Terminology: Towards a systematic integration of semantics and metadata (2)

Maria Recort Ruiz
International Labour Office – *Official Meetings, Documentation and Relations Department*

Denis Dechandon
Publications Office of the European Union – *Directorate A – Information Management*

A common project – overview

▶ **Challenges to be tackled**

- Accessing information, enhancing transparency, ensuring inclusiveness, building on multilingualism

▶ **The basis of our project or when terminology meets the multilingual semantic web**

- Publications Office of the European Union – EuroVoc (VocBench)
- International Labour Office – ILOTERM/Thesaurus/Taxonomy (PoolParty)

▶ **Objectives**

- Improving quality of terminology and semantic assets
- Moving to a systematic use of semantics in the translation field
- Building on linked data

Terminology: Towards a systematic integration of semantics and metadata (2)

▶ op.europa.eu
▶ ilo.org

A common project – the rationale

▟▟ Building on synergies to increase the return on investment

▟▟ Linking terminology collections with controlled vocabularies and ontologies to

- Bring structure to unstructured data (annotation, NLP, ML, AI)
- Enrich metadata assets and terminology collections
- Enhance interoperability, data quality, discoverability and reuse
- Improve semantic search and the exploitation of digital contents

▟▟ Which also has a positive effect on

- Translation work and tools
- Terminology management
- Translation retrievability, reuse and usefulness

Terminology: Towards a systematic integration of semantics and metadata (2)

▶ op.europa.eu
▶ ilo.org

Various assets – categorisation

Terminology collections

- Designed to support the multilingual drafting of documents
- Useful for multilingual drafting, translation and interpretation

Controlled vocabularies

- Lists of standardised terms in a domain established by an authority
- "*On the Semantic Web, vocabularies define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. Vocabularies are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms. In practice, vocabularies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only)."* https://www.w3.org/standards/semanticweb/ontology
- Allow for the categorisation, indexing, and retrieval of information

Terminology: Towards a systematic integration of semantics and metadata (2)

op.europa.eu
ilo.org

Standards

Terminology

- Numerous ISO standards
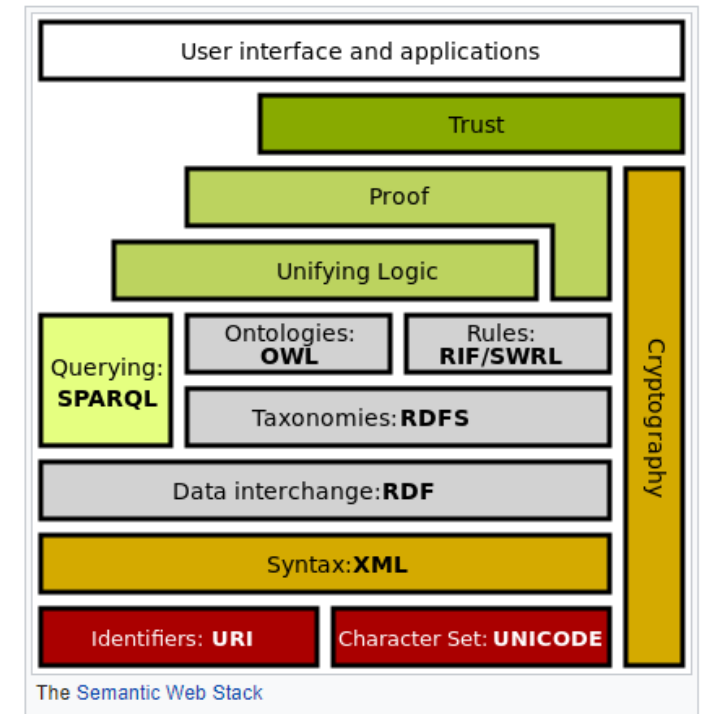- https://termcoord.eu/terminology-iso-standards/

Semantic web

- W3C standards
- ISO 25964-1:2011, Information and documentation — Thesauri and interoperability with other vocabularies —
  Part 1: Thesauri for information retrieval
- ISO 25964-2:2013, Information and documentation — Thesauri and interoperability with other vocabularies —
  Part 2: Interoperability with other vocabularies

The Semantic Web Stack

https://en.wikipedia.org/wiki/Semantic_Web

Terminology: Towards a systematic integration of semantics and metadata (2)

