CLLD – Cross-Linguistic Linked Data

Robert Forkel
Department of Linguistics, Max Planck Institute for Evolutionary Anthropology

http://clld.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: `clld` - 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

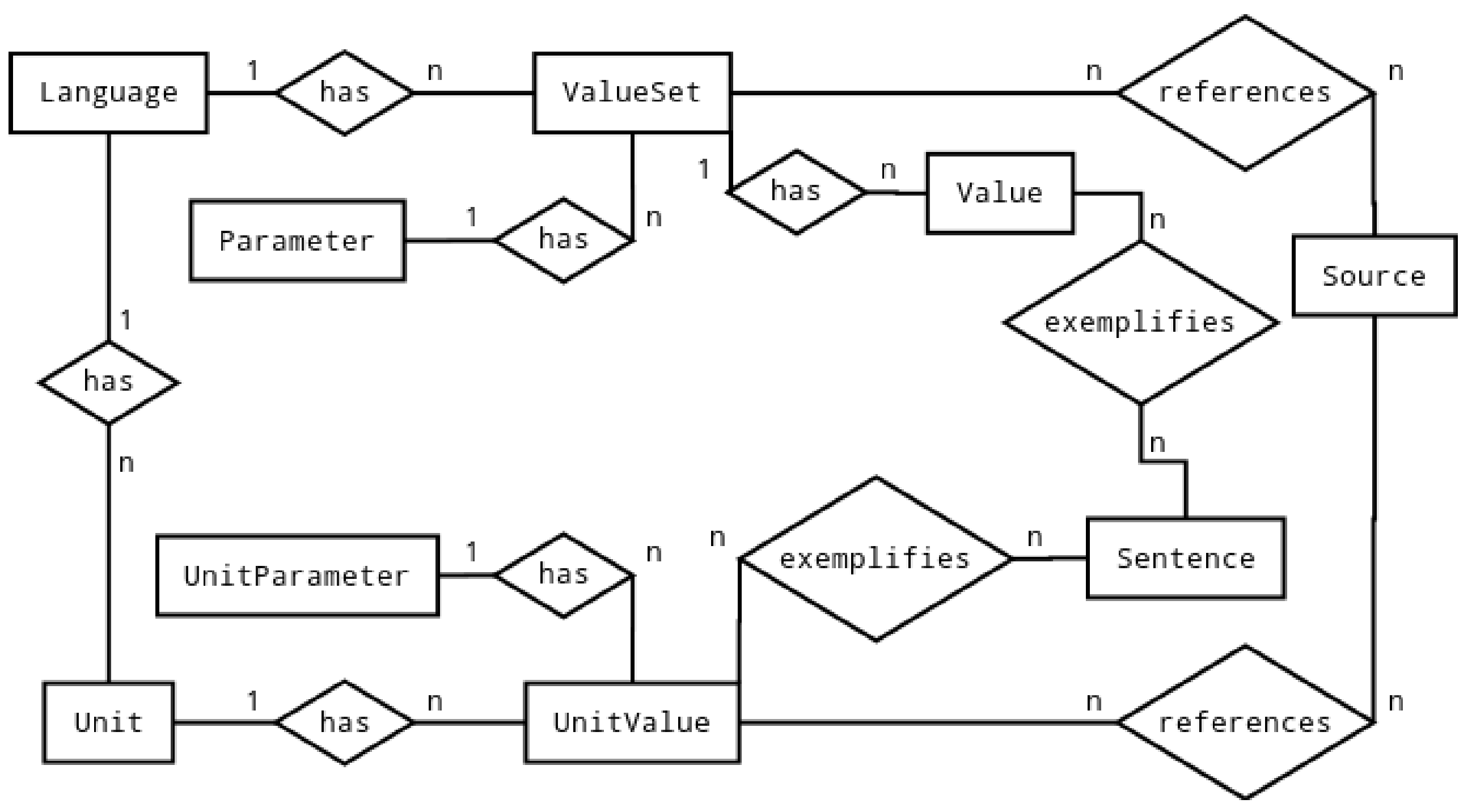
To maximize resuability

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

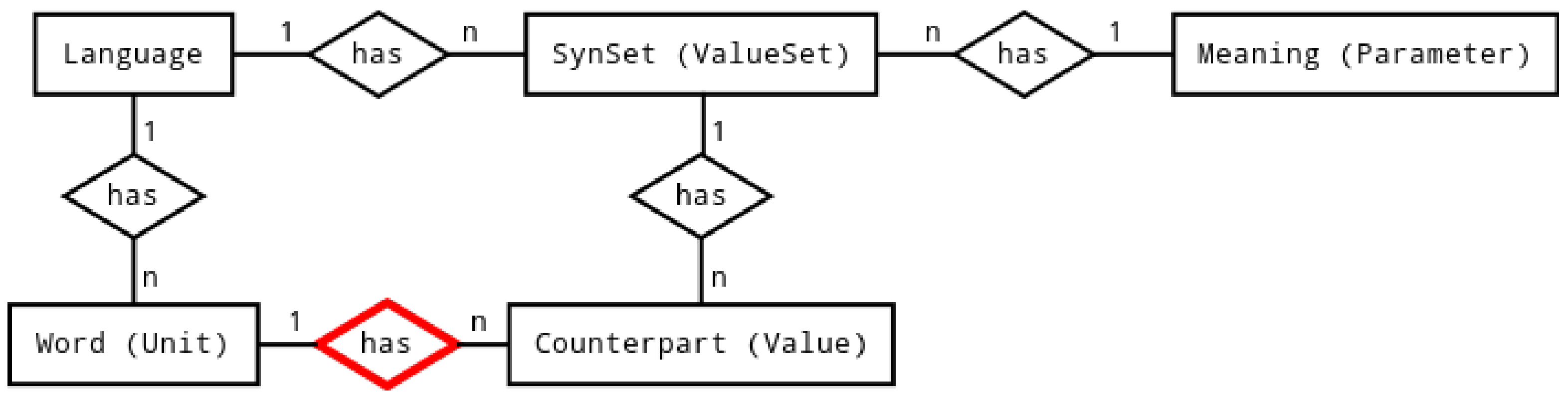
The `clld` data model

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:



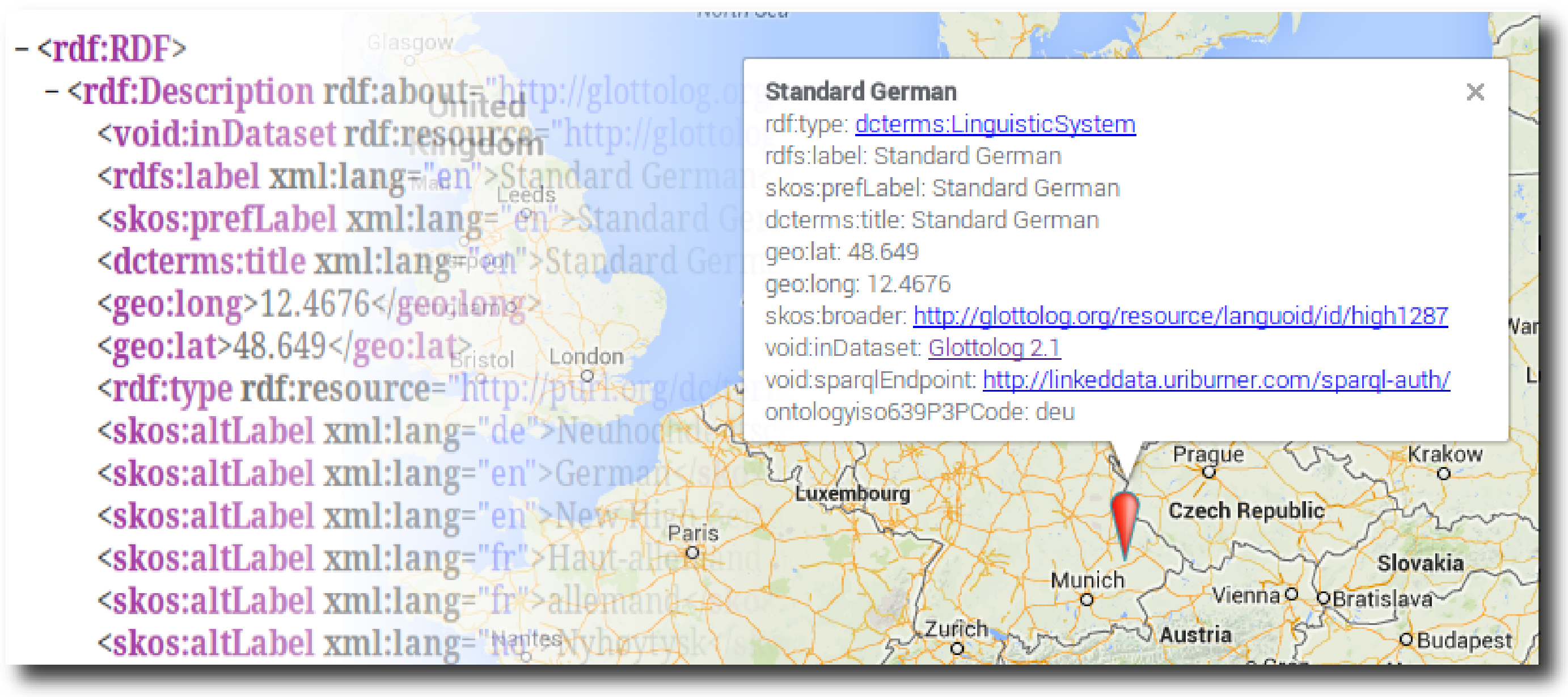
Linked Data - the `clld` API

- ✓ 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.

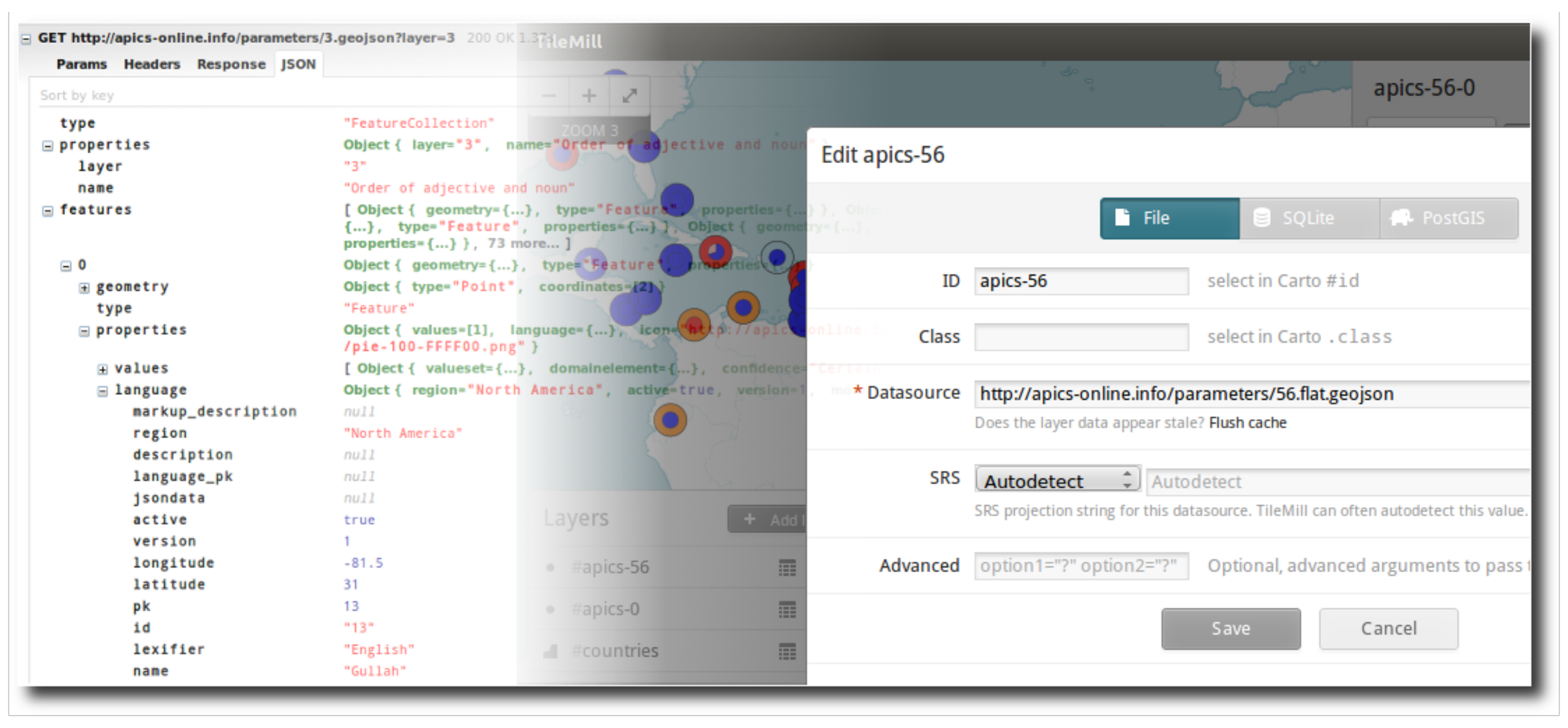
Linked Data Explorer accessing Glottolog Linked Data serialized as RDF/XML.



```
- <rdf:RDF>
  - <rdf:Description rdf:about="http://glottolog.o
      <void:inDataset rdf:resource="http://glottolo
      <rdfs:label xml:lang="en">Standard German
      <skos:prefLabel xml:lang="en">Standard G
      <dcterms:title xml:lang="en">Standard Ger
      <geo:long>12.4676</geo:long>
      <geo:lat>48.649</geo:lat>
      <rdf:type rdf:resource="http://purl.org/dc/
      <skos:altLabel xml:lang="de">Neuhoch
      <skos:altLabel xml:lang="en">German
      <skos:altLabel xml:lang="en">New High
      <skos:altLabel xml:lang="fr">Haut-allemand
      <skos:altLabel xml:lang="fr">allemand
      <skos:altLabel xml:lang="no">Nyhøytysk
```

