Language-related Linked (Open) Data for Knowledge Solutions, Artificial Intelligence and more

Christian Lieske, (SAP SE), Dr. Felix Sasaki (Cornelsen GmbH)
September, 2019

https://2019.semantics.cc/language-related-linked-open-data-knowledge-solutions-artificial-intelligence-and-more

Session 5.2

Wednesday, September 11, 11:45 am

Hall 2

Cornelsen

SEMANTiCS Karlsruhe 2019

THE BEST RUN SAP

Programme Announcement (includes last paragraph of Submission)

https://2019.semantics.cc/language-related-linked-open-data-knowledge-solutions-artificial-intelligence-and-more

Language-related Linked (Open) Data for Knowledge Solutions, Artificial Intelligence and more

This presentation sketches exploratory work on Language-related Linked Data. In particular, it reports on work and networks in the realm of schema.org, Wikidata, and Linguistic Linked Data. This work likely is of interest to companies, governmental institutions, researchers, and other constituencies involved in or in need for assistive solutions related to content in general and knowledge in particular. Examples for this include content syndication, multi-modal chatbot-based applications, and various Artificial Intelligence/Machine Learning areas.

Introductory Text from Submission

Many Linked (Open) Data scenarios and applications relate to human language, and human language technology (HLT). The relationship is twofold: On the one hand, human language is analyzed by HLT to yield Linked Data (see e.g. the "generation" of DBpedia from Wikipedia). On the other hand, HLT uses Linked Data in various processing contexts (see e.g. the use of Wikidata for Named Entity Linking).

Currently, activities related to a special area of Language-related Linked Data is picking up speed: Linguistic/Lexicographic Linked Data. With this move, the existing intuitive tools for end users, and the powerful interfaces for programmers make Linked Data relevant for even more usage scenarios – amongst others in the realm knowledge discovery, content syndication and enrichment, terminology work, and translation.

Possible Topics mentioned in Submission

Basics and exploratory work related to language and schema.org

- Which role can terminological assets play in today's Web – especially concerning the search context?
- Does Schema.org already include everything that is needed to realize cross-lingual scenarios easily?

Basics and exploratory work related to language and Wikidata

- Wikidata usage scenarios, and parts of the Wikidata tooling related to all things language (e.g. extraction of terminology for a certain domain, or generation of multilingual dictionaries).
- The Wikibase Lexeme extension (allow for example senses, inflections, compounding to be captured).

Basics and exploratory work related to Language-related/Linguistic Linked Data

- Data sets (textual, multimodal/multimedia and lexical data, grammars, language models, etc.) and tools/ technologies/services used for their processing
- Sample applications
- (European) spaces to watch

Storyline // Slides

Basics/In a Nutshell

- Linked data
  - Universal API
- Language-related data
  - LLOD cloud
- Knowledge Solutions in a Nutshell
  - Search and Rich Snippets

Schema.org

- What is it?
  - Voc in search
- Scenario
  - Terminology
  - Aside: language-related vocs/ formats
- Gap
- Closing the gap

Wikidata

- Scenarios without Lexeme extension
  - Bi-lingual for a domain
- Lexeme extension
  - Data model (graphic and ontolex)
- Scenarios with Lexeme extension
  - Etymology, compounding

LLOD

- Produce or use
- Agreed vocs/ formats
- Apps (naisc?, kahoot?)
- Projects

In a Nutshell – Linked Open Data (1/2)

Universal network of units with explicit meaning to bridge semantic gaps (like not knowing that a string of numbers is a price)

Semantic Web architecture (e.g. Resource Description Framework for standardized meaning representation); Data on a Solid Stack

Enable or facilitate computers and people to work better in cooperation

30 percent of HTML pages contain structured data

<span itemprop="price“
content="1000.00">1,000.00

https://schema.org/price

In a Nutshell – Linked Open Data (2/2)

Subject

Predicate

Object

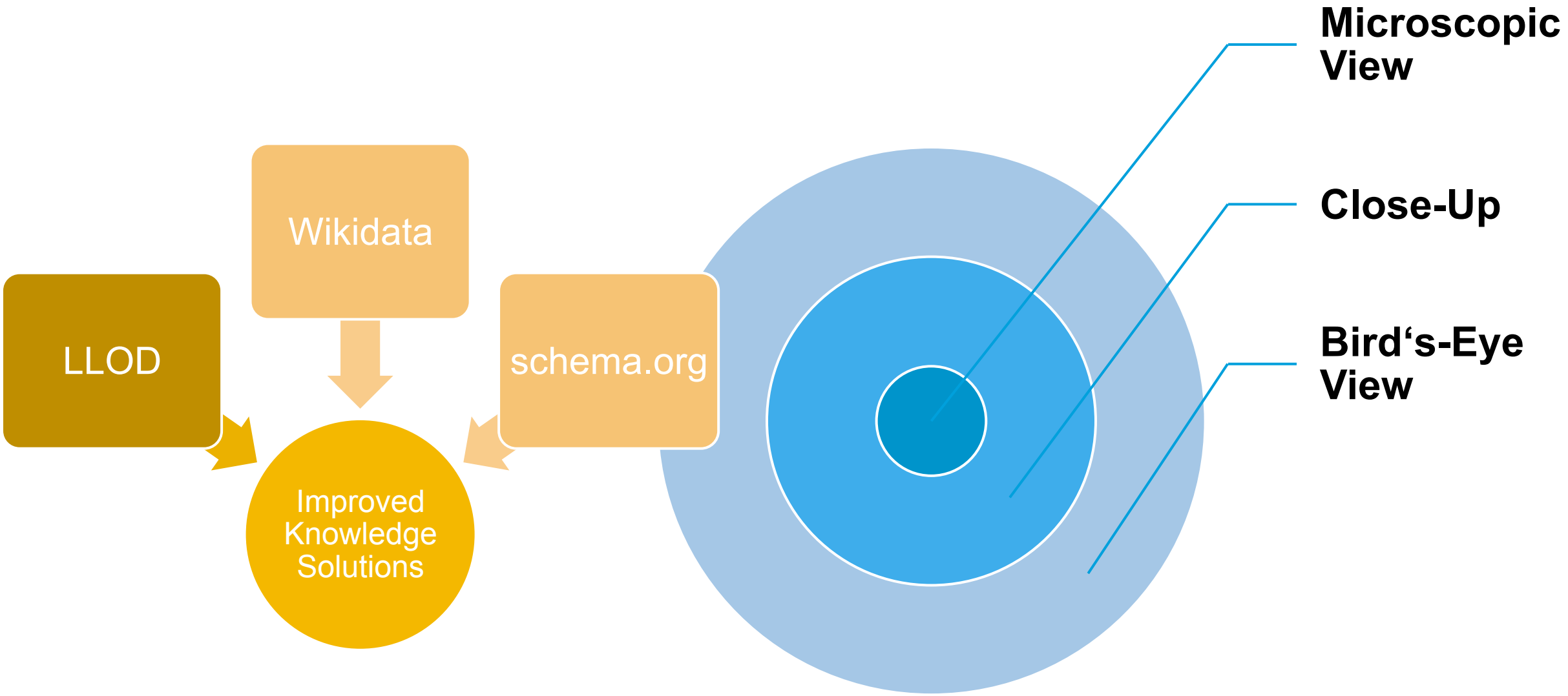
Conceptualization/model = explicit, machine understandable, consensus

