



Scotland's National Collections and the Digital Humanities, Edinburgh, 14/02/2014

PROJECT OVERVIEW

- JISC/SSHRC Digging into Data Challenge II
- Jan 2012 - Dec 2013
- Text mining, data extraction and information visualisation to explore big historical datasets.
- Focus on how commodities were traded across the globe in the 19th century.
- Help historians to discover novel patterns and explore new research questions.

PROJECT TEAM



Ewan Klein, Bea Alex, Claire Grover, Richard Tobin: *text mining*



Colin Coates, Jim Clifford: *historical analysis*



UNIVERSITY OF SASKATCHEWAN



James Reid, Nicola Osborne : *data management, social media*



University of St Andrews

Aaron Quigley, Uta Hinrichs: *information visualisation*

TRADITIONAL HISTORICAL RESEARCH



Map showing the areas where mahogany is grown

Gillow and the Use of Mahogany in the Eighteenth Century, Adam Bowett, Regional Furniture, v.XII, 1998.

Global Fats Supply 1894-98



DOCUMENT COLLECTIONS

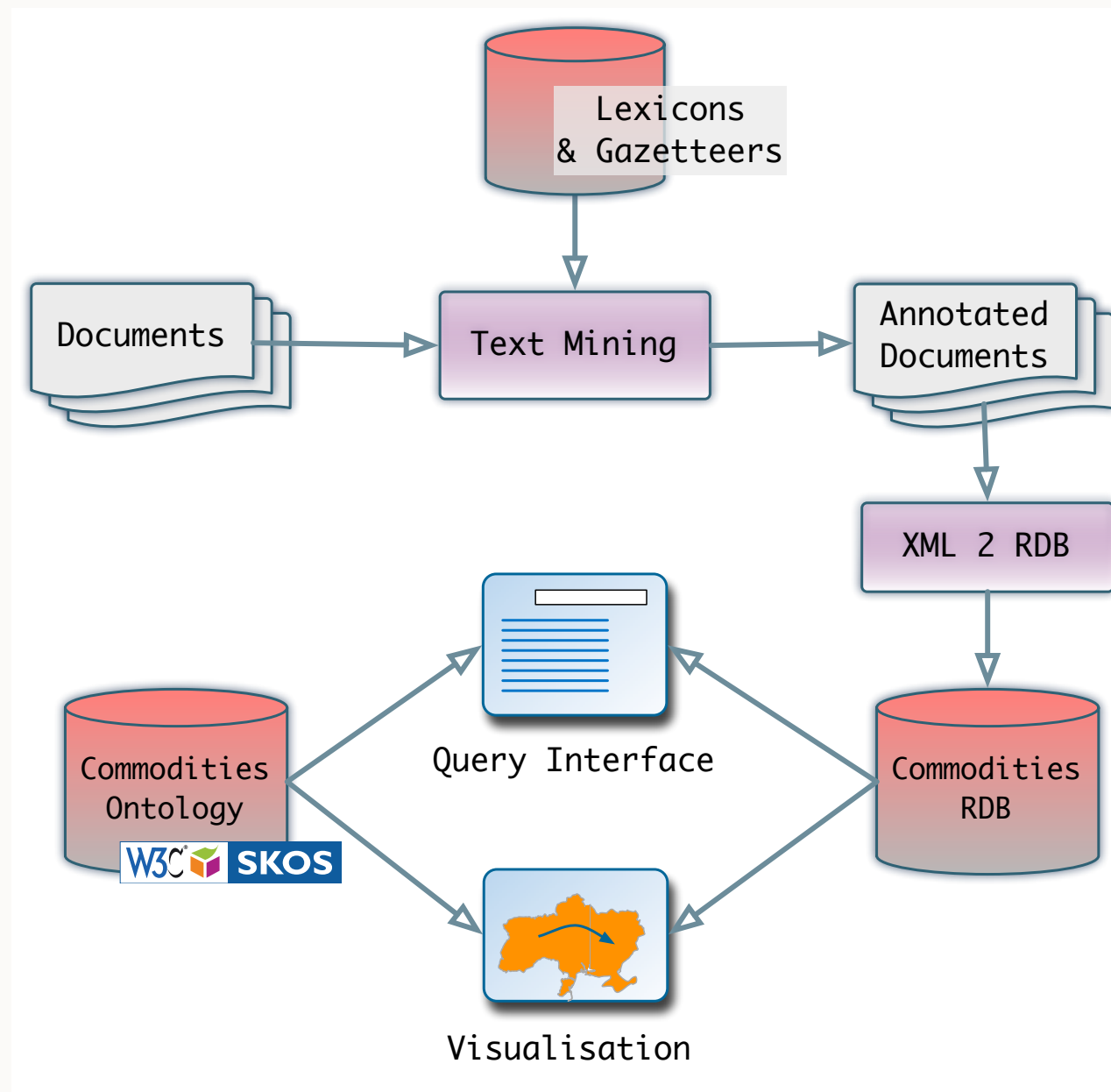
Collection	# of Documents	# of Images
House of Commons Parliamentary Papers (ProQuest)	118,526	6,448,739
Early Canadiana Online	83,016	3,938,758
Directors' Letters of Correspondence (Kew)	14,340	n/ a
Confidential Prints (Adam Matthews)	1,315	140,010
Foreign and Commonwealth Office Collection	1,000	41,611
Asia and the West (Gale)	4,725	948,773 (OCRRed: 450,841)

