Presentation for the

MULTI-TASK
MULTI-LINGUAL
MULTI-MODAL

plenary meeting, Lisbon, 2021-10-06

The COST Action “European network for Web-centred linguistic data science”

EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Nexus Linguarum

CA18209 -European network for Web-centred linguistic data science

Jorge Gracia (University of Zaragoza)
Thierry Declerck, (DFKI GmbH, Saarland Informatics Campus)

06/10/2021

NexusLinguarum participants (as of Summer 2021)

208 researchers from 42 countries

36/38 COST Members

Albania, Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, the Republic of Moldova, Montenegro, The Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, United Kingdom.

1 Cooperating Member

Israel

1 Specific organisation

Translation Centre for the Bodies of the European Union (Luxembourg)

3 NNC and 2 IPC

Georgia, Belarus, Kosovo (UNSCR 1244/1999), USA, Singapore

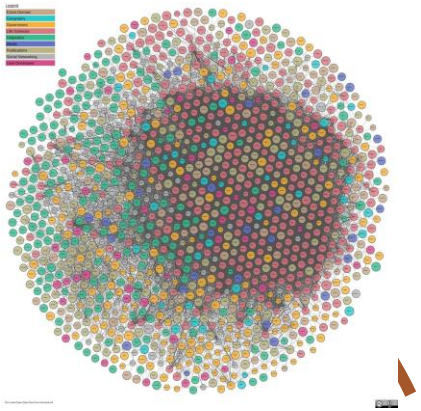
- MC participants: 134
- WG participants (not in MC): 74
- TOTAL NexusLinguarum participants: **208**

NexusLinguarum - Main Challenge

“Promote synergies across Europe between linguists, computer scientists, terminologists, language professionals, and other stakeholders in industry and society, in order to investigate and extend the area of ***linguistic data science****, through the construction of an* ***ecosystem*** *of* ***multilingual*** *and* ***semantically interoperable linguistic data*** *at Web scale.”*

NexusLinguarum - concept

Language resources



Linked data

Nexus Linguarum

Data analysis

Natural Language Processing

Some key aspects

- Linked Data as a core technology
- Multilinguality
- Low-resource and minority languages
- Establishing a network of experts
- Collaboratation with international fora, organisations and projects
- Working out a common curriculum to train a new generation of researchers in the area

What is Linguistic Data Science?

