The `clld` toolkit

Robert Forkel and Sebastian Bank

Max Planck Institute for Evolutionary Anthropology, Leipzig

October 7, 2014

Outline

The CLLD project: Overview

Funded by the Max Planck Society for 4 years.

Creates infrastructure for publishing cross-linguistic datasets, including

- organization: a publication platform **`http://clld.org`** supporting two publication models:
  - Standalone databases following an ”edited series” model, like WALS, WOLD, ...
  - Two journals for cross-linguistic datasets
- infrastructure: Glottolog, a language catalog and comprehensive bibliography
- technology: the `clld` toolkit powering our applications

The CLLD project: Datasets

Typological:

- WALS - the World Atlas of Language Structures - a database of structural properties of more than 2600 languages
- APiCS - the Atlas of Pidgin and Creole Language Structures
- SAILS - the South American Indigenous Language Structures
- PHOIBLE - a repository of cross-linguistic phonological inventory data

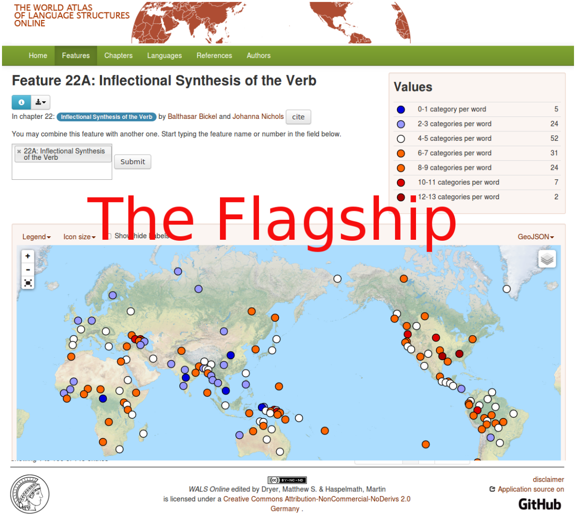
Lexical:

- WOLD - the World Loanword Database contains vocabularies of 41 languages from around the world annotated for loanword status
- Tsammalex - a multilingual lexical database on plants and animals
- IDS - the Intercontinental Dictionary Series (to be published in CLLD in 2014)
- ASJP - the Automated Similarity Judgement Project (to be published in 2014)

Encyclopedic:

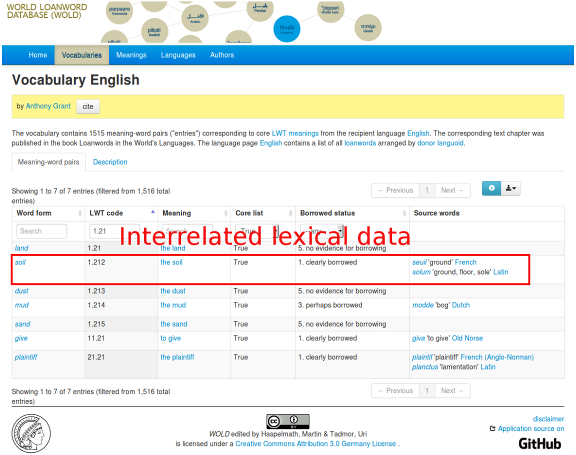
- Glottolog - a language catalog and comprehensive bibliography

The CLLD project: WALS



The Flagship

The CLLD project: WOLD



WORLD LOANWORD DATABASE (WOLD)

Home | Vocabularies | Meanings | Languages | Authors

Vocabulary English

by Anthony Grant cite

The vocabulary contains 1515 meaning-word pairs ("entries") corresponding to core LWT meanings from the recipient language English. The corresponding text chapter was published in the book Loanwords in the World's Languages. The language page English contains a list of all loanwords arranged by donor languoid.

Meaning-word pairs | Description

Showing 1 to 7 of 7 entries (filtered from 1,516 total entries)

← Previous 1 Next →

Interrelated lexical data

Word form	LWT code	Meaning	Core list	Borrowed status	Source words
Search	1.21	Search	True	--any--	
land	1.21	the land	True	5. no evidence for borrowing	
soil	1.212	the soil	True	1. clearly borrowed	*seuil* 'ground' French *solum* 'ground, floor, sole' Latin
dust	1.213	the dust	True	5. no evidence for borrowing	
mud	1.214	the mud	True	3. perhaps borrowed	*modde* 'bog' Dutch
sand	1.215	the sand	True	5. no evidence for borrowing	
give	11.21	to give	True	1. clearly borrowed	*giva* 'to give' Old Norse
plaintiff	21.21	the plaintiff	True	1. clearly borrowed	*plaintif* 'plaintiff' French (Anglo-Norman) *planctus* 'lamentation' Latin

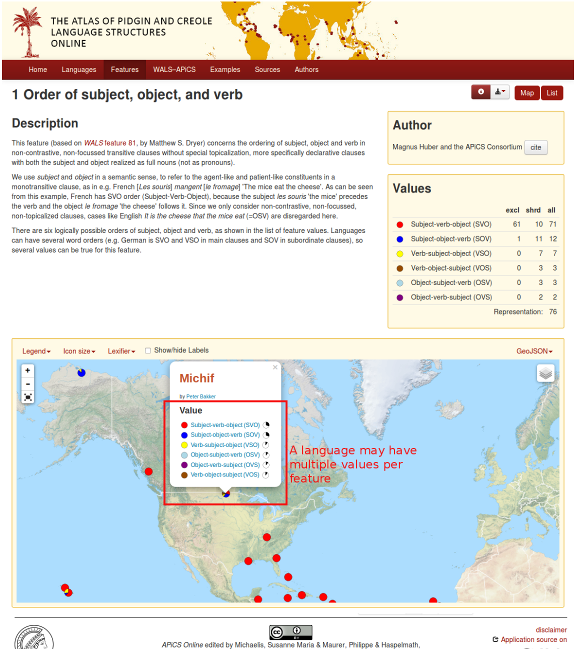
Showing 1 to 7 of 7 entries (filtered from 1,516 total entries)

