Modelling and Publishing the Chinese Information Retrieval Lexicon with VocBench

Yikun Han

Sichuan University, China

hanyikun1@stu.scu.edu.cn

Shimin Yan

Sichuan University, China

yanshimin@stu.scu.edu.cn

Wei Fan

Supervisor

Sichuan University, China

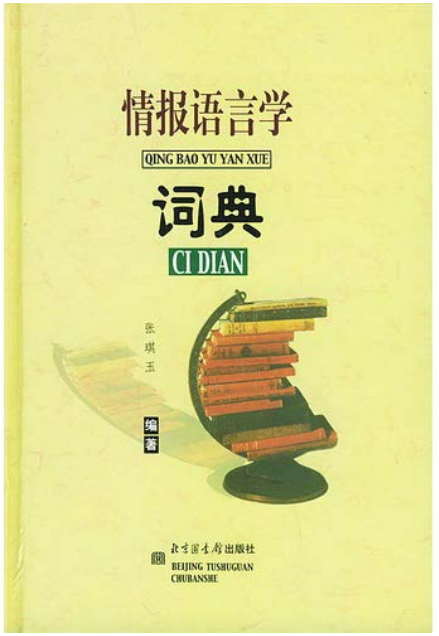
fanw@scu.edu.cn

DCMI 2021 VIRTUAL

Online, 2021 October 4-15

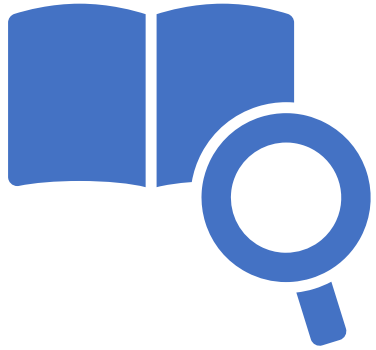
Outline

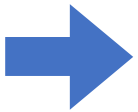
1. Motivation
2. Lexicon data modelling
   - 2.1 SKOS-based modelling
   - 2.2 Ontolex-Lemon
3. VocBench semantic publishing
   - 3.1 Sheet2RDF
   - 3.2 Concept Module
   - 3.3 Semantic Search
4. Future prospects
5. Acknowledgement



Chinese Information Retrieval Lexicon
Beijing Library, 2000

Digitizing Lexicons

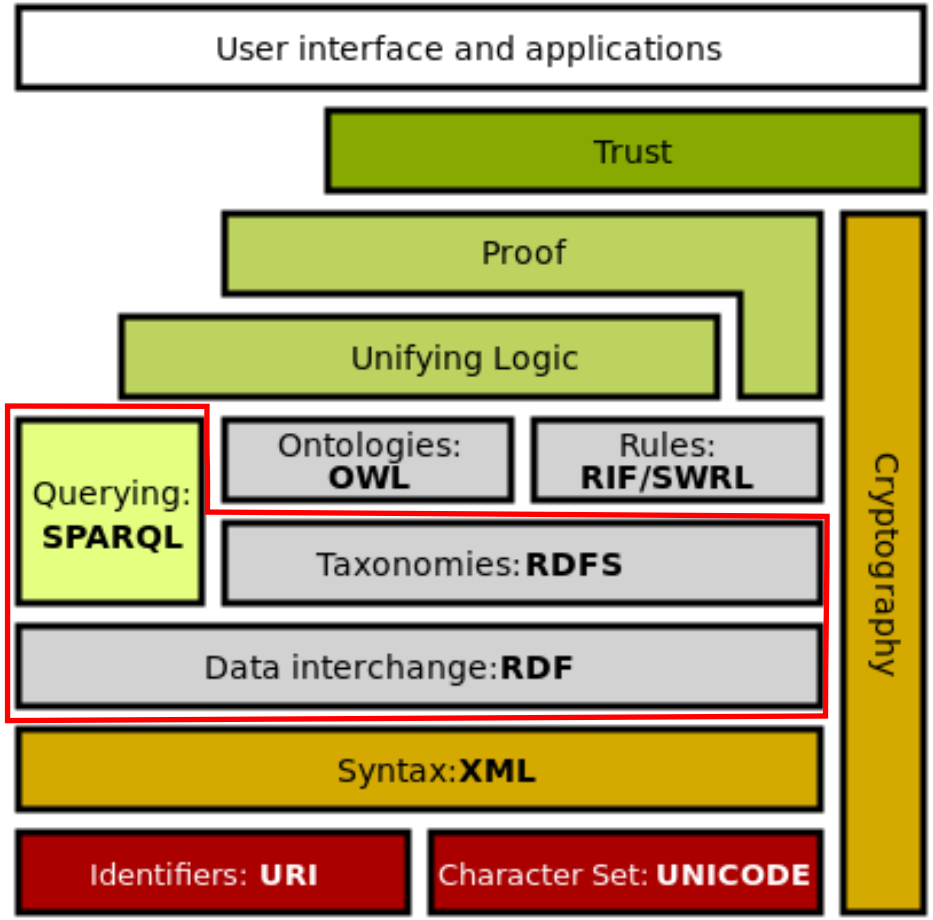




Paper-based Lexicon	Web-based Lexicon
• Outdated • Linear • Browse	• Up-to-date • Network • Semantic search

