Text Mining Diverse Data

David Milward
Linguamatics

II-SDV 2013

Linguamatics

Overview

- Introduction to I2E Agile Text Mining
- Search and extract from diverse data sources
  - How can we search across differently structured data?
    - unstructured
    - semi-structured
    - tabular
  - How do we extend our vocabularies?
    - to different domains
    - to different languages

Whatever the Content...

- ... I2E can mine and extract with precision

Scientific literature

Gene/Protein	Biomarker	Disease		Doc		Hit
TRIM25	as a new potential biomarker for	Breast cancer	1	16144914	1	Estrogen-responsive finger protein as a new potential biomarker for breast cancer.
	is a significant prognostic factor in	Breast cancer	1	16144914	1	CONCLUSIONS: Our data suggest that Efp immunoreactivity is a significant prognostic factor in breast cancer patients.
	was significantly correlated with	Breast cancer	1	16144914	1	Moreover, Efp immunoreactivity was significantly correlated with poor prognosis of breast cancer patients, and multivariate analyses of disease-free survival and overall survival for 151 breast cancer patients showed that Efp immunoreactivity was the independent marker.
BCL2	is a prognostic marker in	Breast cancer	1	16638854	1	Bcl-2 is a prognostic marker in breast cancer independently of the Nottingham Prognostic Index.
	is an independent predictor of	Breast cancer	1	16638854	1	CONCLUSION: Bcl-2 is an independent predictor of breast cancer outcome and seems to be useful as a prognostic adjunct to the NPI, particularly in the first 5 years after diagnosis.

Twitter

Intention	Treatment	User		Doc		Hit
Just got	my flu shot	@4momama	1	217	1	@4momama... Just got my flu shot and in meeting for the ...
I got	my flu shot	@liilii89	2	657	1	@liilii89... I got my flu shot today!
got	my flu shot	@aaaaaaannie	1	658	1	@aaaaaaannie... got my flu shot but now dizzy from the ...
Getting	my flu shot	@amytnguyen	1	737	1	@amytnguyen... Getting my flu shot on
	a flu shot	@oregonian	3	650	1	@oregonian... Getting a flu shot (or two)?
	flu shot	@bobbartz	1	767	1	@bobbartz... Getting flu shot
	the flu shot	@akronbabygirl	1	262	1	@akronbabygirl... Getting the flu shot?
	flu shots	@lauraann9	1	180	1	@lauraann9... Getting flu shots and having blood drawn is ...
	my free flu shot	@brittag10	1	752	1	@brittag10... Getting my free flu shot!

Finding the Right Documents

Terminologies

- Search for synonyms ***and*** children e.g.
  - Cancer, Malignant neoplasm, Malignancy …
  - Leukaemia, Lymphoma, Astrocytoma …

Linguistics

Purified recombinant p42 MAPK | was found to phosphorylate | Wee1 .

Precise Negations

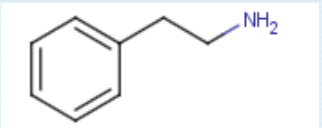
- e.g. “pressure” but not preceded by “blood”

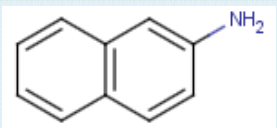
Regular Expressions

- e.g. microRNA: let-?\d+.* mirn?a?-?\d+.*

Chemical Substructure

- e.g.





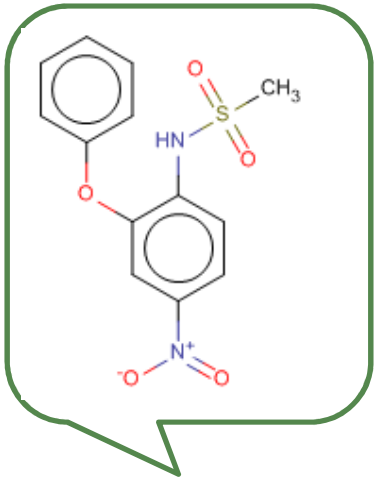
High Throughput

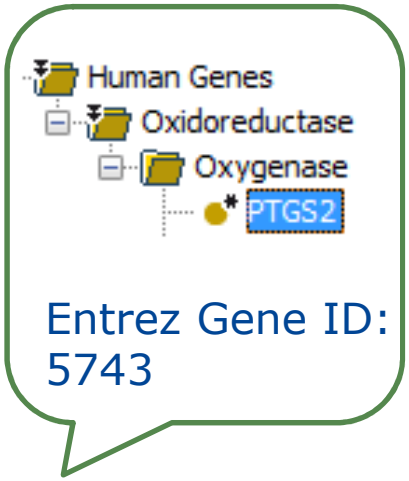
- Simultaneous processing of large numbers of items e.g.
  - 500 genes from microarray experiment