← Previous 1 Next →

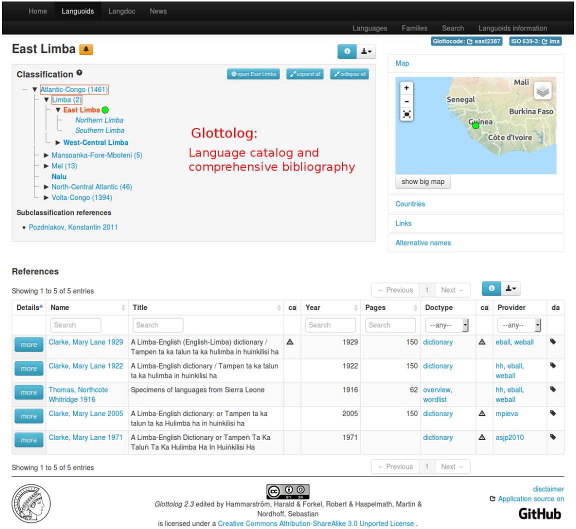
WOLD edited by Haspelmath, Martin & Tadmor, Uri
is licensed under a Creative Commons Attribution 3.0 Germany License .

disclaimer
Application source on GitHub

The CLLD project: APiCS



The CLLD project: Glottolog



Home Languoids Langdoc News

Languages Families Search Languoids information

Glottocode: east2387 ISO 639-3: lma

East Limba

Classification

open East Limba · expand all · collapse all

- Atlantic-Congo (1461)
  - Limba (2)
    - East Limba
      - Northern Limba
      - Southern Limba
    - West-Central Limba
  - Mansoanka-Fore-Mboteni (5)
  - Mel (13)
  - Nalu
  - North-Central Atlantic (46)
  - Volta-Congo (1394)

Subclassification references

- Pozdniakov, Konstantin 2011

Glottolog:

Language catalog and comprehensive bibliography

Map

show big map

Countries

Links

Alternative names

References

Showing 1 to 5 of 5 entries ← Previous 1 Next →

Details	Name	Title	ca	Year	Pages	Doctype	ca	Provider	da
	Search	Search		Search	Search	--any--		--any--	
more	Clarke, Mary Lane 1929	A Limba-English (English-Limba) dictionary / Tampen ta ka talun ta ka hulimba in huinkilisi ha	⚠	1929	150	dictionary	⚠	eball, weball	
more	Clarke, Mary Lane 1922	A Limba-English dictionary / Tampen ta ka talun ta ka hulimba in huinkilisi ha		1922	150	dictionary		hh, eball, weball	
more	Thomas, Northcote Whitridge 1916	Specimens of languages from Sierra Leone		1916	62	overview, wordlist		hh, eball, weball	
more	Clarke, Mary Lane 2005	A Limba-English dictionary: or Tampen ta ka talun ta ka Hulimba ha in huinkilisi ha		2005	150	dictionary	⚠	mpieva	
more	Clarke, Mary Lane 1971	A Limba-English Dictionary or Tampeǹ Ta Ka Taluǹ Ta Ka Hulimba Ha In Huiǹkilisi Ha		1971		dictionary	⚠	asjp2010	

Showing 1 to 5 of 5 entries ← Previous 1 Next →

Glottolog 2.3 edited by Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Nordhoff, Sebastian

is licensed under a Creative Commons Attribution-ShareAlike 3.0 Unported License .

disclaimer

Application source on GitHub

The CLLD project: AfBo

AfBo: A world-wide survey of affix borrowing

Home About Map Languages Affix functions References

About Legal Download Contact

Latin affixes in Basque

Subjects are recipient/donor pairs rather than single languages

Summary

Affix function	number of borrowed affixes
adjectivizer	2
augmentative	2
diminutive	2
gender (human)	1
nominal derivation (miscellaneous)	4
nominalizer: abstract	17
nominalizer: agent	4
nominalizer: social group	2
privative	1

Recipient language:	Basque
Donor language:	Latin
Reliability of borrowed status/affixhood:	high
Borrowed affixes:	35
Interrelated affixes:	35

Description

Information and examples are from Segura Munguía and Etxebarria Ayesta (1996) and Hualde and Urbina (2003). See also Haase (1992: 48–51), who focuses on French and Gascon influence on the Basque variety of lower Navarra, and Mujika (1982).