Standard German
rdf:type: dcterms:LinguisticSystem
rdfs:label: Standard German
skos:prefLabel: Standard German
dcterms:title: Standard German
geo:lat: 48.649
geo:long: 12.4676
skos:broader: http://glottolog.org/resource/languoid/id/high1287
void:inDataset: Glottolog 2.1
void:sparqlEndpoint: http://linkeddata.uriburner.com/sparql-auth/
ontologyiso639P3PCode: deu

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



GET http://apics-online.info/parameters/3.geojson?layer=3 200 OK

```
type                    "FeatureCollection"
properties              Object { layer="3", name="Order of adjective and noun" }
  layer                 "3"
  name                  "Order of adjective and noun"
features                [ Object { geometry={...}, type="Feature", properties={...} }, 73 more... ]
  0                     Object { geometry={...}, type="Feature", properties=... }
    geometry            Object { type="Point", coordinates=[2] }
    type                "Feature"
    properties          Object { values=[1], language={...}, icon="http://apics-online.i.../pie-100-FFFF00.png" }
      values            [ Object { valueset={...}, domainelement={...}, confidence="Certai... } ]
      language          Object { region="North America", active=true, version=1, ... }
        markup_description  null
        region          "North America"
        description     null
        language_pk     null
        jsondata        null
        active          true
        version         1
        longitude       -81.5
        latitude        31
        pk              13
        id              "13"
        lexifier        "English"
        name            "Gullah"
```

Edit apics-56
ID: apics-56 — select in Carto #id
Class: — select in Carto .class
* Datasource: http://apics-online.info/parameters/56.flat.geojson
Does the layer data appear stale? Flush cache
SRS: Autodetect — Autodetect
SRS projection string for this datasource. TileMill can often autodetect this value.
Advanced: option1="?" option2="?" — Optional, advanced arguments to pass

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 `clld` 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.