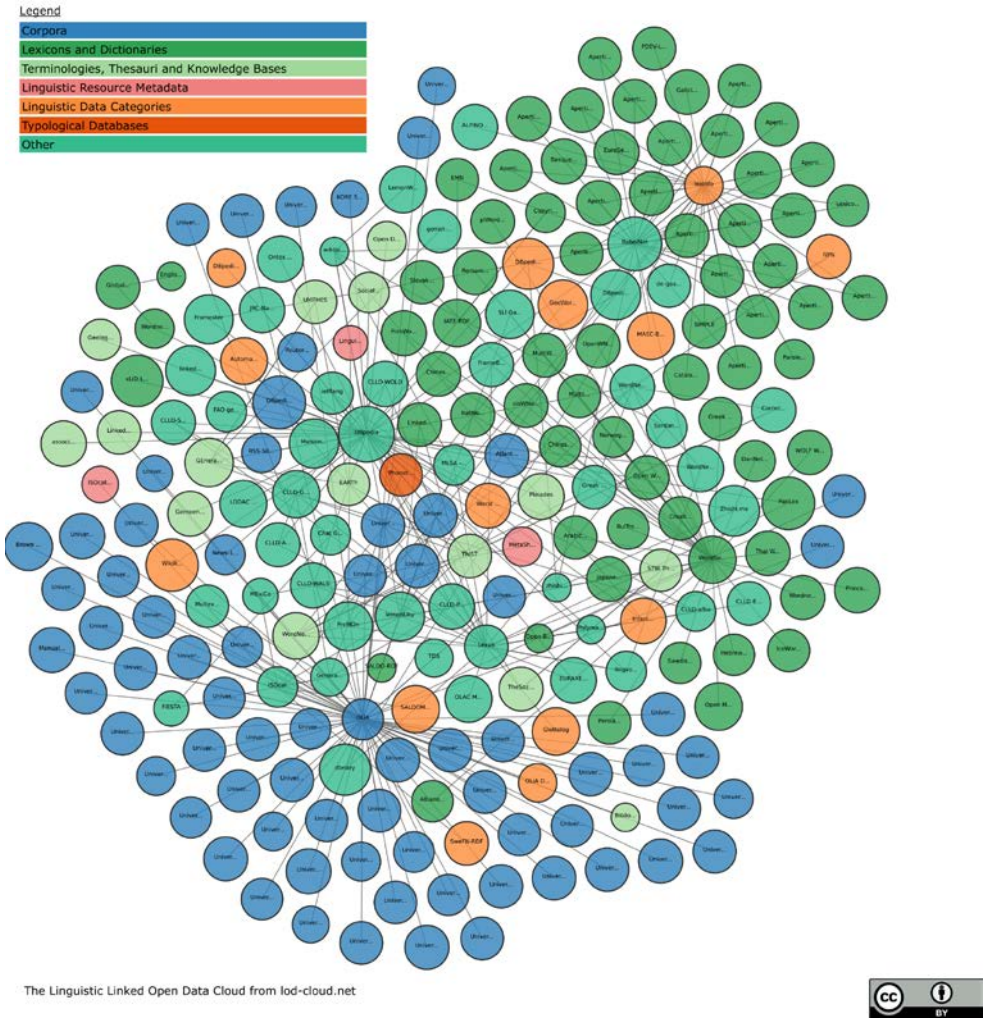
Semantic Web & Linked Data

Semantic web layer cake



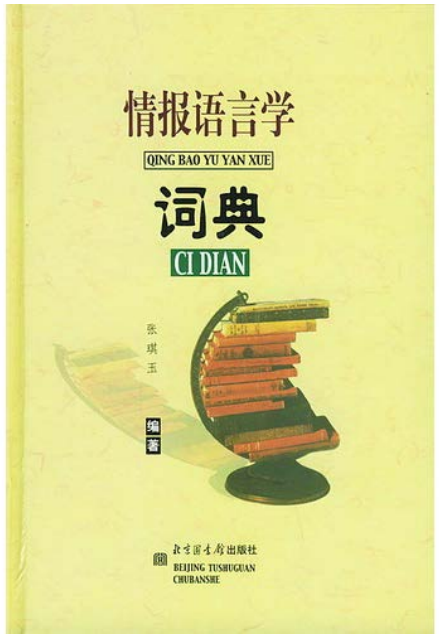
https://en.wikipedia.org/wiki/Semantic_Web_Stack

turn into



https://lod-cloud.net/

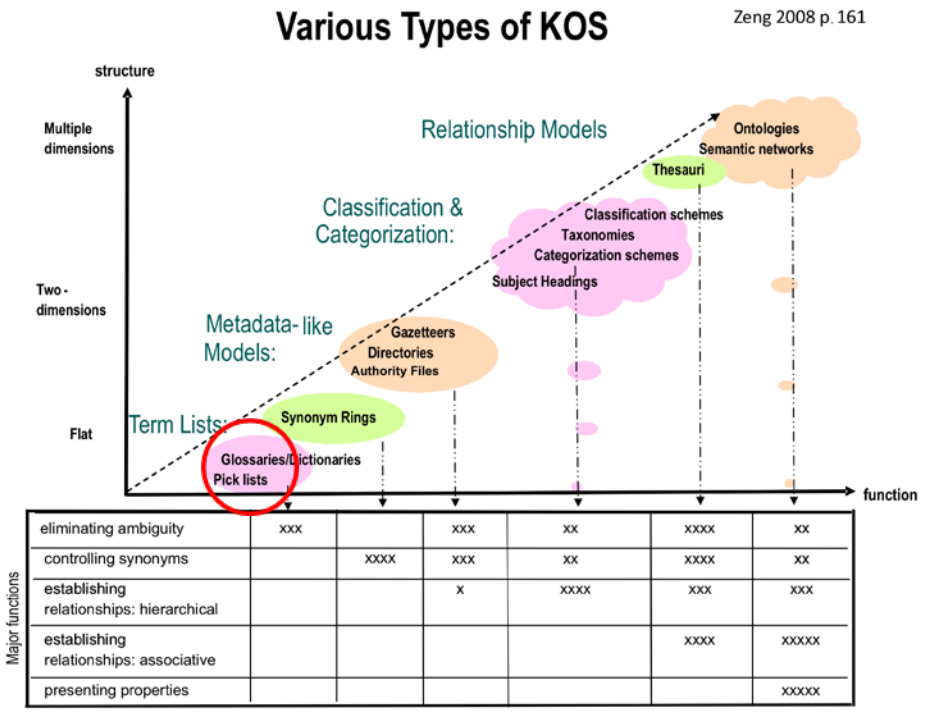
The Chinese Information Retrieval Lexicon