2 diminutive suffixes

-*ila*, -*ilo* 'diminutive', e.g. *neskatila* 'little girl' (from *neska* 'girl'), *leihatila* 'little window' (from *leiho* 'window'), *andrakila* 'doll' (from *andra* 'woman'), *astokilo* 'little donkey' (from *asto* 'donkey'), *gizonilo* 'little man' (from *gizon* 'man') (Hualde 2003a: 331) (see also Haase 1992: 49; Segura Munguía and Etxebarria Ayesta 1996: 84, 89)

-*(i)no*, -*(i)na*, -*(i)ño*, -*(i)ña*, -*ño* 'diminutive', e.g. *emekiñio* 'very softly' (from *emeki* 'softly'), *batño* 'a little one' (from *bat* 'one') (Hualde 2003a: 331; Segura Munguía and Etxebarria Ayesta 1996: 89)

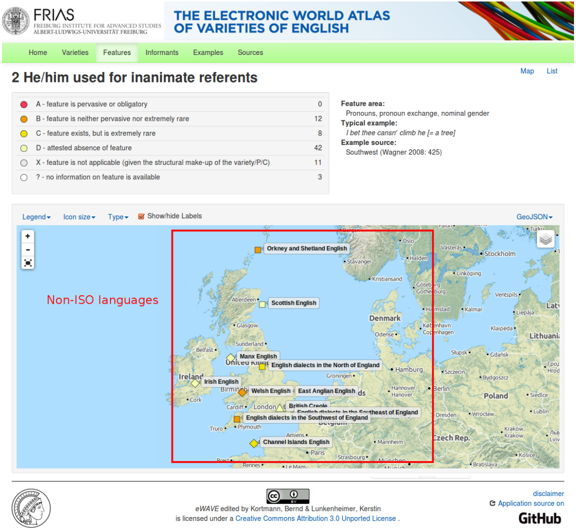
References

Haase, Martin 1992
Sprachkontakt und Sprachwandel im Baskenland : die Einflüsse des Gaskognischen und Französischen auf das Baskische
Hualde, José Ignacio 2003
Segmental phonology
Hualde, José Ignacio 2003
Derivation
Hualde, José Ignacio and Urbina, Jon Ortiz de 2003
A grammar of Basque
Mujika, Luis Mari 1982
Latina eta erromanikoaren eragina euskaran. Euskal lexikoaren azterketa bideetan
Segura Munguía, Santiago and Etxebarria Ayesta, Juan Manuel 1996
Del latín al euskara = Latinetik euskarara
Trask, R.L. 2003
The Noun Phrase: nouns, determiners and modifiers; pronouns and names

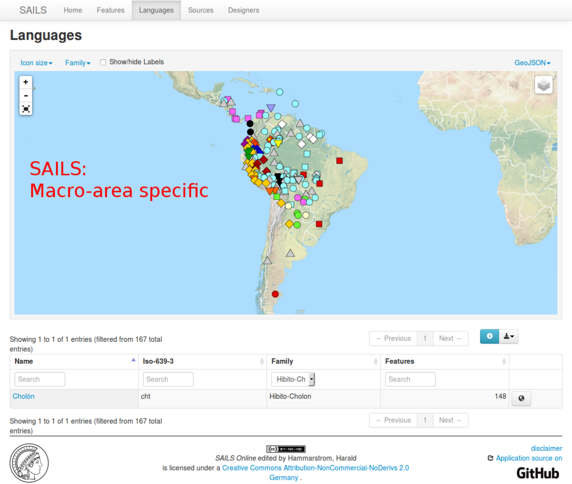
AfBo: A world-wide survey of affix borrowing by Seifart, Frank is licensed under a Creative Commons Attribution 3.0 Unported License .

disclaimer
Application source on GitHub

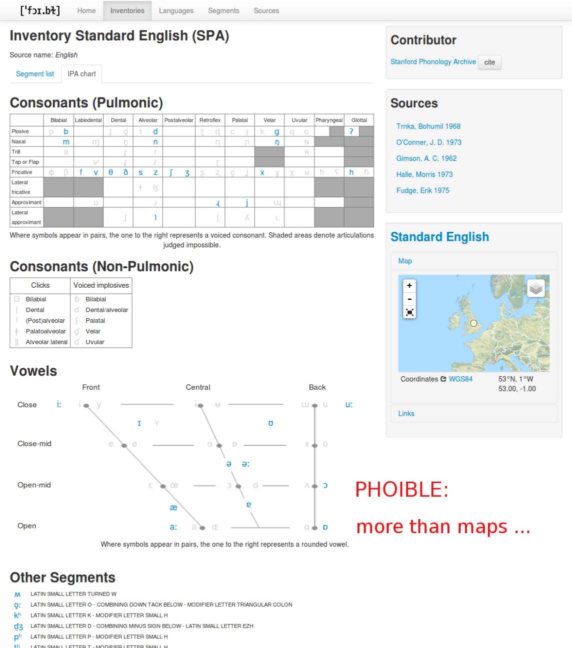
The CLLD project: eWAVE



The CLLD project: SAILS



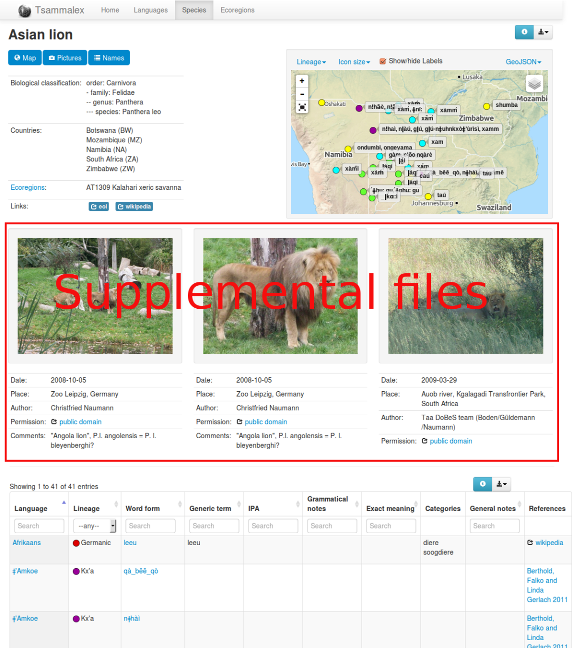
The CLLD project: PHOIBLE



PHOIBLE:

more than maps ...

