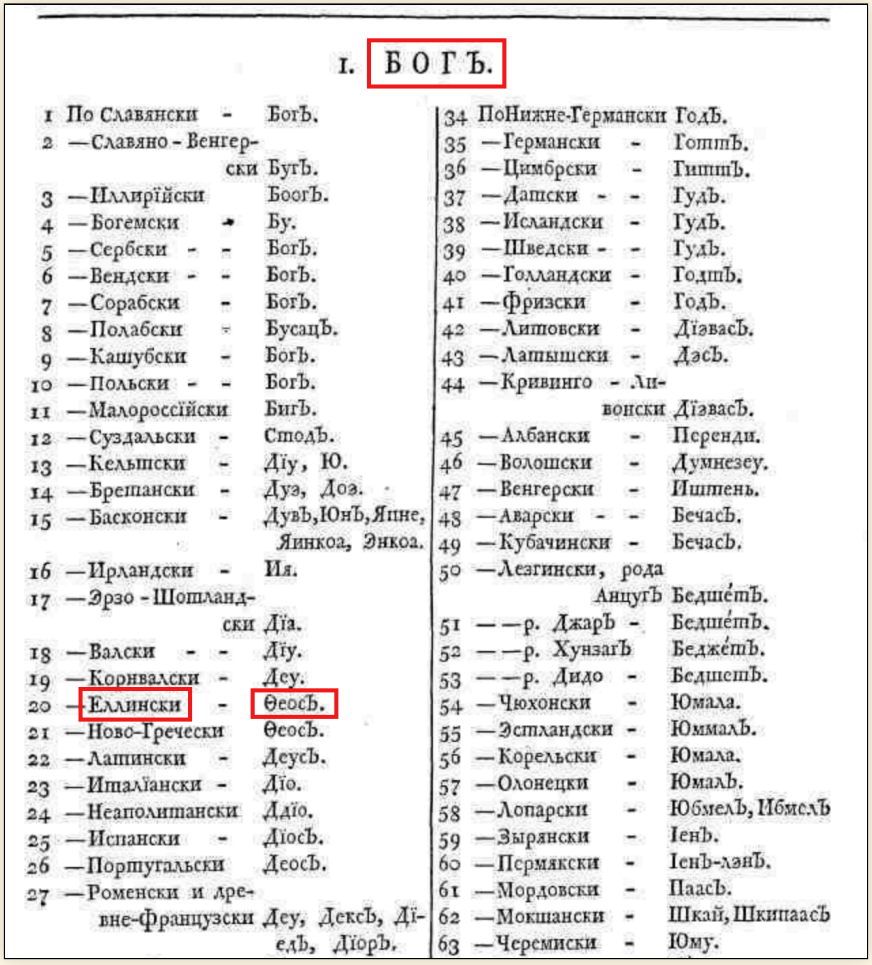
Language Comparison with CLDF

Robert Forkel
Department of Linguistic and Cultural Evolution
MPI EVA, Leipzig

Language Comparison in “Analog” Times

The "Pallas dictionary" from 1787/89



1. БОГЪ.

1	По Славянски -	БогЪ.
2	—Славяно-Венгерски	БугЪ.
3	—Иллирїйски	БоогЪ.
4	—Богемски -	Бу.
5	—Сербски - -	БогЪ.
6	—Вендски - -	БогЪ.
7	—Сорабски -	БогЪ.
8	—Подабски -	БусацЪ.
9	—Кашубски -	БогЪ.
10	—Польски - -	БогЪ.
11	—Малороссїйски	БигЪ.
12	—Суздальски -	СтодЪ.
13	—Кельтски -	Дїу, Ю.
14	—Бретански -	Дуэ, Доэ.
15	—Басконски -	ДувЪ,ЮнЪ,Япне, Яинкоа, Энкоа.
16	—Ирландски -	Ия.
17	—Эрзо-Шотланд-ски	Дїа.
18	—Валски - -	Дїу.
19	—Корнвалски -	Деу.
20	—Еллински -	ѲеосЪ.
21	—Ново-Гречески	ѲеосЪ.
22	—Латински -	ДеусЪ.
23	—Италїански -	Дїо.
24	—Неаполитански	Ддїо.
25	—Испански -	ДїосЪ.
26	—Португальски	ДеосЪ.
27	—Роменски и древне-французски	Деу, ДексЪ, Дїедъ, ДїорЪ.
34	ПоНижне-Германски	ГодЪ.
35	—Германски -	ГоттЪ.
36	—Цимбрски -	ГиттЪ.
37	—Датски - -	ГудЪ.
38	—Исландски -	ГудЪ.
39	—Шведски - -	ГудЪ.
40	—Голландски -	ГодтЪ.
41	—Фризски -	ГодЪ.
42	—Литовски -	ДїэвасЪ.
43	—Латышски -	ДэсЪ.
44	—Кривинго - Лївонски	ДїэвасЪ.
45	—Албански -	Перенди.
46	—Волошски -	Думнезеу.
47	—Венгерски -	Иштень.
48	—Аварски - -	БечасЪ.
49	—Кубачински -	БечасЪ.
50	—Лезгински, рода Анцуг	БедшéтЪ.
51	——р. ДжарЪ -	БедшéтЪ.
52	——р. ХунзагЪ	БеджéтЪ.
53	——р. Дидо -	БедшетЪ.
54	—Чюхонски -	Юмала.
55	—Эстландски -	ЮммалЪ.
56	—Корельски -	Юмала.
57	—Олонецки -	ЮмалЪ.
58	—Лопарски -	ЮбмелЪ, Ибмелъ
59	—Зырянски -	ІенЪ.
60	—Пермякски -	ІенЪ-лэнЪ.
61	—Мордовски -	ПаасЪ.
62	—Мокшански -	Шкай, Шкипаасъ
63	—Черемиски -	Юму.

1. БОГЪ.

1	По Славянски -	Богъ.
2	—Славяно-Венгерски	Бугъ.
3	—Илліріїйски	Боогъ.
4	—Богемски -	Бу.
34	ПоНижне-Германски	Годъ.
35	—Германски -	Готтъ.
36	—Цимбрски -	Гиттъ.
37	—Дашски - -	Гудъ.
38	—Исландски -	Гудъ.

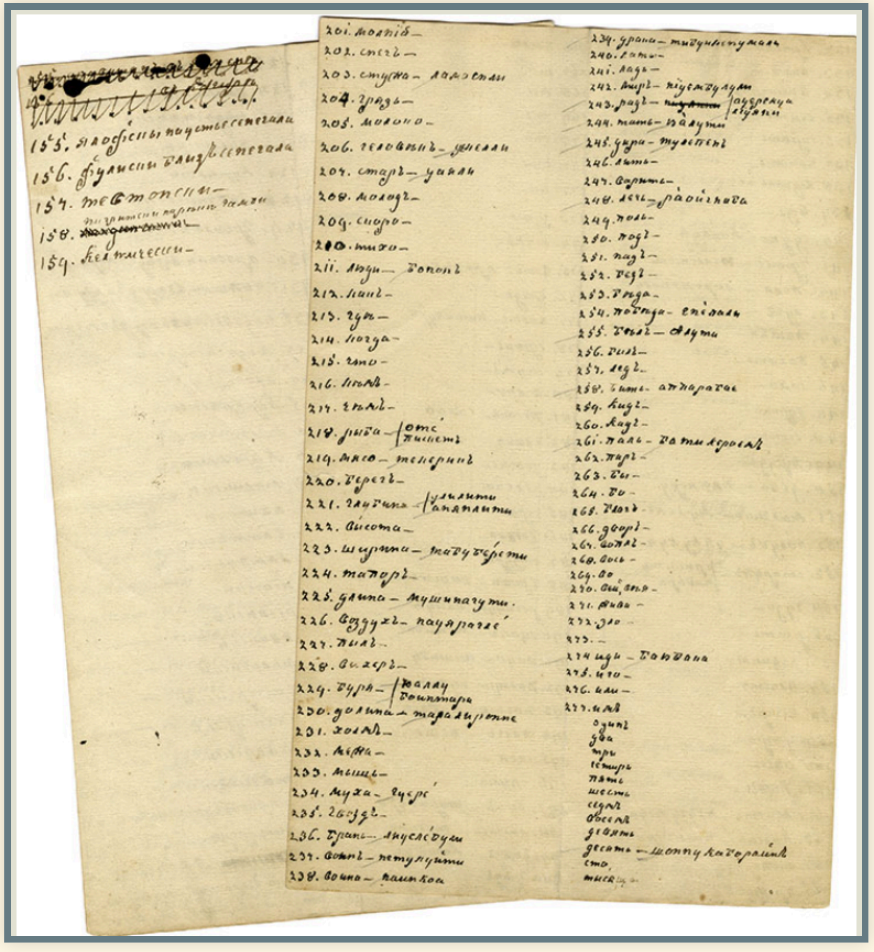
God (60).

Number in General List.	AGGLUTINATIVE NON-INDIAN LANGUAGES.	
	Japanese . .	*Kami*
	Ainu . .	*Kamui*
	Korean . .	*Hànânim*
	Turkī . .	*Tīngrī, Auγān, Bīrī-bār*
	Manchu . .	*Àbkà-i ejen* (Lord of heaven)
	Mongolian . .	*Tegri*
	Saukpà . .	...
	Basque . .	*Jaungoiko, Jainko, Dzipo*

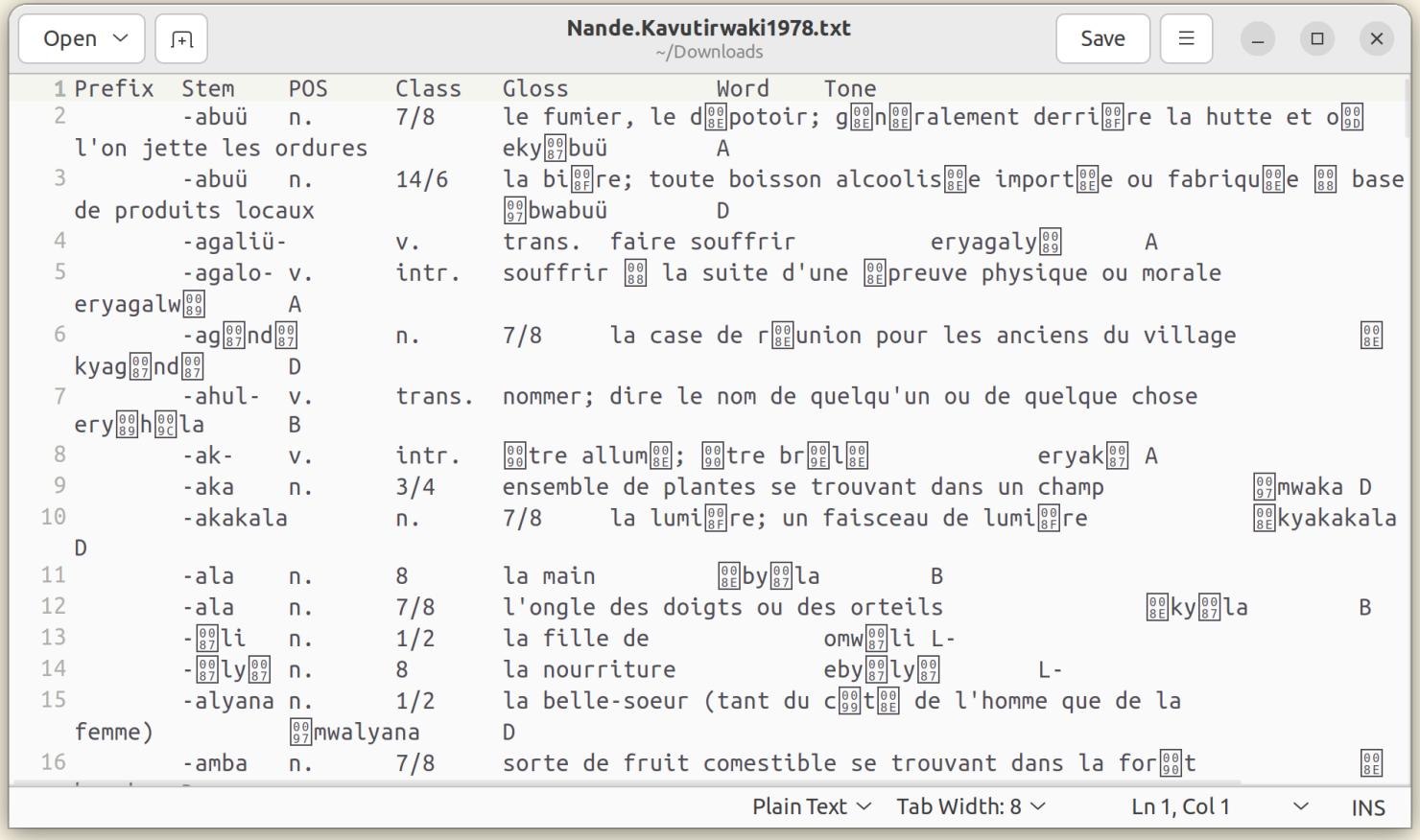
Number in General List.	KAREN LANGUAGES.	
35.	Pwo, literary .	*–Ka –s'ā /ywa*
	„ Bassein }	*Yuà, Chai'*
	„ Maulmein }	
36.	Taungθu . .	*P'rà*
34.	Sgâ, literary .	*–Ka \sa*
	„ spoken .	*G'sà yuà*
32.	Bwè . .	*Bi je uà*
41a.	Wewaw . .	*G'zà δwà*

