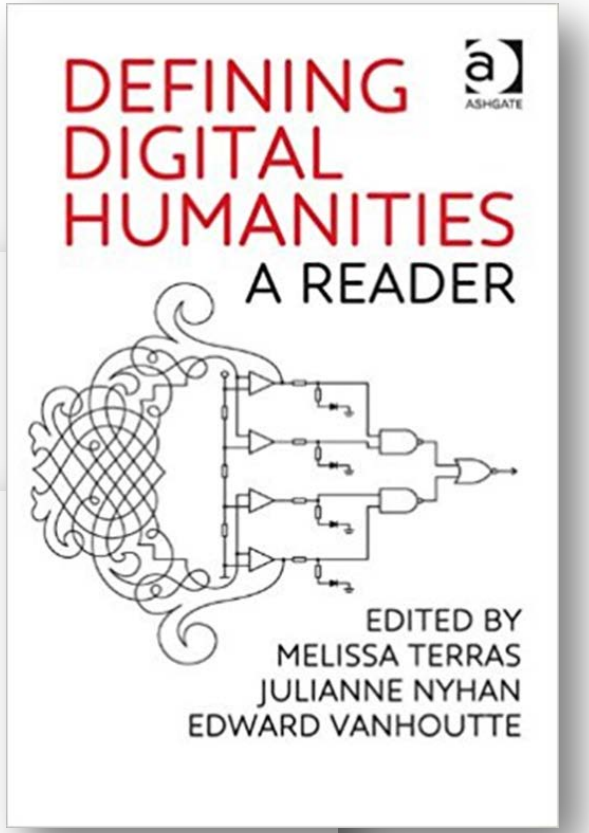
Natural Language Processing and Digital Humanities

Núria Bel

25-04-2018

upf. Universitat Pompeu Fabra Barcelona

https://whatisdigitalhumanities.com/



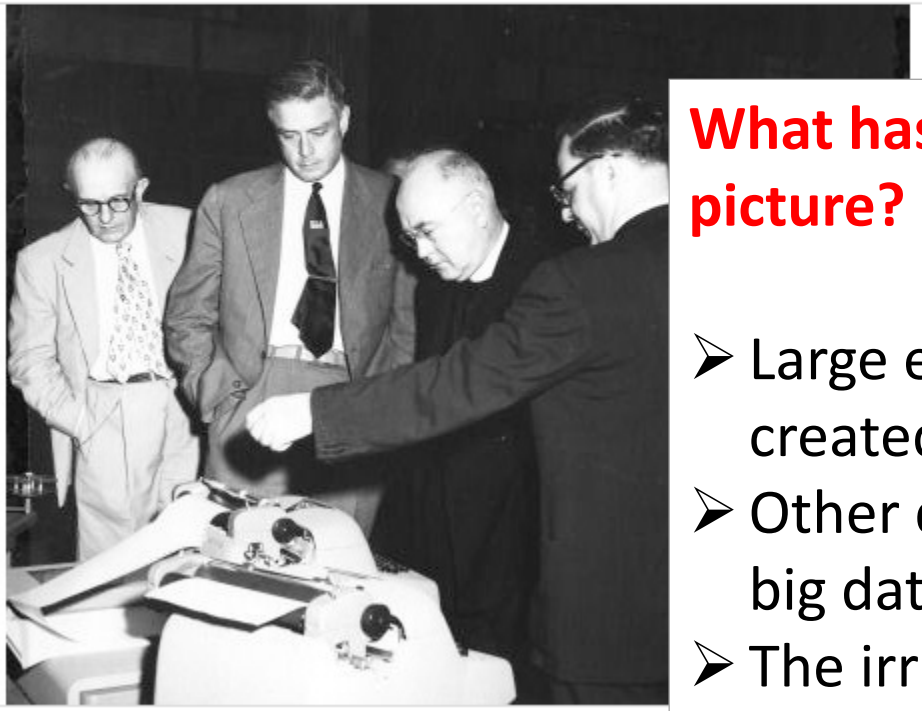
What Is Digital Humanities?

What Is Digital Humanities?

Doing research with digital materials and tools in a collaborative and open fashion. Not everyone needs to be a builder, though working with someone who can helps! *James Baker*

NB: Refresh the page to get a new definition. Quotes were pulled from participants from the Day of DH between 2009-2014. As of January 2015, the database contains 817 rows and randomly selects a quote each time the page is loaded. If you want to do something cool with the data, I am providing a download for the CSV I compiled here.

Italian Roberto Busa is considered the pioneer of Computational Linguistics. In 1946 he proposed a revolutionary idea to IBM: using computers to study texts, in particular the collected works of St. Thomas Aquinas. IBM decided to bet on the future.



New York - IBM World Headquarters: Presentation of the first resul
Thomisticus" project

http://www.corpusthomisticum.org/it/ind

Computers and the Humanities

Coverage: 1966-2004 (Vol. 1, No. 1 - Vol. 38, No. 4)
Published by: Springer

Title History (What is a title history?)

2005-2014 -

What has changed since 2004, when DH comes into the picture?

- Large efforts in digitalization that made many objects created by humans available for software tools.
- Other disciplines being hit by technologies that handle big data
- The irruption of Geographical Information Systems, GIS, and other visualization tools that have been used to easy 'interpretation'

Barcelona

Summary of DH research

Research in humanities is often related to
the interpretation of objects created by humans

