

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”



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.”*

Some key aspects

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

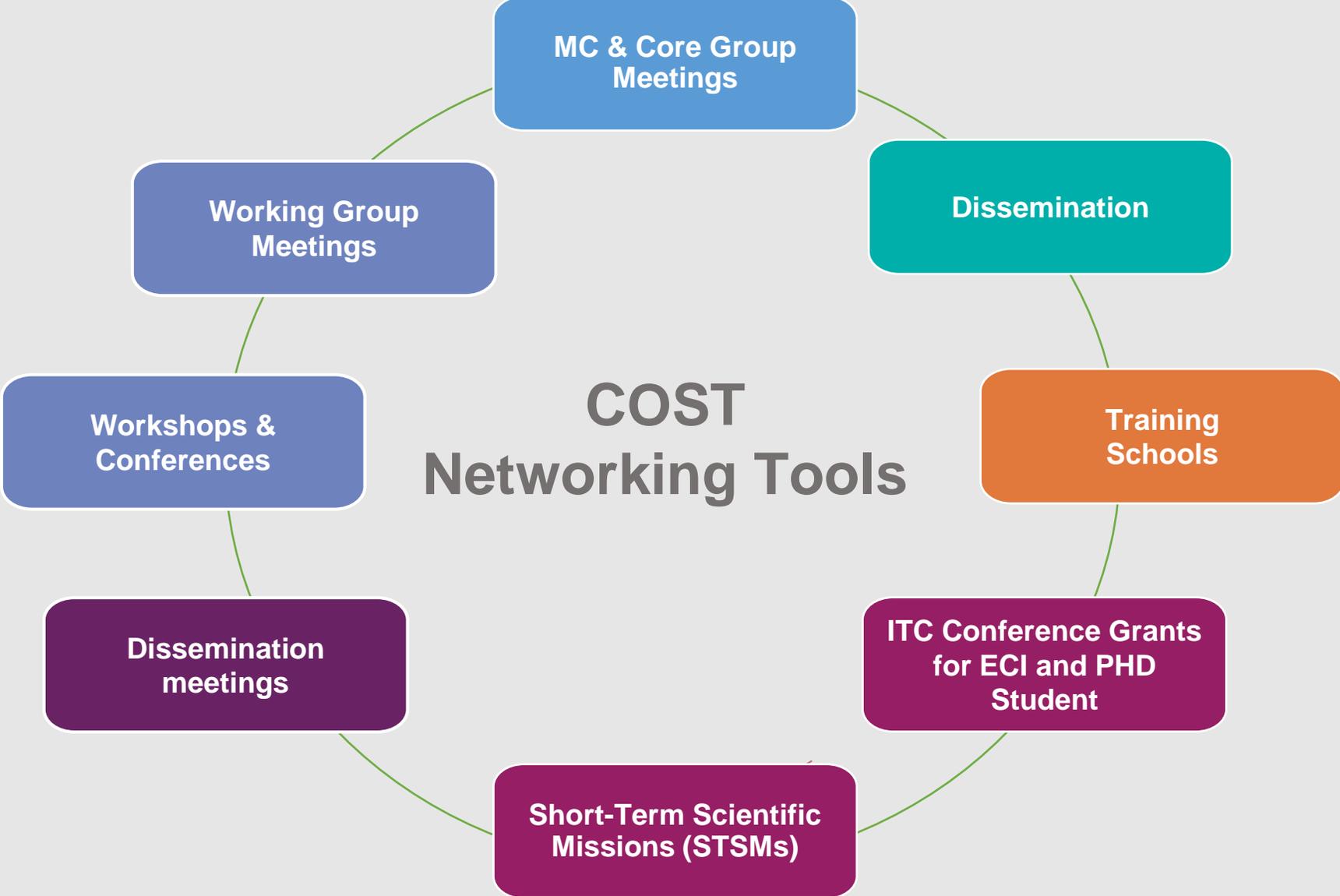
What is Linguistic Data Science?

- 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?