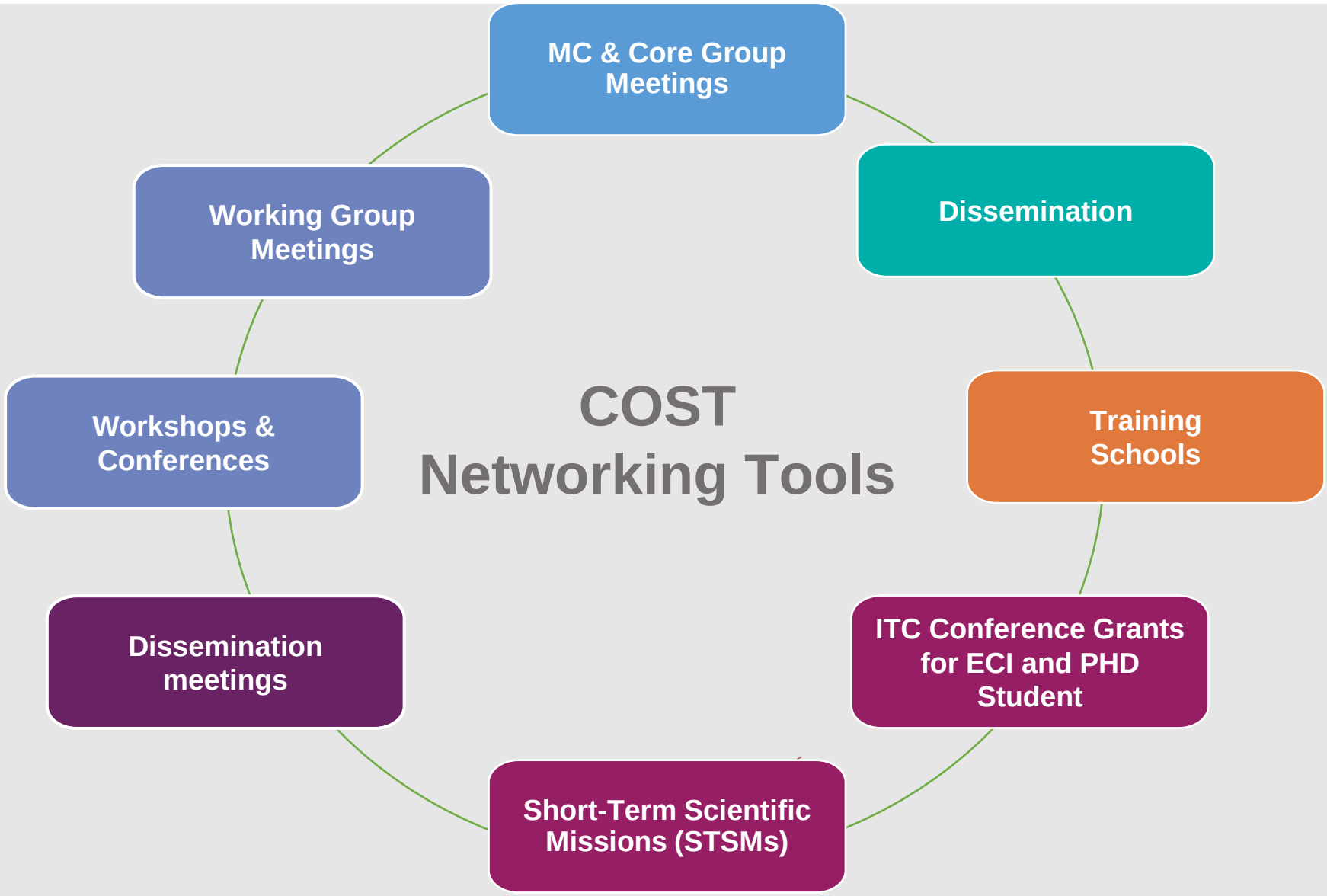
- Subfield of "data science" that provides a formal basis to the analysis, representation, integration and exploitation of **linguistic data** for
  - language analysis (e.g. syntax, morphology, terminology, etc.)
  - language applications (e.g. machine translation, speech recognition, sentiment analysis, etc.).

Working Groups of NexusLinguarum

- WG1 -- Linked data-based language resources
- WG2 -- Linked data -aware NLP services
- WG3 -- Support for linguistic data science
- WG4 -- Use cases and applications
- WG5 -- Management and dissemination

- We see the important role played by Linked Data (and thus resources like Dbpedia, Wikidata etc.) – cooperation with the LOD community central

What is funded by a COST Action?

