Concept-based Organization for semi-automatic Knowledge Inference in Digital Humanities: Modelling and Visualization

Ángel Castellanos, **Ana García-Serrano**, Juan Cigarrán

ETSI Informática, UNED

NLP Group

Natural Language Processing and Information Retrieval Group at UNED

Corpus AGS

- Digital collection of maps, plans and drawings of the *Archivo General de Simancas* (AGS).

  http://www.mcu.es/ccbae/es/consulta/resultados_busqueda.cmd?tipo_busqueda=mapas_planos_dibujos&posicion=1&id=30485

- The data are provided in textual cards (7792 cards), one per each item (map, plan or draw) in the collection.

Corpus AGS

http://www.mcu.es/ccbae/es/consulta/resultados_busqueda.cmd?tipo_busqueda=mapas_planos_dibujos&posicion=1&id=30485

Sección:	Material gráfico AGS
Número de control:	BAB20100018941
Autor:	Huet, Luis
Título:	Plano y perfiles que manifiestan el estado en que se alla la Real obra de Fuerte-Príncipe en 30 d de 1779 [Material gráfico no proyectable] / [rúbrica] Luis Huet
Área de datos:	Escala [ca.1:816], 200 varas reales [=20,6 cm]
Publicación:	Habana, 30 de junio de 1779
Descripción física:	1 plano : ms., col. ; 36,5 x 47 cm
Notas:	Referencias: Mapas, planos y dibujos (Años 1503-1805). Volumen I : p. 576 Tinta y colores a la aguada ocre y encarnado. Explicación con clave alfabética Manuscrito sobre papel. AGS. Secretaría de Guerra, Legajos, 03222. Acompaña a carta y relación delas obras de don Lu al Conde de Ricla de 1 de julio y de la misma fecha del plano
Materia / geográfico:	Fortificaciones-La Habana-Dibujos La Habana-Edificios, estructuras, etc.-Dibujos
Género / forma:	Dibujos de arquitectura-España-S.XVIII

Ficha: 176927

Corpus AGS

- Spanish Project (HAR2012-31117)

  El dibujante ingeniero al servicio de la monarquía hispánica. Siglos XVI-XVIII (DIMH)
  http://dimh.hypotheses.org/

Main goals of the project are:

- Knowledge organization of the cards contents
  - Identification of data relationships
  - Visualization of the results
- In order to support the research of the historian researchers of the project

Corpus DIMH

- The data provided in the textual cards (7792) have been pre-processed in order to:
  - Convert the cards from RDF:DC to XML.
  - Identify the named entities
  - Identify the nominal groups
  - Identify the lemmas
- Not supervised process

Corpus DIMH: Enrichment using Linguistic Resources (not supervised)

<Ficha id="176927">
<Tematica> La Habana-Edificios, estructuras, etc.-Dibujos</Tematica>
<Materia>-La Habana</Materia>
<Materia>-España</Materia>
<Formato>image/jpeg</Formato>
<Idioma>spa</Idioma>
<Materia>La Habana</Materia>
<Materia>-S.XVIII</Materia>
<Notas> AGS. Secretaría de Guerra, Legajos, 03222. Acompaña a carta y relación delas obras de don Luis Huet al Conde de Ricla de 1 de julio y de la misma fecha del plano</Notas>
<Creador>Huet, Luis</Creador>
<Publicacion>1779</Publicacion>
<Notas> Manuscrito sobre papel.</Notas>
<Materia>Cuba-La Habana</Materia>
… </Ficha>

<nes>La_Habana Fortificaciones-Dibujos_Dibujos Dibujos Luis_Huet Huet AGS Cuba España Real_Sociedad Real Conde_de_Ricla Fuerte-Príncipe Luis_La_Habana </nes>
<nes_person>Luis_Huet Huet Fuerte-Príncipe</nes_person>
<nes_organization>Real_Sociedad Real</nes_organization>
<nes_location>La_Habana Cuba España </nes_location>
<nes_misc>Fortificaciones-Dibujos_Dibujos Dibujos AGS Conde_de_Ricla Luis_La_Habana</nes_misc>
<lemas> plano y perfil que manifestar el estado en que se allá el Real obrar de Fuerte - Príncipe en 30 de Junio de 1779 … </lemas>
<sintagmas> plano_perfil real fuerte_principe junio_1779_material_grafico proyectable_huet luis_habana_edificio dibujos
… </Ficha>