Vocabulary = describes and represents an area of concern

Classifies terms that can be used in a particular application

Characterizes entities

Schema.org vocabulary allows search engines to recognize information easier and more precisely


Wikidata
LLOD
schema.org
Improved Knowledge Solutions
Microscopic View
Close-Up
Bird‘s-Eye View

Knowledge Solutions (1/2)

- Information Extraction
- Information Retrieval/Search
- Knowledge Discovery
- Content Enrichment
- Question Answering
- Quality Control

Recognize Entities

Input
- Welcome to Stuttgart

Output
- Stuttgart is an entity (class ”city”)

Recognize facts/statements/relations

Input
- Stuttgart is the capital of Baden-Wurttemberg

Output
- Stuttgart and Baden-Wurttemberg are linked. Stuttgart is ”known” as capital of Baden-Wurttemberg.

Link Entities

Input
- Stuttgart is the capital of Baden-Wurttemberg

Output
- ”Stuttgart” is linked to Wikipedia article about Stuttgart

Disambiguate Entities

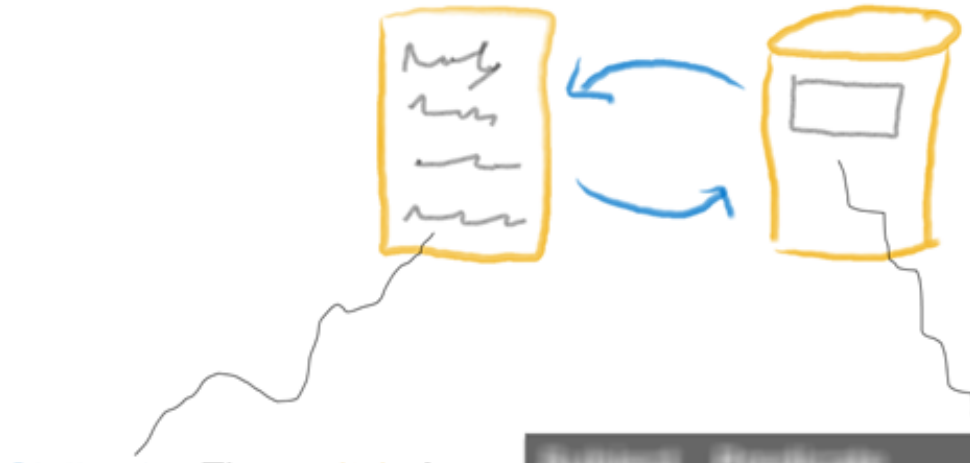
Input:
- Armstrong reached Paris in record time

Output
- ”Armstrong” is identified as cyclist

Knowledge Solutions (2/2)

Generate data (e.g. new entries in knowledge base)

Use data (to enable or improve process)