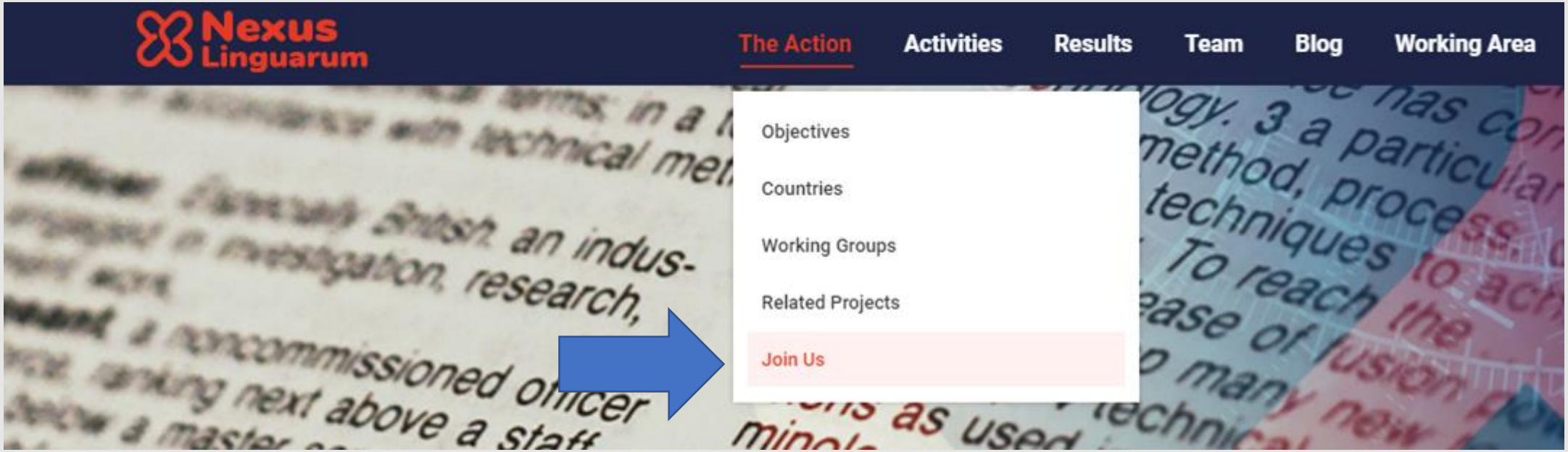


Some events coordinated by NexusLinguarum

- Eurolan 2021 Training School on Linked Data for Linguistics (February 2021, in cooperation with the Romanian Academy, the Research Institute for Artificial Intelligence in Bucharest and the Institute of Computer Science in Iași, as well as the "Alexandru Ioan Cuza" University of Iași, Romania
- 3rd Conference on Language, Data and Knowledge (LDK 2021), 1-4 September in Zaragoza
- Workshop on Deep Learning and Neural Approaches for Linguistic Data, September 30 in Skopje
- A one week lecturing session within the Lisbon Summer School in Linguistics (5-9 July 2021) : Introduction to Linked Open Data in Linguistics
- A tutorial on LLOD – Linguistic Linked Open Data at TALN-RECITAL 2021, Lille (28.06.21)
- …

How to join NexusLinguarum?

https://nexuslinguarum.eu/



Related projects

Prêt-à-LLOD

elexis european lexicographic infrastructure

ISA²

LiODi Linked Open Dictionaries

The World Wide Web

- Focuses on *documents* (web of hypertext, written in HTML)
- Links are established between those documents
- Humans can extract and interpret the meaning of the content in those documents...but this is not so easy for machines