- The theoretical system of information retrieval in Chinese LIS field
- Complement to natural language retrieval
- Similar to glossary

Prof. Zhang, Qiyu.
1930-2017

Lexicon in KOS landscape



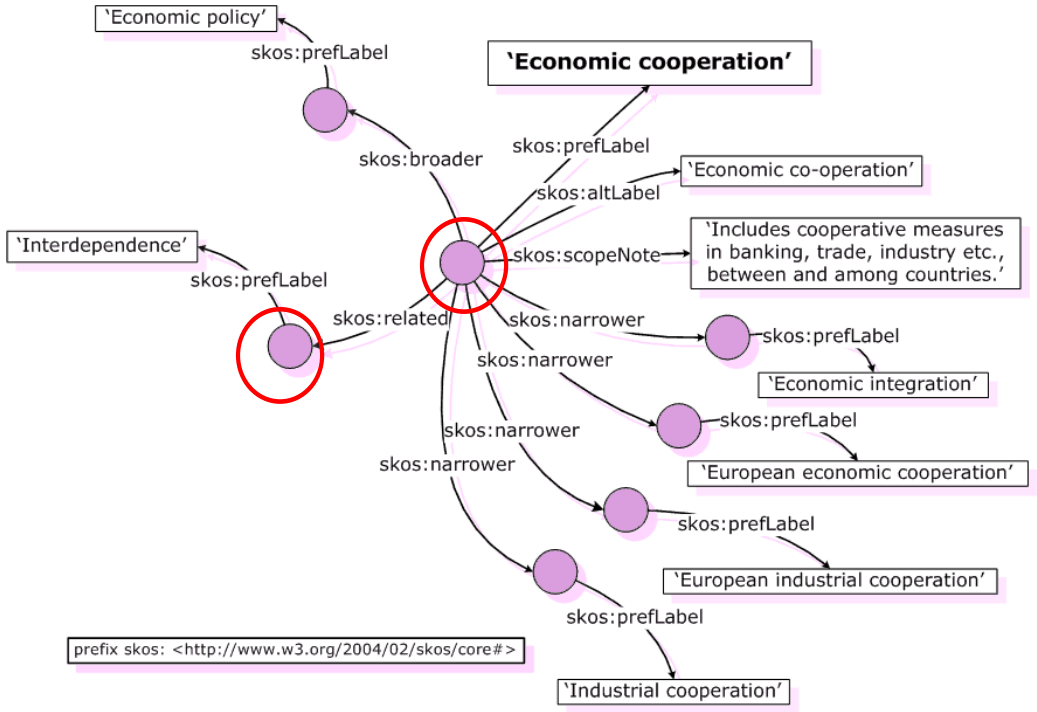
An overview of the structures and functions of KOSs (Zeng 2008 p.161)

- **Knowledge organization system**
  - *Knowledge organization system is a generic term used for referring to a wide range of items (e.g. subject headings, thesauri, classification schemes and ontologies), which have been conceived with respect to different purposes, in distinct historical moments.*

https://www.isko.org/cyclo/kos.htm

Specific type of KOSs

Simple Knowledge Organization System



https://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/

- **RDF-based vocabulary**
  - W3C recommendation
  - Designed for KOS
  - Concept-centric
  - Linked and integrated

SKOS-based Lexicon Modelling

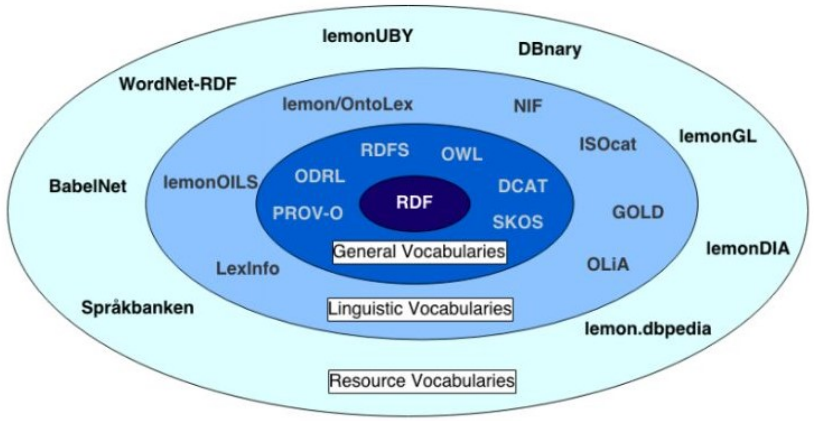
- **Concept-based modelling**
  - Concept & term: lexical labels
  - Concept & content: documentation properties
  - Concept hierarchy: semantic relations
  - Concept association: semantic relations

