Enabling Digital history:

Text mining big historical document collections on trade in the British Empire

THE UNIVERSITY of EDINBURGH informatics

Beatrice Alex

balex@inf.ed.ac.uk

Trading Consequences

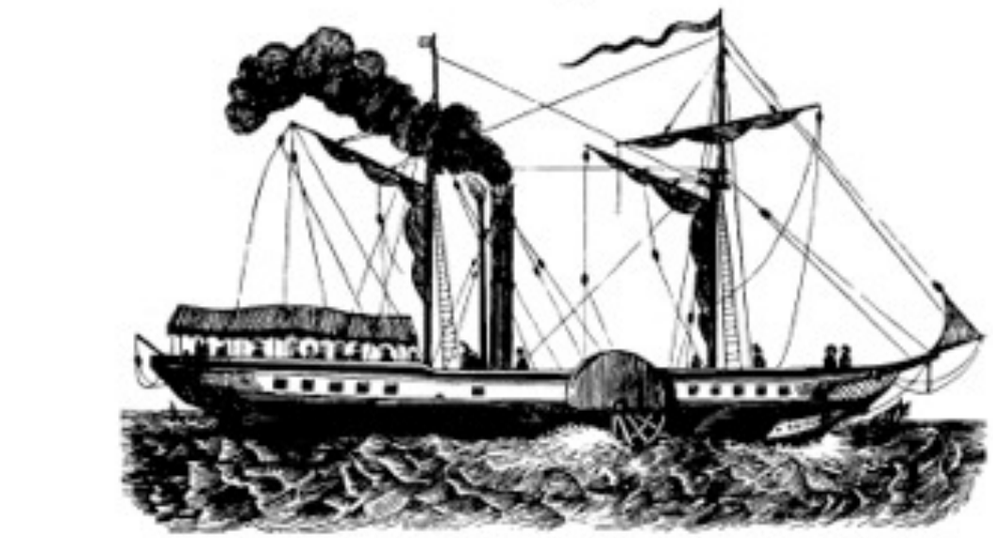
PQIS All Team Meeting, ProQuest, April 23rd 2014

TEXT MINING

- Describes a set of linguistic, statistical and/or machine learning techniques that model and structure the information content of textual resources.
- Turns unstructured text into structured data (e.g. relational database or linked data).
- Is very useful for analysing large text collections automatically (overcoming data paralysis).
- Goal in DHSS research: By analysing large amounts of textual data, help HSS scholars to discover novel patterns and explore hypotheses.

TYPES OF ANALYSES

- **Named entity recognition.**
- **Grounding, e.g. geo-referencing.**
- **Relation extraction.**
- Clustering, e.g. topic modelling.
- Sentiment analysis.



Trading Consequences

Digging into Data II

PQIS All Team Meeting, ProQuest, April 23rd 2014

PROJECT TEAM

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

Colin Coates, Andrew Watson: *historical analysis*

Jim Clifford: *historical analysis*

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

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

PROJECT GOALS

- 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.

DOCUMENT COLLECTIONS

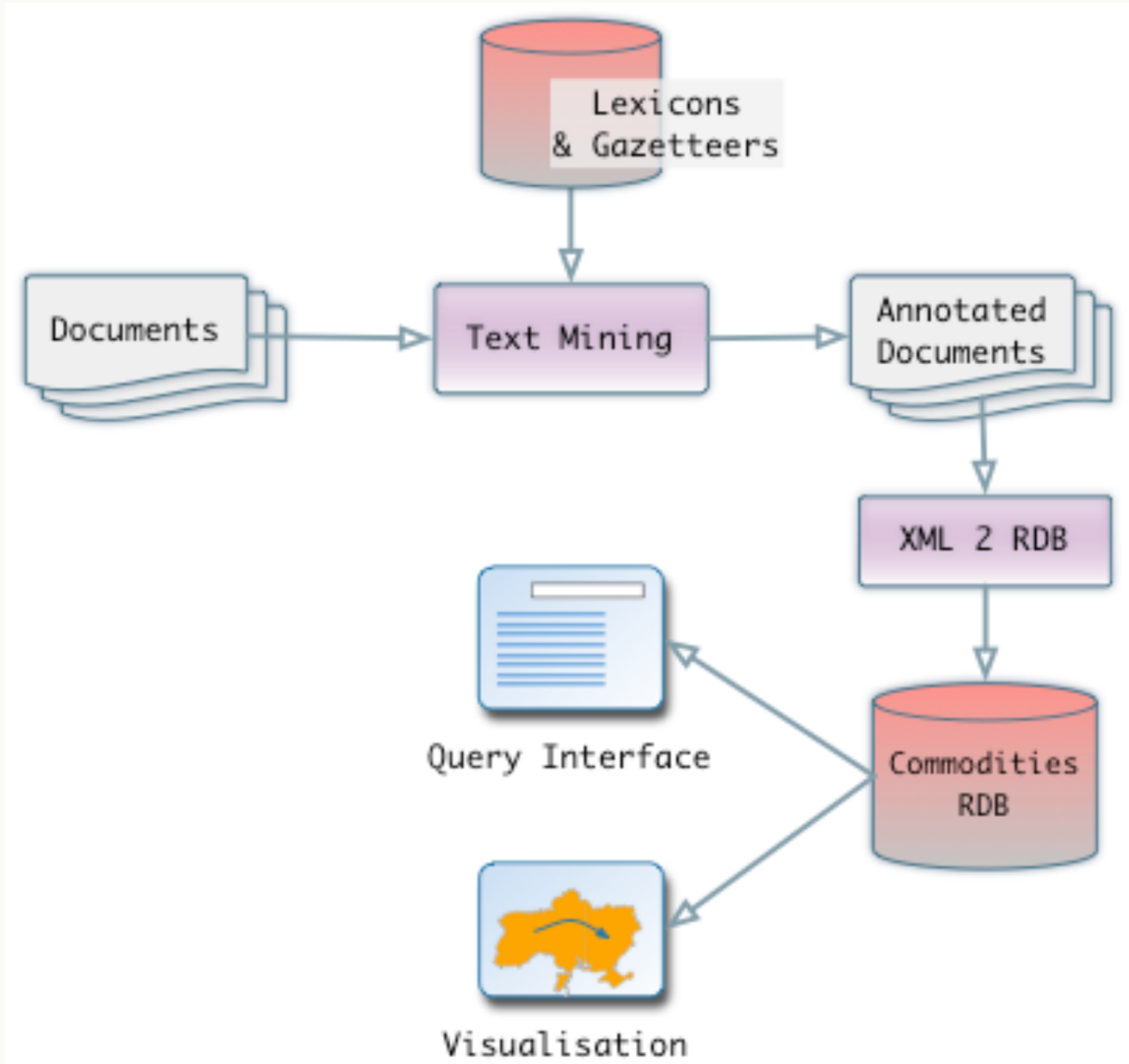
Collection	# of Documents	# of Images
House of Commons Parliamentary Papers (ProQuest)	118,526	6,448,739
Early Canadiana Online	83,016	3,938,758
Directors' Letters of Correspondence (Kew)	14,340	n/a
Confidential Prints (Adam Matthews)	1,315	140,010
Foreign and Commonwealth Office Collection	1,000	41,611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

DOCUMENT COLLECTIONS

Collection	# of Documents	# of Images
House of Commons Parliamentary Collection		
Early C		
Dire Corres		
Confide		
Comm		
Asia and the West (Gale)	4,725	948,773 (OCRed: 450,841)

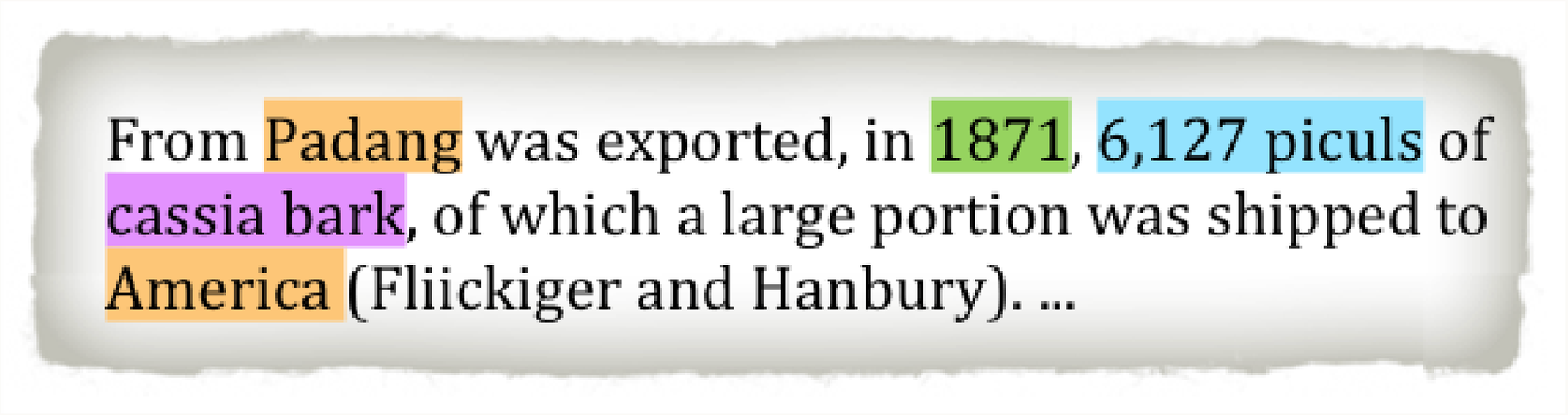
Over 10 million document pages,
Over 7 billion word tokens.

