

Bootstrapping a historical commodities lexicon with SKOS and DBpedia.

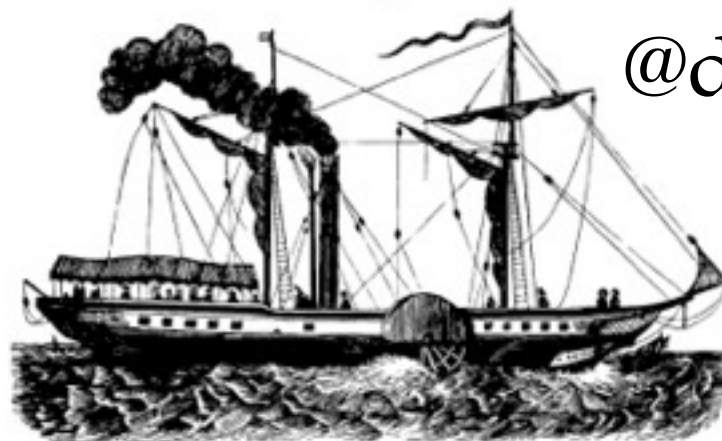
Ewan Klein, Beatrice Alex, Jim Clifford



THE UNIVERSITY of EDINBURGH
informatics



UNIVERSITY OF
SASKATCHEWAN



@digtrade

Trading Consequences

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, Beatrice Alex, Claire Grover, Richard Tobin: *text mining*



Colin Coates, Andrew Watson: *historical analysis*



Jim Clifford: *historical analysis*

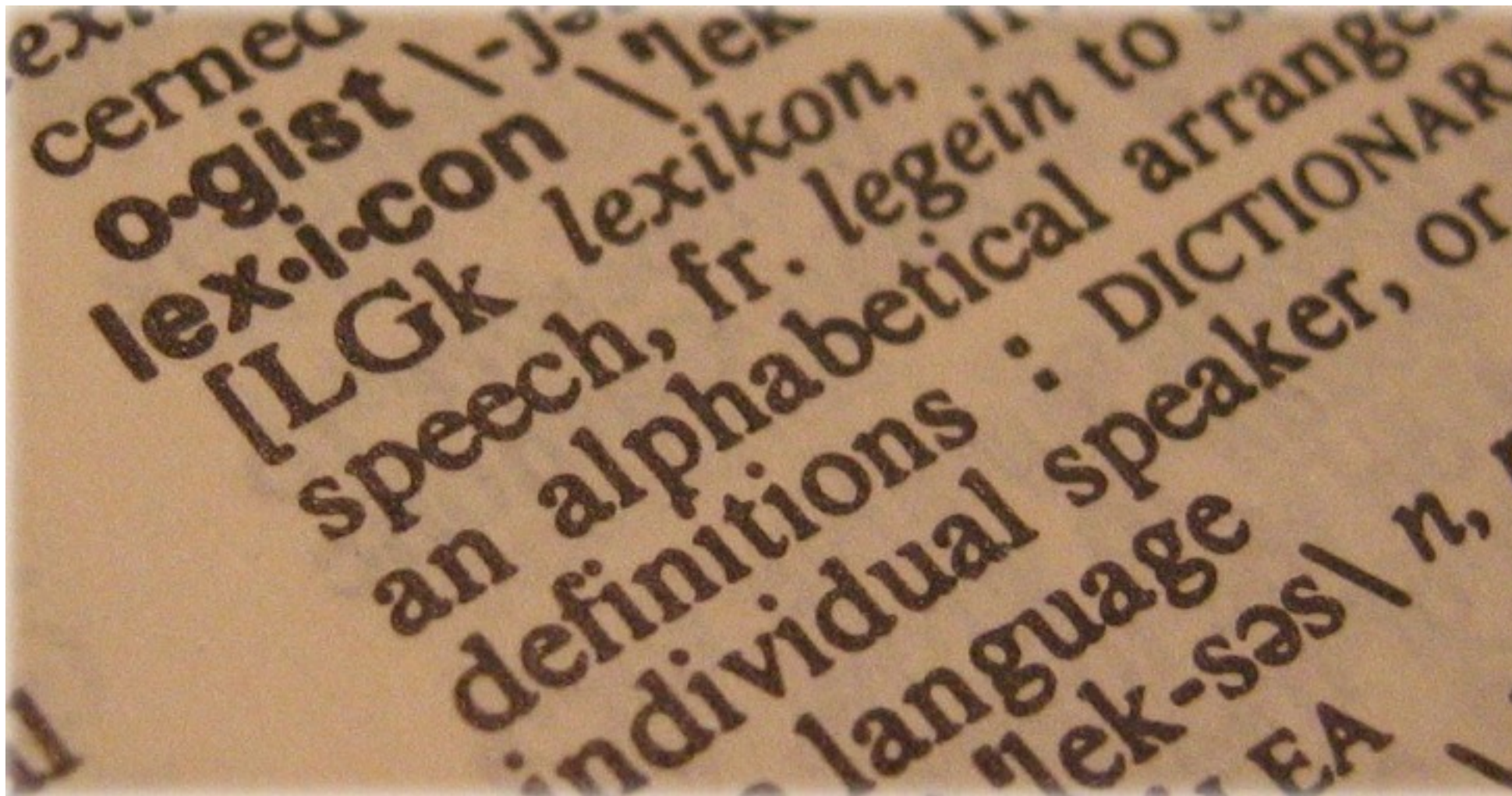


James Reid, Nicola Osborne: *data management, social media*



Aaron Quigley, Uta Hinrichs: *information visualisation*

COMMODITY LEXICON CREATION



SEED SET

Imports.		Imports.	
Animals living <i>etc.</i>		Sheep and Lambs	
1875.		1875.	
COUNTRIES FROM WHICH IMPORTED		COUNTRIES FROM WHICH IMPORTED	
From	QUANTITIES Number	From	QUANTITIES Number
Russia—Northern Ports	1712	Austria	2
" Southern Ports	3	" Bohemia	2
Sweden	55770	" S. T. South Africa—Cape of Good Hope	11
Norway	339479	" Holland	11
Denmark	119763	East Coast of Africa—Portuguese Possessions	11
Germany	662720	" Native States	11
Netherlands	414320	Algeria	11
Belgium	165861	Nubia	11
Channel Islands	9382	Barbary (Tunisia)	11
France	22406	Manilla	11
Portugal	107	Andalus (Spain)	11
" Azores	115	Aden	11
" Madeira	11	Perth	11
Spain	5	India—Bombay and Sindia	11
" Canary Islands	11	" Madras	11
Gibraltar	11	" Bengal and Orissa	11
Italy	11	" Straits Settlements	11
American Territories	11	" Ceylon	11
Mexico and Central	11	" British East Indies	11
Central (including Indian Islands)	11	India—French Possessions	11
Turkey—European	11	" Portuguese Possessions	11
" Asiatic Turkey, Syria, and El Hadjaz	11	" Dutch Possessions—Java	11
" Wallachia and Moldavia	11	" Other Possessions	11
Egypt	11	" Spanish Possessions—Philippine Islands	11
Tripoli and Tunis	11	" Siam	11
Algeria	11	" Szechuan, Chekiang, and Tientsin	11
Morocco	11	China	11
Spanish Ports in Northern Africa	11	" Hong Kong	11
Sardinia	11	" Macao	11
British West Africa	11	Japan	11
The Gold Coast	11		
Fornaco Pt.	11		
West Africa—Portuguese	11		
" Not designated	11		

- Seed set from customs import records.