Ficha: 176927

Main Goal

- End-Users interested in **hidden data** and/or **data relationships**

  - Knowledge Extraction and
  - Latent Organization Discovering

- Available techniques and technologies
  - Ontologies and LOD resources
  - Quantitative approaches

Motivation of the Work

Probabilistic techniques as Latent Dirichlet Allocation (LDA), have become almost a standard for the **Content Organization** in the Digital Humanities (DH) (Meeks & Weingart 2012, Yin et al. 2013, …)

It suffers from:

- the need to fix the number of topics to be detected or
- the non-trivial interpretation by the humanists.

Our approach is based on the Formal Concept Analysis (FCA) that takes advantage of a well-founded mathematical background

FCA for Content Organization

- **Formal Concept Analysis is a priori solution because:**
  - Allows the organization of objects according to their shared attributes into a generalization-specification relationship.
  - Organize the latent structure according to the shared terms in thematic-based concepts.
  - Generate a hierarchical structure (Lattice) allowing its navigation and visualization**.**

But, We need to develop our own **framework**

FCA at Glance

Formal Context *is* a trip (G, M, I), where

- G is a set of (formal)
- M a set of (formal) d
- I

i.e. ($I \subseteq$ ead as: the object g has t

The main con ***rmal Concept***, a pair (A, B) w G is a se (*extent*) and $B \subseteq M$ is t mal set of attribut *nt*) shared by all the obj s in A.

A *Formal Context* is an incidence matrix that indicates whether or not an attribute is related to an object.

FCA at a Glance

Formal Context is a tri… $:= (G, M, I)$, where

- G is a set of (forma… *ts*,
- M a set of (form… s and

A Formal Concept is a set of objects sharing a set of attributes

i.e. (I … is read as: the object g has …

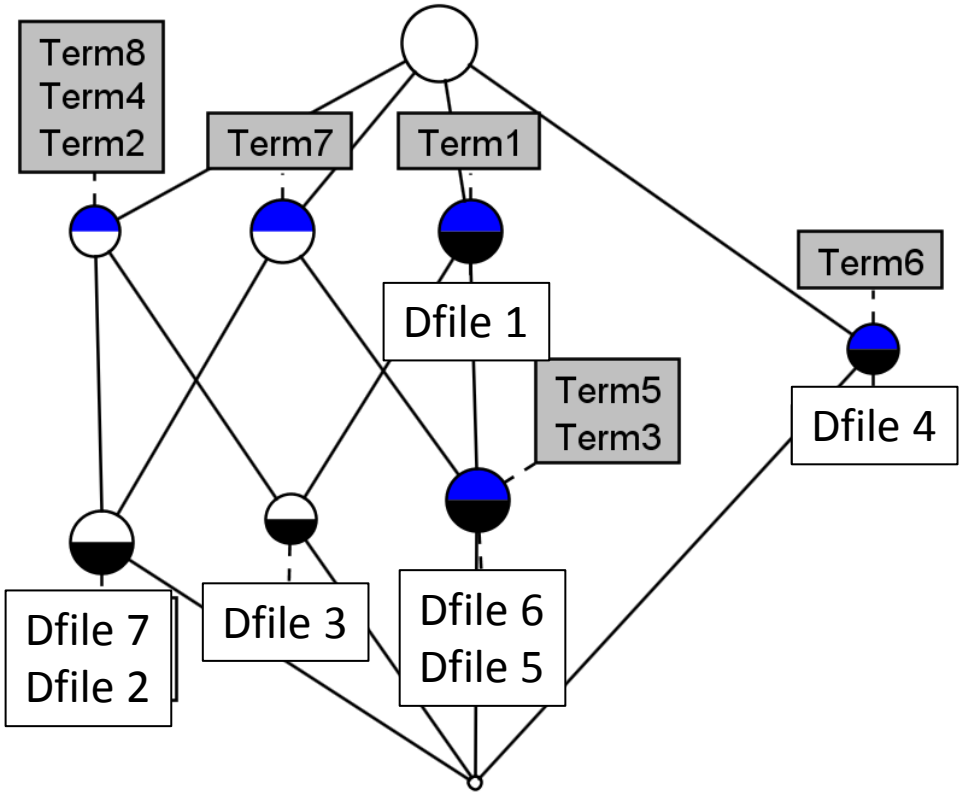
The main … the t… ***Formal Concept***, a pair $(A, B$… ere $A \subseteq G$ is a set … cts (*extent*) and $B \subseteq M$… the maximal set of attributes (…*tent*) shared by all the objects in A.

