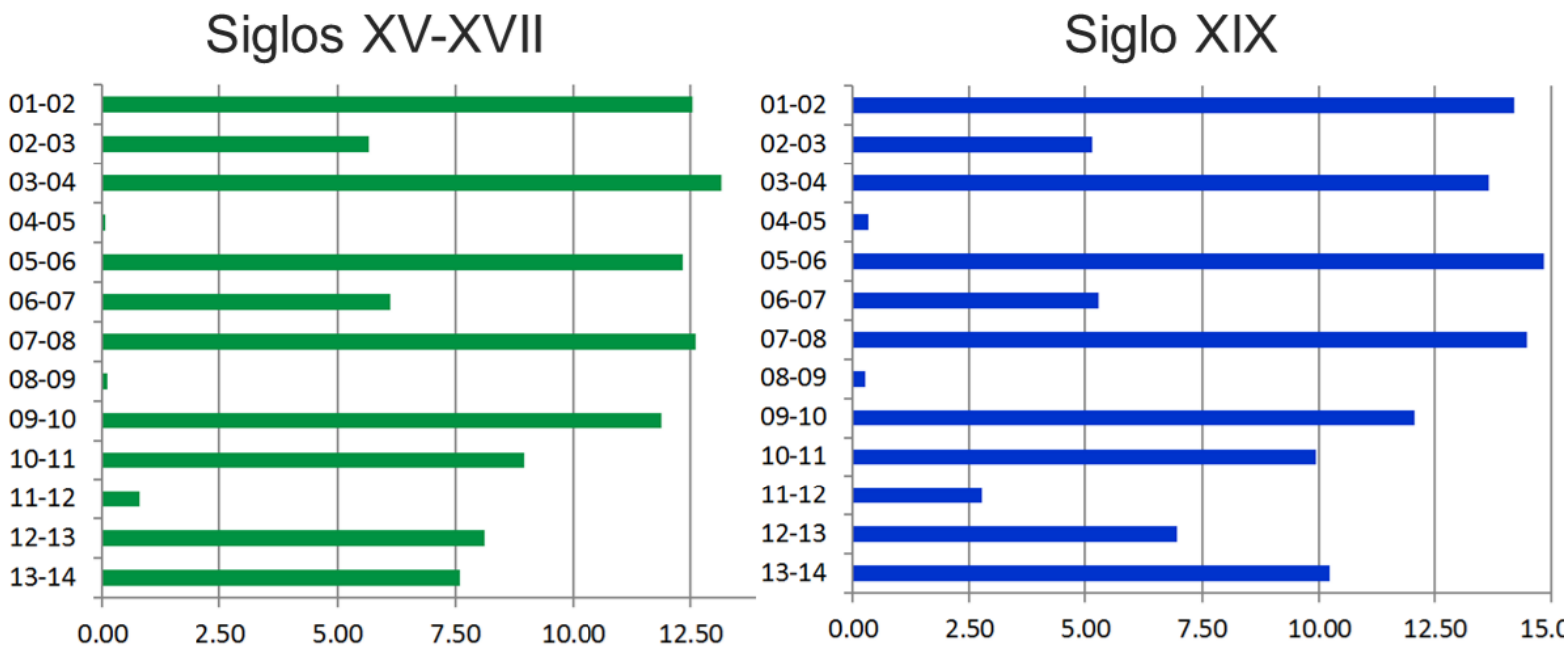
4

Digital Methods: Five Challenges

Bernhard Rieder and Theo Röhle

While the use of computers for humanities and social science research has a long history[1], the immense success of networked personal computing has made both physical machines and software more accessible to scholars. But even more importantly, digital artifacts now populate every corner of post-industrial societies. This means that besides the study of non-digital objects and phenomena with the help of computers, there now is a continuously expanding space of cultural production and social interaction riddled by machine mediation, which has been, from the beginning, tied to digital schemes and formats. An obvious effect of this expansion has been the explosion of material available in digital form. 'Traditional' cultural artifacts like books or movies, 'native' digital forms such as software programs, online publications or computer games, and

Interpretación de anotaciones automáticas



Porcentaje de encabalgamientos por par de versos en sonetos

Resumen

Gracias

Pablo Ruiz Fabo
pablo.ruiz@linhd.uned.es
Helena Bermudez Sabel
helena.sabel@linhd.uned.es