Linguamatics

From surface form to semantics

Identifying entities and relations

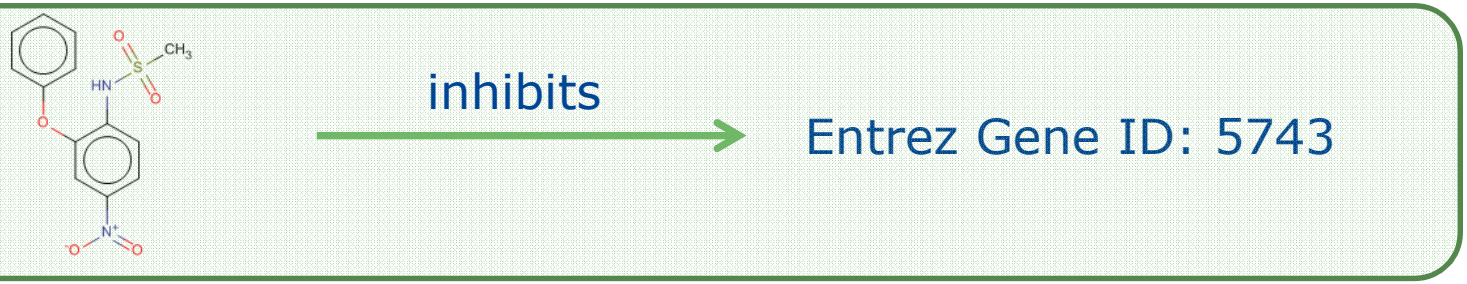




"Among them, nimesulide, a selective COX2 inhibitor, …"

inhibits

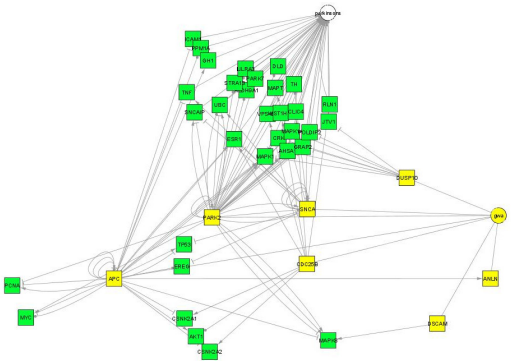
Linguistics to establish relationships



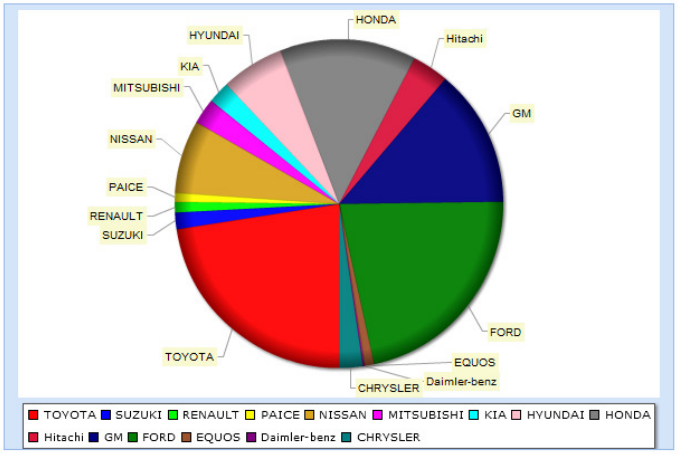
Linguamatics

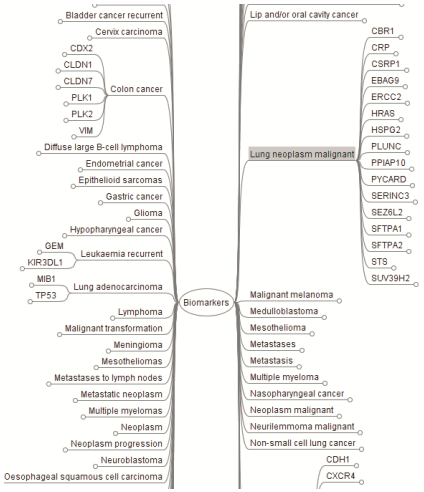
Clustering Results for More Efficient Review