LaTeX 2014, Gothenburg, April 26th 2014

SEED SET

16	Ammunition - Shot, Large and Small of Iron
17	Ammunition - Rockets and other combustibles for puposes of war and Ammunition unenumerated
18	Animals Living - Asses
19	Animals Living - Goats
20	Animals Living - Kids
21	Animals Living - Oxen and Bulls
22	Animals Living - Cows
23	Animals Living - Calves
24	Animals Living - Horses, Mares, Geldings, Colts and Foals
25	Animals Living - Mules
26	Animals Living - Sheep
27	Animals Living - Lambs
28	Animals Living - Swine and Hogs
29	Animals Living - Pigs (sucking)
30	Animals Living - Unenmumerated
31	Annatto - Roll
32	Annatto - Flag
33	Antimony - Ore of

- Seed set from customs import records.

SEED SET

16	Ammunition - Shot, Large and Small of Iron
17	Ammunition - Rockets and other combustibles for puposes of war and Ammunition unenumerated
18	Animals Living - Asses
19	Animals Living - Goats
20	Animals Living - Kids
21	Animals Living - Oxen and Bulls
22	Animals Living - Cows
23	Animals Living - Calves
24	Animals Living - Horses, Mares, Geldings, Colts and Foals
25	Animals Living - Mules
26	Animals Living - Sheep
27	Animals Living - Lambs
28	Animals Living - Swine and Hogs
29	Animals Living - Pigs (sucking)
30	Animals Living - Unenmumerated
31	Annatto - Roll
32	Annatto - Flag
33	Antimony - Ore of

- Seed set from customs import records.

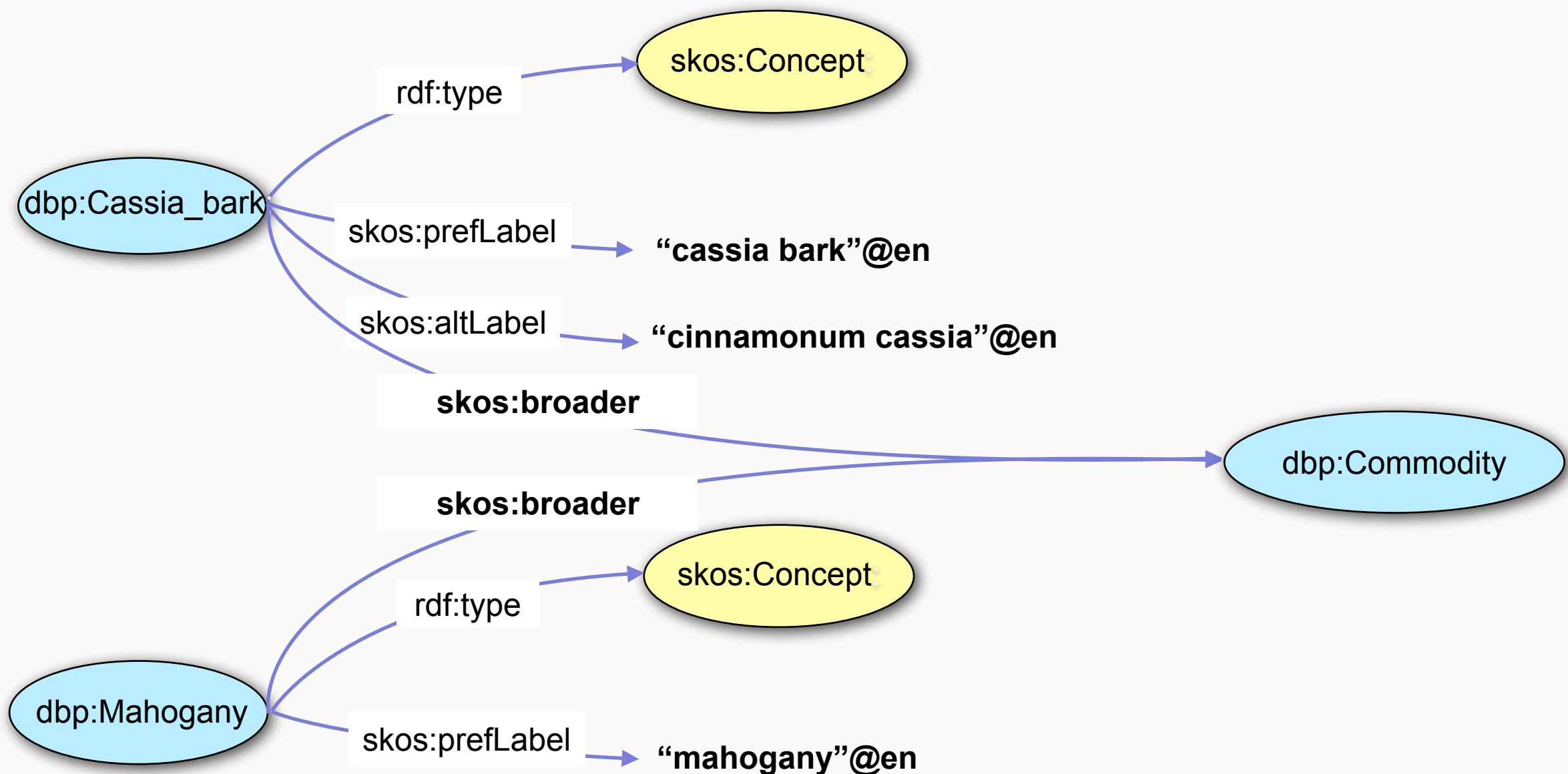
STRUCTURE

- How should synonyms be represented?
 - donkey ~ ass
- How should commodity mentions be grounded?
 - cinnamon -> cinnamonum verum
-> cinnamonum cassia
- How do we group commodities together by type?
 - lemons, limes, oranges -> citrus fruit

SKOS

- Simple Knowledge Organisation System
 - A W3C initiative for the representation of thesauri, classification schemes, taxonomies etc.
 - A standard way to represent knowledge organisation systems using the Resource Description Framework.
 - Looser semantics than strict hierarchies.