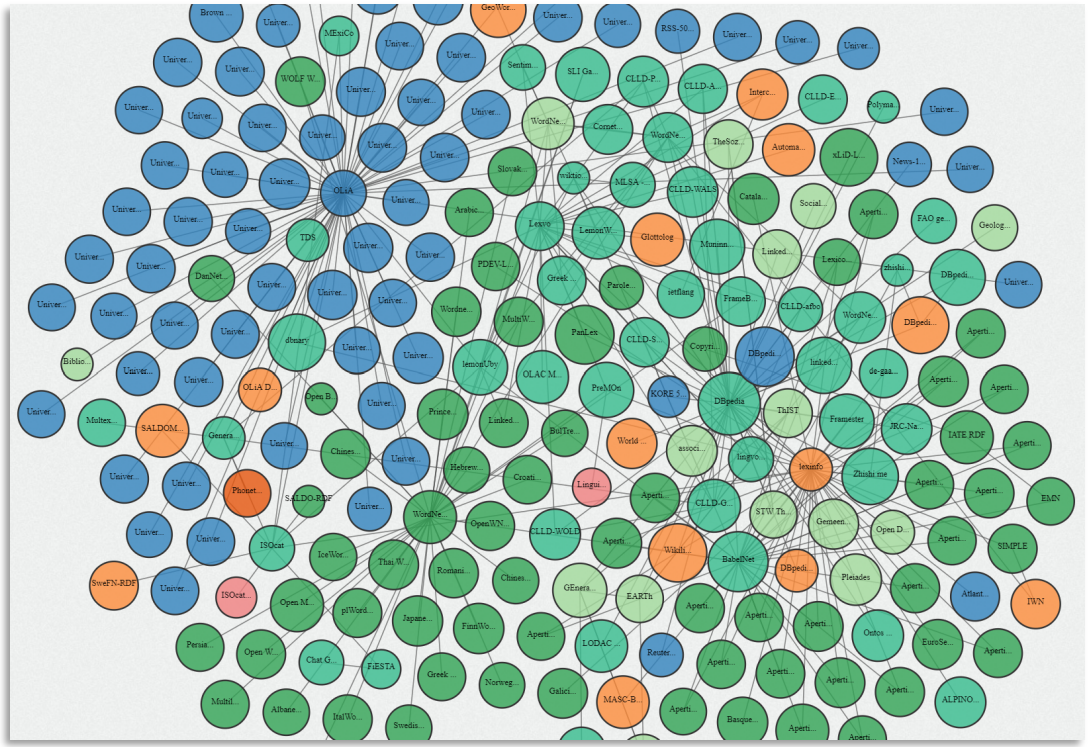
Stuttgart ... The capital of Baden-Wurttemberg ... Area code 0712 ... Nice sights include the Venusberg, ...

[illegible]	[illegible]	[illegible]
Stuttgart	is_capital_of	Baden-Wurttemberg
Stuttgart	has_area code	0711
Bonn	has_sight	Venusberg
…	…	…

Bird's-Eye View *Linguistic Linked Open Data (LLOD)*

"Method and interdisciplinary community concerned with creating, sharing, and (re-)using language resources in accordance with Linked Data principles"

Source: https://en.wikipedia.org/wiki/Linguistic_Linked_Open_Data



Source: http://linguistic-lod.org/

Close-Up *Linguistic Linked Open Data (LLOD)*

Data

- **Categories**: Lexicons, Dictionaries, Terminologies, Ontologies, Thesauri, Knowledge Bases, Linguistic Resource Metadata, Linguistic Data Categories, ...
- **Examples**: OLiA, WordNet, DBpedia, lexinfo.net, BabelNet, lexvo.org, DataCatReg, UBY, Glottolog, METAShare LingHub …

Models/formats/vocabularies

- Ontolex-Lemon (with modules like e.g. *vartrans*, *lexicography*)
- NLP Interchange Format (NIF)

Tools/Technical Services

- LIDER tools and FREME stack (entity recognition and linking, creation of eBooks, annotatation of text with terminology, conversion of terminology into RDF, …)
- DBpedia Spotlight, AIDA, Babelfy
- Terminoteca RDF, Apertium RDF
- WebBrain, MENTA, Klint, PIKES, KnEWS

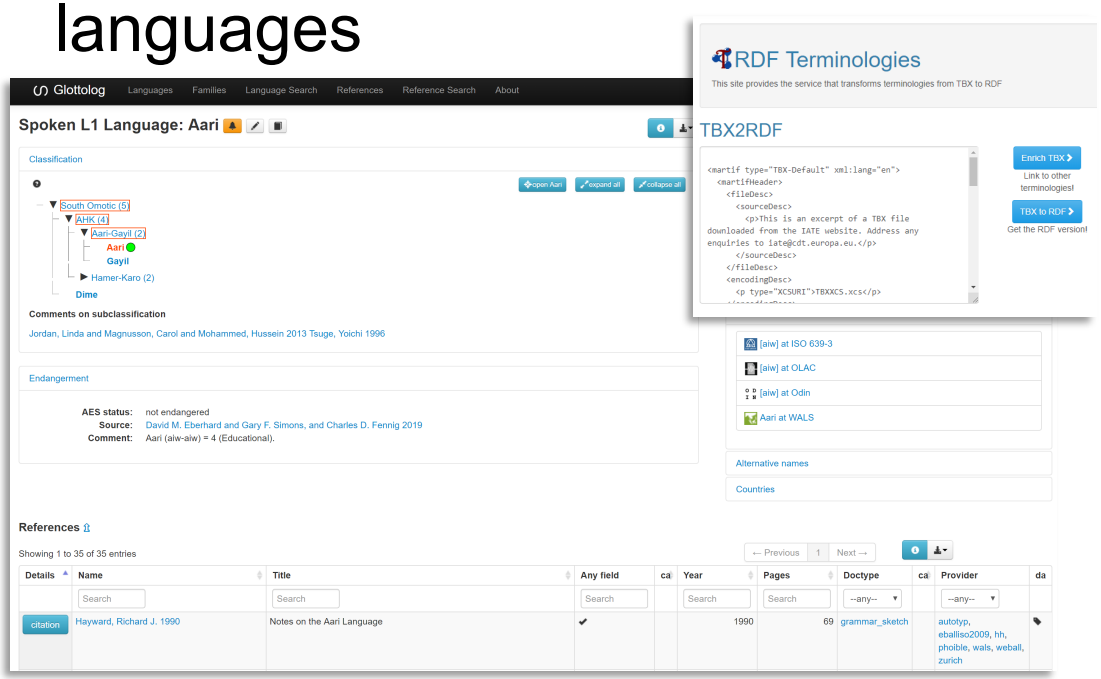
Community

- W3C Community Groups (e.g. W3C Linked Data for Language Resources community group)
- Summer Schools and Datathons
- Conferences (e.g. Language, Data and Knowledge with shared tasks such as *Translation Inference across Dicitionairies*)
- Research Projects (e.g. FREME, LIDER, LiODi, Prêt-à-LLOD, ELEXIS and Lynx)