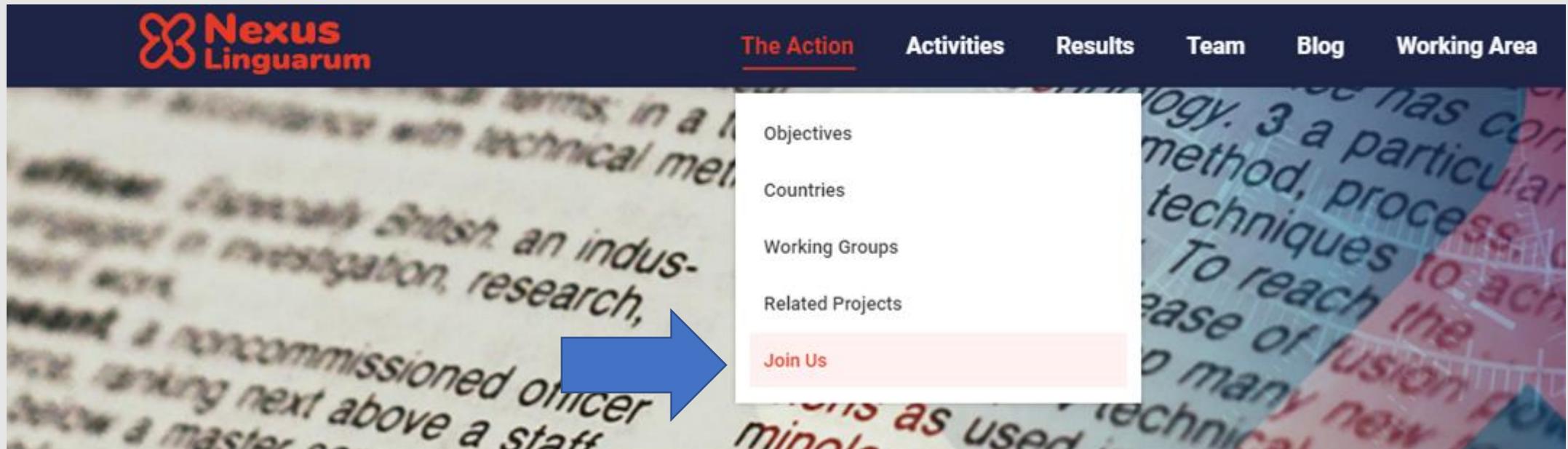


Some events coordinated by NexusLinguarum

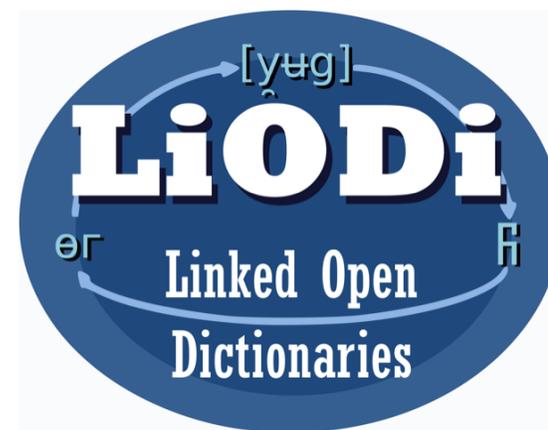
- Eurolan 2021 Training School on Linked Data for Linguistics (February 2021, in cooperation with the Romanian Academy, the Research Institute for Artificial Intelligence in Bucharest and the Institute of Computer Science in Iași, as well as the “Alexandru Ioan Cuza” University of Iași, Romania)
- 3rd 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



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

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)

<https://dbpedia.org/page/Lisbon>

DBpedia Browse using - Formats - Faceted Browser Sparql Endpoint

About: Lisbon

An Entry of Type : Capital city, from Named Graph : <http://dbpedia.org>, within Data Space : dbpedia.org