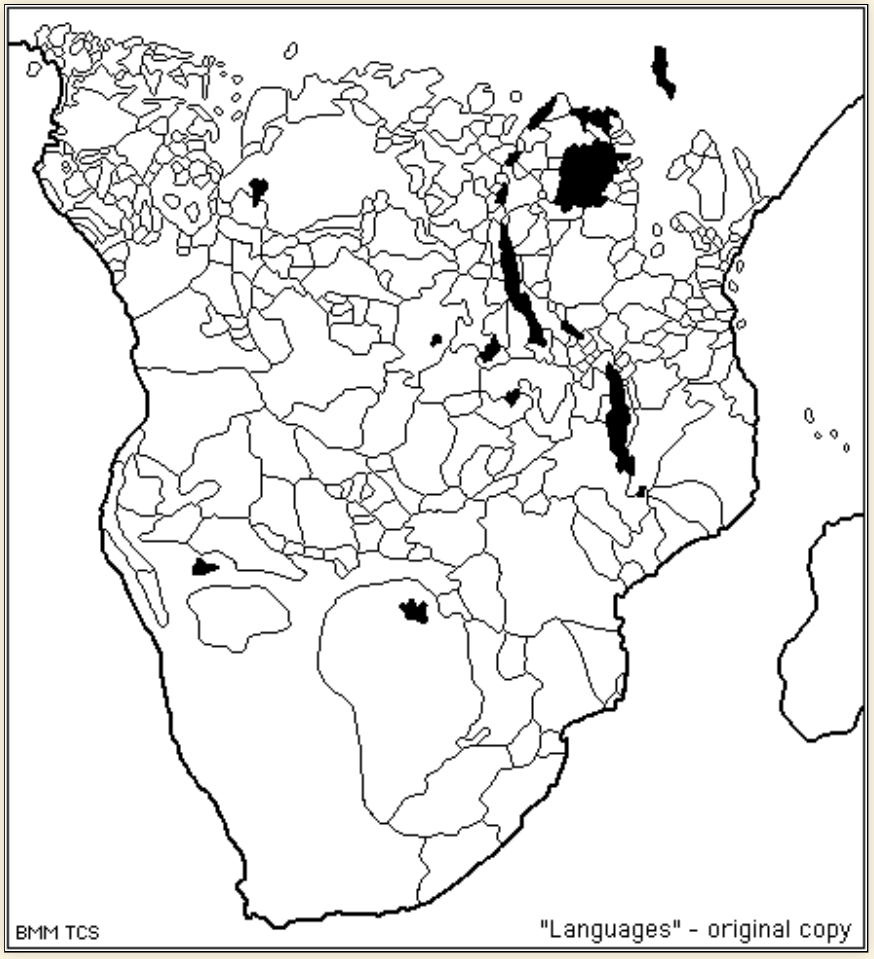
Number in General List.		
123.	Abor . .	...
124.	Miri . . .	...
125.	Dafla . .	*Ūī*
126.	Mišmi, Digārū .	*Nin-ya* (?)
	Mijū . .	*Se-lap*
	Lolo-Mos'o Group.	
	Si-hia . .	*K-ñum, g-ñum* (Heaven)
273.	Lolo, /Ñ[1] . .	*/M[u] \se /p'a*
	A-hi . .	*Mu° sa° p'o°*

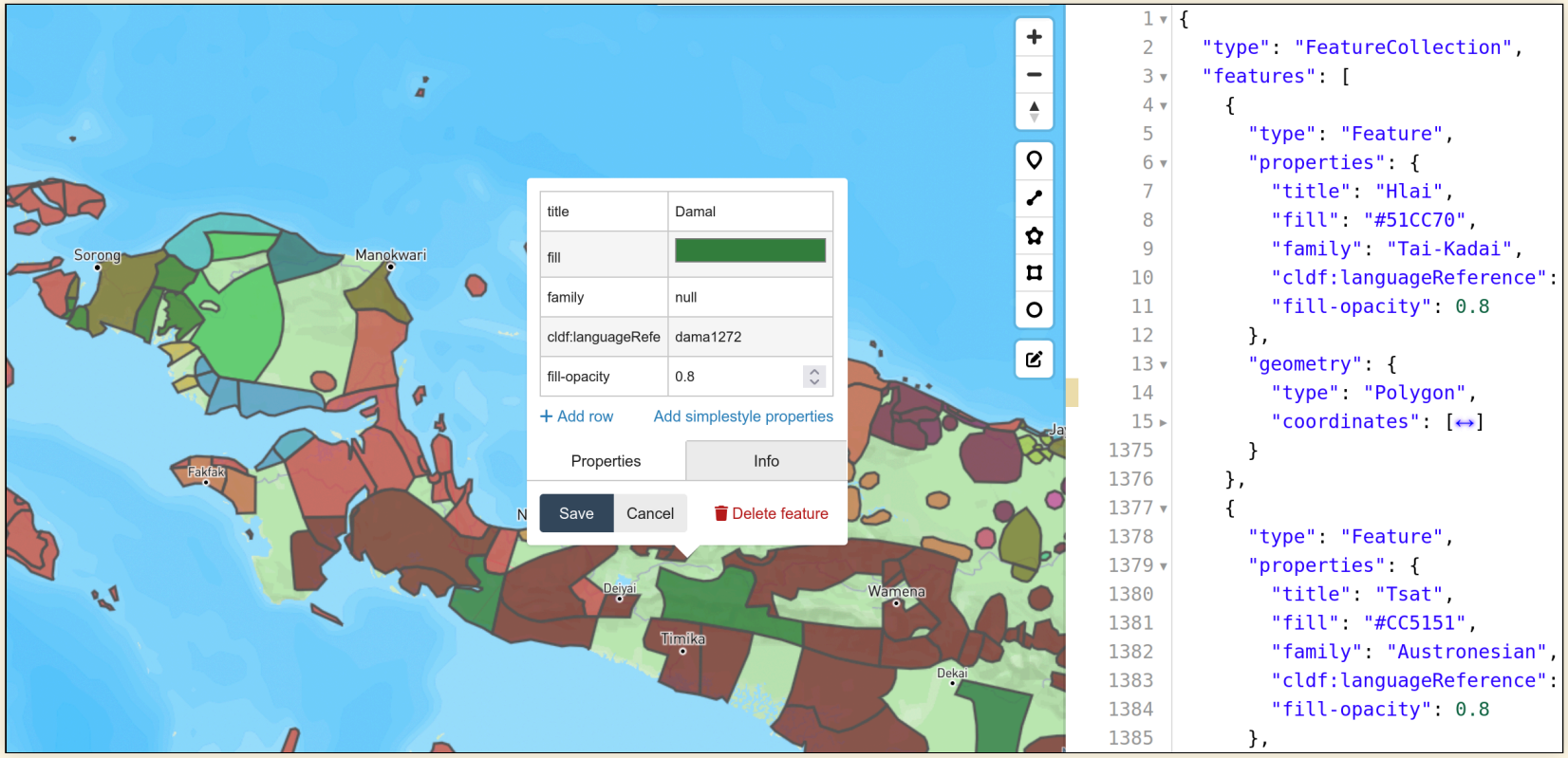
Data curation
requires "comparability", too.



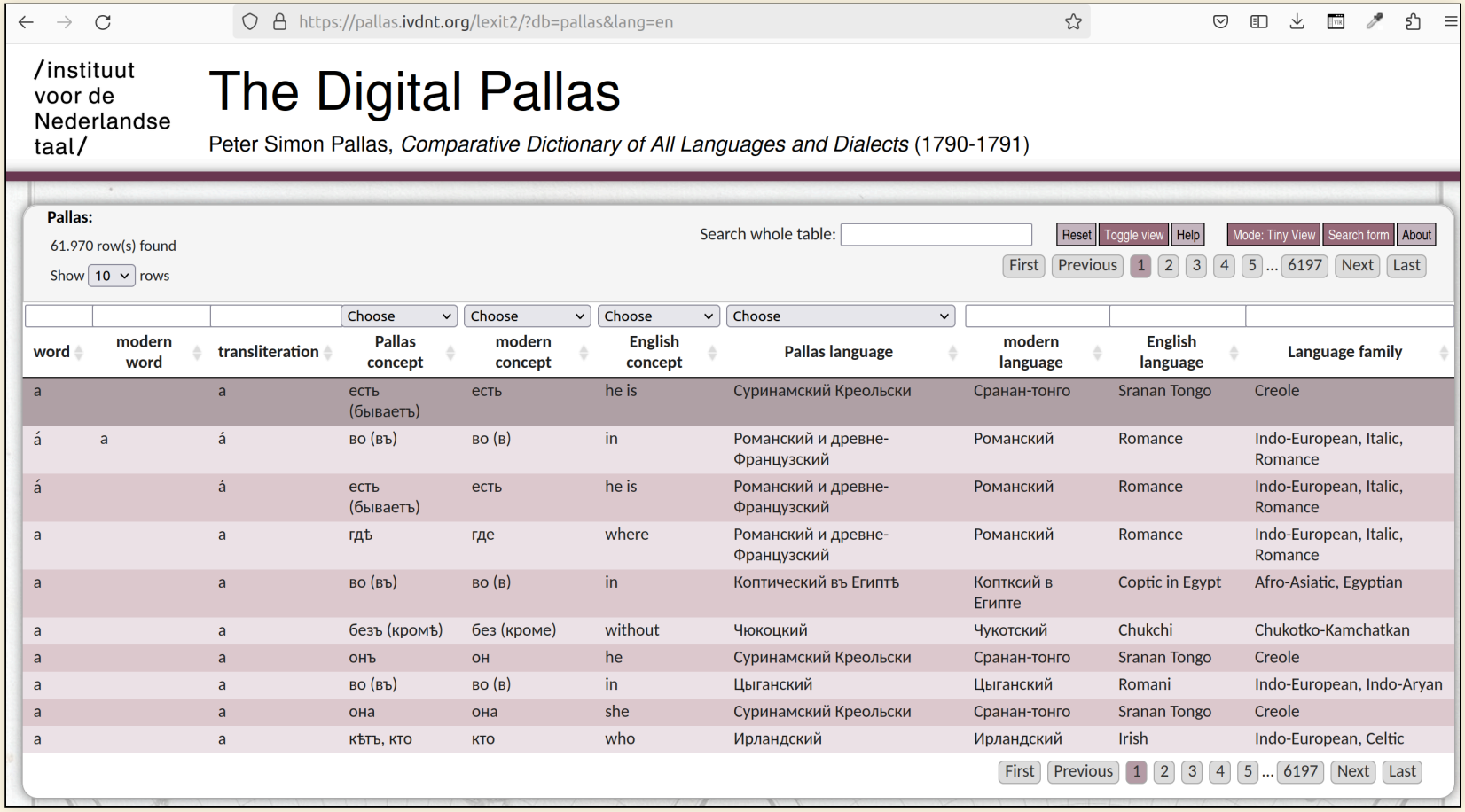
Going “Digital”


Open
Nande.Kavutirwaki1978.txt
~/Downloads
Save
Prefix Stem POS Class Gloss Word Tone
-abuü n. 7/8 le fumier, le d potoir; g n ralement derri re la hutte et o l'on jette les ordures eky buü A
-abuü n. 14/6 la bi re; toute boisson alcoolis e import e ou fabriqu e base de produits locaux bwabuü D
-agaliü- v. trans. faire souffrir eryagaly A
-agalo- v. intr. souffrir la suite d'une preuve physique ou morale eryagalw A
-ag nd n. 7/8 la case de r union pour les anciens du village kyag nd D
-ahul- v. trans. nommer; dire le nom de quelqu'un ou de quelque chose ery h la B
-ak- v. intr. tre allum ; tre br l eryak A
-aka n. 3/4 ensemble de plantes se trouvant dans un champ mwaka D
-akakala n. 7/8 la lumi re; un faisceau de lumi re kyakakala D
-ala n. 8 la main by la B
-ala n. 7/8 l'ongle des doigts ou des orteils ky la B
- li n. 1/2 la fille de omw li L-
- ly n. 8 la nourriture eby ly L-
-alyana n. 1/2 la belle-soeur (tant du c t de l'homme que de la femme) mwalyana D
-amba n. 7/8 sorte de fruit comestible se trouvant dans la for t
Plain Text
Tab Width: 8
Ln 1, Col 1
INS


BMM TCS
"Languages" - original copy


Sorong
Manokwari
Fakfak
Deiyai
Timika
Wamena
Dekai
title
Damal
fill
family
null
cldf:languageRefe
dama1272
fill-opacity
0.8
+ Add row
Add simplestyle properties
Properties
Info
Save
Cancel
Delete feature
{
"type": "FeatureCollection",
"features": [
{
"type": "Feature",
"properties": {
"title": "Hlai",
"fill": "#51CC70",
"family": "Tai-Kadai",
"cldf:languageReference":
"fill-opacity": 0.8
},
"geometry": {
"type": "Polygon",
"coordinates": [↔]
}
},
{
"type": "Feature",
"properties": {
"title": "Tsat",
"fill": "#CC5151",
"family": "Austronesian",
"cldf:languageReference":
"fill-opacity": 0.8
},

Going On-Line


https://pallas.ivdnt.org/lexit2/?db=pallas&lang=en
/instituut voor de Nederlandse taal/
The Digital Pallas
Peter Simon Pallas, Comparative Dictionary of All Languages and Dialects (1790-1791)
Pallas:
61.970 row(s) found
Show 10 rows
Search whole table:
Reset
Toggle view
Help
Mode: Tiny View
Search form
About
First
Previous
1
2
3
4
5
...
6197
Next
Last
Choose
word
modern word
transliteration
Pallas concept
modern concept
English concept
Pallas language
modern language
English language
Language family
а | | а | есть (бываетъ) | есть | he is | Суринамский Креольски | Сранан-тонго | Sranan Tongo | Creole
á | а | á | во (въ) | во (в) | in | Романский и древне-Французский | Романский | Romance | Indo-European, Italic, Romance
á | | á | есть (бываетъ) | есть | he is | Романский и древне-Французский | Романский | Romance | Indo-European, Italic, Romance
а | | а | гдѣ | где | where | Романский и древне-Французский | Романский | Romance | Indo-European, Italic, Romance
а | | а | во (въ) | во (в) | in | Коптический въ Египтѣ | Коптксий в Египте | Coptic in Egypt | Afro-Asiatic, Egyptian
а | | а | безъ (кромѣ) | без (кроме) | without | Чюкоцкий | Чукотский | Chukchi | Chukotko-Kamchatkan
а | | а | онъ | он | he | Суринамский Креольски | Сранан-тонго | Sranan Tongo | Creole
а | | а | во (въ) | во (в) | in | Цыганский | Цыганский | Romani | Indo-European, Indo-Aryan
а | | а | она | она | she | Суринамский Креольски | Сранан-тонго | Sranan Tongo | Creole
а | | а | кѣтъ, кто | кто | who | Ирландский | Ирландский | Irish | Indo-European, Celtic
First
Previous
1
2
3
4
5
...
6197
Next
Last