Such objects are mainly texts for literature and history, for instance

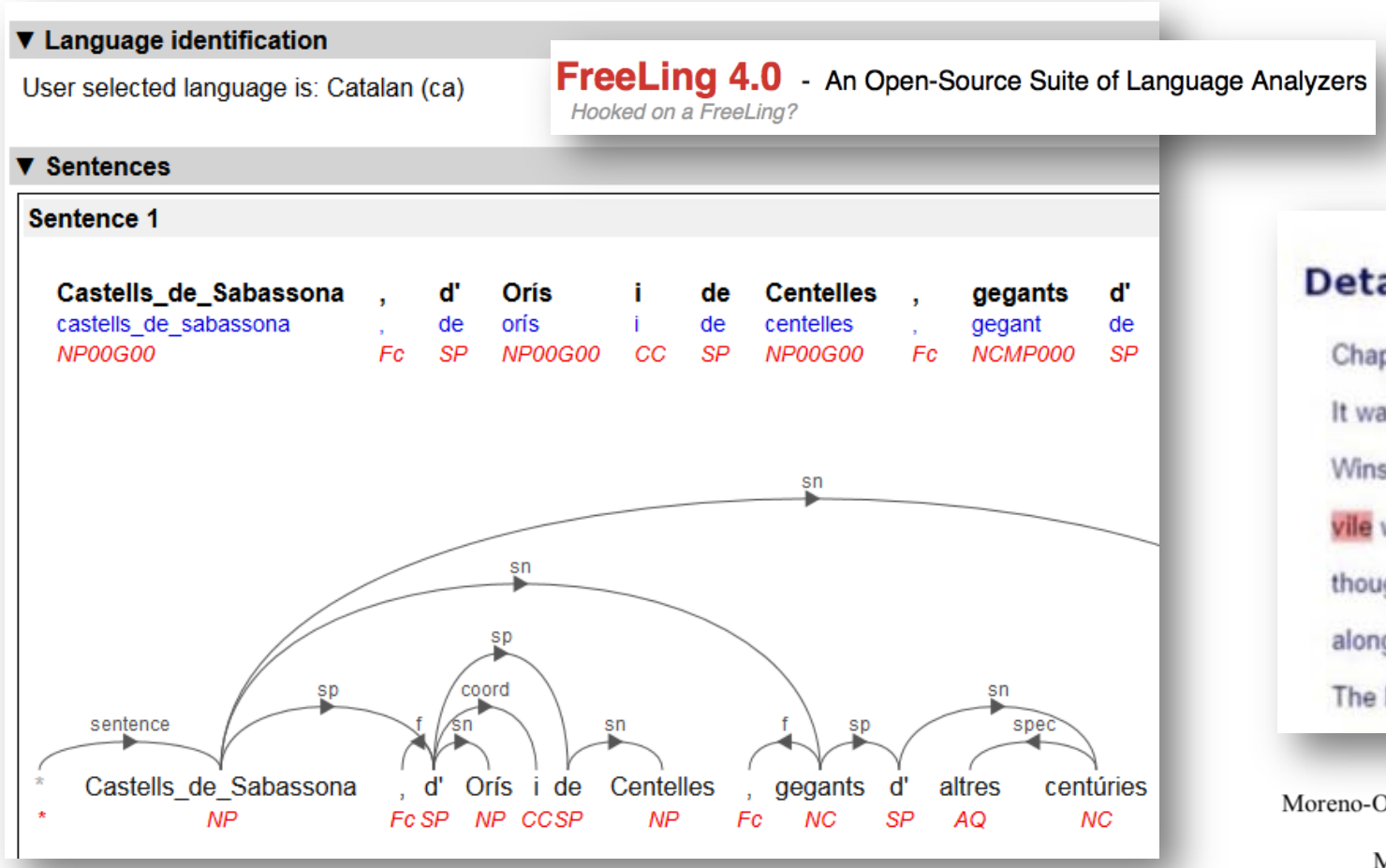
Research has to do with:

- Finding ... Texts
- Finding ... Data in texts
- Finding ... Patterns and relations in data, so that ..
- Making a particular interpretation in a sound and evident way is possible

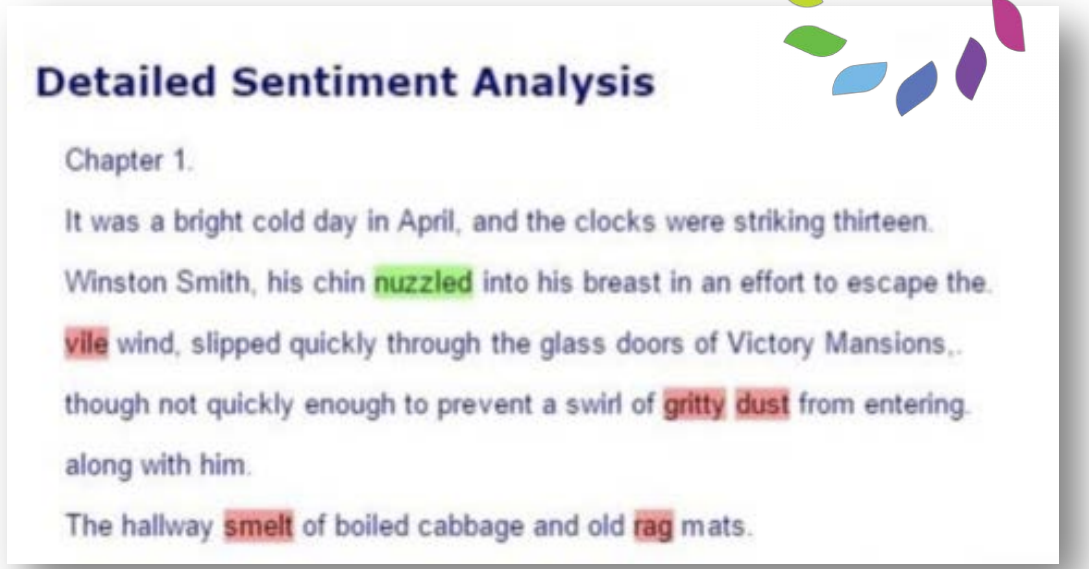
Natural Language Processing (NLP) are available tools for …

- Finding data in **large quantities of texts**
- Finding and **extracting patterns and correlations**
- Representing data to enable their analysis

NLP tools find names, relations, opinions …



Lluís Padró and Evgeny Stanilovsky.
FreeLing 3.0: Towards Wider Multilinguality
Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA.