EXAMPLE



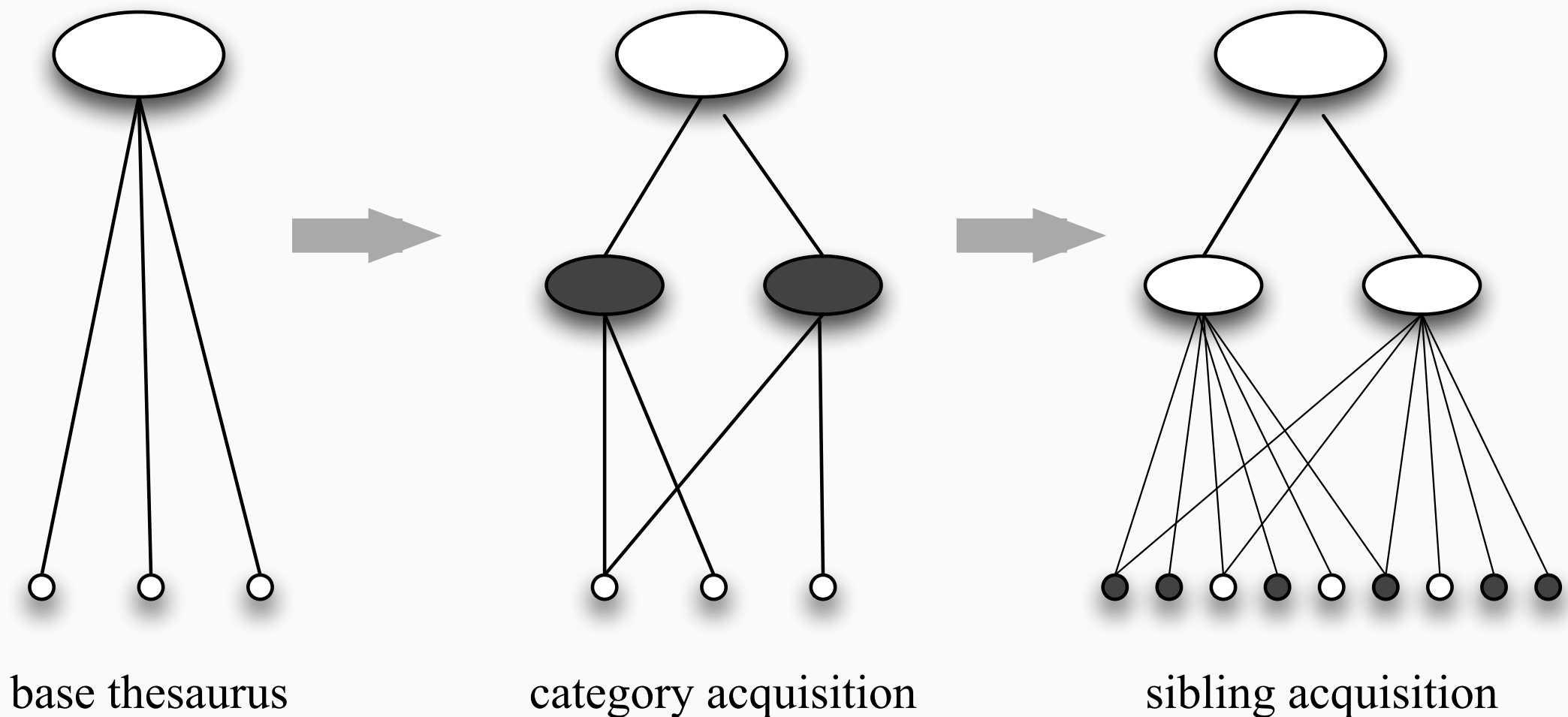
SEED SET IN SKOS

Concept	prefLabel	altLabel
dbp:Cork_(material)	cork	
dbp:Cornmeal	cornmeal	indian corn meal, corn meal
dbp:Cotton	cotton	cotton fiber
dbp:Cotton_seed	cotton seed	
dbp:Cowry	cowry	cowrie
dbp:Coypu	coypu	nutria, river rat
dbp:Cranberry	cranberry	
dbp:Croton_cascarilla	croton cascarilla	cascarilla
dbp:Croton_oil	croton oil	
dbp:Cubeb	cubeb	cubib, Java pepper
dbp:Culm	culm	
dbp:Dammar_gum	dammar gum	gum dammar
dbp:Deer	deer	
dbp:Dipsacus	dipsacus	teasel
dbp:Domestic_sheep	domestic sheep	
dbp:Donkey	donkey	ass
dbp:Dracaena_cinnabari	dracaena cinnabari	sanguis draconis, gum dragon's blood