ARCHITECTURE



MINED INFORMATION

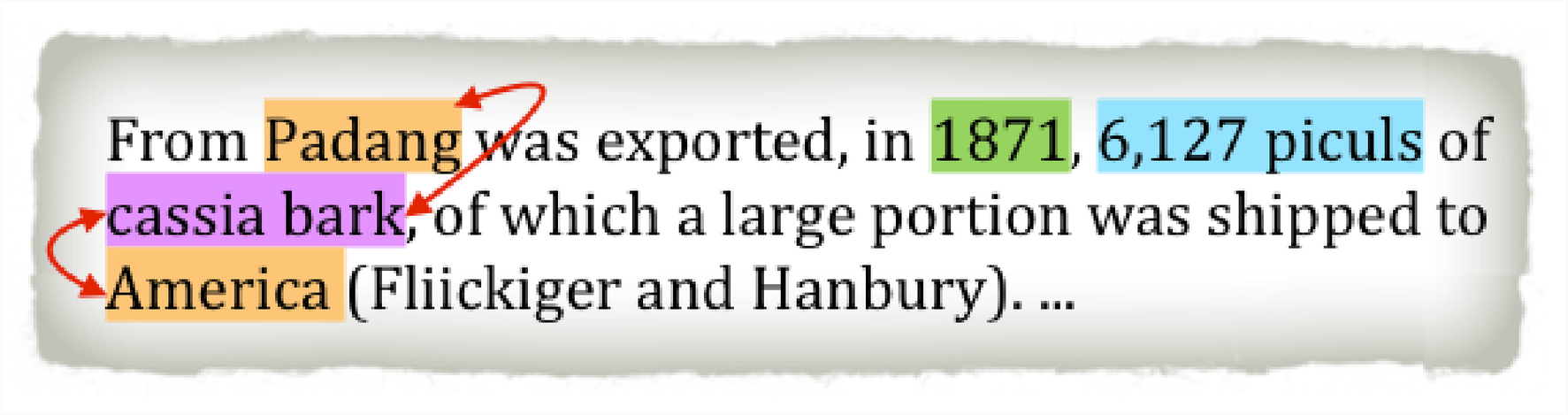
- Example sentence:



- 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:



- 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

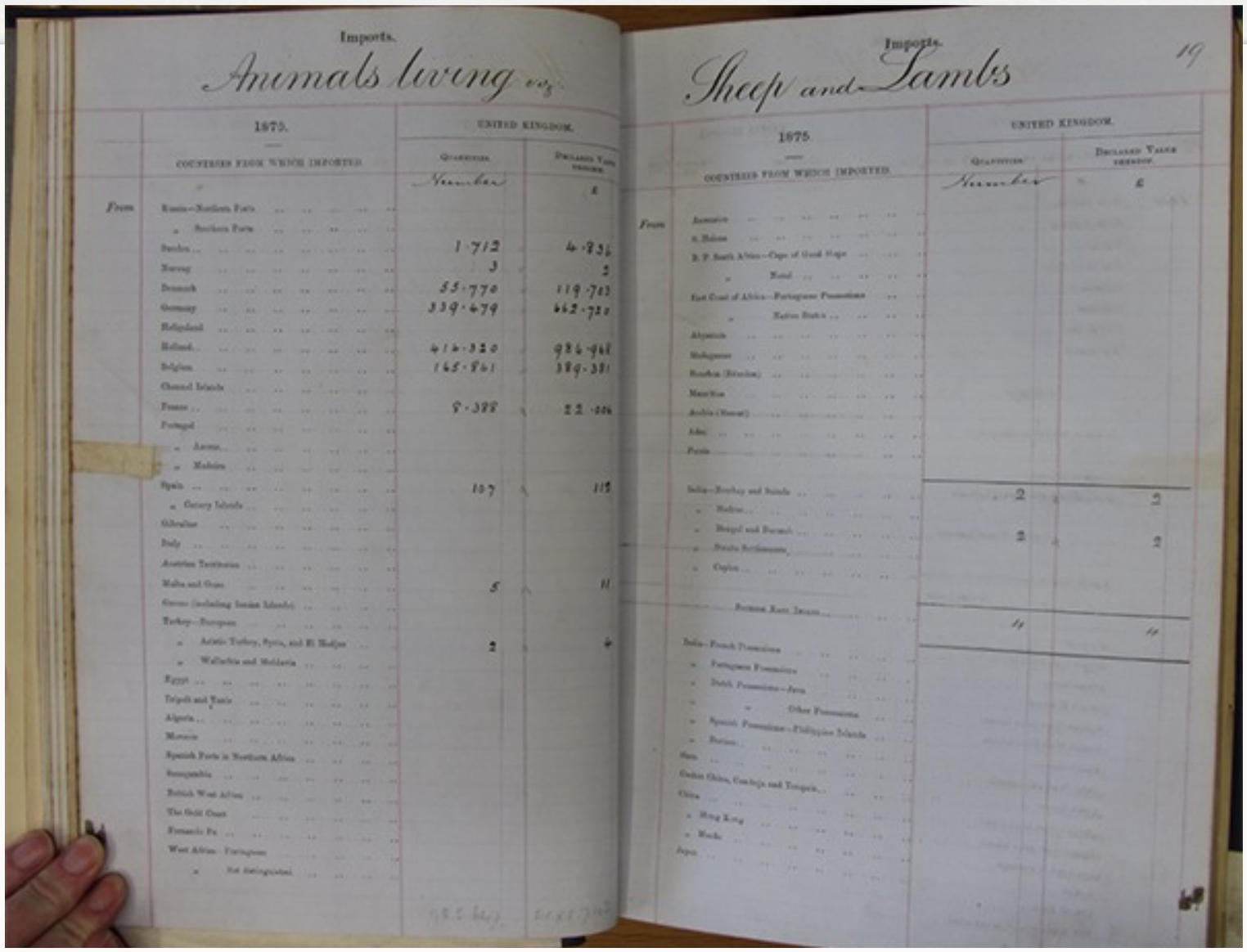
EDINBURGH GEOPARSER

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

PQIS All Team Meeting, ProQuest, April 23rd 2014

COMMODITY LEXICON



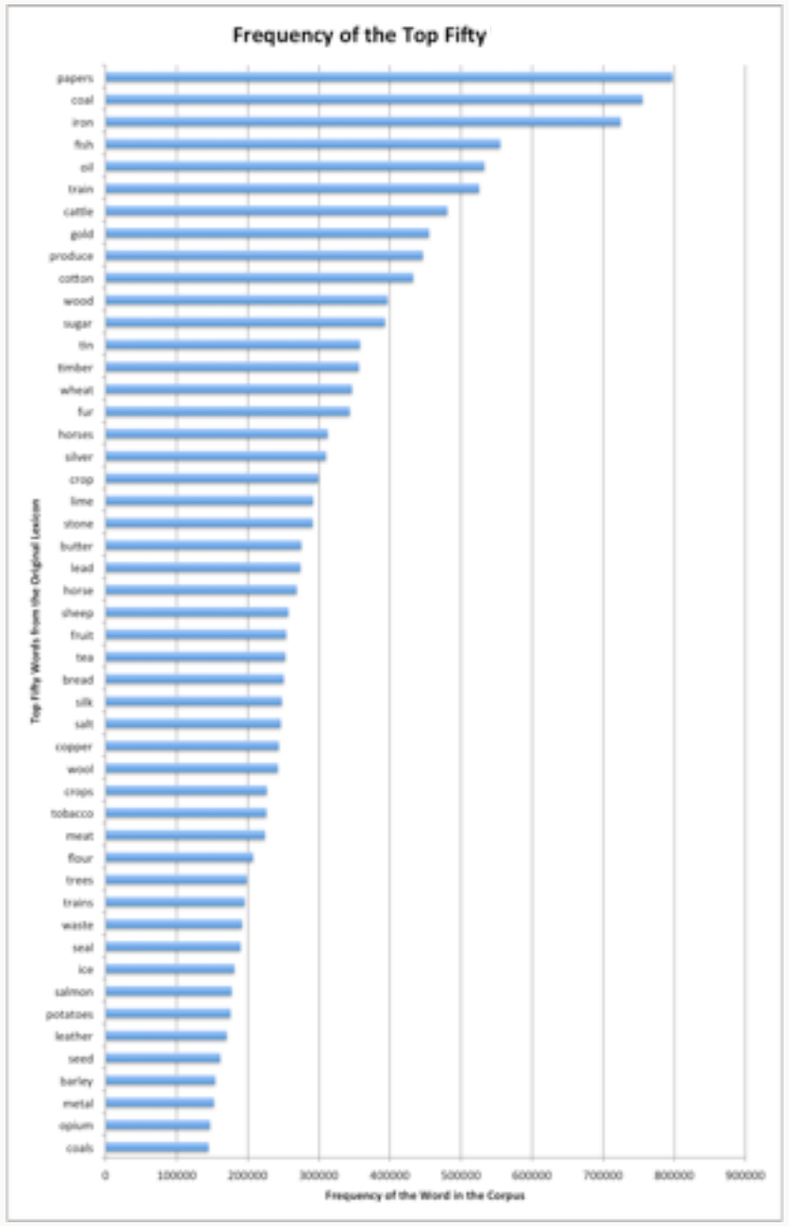
- Seed set from customs import records.