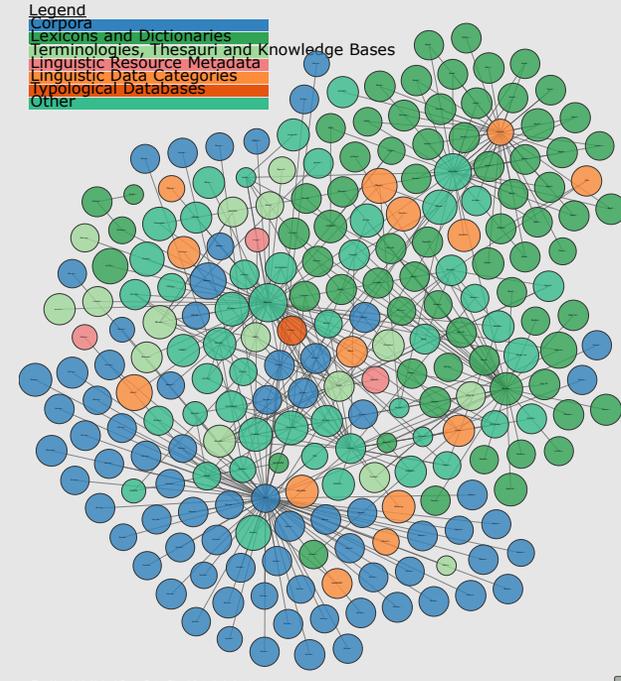
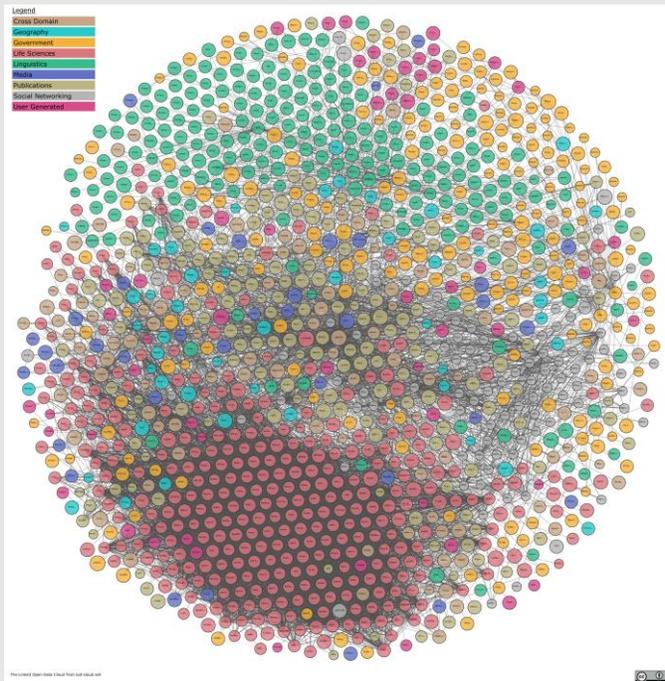
Lisbon (, Portuguese: Lisboa; [liʒˈboɐ] ()) is the capital and the largest city of Portugal, with an estimated population of 505,526 within its administrative limits in an area of 100.05 km2. Lisbon's urban area extends beyond the city's administrative limits with a population of around 2.8 million people, being the 10th-most populous urban area in the European Union. About 3 million people live in the Lisbon metropolitan area, which represents approximately 27% of the country's population. It is mainland Europe's westernmost capital city and the only one along the Atlantic coast. Lisbon lies in the western Iberian Peninsula on the Atlantic Ocean and the River Tagus. The westernmost portions of its metro area, the Portuguese Riviera, form the westernmost point of Continental Europe, culminating at Cabo da Roca. Lisbon is recognised as an alpha-level global city because of its importance in finance, commerce, media, entertainment, arts, international trade, education and tourism. Lisbon is one of two Portuguese cities (alongside Porto) to be recognised as a global city. It is one of the major economic centres on the continent, with a growing financial sector and one of the largest container ports on Europe's Atlantic coast. Additionally, Humberto Delgado Airport served 29 million passengers in 2018, being the busiest airport in Africa. Pendular links the main cities of Portugal to Lisbon. The city is the 9th-most-visited city in Southern Europe, after Rome, Istanbul, Barcelona, Milan, Venice, Madrid, Florence and Athens, with 3,320,300 tourists in 2017. The Lisbon region has a higher GDP PPP per capita than any other region in Portugal. Its GDP amounts to US\$96.3 billion and thus \$32,434 per capita. The city occupies the 40th place of highest gross earnings in the world. Most of the headquarters of multinational corporations in Portugal are located in the Lisbon area. It is also the political centre of the country, as its seat of government and residence of the head of state. Lisbon is one of the oldest cities in the world, and the second-oldest European capital city (after Athens), predating other modern European capitals by centuries. Julius Caesar made it a municipium called Felicitas Julia, adding to the name Olisipo. Ruled by a series of Germanic tribes from the 5th century, it was captured by the Moors in the 8th century. In 1147, the Crusaders under Afonso Henriques reconquered the city and since then it has been the political, economic and cultural center of Portugal. (en)