DOCUMENT COLLECTIONS

Collection	# of Documents	# of Images
House of Commons Parliamentary Papers		
Early Census		
Directorial Correspondence		
Confidential Papers		
Commons Papers		
Asia and the West (Gale)	4,725	948,773 (OCR'd: 450,841)

Over 10 million document pages,
Over 7 billion word tokens.

SYSTEM



MINED INFORMATION

- Example sentence:

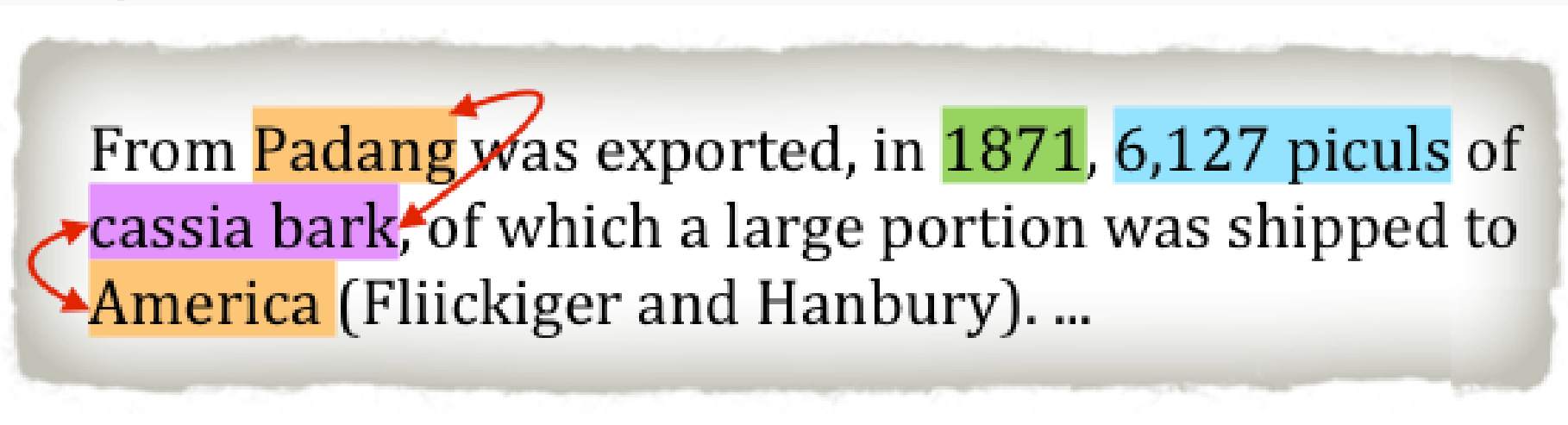
From Padang was exported, in 1871, 6,127 piculs of cassia bark, of which a large portion was shipped to America (Flickiger and Hanbury). ...

