CLLD – Cross-Linguistic Linked Data

Robert Forkel

Department of Linguistics, Max Planck Institute for Evolutionary Anthropology

Goal

Help collecting the world's language diversity heritage by providing interoperable data publication structures

Linguistic databases on the web

Observations

- ✓ Almost all quantitative papers at ALT 10 used WALS data.
- ✓ Many typologists at ALT 10 have their own typological database.

CLLD – The strategy

Since reuse tends to be the determining factor that keeps resources from vanishing, we want to bridge the gap between data collection and data reuse by

- ✓ publishing databases thereby incentivizing researchers through recognition;
- ✓ using technology that maximizes exposure of our data in the emerging web of data.

CLLD – The implementation

This twofold strategy is implemented by three service components:

- ✓ infrastructural: Glottolog - a comprehensive language catalog and bibliography,
- ✓ structural: Dictionaria – a dictionary journal and CrossGram Journal – a journal publishing typological databases,
- ✓ technological: `clld` - a software platform for implementing linguistic database applications, which will be used to serve standalone database publications like IDS, WOLD, ASJP, WALS, APiCS, eWAVE, ValPal as well as the journals.

To maximize resuability

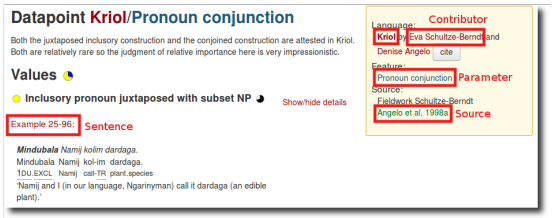
- ✓ we provide the data under Open Data Licenses,
- ✓ and the platform as Open Source software under a free license.

The data model

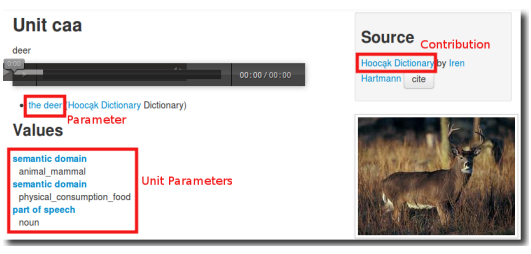
This implementation plan is realistic, because the underlying data model for the target databases is both reasonably simple and abstract enough to cover many use cases.

The data model comprises the following entities: Dataset, Contribution, Contributor, Language, Parameter, Value, UnitParameter, Unit, Sentence, Source.

Value – an APiCS datapoint



Unit – a Dictionaria word

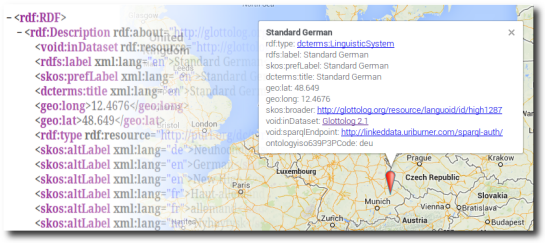


Linked Data - disseminating data in standard formats

- ✓ Defines a unified data access protocol for the web.
- ✓ Well-suited for distributed data providers
  - identifiers are URLs which are globally unique,
  - RDF and OWL provide the vocabulary to merge resources.
- ✓ Provides an easy to implement lowest level of service in a graceful degradation scenario

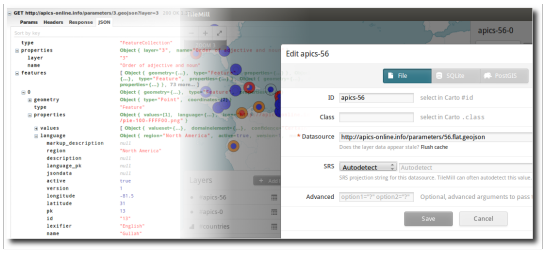
Use off-the-shelf tools to explore a dataset

Linked Data Explorer accessing Glottolog Linked Data serialized as RDF/XML.

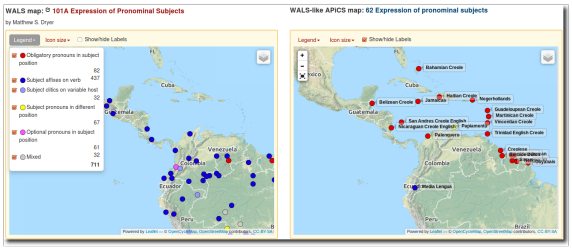


Use off-the-shelf tools to transform a dataset

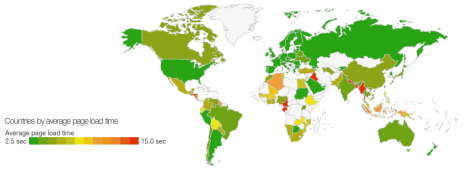
The map-making software Tilemill accessing APiCS data in GeoJSON format.



Comparing WALS and APiCS data pulled in as GeoJSON



Mission accomplished? – A week's Visitors to Glottolog



Printed at Universitätsrechenzentrum Leipzig