Property	Value
dbo:PopulatedPlace/areaMetro	• 3015.24
dbo:PopulatedPlace/areaTotal	• 100.05
dbo:PopulatedPlace/areaUrban	• 1376.0
dbo:abstract	• Lisbon (, Portuguese: Lisboa; [liʒˈboɐ] ()) is the capital and the largest city of Portugal, with an estimated population of 505,526 within its administrative limits in an area of 100.05 km2. Lisbon's urban area extends beyond the city's administrative limits with a population of around 2.8 million people, being the 10th-most populous urban area in the European Union. About 3 million people live in the Lisbon metropolitan area, which represents approximately 27% of the country's population. It is mainland Europe's westernmost capital city and the only one along the Atlantic coast. Lisbon lies in the western Iberian Peninsula on the Atlantic Ocean and the River Tagus. The westernmost portions of its metro area, the Portuguese Riviera, form the westernmost point of Continental Europe, culminating at Cabo da Roca. Lisbon is recognised as an alpha-level global city because of its importance in finance, commerce, media, entertainment, arts, international trade, education and tourism. Lisbon is one of two Portuguese cities (alongside Porto) to be recognised as a global city. It is one of the major economic centres on the continent, with a growing financial sector and one of the largest container ports on Europe's Atlantic coast. Additionally, Humberto Delgado Airport served 29 million passengers in 2018, being the busiest airport in Africa. Pendular links the main cities of Portugal to Lisbon. The city is the 9th-most-visited city in Southern Europe, after Rome, Istanbul, Barcelona, Milan, Venice, Madrid, Florence and Athens, with 3,320,300 tourists in 2017. The Lisbon region has a higher GDP PPP per capita than any other region in Portugal. Its GDP amounts to US\$96.3 billion and thus \$32,434 per capita. The city occupies the 40th place of highest gross earnings in the world. Most of the headquarters of multinational corporations in Portugal are located in the Lisbon area. It is also the political centre of the country, as its seat of government and residence of the head of state. Lisbon is one of the oldest cities in the world, and the second-oldest European capital city (after Athens), predating other modern European capitals by centuries. Julius Caesar made it a municipium called Felicitas Julia, adding to the name Olisipo. Ruled by a series of Germanic tribes from the 5th century, it was captured by the Moors in the 8th century. In 1147, the Crusaders under Afonso Henriques reconquered the city and since then it has been the political, economic and cultural center of Portugal. (en)
dbo:areaCode	• (+351) 21 XXX-XXXX
dbo:areaMetro	• 3015240000.000000 (xsd:double)
dbo:areaTotal	• 100050000.000000 (xsd:double)
dbo:areaUrban	• 1376000000.000000 (xsd:double)
dbo:country	• dbr:Portugal
dbo:demonym	• Lisboaeta (en) • Olisiponense (en) • Alfacinha (colloquial) (en)

```
sparql_2021-07-02_17-14-09Z (2) - Editor
Datei Bearbeiten Format Ansicht Hilfe
<?xml version="1.0" encoding="utf-8" ?>
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:foaf="http://xmlns.com/foaf/0.1/"
  xmlns:skos="http://www.w3.org/2004/02/skos/core#"
  xmlns:dbp="http://dbpedia.org/property/"
  xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#"
  xmlns:dbo="http://dbpedia.org/ontology/"
  xmlns:dct="http://purl.org/dc/terms/"
  xmlns:georss="http://www.georss.org/georss/"
  xmlns:schema="http://schema.org/"
  xmlns:prov="http://www.w3.org/ns/prov#"
  xmlns:ns12="http://dbpedia.org/ontology/PopulatedPlace/" >
  <rdf:Description rdf:about="http://dbpedia.org/resource/1911_Portuguese_Constituent_National_Assembly_election">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1911_in_South_Africa">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1914_in_sports">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1919">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1919_in_science">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1919_in_the_United_States">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1921">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1943_in_aviation">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1944_in_Ireland">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1946_Birthday_Honours">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1946_in_aviation">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1947_in_Portugal">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  <rdf:Description rdf:about="http://dbpedia.org/resource/1947_in_aviation">
    <dbo:wikiPageWikiLink rdf:resource="http://dbpedia.org/resource/Lisbon" />
  </rdf:Description>
  </rdf:Description>
```

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.