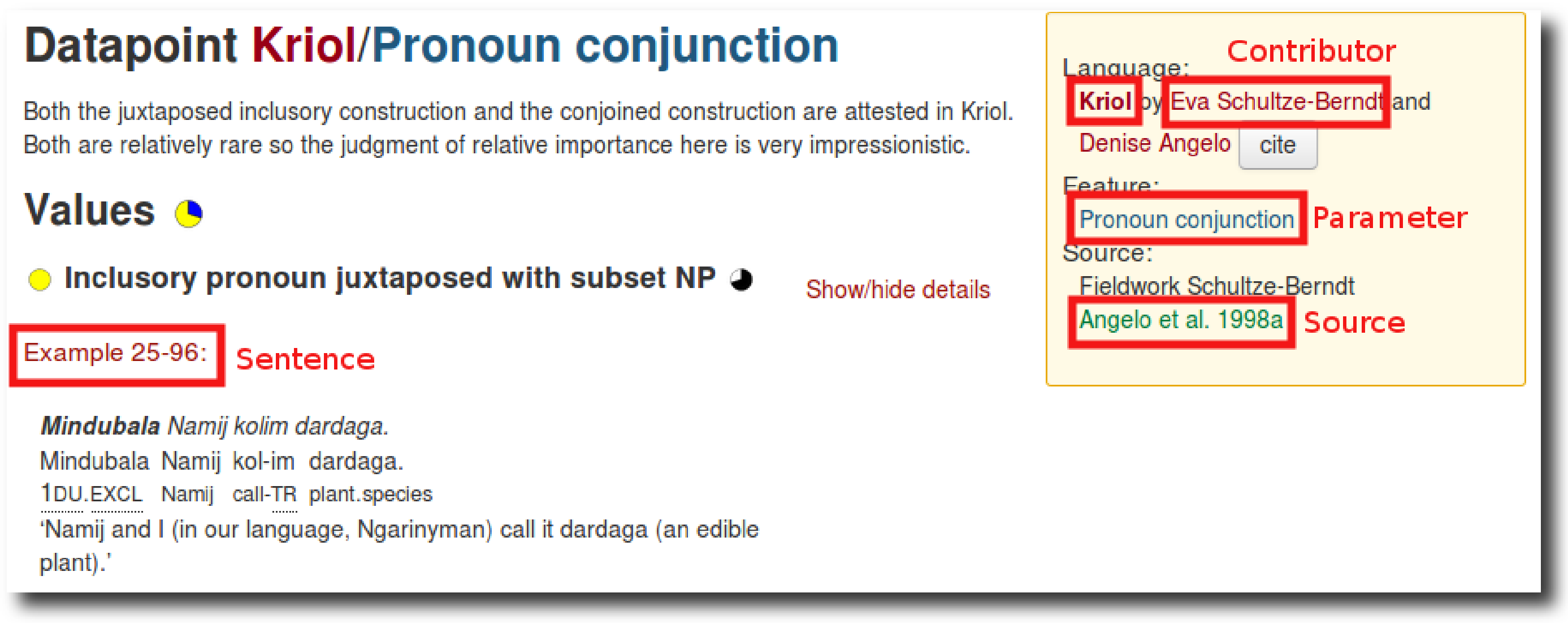
So while Linked Data is still not the way researchers actually do or want to access data (at least if they can get away with csv instead), there's something to be gained for the developer: A stable API across phases of deployment which can be used by any additional services built on top of the data.

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 `clld` data browser

A reference implementation for visualising CLLD datasets.

Viewing a valueset of APiCS Online:



Datapoint Kriol/Pronoun conjunction

Both the juxtaposed inclusory construction and the conjoined construction are attested in Kriol. Both are relatively rare so the judgment of relative importance here is very impressionistic.

