Towards a Systematic Integration of Semantics and Metadata

Denis Dechandon, Anikó Gerencsér, Maria Recort Ruiz

Publications Office of the EU
Directorate A – Information Management
International Labour Office
Official Meetings, Documentation and Relations Department

Translating and the Computer 41, London 21-22/11/2019

Outline

- Challenges and objectives of knowledge sharing
- Technologies and standards
- Our project
- Intermediary results and next steps
- The bigger picture of the digital age

CHALLENGES AND OBJECTIVES OF KNOWLEDGE SHARING

Knowledge sharing: Common challenges and objectives

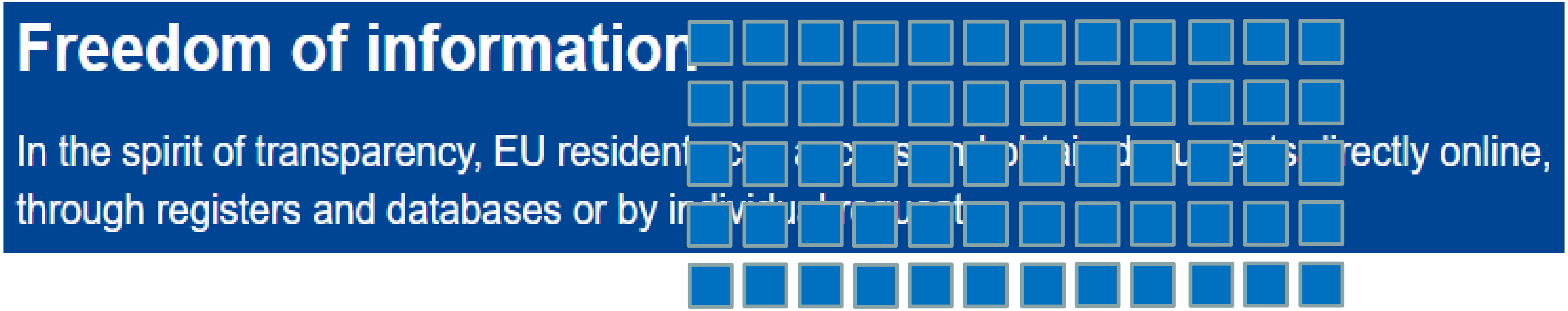
- Getting information
- Transparency
- Inclusiveness
- **Accessing information**
- **Multilingualism**

Knowledge sharing: Common challenges and objectives - Categorisation

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

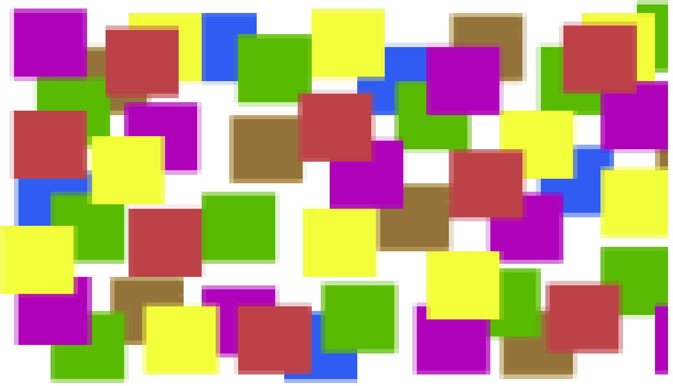
- Controlled vocabularies
  - Lists of standardised **terms** in a **domain** established by an authority
  - Allow for the categorisation, indexing, and retrieval of information

Knowledge sharing: Common challenges and objectives



Freedom of information

In the spirit of transparency, EU resident [illegible] rectly online, through registers and databases or by in [illegible]



We have to make our single market fit for the digital age, we need to make the most of #AI and #bigdata, we have to improve on cybersecurity and we have to work hard for our technological sovereignty. #EUstrivesformore