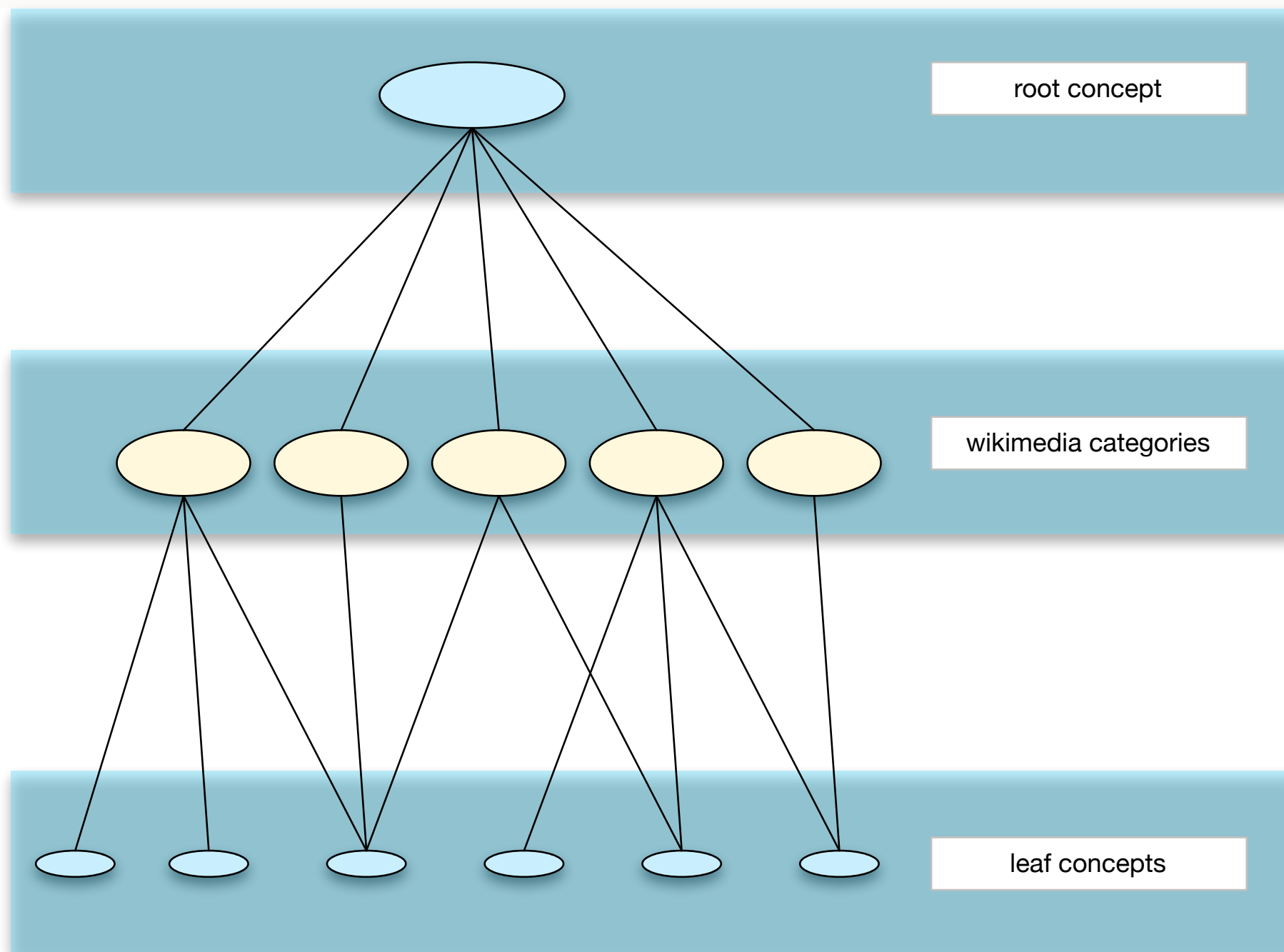
EXAMPLE

	Sustainable forestry · Timber recycling · Wildfire · Wilding	
Industries	Forest (coppicing · farming · gardening · logging) · Manufacturing (lumber · plywood · pulp and paper · sawmilling) · Products (biochar · biomass · charcoal · non-timber · palm oil · rayon · rubber · tanbark) · Transport · Tree farms (Christmas) · Wood (engineered · fuel · mahogany · teak) · Woodworking	
Occupations	Forester · Arborist · Bucker · Choker setter · Ecologist · Firefighter (handcrew · hotshot · lookout · smokejumper) · Log driver · Log scaler · Lumberjack (urban) · Ranger · Resin tapper · Rubber tapper · Shingle weaver · Timber cruiser · Tree planter · Wood process engineer	
Laws and organizations	FAO (FIC · <i>Unasylva</i>) · Forest governance (Forest Principles · Montreal Process · UN Forum) · Forest law (enforcement · forest types) · ITTO · IUFRO · Ministries · Museums · Research institutes · Societies · Technical schools · Universities and colleges (historic)	
Events and initiatives	Arbor Day · Big Tree Plant (UK) · Billion Tree Campaign · Great Green Wall (Africa) · International Day of Forests · International Year of Forests · Million Tree Initiative · Three-North Shelter Forest Program (China) · World Forestry Congress	
 Forestry portal ·  Category (by continent · by country · lists) ·  Commons ·  WikiProject		
Categories: Antiques Furniture Wood Meliaceae History of forestry Forestry in Central America Forestry in Asia Plant common names		

SIBLING ACQUISITION



HIERARCHY



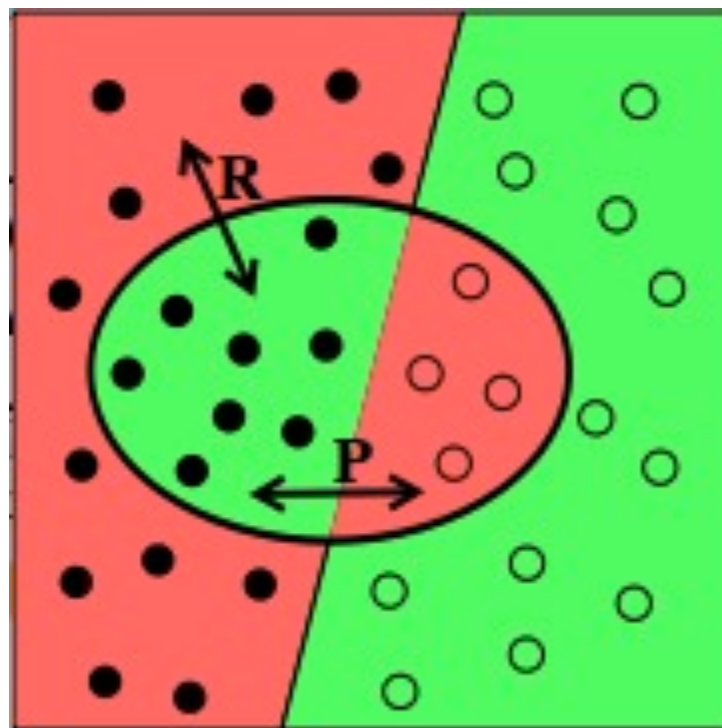
LEXICON IN XML

```
<lex>
  ...
  <lex category="Rubber|Nonwoven_fabrics" concept="Natural_rubber" word="caoutchouc"/>
  <lex category="Rubber|Nonwoven_fabrics" concept="Natural_rubber" word="indian rubber"/>
  <lex category="Rubber|Nonwoven_fabrics" concept="Natural_rubber" word="rubber"/>
  ...
</lex>
```


LEXICON BOOTSTRAPPING

Seed lexicon	319 concepts
Extended lexicon	16,928 concepts
With pluralisation of single word entries	20,476 entries

EVALUATION



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

DOCUMENT COLLECTIONS

Collection	# of Documents	# of Images
House of Commons Parliamentary Papers (1701-1801)		
Early Case Reports		
Direct Correspondence		
Confidential Manuscripts		
For Common Collection		

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

INTERMEDIATE RESULTS

- Lexicon with 20,476 entries and 16,928 concepts.
- Need to evaluate lexicon precision and recall.
- Commodity recognition using rule-based (context and linguistically sensitive) matching.
- Frequency distribution of all commodities detected in our data (31,169,104 in 7 billion words).
- Found 5,841 different commodities (belonging to 4,466 concepts) in the data: 28.5% (26.4%) of commodities in the lexicon.

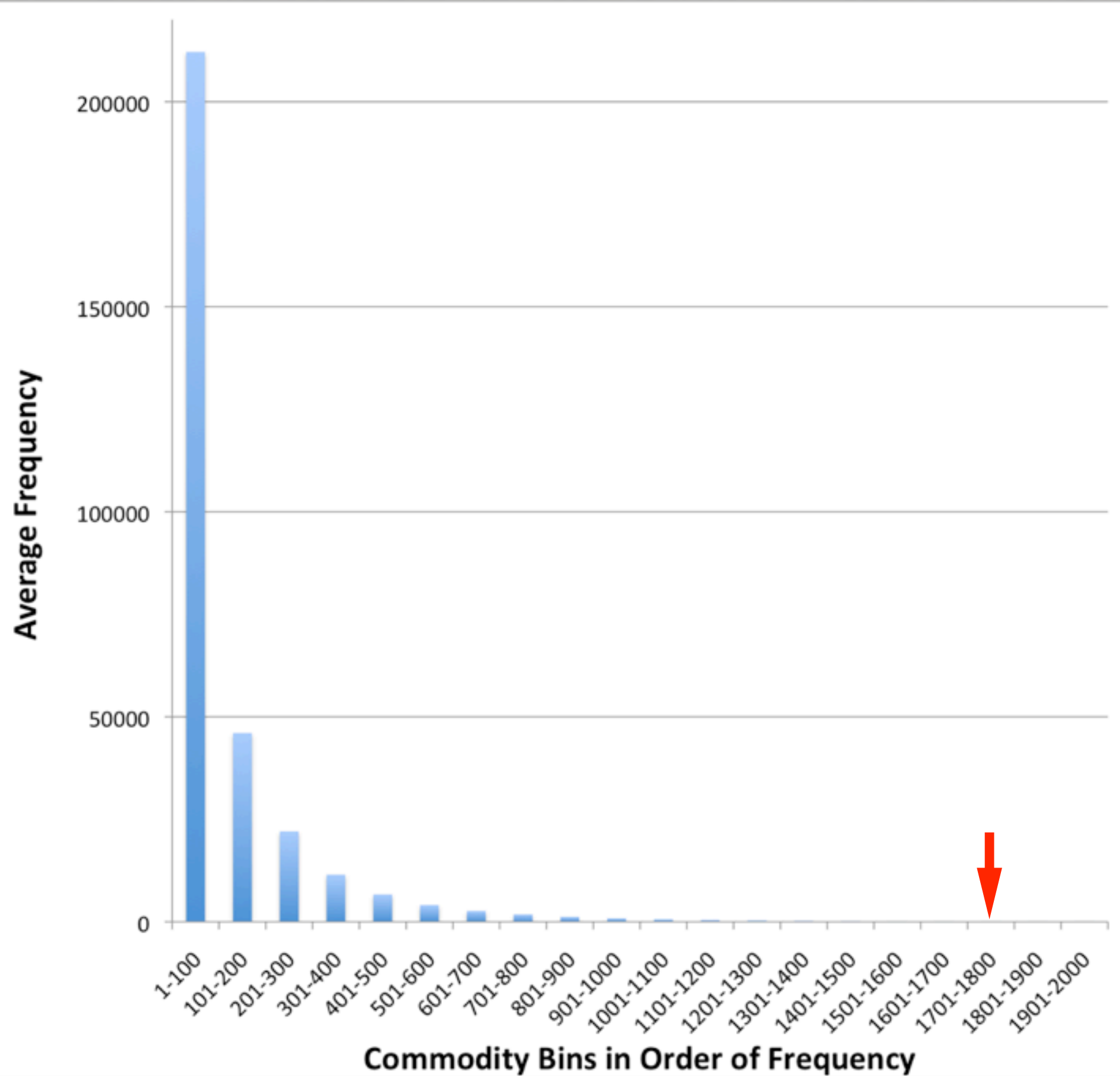
EVALUATION

- How well does our commodity recognition perform on a random test set?
- Indirect evaluation using annotated gold standard:
 - Let human annotator mark up commodities in 120 documents manually.
 - Compared that against the text mining output.

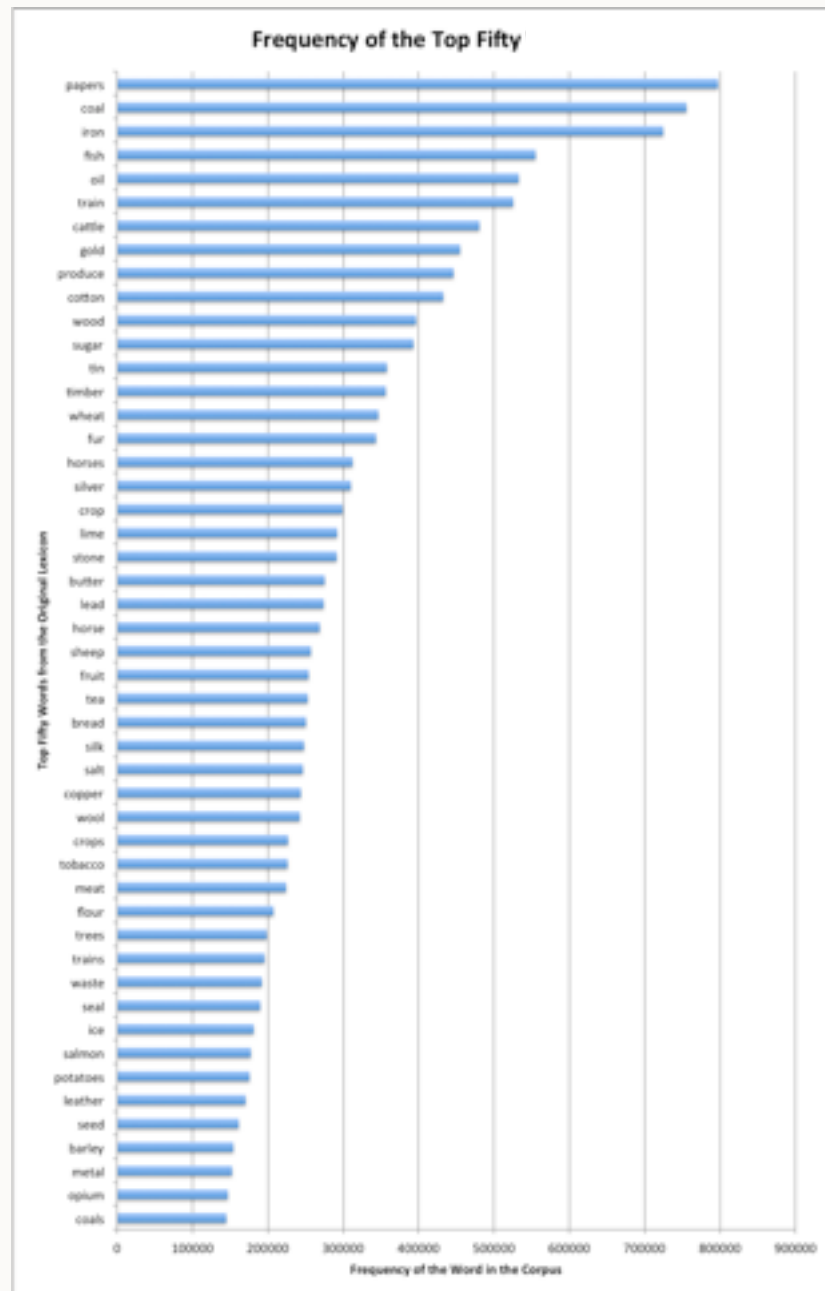
PROTOTYPE EVALUATION

	Evaluation	TP	FP	FN	P	R	F-score
Text Mining Prototype	Strict	616	431	491	0.59	0.56	0.57
	Lax boundaries	791	256	316	0.76	0.71	0.73
IAA	Strict	283	112	109	0.72	0.72	0.72
	Lax boundaries	314	81	80	0.78	0.80	0.80

- Error analysis showed errors in the lexicon and boundary errors affect precision.
- Boundary errors, OCR errors and spelling variations affect recall.