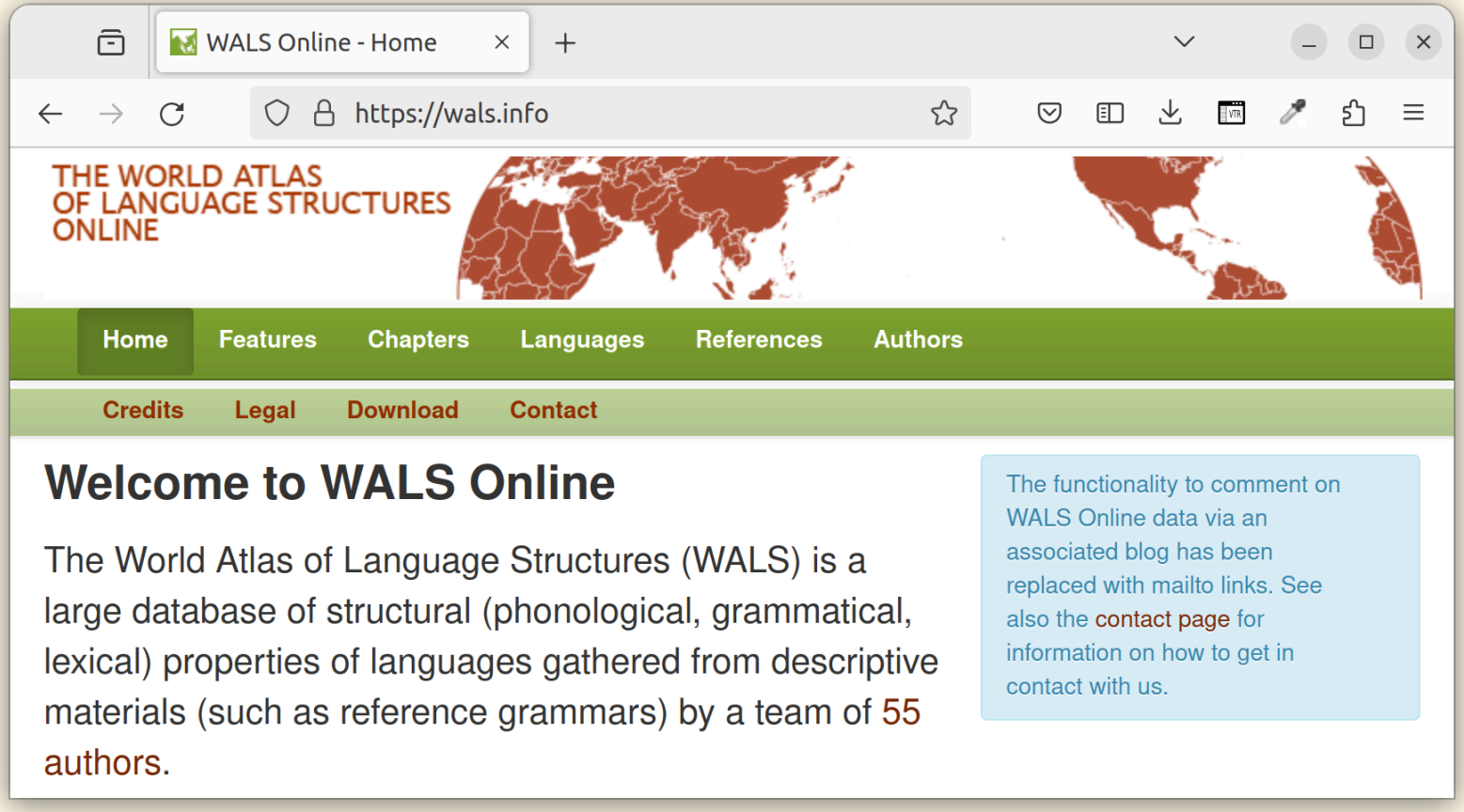
The Bantuists' Manifesto

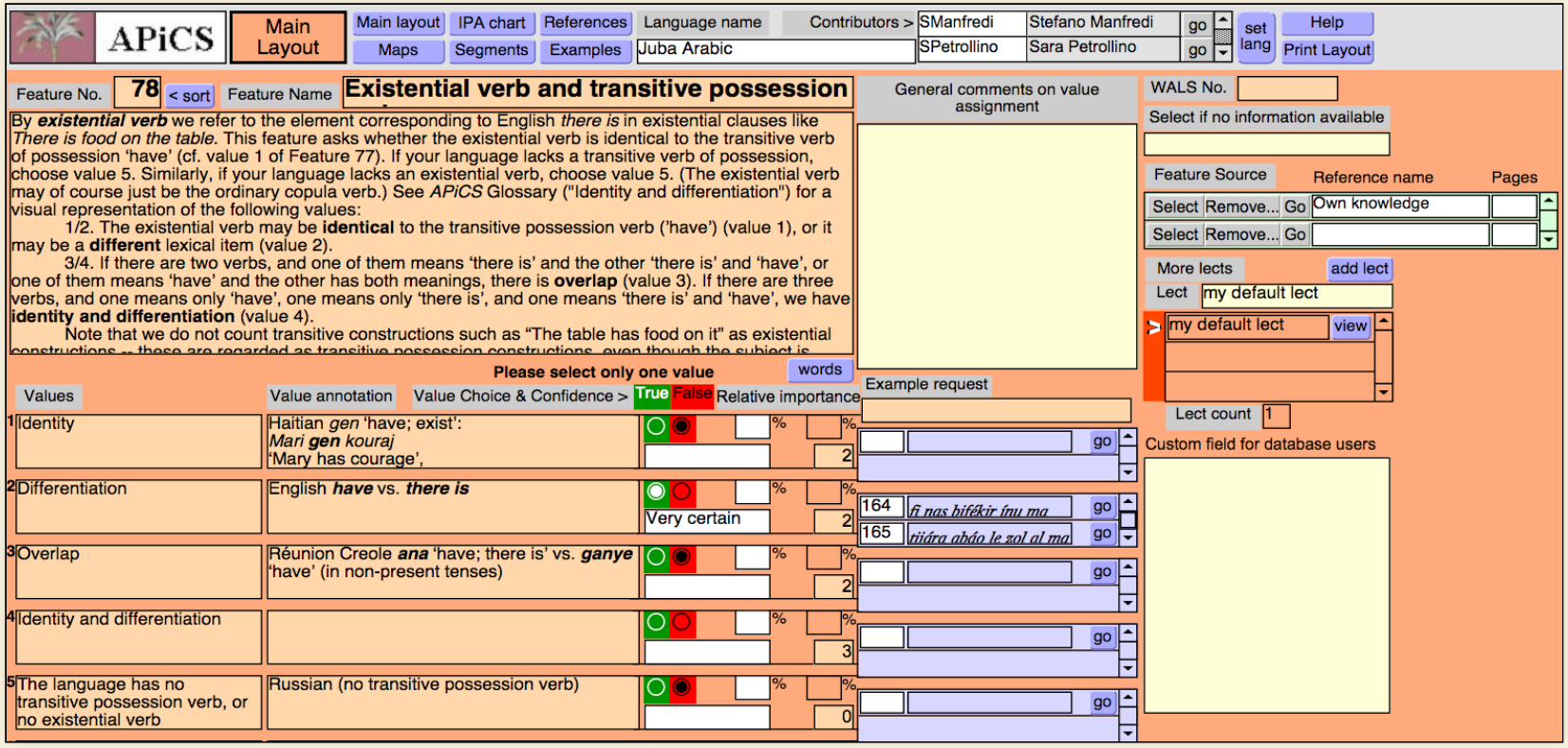
We, the undersigned researchers in African languages, therefore agree:

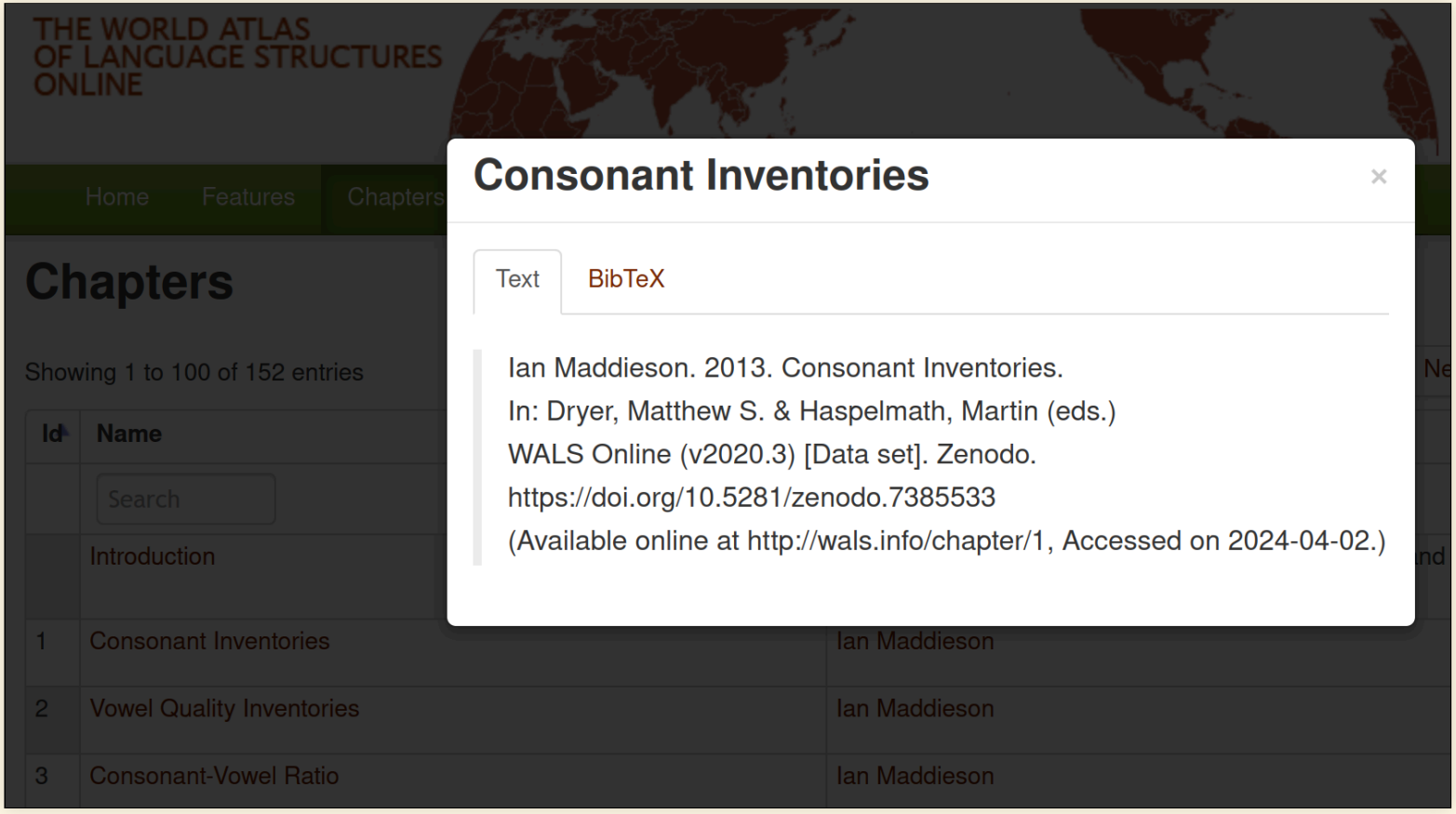
- to provide our data in electronic form for use by others […];
- to acknowledge the use we make of the data provided to use by others […];
- to hold others harmless for any errors or misuse of the data;
- and to make no commercial use of such data […].

The Bantuists’ Manifesto

A CC-BY-NC license predating the Creative Commons initiative by 5 years - and still a rarity for linguistic online publications like the “Digital Pallas” …


WALS Online - Home
https://wals.info
THE WORLD ATLAS
OF LANGUAGE STRUCTURES
ONLINE
Home
Features
Chapters
Languages
References
Authors
Credits
Legal
Download
Contact
Welcome to WALS Online
The World Atlas of Language Structures (WALS) is a large database of structural (phonological, grammatical, lexical) properties of languages gathered from descriptive materials (such as reference grammars) by a team of 55 authors.
The functionality to comment on WALS Online data via an associated blog has been replaced with mailto links. See also the contact page for information on how to get in contact with us.


APiCS
Main Layout
Main layout
IPA chart
References
Language name
Contributors >
SManfredi
Stefano Manfredi
go
Maps
Segments
Examples
Juba Arabic
SPetrollino
Sara Petrollino
go
set lang
Help
Print Layout
Feature No. 78 < sort Feature Name Existential verb and transitive possession
General comments on value assignment
WALS No.
Select if no information available
Feature Source
Reference name
Pages
Select Remove... Go Own knowledge
Select Remove... Go
More lects
add lect
Lect my default lect
my default lect view
By existential verb we refer to the element corresponding to English there is in existential clauses like There is food on the table. This feature asks whether the existential verb is identical to the transitive verb of possession 'have' (cf. value 1 of Feature 77). If your language lacks a transitive verb of possession, choose value 5. Similarly, if your language lacks an existential verb, choose value 5. (The existential verb may of course just be the ordinary copula verb.) See APiCS Glossary ("Identity and differentiation") for a visual representation of the following values:
1/2. The existential verb may be identical to the transitive possession verb ('have') (value 1), or it may be a different lexical item (value 2).
3/4. If there are two verbs, and one of them means 'there is' and the other 'there is' and 'have', or one of them means 'have' and the other has both meanings, there is overlap (value 3). If there are three verbs, and one means only 'have', one means only 'there is', and one means 'there is' and 'have', we have identity and differentiation (value 4).
Note that we do not count transitive constructions such as "The table has food on it" as existential
Please select only one value
words
Example request
Lect count 1
Custom field for database users
Values
Value annotation
Value Choice & Confidence >
True
False
Relative importance
1 Identity
Haitian gen 'have; exist':
Mari gen kouraj
'Mary has courage',
2
2 Differentiation
English have vs. there is
Very certain
2
164 fi nas bifékir ínu ma go
165 tiiára abáo le zol al ma go
3 Overlap
Réunion Creole ana 'have; there is' vs. ganye 'have' (in non-present tenses)
2
4 Identity and differentiation
3
5 The language has no transitive possession verb, or no existential verb
Russian (no transitive possession verb)
0