TECHNOLOGIES AND STANDARDS

Semantic web, technologies and standards

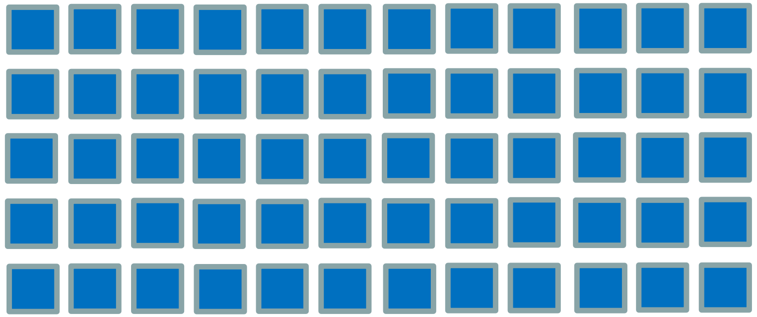
- “I have a dream for the Web [in which computers] become capable of **analyzing all the data on the Web** […] [T]he day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by **machines talking to machines**. The "intelligent agents" people have touted for ages will finally materialize.” *

Resource Description Framework (RDF)

OWL Web Ontology Language

Terse RDF Triple Language

SKOS Simple Knowledge Organization System



- Controlled vocabularies

* Berners-Lee, Tim; Fischetti, Mark (1999). Weaving the Web. HarperSanFrancisco. Chapter 12. ISBN 978-0-06-251587-2.

Semantic web, terminology and corpora – Building bridges

- ISO standards
- InterActive Terminology for Europe (IATE) and EUR-Lex (EU law)
  - Concept domains based on EuroVoc content

    EuroVoc in IATE: disambiguation
  - Benefitting from terminologists' work

    IATE in EuroVoc: lexicalisations and definitions
  - Descriptions of the corpus contents

    EuroVoc in EUR-Lex: a step towards knowledge management
- **Metadata** are everywhere in structured data
- **Terminology** is everywhere in **un**structured data