Lexicon elements	SKOS vocabularies
lexicon	*skos:ConceptScheme*(Class)(URI)
A-Z index	*skos:Collection*(Class)(URI)
identifier & preferred term	*skos:Concept*(Class)(URI)
preferred term	*skos:prefLabel*(Property)(Literal)
entry term	*skos:altLabel*(Property)(Literal)
content	*skos:definition/skos:example/skos:scope Note*(Property)(Literal)
hierarchy	*skos:broader/skos:narrower*(Property)(URI)
associative relationship	*skos:related*(Property)(URI)

Correspondence table

Ontological modelling

http://lider-project.eu/lider-project.eu/sites/default/files/referencecards/How-to-publish-linguistic-linked-data-Reference-Card.pdf

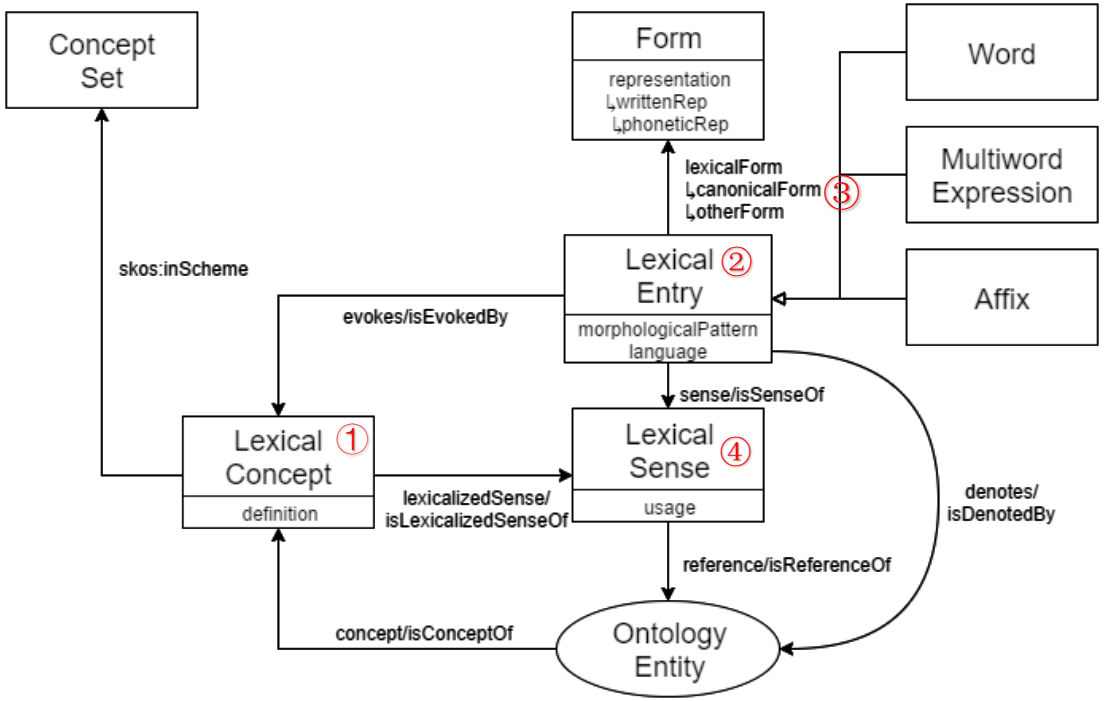


标引方式 □0077

在标引过程中对文献主题的取舍选择方式，决定提供哪些检索途径。主要有：(1)整体标引，是指对一部书或一篇文章的整体主题用一个标识来进行概括性标引；(2)全面标引，是指对一篇文献的各个局部主题或构成完整主题的各个主题因素分别标引。如有必要，同时对整体主题作概括性标引；如无必要，也可不再 对 整体主题作标引；(3)补充标引，是指除了对一篇文献的整体主题作概括性标引外，又对个别重要的局部主题或主题因素作单独标引；(4)重点标引（又称部分标引、局部标引、对口标引），是指仅仅选择一篇文献中对本单位服务对象有情报价值的个别局部主题作标引。

- **Why not SKOS?**
  - General purpose for KOS
  - Unsegmented definition
- **Why Ontolex-Lemon?**
  - Lexical model
  - Multiple senses

Ontolex-Lemon core model and encoding



https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

标题 □0056

(1)指标题词；(2)指文章的题名。

Heading □0056

(1)Refers to header; (2)Refers to the title of the article.

