Language Technology Tools for supporting the Multilingual (Semantic) Web

Thierry Declerck, DFKI GmbH, LT-Lab

Max Silberztein, Université de Franche-Comté

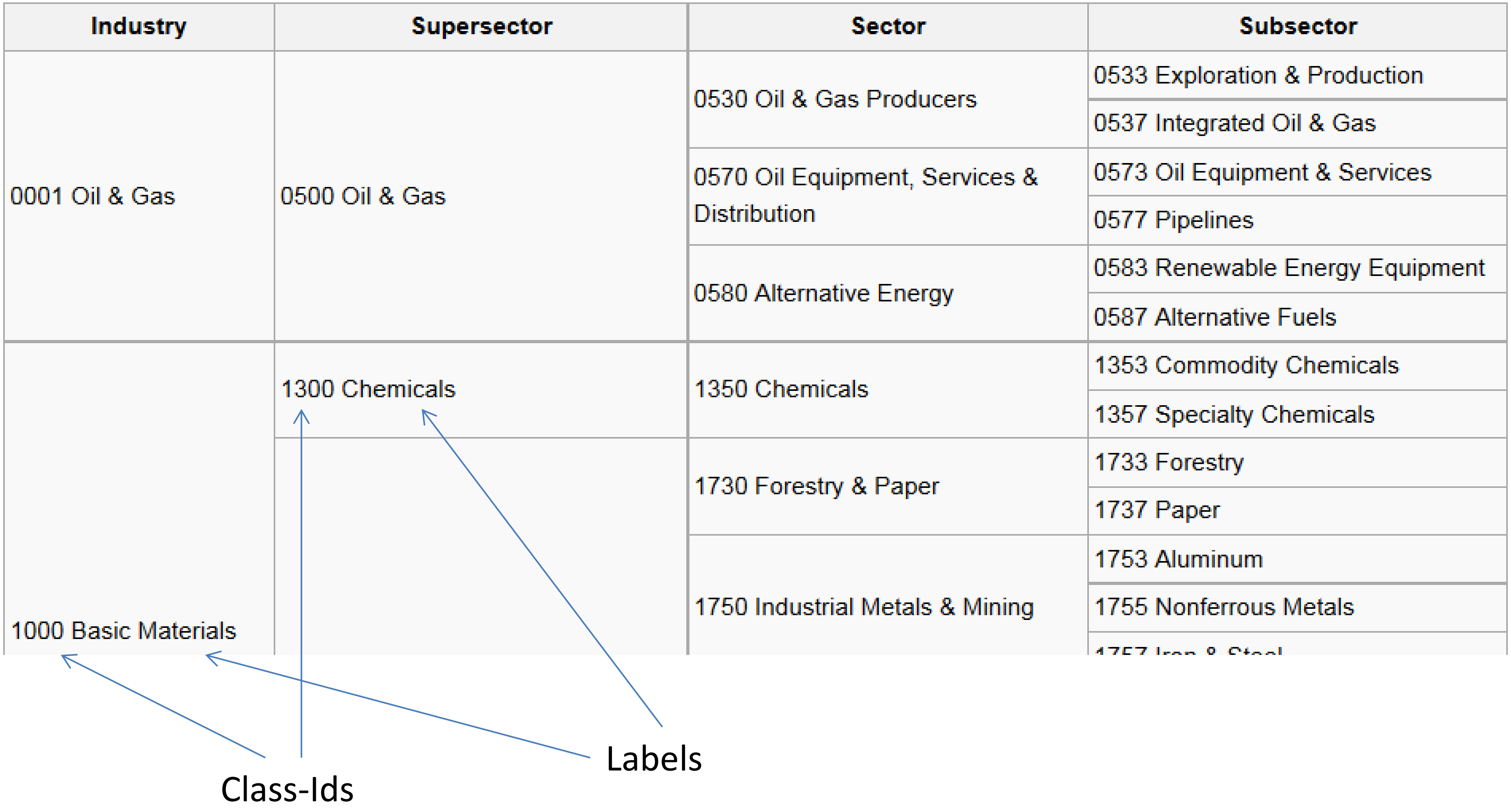
The Web is (partly) Multilingual

- Examples:
  - Multilingual pages
  - Online multilingual dictionaries
  - Online translation tools
  - …
- Differences in term of languages covered
- Not every document available in many languages
- Only few cross-lingual access supported

Multilingual Semantic Resources

- Semantic Resources are also available on the Web, which are including multilingual domain specific terms. Examples:
  - TheSoz (Thesaurus Sozialwissenschaften, 8.000 descriptors in English, French, German – plus other multilingual information
  - GICS (Global Industry Classification Standard, 8 languages) or ICB (Industry Classification Benchmark, 14 languages)
  - Gemet (GEneral Multilingual Environmental Thesaurus, 33 languages)
- Some of those resources have to be mapped first to RDF or SKOS in order to be used in Semantic Web/Linked Data scenarios

Detailed example: GICS



Industry	Supersector	Sector	Subsector
0001 Oil & Gas	0500 Oil & Gas	0530 Oil & Gas Producers	0533 Exploration & Production
			0537 Integrated Oil & Gas
		0570 Oil Equipment, Services & Distribution	0573 Oil Equipment & Services
			0577 Pipelines
		0580 Alternative Energy	0583 Renewable Energy Equipment
			0587 Alternative Fuels
1000 Basic Materials	1300 Chemicals	1350 Chemicals	1353 Commodity Chemicals
			1357 Specialty Chemicals
		1730 Forestry & Paper	1733 Forestry
			1737 Paper
		1750 Industrial Metals & Mining	1753 Aluminum
			1755 Nonferrous Metals
			1757 Iron & Steel