Microscopic View *Linguistic Linked Open Data (LLOD)*

Exemplary Tools/Technical Services

- LIDER terminology conversion
- Glottolog info on lesser known languages



Exemplary LLOD achievements (papers from Language, Data and Knowledge 2019)

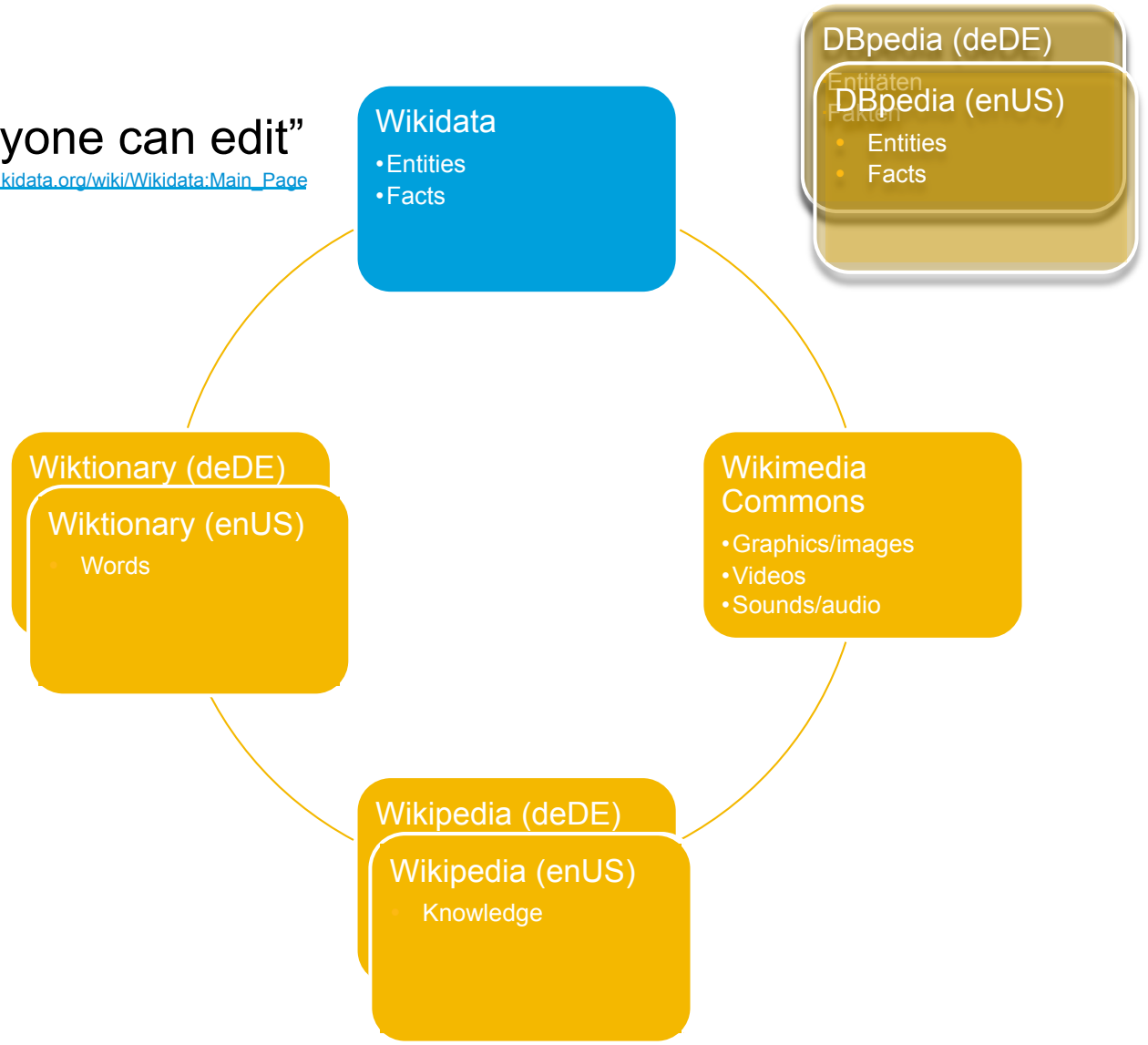
- Name Variants for Improving Entity Discovery and Linking
- Inflection-Tolerant Ontology-Based Named Entity Recognition for Real-Time Applications
- Efficient Harmonization of Concurrent Tokenization and Textual Variation
- Opening Digitized Newspapers Corpora: Europeana's Full-text Data Interoperability Case

Bird's-Eye View *Wikidata*

"The free knowledge base that anyone can edit"

Source: https://www.wikidata.org/wiki/Wikidata:Main_Page

Source: https://commons.wikimedia.org/wiki/File:Wikidata-logo-en.svg



Close-Up *Wikidata*

Language in general

- Any item in Wikidata can relate to language: label(s), description(s)

Language in detail

- Special item type (lexeme) allows to capture information on *gender, conjugation, inflection, derivation, antonym, false friend, phonetics, dictionaries/databases, colloquial form* and much more compounding
- Aligns with Ontolex-lemon

Beijing (Q956)