Moreno-Ortiz, A. (2016). Lingmotif 1.0 [Computer Software]. Málaga, Spain: Universidad de Málaga. Retrieved from Tecnolengua website: http://tecnolengua.uma.es/lingmotif

upf. Universitat Pompeu Fabra Barcelona

Finding names. Examples

Geographical Information Systems (GIS) and NLP Named Entity Recognition & Desambiguation (NER)

The task is:

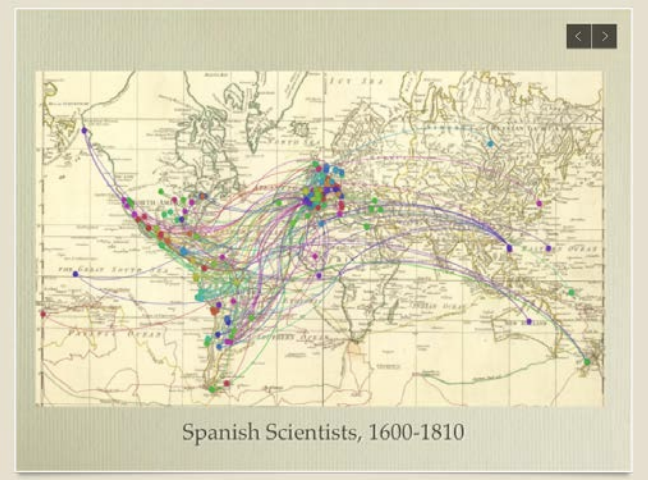
- identifying an entity mention (a proper noun) and
- identifying to which entity it refers to.

Geographical Names (assigning coordinates, for instance)

Person Names (linking to VIAF, Wikipedia, for instance)

A1 | Place

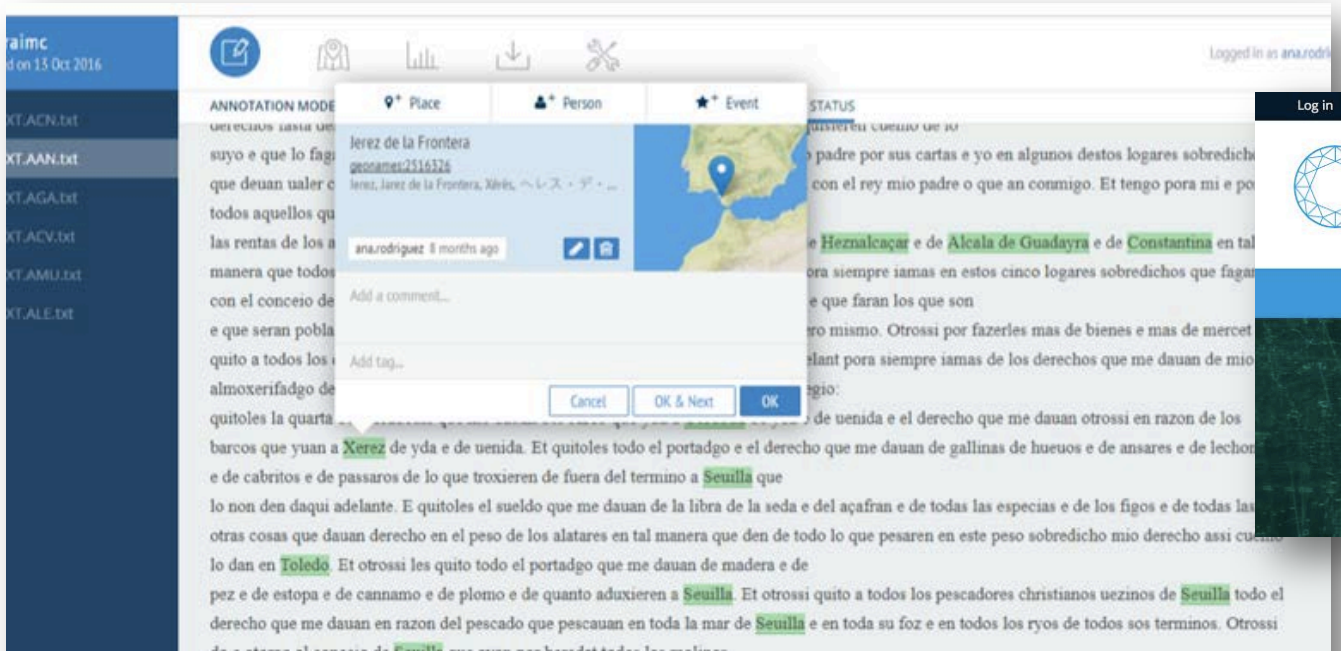
	A	B	C
1	Place	Coordinates	
2	Amsterdam	52.373056, 4.892222	
3	Ankara	39.916667, 32.833333	
4	Arlington	38.880278, -77.108333	
5	Athens	37.966667, 23.716667	
6	Baden Baden	48.762778, 8.240833	
7	Belgrade	44.816667, 20.466667	
8	Berlin	52.516667, 13.383333	
9	Bern	46.95, 7.45	
10	Boston	42.358056, -71.063611	
11	Brussels	50.85, 4.35	
12	Bucharest	44.4325, 26.103889	
13	Budapest	47.471944, 19.050278	
14	Bristol	51.45, -2.583333	