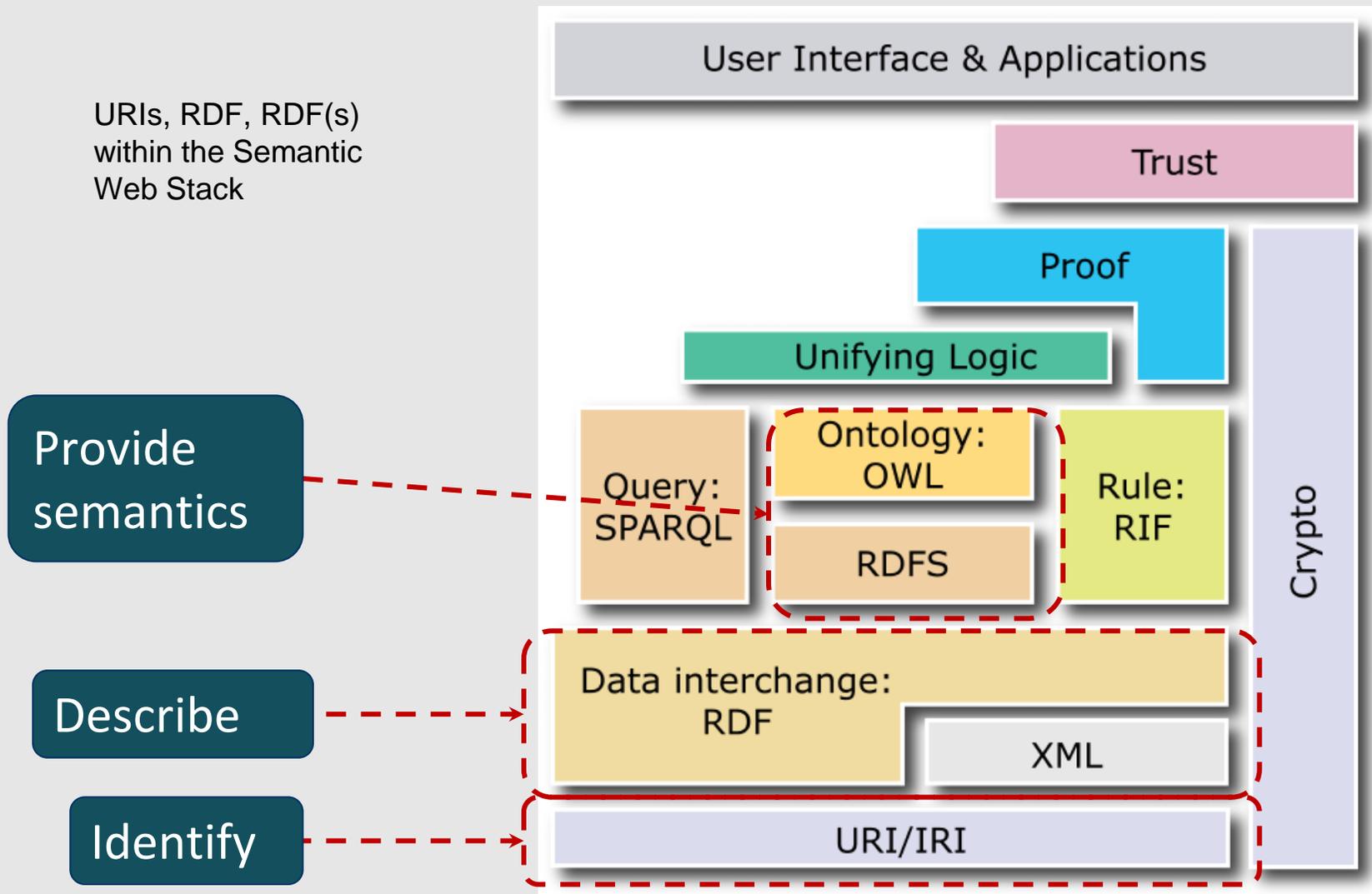


<http://www.w3.org/DesignIssues/LinkedData.html>



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