▶ op.europa.eu
▶ ilo.org

Phases & methodology

Basis

- Building on standards and applying them

Phases 1 and 2 (2019-2020)

- Automatic aligner (lexical vs. semantic) and human validation of suggested mappings

Phase 3 (2021)

- Comparison
- Harmonisation of metadata

Phase 4 (2022)

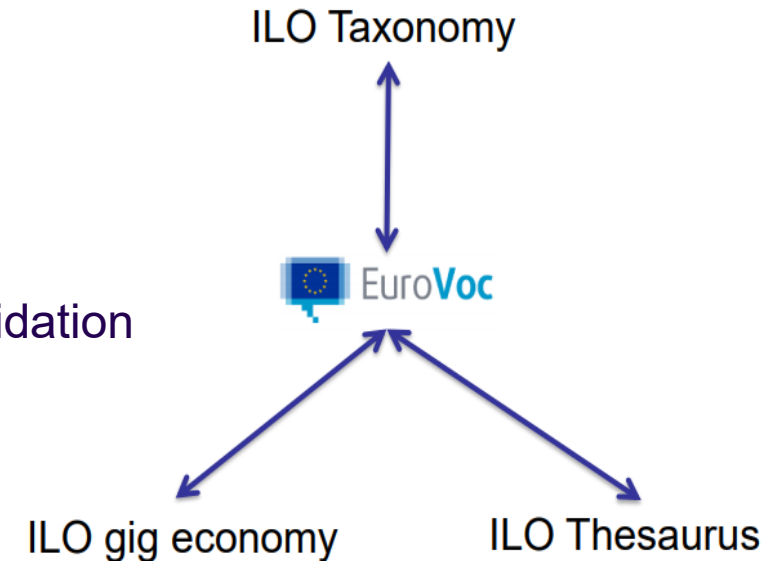- Structuring unstructured data

One step beyond

- Harmonising metadata in tools used throughout the Authoring-Translation-Publishing chain
- Using metadata/semantics systematically
- Creation automation
- Streamlining efforts
- Structuring unstructured data
  - New source files
  - Bulk of existing source files and translations

Metadata   are everywhere in   structured data.
Terminology is   everywhere in unstructured data.

Terminology: Towards a systematic integration of semantics and metadata (2)

▶ op.europa.eu
▶ ilo.org

Phase 1 – Achievements

▰▰ Automatic (lexical) alignment ILO/EuroVoc vocabularies and manual validation

- Definition of vocabulary quality improvements
- 5* LOD, labels, definitions, etc.

▰▰ Alignment of ILO vocabularies

- Definition of vocabulary quality improvements
- Granularity

Terminology: Towards a systematic integration of semantics and metadata (2)

op.europa.eu
ilo.org
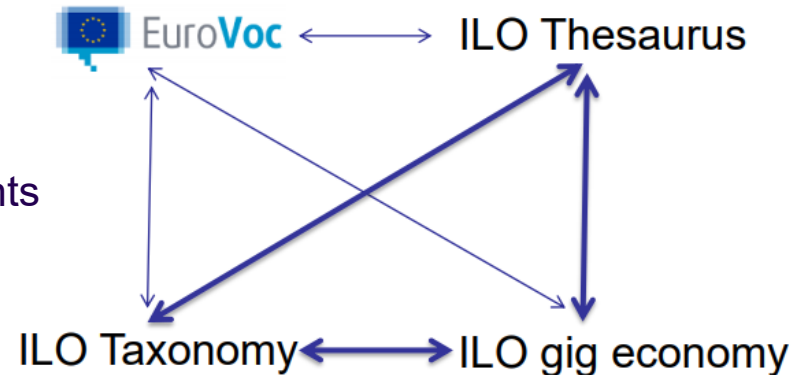
An enhanced interoperability – example

Country name: United Kingdom vs. United Kingdom of Great Britain and Northern Ireland

United Kingdom of Great Britain and Northern Ireland   24-10-45
http://ukun.fco.gov.uk/en/
885 Second Avenue New York, N.Y. 10017
(212) 745-9200
11 June (Celebration of the birthday of Her Majesty the Queen)
UN Statistics on United Kingdom of Great Britain and Northern Ireland

- EuroVoc uses **United Kingdom** as a preferred label, and the full name as a non-preferred concept (UF)
- The ILO Thesaurus, uses United Kingdom of Great Britain and Northern Ireland as a preferred label instead

