



# CLLD – Cross-Linguistic Linked Data

Robert Forkel

Department of Linguistics, Max Planck Institute for Evolutionary Anthropology

<http://cld.org>

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

## Cross-linguistic databases on the web

While several archives for cross-linguistic data exist (among them the DOBES archive), published databases, i.e. freely accessible, citable datasets are few and far between. This is the case despite the fact that many linguists collect lexical or typological datasets, serving as primary sources for their own publications.

## CLLD – The strategy

Since reuse tends to be the determining factor in keeping 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 JCLD – a journal publishing typological databases,
- ✓ technological: c1ld – a software platform for implementing linguistic database applications like Glottolog and the journals, but also to serve standalone datasets like
  - WALS - The World Atlas of Language Structures,
  - APiCS - The Atlas of Pidgin and Creole Language Structures,
  - WOLD - The World Loanword Database

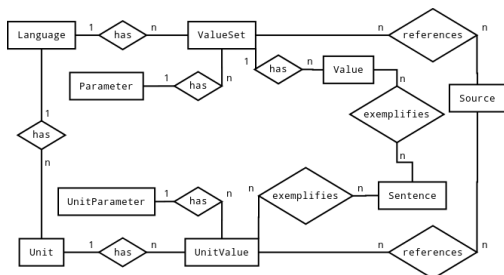
To maximize resuability

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

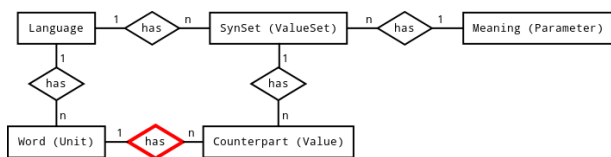
## The c1ld data model

It turns out that many linguistic datasets can be modelled using a small set of concepts.

The core data model:



The WOLD incarnation of this data model:



## Linked Data - the c1ld API

- ✓ 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

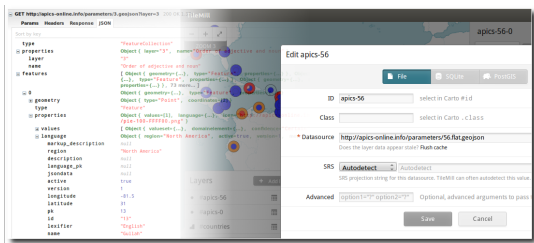
## Emerging API between data publications and tools

Linked Data does already provide a data access protocol, i.e. the first part of an API to be used by tools which analyze, visualize or otherwise reuse data.

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



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



## Graceful degradation of service

Providing access to datasets following Linked Data principles can be rather easy, e.g. by serving static files from an HTTP server. On the other hand we argue that this level of data access can and should be sufficient to establish an API, i.e. to sustain an infrastructure of tools on top of the data. Thus, we use this emerging API as definition of a minimal level of service which is easy to uphold.

The c1ld framework will provide an "emergency exit" feature, which will create a set of files in an appropriate directory structure to be put on a vanilla webserver. This can be done by enumerating the resource types, instances and available representations.

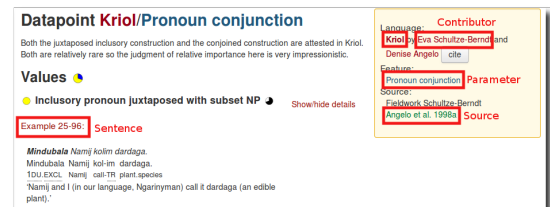
So while Linked Data is still not the way researchers actually do or want to access data (at least if they can get away with csv instead), there's something to be gained for the developer: A stable API across phases of deployment which can be used by any additional services built on top of the data.

Since maintaining the minimal level of service is rather easy, providing sustainable services becomes much more likely because scenarios like transfer of ownership become feasible.

## The c1ld data browser

A reference implementation for visualising CLLD datasets.

Viewing a valueset of APiCS Online:



Viewing a Dictionaria word:

