5TH WORKSHOP ON LINKED DATA IN LINGUISTICS (LDL-2016)
MANAGING, BUILDING AND USING LINKED LANGUAGE RESOURCES

09:00 Opening and Introduction
09:30 Keynote: Damir Cavar (Indiana University / LINGUIST List)
   10:30 Coffee Break
11:00 Session I: Dictionaries and Ontologies
   12:40 Lunch Break
14:00 Poster Session
14:40 Session II: Building Linked Language Resources
   16:00 Coffee Break
16:30 Session III: Using Linked Language Resources
17:40 Discussion and Closing

LREC 2016, May 24th, 2016, Portorož, Slovenia
A BRIEF INTRODUCTION TO
LINKED DATA IN LINGUISTICS
RESOURCES AND APPLICATIONS
LANGUAGE RESOURCES, 2010 AD

- used in natural language processing, scientific research, language documentation, ...
- accessibility challenges
  - different formats, schemes
  - distributed
  - dispersed metadata collections
  
  ⇒ require common specifications to represent, share, access and register language resources

Linguistic Linked Open Data (LLOD) & LLOD cloud
by the time, long been recognized as a problem

- accessibility challenges
  + different formats, schemes
  + distributed
  + dispersed metadata collections

hence, several proposals to address them

- Independently, different groups considered using RDF/OWL as a (local) solution, e.g., for
  + annotations (GOLD, ISOcat, OLiA)

(Farrar & Langendoen 2003, Ide & Wright 2004, Schmidt et al. 2006)
by the time, long been recognized as a problem

- accessibility challenges
  - different formats, schemes
  - distributed
  - dispersed metadata collections

- Independently, different groups considered using RDF/OWL as a (local) solution, e.g., for
  - NLP pipelines

by the time, long been recognized as a problem

- accessibility challenges
  - different formats, schemes
  - distributed
  - dispersed metadata collections

- Independently, different groups considered using RDF/OWL as a (local) solution, e.g., for
  - integrating typological data bases (TDS)

(Saulwick et al. 2005, Dimitriades et al. 2010)
by the time, long been recognized as a problem

accessibility challenges
+ different formats, schemes
+ distributed
+ dispersed metadata collections

 Independently, different groups considered using RDF/OWL as a (local) solution, e.g., for
+ modelling and querying multi-layer corpora
(Cassidy 2010, Chiarcos et al. 2008, Rehm et al. 2008)

hence, several proposals to address them
by the time, long been recognized as a problem

- accessibility challenges
  - different formats, schemes
  - distributed
  - dispersed metadata collections

hence, several proposals to address them

Independently, different groups considered using RDF/OWL as a (local) solution, e.g., for
  - integrating corpus and dictionary data

(Burchardt et al. 2008, Mazziotta et al. 2010)
by the time, long been recognized as a problem

- accessibility challenges
  - different formats, schemes
  - distributed
  - dispersed metadata collections

hence, several proposals to address them

Independently, different groups considered using RDF/OWL as a (local) solution

lexical resources long provided by the SemWeb

(Gangemi et al. 2003, Buitelaar et al. 2006)
by the time, long been recognized as a problem

- accessibility challenges
  - different formats, schemes
  - distributed
  - dispersed metadata collections

hence, several proposals to address them

- Independently, different groups considered using RDF/OWL as a (local) solution

... not actually linked
1. Use URIs as names for things
2. Use HTTP URIs, so people can look up the names
3. Provide data at the URIs, using standards like RDF and SPARQL
4. Include links to other URIs, so people can discover more things

=: Linked Data

http://www.w3.org/DesignIssues/LinkedData.html
LINKED DATA FOR LINGUISTICS

- Representation and modelling
- Dynamic Import
- Structural interoperability
- Conceptual interoperability
- Federation
- Community and ecosystem

Chiarcos et al. (2013)
LINKED DATA FOR LINGUISTICS

- Representation and modelling
- Dynamic Import
- Structural interoperability
- Conceptual interoperability
- Federation
- Community and ecosystem
INTEROPERABILITY BY LINKED DATA

- Structural interoperability
  - same representation formalism for different data sets

- Conceptual interoperability
  - same vocabularies for different data sets

- Federation
  - procedure for remote calls to external end points
INTEROPERABILITY BY LINKED DATA

- Structural interoperability
  + *same query language* for different data sets

- Conceptual interoperability
  + *same query* for different data sets

- Federation
  + *a single query* for different, distributed data sets

(simplified)
 Representation and modelling
 Dynamic Import
 Structural interoperability
 Conceptual interoperability
 Federation
 Community and ecosystem
by the time, long been recognized as a problem

- accessibility challenges
  - different formats, schemes
  - distributed
  - dispersed metadata collections