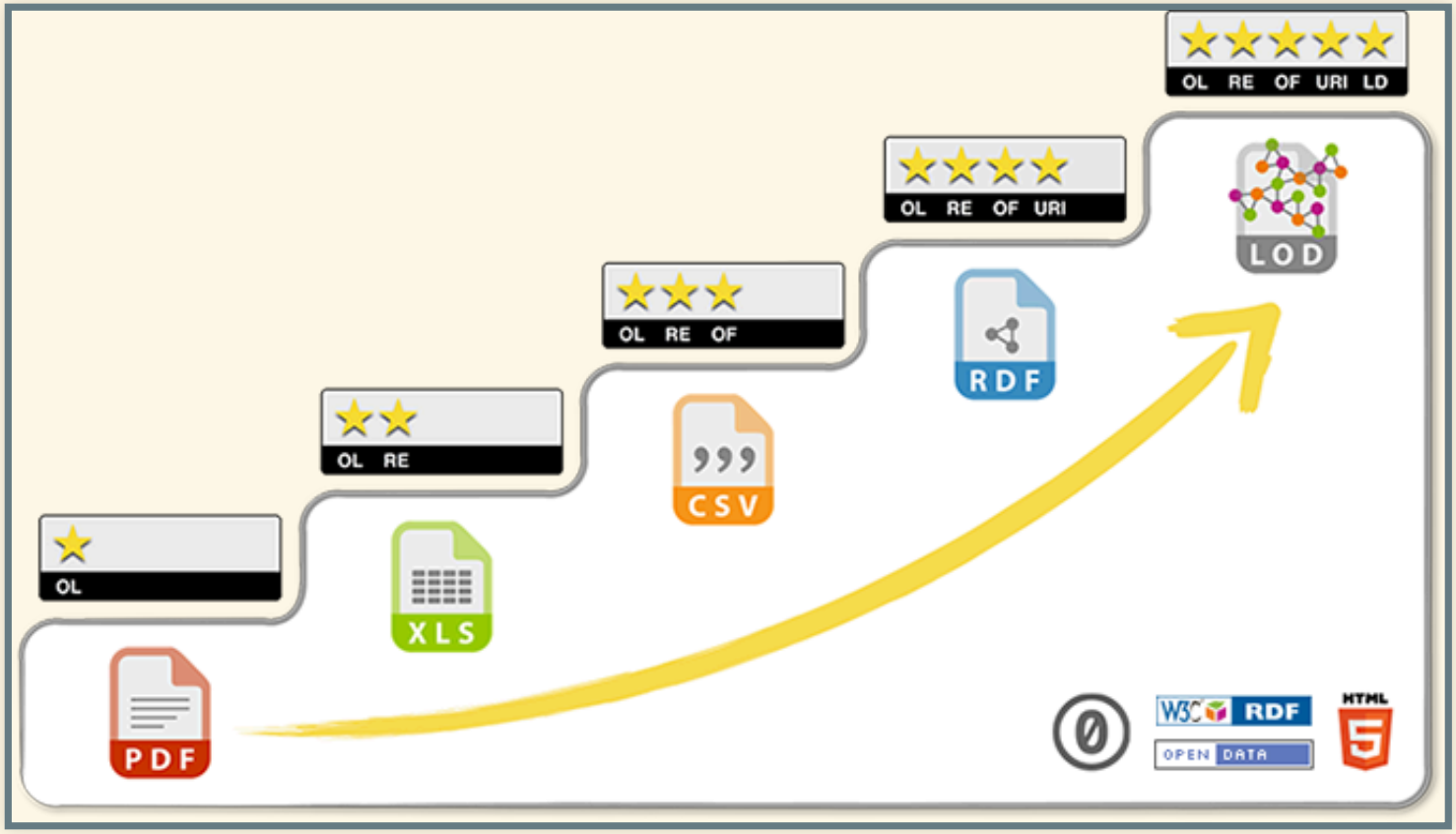

THE WORLD ATLAS
OF LANGUAGE STRUCTURES
ONLINE
Home
Features
Chapters
Chapters
Showing 1 to 100 of 152 entries
Id
Name
Search
Introduction
1
Consonant Inventories
Ian Maddieson
2
Vowel Quality Inventories
Ian Maddieson
3
Consonant-Vowel Ratio
Ian Maddieson
Consonant Inventories
Text
BibTeX
Ian Maddieson. 2013. Consonant Inventories.
In: Dryer, Matthew S. & Haspelmath, Martin (eds.)
WALS Online (v2020.3) [Data set]. Zenodo.
https://doi.org/10.5281/zenodo.7385533
(Available online at http://wals.info/chapter/1, Accessed on 2024-04-02.)

The Heyday of Linguistic Web Apps

The Semantic Web

I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web […]. A "Semantic Web", which makes this possible, has yet to emerge, but when it does, the day-to-day mechanisms of […] our daily lives will be handled by machines talking to machines. The "intelligent agents" people have touted for ages will finally materialize

Tim Berners-Lee, 1998


OL
OL RE
OL RE OF
OL RE OF URI
OL RE OF URI LD
PDF
XLS
CSV
RDF
LOD
W3C RDF
OPEN DATA
HTML
5

Cool URIs don't change

What makes a cool URI?
A cool URI is one which does not change.
What sorts of URI change?
URIs don't change: people change them.

Ruby On Rails In 60 Minutes

60

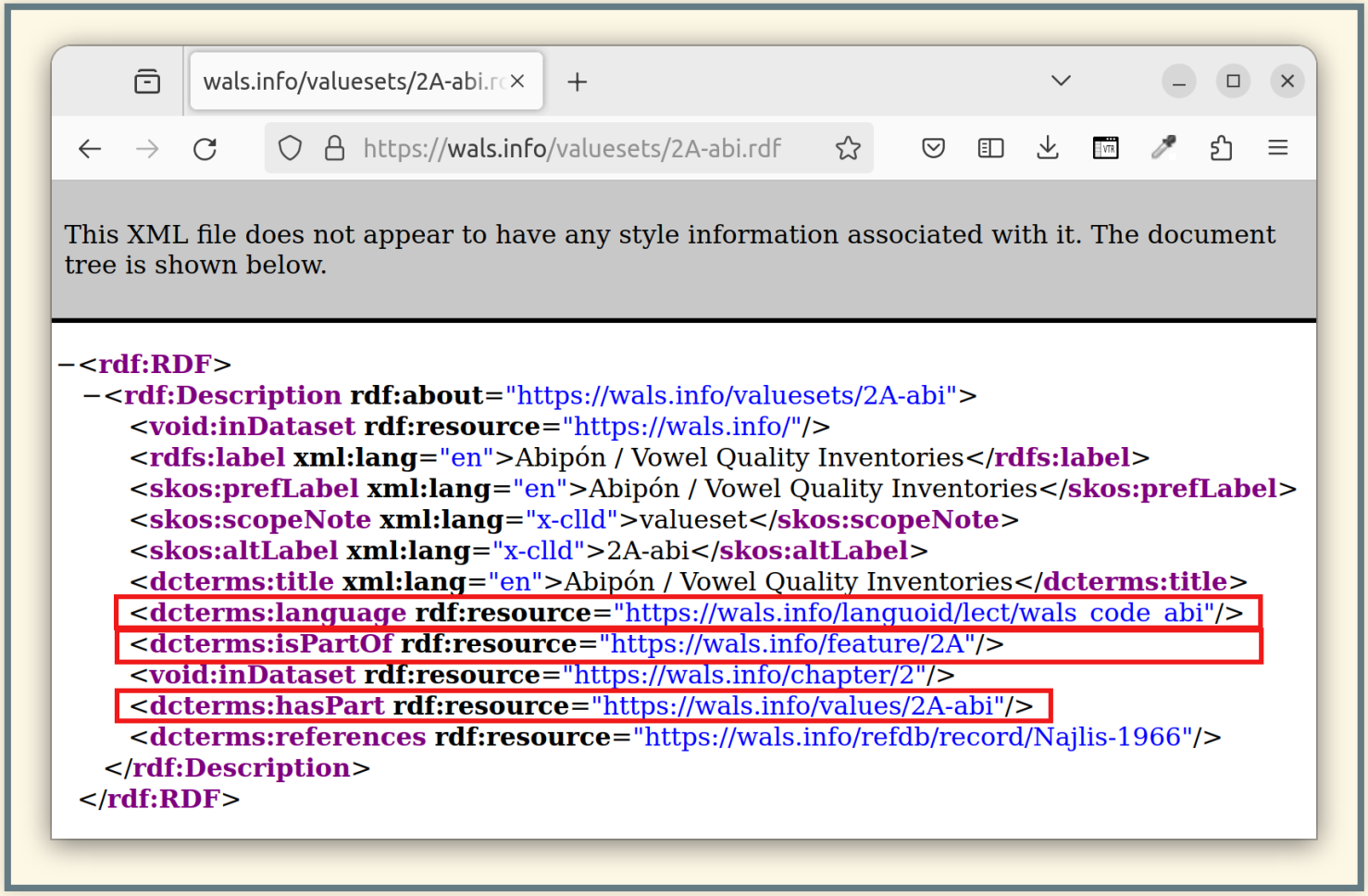
RUBY

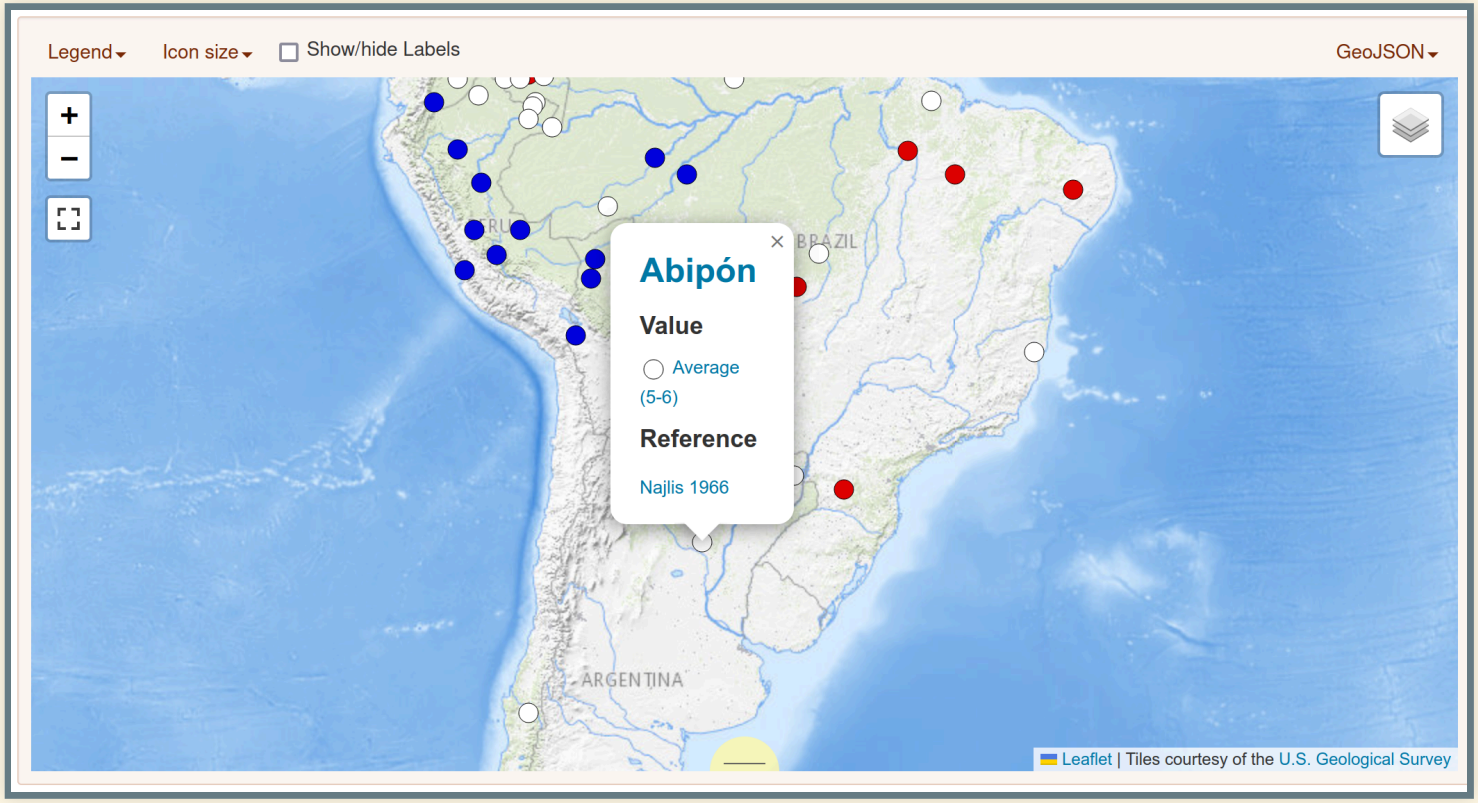
RAILS

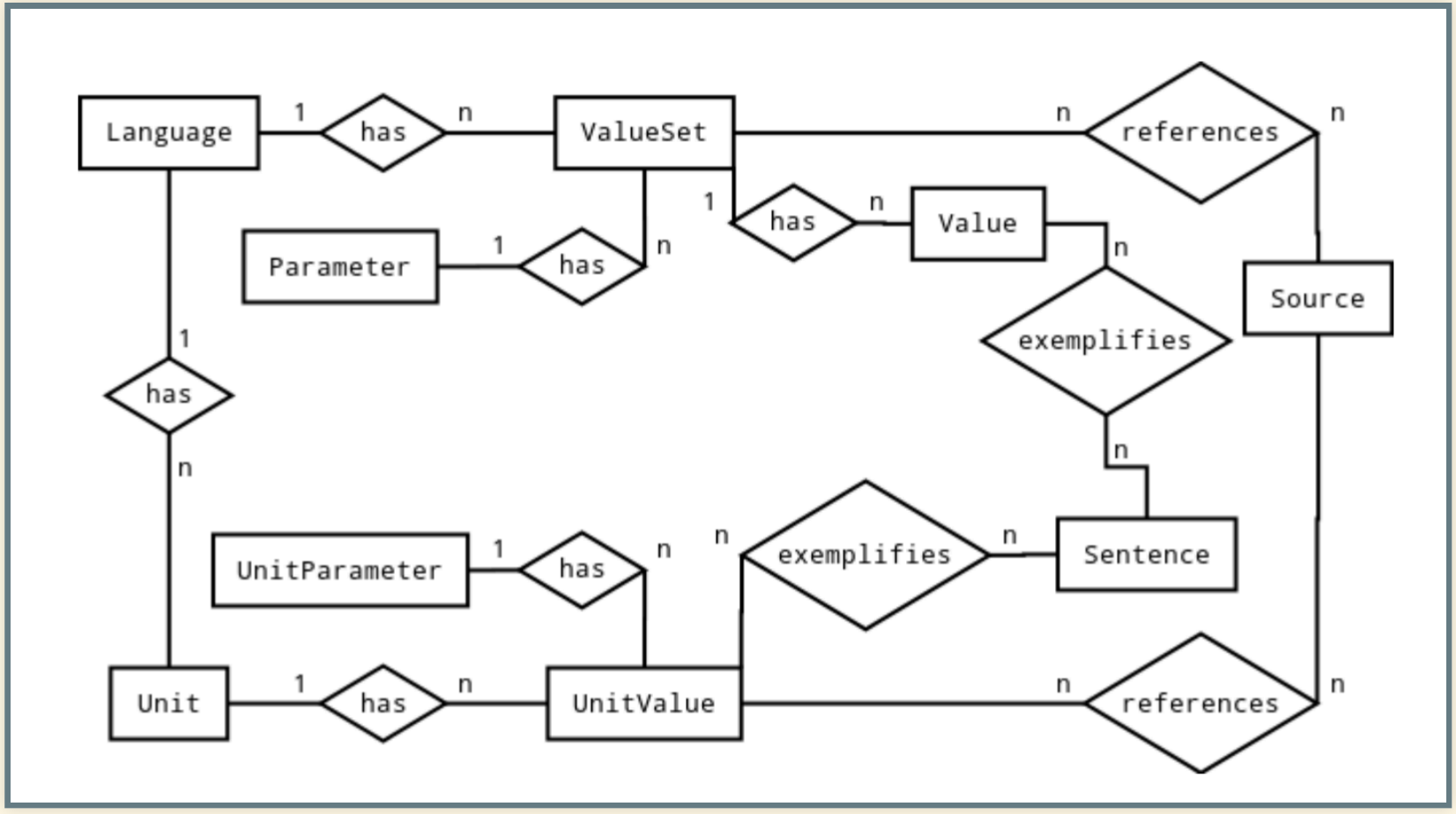
MIN

This lead to projects like

- **GOLD**, the “General Ontology for Linguistic Description” - minting URIs for Linguistic concepts
- **LLOD** - the “Linguistic Linked Open Data” initiative
- **CLLD** - the Cross-Linguistic Linked Data project