LEXICON CREATION

Seed lexicon	~600
Extended lexicon	~17,000
With pluralisation of single word entries	~20,500

Bootstrapping a historical commodities lexicon with SKOS and DBpedia. Klein, Alex & Clifford, LaTeCH 2014.

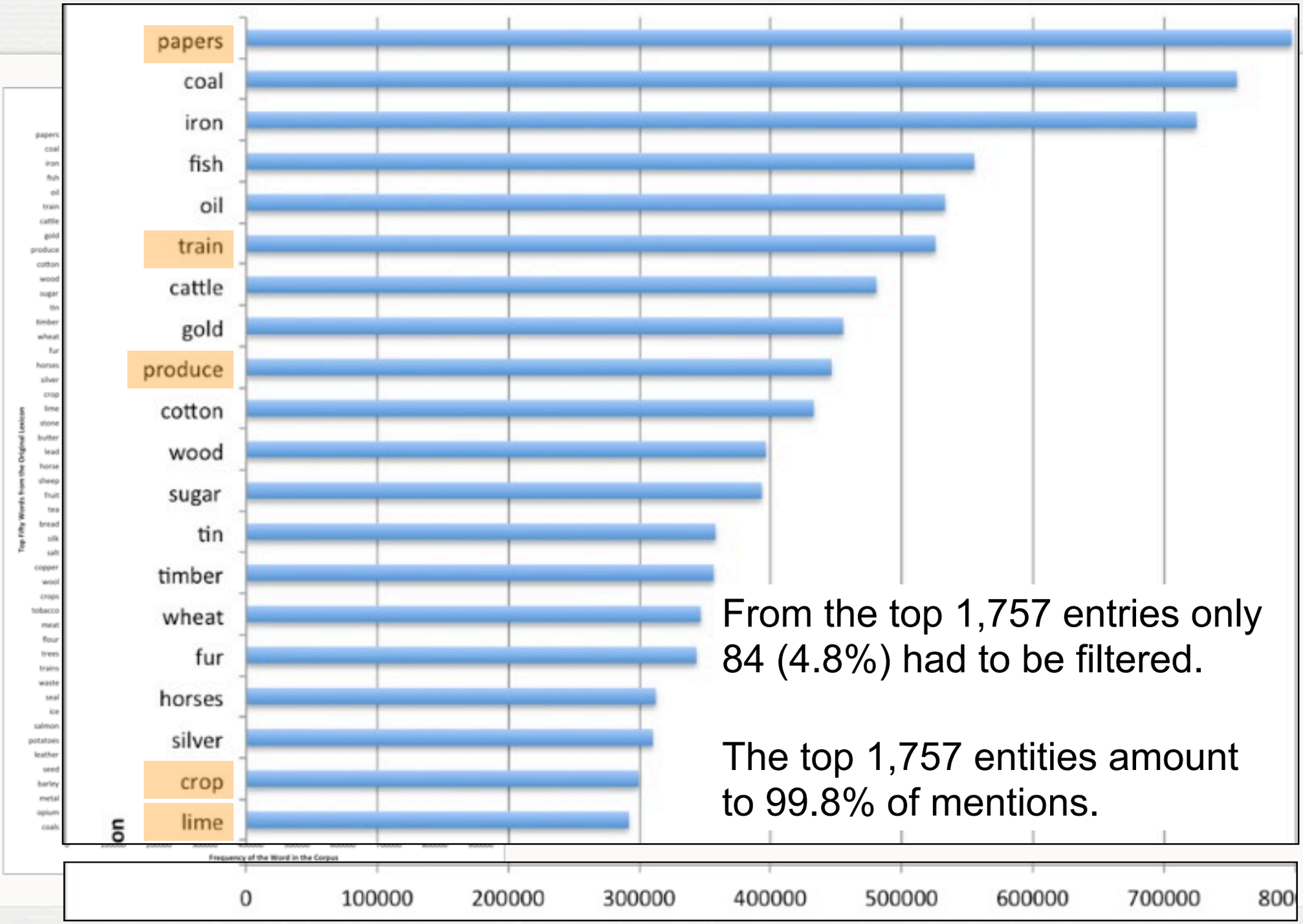
LEXICON CLEAN-UP



From the top 1,757 entries only 84 (4.8%) had to be filtered.

The top 1,757 entities amount to 99.8% of mentions.

LEXICON CLEAN-UP



From the top 1,757 entries only 84 (4.8%) had to be filtered.

The top 1,757 entities amount to 99.8% of mentions.

NOISY DATA

- Optical character recognition contains many errors and often the structure of the page layout is lost.
  - Sophistication of the OCR engine and scanning equipment.
  - Quality of the original print and paper.
  - Use of historical language.
  - Information in page margins (header, page numbers, etc.).
  - Information in tables.
  - Language of the text.

FIXING NOISY DATA

- Text normalisation and correction:
- End-of-line soft hyphen removal
  - Dehyphen all token-splitting hyphens using a dictionary-based approach.
- “False f”-to-s conversion
  - Convert all false f characters to s using a corpus.
- Example: reduced number of words unrecognised by spell checker from 61 to 21 -> 67%, on average 12% reduction in word error rate in a random sample (Alex et al., 2012).

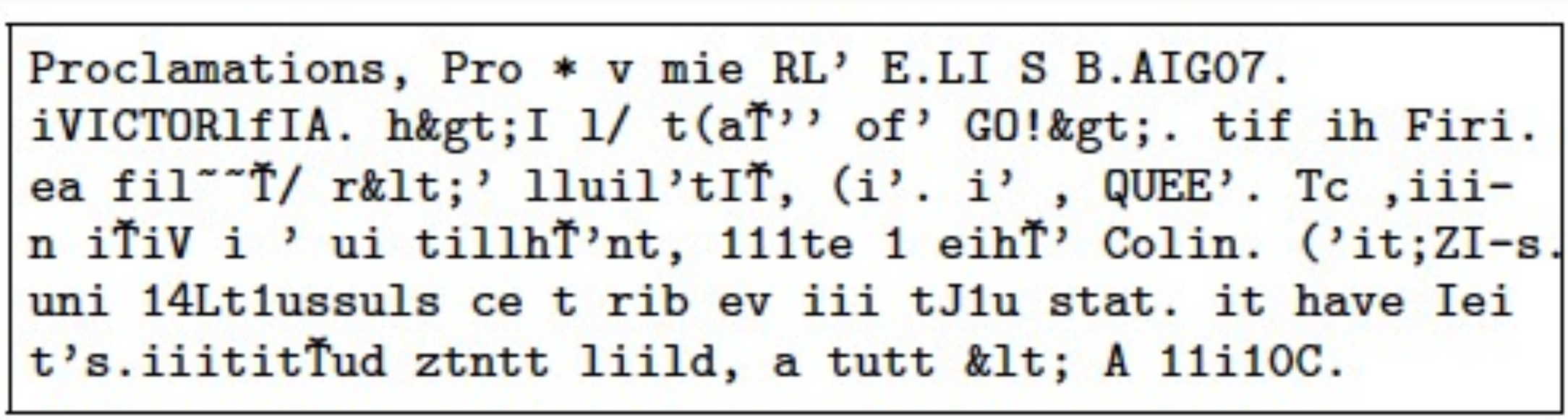
FIXING NOISY DATA

```
<?xml version="1.0" encoding="UTF-8"?>
<article id="10.2307/60227644">
 <page> <![CDATA[THE HISTORY OP THE POLITICKS OF GREAT BRITAIN AND FRANCE, VINDI GATED FROM A LATE ATTACK OF
MR. WILLIAM BELSHAM. BY HERBERT MARSH, B. D. F. R. S. and tellow or st. John's college, Cambridge. Ecntmn:
PRINTED FOR JOHN STOCKDALE, PICCADILLY. 1801. t35)lvjf~ Udf4~ P.]]> </page>
<page> <![CDATA[T, G1IXET, Printer.]]> </page>
<page> <![CDATA[' INTRODUCTION, AS the following Vindication may fall into the hands of perfons who have
never read the Hiftory of the Politicks of Great Britain and France, it will not be improper, before I enter
on my Defence, to ftate the principal facts, which were fuccef- fively proved by authentic documents, in the
fixteen chapters, of which that wrork is compofed. - r 1. In the celebrated conference at PiUnitz; in Auguft,
i;gi, the Britifh Government took not the rnoft diftant part: and if.-any treaty was concluded there, which
is itfelf a matter of great doubt, the Britifh Go¬ vernment not only never acceded to it, but was,never
apprifed even of its contents.- Further, when the Britiih Government was requefted in 1701 to join a
coalition againft France, it gave a pofitive and unequivocal refufal. B 2 2. Toward]]> </page>
<page> <![CDATA[4 2. Toward the clofe of the fame year the valuable colony of St. Domingo was pre¬ served to
France by the timely affiftance fent by Lord Effingham, then Governor of Jamaica : and the Britifh. Cabinet
fignified through its AmbafTador at Paris to the French Government, that it fully approved of Lord
Effingham's conduct.. At the fame time, true to the ftri&eir. principles of ho¬ nour and neutrality, it
refufed the advan- tageous offer made by the French colonifts, who were highly diflatisfied with the Na¬
tional AfTembly, to furrender the French part of St. Domingo to the Crown of Bri¬ tain. And thefe a6ls of
generofity were re* paid by France with the utmoft ingrati¬ tude. 3. When Louis XVI. formally accepted the
new conflitution, in September, 17Q1, and fent circular letters to the different Courts of Europe fignifying
his affent, the Court of Great Britain was one of the firft which returned an anfwer ; and the anfwer was
couched in very refpeclful terms, where¬ as fome other courts either did not anfwer at HfWta]]> </page>
...
 </article>
```