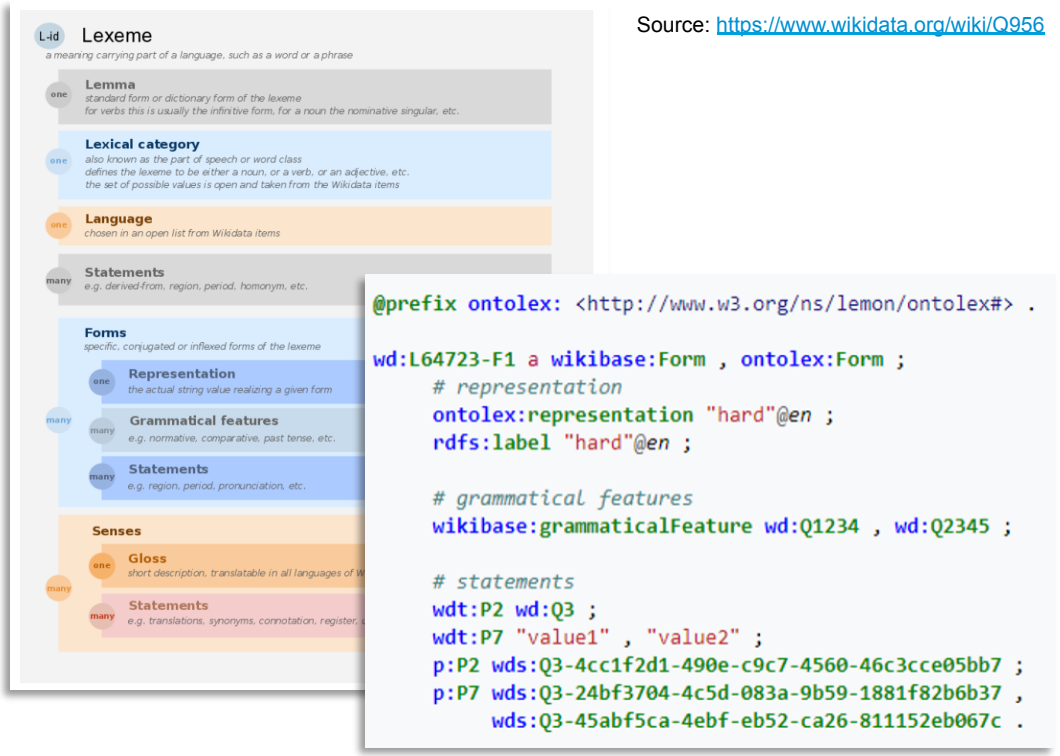
capital of China

Peking | Beiping | Peiping | Yanjing | Zhongdu | Khanbaliq | BJ

In more languages Configure

Language	Label	Description	Also known as
English	Beijing	capital of China	Peking Beiping Peiping Yanjing Zhongdu Khanbaliq BJ
German	Peking	Hauptstadt der Volksrepublik China	Beijing
French	Pékin	capitale de la Chine	Beijing

Source: https://www.wikidata.org/wiki/Q956




```
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .

wd:L64723-F1 a wikibase:Form , ontolex:Form ;
     # representation
     ontolex:representation "hard"@en ;
     rdfs:label "hard"@en ;

     # grammatical features
     wikibase:grammaticalFeature wd:Q1234 , wd:Q2345 ;

     # statements
     wdt:P2 wd:Q3 ;
     wdt:P7 "value1" , "value2" ;
     p:P2 wds:Q3-4cc1f2d1-490e-c9c7-4560-46c3cce05bb7 ;
     p:P7 wds:Q3-24bf3704-4c5d-083a-9b59-1881f82b6b37 ,
          wds:Q3-45abf5ca-4ebf-eb52-ca26-811152eb067c .
```

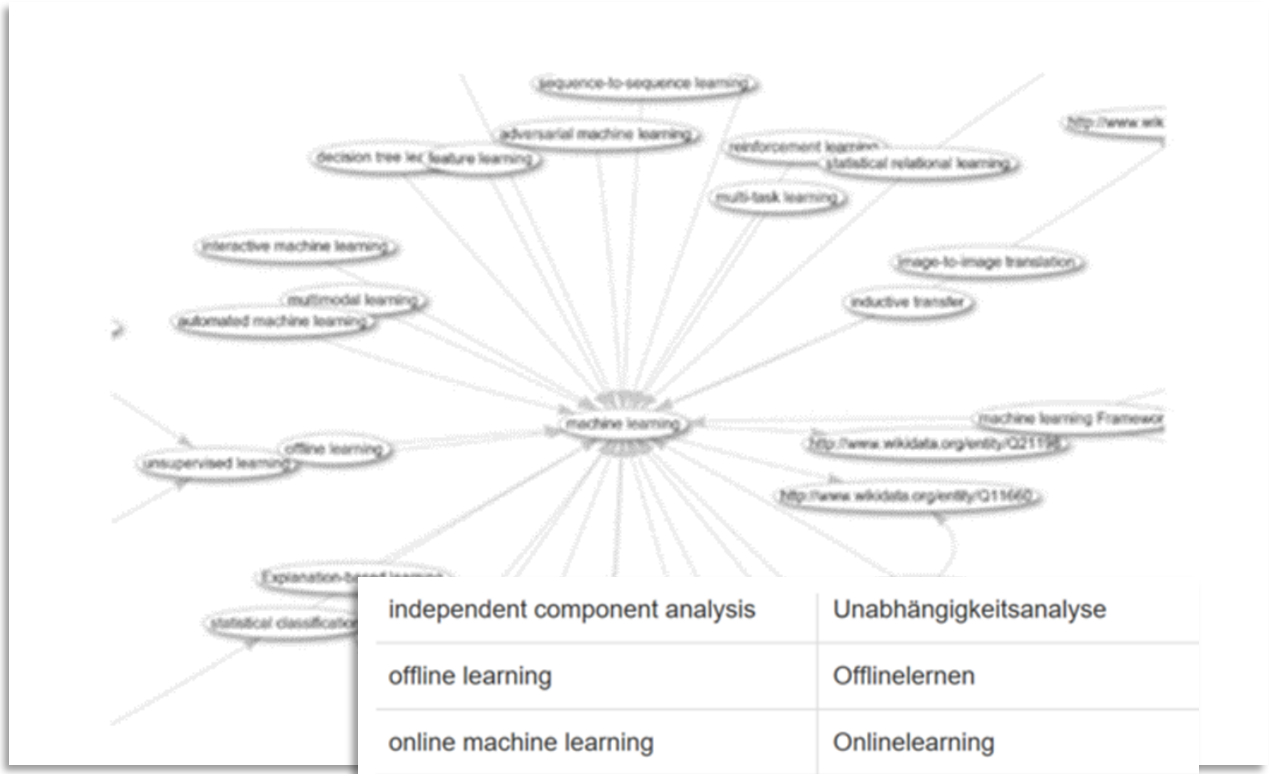
Source: https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation

Source: https://www.mediawiki.org/wiki/Extension:WikibaseLexeme/RDF_mapping

Microscopic View *Wikidata (1/2)*

Sample usage scenarios

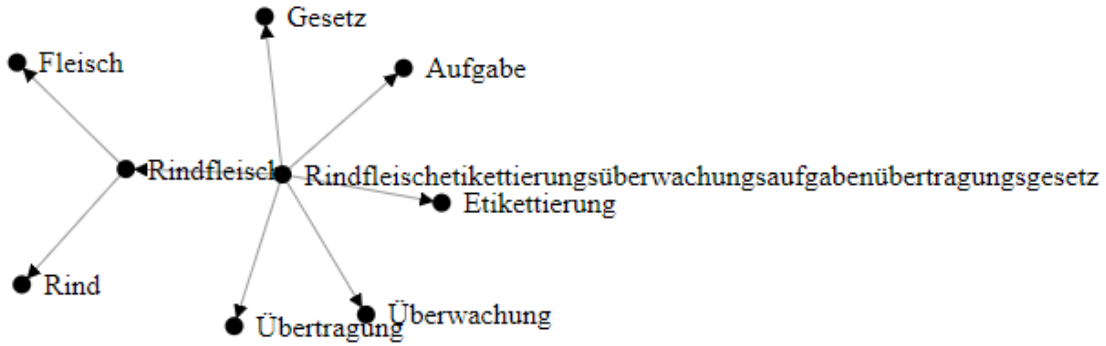
- Generate ontology and translation proposals for the English terms used in the ontology
- Show the parts of a compound word
- Visualize etymological information about a word

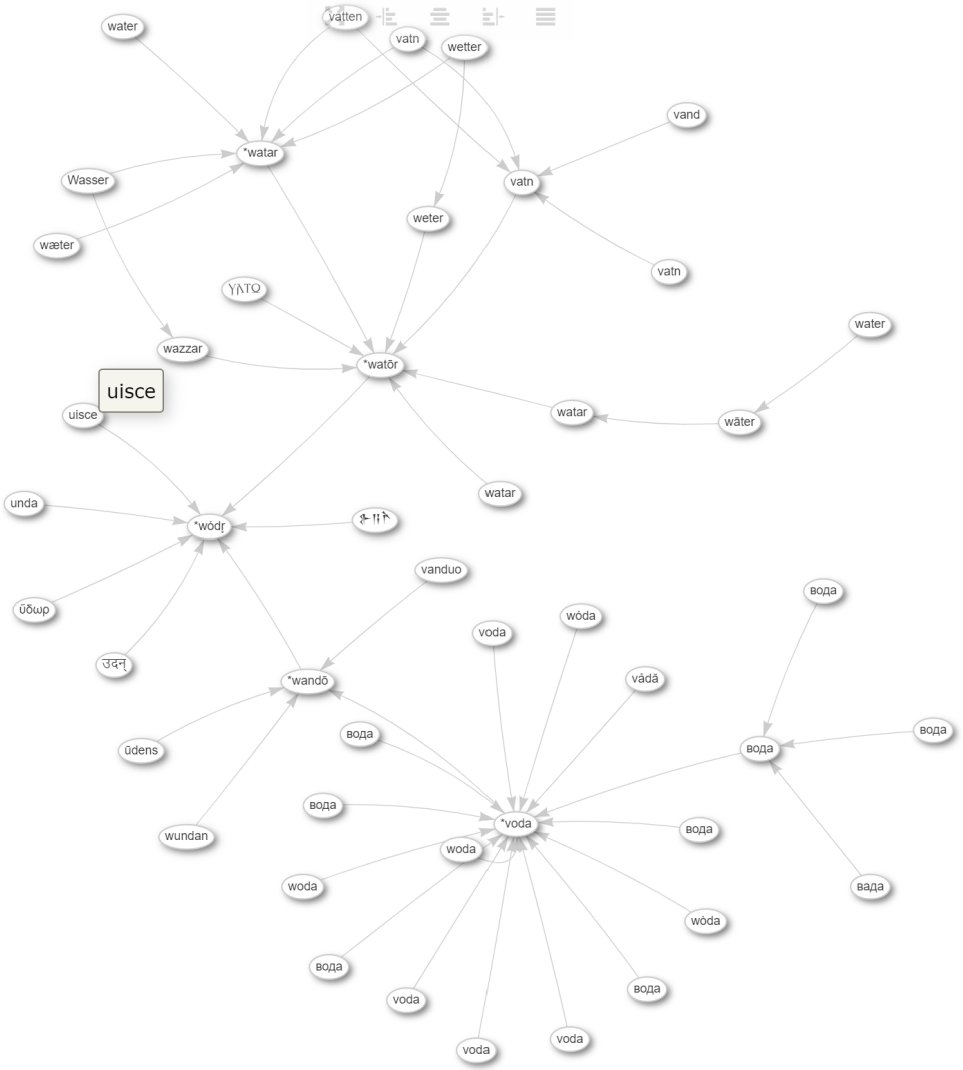


independent component analysis	Unabhängigkeitsanalyse
offline learning	Offlinelernen
online machine learning	Onlinelearning
pattern recognition	Mustererkennung
reinforcement learning	Bestärkendes Lernen
statistical classification	Klassifikationsverfahren
supervised learning	Überwachtes Lernen

http://www.tcworld.info/e-magazine/technical-communication/article/wikidata-at-work/news/education-training/