Labels

Class-Ids

Similar: GICS – showing multilingual labels

1010 Energy (Energía / Energie /…)

- 101010 Energy Equipment & Services (Equipos y Servicios de Energía / Energiezubehör und -dienste /…)
  - 10101010 Oil & Gas Drilling (Perforación de Pozos Petrolíferos y Gasíferos / Erdöl- & Erdgasförderung /…)
    - Drilling contractors or owners of drilling rigs that contract their services for drilling wells
    - Contratistas de perforación o propietarios de torres de perforación que contratan sus servicios para perforar pozos.
    - Anbieter von Bohrdiensten oder Eigentümer von Ölförder- und -bohrausrüstungen, die ihre Bohrdienste anbieten

Towards a Multilingual Linguistic Semantic Web

- Work in Monnet project; also at the basis of the Lemon representation of multilingual content of ontologies, see poster by John McCrae at this workshop and www.monnet-project.eu. A starting point of this development: Paul Buitelaar et al., LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies
- Development of the Linguistic Linked Open Data (LLOD, http://nlp2rdf.lod2.eu/OWLG/llod/llod.png
- Need for a combination of NLP tools and Semantic Representation, for semantic annotation of textual (web) documents. 2 Steps:
  - Linguistic analysis of labels of konowledge sources, results of which to be stored as linguistically analysed labels of elements of knowledge sources (using Lemon as representational means)
  - Application of this combined set of linguistic and semantic data to texts, for a semantic annotation.
- Retrieval of multilingual equivalents of detected semantic objects in text not by applying (only) machine translation algorithms, but by displaying the labels in other languages

Test with NooJ

- NooJ is a development environment used to construct large-coverage formalized descriptions of natural languages. See www.nooj4nlp.net/
- NooJ supplies tools to describe inflectional and derivational morphology, terminological and spelling variations, vocabulary (simple words, multi-word units and frozen expressions), semi-frozen phenomena (local grammars), syntax (grammars for phrases and full sentences) and semantics (named entity recognition, transformational analysis).
- NooJ is also used as a corpus processing system: it allows users to process sets of (thousands of) text files. Typical operations include indexing morpho-syntactic patterns, frozen or semi-frozen expressions (e.g. technical expressions), lemmatized concordances and performing various statistical studies of the results.
- New version as open source very soon available as the result of the CESAR project (a satellite project of META-NET): Max Silberztein; Tamás Váradi; Marko Tadic‡ Open source multi-platform NooJ for NLP, Coling 2012

NLP Analysis of Labels

- Oil & Gas Drilling
  - [NP [Noun Conj Noun Noun]]
- Perforación de Pozos Petrolíferos y Gasíferos
  - [NP [Noun Prep Noun Adj Conj Adj]]
- Erdöl- & Erdgasförderung
  - [NP [Noun Conj Noun]]
- *Leading to language specific patterns for term recognitions in text*
  - *but need for prior harmonization (i.e „&“ => „and“, ellipsis resolution, etc)*

Terminological Expansion of Labels

- Goal: Supporting this way higher coverage of Ontology-Based Information Extraction (OBIE). Example: Erdöl- & Erdgasförderung (*Oil & Gas Drilling*),as the prefLabel, generating automatically alternative Labels:
    - Erdölförderung und Erdgasförderung (*Oil Drilling & Gas Drilling*)
    - Erdölförderung / Ölförderung
    - Erdgasförderung / Gasförderung
    - Förderung von Erdöl / Drilling oil wells
    - Fördertung von Erdas / Drilling gas wells
- Domain Specific Class Ids plus prefLabel and altLabel(s) can be encoded in NooJ grammars

Cross-Lingual Terms Expansion

- Apply the ellipsis resolution cross-lingually to all labels in other languages corresponding to a German hyphen compound
  - Perforación de Pozos Petrolíferos y Gasíferos
    - Perforación de Pozos Petrolíferos y Perforación de Pozos Gasíferos
  - Бурение нефтяных и газовых скважин
    - Бурение нефтяных#скважин и Бурение газовых скважин

- Need for a check due to language specific morpho-syntactic properties

Automatic Generation of OBIE grammars

- Work by Declerck and Buitelaar et al in Monnet (example in NooJ)
  - Input: Ontology/Taxonomy Elements together with prefLabels and altLabels (Either in *Lemon* or directly in NooJ Format)
  - Output: A NooJ grammar that can be directly applied to text.
    –

Application of OBIE to Text

- “VUELING es la segunda mayor aerolínea española”

<GICS ID="20302010" LABEL="Líneas_Aéreas">

<ICB Label="Líneas_aéreas" ID="5751" LEV3="5750" LEV2="5700" LEV1="5000">

The system can also display all the corresponding terms in the other available languages

Aknowledgments

- Thanks to the MLW project for the invitation to present our work
- Thanks to Paul Buitelaar and the Monnet project for inspiring discussions
- Thanks to Piroska Lendvai for introducing me to NooJ and for the joint work on multilingual labels, also in the context of Digital Humanities
- Thanks to Dagmar Gromann for her very productive cooperation on the relation between Terminology and Ontologies