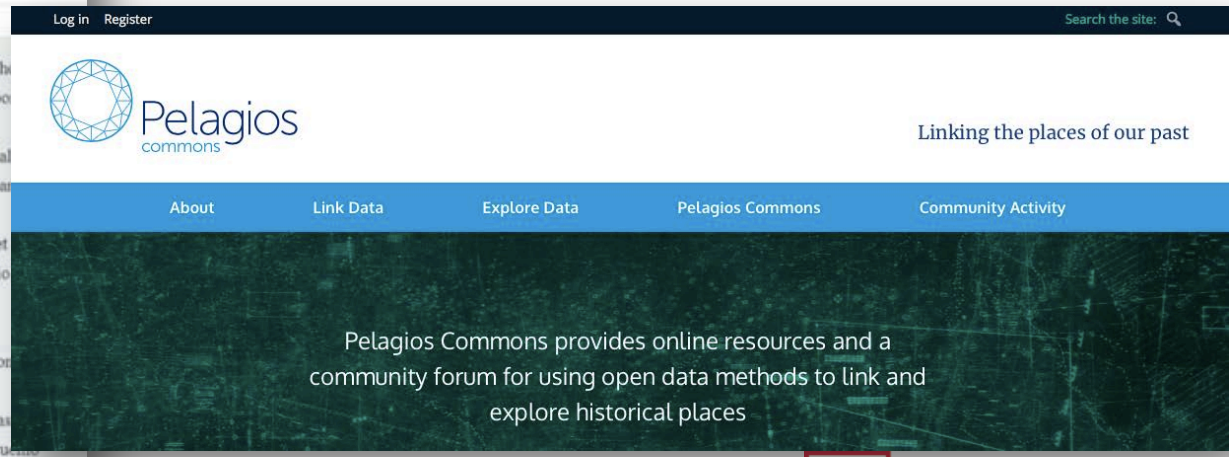


GIS and NER (1)
Pelagios Mediaeval Iberia http://commons.pelagios.org

Annotating texts with Identification of Named Entities

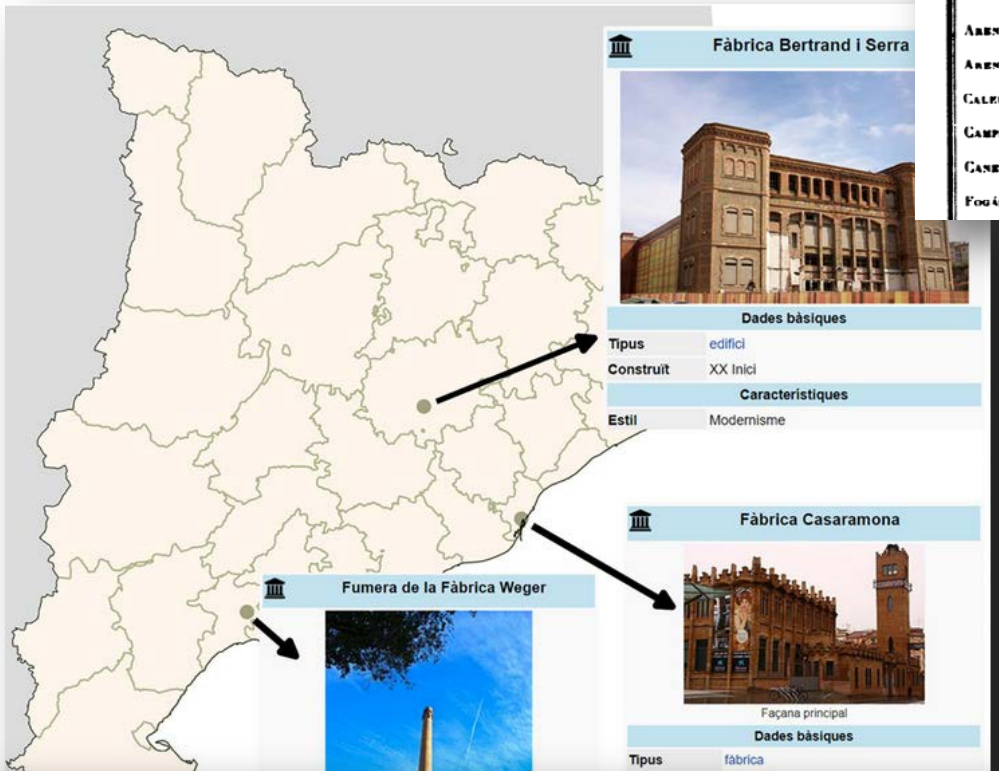
- NER with FreeLing
- Relate and extend with LOD, dbpedia...
- And geolocation with Geographic Infomation Systems (GIS)





upf. Universitat Pompeu Fabra Barcelona

GIS and NER (2)



AYUNTAMIENTOS.	CÉDULAS DE INSCRIPCION.	HABITANTES.	POR NATURALEZA. NACIONALES. Establecidos.	NACIONALES. Transeuntes.	EXTRANJEROS. Establecidos.	EXTRANJEROS. Transeuntes.	POR SEXO.	POR ESTADO CIVIL. Solteros.	Casados.	Viudos.	Menos de 1 año.	De 1 á 5.	De 6 á 10.	De 11 á 15.	De 16 á 19.
PARTIDO DE ARENYS DE MAR.															
Arenys de mar	1.105	5.219 Varones.	2.328	28	2	5	2.363	1.363	877	123	70	297	276	248	133
		5.219 Hembras.	2.827	29	..	..	2.856	1.595	963	298	72	267	258	257	232
Arenys de munt	601	3.305 Varones.	1.642	24	..	..	1.666	1.003	616	47	50	208	199	171	122
		3.305 Hembras.	1.624	15	..	..	1.639	907	619	113	55	203	160	188	152
Calella	774	3.526 Varones.	1.585	38	1	..	1.624	915	611	68	48	183	158	161	106
		3.526 Hembras.	1.882	19	1	..	1.902	1.030	684	188	37	170	196	190	183
Campins	72	373 Varones.	209	4	..	..	213	134	73	6	5	25	27	16	13
		373 Hembras.	160	..	..	..	160	80	70	10	7	12	13	15	8
Canet de mar	724	3.232 Varones.	1.518	..	..	..	1.538	937	541	57	48	161	190	138	121
		3.232 Hembras.	1.694	..	..	..	1.694	1.006	564	124	42	170	205	130	116
Fogás de Tordera	143	707 Varones.	319	71	..	4	394	247	129	18	12	38	47	41	26
		707 Hembras.	292	18	..	3	313	169	125	19	9	45	38	28	20

ORIGINAL RESEARCH ARTICLE
Front. Digit. Humanit., 04 October 2017 | https://doi.org/10.3389/fdigh.2017.00019

Heritage As a Source of Studies into Industrial History: Using Digital Tools to Explore the Geography of the Industrialization

Guillermo Esteban-Oliver*, Adrià San José* and Jordi Martí-Henneberg*

Department of Geography and Sociology, University of Lleida, Lleida, Spain

The main objective of this article is to explore the possibility of combining two very different sources in order to study the distribution of industrial activity throughout history. The traditional primary sources to use for this purpose are the official censuses on population and economic activity that have been conducted in the majority of countries since the mid-nineteenth century. However, the majority of these lack detail at the regional level and also with respect to the types of professional occupations that they quantify. In order to complement and profile these census data, we propose the use of another type of information which can also be quantified, but whose characteristics are very different. We refer to the industrial heritage sites identified in digital format in a given territory, which in this case is Catalonia, Spain. This innovative dataset was obtained using digital tools such as web scraping and data mining techniques. This type of historical information was used to check whether it is reliable and valid for interpreting the spatial impact of the introduction of industrial activity. The article also shows that the systematic identification of elements of industrial heritage offers a new and very useful source of information for interpreting the history of industrial geography.