Values

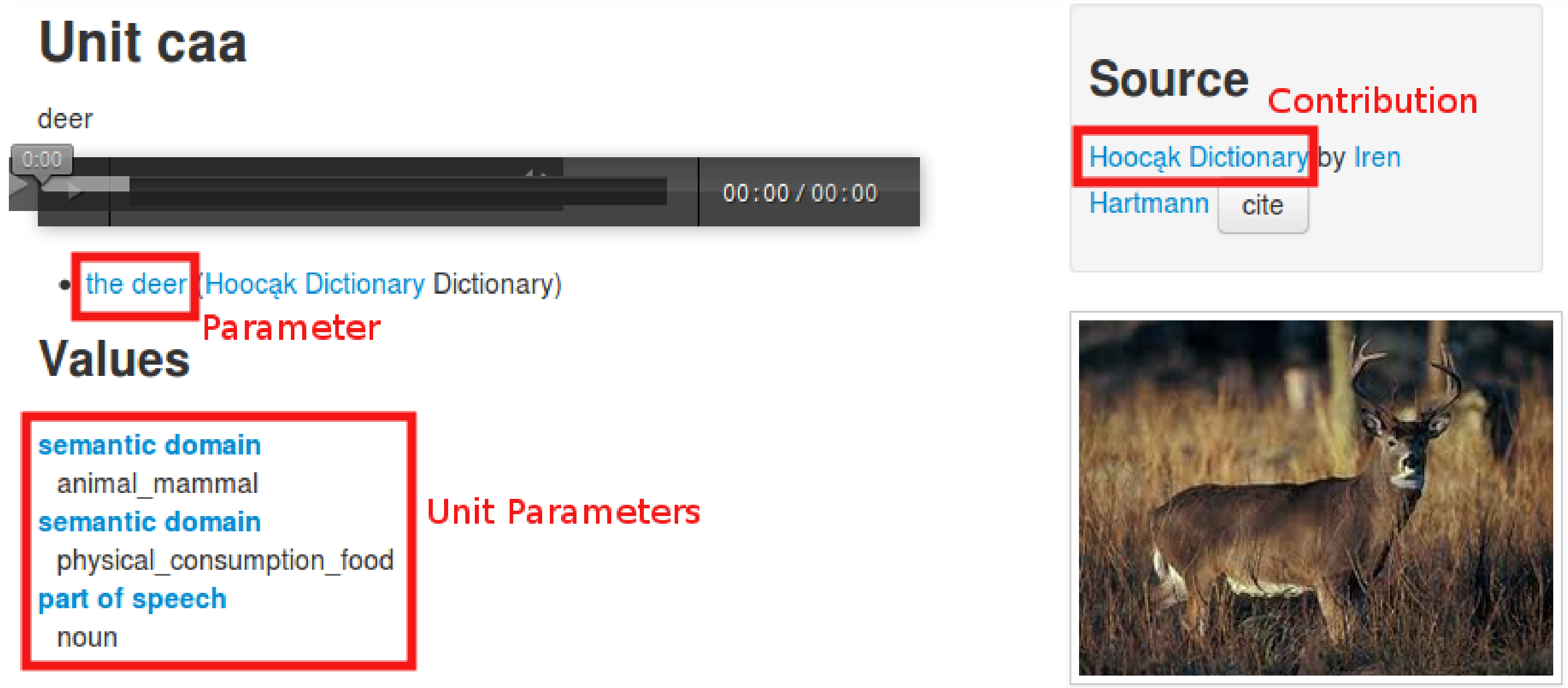
● Inclusory pronoun juxtaposed with subset NP — Show/hide details

Example 25-96: — Sentence

Mindubala Namij kolim dardaga.
Mindubala Namij kol-im dardaga.
1DU.EXCL Namij call-TR plant.species
'Namij and I (in our language, Ngarinyman) call it dardaga (an edible plant).'

Language: Kriol by Eva Schultze-Berndt and Denise Angelo — cite — Contributor
Feature: Pronoun conjunction — Parameter
Source: Fieldwork Schultze-Berndt; Angelo et al. 1998a — Source

Viewing a Dictionaria word:



Unit caa

deer

00:00 / 00:00

- the deer (Hoocąk Dictionary Dictionary) — Parameter

Values

semantic domain
animal_mammal
semantic domain
physical_consumption_food
part of speech
noun
— Unit Parameters

Source — Contribution
Hoocąk Dictionary by Iren Hartmann — cite