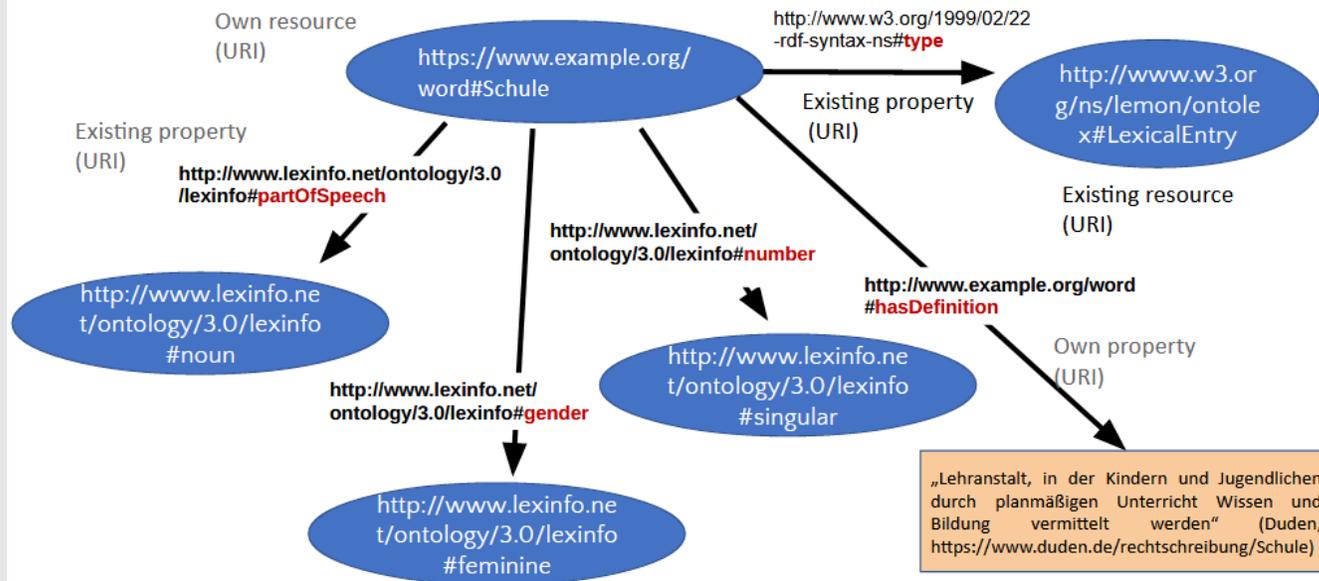
Introduction to Linked Data (Julia Bosque-Gil & Thierry Declerck) -- 6