EDITED BY
Josep Lladós
Universitat Autònoma de Barcelona, Spain

REVIEWED BY
Andreas Degkwitz
Humboldt-Universität zu Berlin, Germany
Joana Maria Pujadas Mora
Centre for Demographic Studies, Spain

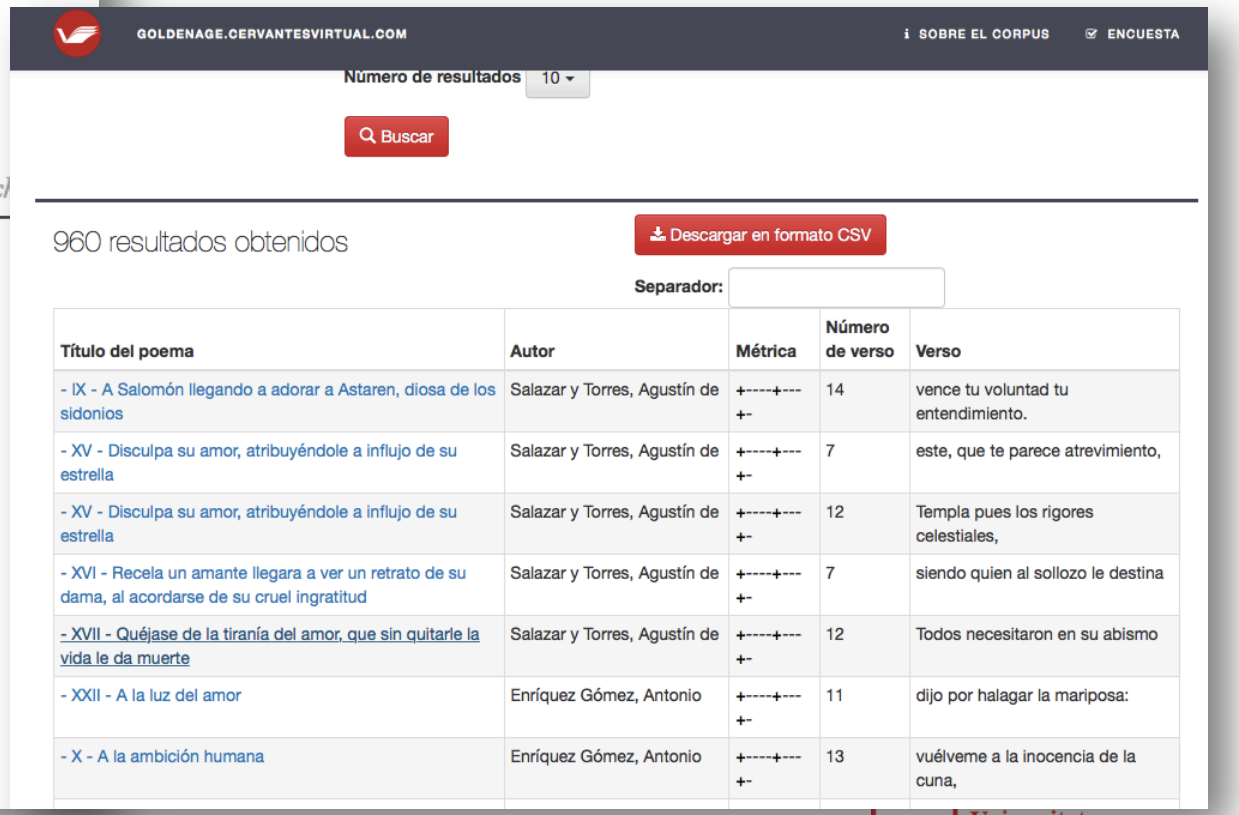
Materials and Methods

The methodology used to develop this research consisted of an analysis to correlate industrial occupations with data about industrial heritage sites. We will start by first describing the methodology that allowed us to compile a database for industrial occupations in 1860. Then we will describe how we created a complementary database on industrial sites. As a result, we were able to analyze whether the two datasets were correlated and the extent to which they were able to provide detailed georeferenced data.[6]