Introduction to Linked Data (Julia Bosque-Gil & Thierry Declerck) -- 2

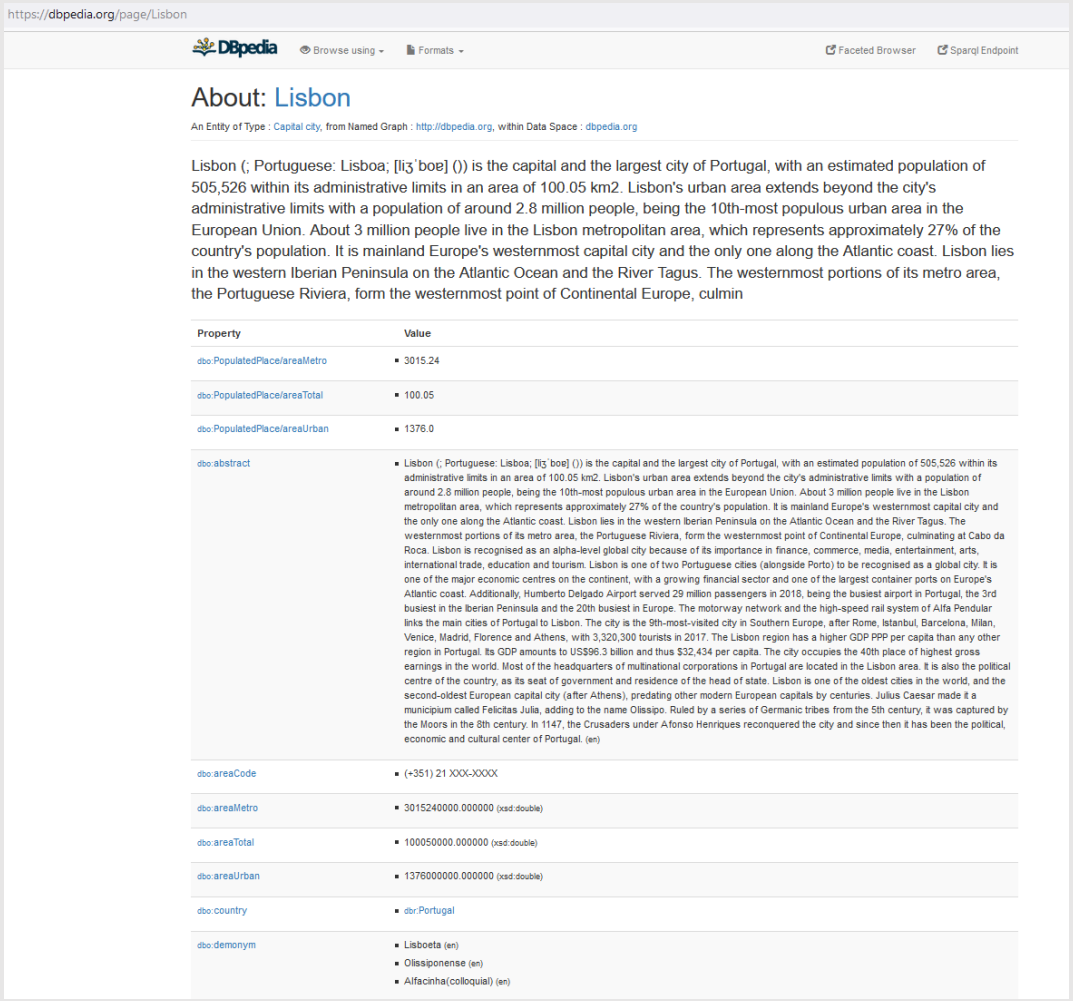
The World Wide Web

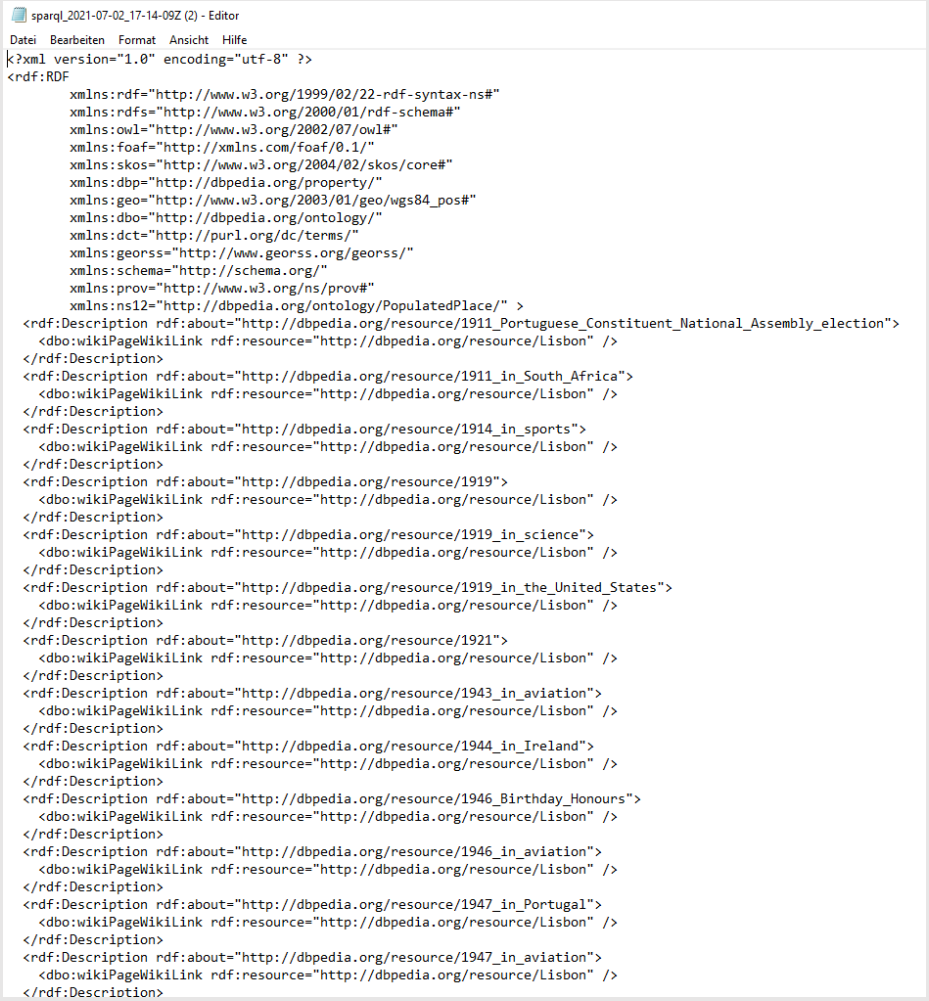
The Web of Data

- Documents
  - → Focuses on data, which are described in RDF
- Links are established between those documents
  - → Links are established between the data: Easier sharing and discovery
- Humans can extract and interpret the meaning of the content in those documents...but this is not so easy for machines
  - → Meaning processable by machines