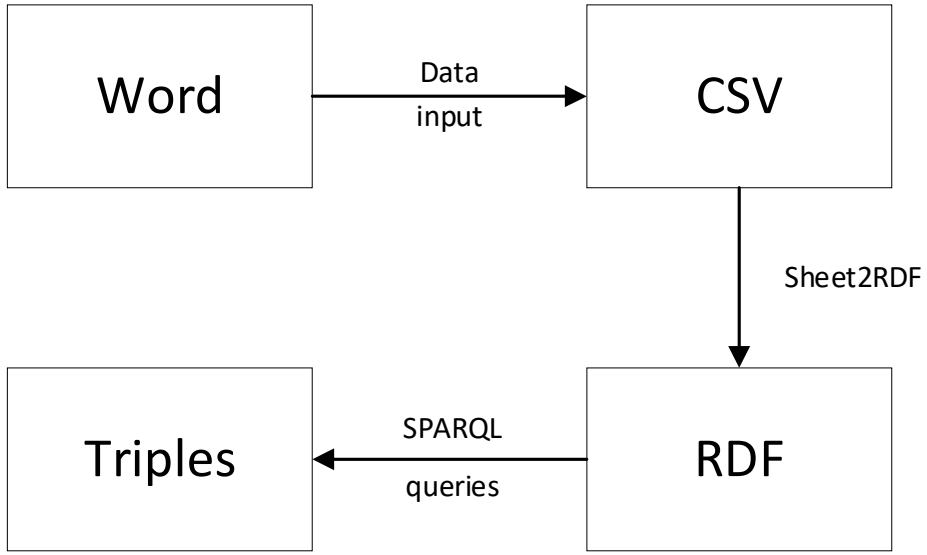
```
<C0056> ①
    a ontolex:LexicalConcept ;
    ontolex:lexicalizedSense <S0056> ;
    ontolex:lexicalizedSense <S0056_1> ;
    ontolex:lexicalizedSense <S0056_2> ;
    ontolex:isEvokedBy <E0056> ;
    ontolex:isEvokedBy <E0056_1> ;
    ontolex:isEvokedBy <E0056_2> .

<E0056> ②
    a ontolex:LexicalEntry ;
    ontolex:sense <S0056> ;
    ontolex:evokes <C0056> ;
    ontolex:canonicalForm <F0056_C> .

<E0056_1> ②
    a ontolex:LexicalEntry ;
    ontolex:sense <S0056_1> ;
    ontolex:evokes <C0056> .

<E0056_2> ②
    a ontolex:LexicalEntry ;
    ontolex:sense <S0056_2> ;
    ontolex:evokes <C0056> .
```

```
<F0056_C> ③
    a ontolex:canonicalForm ;
    ontolex:writtenRep "标题"@zh ;
    ontolex:writtenRep "Heading"@en ;
    ontolex:phoneticRep "biaoti"@zh-fonipa ;
    ontolex:phoneticRep "ˈhedɪŋ"@en-GB-fonipa .

<S0056> ④
    a ontolex:LexicalSense ;
    ontolex:isSenseOf <E0056> ;
    ontolex:isLexicalizedSenseOf <C0056> ;
    ontolex:usage[
    rdf:value "(1)指标题词；(2)指文章的题名。"@zh
    rdf:value "(1)Refers to header; (2)Refers to the title of the article."@en
    ] .

<S0056_1> ④
    a ontolex:LexicalSense ;
    ontolex:isSenseOf <E0056_1> ;
    ontolex:isLexicalizedSenseOf <C0056> ;
    ontolex:usage[
    rdf:value "(1)指标题词；"@zh
    rdf:value "(1)Refers to header; "@en
    ] .

<S0056_2> ④
    a ontolex:LexicalSense ;
    ontolex:isSenseOf <E0056_2> ;
    ontolex:isLexicalizedSenseOf <C0056> ;
    ontolex:usage[
    rdf:value "(2)指文章的题名。"@zh
    rdf:value "(2)Refers to the title of the article."@en
    ] .
```

- **Entry-concept-sense**
  - Many-to-one-to-many
  - Semantic triangle

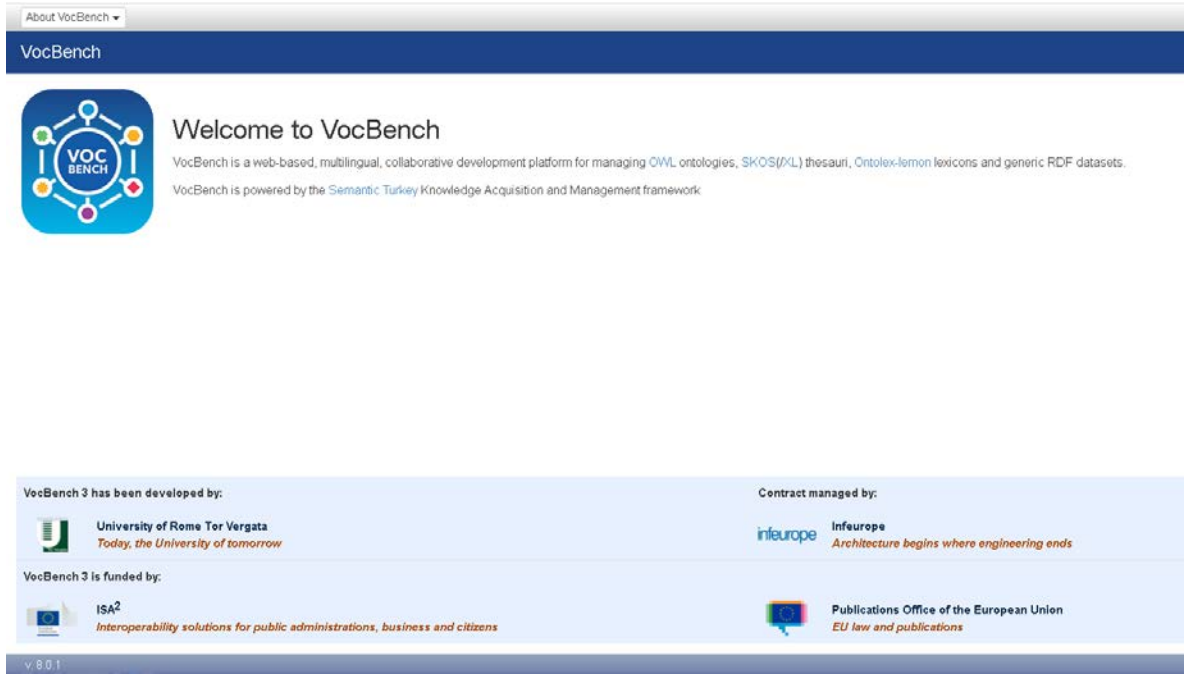
Data Conversion

From Modelling to Publishing



- **Requirement:**
  - A plug-in for transformation of datasheets into RDF

VocBench



Home Page: http://art.uniroma2.it/

- **Introduction**
  - Semantic web platform
  - Collaborative editing
  - Multilingual terminology