wals.info/valuesets/2A-abi.r
https://wals.info/valuesets/2A-abi.rdf
This XML file does not appear to have any style information associated with it. The document tree is shown below.
−<rdf:RDF>
−<rdf:Description rdf:about="https://wals.info/valuesets/2A-abi">
<void:inDataset rdf:resource="https://wals.info/"/>
<rdfs:label xml:lang="en">Abipón / Vowel Quality Inventories</rdfs:label>
<skos:prefLabel xml:lang="en">Abipón / Vowel Quality Inventories</skos:prefLabel>
<skos:scopeNote xml:lang="x-clld">valueset</skos:scopeNote>
<skos:altLabel xml:lang="x-clld">2A-abi</skos:altLabel>
<dcterms:title xml:lang="en">Abipón / Vowel Quality Inventories</dcterms:title>
<dcterms:language rdf:resource="https://wals.info/languoid/lect/wals_code_abi"/>
<dcterms:isPartOf rdf:resource="https://wals.info/feature/2A"/>
<void:inDataset rdf:resource="https://wals.info/chapter/2"/>
<dcterms:hasPart rdf:resource="https://wals.info/values/2A-abi"/>
<dcterms:references rdf:resource="https://wals.info/refdb/record/Najlis-1966"/>
</rdf:Description>
</rdf:RDF>


Legend
Icon size
Show/hide Labels
GeoJSON
+
−
BRAZIL
ARGENTINA
Abipón
Value
Average (5-6)
Reference
Najlis 1966
Leaflet | Tiles courtesy of the U.S. Geological Survey

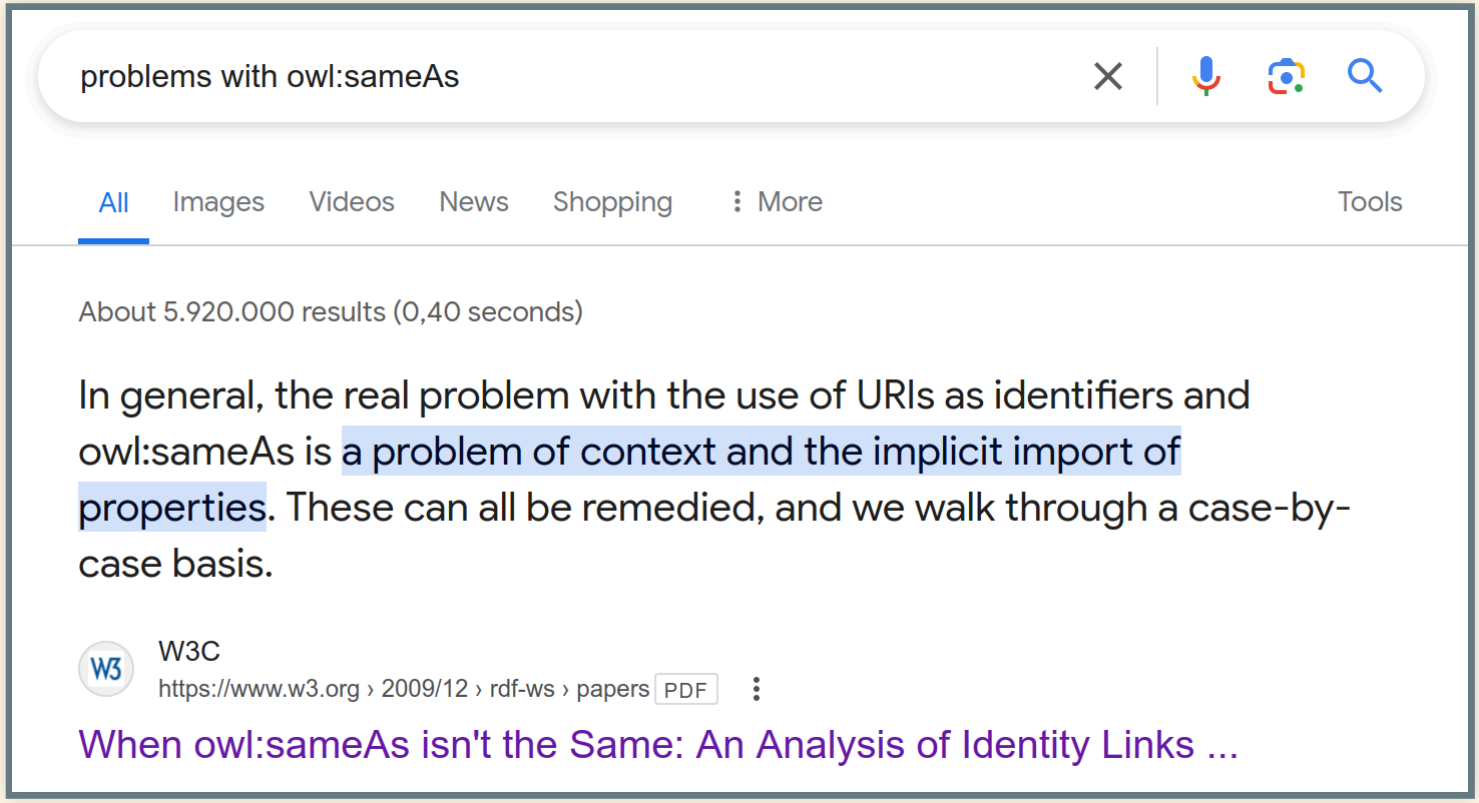

Language
1
has
n
ValueSet
n
references
n
Parameter
1
has
n
1
has
n
Value
n
Source
exemplifies
1
has
n
n
UnitParameter
1
has
n
n
exemplifies
n
Sentence
Unit
1
has
n
UnitValue
n
references
n

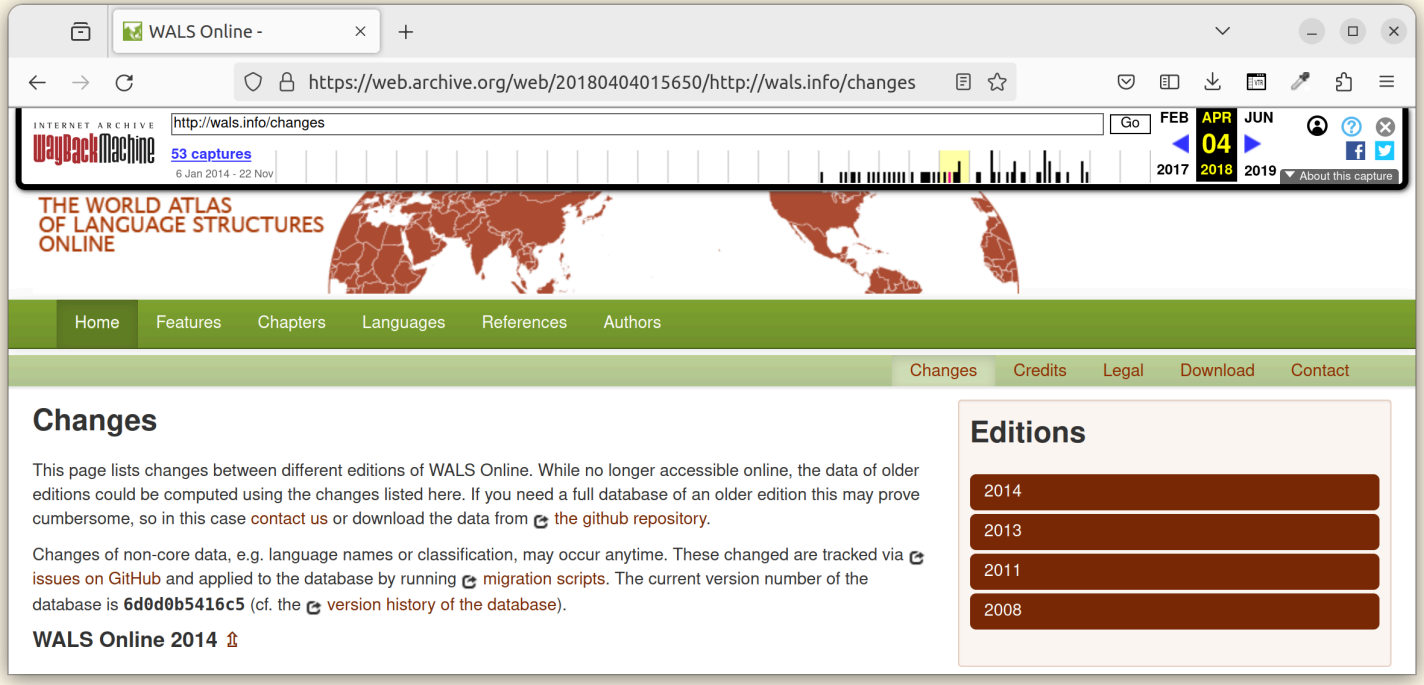
- *As demonstrated [by the `clld` framework], a standard software stack is*
  useful.
- *But software has a half-life of less than 10 years.*
- ***Next step is essential: extract a domain specific API which can become standard.***

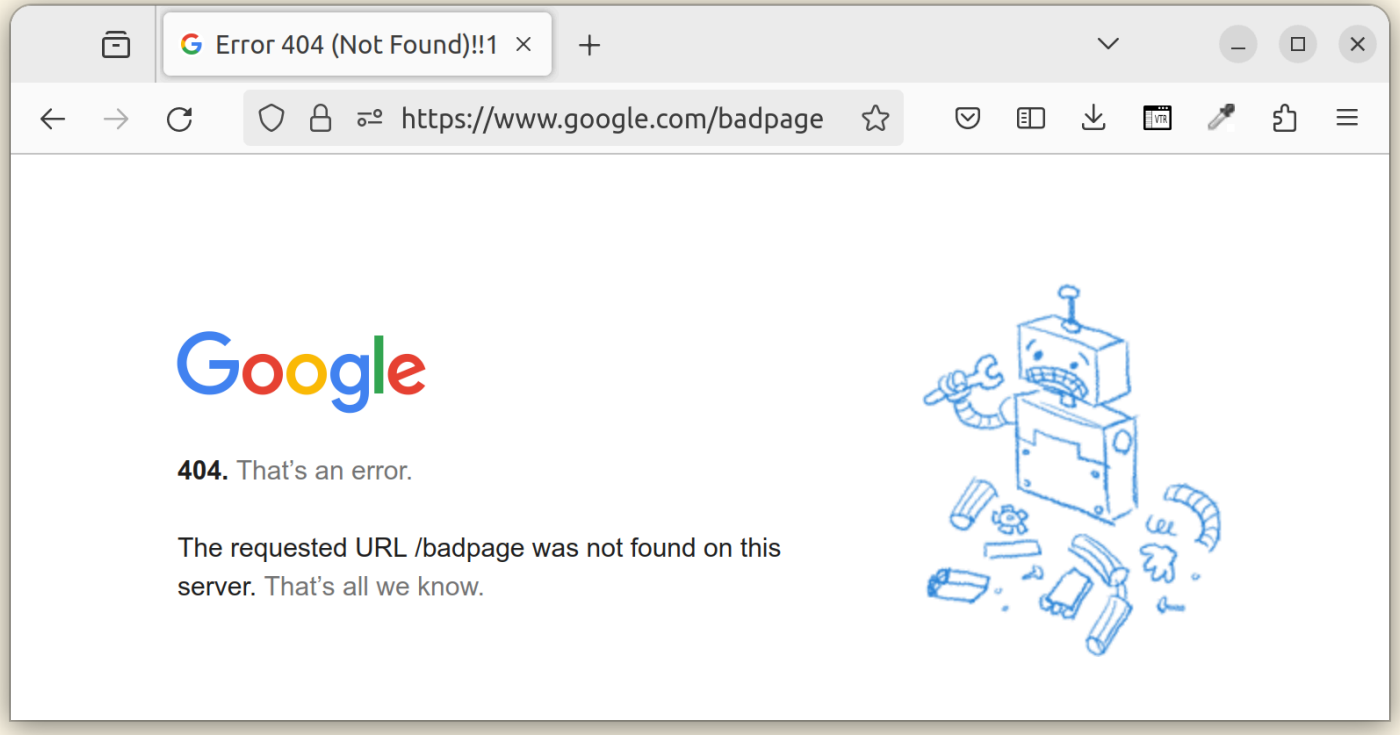
Forkel & Bank, 2014

Cracks in the Facade

- Inconsistency
- Versioning
- Maintainability


problems with owl:sameAs
All
Images
Videos
News
Shopping
More
Tools
About 5.920.000 results (0,40 seconds)
In general, the real problem with the use of URIs as identifiers and owl:sameAs is a problem of context and the implicit import of properties. These can all be remedied, and we walk through a case-by-case basis.
W3C
https://www.w3.org › 2009/12 › rdf-ws › papers PDF
When owl:sameAs isn't the Same: An Analysis of Identity Links ...