- Focus on documents with the necessary information
- Don’t need to read every document supporting commonly known information



Entrez Genes		Doc
ERBB2	470	15870086
BRCA1	248	15564800
BRCA2	160	15986445
PGR	129	15272277
EGFR	101	16280056
VEGFA	76	15897560
TP53	73	15583825
INS	68	15805581





Linguamatics

Generating New Hypotheses

- Generate new knowledge by linking information from different documents e.g. for drug repurposing

Class1	Relation	Class2	Relation	Class2		Doc		Hit		Doc		Hit
thalidomide	inhibit	VEGFA	prognostic	Carcinoma, Non-Small-Cell Lung	7	21693319	1	Moreover, thalidomide significantly inhibited the production of VEGF and ICAM-1 in serum (P < .05).	9	21964530	1	These data indicate that elevated serum VEGF is a negative prognostic variable in NSCLC.
		NFKB1	sensitize	Carcinoma, Non-Small-Cell Lung	5	16604421	2	The inhibitory effects of thalidomide on the NFkappaB signaling cascade and fibrosis markers including alpha-smooth muscle actin (alpha-SMA) and collagen, were assessed.	3	11016336	2	Inhibition of NF-kappaB sensitizes non-small cell lung cancer cells to chemotherapy-induced apoptosis.

Linguamatics

Exploiting Structure

Structure in Text Mining

- If there is existing structure, want to exploit it e.g.
  - Search within specific fields, or exclude certain fields
    - Claims vs. Description
    - References
    - Tables/Table Cells
  - Restrict matching to avoid noise e.g. don't try to find genes or proteins in the Address field
- May introduce structure using pre-processing e.g.
  - Use OCR to get tabular structure from a PDF
  - Split Claims sections in WIPO patents into the individual Claims
  - Split Criteria in ClinicalTrials.gov into Inclusion and Exclusion criteria

Structured Output Extracted from a Table

- Output a whole table
  - Table description
  - Column/row heading
  - Cell value

TABLE 2. Effects of Short-Term Placebo, Atorvastatin, Irbesartan, and Atortastatin Plus Irbesartan Treatments	Diastolic blood pressure, mm Hg	4 Days of Atorvastatin	82.4±5.5
TABLE 2. Effects of Short-Term Placebo, Atorvastatin, Irbesartan, and Atortastatin Plus Irbesartan Treatments	Diastolic blood pressure, mm Hg	4 Days of Atorvastatin+Irbesartan	82.2±6.4
TABLE 2. Effects of Short-Term Placebo, Atorvastatin, Irbesartan, and Atortastatin Plus Irbesartan Treatments	Diastolic blood pressure, mm Hg	4 Days of Irbesartan	84.3±2.9
TABLE 2. Effects of Short-Term Placebo, Atorvastatin, Irbesartan, and Atortastatin Plus Irbesartan Treatments	Diastolic blood pressure, mm Hg	4 Days of Placebo	84.2±5.1
TABLE 2. Effects of Short-Term Placebo, Atorvastatin, Irbesartan, and esartan	Diastolic blood pressure, mm Hg	Baseline	82.5±4.7

Pharmacologic Substa..	HDL	Cell Content	Units	Units	Table
Atorvastatin	HDL-C	1.1 ±0.6	mmol	l	TABLE 2. Effects of Short-Term Placebo, Atorvastatin, Irbesartan, and Atortastatin Plus Irbesartan Treatments
Irbesartan	HDL-C	1.2±0.5	mmol	l	TABLE 2. Effects of Short-Term Placebo, Atorvastatin, Irbesartan, and Atortastatin Plus Irbesartan Treatments

- Find specific values
  - use ontologies to find the correct drug, however it is expressed
  - find measurement units

Linguamatics

Linking from Definitions to Table Values

Connecting information found in different parts of the document for example finding a compound as “Example 12” in a patent and linking to a table where numerical data is reported

EXAMPLE 12

Patent document

2-(R)-2-(2-Amino-2-methylpropionylamino)-3-(2,4-difluorophenyl)methoxy propionic acid N-[5-(4-chlorophenyl)-3,3-dimethyl-1,1-dioxo-2,3-dihydroisothiazol-4-ylmethyl]-N-ethylamide Hydrochloride (Z=2,4-F_2)

...