LEXICON PRECISION

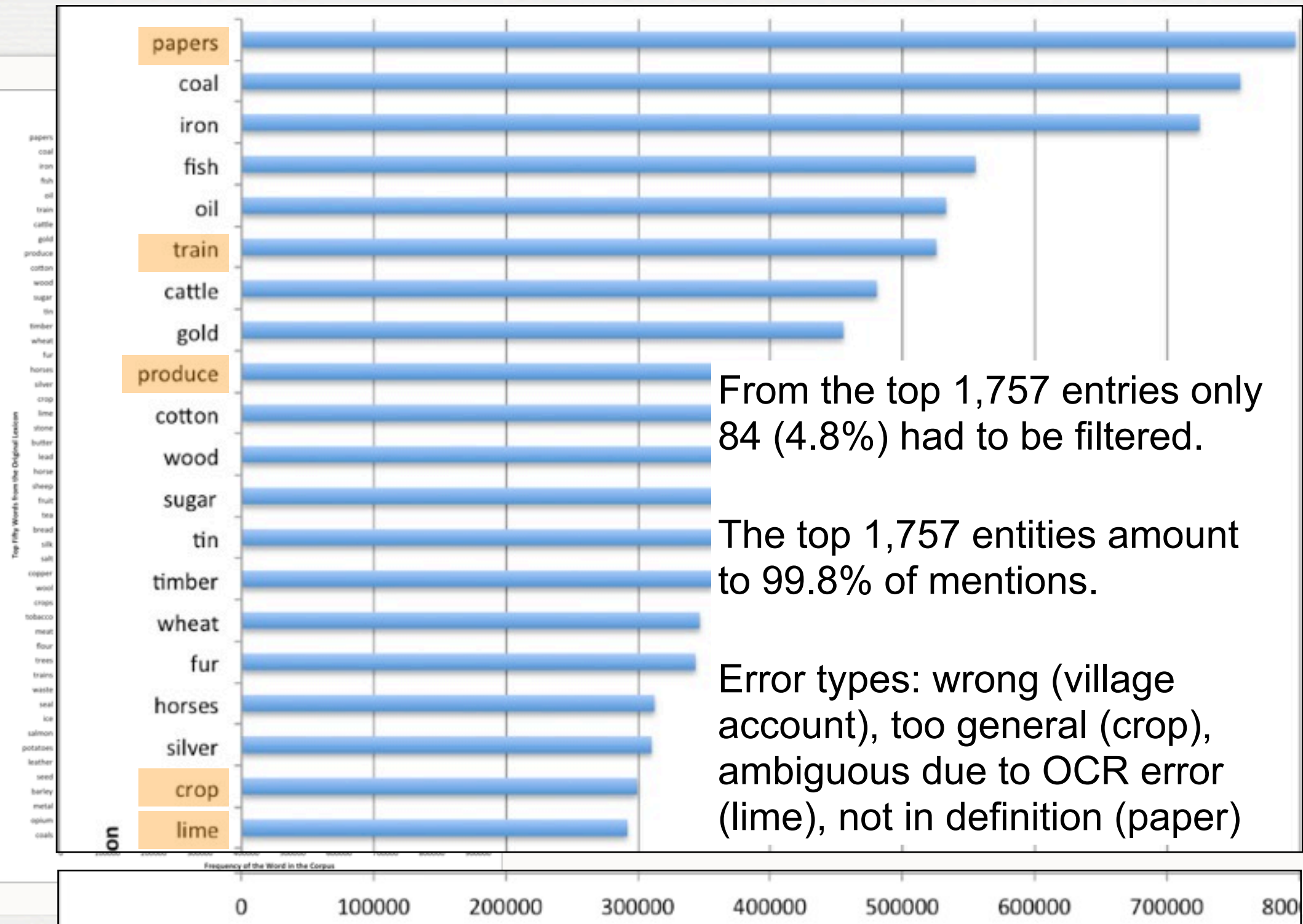


From the top 1,757 entries only 84 (4.8%) had to be filtered.

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

Error types: wrong (village account), too general (crop), ambiguous due to OCR error (lime), not in definition (paper)

LEXICON PRECISION



FALSE NEGATIVES

- Hand annotated texts contain 1,107 commodity mentions (506 different entities).
- 178 entities (683 mentions) are in the first version of the expanded lexicon.
- **329 terms (424 mentions) are not in the lexicon:**
 - 110 (115 mentions) contain OCR errors, approx. 10% of all commodity mentions.
 - 160 commodities are missing, 59 should not be added.

IMPROVING RECALL

tree-s-LeftBigrams		export-s-LeftBigrams	
19386 the		3286 imports	
10546 of		2991 general	
8932 ,		1928 total	
4902 fruit		1282 tho	
4335 a		820 colonial	
3880		618 principal	
2705 and		414 regards	
2234 young		408 *	
2130 destroying		396 grain	
1785 other		387 staple	
1695 -		341 chief	
1500 with		298 j	
1371 forest		253 ..	
1366 timber		228 and	
1304 to		227 kingdom	
1043 mulberry		193 -"	
1031 olive		176 markets	
986 these		171 articles	
985 ;		161 foreign	
972 tho		156 ^	
964 any		148 s.	
921 large		147 \	
916 this		146 wine	
881 palm		146 -f	
819 or		134 i	
780 oak		134 duties	
773 apple		133 trade	
751 rubber			
642 such			
617 orange			

IMPROVEMENTS

- i. Removing terms based on frequency analysis
- ii. Boundary extension rules
- iii. Adding terms based on bigram analysis
- iv. Combination of i-v (with new lexicon: 17,247 concepts and 22,723 entries)