WALS Online -
https://web.archive.org/web/20180404015650/http://wals.info/changes
http://wals.info/changes
Go
53 captures
6 Jan 2014 - 22 Nov
FEB
APR
JUN
04
2017
2018
2019
About this capture
THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE
Home
Features
Chapters
Languages
References
Authors
Changes
Credits
Legal
Download
Contact
Changes
This page lists changes between different editions of WALS Online. While no longer accessible online, the data of older editions could be computed using the changes listed here. If you need a full database of an older edition this may prove cumbersome, so in this case contact us or download the data from the github repository.
Changes of non-core data, e.g. language names or classification, may occur anytime. These changed are tracked via issues on GitHub and applied to the database by running migration scripts. The current version number of the database is 6d0d0b5416c5 (cf. the version history of the database).
WALS Online 2014
Editions
2014
2013
2011
2008

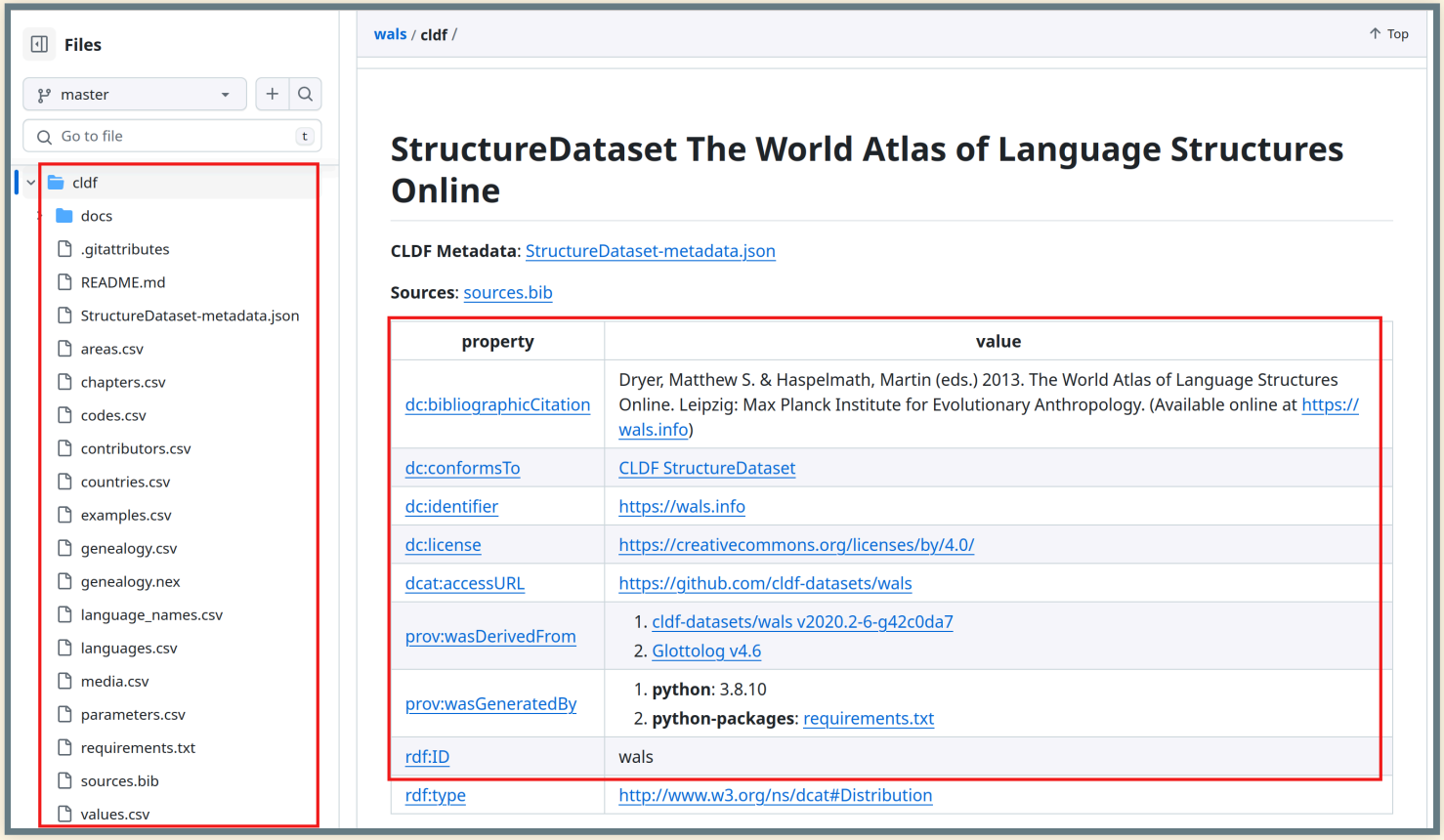

Error 404 (Not Found)!!1
https://www.google.com/badpage
Google
404. That's an error.
The requested URL /badpage was not found on this server. That's all we know.

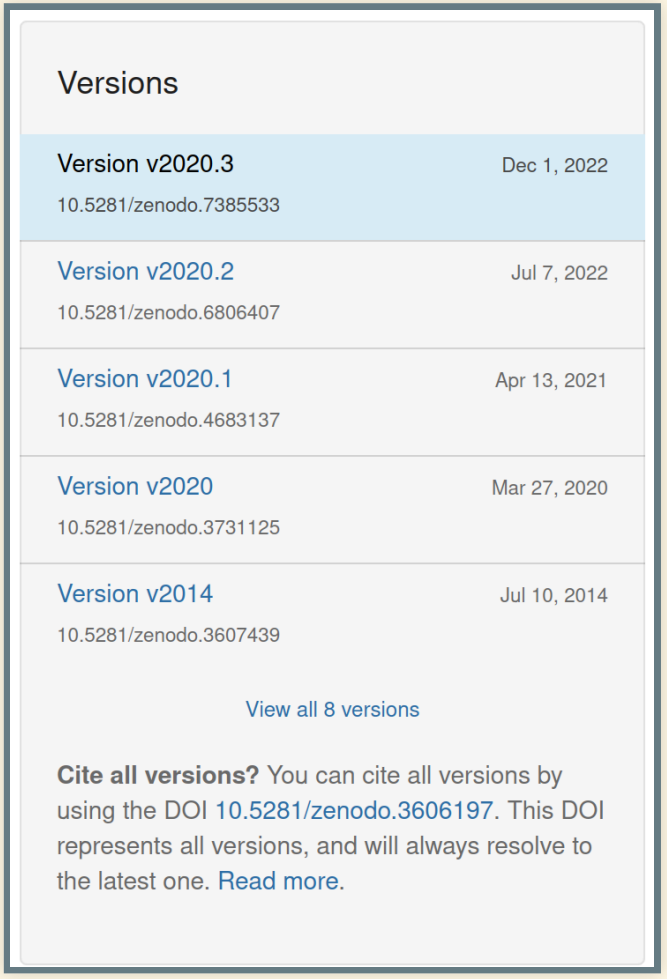
Linked Data as persistent API is not maintainable

(Linked Data as data integration paradigm is still useful, though.)

CLDF - a maintainable API for Linguistic Data

- on-disk format
- line-based text files, suitable for version control (CSV, JSON, BibTeX)
- machine-readable metadata
- Linked Data still in the mix as ontology and links to reference catalogs!


Files
master
Go to file
t
cldf
docs
.gitattributes
README.md
StructureDataset-metadata.json
areas.csv
chapters.csv
codes.csv
contributors.csv
countries.csv
examples.csv
genealogy.csv
genealogy.nex
language_names.csv
languages.csv
media.csv
parameters.csv
requirements.txt
sources.bib
values.csv
wals / cldf /
Top
StructureDataset The World Atlas of Language Structures Online
CLDF Metadata: StructureDataset-metadata.json
Sources: sources.bib
property
value
dc:bibliographicCitation
Dryer, Matthew S. & Haspelmath, Martin (eds.) 2013. The World Atlas of Language Structures Online. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at https://wals.info)
dc:conformsTo
CLDF StructureDataset
dc:identifier
https://wals.info
dc:license
https://creativecommons.org/licenses/by/4.0/
dcat:accessURL
https://github.com/cldf-datasets/wals
prov:wasDerivedFrom
1. cldf-datasets/wals v2020.2-6-g42c0da7
2. Glottolog v4.6
prov:wasGeneratedBy
1. python: 3.8.10
2. python-packages: requirements.txt
rdf:ID
wals
rdf:type
http://www.w3.org/ns/dcat#Distribution


Versions
Version v2020.3
Dec 1, 2022
10.5281/zenodo.7385533
Version v2020.2
Jul 7, 2022
10.5281/zenodo.6806407
Version v2020.1
Apr 13, 2021
10.5281/zenodo.4683137
Version v2020
Mar 27, 2020
10.5281/zenodo.3731125
Version v2014
Jul 10, 2014
10.5281/zenodo.3607439
View all 8 versions
Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.3606197. This DOI represents all versions, and will always resolve to the latest one. Read more.

Another learning process

should fit in one file

vs.

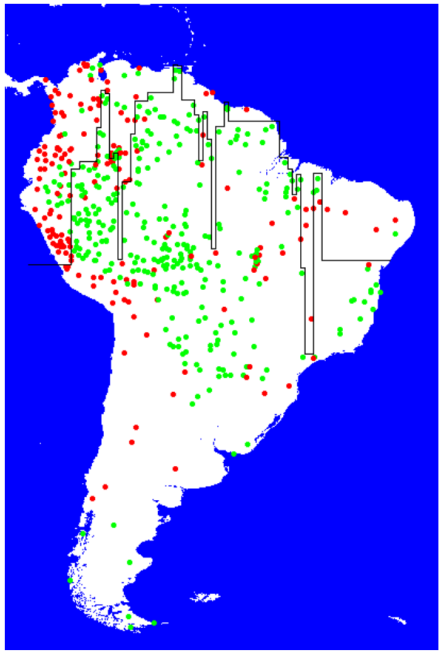
anticipate the need to use multiple tables

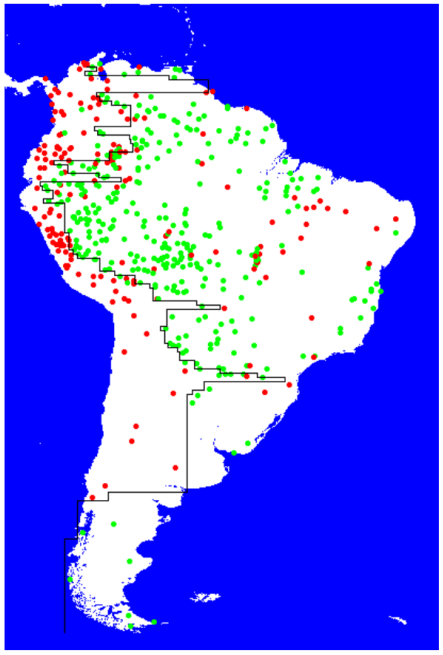
What does API mean?

- CLDF allows decoupling of dataset curation and analysis tool development!
- Thus, allows replacing the browser as the primary way to interact with data.
- (But `clld` apps are still *one* type of analysis tool for CLDF data).

Language comparison with CLDF

Optimal Isogloss Lines: Outcome



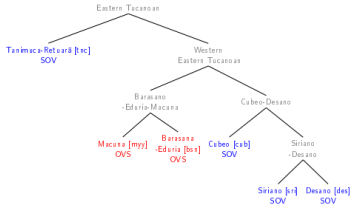


- The optimal line east-west has error 513.125 but
- the optimal line north-south only has 241.875

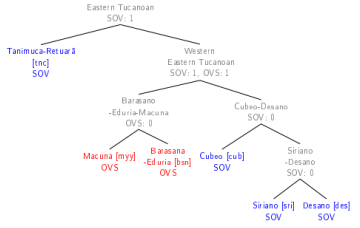
Parsimony Reconstruct: Procedure & Output

To each internal node, reconstruct the value that **minimizes** *the* **total number of changes** *required*

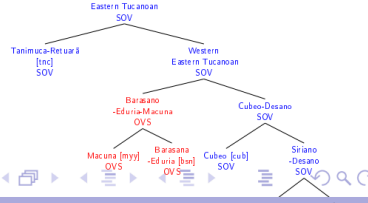
1. Input (a tree and values at the leaves)



2. For each internal node, starting near the leaves, calculate the minimum number of changes required below it for each possible reconstructed value



3. Reconstruct that which yield the minimum total number of changes in the tree



Language identification

CLDF allows unambiguous identifiation of languoids in a dataset via

- the `LanguageTable` component and
- the `glottocode` property

Bonkiman (Madang/Finisterre; also Yupna?) Claassen & McElhanon 1970, McElhanon 1978, W&H 1981, Z'graggen 1975

Borai (E Bird's Head; also Mansim) Barrs 1978, Voorhoeve 1975, W&H 1981; Borai-Hattam W&H 1981

Border d of Komba W&H 1981; Border Stock W&H 1981; Border-Tor-Lake Plain > Northern SS

Borebo d of Magi Thomson 1975, 1975

Bori (Musa R, O

Boro MacDonalc

Boroborogo (Ne
Boromeso, Borur

Bosavi (Fly R) S
Family Shaw

Bosilewa (AN, F

Boskien (Madan

Bosmun (W Mac
Bosngun) Bla
graggen 19

Bosngun (*see* B

Botin (S of Angc
1989, Wade

Bougainville lar
Kariks et al 1
1975, Tryon

Bo Ung (d of Ka

Bouye (W Sepik
Boven-Bian > Bi

Bowai (AN; Mar

Boyowan (AN;
1949, n.d.