Finding patters. Examples

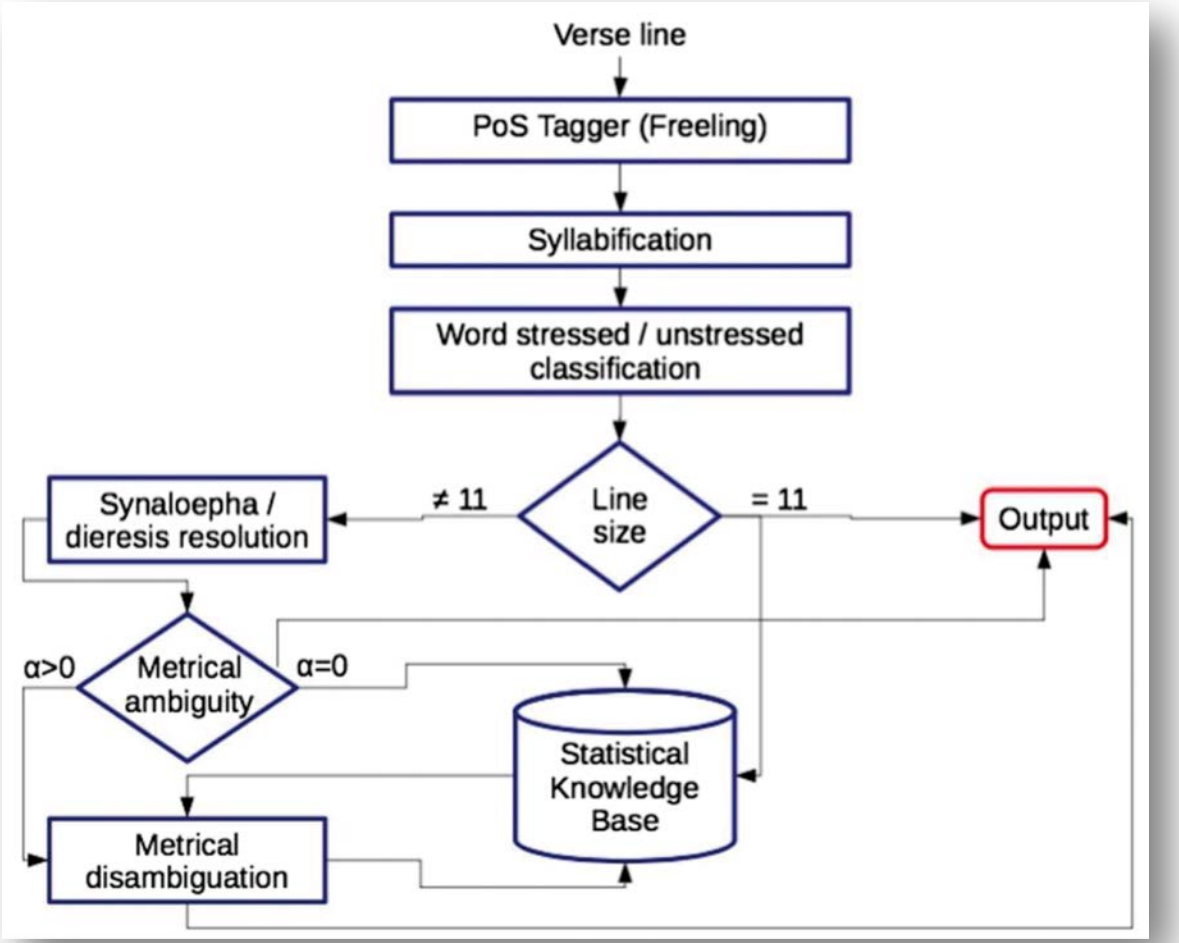
NLP for finding patterns

The corpus has more than 5,000 sonets (about 71,000 lines)



upf. Universitat Pompeu Fabra Barcelona

Borja Navarro Colorado (2018) "A Metrical Scansion System for Fixed-Metre Spanish Poetry", *Digital Scholarship in the Humanities*, Volume 33 (1), pages 112–127



NLP for Sentiment Analysis. Example

NLP
Sentiment Analysis & Text Classification

Figure 1: Communities in Reddit: each node represents a community. Red nodes initiate more conflicts, while blue nodes do not. Communities are embedded using user-community information, as described in Section 6. Figure best viewed in color.

Community Interaction and Conflict on the Web

Srijan Kumar
Stanford University, USA
srijan@cs.stanford.edu

William L. Hamilton
Stanford University, USA
wleif@stanford.edu

Jure Leskovec
Stanford University, USA
jure@cs.stanford.edu

Dan Jurafsky
Stanford University, USA
jurafsky@stanford.edu

It analyzed 40 months of data, containing 1.8 billion comments made by over 100 million users across 36,000 communities

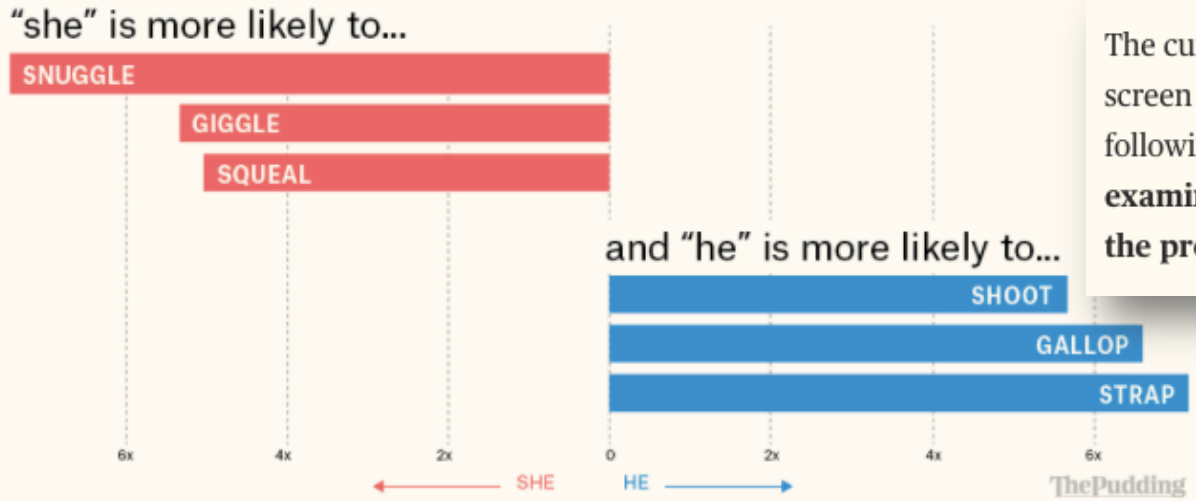
Srijan Kumar, William L. Hamilton, Jure Leskovec, Dan Jurafsky. 2018.Community Interaction and Conflict on the Web. Proceedings of the Web Conference (WWW). 2018 (to appear).

upf. Universitat Pompeu Fabra Barcelona

NLP for Data Analysis. Examples

NLP for data analysis

In screen direction from film scripts,



The curious data here is less what Rose says ("I'm flying") and more what the screen direction prescribes ("she smiles dreamily," "he pushes against her"). In the following analysis, we go deep on screen direction to understand gender tropes. **We examined 2,000 scripts and broke down every screen direction mapped to the pronouns "she" and "he."**

The R code behind this analysis in publicly available on GitHub.