- Normalised and grounded entities:

- commodity: cassia bark [concept: Cinnamomum cassia]
- date: 1871 (year=1871)
- location: Padang (lat=-0.94924;long=100.35427;country=ID)
- location: America (lat=39.76;long=-98.50;country=n/a)
- quantity + unit: 6,127 piculs

MINED INFORMATION

- Example sentence:



From Padang was exported, in 1871, 6,127 piculs of cassia bark, of which a large portion was shipped to America (Flickiger and Hanbury). ...

The diagram shows the sentence with several entities highlighted in colored boxes: 'Padang' (orange), '1871' (green), '6,127 piculs' (blue), 'cassia bark' (purple), and 'America' (orange). Red arrows indicate relationships: one arrow points from 'Padang' to 'cassia bark', and another points from 'cassia bark' to 'America'.

- Extracted entity attributes and relations:

- origin location: Padang
- destination location: America
- commodity–date relation: cassia bark – 1871
- commodity–location relation: cassia bark – Padang
- commodity–location relation: cassia bark – America

EDINBURGH GEOPARSER



only 50 objects displayed, zoom in or deselect some features

	Name	country	feature	km to center
1	Ciudad Victoria	Mexico	seat of a first-order administrative division	11435.01 km
2	Victoria	Seychelles	capital of a political entity	5126.1 km
3	Victoria	Canada	seat of a first-order administrative division	11180.98 km
4	State of Victoria	Australia	first-order administrative division	14782.65 km
5	Hong Kong	Hong Kong	capital of a political entity	9688.28 km
6	Victoria	Malaysia	seat of a first-order administrative division	10558.79 km
7	Durango	Mexico	seat of a first-order administrative division	11882.99 km
8	Victoria	Malta	seat of a first-order administrative division	1494.57 km
9	Victoria	Honduras	second-order administrative division	10832.96 km
10	Victoria	United States	seat of a second-order administrative division	10954.12 km



OCR ERRORS

Proclamations, Pro * v mie RL' E.LI S B.AIG07.
iVICTORlfIA. h>I l/ t(a' of' GO!>. tif ih Firi.
ea fil~/ r<' lluil'tI', (i'. i' , QUEE'. Tc ,iii-
n iVi i ' ui tillh't, 111te 1 eih' Colin. ('it;ZI-s.
uni 14Lt1ussuls ce t rib ev iii tJlu stat. it have lei
t's.iiitit'ud ztntt liild, a tutt < A 11i10C.

Extract of Early Canadiana Online document 9_00952_3, p. vi.

OCR ERRORS



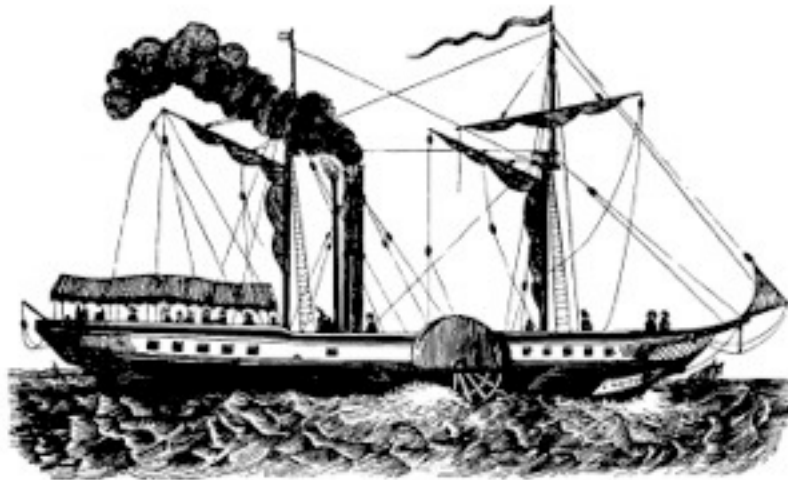
OCR ERRORS



LESSONS LEARNED

- Importance of two-way collaboration between technology and humanities expert in digital HSS projects.
- Value of iterative development and rapid prototyping.
- Geo-referencing text is very important for historical analysis.
- Most OCR errors are noise in big data but HSS scholars need to be made more aware of OCR errors affecting their search results for historical collections.

THANK YOU



Trading Consequences

- Contact: balex@inf.ed.ac.uk
- Website: <http://tradingconsequences.blogs.edina.ac.uk/>
- Online user interface launch: 28/02/2014.