OUR PROJECT

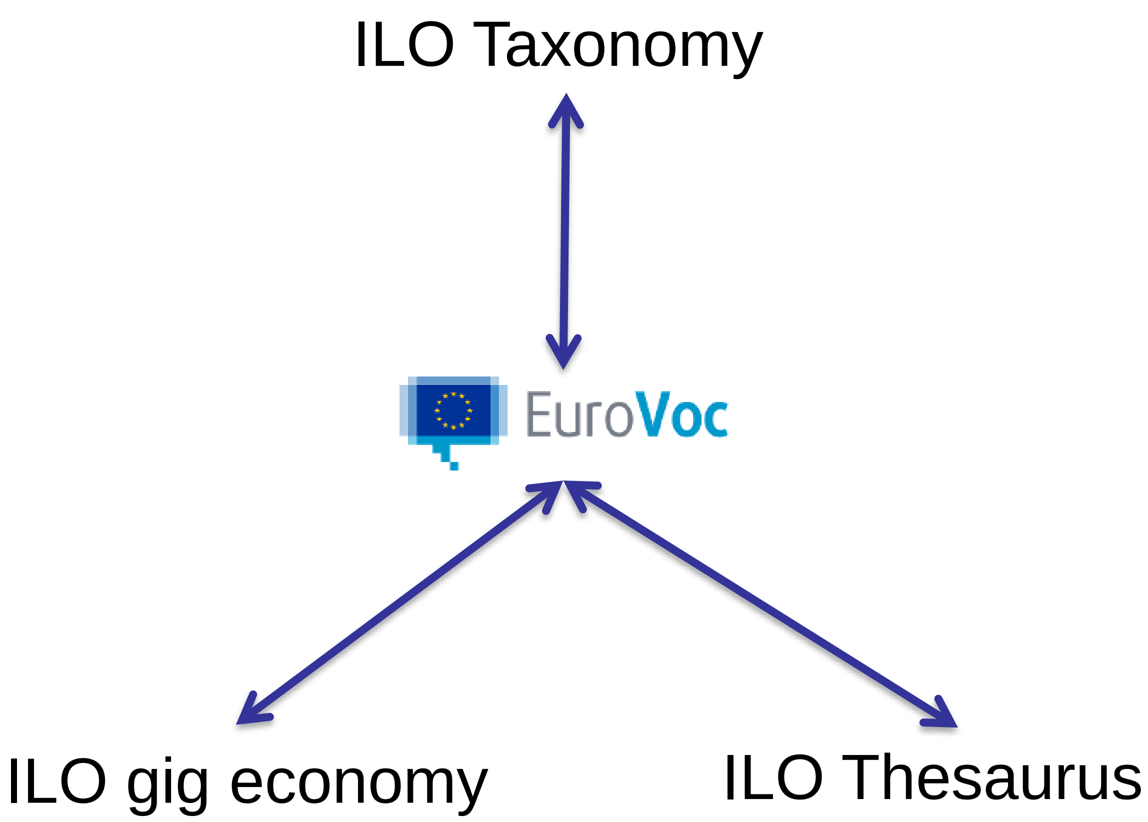
Our project: Phase 1

- Asset/collection identification and file processing
- Use of semantic technologies (VocBench3)
- Content validation
- Automatic lexical alignment of vocabularies
- Checking and validation by vocabulary managers
- Definition of vocabulary quality improvements
    - Definitions and translations
    - Relationships/structure/merging/extensions

Pre-processing and processing (meta)datasets

- Formats
  - From Excel and XML to RDF/XML (SKOS, SKOS XL)
- Content checking
  - Languages, relationships, notes, etc.
- Option definitions
  - Paths to source and target
  - Transformation
  - Comparator
  - Link types

Aligning vocabularies

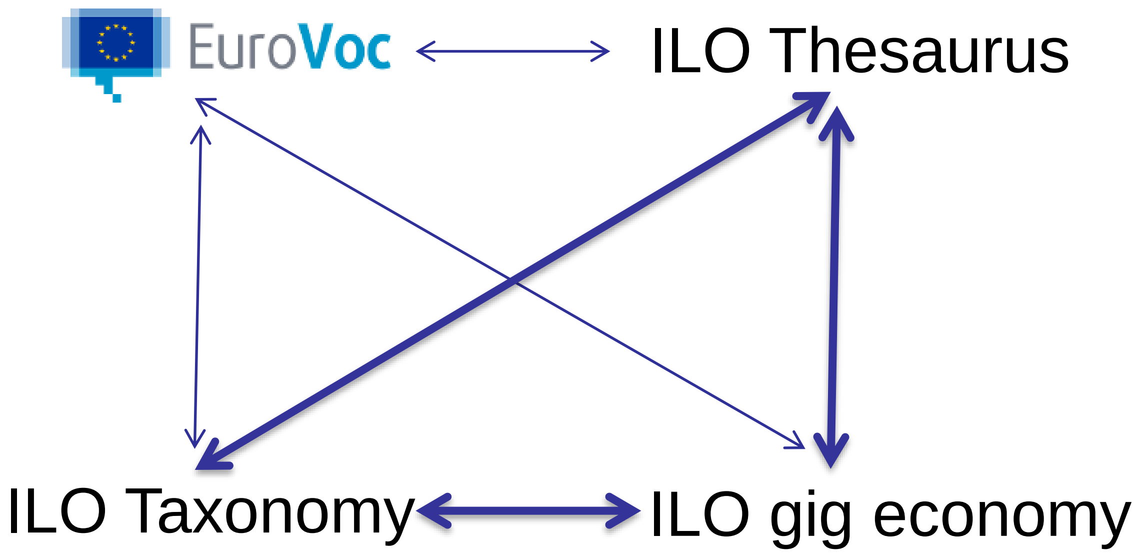


ILO Taxonomy

EuroVoc

ILO gig economy

ILO Thesaurus



EuroVoc

ILO Thesaurus

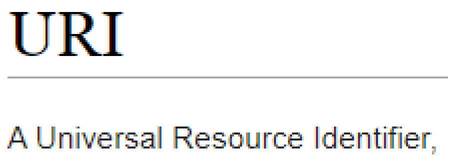
ILO Taxonomy

ILO gig economy

INTERMEDIARY RESULTS AND NEXT STEPS

Preliminary conclusions

- ILO assets
  - Some inconsistencies (content level)
    - Language misbalance
    - Use of notes
    - Granularity
  - No link to semantic technologies (identifiers, standards)
  - Integration levels
- EuroVoc
  - A few inconsistencies (translated prefLabels)
- Options
  - Various language versions aligned
  - At prefLabel level
  - Equality, *skos:exactMatch*