Microscopic View *Wikidata (2/2)*



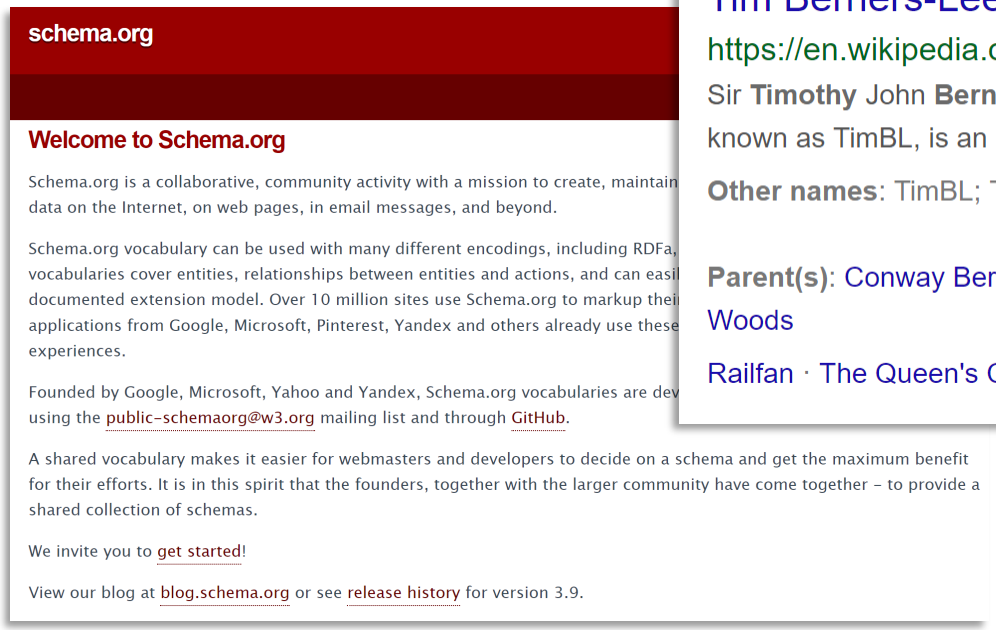


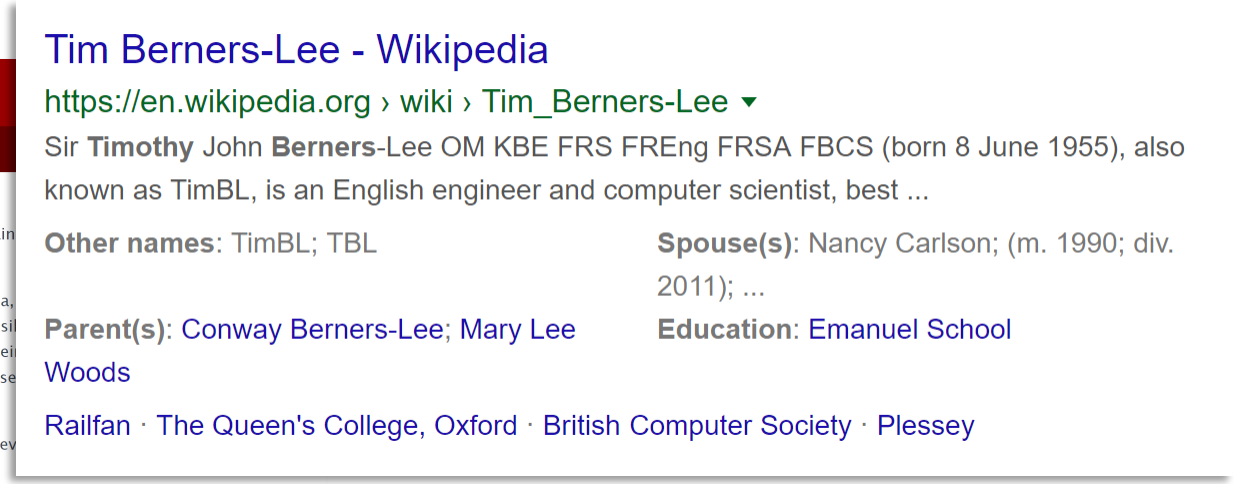
Bird's-Eye View *schema.org*

"Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond."

Source: https://schema.org/

Schema.org vocabulary allows search engines to recognize information easier and more precisely; it can power (rich) snippets