The CLLD project: Tsammalex



Tsammalex Home Languages Species Ecoregions

Asian lion

Map | Pictures | Names

Biological classification:	order: Carnivora - family: Felidae -- genus: Panthera --- species: Panthera leo
Countries:	Botswana (BW) Mozambique (MZ) Namibia (NA) South Africa (ZA) Zimbabwe (ZW)
Ecoregions:	AT1309 Kalahari xeric savanna
Links:	eol wikipedia

Lineage Icon size Show/hide Labels GeoJSON

Supplemental files

Date:	2008-10-05
Place:	Zoo Leipzig, Germany
Author:	Christfried Naumann
Permission:	public domain
Comments:	"Angola lion", P.l. angolensis = P. l. bleyenberghi?

Date:	2008-10-05
Place:	Zoo Leipzig, Germany
Author:	Christfried Naumann
Permission:	public domain
Comments:	"Angola lion", P.l. angolensis = P. l. bleyenberghi?

Date:	2009-03-29
Place:	Auob river, Kgalagadi Transfrontier Park, South Africa
Author:	Taa DoBeS team (Boden/Güldemann /Naumann)
Permission:	public domain

Showing 1 to 41 of 41 entries

Language	Lineage	Word form	Generic term	IPA	Grammatical notes	Exact meaning	Categories	General notes	References
Search	--any--	Search	Search	Search	Search	Search		Search	
Afrikaans	Germanic	leeu	leeu				diere soogdiere		wikipedia
ǂAmkoe	Kx'a	qà_bēē_qò							Berthold, Falko and Linda Gerlach 2011
ǂAmkoe	Kx'a	nǂhài							Berthold, Falko and Linda Gerlach 2011

The CLLD project: Where's my dataset?

Have a dataset in need of publication and presentation on the web?

- Submit to Harald's **J**ournal of **C**ross-**L**inguistic **D**atabases or
- submit to Martin's edited series of cross-linguistic databases **clld.org** or
- get a seasoned python programmer for a month to build your own app on top of the `clld` toolkit!

```
robert@astroman:/tmp/phoible$ cloc --exclude-dir=tests,data phoible/
      38 text files.
      36 unique files.
      28 files ignored.
-------------------------------------------------------------------------------
Language                     files          blank        comment           code
-------------------------------------------------------------------------------
Python                          17            230            173            954
CSS                              1             25             49            159
Javascript                       1              1              0              0
-------------------------------------------------------------------------------
SUM:                            19            256            222           1113
-------------------------------------------------------------------------------
```

The `clld` toolkit: Motivation

Survey databases are all alike.

Can we extract functionality needed to build WALS, WOLD, and APiCS into a reusable piece of software?
Design goals:

- There must be a core database model, which allows for as much shared functionality as possible.
- User interfaces of applications must be fully customizable.
- It must be easy to re-implement legacy applications using the framework.
- Optimize for maintainability, i.e. minimize lines-of-code for apps built with the framework.
- Find the right level of abstraction!

`clld`: A CMS for cross-linguistic data

The `clld` toolkit is an open source Python package hosted on GitHub providing

- an extensible core data model
- a web application framework
  - powering all CLLD databases
  - providing a basic API built on Linked Data principles
  - ”reference implementation” of a dataset browser
  - `clld` apps are web applications built as small layer of code on top of the `clld` framework.
  - `clld` works with python 2.7 and 3.4 and has a test suite with 100% coverage.

Intermezzo: Disambiguation

- CLLD: The project.
- clld.org: The publisher/brand.
- `clld`: The software, aka toolkit, aka framework.
- `clld` app: A web application built using the `clld` framework.

In the remainder of this presentation we will talk about the latter two.

`clld` data model: Design

The design of the data model was guided by three principles:

- All the target datasets have to “fit in” without loss.
- The data model must be as abstract as necessary, as concrete as possible.
- The data model must be extensible.

`clld` data model: Entities

- **Dataset** holds metadata about a dataset like license and publisher information.
- **Language** may be a languoid (Glottolog) or doculect (ASJP).
- **Parameter** a feature that can be determined and coded for a language – e.g. a word meaning, or a typological feature.
- **ValueSet** set of values measured/observed/recorded for one language and one parameter, i.e. the points in the Language-Parameter-matrix.
- **Value** a single measurement (different types of scales can be modeled using custom attributes).
- **Unit** parts of a language system that are annotated, such as sounds, words or constructions.
- **UnitParameter** a feature that can be determined for a unit.
- **UnitValue** measurement for one unit and one unitparameter.
- **Contribution** ValueSets can be partitioned into separate contributions sharing provenance.