FIXING NOISY DATA

```
<document>
<meta>
<attr name="docid">10.2307/60227644</attr>
</meta>
<text>
 <p>THE HISTORY OP THE POLITICKS OF GREAT BRITAIN AND FRANCE, VINDI GATED FROM A LATE ATTACK OF MR. WILLIAM BELSHAM. BY HERBERT MARSH, B. D. F. R. S. and tellow or st. John's college, Cambridge. Ecntmn: PRINTED FOR JOHN STOCKDALE, PICCADILLY. 1801. t35)lvjf~ Udf4~ P.</p>
<p>T, G1IXET, Printer.</p>
<p>' INTRODUCTION, AS the following Vindication may fall into the hands of persons who have never read the History of the Politicks of Great Britain and France, it will not be improper, before I enter on my Defence, to state the principal facts, which were successively proved by authentic documents, in the sixteen chapters, of which that wrork is composed. r 1. In the celebrated conference at PiUnitz; in August, i;gi, the British Government took not the rnoft distant part: and if.-any treaty was concluded there, which is itself a matter of great doubt, the British Government not only never acceded to it, but was,never apprised even of its contents.- Further, when the Britiih Government was requested in 1701 to join a coalition against France, it gave a positive and unequivocal refusal. B 2 2. Toward</p>
<p>4 2. Toward the close of the same year the valuable colony of St. Domingo was preserved to France by the timely assistance sent by Lord Effingham, then Governor of Jamaica : and the British. Cabinet signified through its AmbafTador at Paris to the French Government, that it fully approved of Lord Effingham's conduct.. At the same time, true to the ftri&amp;amp;eir. principles of honour and neutrality, it refused the advantageous offer made by the French colonists, who were highly diflatisfied with the National AfTembly, to surrender the French part of St. Domingo to the Crown of Britain. And these a6ls of generosity were re* paid by France with the utmost ingratitude. 3. When Louis XVI. formally accepted the new conflitution, in September, 17Q1, and sent circular letters to the different Courts of Europe signifying his assent, the Court of Great Britain was one of the first which returned an answer ; and the answer was couched in very refpeclful terms, whereas some other courts either did not answer at HfWta</p>
 </text>
</document>
```

OCR ERRORS


Proclamations, Pro * v mie RL’ E.LI S B.AIGO7.
iVICTORlfIA. h>I l/ t(aŤ’’ of’ GO!>. tif ih Firi.
ea fil˜˜Ť/ r<’ lluil’tIŤ, (i’. i’ , QUEE’. Tc ,iii-
n iŤiV i ’ ui tillhŤ’nt, 111te 1 eihŤ’ Colin. (’it;ZI-s.
uni 14Lt1ussuls ce t rib ev iii tJ1u stat. it have Iei
t’s.iiititŤud ztntt liild, a tutt < A 11i10C.

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

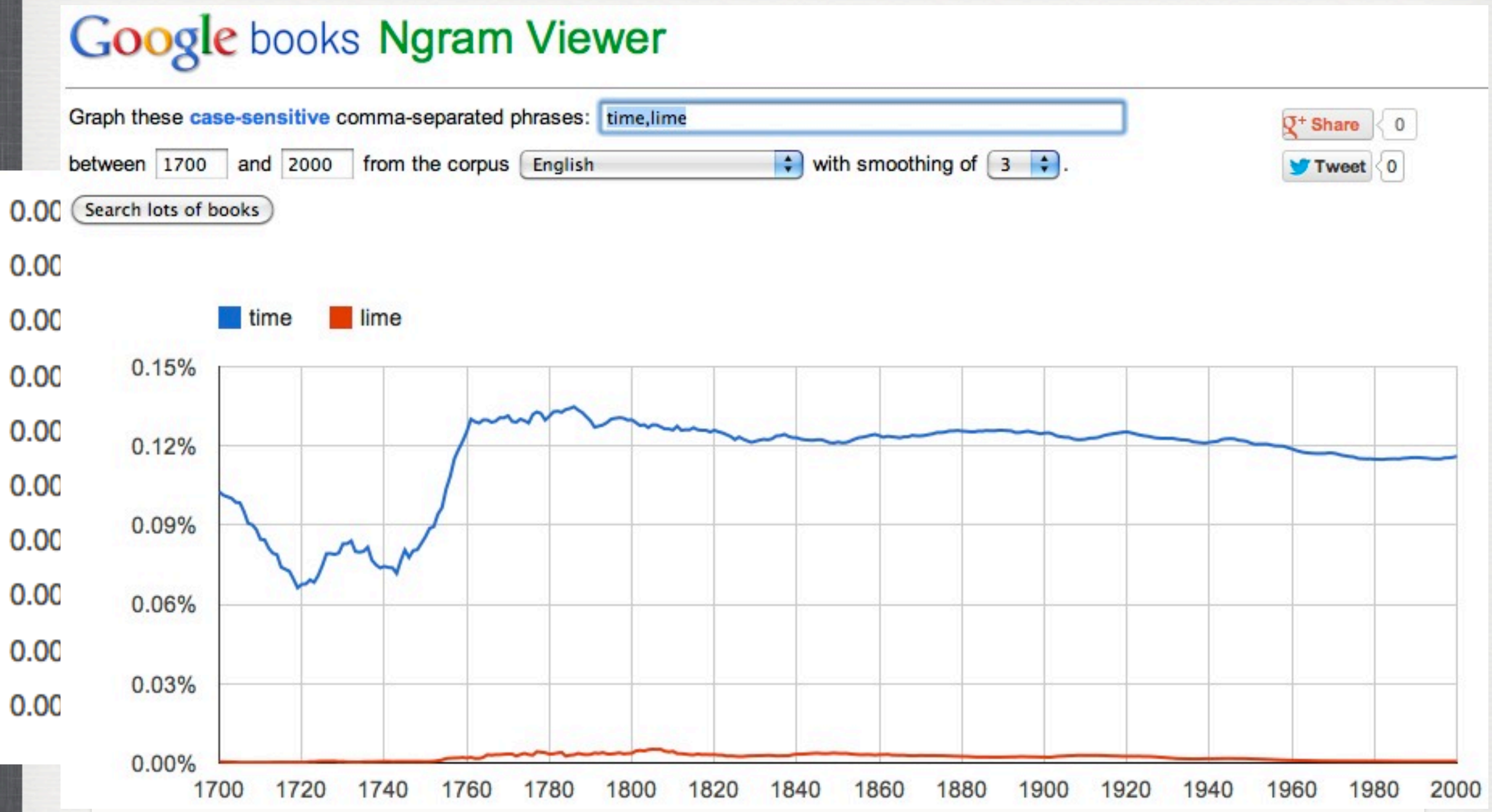
OCR ERRORS


0.000000600%
0.000000550%
0.000000500%
0.000000450%
0.000000400%
0.000000350%
0.000000300%
0.000000250%
0.000000200%
0.000000150%
0.000000100%
0.000000050%
0.000000000%
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
cattlc
Cattlc