URI

A Universal Resource Identifier,

Next steps: Phase 2

- Finalising the improvements of all selected assets
- Implementing semantic technologies for all assets
- Disseminating and implementing validated mappings

- Automatic (semantic) alignment (same vocabularies)
- Streamlining efforts to ensure
  - Consistency
  - Understanding inside and between organisations
- Enriching vocabulary content
  - Inserting mappings
  - Structuring terminology asset(s)
  - Disseminating and implementing validated mappings (improving website search features)

THE BIGGER PICTURE OF THE DIGITAL AGE

Why data is important for tomorrow and… today?

- A common need
  - Understanding on which basis decisions are made
  - Accessing information

Not tomorrow but already today

- Millions of words translated every year
- Loss of information

- No human being can read, keep in mind, analyse, find or easily re-use these huge amounts of information
- Computers can (structured contents)
- Information can produce information thanks to A.I.

And beyond

- Transforming or enriching a textual corpus (semantic annotations, NLP, ML, AI) to
  - Make human-readable documents accessible to machines and apps
  - Create knowledge graphs

- Semantic annotations, e.g.
  - EUR-Lex
  - ELI project (today and soon)
    - insertion of descriptors at the document level
    - (and soon) at the paragraph level
  - At sentence / segment level?

In a nutshell

- Build on **synergies to increase the return on investment**

- Link 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**

Contacts

For more information please contact:

- denis.dechandon@publications.europa.eu
- aniko.gerencser@publications.europa.eu
- recortruiz@ilo.org