https://juliasilge.com/about
https://pudding.cool/2017/08/screen-direction/

upf. Universitat Pompeu Fabra Barcelona

Word embeddings quantify 100 years of gender and ethnic stereotypes

Garg, N., Schiebinger, L., Jurafsky, D. & Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. PNAS 201720347 (2018). doi:10.1073/pnas.1720347115

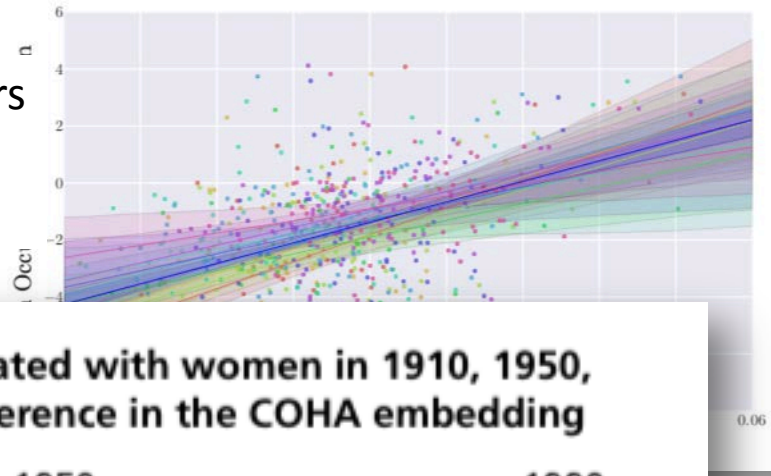


Table 1. The top 10 occupations most closely associat each ethnic group in the Google News embedding

Hispanic	Asian
Housekeeper	Professor
Mason	Official
Artist	Secretary
Janitor	Conductor
Dancer	Physicist
Mechanic	Scientist
Photographer	Chemist
Baker	Tailor
Cashier	Accountant
Driver	Engineer

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

upf. Pompeu Fabra Barcelona

NLP and Digital Humanities?

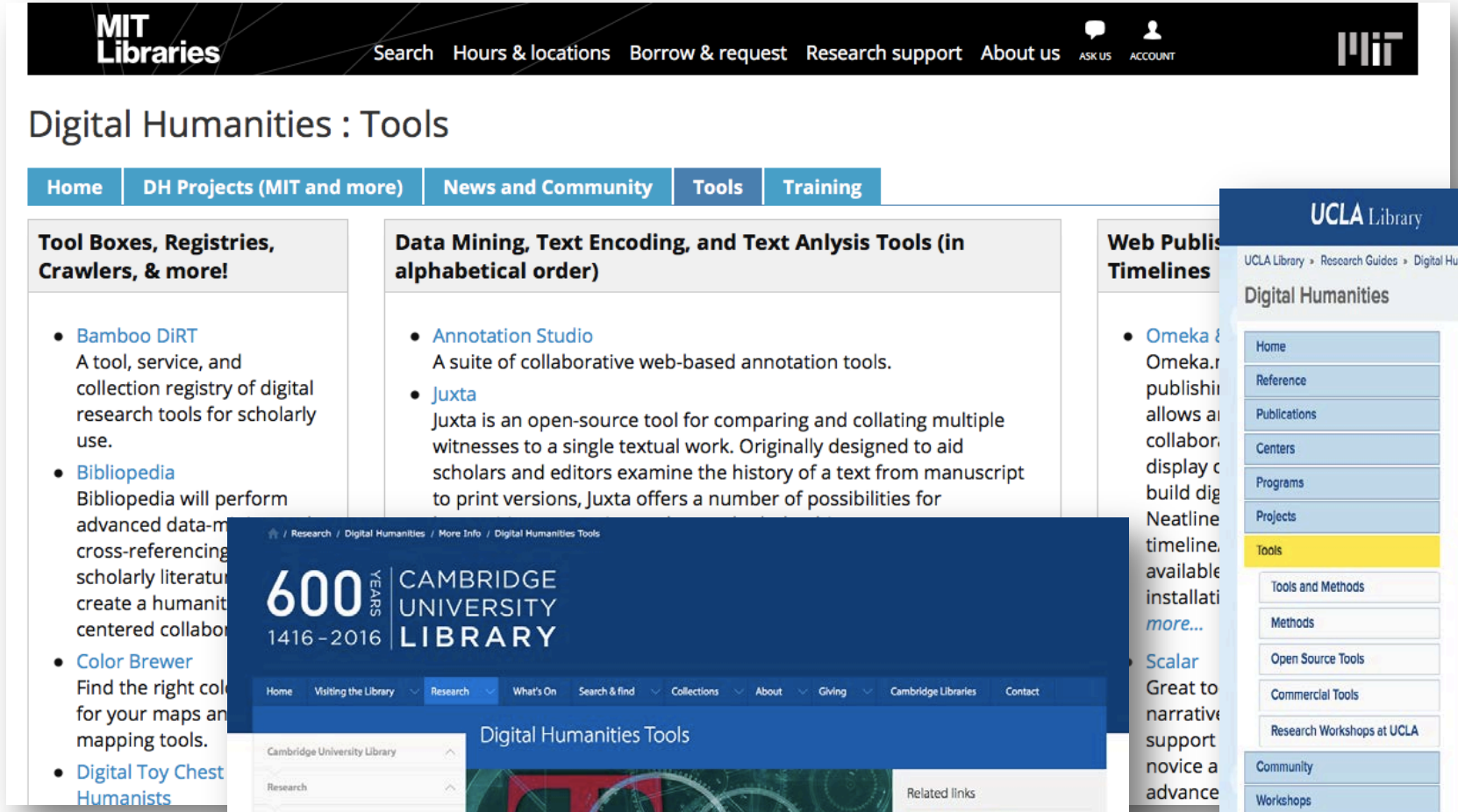
It's impossible to conduct research without software, say 7 out of 10 UK researchers