FCA at a Glance

Formal Concepts can be formally ordered in a **subconcept-superconcept-relation** according to their extents:

where (C, and,
conversely, B) is more
specific tha e proven
to be a *latt*

Since *conce* be
displayed i



FCA application to DIMH corpus

Main goals:

- how FCA performs for a content modelling task and
- whether the obtained model infers new knowledge from the original data.

The steps in the FCA application for content modelling are:

1. **Information Extraction:** It extracts the data in the processed AGS files (DIMH corpus)).

   Selected Metadata: publication, reference, notes, named entities, topic, material and title.

FCA application to DIMH corpus

1. **Information Extraction**
2. **Formal Context Generation:** The *objects* are the AGS files and the *attributes* are the selected terms related to the files.
3. **Formal Context Reduction:** The formal context generated in the previous step includes redundant and not valuable information.

   The formal context reduction takes only those features in the *formal context* that allow the identification of more relationships among the cards, avoiding information loss.

FCA application to DIMH corpus

1. **Information Extraction**
2. **Formal Context Generation**
3. **Formal Context Reduction**
4. **FCA execution:** formal concepts and its hierarchical structure generation.

   The FCA algorithm is based on a self-implemented version of the Next Neighbourhood algorithm (Carpineto & Romano 2004).

Applied Research: Own ToolKit

- Modelling step
  - Library for KLD Pre- Filtering
  - FCA Algorithm (Jbrainded, la4j for huge matrix)
  - Graph Representation
    (Neo4j, graph DB based on Lucene)
  - Navigation Algorithms (Code for graph navigation)

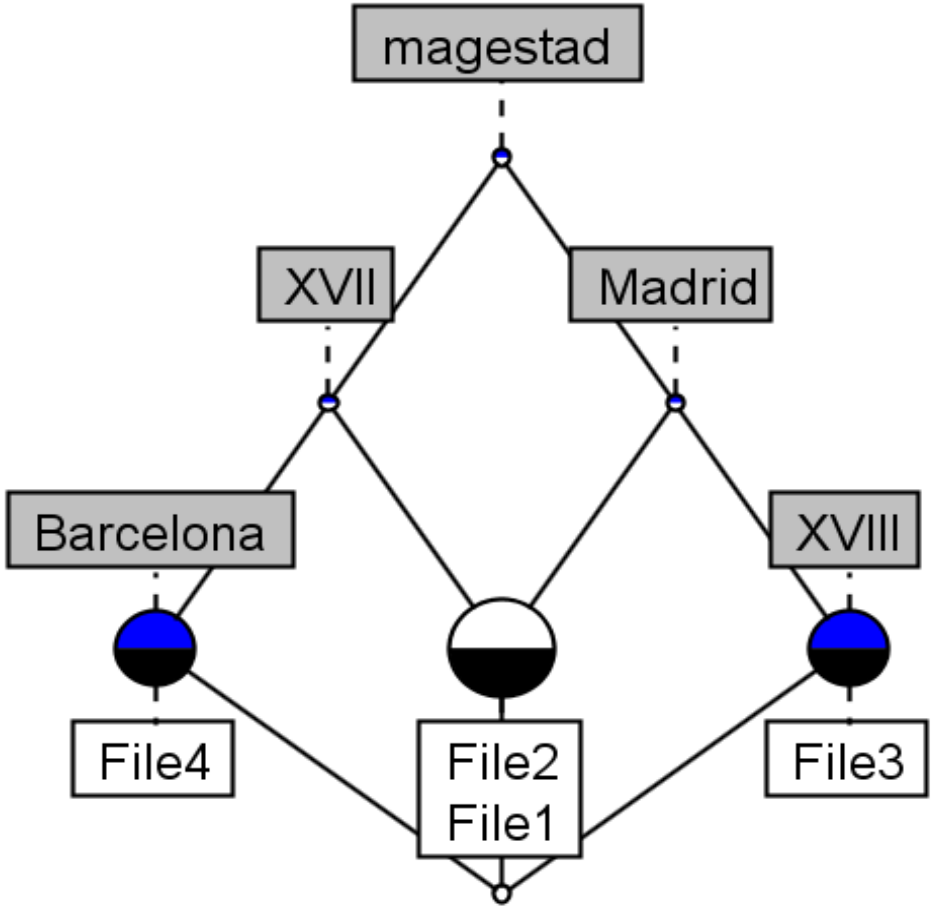
- Interface to facilitate the interaction within the Lattice
  (Alpha version using Google Web Toolkit)