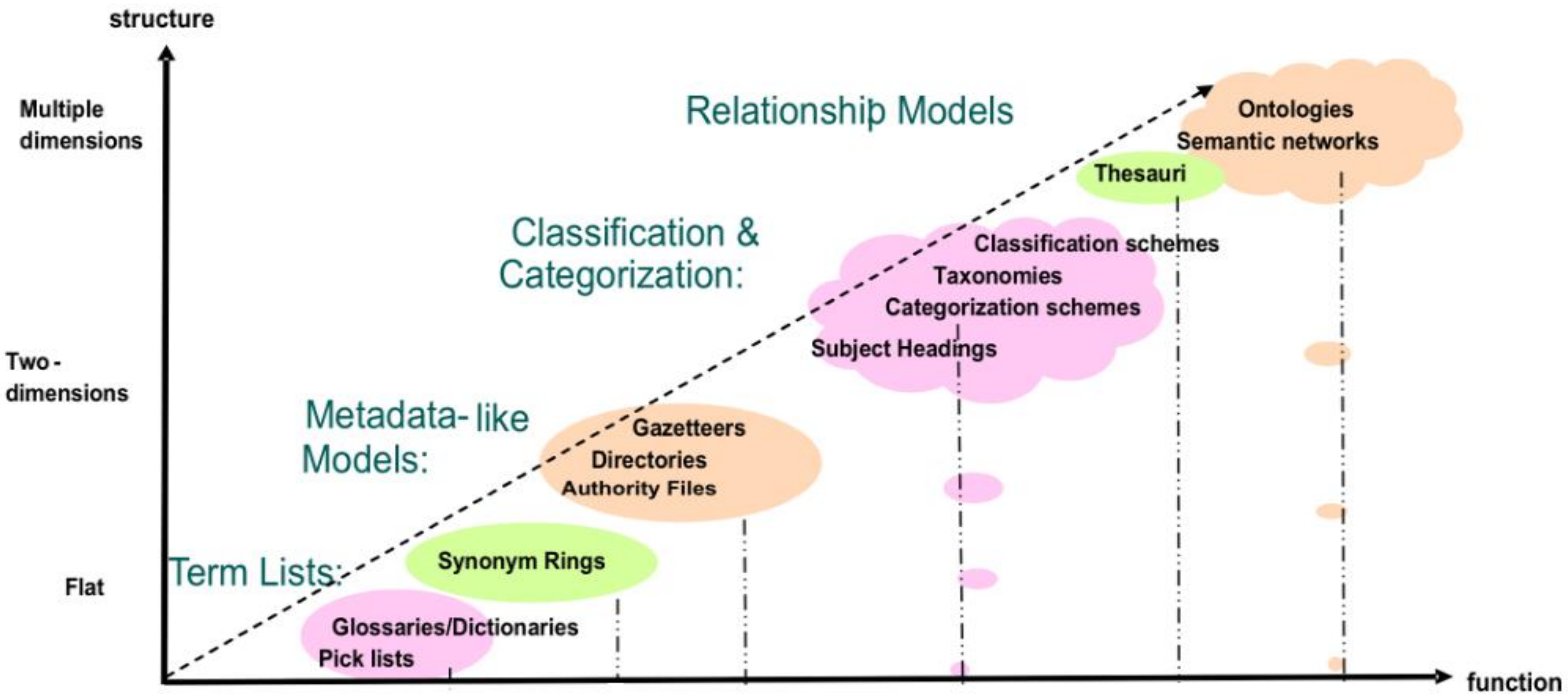
ADDITIONAL INFORMATION

Semantic web, terminology and corpora – Building bridges

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

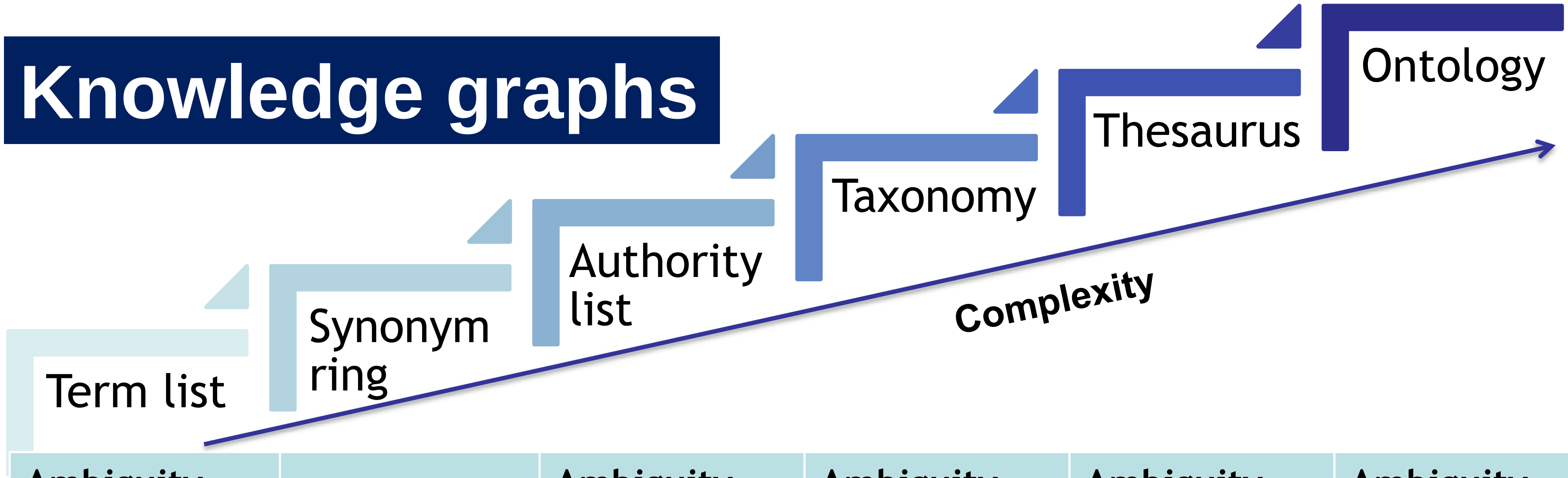
Semantic web, controlled vocabularies, metadata, identifiers

Zeng, Marcia L. 2008. "Knowledge Organization Systems (KOS)". *Knowledge Organization* 35, nos. 2-3: 160–182.



From terminology assets to complex controlled vocabularies

Knowledge graphs



Ambiguity control		**Ambiguity control**	**Ambiguity control**	**Ambiguity control**	**Ambiguity control**
	Synonym control	Synonym control	Synonym control	Synonym control	Synonym control
			Hierarchical relations	Hierarchical relations	Hierarchical relations
				Associative relations	Associative relations
					Properties and classes

- Interoperability
- Discoverability
- Retrieval
- Re-use

Publications Office and ILO, why this collaboration?