Introduction to Linked Data (Julia Bosque-Gil & Thierry Declerck) -- 3

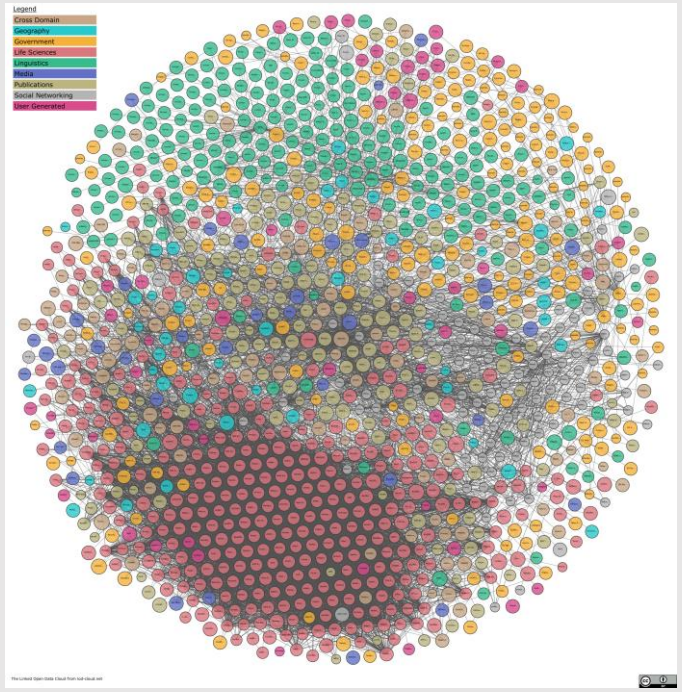
The web of documents vs the web of data: An example from DBpedia. The request for "https://dbpedia.org/**resource**/Lisbon" is leading by default to "https://dbpedia.org/**page**/Lisbon" (human readable), but with a specific request one can be directed to https://dbpedia.org/**data**/Lisbon (machine readable)

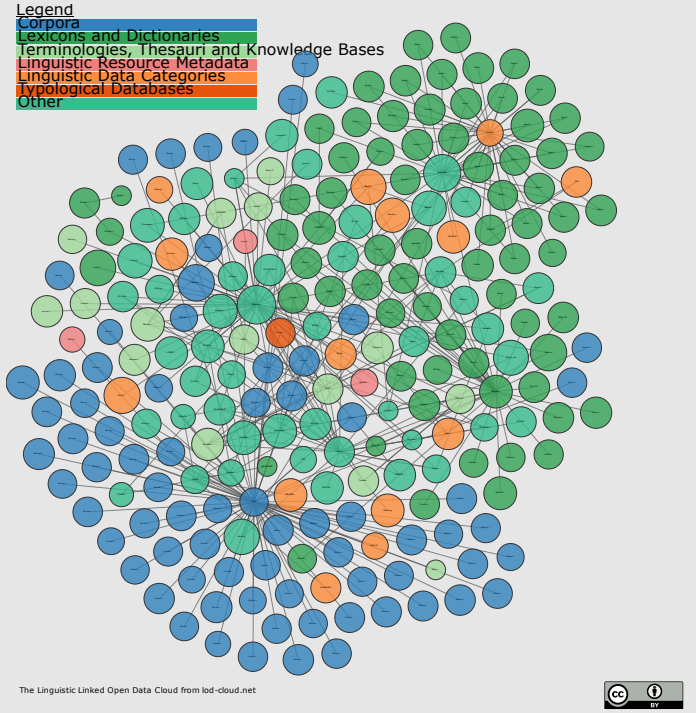




Introduction to Linked Data (Julia Bosque-Gil & Thierry Declerck) -- 4

The web of data is best represented by the Linked Data cloud, while a subset of this cloud is built by the Linguistic Linked Data cloud



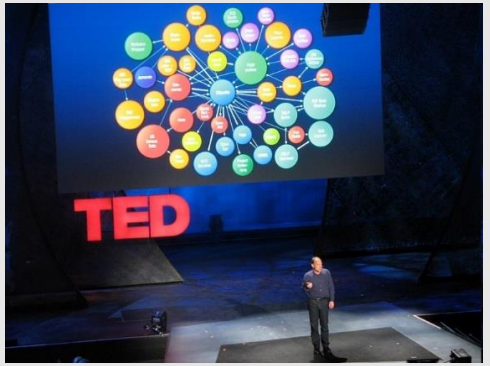


What are the principles behind those clouds, how do we represent the data?