- Tools for Lab experimentation (refinement, evaluation)

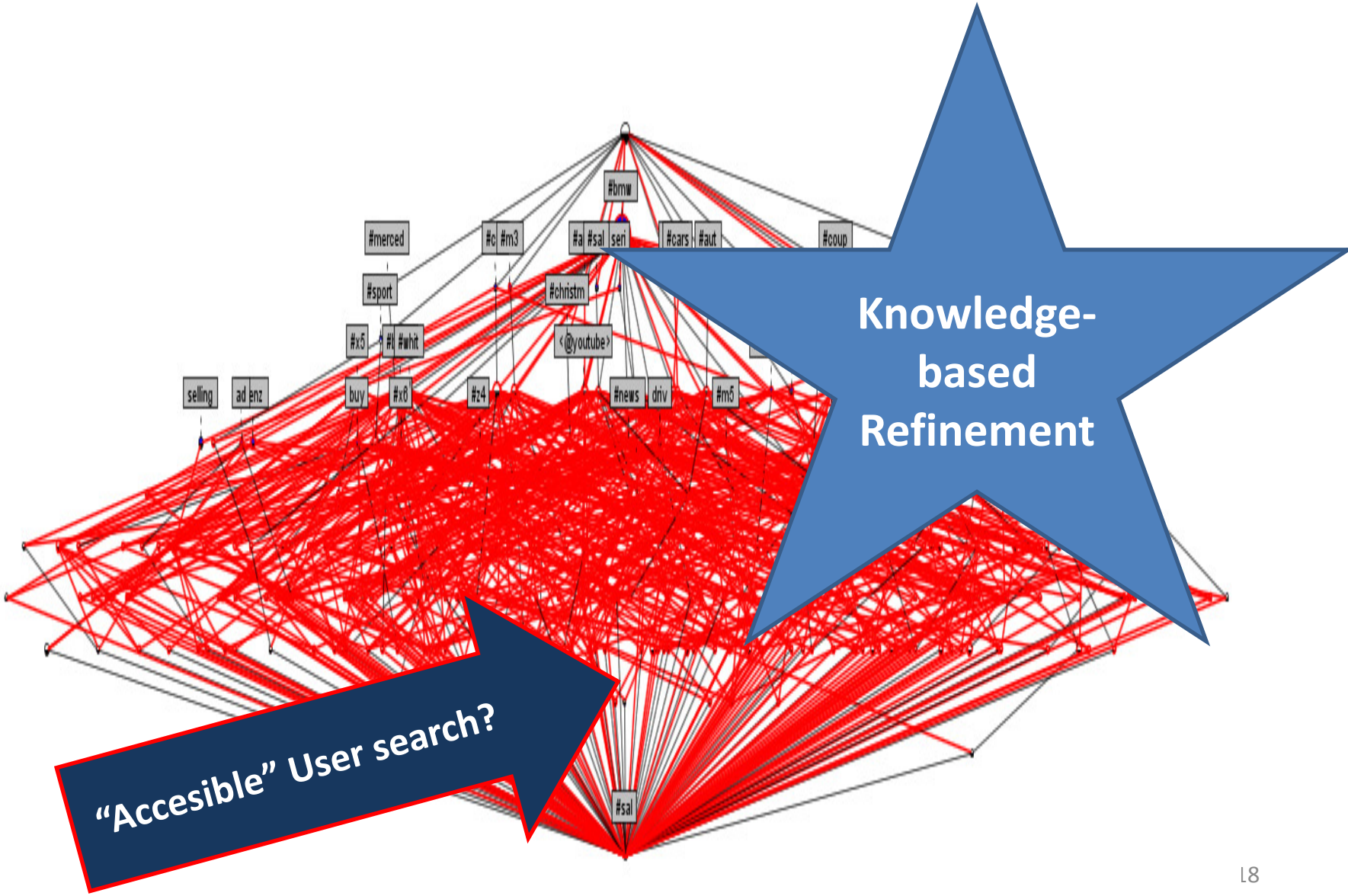
FCA-DIMH at a Glance

FCA components:

- Objects = DIMH Cards
- Attributes = terminology + special features
- Formal Concepts = similar DIMH cards together according to their features