Boze-Giringare
Giringarede

Bragat (Torricell

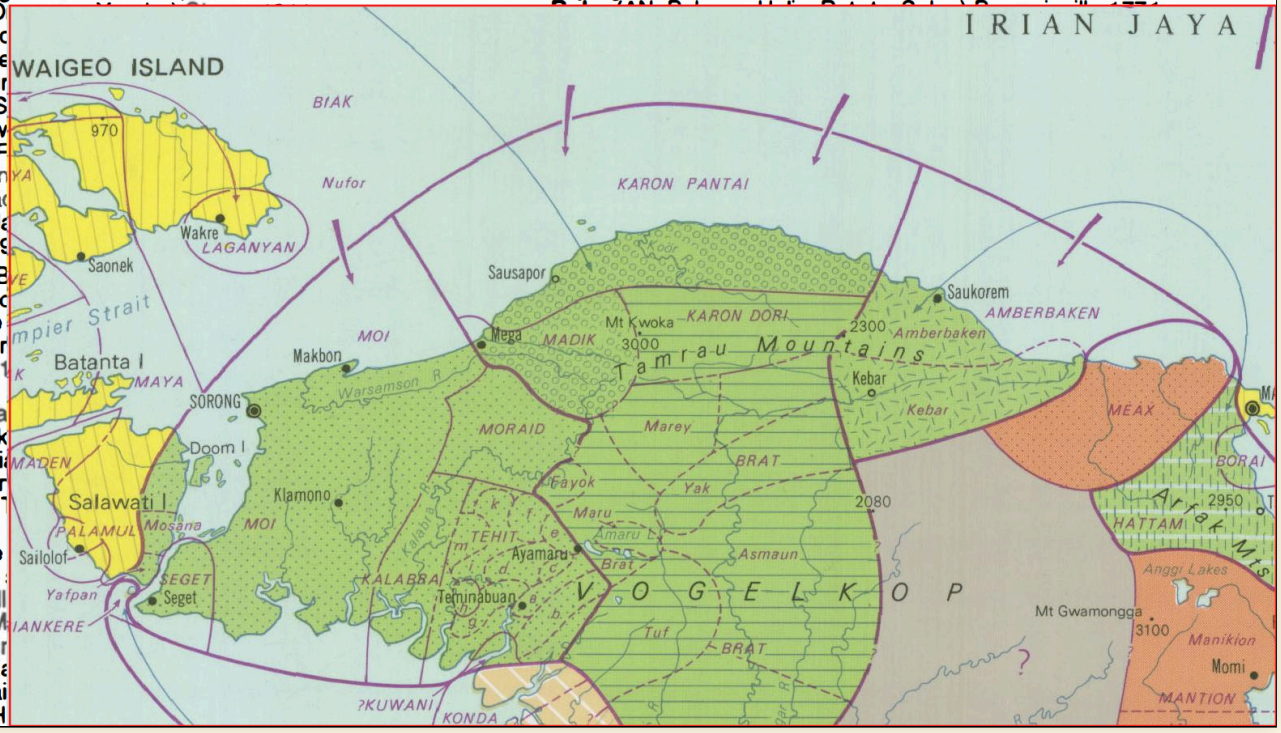
Brahman FS (M

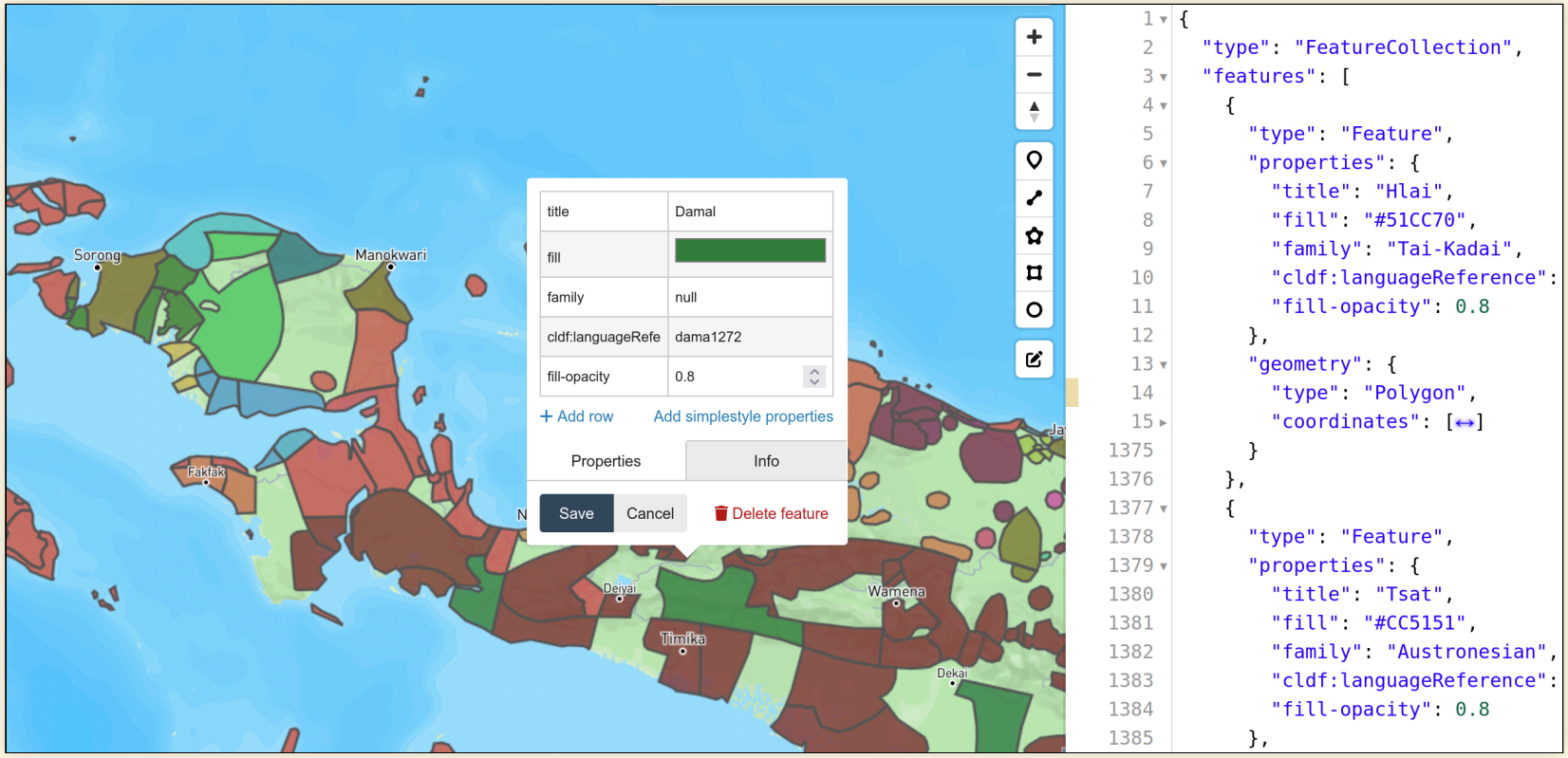
Brat (Central Bir
Gosirago, Ka
see also Mai
of Brat W&H

Longacre 1972, Lugabai 1971, Müller 1944-45, n.d. x 3, Müller & Miltrup 1943-44, n.d., n.d., Nakota 1974, Oliver-Berg 1979, Potu 1974, Rugabai & Griffin 1971, Schmidt 1909, SIL 1971, 1975, Simons 1982, Tauria 1974, Thurnwald 1909, 1912, 1934, 1934, 1937, 1942, n.d., Vaughan 1977, Wheeler 1911; d's W&H 1981; Buin Family W&H 1981

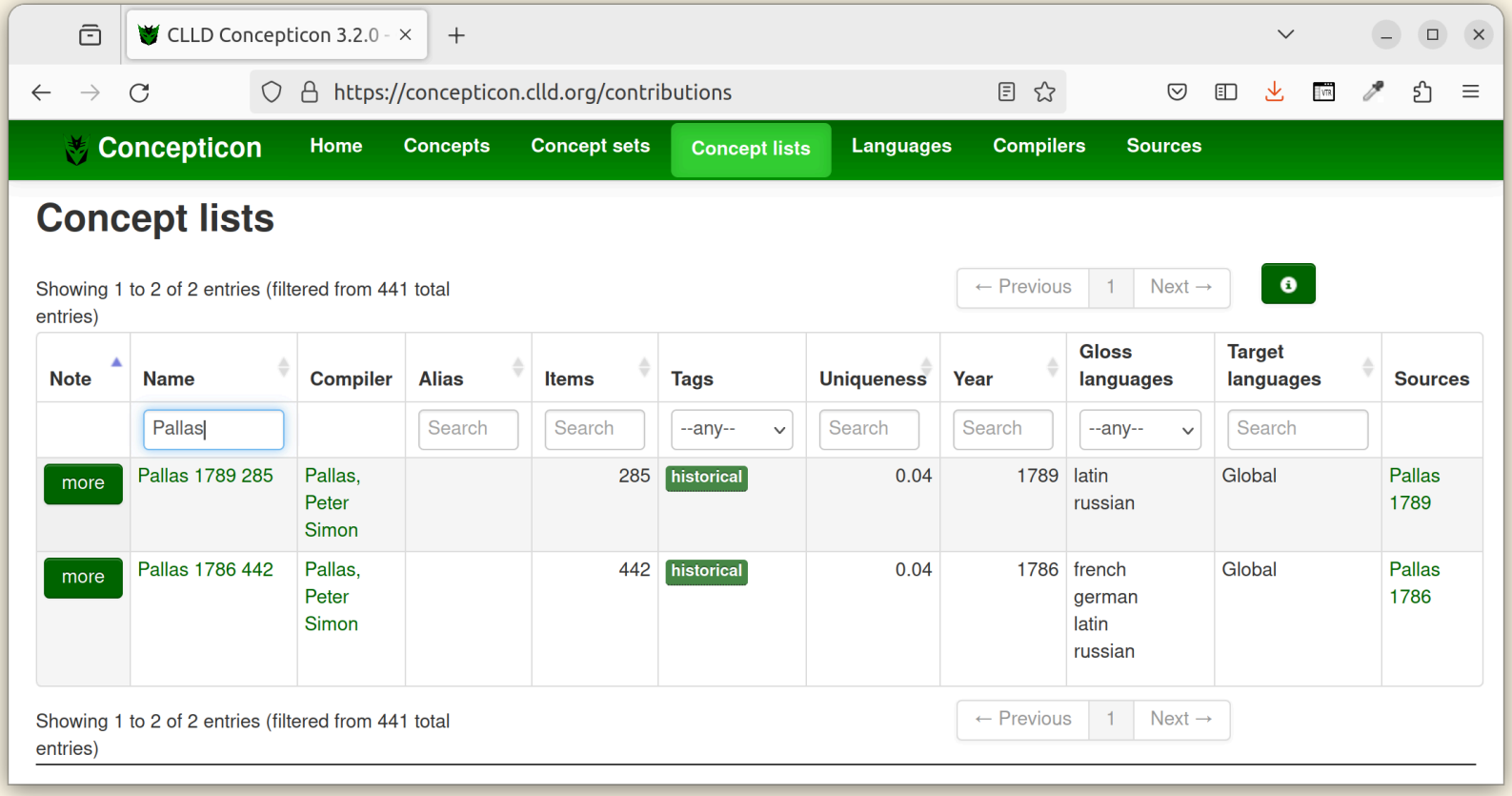
Buiye > Bouye

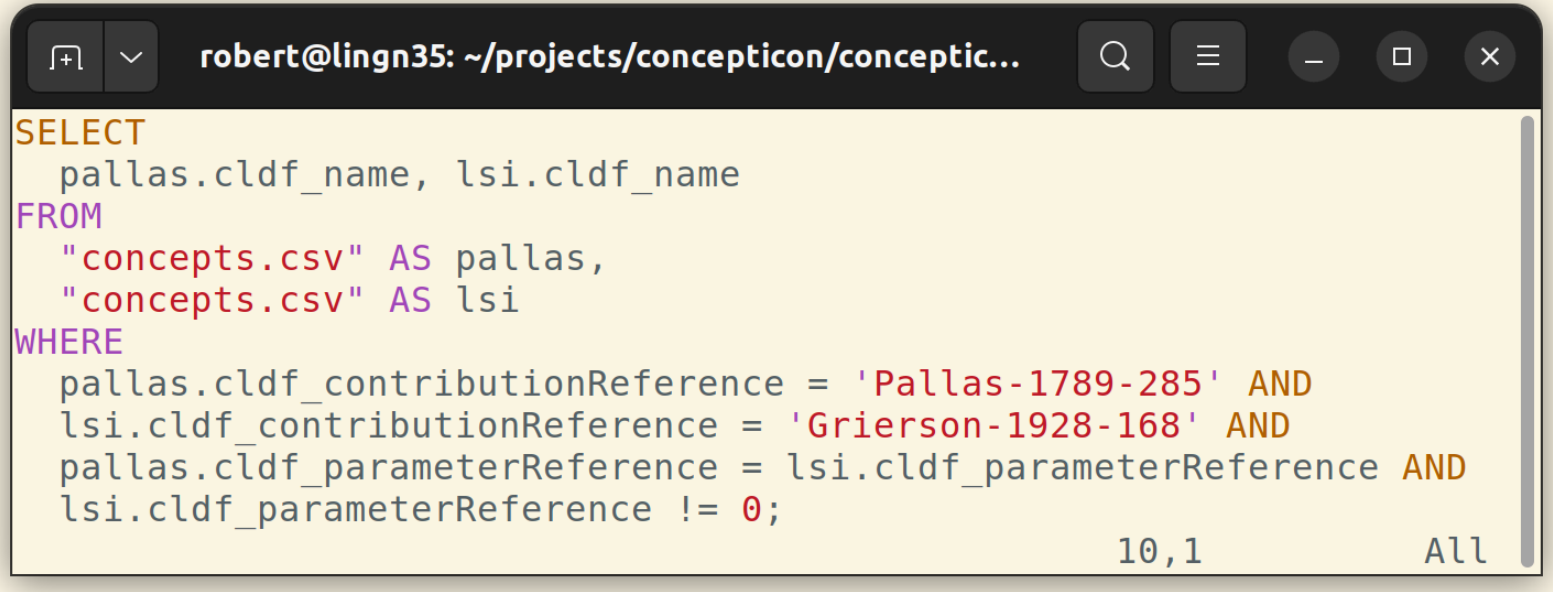
Bujang > Kele




Sorong
Manokwari
Fakfak
Deiyai
Timika
Wamena
Dekai
title
Damal
fill
family
null
cldf:languageRefe
dama1272
fill-opacity
0.8
+ Add row
Add simplestyle properties
Properties
Info
Save
Cancel
Delete feature
{
"type": "FeatureCollection",
"features": [
{
"type": "Feature",
"properties": {
"title": "Hlai",
"fill": "#51CC70",
"family": "Tai-Kadai",
"cldf:languageReference":
"fill-opacity": 0.8
},
"geometry": {
"type": "Polygon",
"coordinates": [↔]
}
},
{
"type": "Feature",
"properties": {
"title": "Tsat",
"fill": "#CC5151",
"family": "Austronesian",
"cldf:languageReference":
"fill-opacity": 0.8
},

Concept identification


CLLD Concepticon 3.2.0
https://concepticon.clld.org/contributions
Concepticon
Home
Concepts
Concept sets
Concept lists
Languages
Compilers
Sources
Concept lists
Showing 1 to 2 of 2 entries (filtered from 441 total entries)
← Previous
1
Next →
Note
Name
Compiler
Alias
Items
Tags
Uniqueness
Year
Gloss languages
Target languages
Sources
Pallas
Search
--any--
more
Pallas 1789 285
Pallas, Peter Simon
285
historical
0.04
1789
latin russian
Global
Pallas 1789
more
Pallas 1786 442
Pallas, Peter Simon
442
historical
0.04
1786
french german latin russian
Global
Pallas 1786
Showing 1 to 2 of 2 entries (filtered from 441 total entries)
← Previous
1
Next →


robert@lingn35: ~/projects/concepticon/conceptic...
SELECT
pallas.cldf_name, lsi.cldf_name
FROM
"concepts.csv" AS pallas,
"concepts.csv" AS lsi
WHERE
pallas.cldf_contributionReference = 'Pallas-1789-285' AND
lsi.cldf_contributionReference = 'Grierson-1928-168' AND
pallas.cldf_parameterReference = lsi.cldf_parameterReference AND
lsi.cldf_parameterReference != 0;
10,1 All