Example 5	0.6
Example 7	1.7
Example 12	8.0
Example 15	2.1

Combined into a row of data in the structured results table

Chemical	Inhibition Metric	Value	Units		Doc
2-(R)-2-(2-Amino-2-methylpropionylamino)-3-(2,4-difluorophenyl)methoxy propionic acid N-[5-(4-chlorophenyl)-3,3-dimethyl-1,1-dioxo-2,3-dihydroisothiazol-4-ylmethyl]-N-ethylamide Hydrochloride (Z=2,4-F 2)	EC50	8.0	nM	2	US-7396846-B2

Patent Data from IFI Claims Direct

Linguamatics

Virtual Data Integration

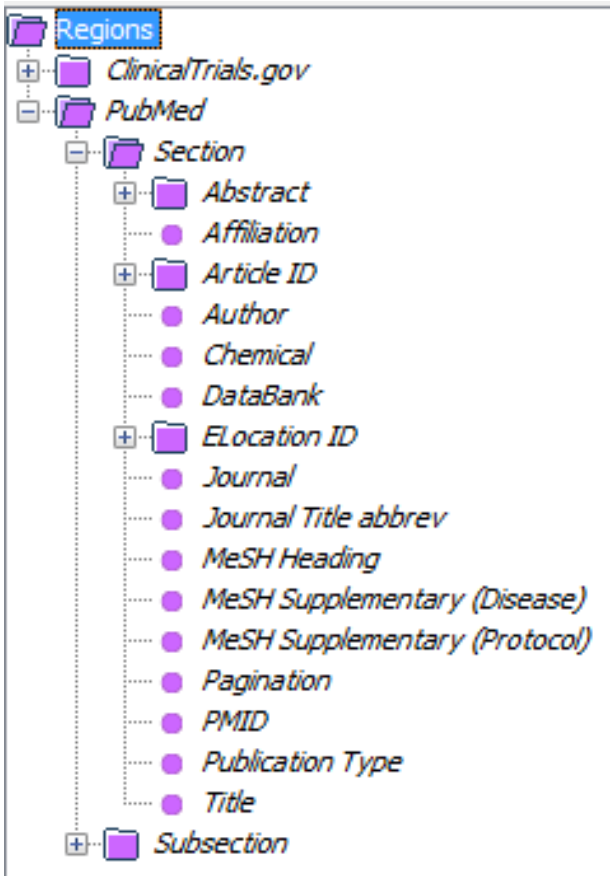
Linguamatics

Motivations for Text Mining Diverse Data

- Convenience
  - One text mining query, different data sets
  - Clustering of information from the different sources
- New knowledge
  - Text mining allows us to discover new knowledge, and make connections where information is from different documents
  - Text mining diverse data means we can make connections, not just between documents of the same kind, but also across different kinds of data

Virtual Data Integration with Structured Data

- Key here is that the precise structure is still preserved
  - It is not a case of only having access to the common fields
  - Can multi pick different fields e.g. search for a word in either the Claims of a patent or the Abstract of a scientific article



Virtual Data Integration: Examples

- ClinicalTrials.gov and MEDLINE abstracts
  - Same search, data returned from all sources

214984	3	Cytologic evaluation of experimental type 2 herpes simplex infection in mice.
NCT00947687	2	- Have not been vaccinated for influenza virus since 2006 (as determined in the medical history) or had a known influenza infection in the current season, defined as in the last 12 months.
14634106	2	Murine infection with the Gram-positive intracellular bacterium Listeria monocytogenes activates CD8(+) T cells that recognize bacterially derived N-formyl methionine peptides in the context of H2-M3 MHC class Ib molecules.

- Patents and published content e.g.
  - Find differences in trends
- Clinical records and clinical trials
  - What clinical trials are available for my patient
- Clinical records and FDA Drug Labels
  - Are prescriptions consistent with drug labelling?

Virtual Data Integration: EHR + Drug Labels

- Identify patients who have conditions that are contraindicated for the drugs they are taking, or patients who are showing known side effects
- One data source is the electronic health records
- One data source is FDA Drug Labels