Publishing

- **Applications**
  - AGROVOC Multilingual Thesaurus
  - EuroVoc Thesaurus
  - InforMEA

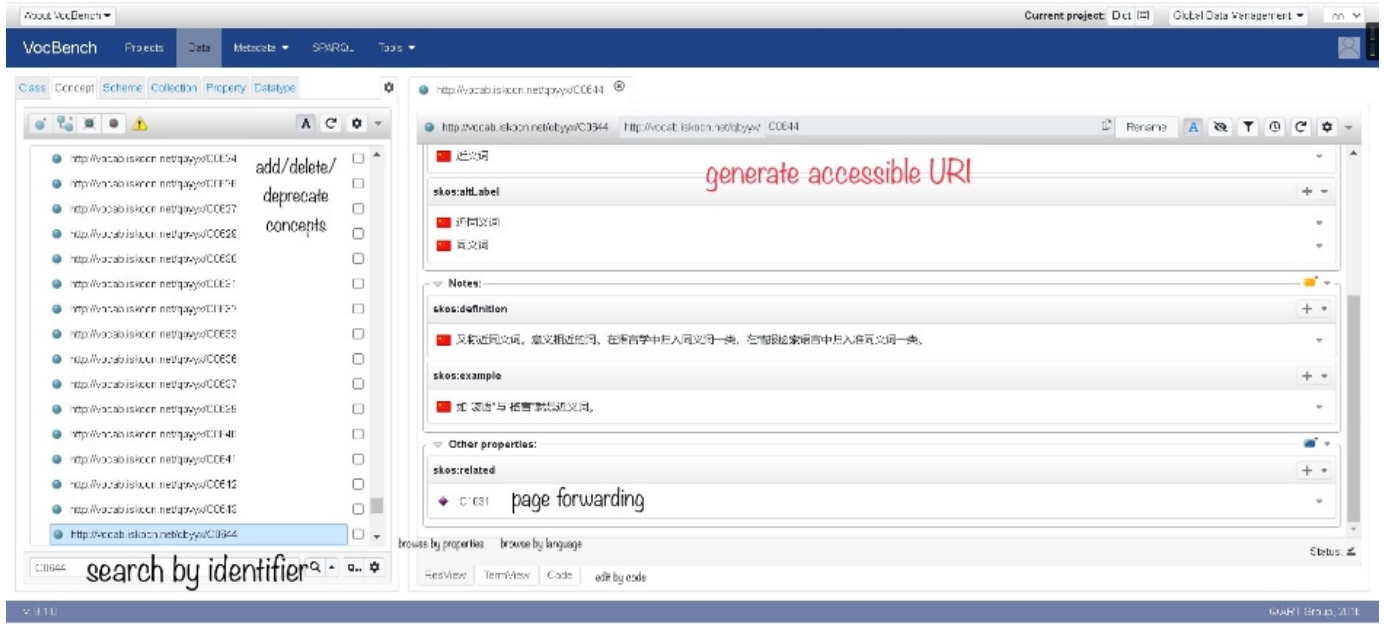
Sheet2RDF conversion

- **Subject mapping**
  - Header: *ontolex: Lexical Entry*
  - Header-based type
- **Pearl elements**
  - Nodes
  - Graph

Property	Accessbility
Header	optional
PEARL feature	read-only
Node ID	read-only
Type	optional

Concept module

- **Data update requirement**
  - Translation to different languages
  - Correspondence with other schemes
  - Metadata addition
  - Generation of new concepts
  - Change of lexical sense
  - Newfound association between concepts