'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

A Duden lexicographic Entry in RDF



- **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

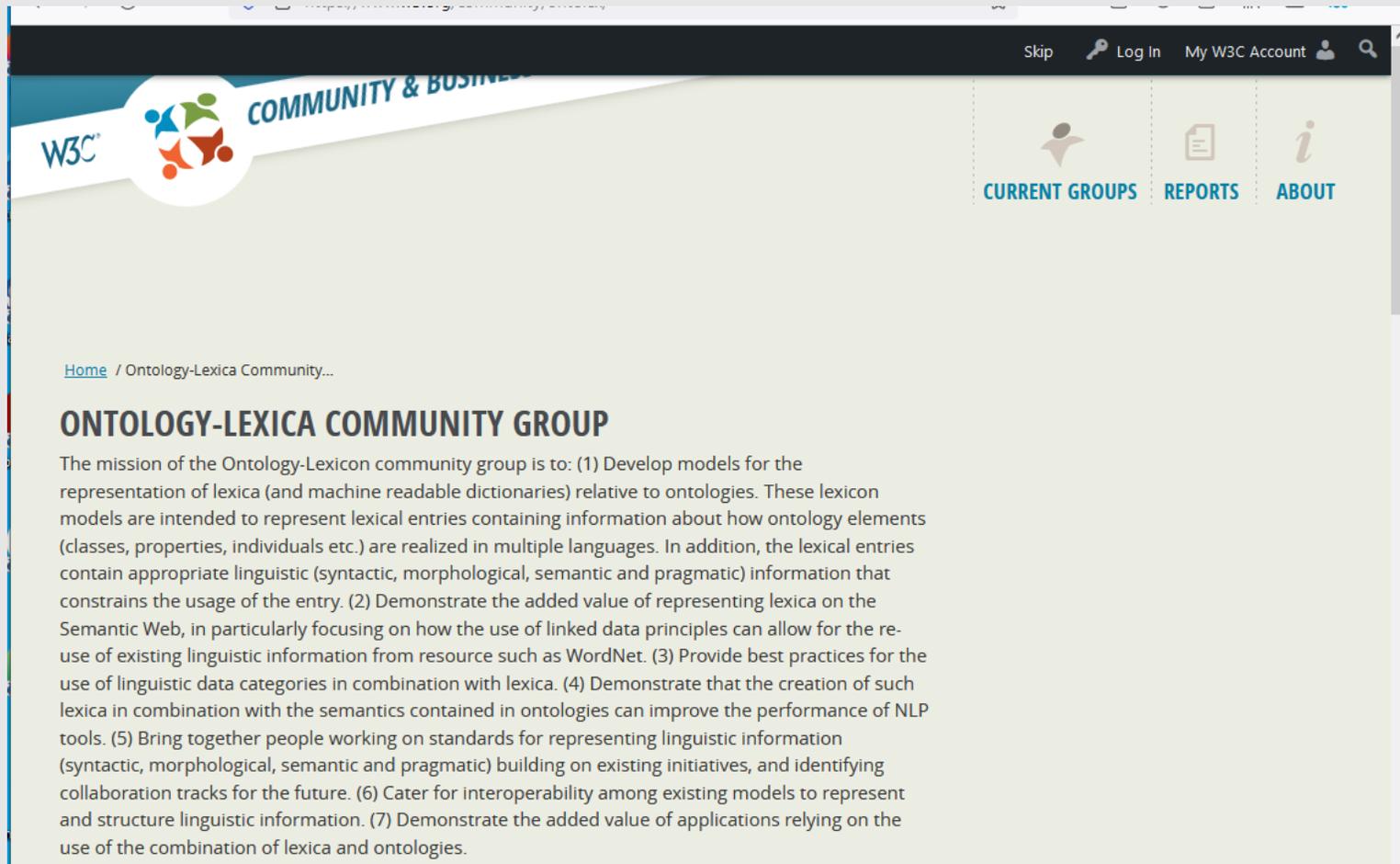
- The OntoLex-Lemon model (<https://www.w3.org/2016/05/ontolex/>) is subdivided in 6 modules (to be presented in detail in a separated slide set, by John McCrae)
 - The core module: Ontology-lexicon interface (ontolex) -- <https://www.w3.org/2016/05/ontolex/#core>
 - The Syntax and Semantics (synsem) 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/>



The screenshot shows the W3C website for the Ontology-Lexica Community Group. The page features a header with the W3C logo and a navigation menu with links for 'CURRENT GROUPS', 'REPORTS', and 'ABOUT'. The main content area includes a breadcrumb trail 'Home / Ontology-Lexica Community...' and a section titled 'ONTOLOGY-LEXICA COMMUNITY GROUP'. The text describes the group's mission, which is to develop models for the representation of lexica relative to ontologies, demonstrate the value of representing lexica on the Semantic Web, provide best practices for the use of linguistic data categories, demonstrate that the creation of such lexica can improve NLP tools, bring together people working on standards for representing linguistic information, cater for interoperability among existing models, and demonstrate the value of applications relying on the use of the combination of lexica and ontologies.

W3C COMMUNITY & BUSINESS

Skip Log In My W3C Account

CURRENT GROUPS REPORTS ABOUT

[Home](#) / Ontology-Lexica Community...

ONTOLOGY-LEXICA COMMUNITY GROUP

The mission of the Ontology-Lexicon community group is to: (1) Develop models for the representation of lexica (and machine readable dictionaries) relative to ontologies. These lexicon models are intended to represent lexical entries containing information about how ontology elements (classes, properties, individuals etc.) are realized in multiple languages. In addition, the lexical entries contain appropriate linguistic (syntactic, morphological, semantic and pragmatic) information that constrains the usage of the entry. (2) Demonstrate the added value of representing lexica on the Semantic Web, in particularly focusing on how the use of linked data principles can allow for the re-use of existing linguistic information from resource such as WordNet. (3) Provide best practices for the use of linguistic data categories in combination with lexica. (4) Demonstrate that the creation of such lexica in combination with the semantics contained in ontologies can improve the performance of NLP tools. (5) Bring together people working on standards for representing linguistic information (syntactic, morphological, semantic and pragmatic) building on existing initiatives, and identifying collaboration tracks for the future. (6) Cater for interoperability among existing models to represent and structure linguistic information. (7) Demonstrate the added value of applications relying on the use of the combination of lexica and ontologies.