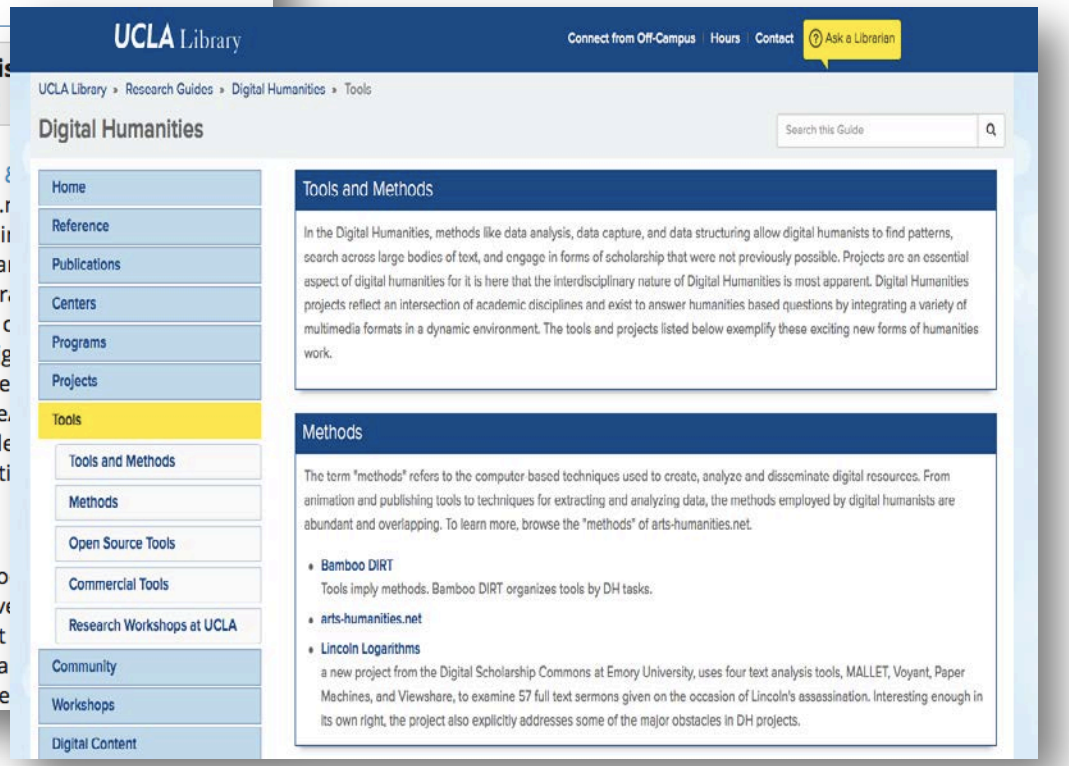
The Software Sustainability Institute, University of Edinburgh

(http://www.software.ac.uk/blog/2014-12-04-its-impossible-conduct-research-without-software-say-7-out-10-uk-researchers)

2014	Do you use research software?			Do you develop your own research software?		
Humanities & Language Based Studies & archaeology	Yes	21	55,3%	Yes	5	13,2%
	No	17	44,7%	No	33	86,8%
Medicine, Dentistry & Health	Yes	55	96,5%	Yes	24	42,1%
	No	2	3,5%	No	33	57,9%
Social Studies	Yes	38	86,4%	Yes	15	34,1%
	No	6	13,6%	No	29	65,9%

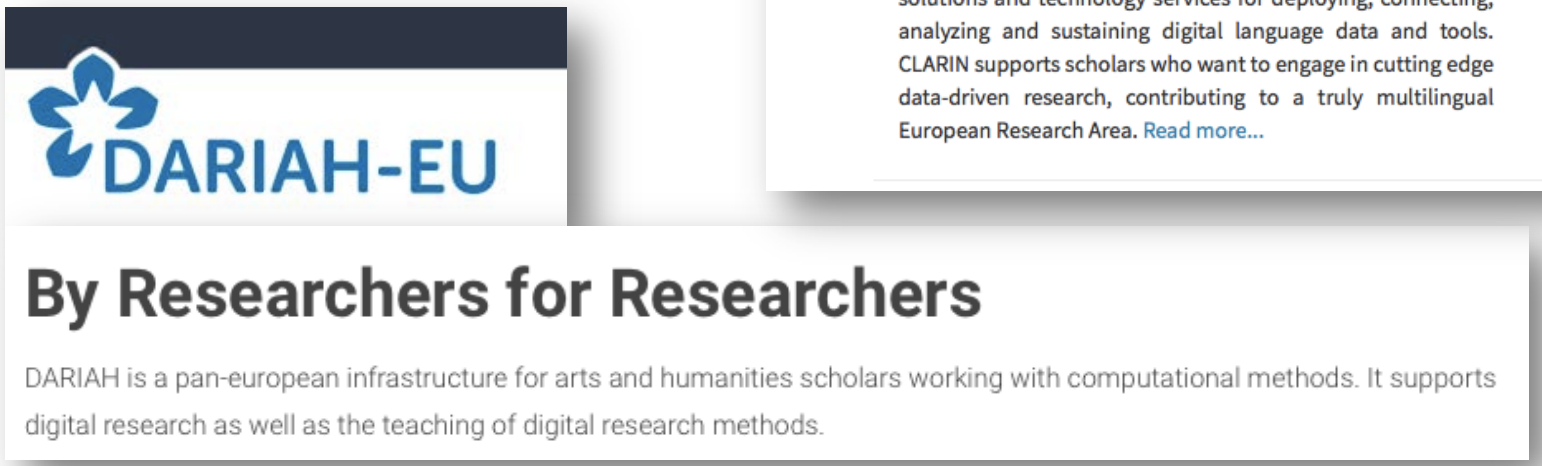
Researchers need information about tools ....





About research infrastructures …

https://www.clarin.eu



https://www.dariah.eu

Summary of DH research and NLP

Research in humanities is often related to
the interpretation of objects created by humans

Such objects are mainly texts for literature and history, for instance

Research has to do with:

- Finding ... Texts ???????

✓Finding ... Data in texts

✓Finding ... Patterns and relations in data, so that ..

✓Making a particular interpretation, sound and evident, is possible

Finding texts: Researchers need downloadable and processable DATA as collections …