Introduction to Linked Data (Julia Bosque-Gil & Thierry Declerck) -- 5

1. *Use **URIs** as names for things*
2. *Use **HTTP URIs** so that people can look up those names*
3. *When someone looks up a URI, provide useful **information** (using the standards: RDF*, SPARQL)*
4. *Include **links** to other URIs, so that they can discover more things.*

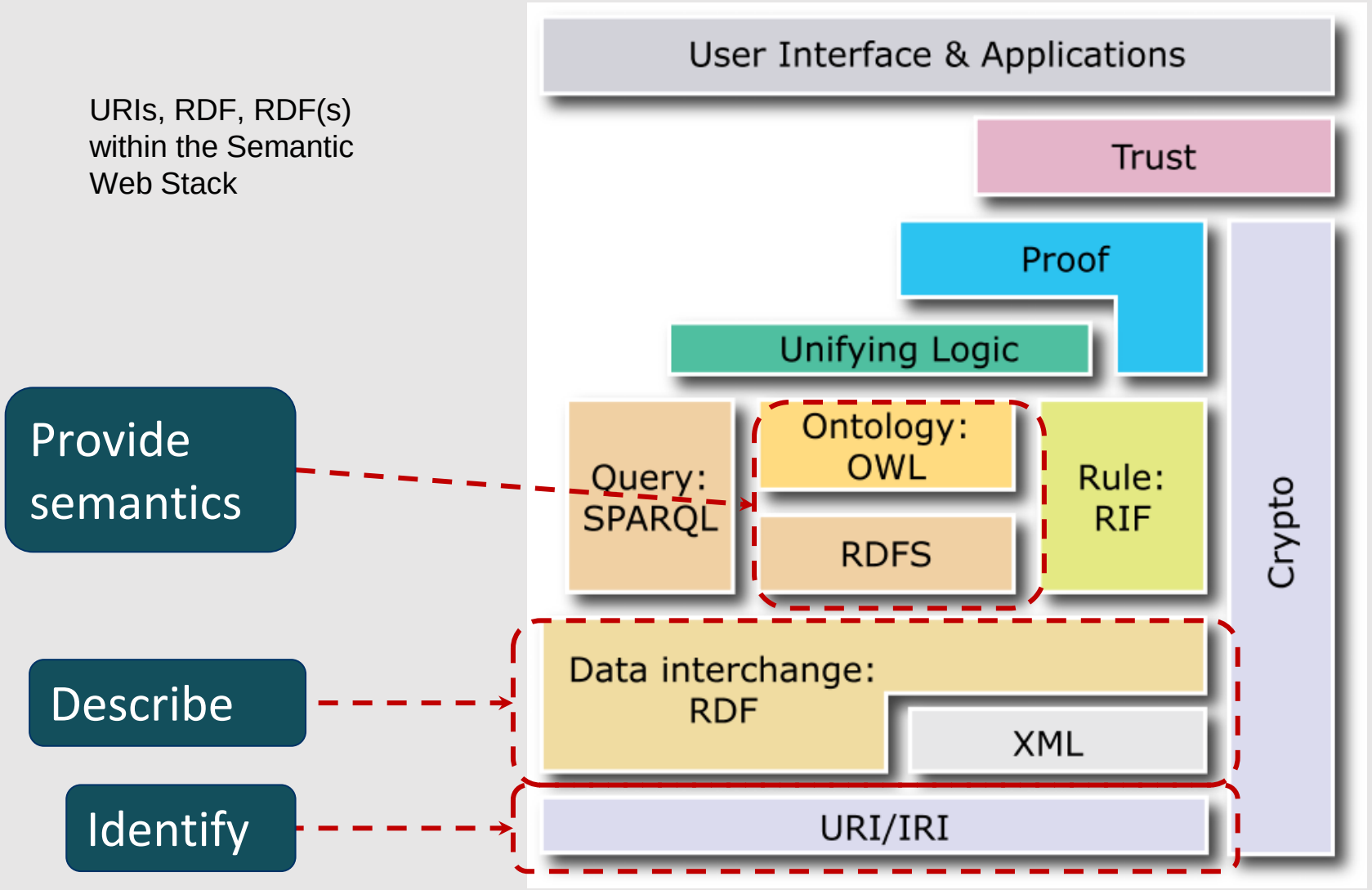
http://www.w3.org/DesignIssues/LinkedData.html