- But these activities were not coordinated
  - and in particular, RDF was used, but resources rarely linked to other resources in the web of data... and RDF as a solution...
LINGUISTIC LINKED OPEN DATA (LLOD) CLOUD

- a collection of linguistic resources
  + published under open licenses
  + as linked data
  + decentralized developed and maintained
  + meta data at datahub.io => linghub.org
    => cloud diagram
  + developed as a community effort in the context of the Open Linguistics Working Group of the Open Knowledge Foundation
OPEN LINGUISTICS WORKING GROUP (OWLG)

- founded in Oct 2010 in Berlin, Germany
  + Open Knowledge Foundation
- open network of individuals interested in
  + linguistic resources and/or
  + their publication under open licenses
- multi-disciplinary
  + NLP/CL, typology/language documentation, SW, ...
- infrastructure
  + mailing list, web site/blog, wiki
  + http://linguistics.okfn.org
OWLG Activities
(http://linguistics.okfn.org)

- Promoting open linguistic resources
  - Raising awareness, collecting metadata
    (datahub.io)
- Facilitating wide-range community activities
  - Workshops, mailing list, publications
    - Linked Data in Linguistics (LDL)
    - Multilingual Linked Open Data for Enterprises (MLODE)
    - Linked Data in Linguistic Typology (LDLT)
    - Knowledge Extraction and Knowledge Integration (KEKI)
initially, we maintained a list of open or representative resources
+ in Jan 2011, we marked possible synergies
merely a vision
+ includes non-open resources as placeholders for other resources to come
+ not physically realized
a strong metaphor brought to a new community
Representing and Linking Language Data and Language Metadata
not yet fully implemented
Multilingual Linked Open Data for Enterprises
+ first instance of the LLOD cloud built
+ workshop & hackathon
  ❏ authors provide data
  ❏ data conversion, metadata update at http://datahub.io

automatically generated diagram
+ original LOD scripts
+ from datahub
- more data sets
- manually categorized and colored
- intensified community work
  + W3C Community Groups Ontology-Lexica (OntoLex)
  + LIDER (EU FP7)
more data sets
more rigid criteria
community work
W3C CGs
QtLeap (EU FP7)
Aug 2014
top-level category in the LOD diagram

Linguistic Linked Open Data (LLOD) cloud diagram

April 2014
CC-BY Open Linguistics Working Group
(http://linguistics.okfn.org/llod)
created in preparation of the 3rd Linked Data in Linguistics Workshop (LDL-2014)

LEXICON:
lexical-semantic resources (LSRs)
- general semantic knowledge bases
- lexical resources with grammar information

CORPUS:
collections of language samples
- annotated corpora

METADATA:
information about language and language resources
- information about language resources (incl. bibliography)
- linguistic terminology repositories
- databases of language features (e.g., from typology)
BUILDING THE CLOUD: 2014+

- building infrastructure
  - dynamically generated diagram
    - linguistic-lod.org
  - improved meta data management (linghub.org)
    - restructures datahub content
- clarifying terminology
  - defining linguistic relevance
  - linguistic resource categories
- improving quality and coverage
  - enforcing more rigid constraints on LLOD „bubbles“
GETTING FASHIONABLE: 2015+

- **LREC-2014**
  - „the new hot topic in our community“

- **6 LLOD summer events**
  - 4th Multilingual Semantic Web Workshop (Portoroz, Slovenia, June 2015)
  - 1st Summer Datathon on LLOD (Spain, June 2015)
  - EUROLAN-2015 “Linguistic Linked Open Data” (Sibiu, Romania, July 2015)
  - LLOD workshop at the LSA Summer Institute (Chicago, US, July 2015)
  - LDL-2015 at ACL-IJCNLP 2015 (Beijing, July 2015)
  - NLP&LOD2 at RaNLP-2015 (Bulgaria, Sep 2015)
BUILDING THE CLOUD: LDL-2016

linguistic-lod.org

The LLDD diagram is maintained by the OKFN Working Group on Linguistics and provided under the Creative Commons Attribution 3.0 Unported (CC BY 3.0).
LONG-TERM DEVELOPMENTS

- 2010+
  - Linking Communities
    - NLP, Semantic Web, language documentation/typology, ...
    - pan-European => international
LONG-TERM DEVELOPMENTS

2010+ Linking Communities
2012+ Linking Resources

- emergence of consensus vocabularies and terminology bases
  - dictionaries
  - language identifiers
  - annotation terminology
  - NLP output

- now, we need to demonstrate usability
LONG-TERM DEVELOPMENTS

- **2010+** Linking Communities
- **2012+** Linking Resources
- **2015+** Broadening the Scope
  - Increasing focus on application
    - In NLP and linguistics
  - ... and dissemination
    - Summer schools and datathons
  - Reaching out to other communities
    - North America (LSA-LLOD workshop)
    - Asia & Pacific (LDL-2015 @ ACL-IJCNLP)
LONG-TERM DEVELOPMENTS

- **2010+** Linking Communities
- **2012+** Linking Resources
- **2015+** Broadening the Scope
- **2016+** Consolidation & Application?