SYSTEM PERFORMANCE

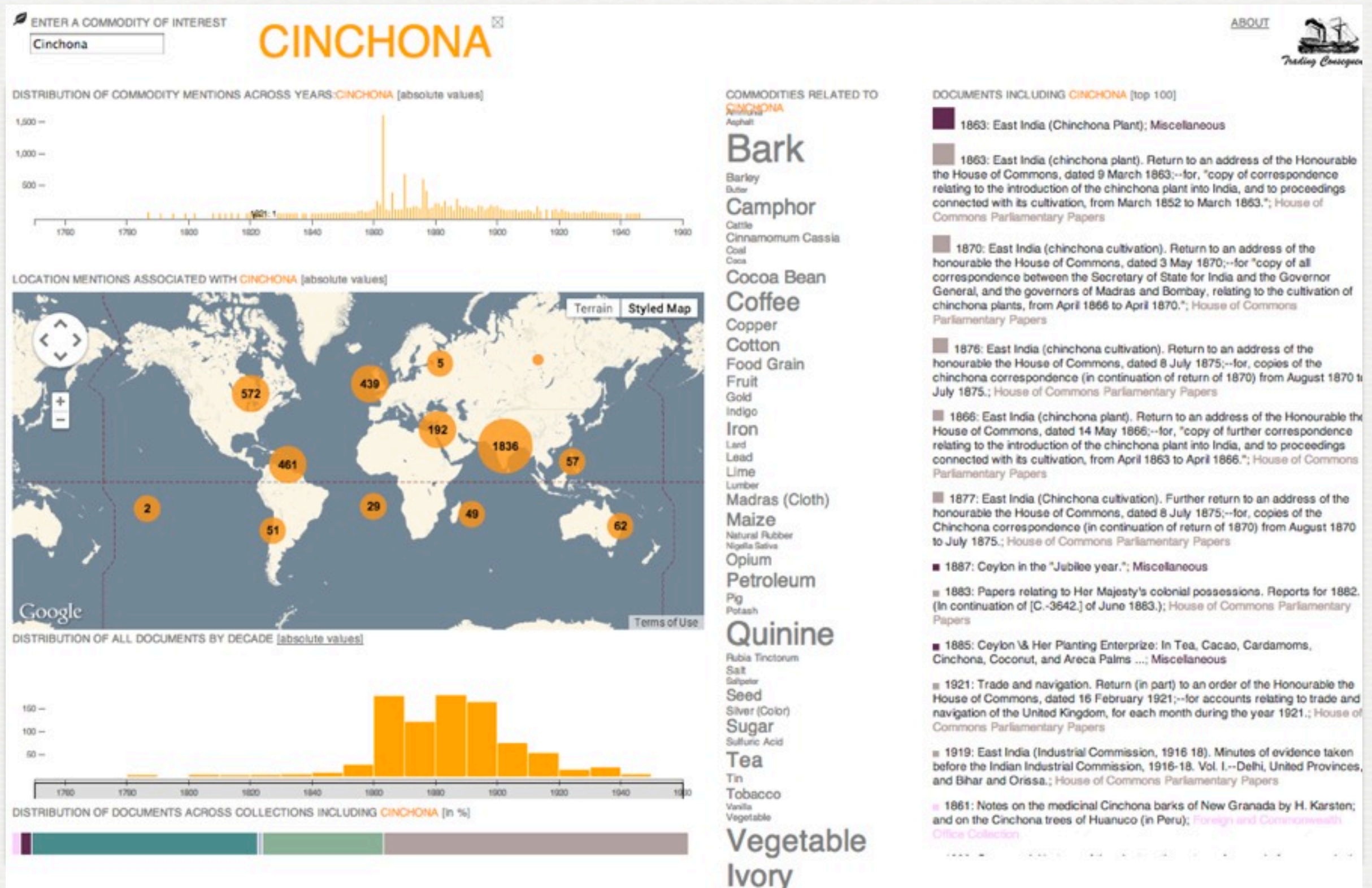
	Evaluation	TP	FP	FN	P	R	F-score
Text Mining Prototype	Strict	616	431	491	0.59	0.56	0.57
	Lax boundaries	791	256	316	0.76	0.71	0.73
IAA	Strict	283	112	109	0.72	0.72	0.72
	Lax boundaries	314	81	80	0.78	0.80	0.80

	Evaluation	TP	FP	FN	P	R	F-score
Text Mining Prototype	Strict	616	431	491	0.59	0.56	0.57
	Lax	791	256	316	0.76	0.71	0.73
(i) Removal of lexicon errors	Strict	603	331	504	0.65	0.54	0.59
	Lax	765	169	342	0.82	0.69	0.75
(ii) Context Rules	Strict	664	483	443	0.58	0.60	0.59
	Lax	777	370	330	0.68	0.70	0.69
(iii) Bigram-based additions	Strict	673	441	434	0.60	0.61	0.61
	Lax	855	259	252	0.77	0.77	0.77
Modified Lexicon: combination of (i)–(iii)	Strict	652	353	455	0.65	0.59	0.62
	Lax	792	213	315	0.79	0.72	0.75

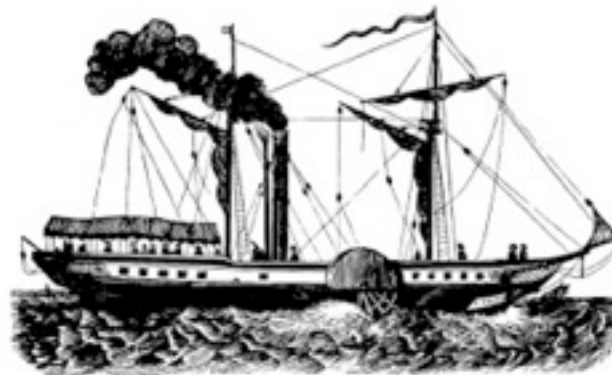
LESSONS LEARNED

- SKOS is useful for organising a lexicon.
- We developed a method for bootstrapping from a seed set using categorial similarity of other entities.
- Expert knowledge and historians' input was important for optimisation.
- Bootstrapping a lexicon and text mining are not error free (but even human experts can disagree).

USER INTERFACE



THANK YOU



Trading Consequences

- 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

TRADITIONAL HISTORICAL RESEARCH



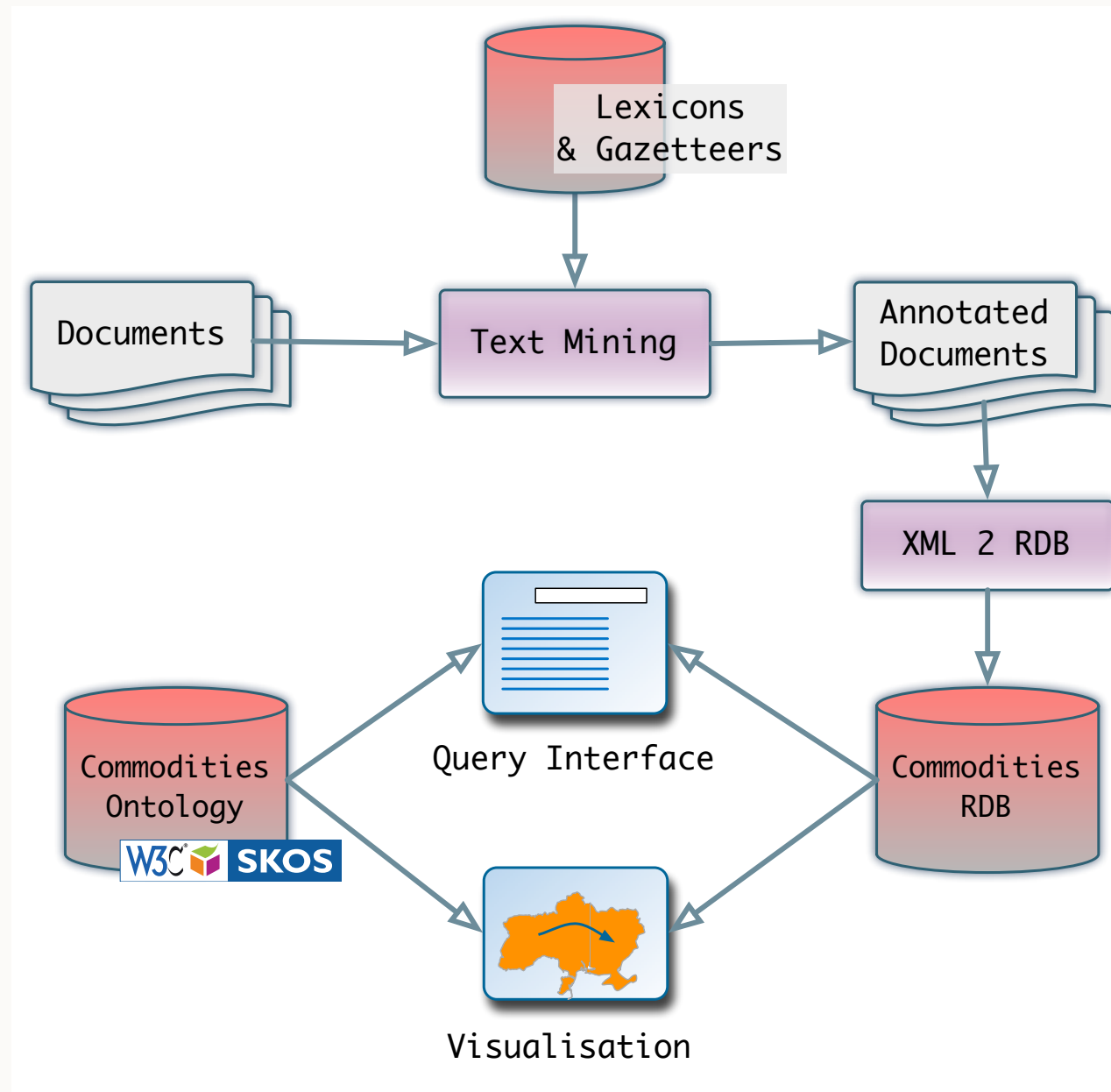
Map showing the areas where mahogany is grown