Consult the inspiring TED-2009 talk by Tim Berners Lee
https://www.ted.com/talks/tim_berners_lee_the_next_web?language=e

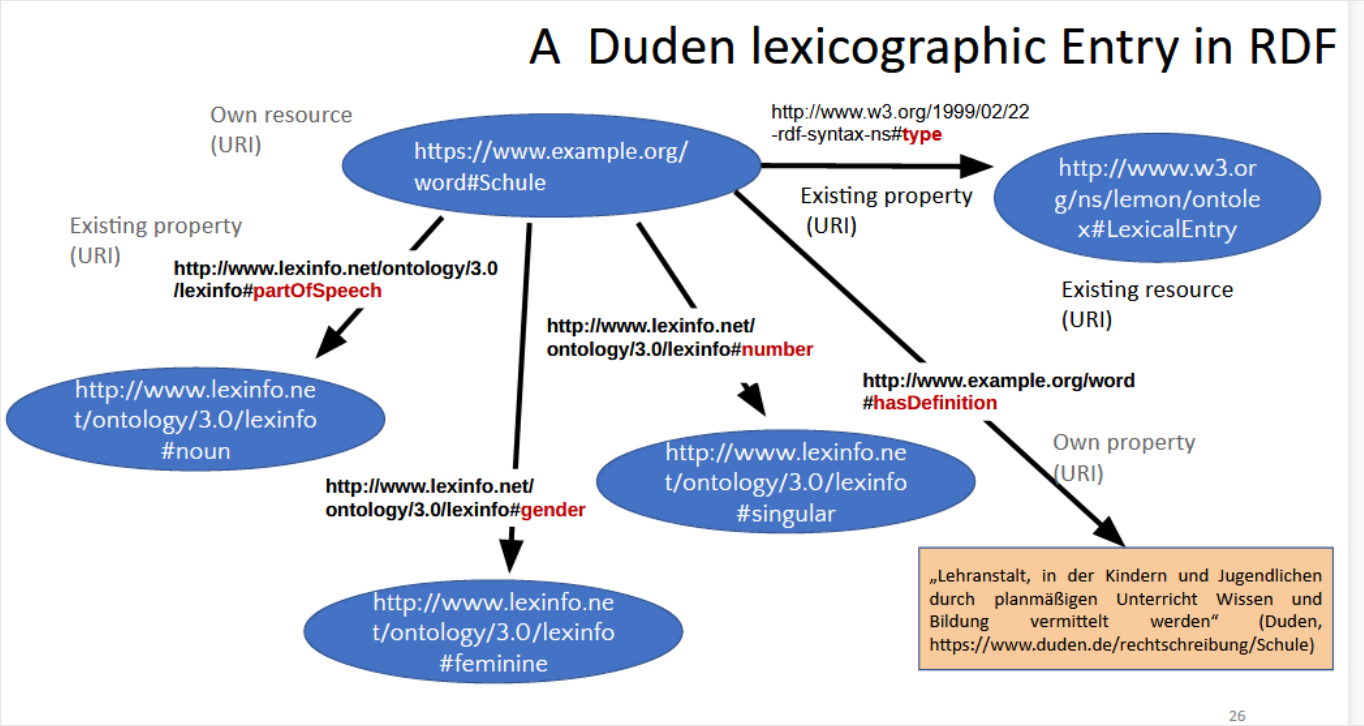
Introduction to Linked Data (Julia Bosque-Gil & Thierry Declerck) -- 6

URIs, RDF, RDF(s) within the Semantic Web Stack



'A visual representation of the semantic web's structure, often referred to as "layer cake"', taken from: http://webservices.itcs.umich.edu/mediawiki/DigitalRhetoricCollaborative/index.php/Image:Semantic_Web_Stack.png