Semantic search powered by SPARQL

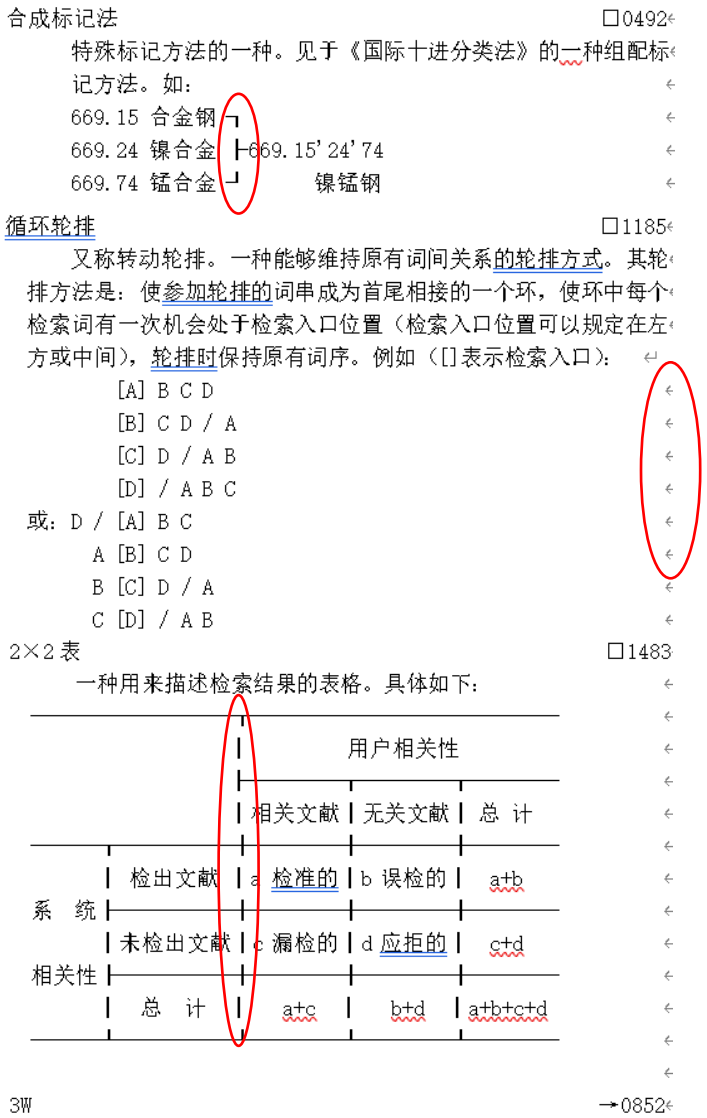
```
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
SELECT DISTINCT ?relationalType ?otherConcept
WHERE {
   BIND(<http://vocab.iskocn.net/qbyyx/C0100> AS ?LexicalConcept)
   {
   ?LexicalConcept ontolex:conceptRel ?otherConcept .
   BIND("conceptRel" AS ?relationalType )
   } UNION {
   ?LexicalConcept ontolex:isEvokedBy ?LexicalEntry .
   ?LexicalEntry ontolex:lexicalRel ?otherEntry .
   ?otherEntry ontolex:evokes ?otherConcept .
   BIND("lexicalRel" AS ?relationalType )
   }
}
```

- **Relational queries**
  - Select all related lexical concepts

```
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
SELECT (count(?otherForm)+count(?canonicalForm)as ?count)
WHERE {
   ?subject a ontolex:LexicalConcept .
   ?subject ontolex:isEvokedBy ?LexicalEntry .
   ?LexicalEntry ontolex:canonicalForm ?canonicalForm .
   ?LexicalEntry ontolex:otherForm ?otherForm .
}
```

- **Quantitative queries**
  - Counting all the labels in the lexicon

Issues



合成标记法 □0492

特殊标记方法的一种。见于《国际十进分类法》的一种组配标记方法。如：

669.15 合金钢 ┐
669.24 镍合金 ├669.15'24'74
669.74 锰合金 ┘ 镍锰钢

循环轮排 □1185

又称转动轮排。一种能够维持原有词间关系的轮排方式。其轮排方法是：使参加轮排的词串成为首尾相接的一个环，使环中每个检索词有一次机会处于检索入口位置（检索入口位置可以规定在左方或中间），轮排时保持原有词序。例如（[]表示检索入口）：