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

Global Fats Supply 1894-98



SYSTEM



MINED INFORMATION

- Example sentence:

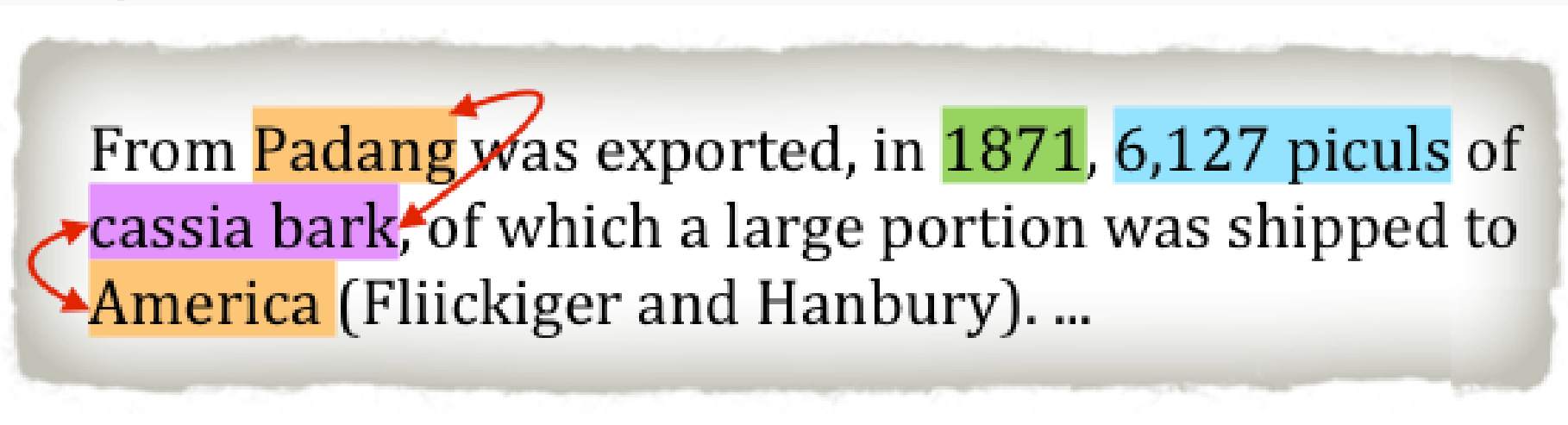
From Padang was exported, in 1871, 6,127 piculs of cassia bark, of which a large portion was shipped to America (Flickiger and Hanbury). ...

- 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

SIBLING ACQUISITION

B

- Bilinga (wood)
- Bird's eye figure
- Bloodwood
- Board foot
- Bog-wood
- Bulnesia sarmientoi
- Burl

C

- Calamander wood
- Cedar wood
- Certified wood
- Chlorocardium rodiei
- Cinnebar
- Cocobolo
- Coconut timber
- Sapele

D

- Dalbergia melanoxylon
- Diamond willow

- Heart pine
- Hickory

I

- Intraspecific antagonism
- Iroko
- Ironwood

J

- Janka hardness test

K

- Kingwood (wood)

L

- Lacewood
- Lignum nephriticum
- Lignum vitae
- Log driving
- Lyptus

M

- Mahogany

T

- Teak
- Teak furniture
- Thyine wood
- Tigerwood
- Timber slide
- Tonewood
- List of Indian timber trees
- Tulipwood
- Tylosis (botany)


V

- Vessel element

W

- White wax wood
- Wood Awards
- Wood flour
- Wood grain
- Wood lagging
- Wood preservation
- Wood processing

EXAMPLE



WIKIPEDIA
The Free Encyclopedia

Navigation


- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia
- Wikimedia Shop

Interaction

- Help
- About Wikipedia
- Community portal
- Recent changes
- Contact page


Toolbox

- What links here

Create account  Log in

Article [Talk](#)

[Read](#) [Edit](#) [View history](#)

Search 


Mahogany


From Wikipedia, the free encyclopedia

This article is about the timber. For other uses, see [Mahogany \(disambiguation\)](#).

Mahogany refers to the straight-grained, reddish-brown [timber](#) of three [tropical hardwood species](#) of the [genus *Swietenia*](#), part of the [chinaberry family](#), *Meliaceae*, indigenous to the [Americas](#).^[1] The three species are:

- **Honduran** or **big-leaf mahogany** (*[Swietenia macrophylla](#)*), with a range from Mexico to southern [Amazonia](#) in [Brazil](#), the most widespread species of mahogany and the only true mahogany species commercially grown today.^[1]
- **West Indian** or **Cuban mahogany** (*[Swietenia mahagoni](#)*), native to [southern Florida](#) and the [Caribbean](#), formerly dominant in the mahogany trade, but not in widespread commercial use since World War II.^[1] Illegal logging of *S. macrophylla*, and its highly destructive environmental effects,^[2] led to the species' placement in 2003 on Appendix II of [Convention on International Trade in Endangered Species](#) (CITES), the first time that a high-volume, high-value tree was listed on Appendix II.^[3]
- *[Swietenia humilis](#)*, a small and often twisted mahogany tree limited to seasonally [dry forests](#) in Pacific Central America that is of limited commercial utility.^[1] Some [botanists](#) believe that *S. humilis* is a mere variant of *S. macrophylla*.^[1]



Honduras mahogany 

CATEGORY ACQUISITION