[https://www.w3.org/2016/05/ontolex/](https://www.w3.org/2016/05/ontolex/)
LONG-TERM DEVELOPMENTS

- **2010+** Linking Communities
- **2012+** Linking Resources
- **2015+** Broadening the Scope
- **2016+** Consolidation & Application?

+ [https://www.w3.org/2016/05/ontolex/](https://www.w3.org/2016/05/ontolex/)
+ communities going LLOD, e.g., OLAC (May 2016)
LONG-TERM DEVELOPMENTS

× 2010+ Linking Communities
× 2012+ Linking Resources
× 2015+ Broadening the Scope
× 2016+ Consolidation & Application?

+ https://www.w3.org/2016/05/ontolex/

+ communities going LLOD
  × increasing number of LLOD-related events
  ∗ e.g., numerous activities at LREC
LONG-TERM DEVELOPMENTS

- 2010+ Linking Communities
- 2012+ Linking Resources
- 2015+ Broadening the Scope
- 2016+ Consolidation & Application?

+ https://www.w3.org/2016/05/ontolex/
+ communities going LLOD
+ applications
  - infrastruct
    - NLP
    - DH
LONG-TERM DEVELOPMENTS

- **2010+** Linking Communities
- **2012+** Linking Resources
- **2015+** Broadening the Scope
- **2016+** Consolidation & Application?

+ challenges remain
  - still a lot of uncertainty in the linguistic community
    - legal issues largely unsolved
    - no central hosting solutions so far
    - (often) limited in-house expertise
LONG-TERM DEVELOPMENTS

- 2010+ Linking Communities
- 2012+ Linking Resources
- 2015+ Broadening the Scope
- 2016+ Consolidation & Application?

+ challenges remain
  - still a lot of uncertainty in the linguistic community
  - reservations in the NLP community declining slowly, only
    - „As for standards, I don’t think these are important“
      (Mark Johnson, Computational linguistics: Where do we go from here?, keynote at the 50th ACL, Jeju, Korea 2012)
    - „What do we need that CoNLL (tsv) doesn’t provide us with?“
LONG-TERM DEVELOPMENTS

- 2010+ Linking Communities
- 2012+ Linking Resources
- 2015+ Broadening the Scope
- 2016+ Consolidation & Application?

+ challenges remain
  - still a lot of uncertainty in the linguistic community
  - reservations in the NLP community declining slowly, only
    - numerous applications in infrastructures
    - very few convincing LLOD-aware algorithms published, so far
LONG-TERM DEVELOPMENTS

- 2010+ Linking Communities
- 2012+ Linking Resources
- 2015+ Broadening the Scope
- 2016+ Consolidation & Application?

+ challenges remain
  - still a lot of uncertainty in the linguistic community
  - reservations in the NLP community declining slowly, only
  - shared vocabularies beyond lexical resources
    - still emerging
LONG-TERM DEVELOPMENTS

- 2010+ Linking Communities
- 2012+ Linking Resources
- 2015+ Broadening the Scope
- 2016+ Consolidation & Application?

+ challenges remain
  - still a lot of uncertainty in the linguistic community
  - reservations in the NLP community declining slowly, only
  - shared vocabularies beyond lexical resources
  - more activity => more coordination required
  - benefits of LLOD are due to using shared resources
LONG-TERM DEVELOPMENTS

× 2010+  Linking Communities
× 2012+  Linking Resources
× 2015+  Broadening the Scope
× 2016+  Consolidation & Application

+ challenges remain
  × still a lot of uncertainty in the linguistic community
  × reservations in the NLP community declining slowly, only
  × shared vocabularies beyond lexical resources
  × more activity => more coordination required

× benefits of LLOD are due to using shared resources

Interested in the challenges? Join the communities! Talk to us!

https://lists.okfn.org/mailman/listinfo/open-linguistics

Specialized W3C Community Groups
09:00 Opening and Introduction

09:30 **Keynote**: Damir Cavar (Indiana University / LINGUIST List)

10:30 Coffee Break

11:00 **Session I**: *Dictionaries and Ontologies*

12:40 Lunch Break

14:00 **Poster** Session

14:40 **Session II**: *Building Linked Language Resources*

16:00 Coffee Break

16:30 **Session III**: *Using Linked Language Resources*

17:40 Discussion and Closing

LREC 2016, May 24th, 2016, Portorož, Slovenia
For lunch, buffet option
+ **must** be booked until 12:00
+ GHB, one floor down
09:30 – 10:30 **Keynote**

Damir Cavar (Indiana University and The LINGUIST List)

*Linked Open Language Data for Language Documentation, Linguistic Research, and Speech and Language Technologies*