- Structure = Formal concepts ordered in a hierarchical structure (lattice)



FCA-DIMH First Results



First Refinement: Attribute Selection

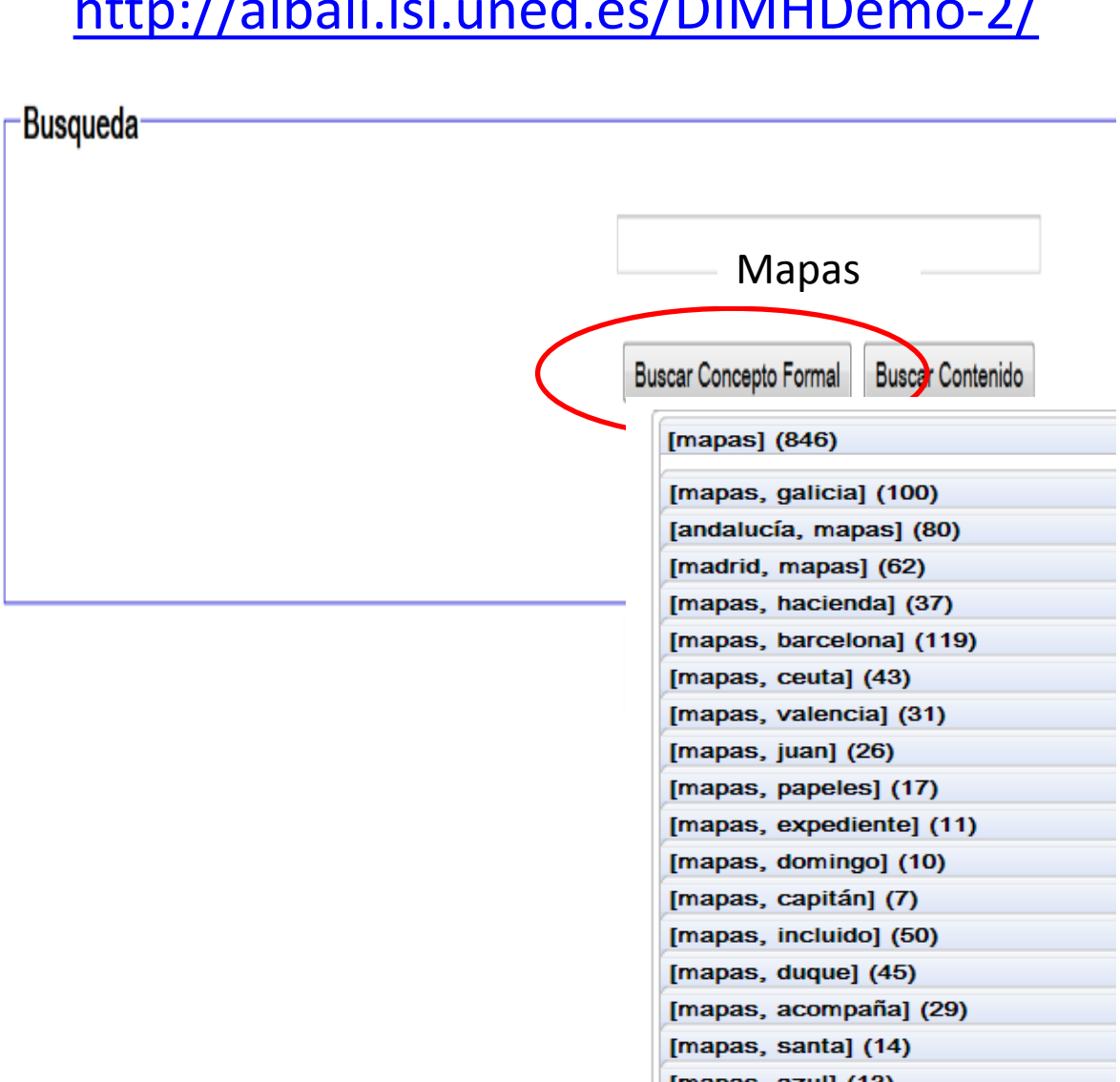
To identify the “interesting” terms, it is applied the Kullback-Leibler Divergence (Kullback & Leibler 1951).