`clld` data model: Relationships

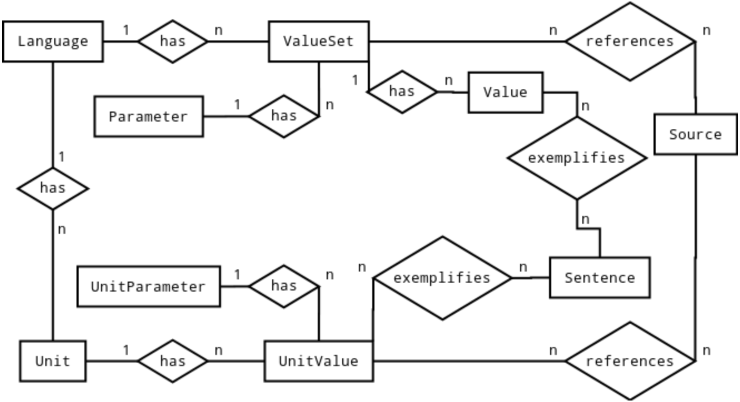


Figure 1: The default `clld` data model. Note: Modelling constructions as Units and features as UnitParameters the case mentioned by Harald fits in.

`clld` data model: Extensibility

`clld` uses *joined table inheritance* as implemented in SQLAlchemy to provide extensibility of the core data model:

- Each core model can be specialised/customized in a `clld` app, adding columns or relationships.

```
@implementer(ILanguage)
class Languoid(Language, CustomModelMixin):
    ...
```

- The ORM (Object Relational Mapper) transparently joins the two corresponding tables when querying, retrieving the specialized object, i.e. the full set of columns.
- Additional models can be added freely, reusing `clld` functionality to enable functionality like versioning, etc.

clld data model: Lexical data

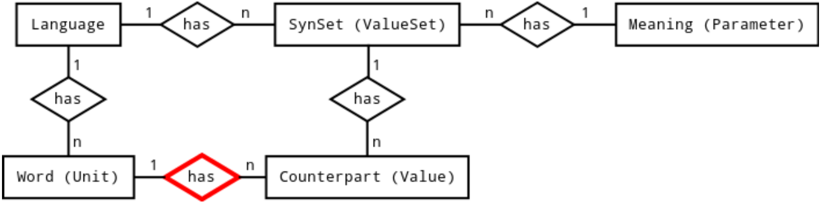


Figure 2: The WOLD instantiation of the data model.

```
@implementer(interfaces.IValue)
class Counterpart(Value, CustomModelMixin):
    ...
    word_pk = Column(Integer, ForeignKey('word.pk'))
    word = relationship(Word, backref='counterparts')
    ...
```

clld data model: Lexical data

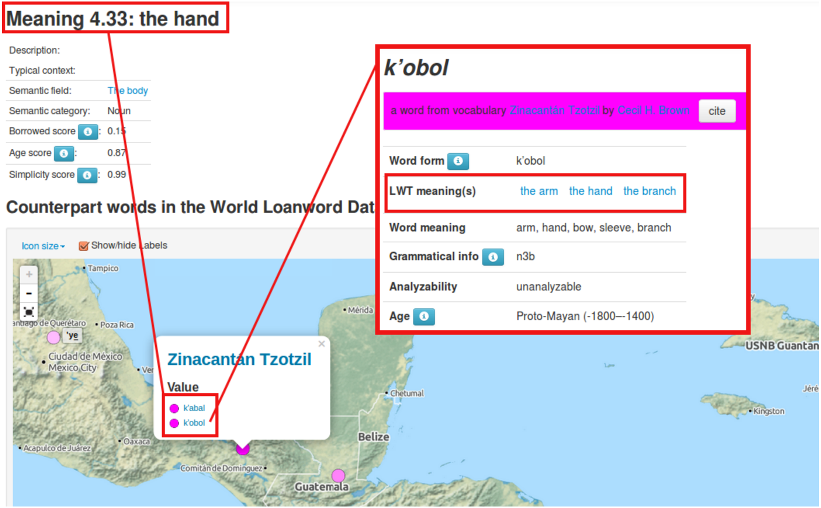


Figure 3: Many-to-many relation between words and meanings in WOLD.

clld data model: Glottolog

Figure 4: In Glottolog genealogy is implemented via a self-referential `father` relation on `Language`.

```
@implementer(ILanguage)
class Languoid(Language, CustomModelMixin):
    ...
    father_pk = Column(Integer, ForeignKey('languoid.pk'))
    children = relationship(
        'Languoid',
        foreign_keys=[father_pk],
        backref=backref('father', remote_side=[pk]))
    ...
```

`clld` resources: Overview

Data done the Web way.

`clld` implements a Resource Oriented Architecture.

- Data model is good basis to support shared behaviour across apps.
- Resource concept makes model entities actionable.
  - Resources are the things we describe and publish.
  - Resources define the level of granularity that is of interest.
- `clld` knows how to display filtered lists of resources of the same type
- and detail views of single resources.