OCR ERRORS


0.00900%
0.00800%
0.00700%
0.00600%
0.00500%
0.00400%
0.00300%
0.00200%
0.00100%
0.00000%
1760
1780
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
cattle (All)
cattlc (All)

OCR ERRORS


Google books Ngram Viewer
Graph these case-sensitive comma-separated phrases: time,lime
between 1700 and 2000 from the corpus English with smoothing of 3 .
Share 0
Tweet 0
Search lots of books
time
lime
0.15%
0.12%
0.09%
0.06%
0.03%
0.00%
1700 1720 1740 1760 1780 1800 1820 1840 1860 1880 1900 1920 1940 1960 1980 2000

HOW NOISY IS TOO NOISY?

qBiu si }S3A:req s,uauuaqsu aq} }Bq} uirepo.ifT
'papua}X3 sSuiav }qSuq Jiaq} qiiM jib ui snnS bbs aqx
'a"3(s aq} tnojj ssfitns q}TM Sni5[ooi si jb}s }S.ii; aqx
'papnaoSB q}Bq naABSjj qS;H °1 ssbui s.uauuaqsu aqx

Extract from document
10.2307/60238580 in FCOC.

Brave Dermot looks up with a spirit all lowly,
And, wiping the brine from his storm-beaten head,
Cries "Mother of Jesus, and saints the most holy,
Befriend!—Sure, the fishermen's mass it was said."

Upon the dark storm, from the fisherman's dwelling,
There shines a faint light like a star through a cloud;
Fast hugging her babe to her bosom high swelling,
Tight clasping her rosary, Mary's low bowed.

She looks through her tears to her Father in heaven,
Then, kissing her child, cries, "Ah, bless'd be that smile!
And bless'd be the saints! now the promise is given!
Sure, mass it was said—he'll he here in a while."

Again and again with fond kisses bedewing
Her baby's soft cheek, as the rosebud as fair;
Again and again her look upward renewing,
This song, sweet and solemn, she breathes on the air:

When upon the stormy deep,
Fishermen their vigils keep,
Wherefore should their lovers weep?
Thou art there.
When the storm is raging high,
And the waves leap to the sky,
Wherefore should we fret and sigh?
Thou dost care.

Stronger than the tempest's rage
Is Thy power from age to age;
Thou didst once its wrath assuage,
And can still.
We will own Thy mighty power,
Even in the darkest hour,
And receive as blessed dower
All Thy will.

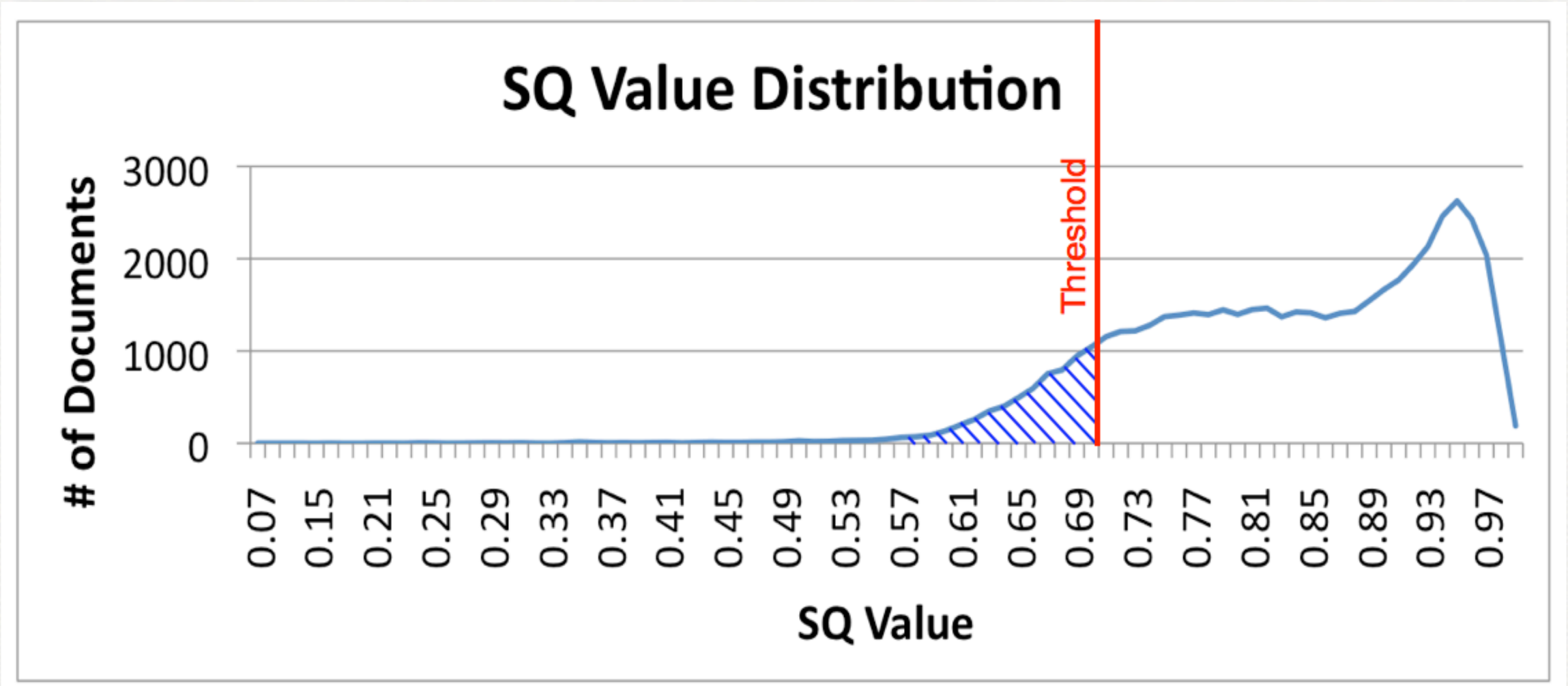
The Fishermen's Mass.

By fair Tralee bay, oh! did you e'er wander,
Where quiet and beauty are blended in one,
Watching the brooklets their pearly drops squander,
Bright gems soon to form for the brow of the sun?
And have you e'er heard the swell of devotion,
Like chorus of angels from depths of the sea,
As mass hath been said by those sons of the ocean,
The fishermen humble of quiet Tralee?

Oh! I have been there, and seen them all kneeling,
Their faces turned upward beneath the bright sky,
And heard their low throbs of worshipping feeling,
And seen pure devotion gush out from each eye;
And oft, 'mid the worlds rough battles and sorrow,
I've thought of that fishermen's mass of the sea,
And pray'd for such faith in brighter to-morrow
As shone from each face in the bay of Tralee.

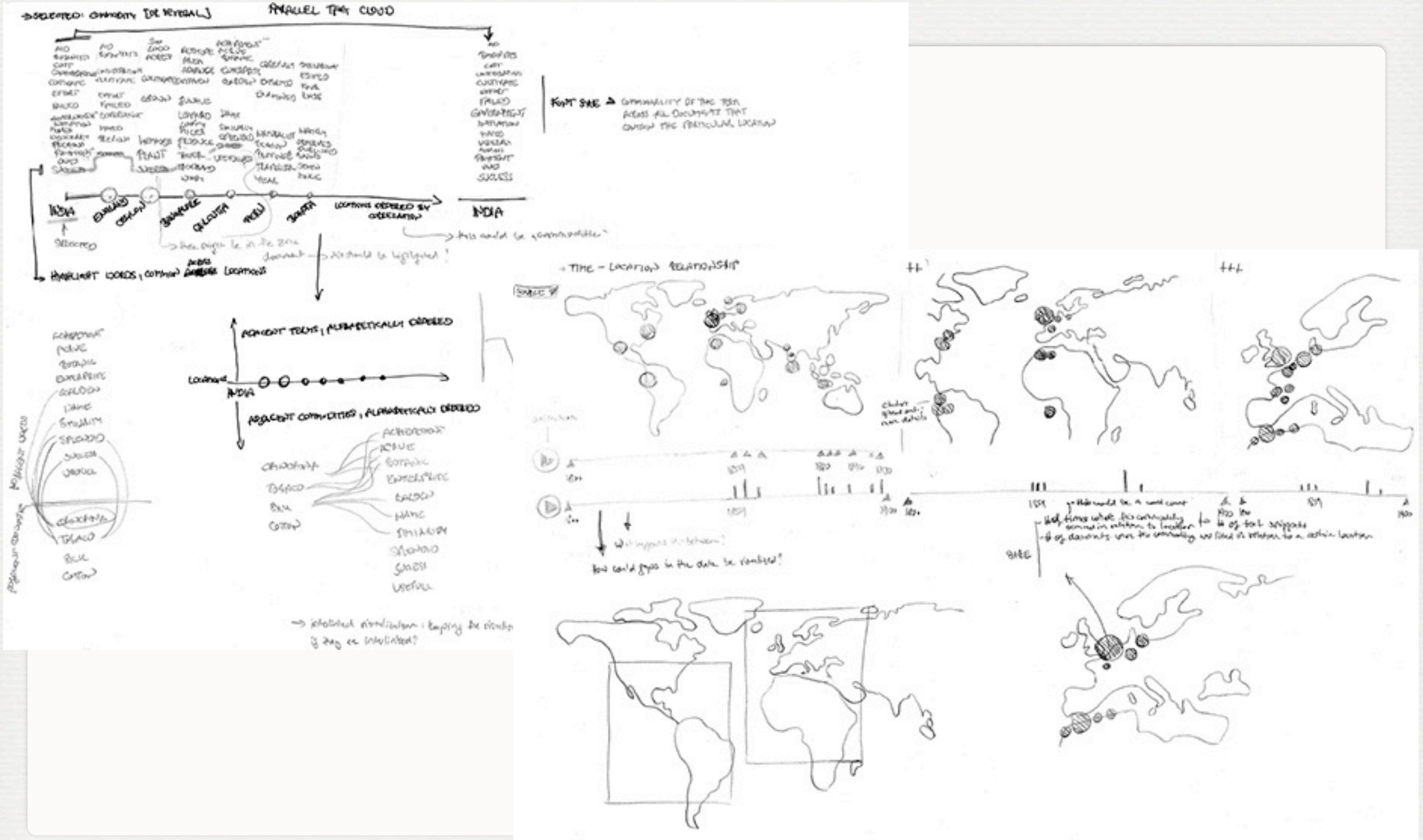
The fishermen's mass to High Heaven hath ascended,
The first star is looking with smiles from the sky,
The sea gulls in air with their bright wings extended,
Proclaim that the fishermen's harvest is nigh.

OCR ERRORS



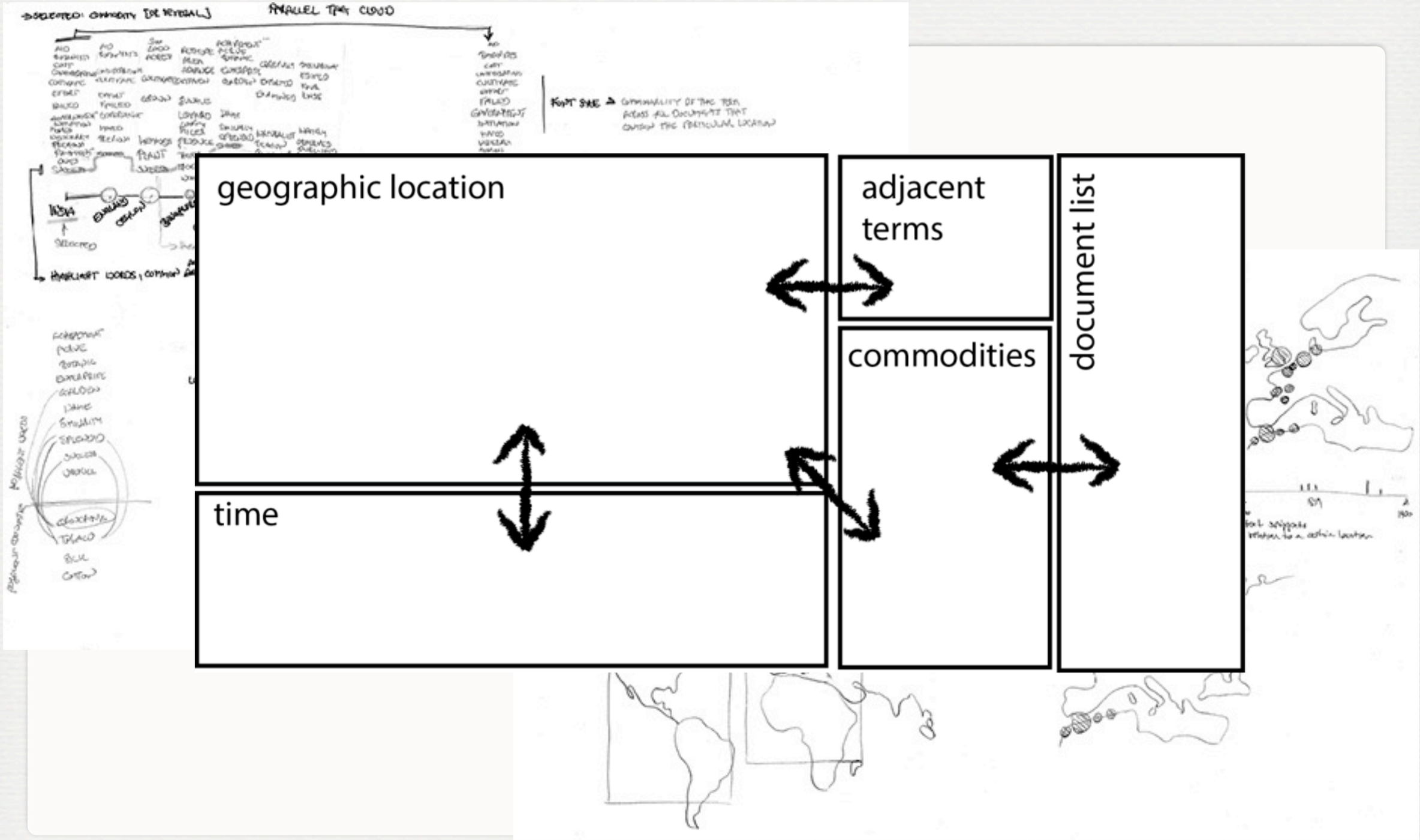
- Study of correlating manual quality ratings of documents with automatic quality scoring (Alex & Burns, DATeCH 2014).

VISUALISATION SKETCHES



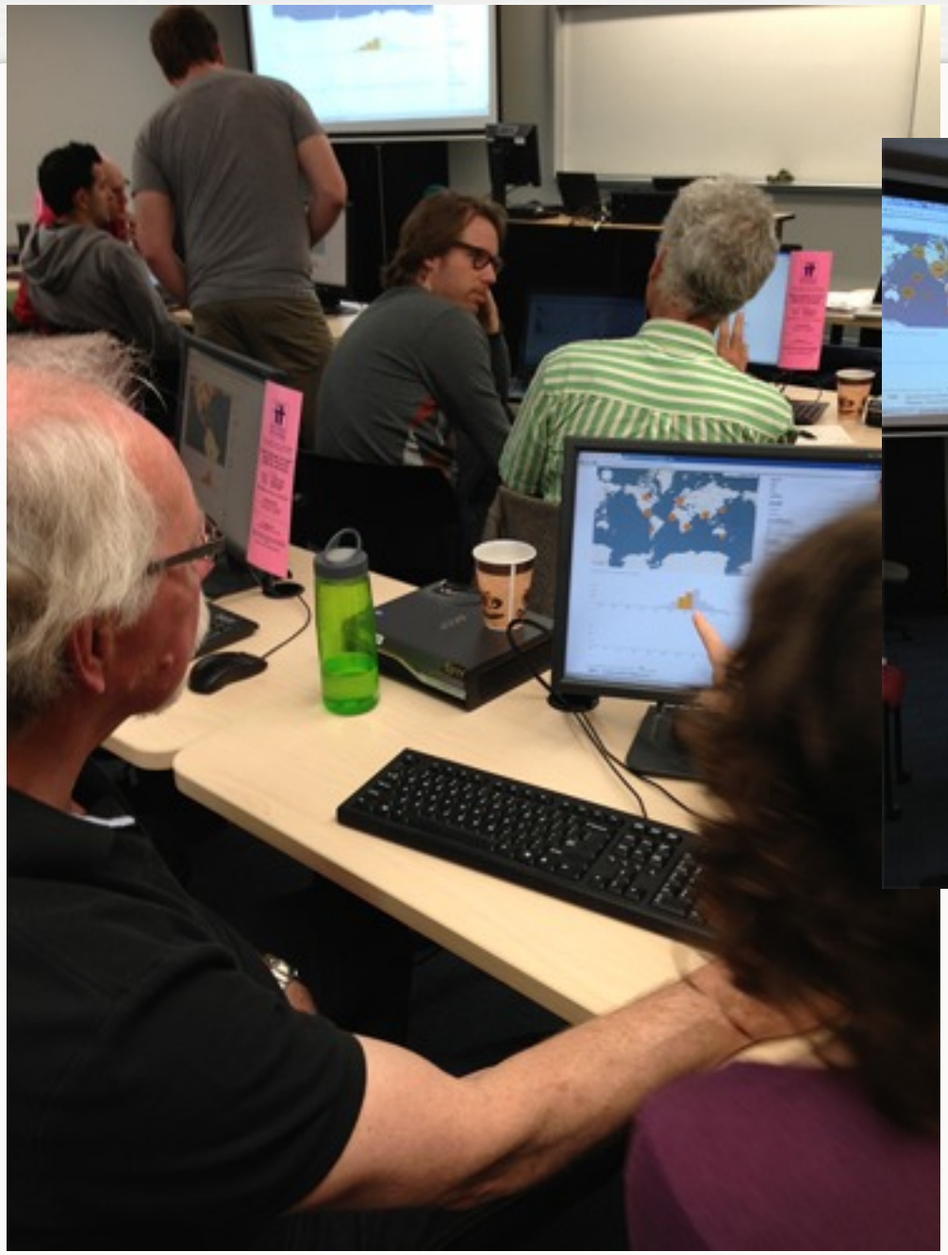
PQIS All Team Meeting, ProQuest, April 23rd 2014