KLD analyses the probability distribution of an attribute over a document and a collection of documents, **to identify those attributes that best represent a document by differentiating it from the rest in the collection.**

	Files	Attributes	Relationships	Formal Concepts
BEFORE KLD	7,792	130	+ 32,000	**+ 25,000**
AFTER KLD	7,792	103	29,874	**17,501**

FCA -DIMH Browser

http://albali.lsi.uned.es/DIMHDemo-2/



FCA – DIMH Contents Browser

Perfil del Castillo d Monzón cortado por los puntos E y F de Poniente a Oriente [Material cartográfico] Monzón (Huesca). Forti

1797

Idioma: spa

España-Aragón-Huesca (Provincia)-Monzón Manuscrito sobre papel. AGS. Secretaría de Guerra, Legajos, 05868. Con oficio del marquÉs de Alos a Juan Manuel Alvarez. Zar dibujos (Años 1508-1962). Volumen II : p. 287 Perfil del Castillo d Monzón cortado por los puntos E y F de Poniente a Oriente [Material cartográfico] Monzón (Huesca). Fortifica Zaragoza Zaragoza Juan_Manuel_Alvarez Perfil_del_Castillo Monzón Huesca Aragón España Zaragoza Zaragoza Mapas_Referencias Mapas Perfil_del_Castillo Alos Juan_

Jaca (Huesca). Fortificaciones. Planos. 1642 Planta del Castillo de Jaca [Material cartográfico]

1642

Idioma: spa

España-Aragón-Huesca (Provincia)-Jaca Manuscrito sobre papel. Tinta AGS. Guerra y Marina, Legajos, 01454. Acompaña a carta del MarquÉs de Távara a S. M., Zaragoza, 2 : p. 605 Jaca (Huesca). Fortificaciones. Planos. 1642 Planta del Castillo de Jaca [Material cartográfico] Huesca Jaca Mapas_Referencias Mapas Aragón España Zaragoza Za Simón_Manuscrito Huesca Jaca Aragón España Zaragoza Zaragoza Mapas_Referencias Mapas Cornacholo MarquÉs_de_Távara Planta_del_Castillo_de_Jaca Simón_Mar

Planta del castillo de Canfranc con los reparos que hay que hacer en Él [Material cartográfico] Canfranc (Huesca). Fortificacio