`clld` resources: Adaption

- ZCA (Zope Component Architecture) provides machinery to register behaviour tied to interfaces, e.g. to resources.
- Resources can be adapted to representations:
  - Glottolog: Language represented as family tree in newick format.
  - ASJP: Contribution serialized in ASJP wordlist format.
  - All lists can be represented as feeds.
- The web pages created by a `clld` app are just resources adapted to HTML.
- These registry entries can be overridden by `clld` apps, e.g. providing custom DataTables, custom map markers, custom maps.
- Again it's about the right level of abstraction: Writing a `clld` app as declarative as possible, just implement adapters.

cld resources: Adaption

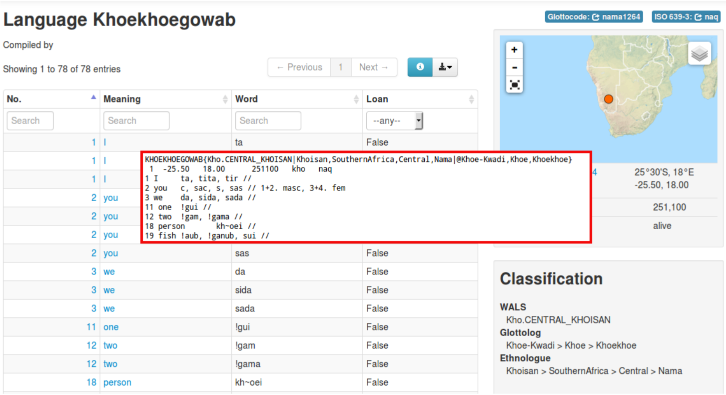
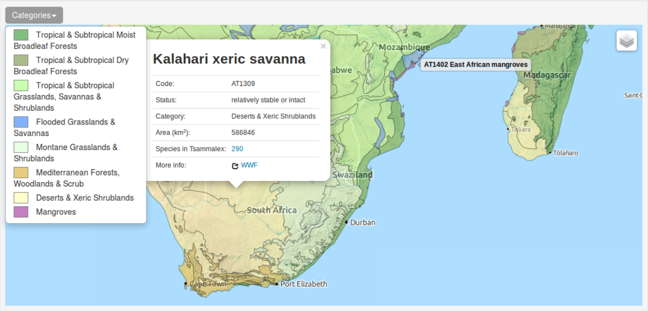


Figure 5: Two adaptions of a Language object in ASJP.

`clld` resources: Extensibility



Showing 1 to 15 of 15 entries (filtered from 114 total entries)

Code	Name	Category	Status		
Search	sava	--any--	--any--		Search
AT0705	East Sudanian savanna	Tropical & Subtropical Grasslands, Savannas & Shrublands	critical or endangered		WWF
AT0707	Guinean forest-savanna mosaic	Tropical & Subtropical Grasslands, Savannas & Shrublands	critical or endangered		WWF

Figure 6: Tsammalex defines a new resource type EcoRegion. EcoRegions behave just like other resources, i.e. they can be listed, bookmarked and associated with maps.

`clld` and Linked Data

- We regard Linked Data principles as rules of best practice for publishing data on the web.
- How do `clld` apps fare with respect to the five-star rating for Linked Data?
  - $*$ Make your stuff available on the web (whatever format).
  - $**$ Make it available as structured data (e.g. excel instead of image scan of a table).
  - $***$ Non-proprietary format (e.g. csv instead of excel).
  - $****$ Use URLs to identify things, so that people can point at your stuff.
  - $*****$ Link your data to other people's data to provide context.

`clld` and Linked Data: three stars

> *Make your stuff available on the web, as structured data in non-proprietary formats.*

- `clld` apps do just that.
- Most CLLD datasets are published under CC-BY, i.e. open, licenses.

cld and Linked Data: three stars

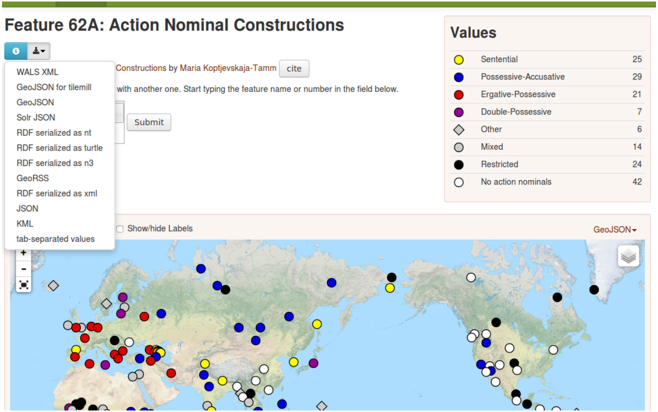


Figure 7: The data of a WALS feature is available in various formats. Note that the map on the page is created by calling the WALS API to retrieve the GeoJSON representation.

`clld` and Linked Data: four stars

Use URLs to identify things, so that people can point at your stuff.

- “People” includes yourself
- forces you to think about the things you want to describe and at which level of granularity
- enables distributed development of data and the basis for merging via globally unique identifiers
- puts coarse provenance information in each identifier

clld and Linked Data: four stars

http://wals.info/valuesets/138A-lat

Figure 8: The level of granularity of the WALS data allows to link comments, history and examples to datapoints.

`clld` and Linked Data: 4-out-of-5 stars

Generally, the usefulness of “4-out-of-5 stars” Linked Data has to be stressed:

- Linked Data as uniform data access API (following the “crawler” paradigm)
- enables distributed databases,
- allows follow-your-nose API discovery (cf. REST),
- plays well with the web at large (Internet archive, bookmarking, google, etc.),
- allows easy hosting (thus helps with sustainability, and is attractive for developers/administrators as well) – which cannot be said about SPARQL endpoints.

`clld` and Linked Data: API and storage format

Publishing Linked Data can be as easy as putting a bunch of files on a web server.