VISUALISATION SKETCHES



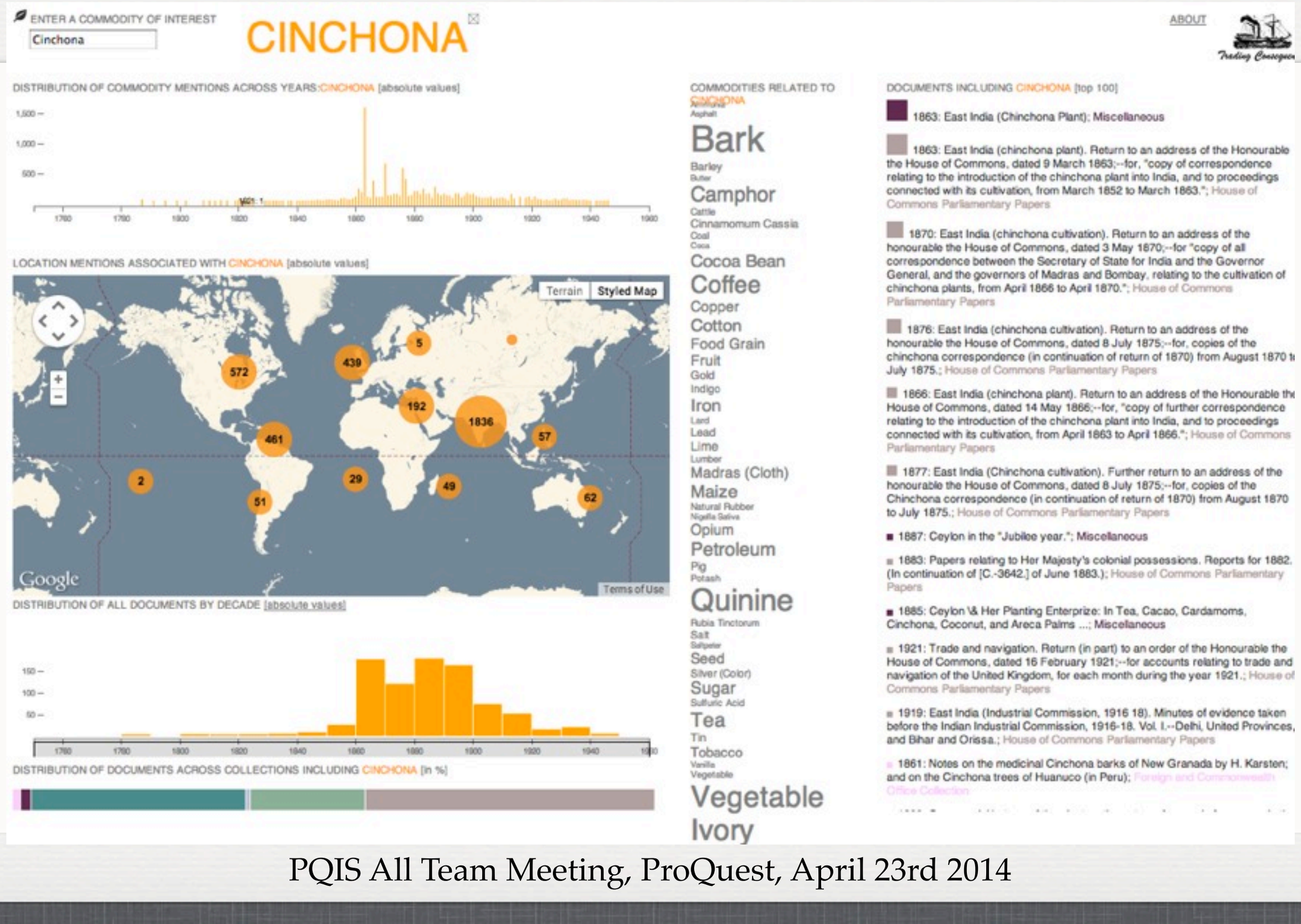
PQIS All Team Meeting, ProQuest, April 23rd 2014

USER WORKSHOP



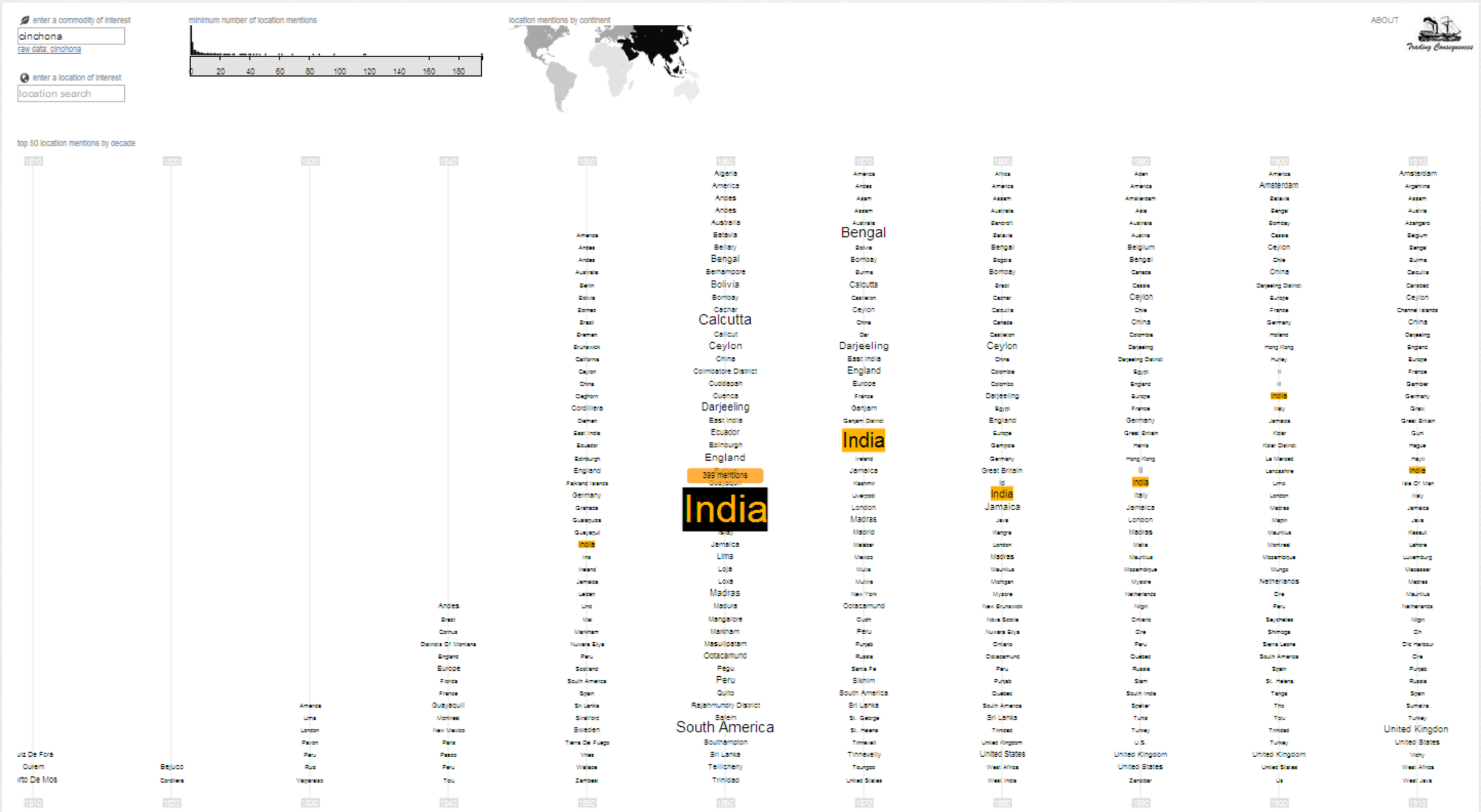
User workshop to improve the functionality of the interface (Hinrichs et al., 2014)