We represent Language Data using Elements of the Semantic Web Stack: A lexicographic example taken from DUDEN



- **Subclasses** of an ontolex:LexicalEntry (owl:Class will be explained tomorrow!)

```
ontolex:Word a rdfs:Class;
        rdfs:subClassOf ontolex:LexicalEntry .
ontolex:Phrase  a rdfs:Class;
        rdfs:subClassOf ontolex:LexicalEntry .
```

- “Schule” as an **instance** of the “Word” class (and by inference, also an instance of the LexicalEntry class)

- From the LexInfo vocabulary we know that **lexinfo:number** is a sub-property of **lexinfo:morphosyntacticProperty**

Models and Vocabularies used in our Representation

- We make use of the OntoLex-Lemon model, which aims "to provide rich linguistic grounding for ontologies. Rich linguistic grounding includes the representation of morphological and syntactic properties of lexical entries as well as the syntax-semantics interface, i.e. the meaning of these lexical entries with respect to an ontology or vocabulary." (https://www.w3.org/2016/05/ontolex/)
- We also make strong use of the Lexinfo vocabulary, which is "data category ontology for OntoLex-Lemon, which provides description of lexicographic resources in RDF relative to ontologies" (https://github.com/ontolex/lexinfo)
- Both are presented in a separated set of slides (by John McCrae), after presenting the W3C activities relevant for NexusLinguarum

W3C Activities relevant for/involving NexusLinguarum

- Ontology Lexica (Ontolex) W3C Community Group (https://www.w3.org/community/ontolex/)
  - The OntoLex-Lemon model (https://www.w3.org/2016/05/ontolex/)
  - Extensions to OntoLex-Lemon
    - The OntoLex Lemon Lexicography Module (Lexicog) – published (https://www.w3.org/2019/09/lexicog/)
    - The OntoLex-Lemon Morphology Module (Morphology) – advanced (https://www.w3.org/community/ontolex/wiki/Morphology)
    - The OntoLex-Lemon *for frequency, attestation and corpus information* (FRaC) – in discussion (https://acoli-repo.github.io/ontolex-frac/)
    - The OntoLex-Lemon Terminology Module (TermLex, suggested name) – at a proposal stage (https://www.w3.org/community/ontolex/wiki/Terminology)
    - Current discussions on how and where to integrate the representation of Sign Languages (lexicons) and Multimodality in general.
- Linked Data for Language Technologies (LD4LT) W3C Community Group (https://www.w3.org/community/ld4lt/)
  - With a recent workshop on linguistic annotation on the Web: https://www.w3.org/community/ld4lt/wiki/LD4LT_Annotaton_Workshop_Zaragoza_2021

Details for OntoLex-Lemon

- The OntoLex-Lemon model (https://www.w3.org/2016/05/ontolex/) is subdivided in 6 modules (to be presented in detail in a separated slide set, by John McCrae)
  - The core module: Ontology-lexicon interface (ontolex) -- https://www.w3.org/2016/05/ontolex/#core
  - The Syntax and Semantics (synsem) module -- https://www.w3.org/2016/05/ontolex/#syntax-and-semantics-synsem
  - The Decomposition (decomp) module -- https://www.w3.org/2016/05/ontolex/#decomposition-decomp
  - The Variation & Translation (vartrans) module -- https://www.w3.org/2016/05/ontolex/#variation-translation-vartrans
  - The Metadata (lime) -- https://www.w3.org/2016/05/ontolex/#metadata-lime

Language Data related Ontologies and Vocabularies relevant for NexusLinguarum

- Lexinfo: data category ontology for OntoLex-Lemon (https://github.com/ontolex/lexinfo)
- OLIA: "The Ontologies of Linguistic Annotation (OLiA) are a repository of linguistic data categories used for
  - corpus annotation,
  - Natural Language Processing (NLP) tools,
  - machine-readable dictionaries,
  - and other linguistic resources" (http://acoli.cs.uni-frankfurt.de/resources/olia/)
- SKOS (*Simple Knowledge Organization System*): "SKOS is an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading lists and taxonomies within the framework of the Semantic Web" (https://www.w3.org/2004/02/skos/)
- …..

Thanks for your attention!

- Interested in joining the discussions within the Ontolex W3C community group?
  https://www.w3.org/community/ontolex/