- `clld` apps will be able to fall back to that, i.e. dumping the resources they serve as static files by enumerating their URL space.
- This allows for a graceful degradation of service:
  - When served from the app, resources will point to a canonical URI using the appropriate HTTP Link header.
  - These URIs will still resolve in the static-files-on-webserver scenario.
  - So when served as static files from a plain HTTP server, most things will still work

`clld` and Linked Data: the 5th star

> *Link your data to other people's data to provide context.*

While HTML provides the prime example of embedding links to provide context, for structured data and common domains RDF models are more useful.

- Again "other people" includes yourself.
- VoID is used to convey basic provenance and license information.
- Typically all statements of linguistic interest (i.e. value assignments) are linked to sources.

`clld` and Linked Data: the 5th star

- Our publication platform does spit out RDF.
- The RDF model for a particular `clld` app can be completely customized.
- But should it?
- Balance between
  - uniform access across CLLD apps and
  - semantic interoperability with existing infrastructure.
  - Is it more useful to model resources as having multiple types or provide mappings?
- Example: Model lexical data using lemon?
- Generally, in terms of user-friendliness, the problem is not a choice of RDF models but consumable formats (csv, Newick, . . .)

`clld` and Linked Data: the 5th star

- Glottolog as hub in the CLLD Linked Data cloud:
  - language catalog (linking in turn to lexvo, dbpedia, etc.), iso639-3 is often not sufficient.
  - shared bibliography
- WOLD as catalog for comparison meanings (cf. Leipzig-Jakarta list) – a *concepticon*, or an *ontology*.
- PHOIBLE may play such a role for phonological segments, e.g. as reference for transcriptions.
- filling in blanks: Identify phonological descriptions for lanfuages missing in PHOIBLE by inspecting Glottolog.
- fill in missing values in WALS for phonological features by looking up PHOIBLE.

`clld` and Linked Data: A workflow for research based on CLLD data

1. Identify suitable datasets.
2. Aggregate the data in a triple store (crawling/importing dumps).
3. Filter data in the triple store (using provenance information, etc.).
4. Export data to suitable format for analysis.

- CLLD and Linked Data will mainly play a role during aggregation of raw data.

`clld` utilities: Versioning/updating/preservation

Several models are possible:

- versioned data in database
- only current data in database, archived older versions (ZENODO)
- updates via database migration scripts (versioned together with the software)

clld utilities: SAILS archived with ZENODO

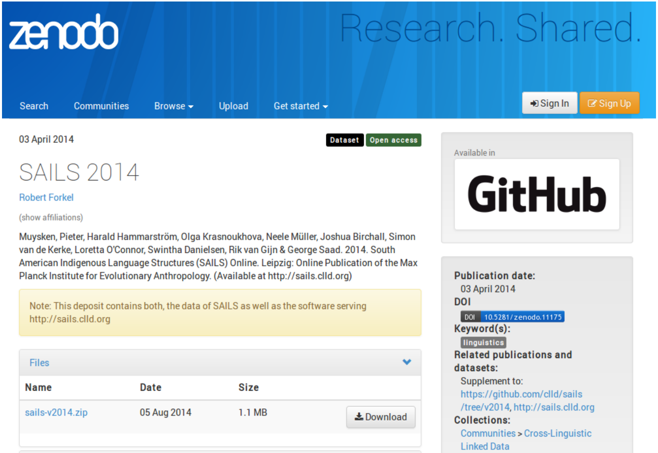


Figure 9: Archiving SAILS with ZENODO means longterm preservation and better citeability via DOI.

Standardization the Microsoft way?

- As demonstrated above, a standard software stack is useful.
- But software has a half-life of less than 10 years.
- Next step is essential: extract a **domain specific** API which can become standard.
  - Linked Data is still lacking in domain specificity.
  - Domain specific means semantic interoperability of linguistic concepts.

Towards a domain specific API: Decoupling database from visualization/analysis

- for OLAC there's OAI-PMH
- for mapping (i.e. leaflet, tilemill) there's GeoJSON
- but then there's RefLex
- and http://phonotactics.anu.edu.au/
- and the WALS Sunburst explorer
- …

`clld` databases on OLAC

OLAC Archive Metrics | Comparative Archive Metrics

(Click column headers to sort)

Archive	Overall Rating	Number of Resources	Number of Resources Online	Distinct Languages	Distinct Linguistic Subfields	Distinct Linguistic Types	Distinct DCMI Types	Average Elements Per Record	Average Encoding Schemes Per Record	Average Metadata Quality Score
Glottolog 2.3	★★★★★	7684	7684	7664	0	1	1	10.0	7.0	9.3
Ethnologue: Languages of the World	★★★★	7480	7480	7479	0	0	1	10.0	7.0	8.3
SIL Language and Culture Archives	★★★★	28448	5467	3080	0	3	5	13.2	8.3	8.9
The LINGUIST List Language Resources	★★★★	2440	0	2430	0	0	1	11.0	7.0	8.4
WALS Online	★★★★★	2621	2621	2420	0	1	1	10.0	7.0	9.3
The Rosetta Project: A Long Now Foundation Library of Human Language	★★★★	6571	6571	2365	3	3	3	18.4	7.5	8.9
WALS Online RefDB	★★★★	7157	7157	2341	7	0	1	11.5	8.3	7.1
PHOIBLE Online	★★★★★	1672	1672	1668	1	1	1	11.0	8.0	9.5
Graduate Institute of Applied Linguistics Library	★★★★	8176	394	1335	23	3	5	14.3	7.2	7.8
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	★★★★★	9266	9189	839	4	3	3	26.7	12.3	9.0

Figure 10: 3 out of the top-ten of OLAC archives by number of distinct languages are based on CLLD datasets.

