Ligue des Bibliothèques Européennes de Recherche
Association of European Research Libraries

Text and Data Mining

Using Cultural Heritage Data: Opportunities and challenges



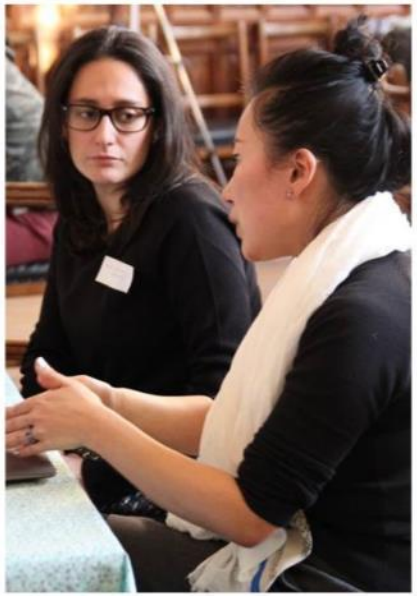
Melanie Imming
EU Projects manager, LIBER

CC BY

FutureTDM
Explore . Analyse . Improve

- Improve uptake of text and data mining (TDM) in the EU
- Raise awareness of TDM
- Develop solutions to barriers together with stakeholders





OPENMINTED

OPEN TEXT AND DATA MINING PLATFORM FOR OPEN SCIENTIFIC CONTENT

- focuses on interoperability across mining services and content providers
- So that researchers can collaboratively create, discover, share and re-use open texts and data

LIBER
Ligue des Bibliothèques Européennes de Recherche
Association of European Research Libraries

TDM Cultural Heritage

Text and Data Mining: How big is big?

Mining:

- More data than you can process yourself in reasonable amount of time
- Data that require computational intervention to make more sense of it all

Not Macro vs Micro

Making use of these techniques, data sets or new methods is not automatically choosing to ‘go big’:

- Can be about one Work of Art
- Not Event History vs Longue Durée

CC BY

Mining Cultural Heritage

What?

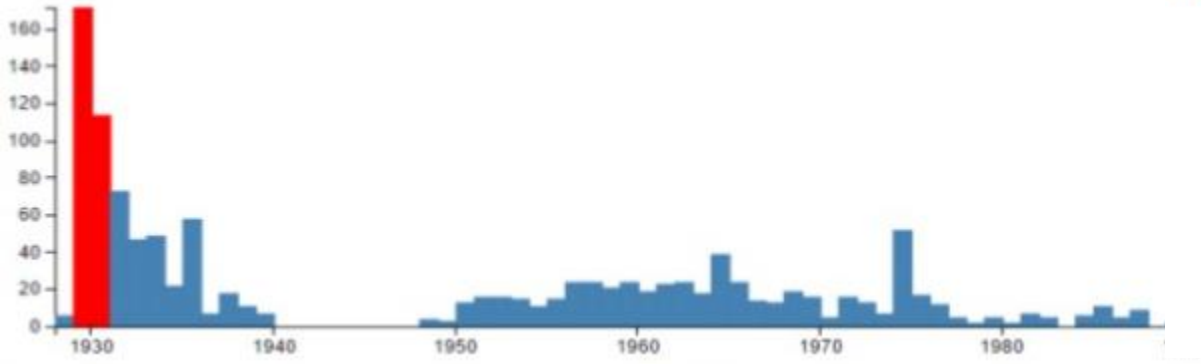
In research projects:

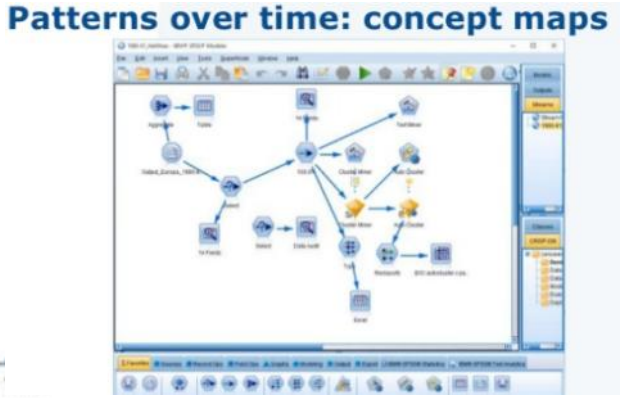
- Basic text mining: e.g. Word Clouds
- Network analysis
- Topic Modelling