- Recurring participation to various events
  - Translating and the Computer
  - JIAMCATT
  - Taxonomy Boot Camp
  - IFLA
- For identical purposes
  - Multilingualism
  - Content indexing
  - Easing the access to information
- Similar vocabulary types…
  - Terminology assets
  - Translation tables
  - Controlled vocabularies

Selected assets

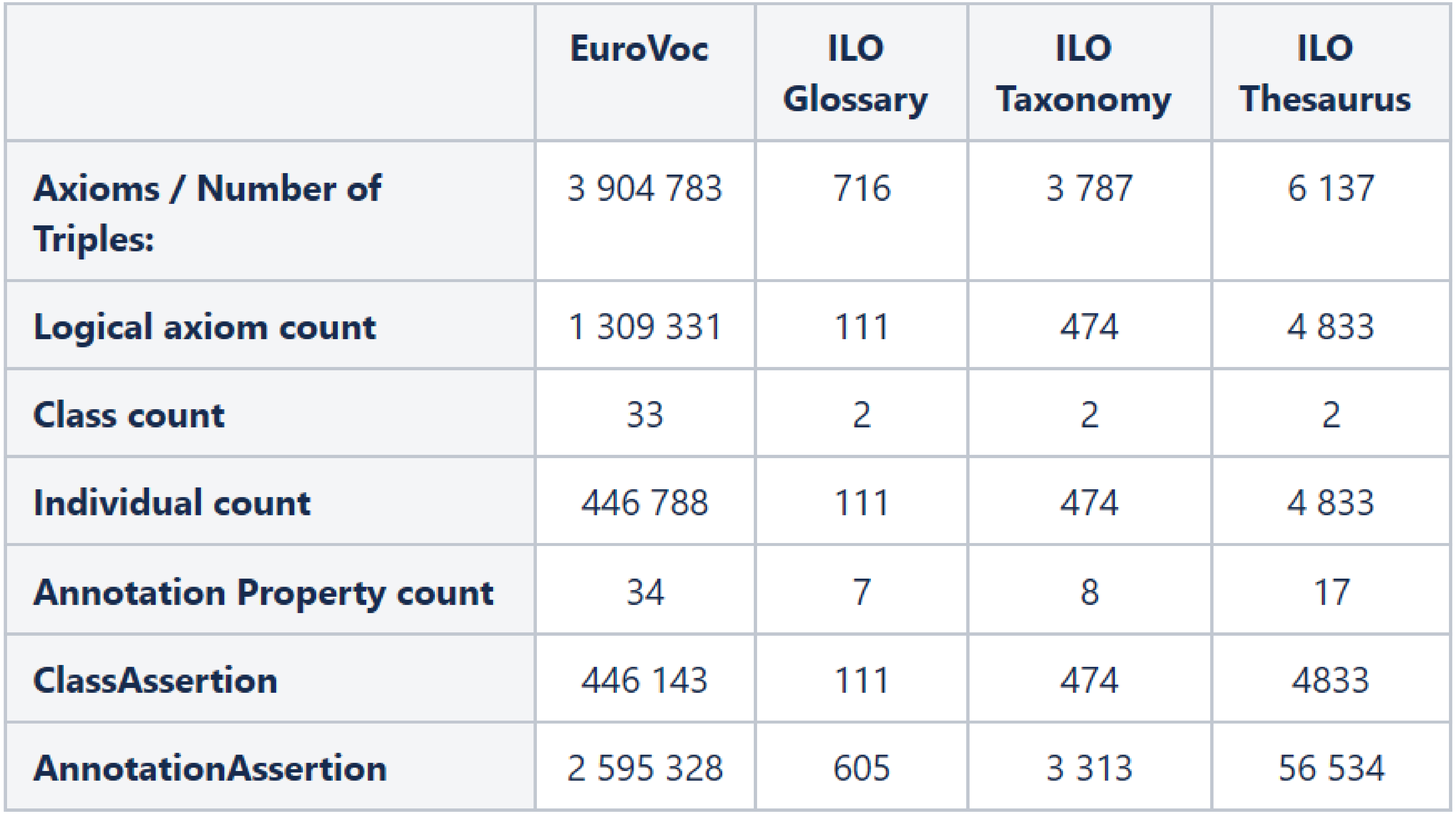
ILO Taxonomy

ILO Thesaurus

EuroVoc

ILO gig economy glossary