Visualization: Phonotactics

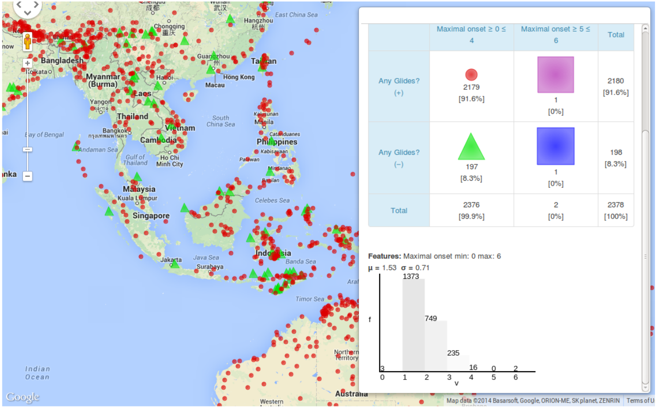


Figure 11: Configurable visualization of phonotactic features of the world's languages.

Visualization: WALS Sunburst Explorer

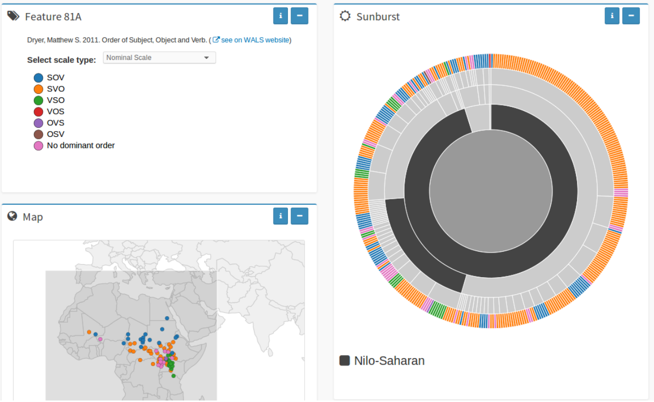
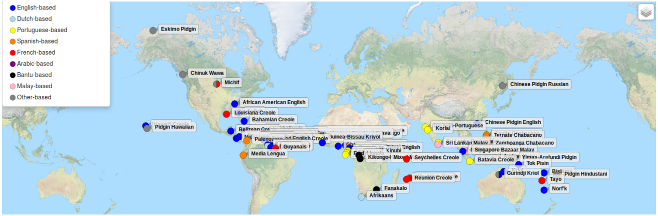


Figure 12: Combined visualization of geolocation, genealogy and coding for a WALS feature.

Semantic interoperability

- Being able to evaluate provenance data during the aggregation of a dataset is useful (e.g. in the ASJP project, some sources of wordlists are regarded as less trustworthy than others).
- Unambiguous identification of languages is required; Glottolog will help with that.
  - Being able to answer the question “which data do we have on a selected sample of languages?” as well as
  - “what sample of languages can we investigate given we need a certain selection of data (lexical, structural, etc.)?”
- For lexical data *lemon* can help to interpret the raw data, i.e. matching senses across languages (cf. Moran and Brümmer 2013).
- The requirements of statistical methods may lead to a standardisation of structural language parameters (features in the WALS sense), but we are not there yet.

Semantic interoperability: Language identification

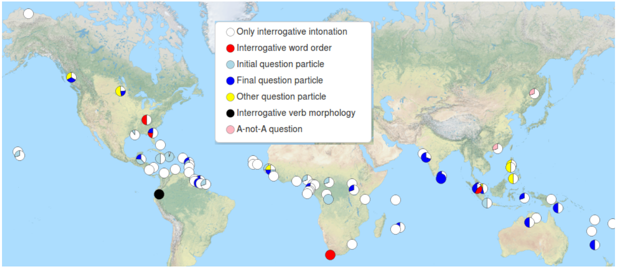


The languages described in APiCS and eWAVE show that iso639-3 is insufficient for language identification.

Semantic interoperability: Limitations

- Generally, useful data formats will be dictated by the needs of the analysis tools (e.g. phylogenetic software),
- so doing analyses directly on the RDF model can not be expected.
- Example APiCS: Interoperability of typological resources is hampered by the difficulty of cross-linguistic categories.

Semantic interoperability: APiCS and WALS



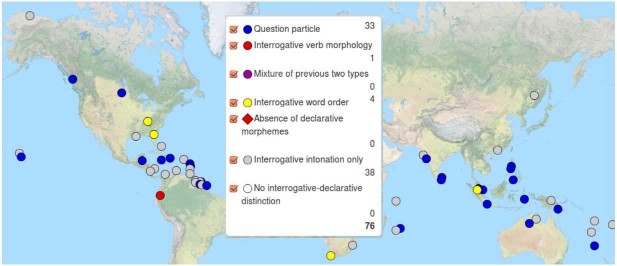


Figure 13: APiCS feature *Polar questions* – original and WALSified.

Towards a domain specific API

Roadmap:

1. ”standardize” on software
2. determine what a proper API would look like (right now!)
   - collect use cases,
   - implement prototypes,
3. specify API – maybe ontologies, maybe RDF models, maybe ling-JSON . . .

http://clld.org

Thank you!