



ANALYSING CSTP REPORTS UNDER THE OECD SEMANTIC ANALYSIS FRAMEWORK

CSTP-TIP Workshop: Semantic Analysis for Innovation Policy
12 March 2018

DKI Team:

Frédéric Abrazian
Mary-Ann Grosset
Jan-Anno Schuur
Thierry Vebr

STI Team:

Andrés Barreneche
Alina Deniau
Michael Keenan
Blandine Serve

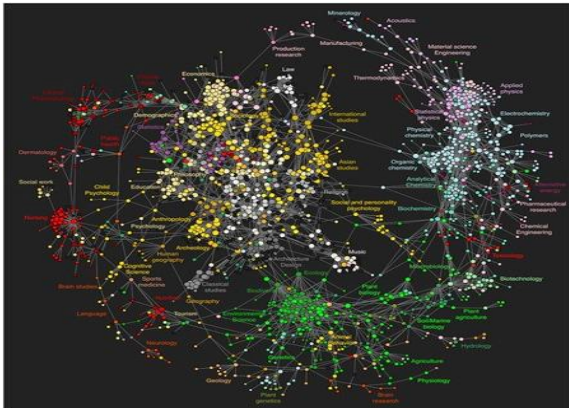


THE OECD SEMANTIC ANALYSIS FRAMEWORK

Breaking the barriers

Taxonomies/Ontologies

Of LANGUAGE and WORDS



Create concepts and link them so that machines can use them as a cloud of knowledge (Semantic layer)

Capture and formalise Subject-Matter-Experts' knowledge

Identify synonyms, related terms in official languages

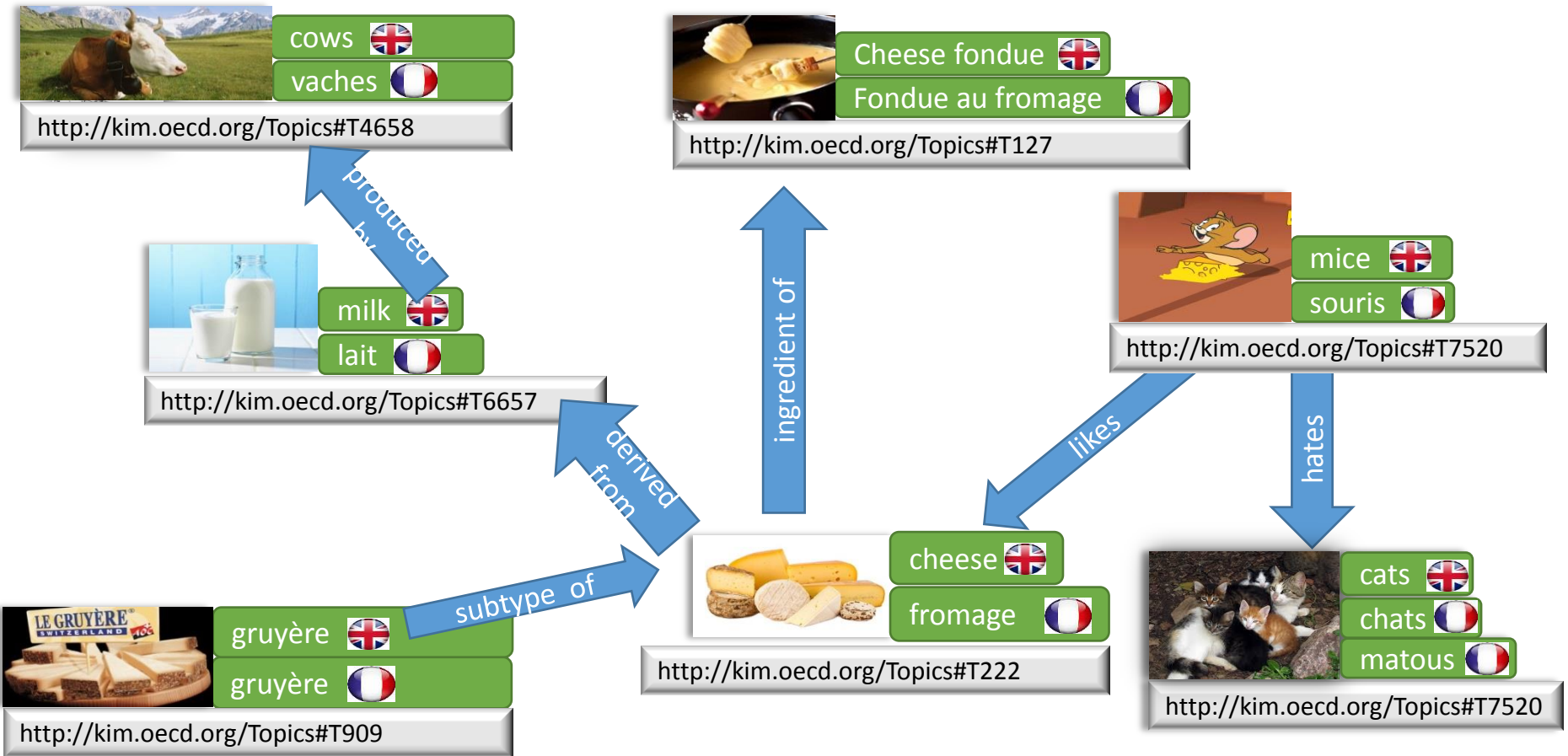
Use URIs not labels

Use in:

Semantic analysis tools

Search and discover tools

Breaking the barriers



How do we create the semantic robots?

- Use lexicon (taxonomies)
- Use text patterns (part of speech)
- Identify relationships (ontologies)
- Test the results on a set of documents
- Debug - Disambiguate
- Test on the complete corpus
- Put in production using Web Services



36 different robots in production



ANALYSING TEXTS

TIP, CSTP AND RESEARCH POLICY



The Making of the IPP Vocabulary

IPP Vocabulary has +1000 unique concepts
relevant within the field of STI policy

MAIN SOURCES USED

- Consultations with STI colleagues
- Oxford Handbook of Innovation
- Semantic analysis of OECD / World Bank flagship publications
 - Ranking of the **10 000** most frequently used terms

Features

- +300 synonyms
- Generic concepts (e.g. firms) are not used for tagging

Limitations

- Flat structure, not taxonomy has been built on top
- Needs updating (e.g. digital economy and inclusive innovation)



Text analysed

Source	Number of items	Text analysed	Time-period
CSTP <u>substantive reports</u>	782	Full text	1993-2017
Subset: TIP <u>substantive reports</u>	100	Full text	1994-2016
Research policy articles	2 527	Titles + abstracts	1988-2016

Output from Luxid system:

- Maximum of 20 IPP topics identified per item.
- For CSTP+TIP documents: A total of 12 390 topics identified.



Treatment of semantic analysis results

- Creation of **word clouds** for topics (50 most recurrent).
- Calculation of **yearly occurrences** of IPP topics identified as keywords.
- **Normalisation** of yearly occurrences by number of items analysed in the respective year (to make occurrences comparable over time).
- **Data aggregation** in time-periods of 8 years: 1993-2000, 2001-2008 and 2009-2017.
- **Ranking** topics in terms of **variance** across the 3 time-periods to identify key trends.