[A] B C D
[B] C D / A
[C] D / A B
[D] / A B C

或：D / [A] B C
A [B] C D
B [C] D / A
C [D] / A B

2×2表 □1483

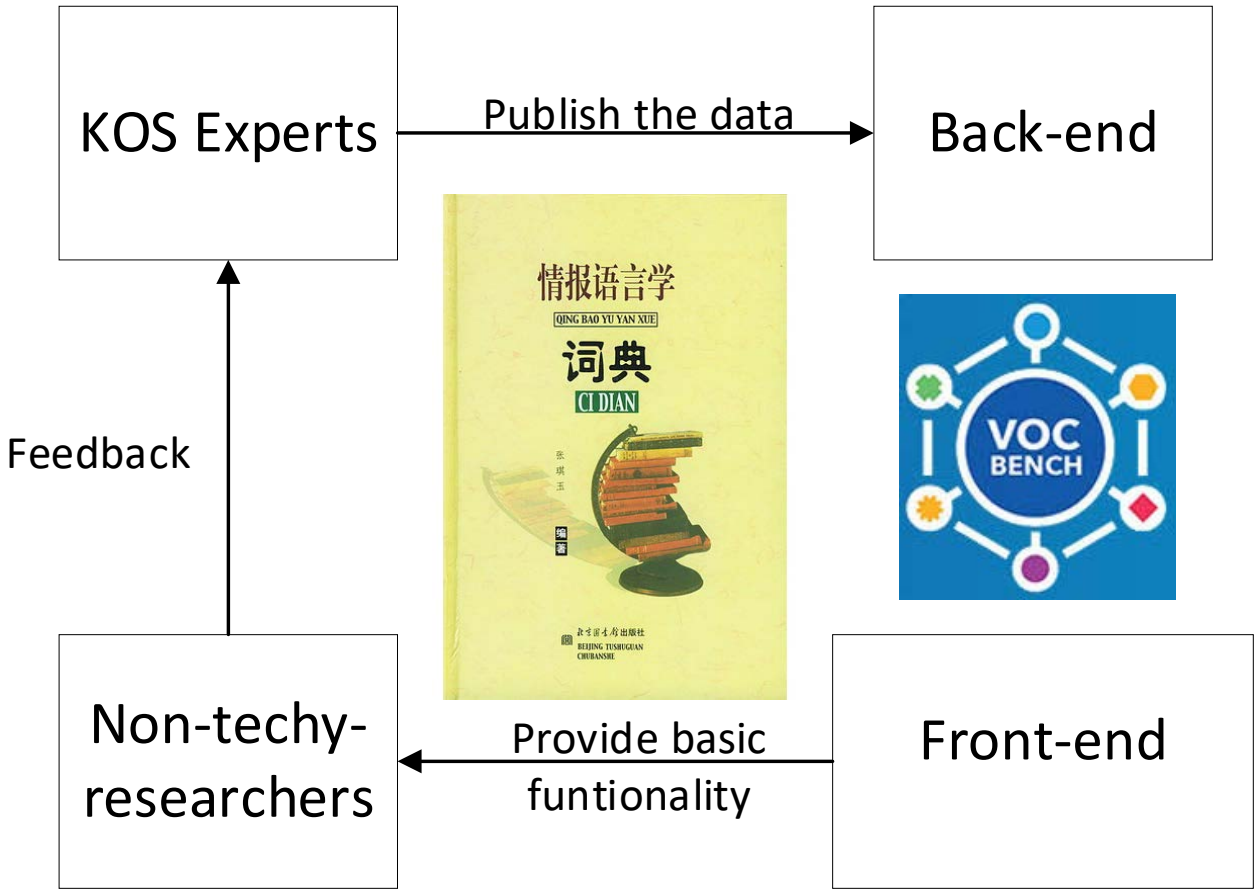
一种用来描述检索结果的表格。具体如下：

		用户相关性		
		相关文献	无关文献	总 计
系统相关性	检出文献	a 检准的	b 误检的	a+b
	未检出文献	c 漏检的	d 应拒的	c+d
	总 计	a+c	b+d	a+b+c+d

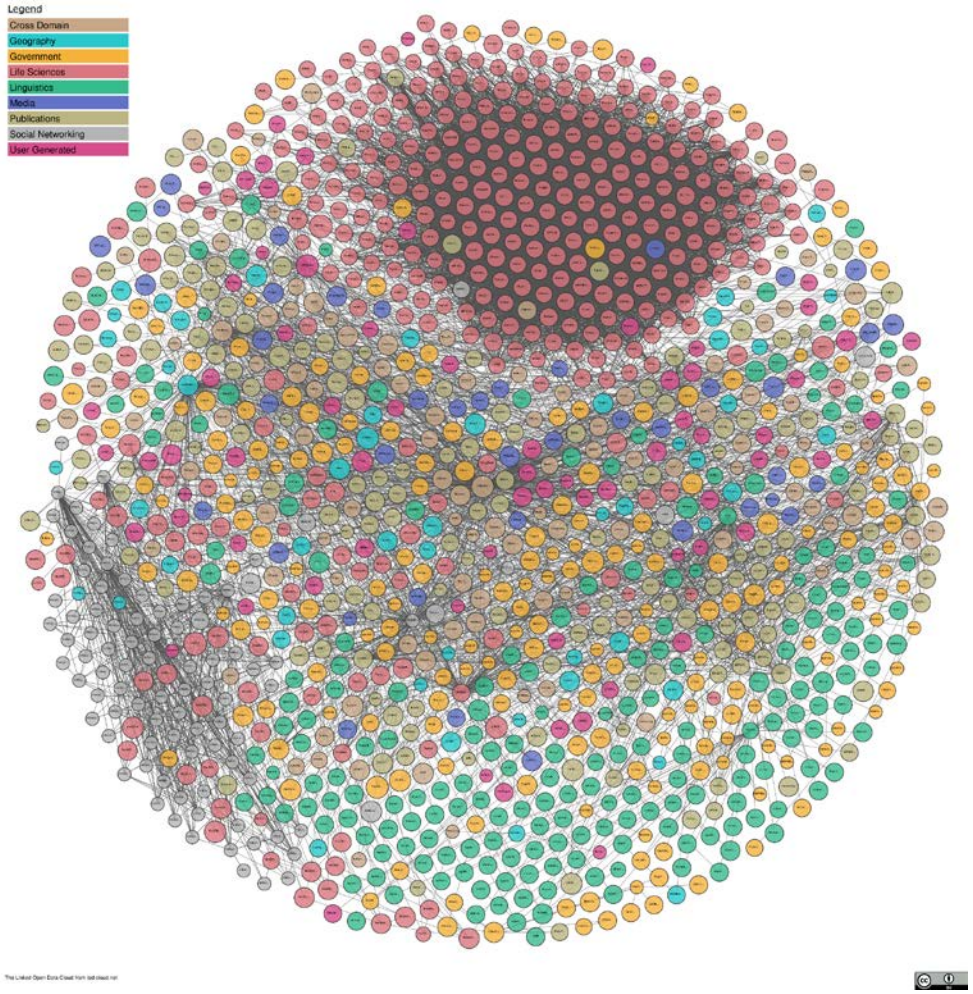
3W →0852

- **Semantic modelling**
  - Subclasses availability
- **RDF encoding**
  - Special symbols
- **RDF publishing**
  - Single-track conversion in VocBench

Future prospects



VocBench-powered



Linked Open Data Cloud

Acknowledgements

This work is a part of the Chinese Information Retrieval Terminology Knowledge Base Project, which is supported by the Chinese Index Society Foundation(No: CSI20A03).

Q&A

hanyikun1@stu.scu.edu.cn
yanshimin@stu.scu.edu.cn

DCMI 2021 VIRTUAL

Online, 2021 October 4-15