BRINGING ARCHIVES ALIVE



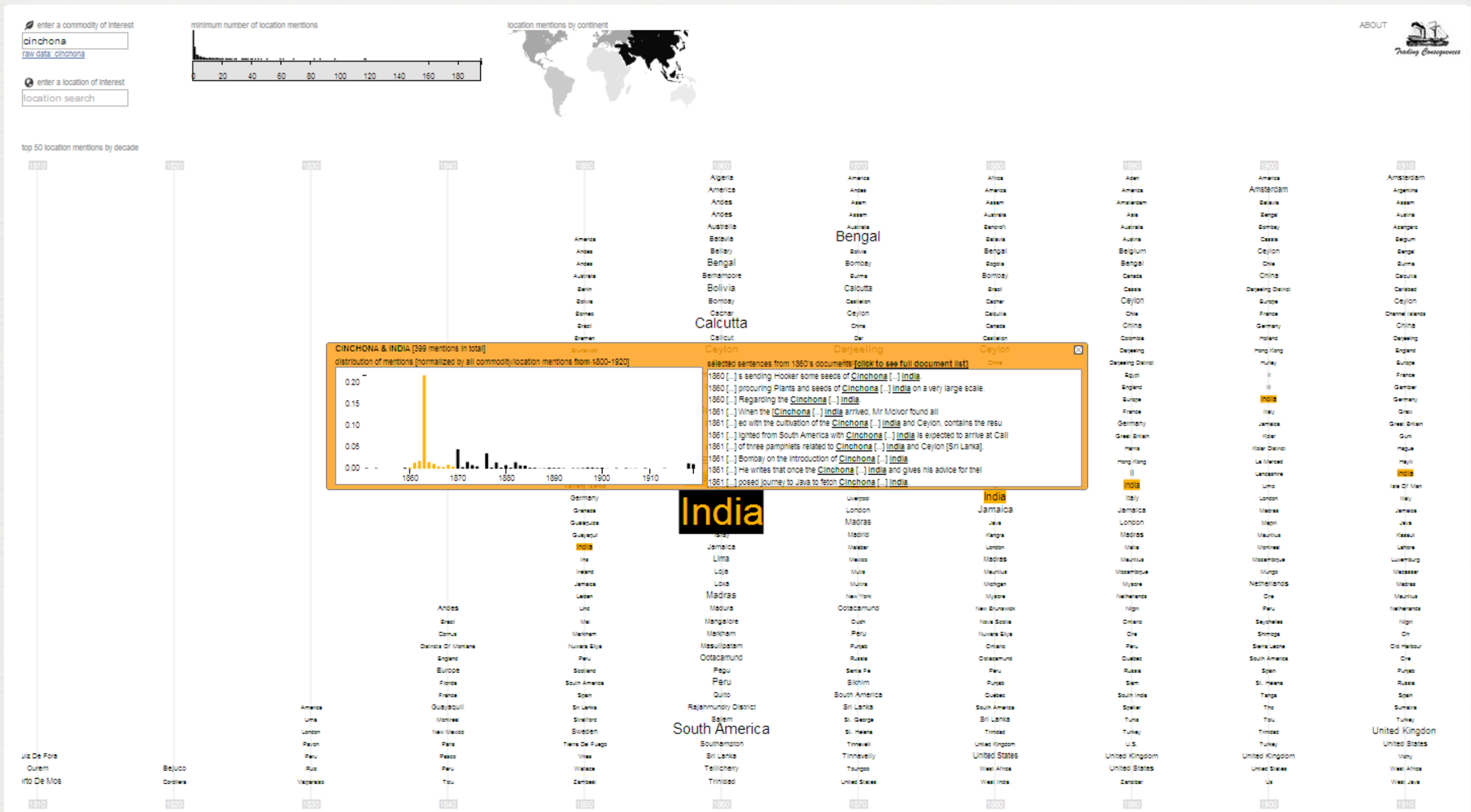
PQIS All Team Meeting, ProQuest, April 23rd 2014

BRINGING ARCHIVES ALIVE



PQIS All Team Meeting, ProQuest, April 23rd 2014

BRINGING ARCHIVES ALIVE

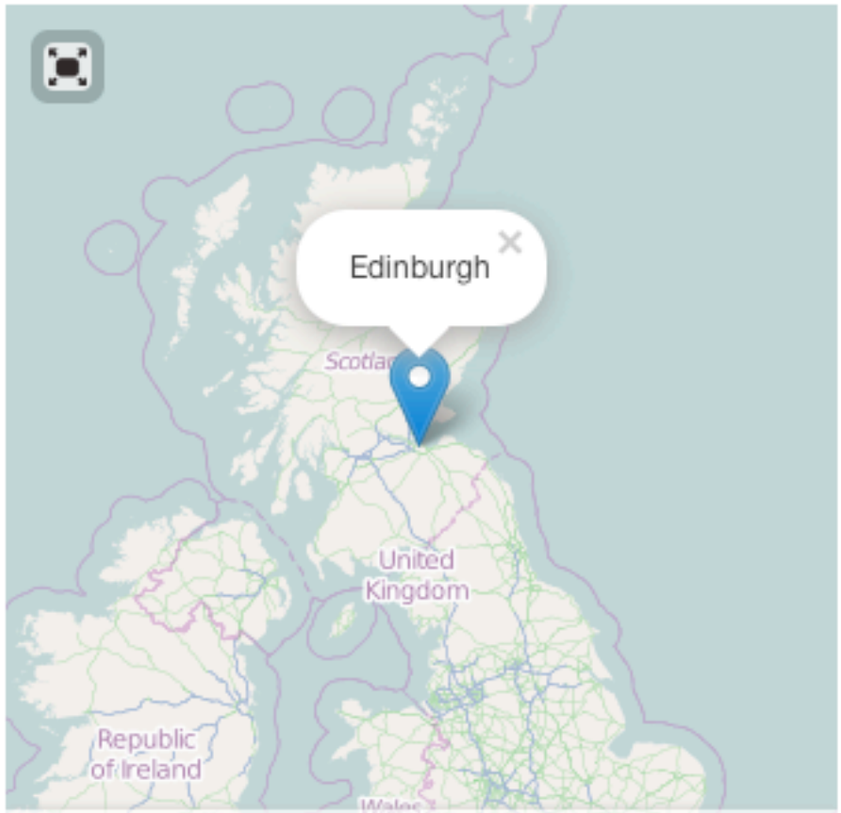


PQIS All Team Meeting, ProQuest, April 23rd 2014

Search by: Commodity | Location

Edinburgh

In Country	GB
Feature Type	Capital Of Top-Level Administrative Division
Population	435,791
GeoNames Entry	View entry



Filter

BY COLLECTION

All

House of Commons Parliamentary Papers (116)

BY DECADE

All

1850s (116)

BY COMMODITY

All

Gold (116)

Documents in which 'Edinburgh' is mentioned in relation to commodities (Page 1 of 1)

Filtered by: Decade (1850) | Collection (House of Commons Parliamentary Papers) | Commodity (Gold)

# Mentions	Document Title
90	Report from the Select Committee on the Bank Acts; together with the proceedings of the committee, minutes of evidence, appendix and index.
4	Twenty-ninth report of the Commissioners of Her Majesty's Woods, Forests and Land Revenues: in obedience to the acts of 10 George IV. (cap. 50), and 2 William IV. (cap. 1).
4	Parliamentary Papers. List of the bills, reports, estimates, and accounts and papers, printed by order of the House of

SUMMARY

- Scholars potentially have access to enormous amounts of data but cannot always easily manage and navigate it.
- Text mining can be applied to process large text collections, enrich existing text with information or pull out trends which can be visualised. It is a way to enable distant reading, even if such technology is not 100% accurate.
- OCR errors in digitised collections can skew results.
- Interdisciplinary setup of Trading Consequences made it more successful for everyone involved. It wouldn't have been possible without the original data.

WHAT CAN PQ DO?

- Sharing OCRed full text data with mining research initiatives similar to Trading Consequences.
- Improve process for arranging legal agreements for sharing this data.
- Enable a feedback mechanism to improve the OCR and ultimately improve search results.

PALIMPSEST: LITERARY EDINBURGH

- Current AHRC big data project: Exploring place in literature by mining and visualising literature set in Edinburgh, (University of Edinburgh, EDINA, University of St. Andrews).

- Aiming to retrieve all out-of-copy-right literature set in Edinburgh.

- Developing a fine-grained gazetteer for Edinburgh to enable geo-referencing on the street and building level.

THANK YOU

- Website: http://tradingconsequences.blogs.edina.ac.uk/
- Demo: http://tcqdev.edina.ac.uk/search/commodity/ , http://tcqdev.edina.ac.uk/vis/tradConVis
- Contact: balex@inf.ed.ac.uk