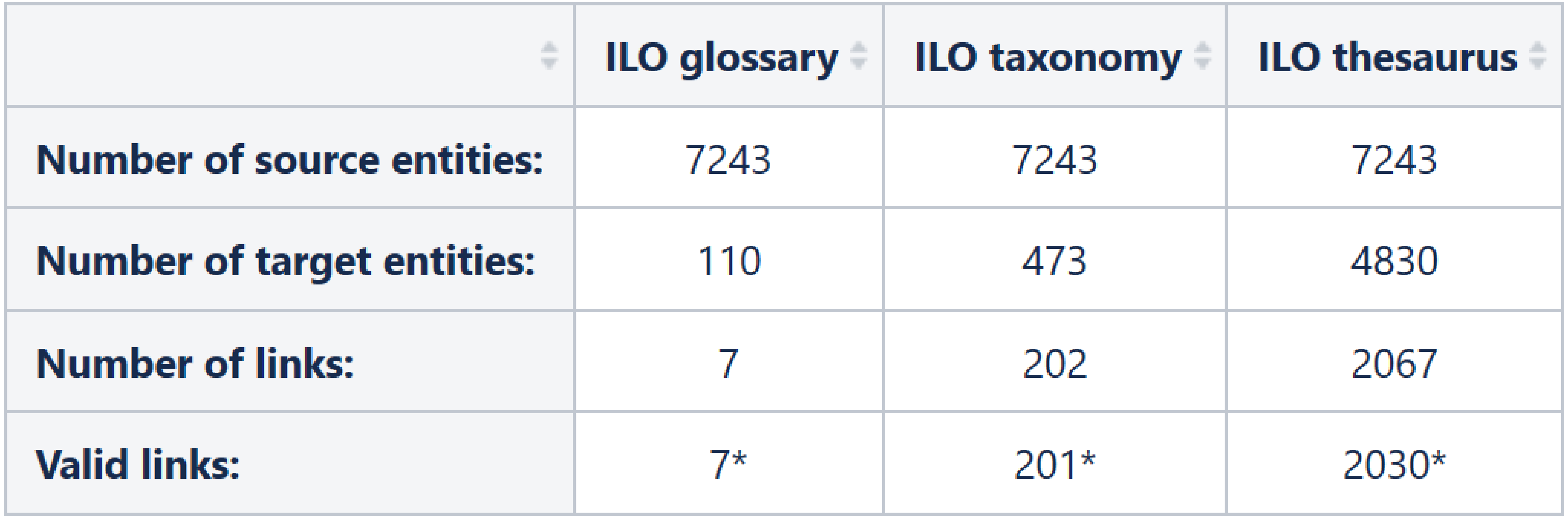
Some figures



	EuroVoc	ILO Glossary	ILO Taxonomy	ILO Thesaurus
Axioms / Number of Triples:	3 904 783	716	3 787	6 137
Logical axiom count	1 309 331	111	474	4 833
Class count	33	2	2	2
Individual count	446 788	111	474	4 833
Annotation Property count	34	7	8	17
ClassAssertion	446 143	111	474	4833
AnnotationAssertion	2 595 328	605	3 313	56 534

Results and observations (EuroVoc as a hub)