Terminology: Towards a systematic integration of semantics and metadata (2)

▶ op.europa.eu
▶ ilo.org

An enhanced consistency – example

EuroVoc	ILO Thesaurus
Labour law is in the facet 44 - Employment and working conditions > 4426 Labour law and labour relations	**Labour law is under 04. Law, human rights, government and politics > 04.01 Law**
Labour legislation is a non-preferred label for Labour law, i.e. they refer to the same concept.	**Labour legislation is a preferred label, with a distinct scope note to distinguish it from labour law**

- While in EuroVoc **labour law** and **labour legislation** are synonyms, or at least refer to the same concept, in the ILO Thesaurus each refer to a different concept.

Terminology: Towards a systematic integration of semantics and metadata (2)

▶ op.europa.eu
▶ ilo.org

Phase 2 – Objectives

▸▸ Finalising the improvements of all selected assets – *ongoing*

▸▸ Implementing semantic technologies for all assets – *done (all but for the gig glossary)*

▸▸ Disseminating and implementing validated mappings – *ongoing & planned*

▸▸ Automatic (semantic) alignment (same vocabularies) – *postponed (2021)*

▸▸ Streamlining efforts to ensure – *ongoing*

- Consistency

- Understanding inside and between organisations

▸▸ Enriching vocabulary content – *ongoing*

- Inserting mappings

- Structuring terminology asset(s)

- Disseminating and implementing validated mappings (improving website search features)

COVID-19

Terminology: Towards a systematic integration of semantics and metadata (2)

▸ op.europa.eu
▸ ilo.org

# Phases 2-4 – Reviewed planning, acceleration and an additional milestone

▸▸ Finalising Phase 2

▸▸ From the creation of a document to its publication through its translation

- Metadata at all stages
- Modelling the Authoring-Translation-Publishing chain
- Adding business rules

▸▸ Structuring unstructured data

- Future documents
- Existing bulk of documents

Terminology: Towards a systematic integration of semantics and metadata (2)

▸ op.europa.eu
▸ ilo.org
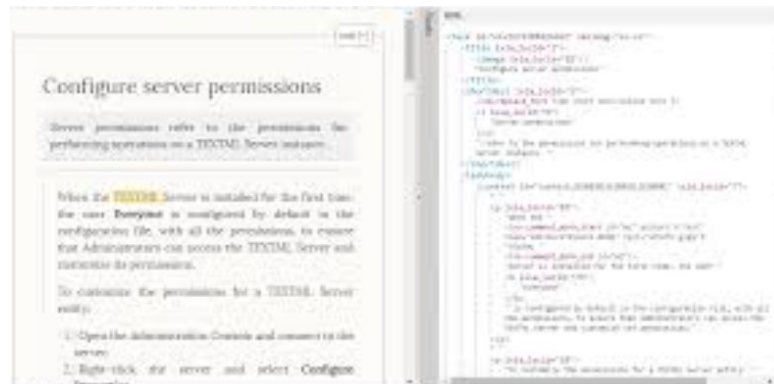
Additional perspectives - Structuring unstructured data

A real example and some more stuff for our workshop *(screenshots to be added)*

op.europa.eu

ilo.org

Towards a systematic integration of semantics and metadata – current achievements, ongoing work and perspectives

▶ Alignment of vocabularies (inter-organisational activity)

▶ Mappings (collaborative work)

▶ Semantic web technologies and tools (integration of metadata)

▶ Vocabulary improvements

▶ Harmonisation and streamlining

▶ Disseminating

▶ Further vocabulary improvements

▶ Disseminating

▶ LOD

▶ Modelling the ATP chain

▶ Structuring unstructured data

▶ New CATT tools

▶ op.europa.eu
▶ ilo.org

Conclusions

- Enhanced interoperability (semantic and technical)

- Ongoing modelling of the ATP chain

- New features for drafters, linguists and translation workflow managers

- Refurbishment of the ATP chain

- Structuring of unstructured data

- Increased reuse of data for different purposes and by various professionals

Terminology: Towards a systematic integration of semantics and metadata (2)

- op.europa.eu
- ilo.org

Contacts

For more information please contact:

recortruiz@ilo.org

denis.dechandon@publications.europa.eu