Source: https://www.google.com/search?hl=en&q=tim%20berners

Source: https://schema.org/

Close-Up *schema.org*

Vocabulary

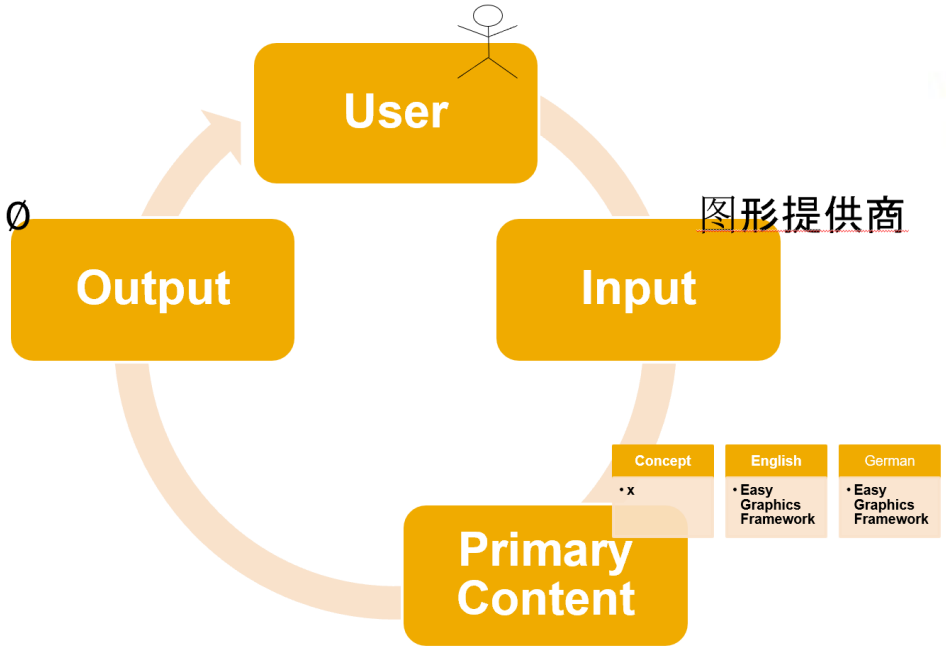
- Ontology of types (e.g. *Creative work* which subsumes *Book*, *Movie*, *MusicRecording*, *Recipe*, *TVSeries* ...)
- Properties (e.g. *award* for *Book*)

Different encodings

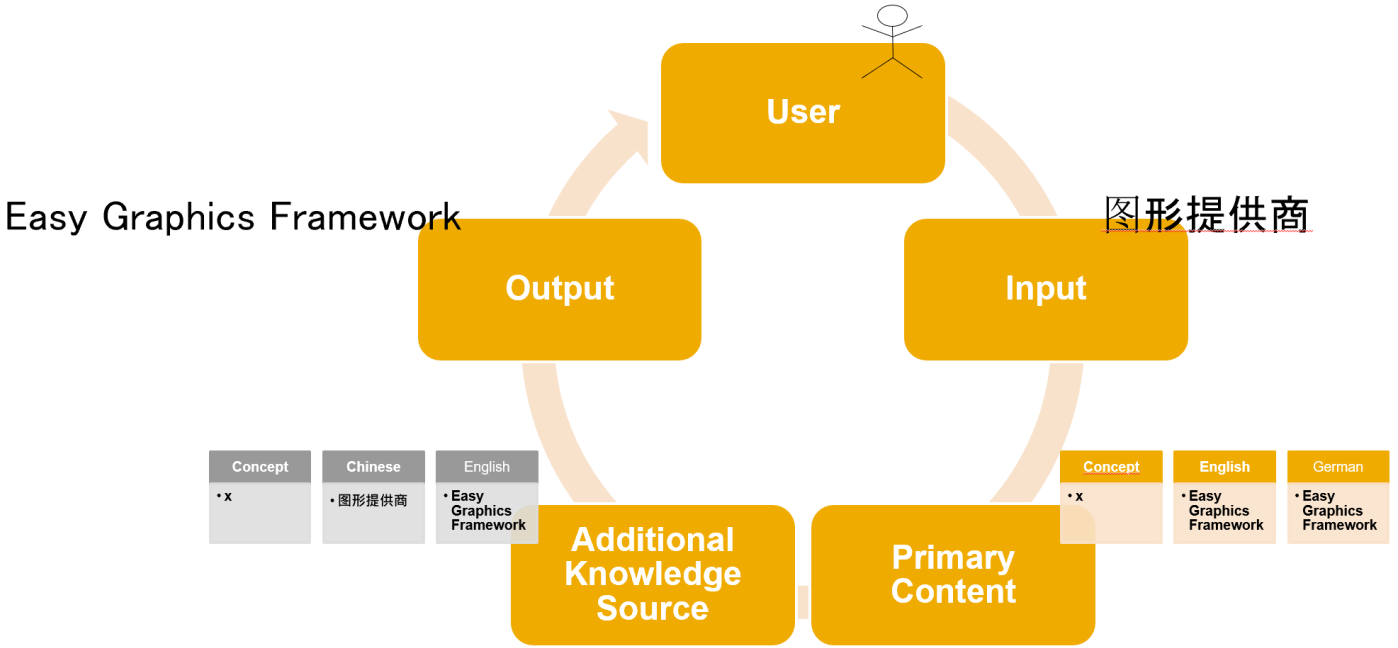
- RDFa
- Microdata
- JSON-LD
- …

```
<script type="application/ld+json">[{
"@context": "http://schema.org/",
"@id": "tid_db6_014D420D507ED411B1360060B03C6BFB-1",
"@type": "CreativeWork",
"inLanguage": "EN",
    "name": "Easy Graphics Framework"},
{
"@context": "http://schema.org/",
"@id": "tid_db6_014D420D507ED411B1360060B03C6BFB-2",
"sameAs" : "tid_db6_014D420D507ED411B1360060B03C6BFB-1",
"@type": "CreativeWork",
"inLanguage": "ZH",
    "name": "图形提供商"}]</script>
```

Microscopic View *schema.org*



图形提供商
www.sap.com/
Known in SAP terminology (English equivalent: Easy Graphics Framework)
A rich snippet could display additional content of your choice (e.g. the definition of a term, domain/ontological information, etc.) here.



http://www.tcworld.info/e-magazine/content-strategies/article/linked-data-and-schemaorg-crossing-the-language-chasm-with-terminological-assets/

Conclusions

Linguistic Linked Open Data (LLOD), Wikidata, and schema.org enable or improve knowledge solutions

Well designed vocabularies/ models

Large entity bases

Powerful tools/ technical services

Communities interested in consensus building, and standardization

Follow us

www.sap.com/contactsap

THE BEST RUN SAP®