1617

Idioma: spa

España-Aragón-Huesca (Provincia)-Canfranc Tinta y color amarillo. Con rotulación Manuscrito sobre papel. AGS. Guerra y Marina, Legajos, 00823. Con carta de Felipe de So 1508-1962). Volumen II : p. 96 Planta del castillo de Canfranc con los reparos que hay que hacer en Él [Material cartográfico] Canfranc (Huesca). Fortificaciones. Planos. 161 Canfranc Huesca Aragón España Mapas_Referencias Mapas Felipe_de_Soria

Jaca (Huesca). Fortificaciones. Planos. 1645 Planta del castillo de Jaca [Material cartográfico]

1645

Idioma: spa

España-Aragón-Huesca (Provincia)-Jaca Manuscrito sobre papel. Dibujo a plumilla. Con rotulación AGS. Guerra y Marina, Legajos, 01594. Con carta de D. Luis Carrillo de T (Años 1503-1805). Volumen I : p. 605 Jaca (Huesca). Fortificaciones. Planos. 1645 Planta del castillo de Jaca [Material cartográfico] Jaca Huesca Mapas_Referencias Mapas España Toledo Toledo Mapas_Referencias Mapas AGS Castillo_de_Jaca Luis_Carrillo

Berdún (Huesca). Mapas generales. 1592 Berdún (Huesca). Fortificaciones. Planos. 1592 Planta de Verdun ; Planta del castill

1592

Idioma: spa

España-Aragón-Huesca (Provincia)-Berdún Tinta. Con rotulación Manuscrito sobre papel AGS. Guerra y Marina, Legajos, 00356, 181. Con carta de D. Alonso de Vargas al R [= 17,5 cm] Referencias: Mapas, planos y dibujos (Años 1503-1805). Volumen I : p. 1000 Berdún (Huesca). Mapas generales. 1592 Berdún (Huesca). Fortificaciones. Planos [Material cartográfico] Berdún Huesca Mapas_Referencias Mapas Mapas Verdun AGS Acosta Aragón España Jaca Tiburzio Vargas D._Alonso Hernando_de_Spannocchi Hu Hernando_de_Spannocchi Berdún Verdun Acosta Tiburzio D._Alonso

User Evaluation

- End-Users Interested in already known Topics/A priori taxonomy of concepts
  - Authors (cited)
  - Measures Units
  - Dates, Places,…

Next solution:

take only the terminology defined by experts!

2nd Refinement: Taxonomy

The inclusion of experts terminology (taxonomy) in the FCA execution lead to a model organizing the data according to it.

¿It is reduced the number of formal concepts and relationships?

	Files	Attributes	Relationships	Formal Concepts
BEFORE KLD	7,792	130	+ 32,000	**+ 25,000**
	7,792	**36**	13,719	**1,197**

Research Support

- By means of the developed visualization, the experts are allowed to explore the data and the inferred relationships, drawing new conclusions about the contents based on their expertise.

- In addition, **associations in the data can be automatically discovered from the lattice structure**.

  Hypothesis: This Knowledge Extraction offers to the experts **the starting point for a deeper analysis** of the discovered implications.

Automatic Association Rules

Automatic knowledge inference from the FCA lattice is carried out in two steps:

- Find the most common feature by the FPGrouth algorithm (Han et al. 2004).
- Find the association rules related to these frequent features by algorithm (Agrawal et al. 1994).

	# Features	# Rules
Named Entities	22	29
Taxonomy	102	133
All	27	121

Results (not yet by the historians)

- First, the rules inferred from the named entities are mostly related to well known locations.
- By means of the taxonomy, more specific rules are obtained.

  plaza → paper plan

  - support: 0.01196 (87/7274)
  - confidence: 0.12850

- Finally, by using all the information, it seems that new unknown information is offered by the rules.

  representation_system material graphic → plans

  - suport: 0.706622176591375 7 (5506/7792)
  - confidence: 0.811373011196229

Next Steps

- New Visualization Methaphore
- Evaluation of the quality of the results
  - Laboratory-based Evaluation???
  - User-based Evaluation

http://dimh.hypotheses.org/

Thanks!!! to Alicia Cámara, DIMH Project Responsible, Departamento de Historia del Arte, UNED

Ana García-Serrano,
Ángel Castellanos,

ETSI Informática, UNED