id	words									topic
1	duitschland	frankrijk	duitsche	fransche	regeering	engeland	politiek	europa	rtali	politieke	politics
2	europa	hel	eu	ke	nen	beeft	heelt	land	vnn	pen	=????
3	man	goed	gaan	vrouw	dag	kwam	menschen	paar	ging	komen	human int.
4	europa	staten	landen	europeesche	internationale	economische	volkenbond	commissie	oostenrijk	conferentie	politics
5	ova	nederland	koloni	ned	holl	buitenland	adam	dam	obl	bank	economy
6	wereld	volk	leven	rusland	staat	land	oorlog	eigen	tijd	russische	politics
7	heer	regeering	nederland	minister	nederlandsche	belgi	leden	voorzitter	partij	vergadering	politics
8	europa	ver	duitsch	wereld	temperatuur	wind	noord	land	mm	lat	weather
9	groote	werk	oude	eenste	film	boek	kunst	tijd	geheel	nieuwe	media/art
10	werden	jaar	koning	groote	kerk	jaren	kwam	plaats	eerste	stad	human int.
11	jaar	millioen	amerika	industrie	productie	groote	ton	nieuwe	handel	prijzen	economy
12	juli	mei	maart	jan	loco	sept	juni	april	dec	oct	time
13	londen	rott	dam	antwerpen	nov	hamburg	rotterdam	new	japan	juni	locations
14	groote	land	plaats	eenste	twee	aantal	groot	maken	amerika	werden	=????
15	europa	dienst	heer	indi	betrekking	amsterdam	klasse	school	verlof	geplaatst	advert
	code	politics	sports	economy	weather	human int.	media/art	time	locations	advert	





Images © prof. dr. Joris van Eijnatten

LIBER
Ligue des Bibliothèques Européennes de Recherche
Association of European Research Libraries

How did newspapers in the twentieth century frame Europe?

Comparitive analysis of cultural patterns in time and space
prof. dr. Joris van Eijnatten


NEDERLANDSE GEKOZEN
TOT „MISS EUROPA"

Vierde was het Franse meisje Geneviève Zanetti. Miss England, de 23-jarige Sonia Hamilton, kreeg de vijfde plaats. Het Amsterdamse meisje gaat nu meedoen aan de Miss World-wedstrijd, die in oktober te Londen wordt gehouden. En wellicht wacht haar nog de wedstrijd om de titel Miss Universe in november in Long Beach (V.S.).

Onmiddellijk na de verkiezing verklapte zij haar geheim aan vijftig journalisten, die om haar heendrongen: „Fietsen, zoals iedereen in Holland doet. Dat houdt mij slank."

Mysterie rond kikvorsman opge
LICHAAM VAN CRA

Toolbox

1 Read stuff (use your eyes)
2 Time line generator (nGram viewers)
3 Semantic tekst mining tool (texcavator)
4 Corpus linguistics (e.g. Antconc, CasualConc, Wordsmith)
5 Topic modelling (e.g. Mallet)
6 Tekst analytics suite (SPSS Modeler)
7 Vector-space modeling (ShiCo)

An Epidemiology of Information:
Data Mining the 1918 Influenza Pandemic



U. of Kentucky

A Digging into Data project:
A Trans-Atlantic Platform for the Social Sciences and Humanities, representing 11 nations from both sides of the Atlantic.

- Harness the power of data mining techniques with interpretive analytics of the humanities and social science
- integrated traditional interpretive analysis (close readings of texts) with dynamic temporal segmentation (topic modeling and segmentation) and tone analysis
- Research can provide methods for understanding the spread of information and the flow of disease in other societies facing the threat of pandemics

Welt der Kinder - Children and their World

KNOWLEDGE OF THE WORLD AND ITS INTERPRETATION IN TEXT BOOKS AND CHILDREN'S LITERATURE, 1850-1918

Prof. Dr. Iryna Gurevych

- Representations and interpretations of the world in the period from 1850 until 1918
- Over 600.000 digitalized pages



"G. B. Wadström unterrichtet einen Negerprinzen" aus: Wilmsen, Friedrich Philipp: Fremde Länder und Völker, Berlin 1815, Frontispiz.

LIBER
Ligue des Bibliothèques Européennes de Recherche
Association of European Research Libraries

CC BY

Welt der Kinder - Children and their World

- Combining an established hermeneutic methodology with innovative methods and technologies
- Close cooperation between historians, information scientists, and computer scientists
- Developing reusable tools for the analysis of large (digital) corpora
- Test model for future similar projects

CC BY

Authorship attribution

Who wrote the lyrics of the *Wilhelmus*, the oldest national anthem in the world?





Mike Kestermont, assistant professor, University of Antwerp

- Stylometry (computational stylistics):

computational algorithms which can automatically identify the authors of anonymous texts through the quantitative analysis of individual writing styles

CC BY

Authorship attribution



The *Wilhelmus* is traditionally ascribed to
Philips of Marnix, Lord of Saint-Aldegonde

By using these computational stylistics, a new possible candidate came up:

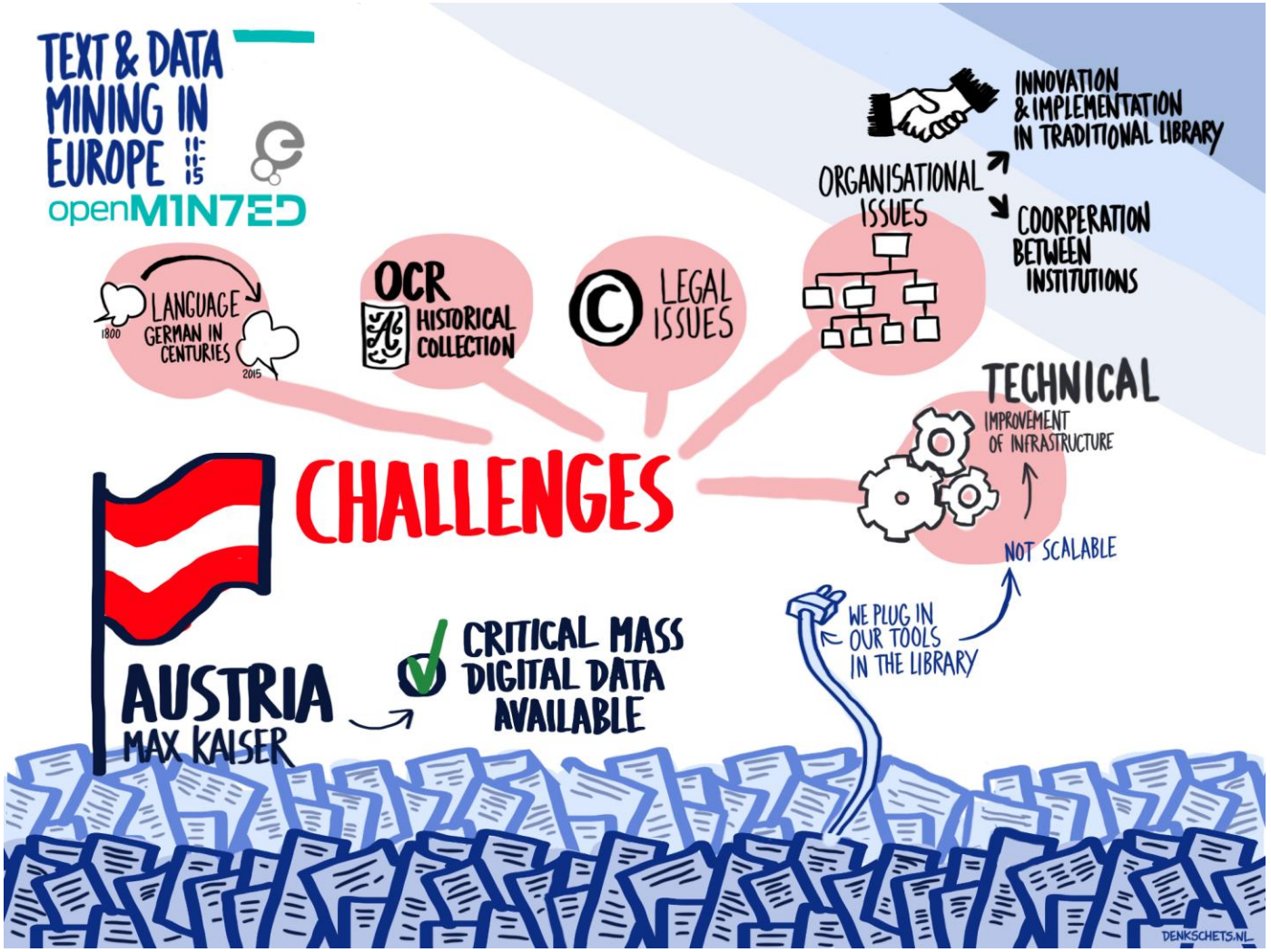


Peter Datheen, a second-rate sixteenth-century poet from French Flanders

Datheen wasn't on the Short List: but he came up when using a control group to validate the method

CC BY

Workshop Nov 2015:
"Text and Data Mining in Europe: Challenges and Action"



Upcoming calls:

Next year:
Deadline: 02 Feb 2017:

CULT-COOP-09-2017: European cultural heritage, access and analysis for a richer interpretation of the past

Humans can easily extract meaning from individual digital assets but are quickly overwhelmed by the sheer number of items which are usually spatially and/or temporally disconnected and of different digital quality. New technologies can be a valuable instrument to process large amounts of data in order to identify new correlations and interpretations and extract new meaning from our cultural and intellectual heritage.

Upcoming calls:

Next year:
Deadline: 02 Feb 2017:

CULT-COOP-06-2017: Participatory approaches and social innovation in culture

A social platform that will bring together relevant heritage stakeholders' representatives from research communities, heritage practitioners from public or private cultural institutions (heritage sites, libraries, archives, museums, and other public or private collections) and organisations (NGOs, associations), as well as policy-makers at European, national, regional or local levels. For improving the excellence of European heritage management and related policy making the platform should also harness the potential of networking among the growing number of European cultural heritage and cultural studies departments at higher education and research institutions.