Continue Naprosyn 500 mg b.i.d., Percocet ... 4. Allergic rhinitis, continue Flonase daily.	56	01e6c0ec-ca2d-4831-bf5a-f48bc76a9ef7	1	NAPROXEN... Anaphylactoid Reactions... This symptom complex typically occurs in asthmatic patients who experience rhinitis with or without nasal polyps, or who exhibit severe, potentially fatal bronchospasm after taking aspirin or other NSAIDs (see CONTRAINDICATIONS #i4i_contraindications_id_16fa73c0-678e-4b91-93f0-f95eb62595dc CONTRAINDICATIONS and PRECAUTIONS: Preexisting Asthma #i4i_precautions_id_94c9606d-6e2c-4cd2-8b62-7f1fee50461b).
The patient states that he has had chest pain and shortness of breath in the past but has not had any currently.... was given a prescription for naproxen 500 b.i.d. p.r.n., dispense ...	50	01e6c0ec-ca2d-4831-bf5a-f48bc76a9ef7	2	NAPROXEN... Information for patients... Although serious CV events can occur without warning symptoms, patients should be alert for the signs and symptoms of chest pain, shortness of breath, weakness, slurring of speech, and should ask for medical advice when observing any indicative sign or symptoms.

CLINICAL NLP ASSOCIATES

Linguamatics

Adapting Terminologies

Linguamatics

Terminology Development

- Agile text mining provides a way of discovering actual language use e.g.
- What terms are synonyms
- What relationships do you get between entities of interest
- What terms are in the same class

Synonym		Doc
CsA	10	21095450
ciclosporin	2	20158284
CSA	2	11669169
Cys	2	17008059
CYA	1	7822939
CyA	1	21169912
CYC	1	20130383
Neoral	1	11260540
Sandimmun	1	2190313
Sandimmune	1	2096278

word		Doc
effective	51	19744410
associated	51	20959578
treat	45	19821411
reduced	21	20419513
used	19	21044451
improves	18	12717587
prevent	16	20234184
reduces	16	20851204
increased	15	20838997
inhibited	14	19825949
induce	13	19734442
inhibit	12	20819433

Terminology Development: Similar Terms

- Distributional features of *apple*
  - being something modified by *red*
  - being something modified by *rotten*
  - being coordinated with *oranges*
  - being coordinated with *bananas*
  - being the object of *eat*
  - being the object of *pick up*
- We can measure how distributionally similar each word or term is to each other in a particular set of documents
- We would expect different kinds of fruits to be similar, but how about vaguer terms e.g. ability

ability:

skill 0.26

capacity 0.24

strength 0.21

talent 0.20

achievement 0.19

competence 0.19

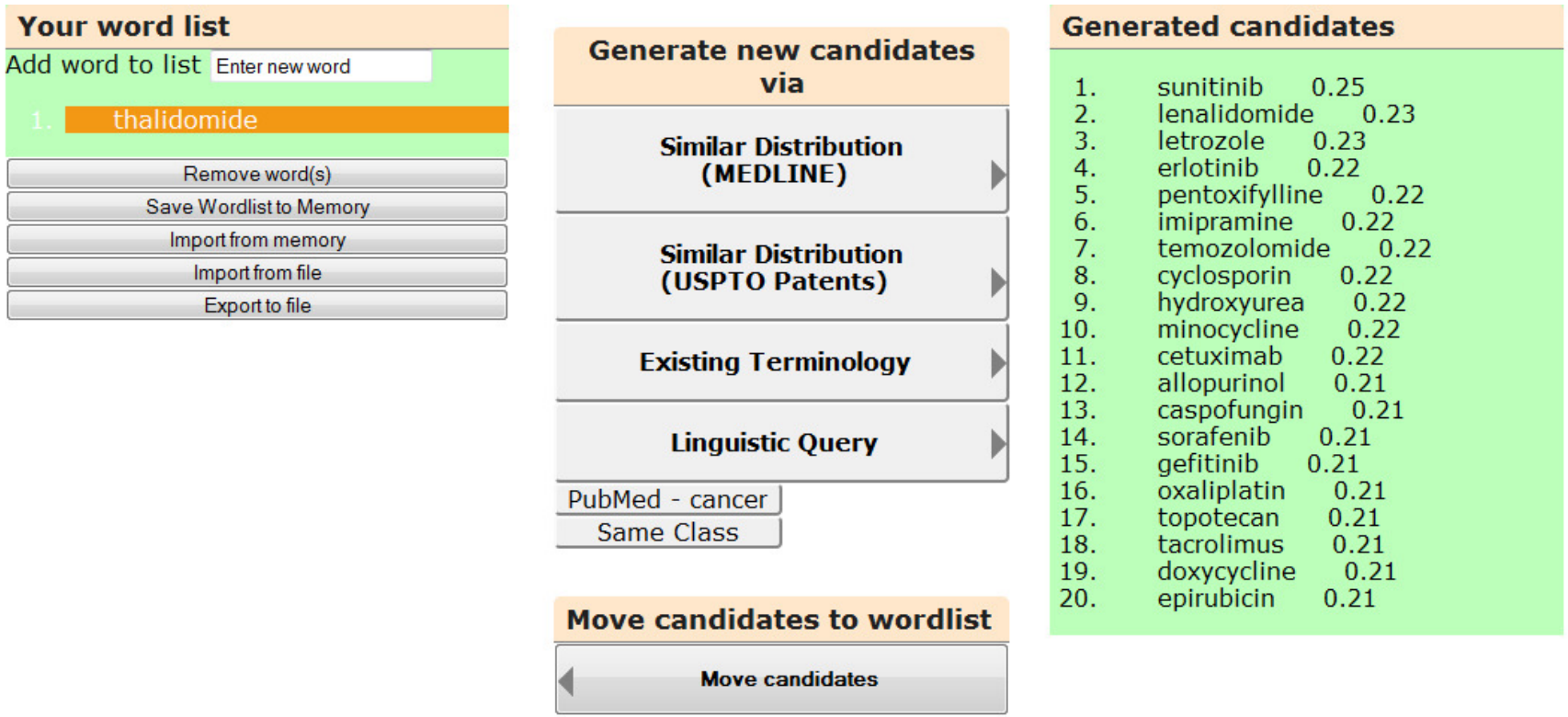
capability 0.19

complexity 0.18

effectiveness 0.18

quality 0.18

Terminology Workbench



Linguamatics

Diverse Data leads to Different Similarities

- Word “fatal” in news vs. scientific abstracts

1.	terrible	0.45
2.	unfortunate	0.36
3.	tragic	0.35
4.	grave	0.28
5.	horrible	0.27
6.	unpleasant	0.27
7.	regrettable	0.23
8.	high-profile	0.22
9.	severe	0.22
10.	minor	0.21
11.	painful	0.2
12.	self-immolation	0.19
13.	sad	0.19
14.	arson	0.18
15.	horrific	0.18
16.	hacking	0.17
17.	untoward	0.16
18.	unexpected	0.16
19.	atrocious	0.14
20.	catastrophic	0.14

1.	ischemic	0.39
2.	ischaemic	0.39
3.	threatening	0.3
4.	fulminant	0.27
5.	hemorrhagic	0.26
6.	non-fatal	0.24
7.	nonfatal	0.24
8.	devastating	0.22
9.	cerebrovascular	0.21
10.	lethal	0.19
11.	thromboembolic	0.18
12.	thrombotic	0.17
13.	haemorrhagic	0.17
14.	embolic	0.17
15.	self-limited	0.15
16.	catastrophic	0.15
17.	accidental	0.15
18.	subacute	0.14
19.	ischemia	0.13
20.	hypoxic	0.13

Multilingual Processing

- English concept name, find occurrences in Chinese text

flourine		Doc		Hit
flourine	1	chinese_example	3	本发明涉及一种特征在于含有氟离子供给化合物、钙离子供给化合物及糖醇且糖醇的1/2以上(重量比)是木糖醇的口腔用组合物。

- Chinese concept name, find occurrences in English or Chinese text

牙齿		Doc		Hit
牙齿	1	chinese_example	8	该组合物在即使存在阴离子表面活性剂的情况下，也能抑制由氟离子和钙离子生成氟化钙沉淀，牙齿的再钙化促进效果好。
				The composition of the even if it exists the anion surface active agent under the condition that can restrain the fluorine ion and calcium ion produce calcium fluoride deposition the teeth and calcium oxide it has good effect of promoting.

Multi-Lingual Terminology Acquisition

m<an>tra

- Most existing terminologies are built for English
- Even when other languages are available the coverage tends to be much lower, often a tenth of the no. synonyms
- If we are to exploit other languages we need better terminologies
- Mantra is an EU FP7 project looking at how to automatically adapt terminologies from one language to another using parallel multi-lingual corpora
- July 2012 – June 2014

Universität Zürich[UZH]

Linguamatics

averbis

Erasmus MC

EMBL-EBI

seit 1558
Friedrich-Schiller-Universität Jena

Conclusions

- Diverse Data provides challenges, but they are solvable
- Once we can search and mine diverse data we can more conveniently
  - perform common searches
  - build connections
  - compare trends
- Can use the data itself to help adapt terminologies to the different data sources, even if these are multi-lingual