■Equality matching

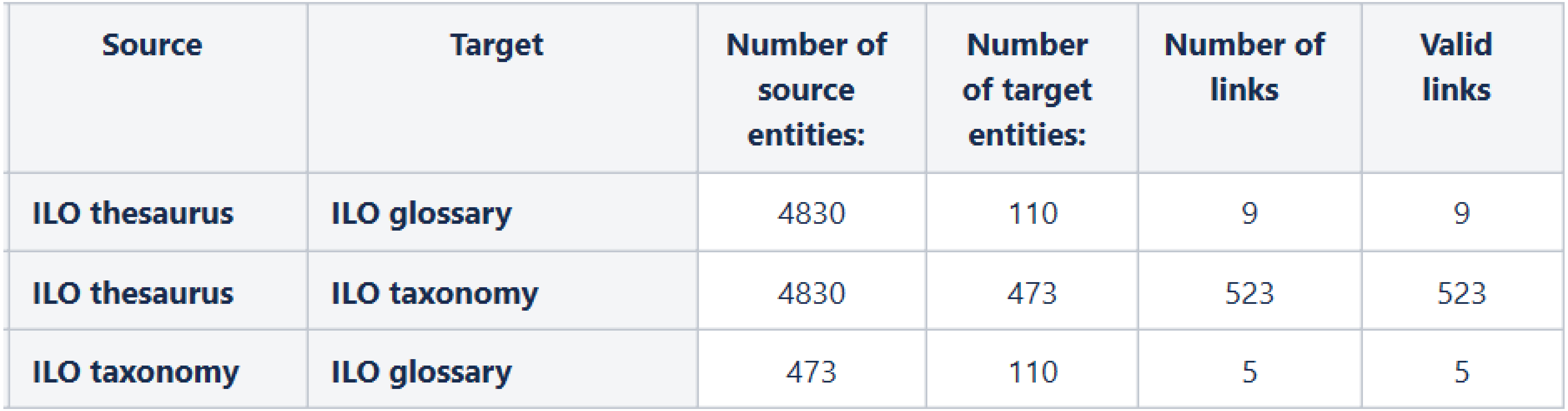


	ILO glossary	ILO taxonomy	ILO thesaurus
Number of source entities:	7243	7243	7243
Number of target entities:	110	473	4830
Number of links:	7	202	2067
Valid links:	7*	201*	2030*

*All links have been manually validated and are correct (except for deviations induced by inconsistencies in the definition of the concept labels either in ILO assets or EuroVoc)

Results and observations (ILO assets only)

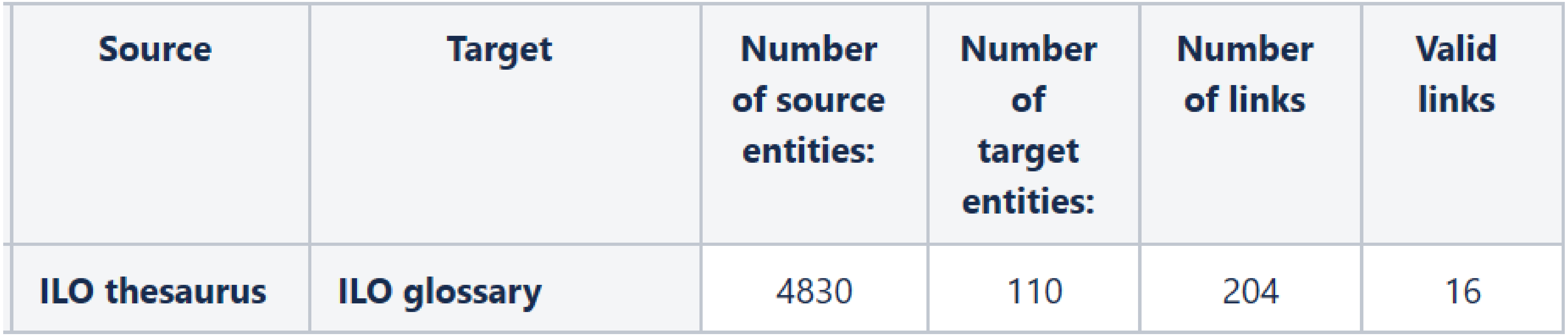
- **Equality matching**



Source	Target	Number of source entities:	Number of target entities:	Number of links	Valid links
ILO thesaurus	**ILO glossary**	4830	110	9	9
ILO thesaurus	**ILO taxonomy**	4830	473	523	523
ILO taxonomy	**ILO glossary**	473	110	5	5

Results and observations (ILO assets only)

- **Equality matching (Jaccard algorithm)**



Source	Target	Number of source entities:	Number of target entities:	Number of links	Valid links
ILO thesaurus	**ILO glossary**	4830	110	204	16