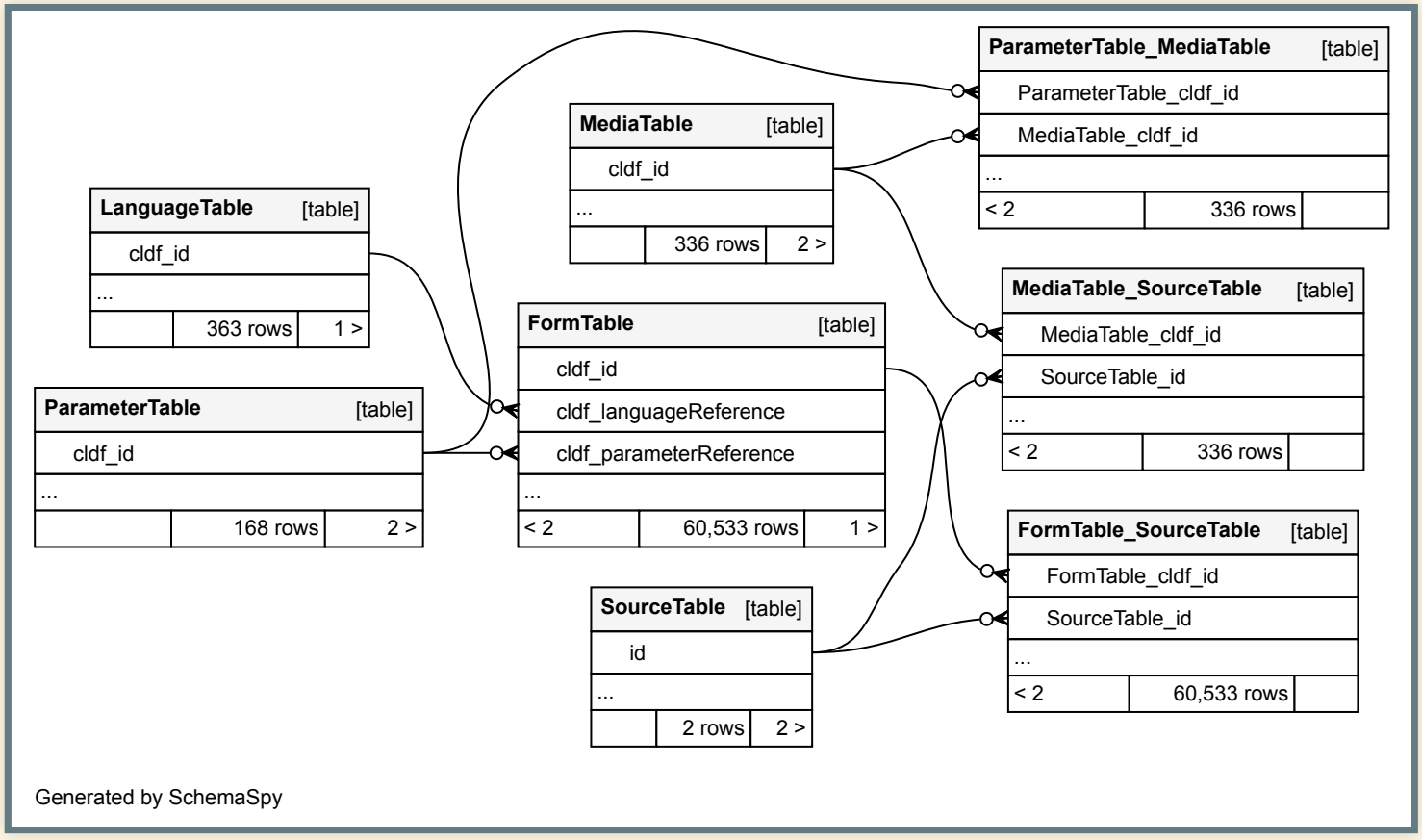
Exploratory analysis

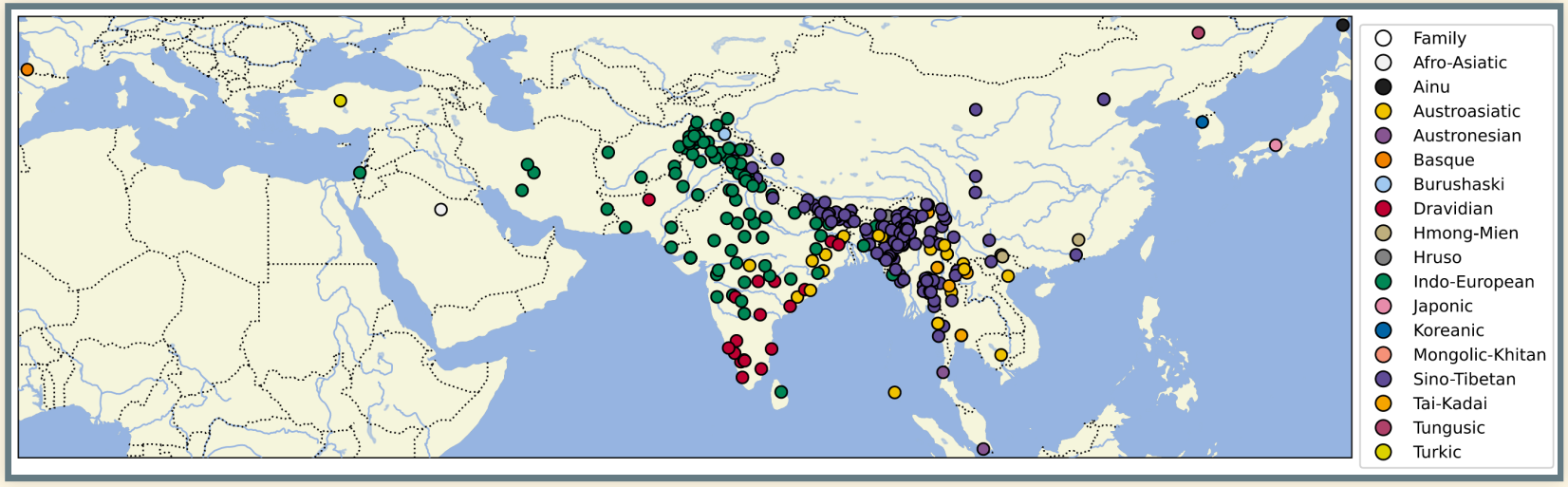
Table values.csv

property	value
dc:conformsTo	CLDF ValueTable
dc:extent	76475

Columns

Name/Property	Datatype	Description
ID	`string`	Primary key
Language_ID	`string`	References languages.csv::ID
Parameter_ID	`string`	References parameters.csv::ID
Value	`string`	
Code_ID	`string`	References codes.csv::ID
Comment	`string`	comments in HTML
Source	list of `string` (separated by `;`)	References sources.bib::BibTeX-key
Example_ID	list of `string` (separated by ` `)	References examples.csv::ID


ParameterTable_MediaTable [table]
ParameterTable_cldf_id
MediaTable_cldf_id
...
< 2
336 rows
MediaTable [table]
cldf_id
...
336 rows
2 >
LanguageTable [table]
cldf_id
...
363 rows
1 >
MediaTable_SourceTable [table]
MediaTable_cldf_id
SourceTable_id
...
< 2
336 rows
FormTable [table]
cldf_id
cldf_languageReference
cldf_parameterReference
...
< 2
60,533 rows
1 >
ParameterTable [table]
cldf_id
...
168 rows
2 >
FormTable_SourceTable [table]
FormTable_cldf_id
SourceTable_id
...
< 2
60,533 rows
SourceTable [table]
id
...
2 rows
2 >
Generated by SchemaSpy


Family
Afro-Asiatic
Ainu
Austroasiatic
Austronesian
Basque
Burushaski
Dravidian
Hmong-Mien
Hruso
Indo-European
Japonic
Koreanic
Mongolic-Khitan
Sino-Tibetan
Tai-Kadai
Tungusic
Turkic

Analysis methods

Cognates

CLDF provides all the semantics necessary for LingPy's cognate detection and alignment algorithms:

- unambiguously identified languages and concepts
- segmented word forms

and the data model to store the computed cognacy relations, thus, the input for phylogenetic analyses, e.g. via BEASTling.

Colexifications

CLDF provides

- the semantics to compute colexifications
- the data model to store the computed networks
- thus, the input for clustering algorithms.

Generic data access: CLDF SQL

- Each CLDF dataset can be loaded into an SQLite database running a simple command from the `pycldf` package.
- The resulting SQL schema allows uniform data access across datasets.

```
ATTACH DATABASE "phoible.sqlite" AS phoible;
ATTACH DATABASE "clts.sqlite" AS clts;
ATTACH DATABASE "lsi.sqlite" AS lsi;
```

```
CREATE TEMP VIEW lsigraphemes AS
SELECT
  DISTINCT grapheme
FROM
(
  WITH split(grapheme, segments) AS (
    SELECT
      '',
      f.cldf_segments || ' '
      FROM lsi.formtable AS f, lsi.languagetable AS l
      WHERE f.cldf_languagereference = l.cldf_id AND l.cldf_glottocode = 'mala1464'
    UNION ALL SELECT
      substr(segments, 0, instr(segments, ' ')),
      substr(segments, instr(segments, ' ') + 1)
    FROM split
    WHERE segments != ''
  ) SELECT grapheme FROM split
  WHERE grapheme!=''
);
```

```
25  CREATE TEMP VIEW phoiblegraphemes AS
26  SELECT
27    DISTINCT c.cltsgrapheme AS grapheme
28  FROM
29  (
30    SELECT v.cldf_value AS grapheme
31    FROM phoible.valuetable AS v
32    WHERE cldf_languagereference = 'mala1464' AND contribution_id = 1762
33  ) AS p
34  JOIN
35  (
36    SELECT phoible.grapheme AS phoiblegrapheme, clts.grapheme AS cltsgrapheme
37    FROM clts."data/graphemes.tsv" AS phoible, clts."data/sounds.tsv" AS clts
38    WHERE phoible.dataset = 'phoible' AND phoible.name = clts.name
39  ) AS c
40  ON c.phoiblegrapheme = p.grapheme;
```

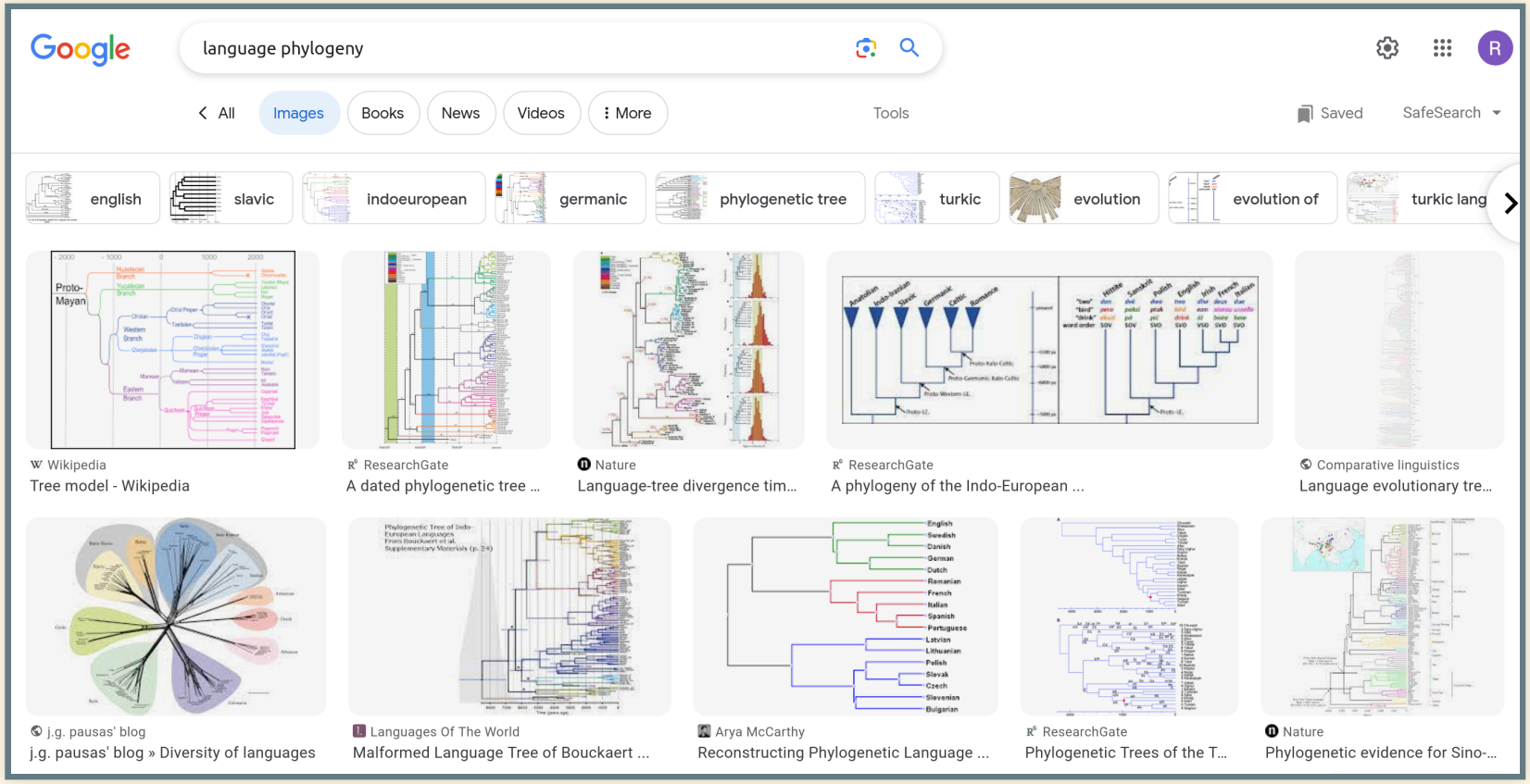
```
42  SELECT lsi.grapheme, 'LSI', clts.name
43  FROM lsigraphemes AS lsi
44  JOIN
45    clts."data/sounds.tsv" AS clts
46  ON clts.grapheme = lsi.grapheme
47  WHERE clts.name LIKE '%vowel' AND lsi.grapheme NOT IN phoiblegraphemes;
48
49  SELECT phoible.grapheme, 'PHOIBLE', clts.name
50  FROM phoiblegraphemes AS phoible
51  JOIN
52      clts."data/sounds.tsv" AS clts
53  ON clts.grapheme = phoible.grapheme
54  WHERE clts.name LIKE '%vowel' AND phoible.grapheme NOT IN lsigraphemes;
```

Glyph	Dataset	CLTS
ʌ	LSI	unrounded_open_mid_back_vowe
ʌː	LSI	long_unrounded_open_mid_back_
a	PHOIBLE	unrounded_open_front_vowel
aː	PHOIBLE	long_unrounded_open_front_vowe
æ	PHOIBLE	unrounded_near_open_front_vowe
ɨ	PHOIBLE	unrounded_close_central_vowel
ʊ	PHOIBLE	rounded_near_close_near_back_v

Data modeling

CLDF is extensible and built to evolve by

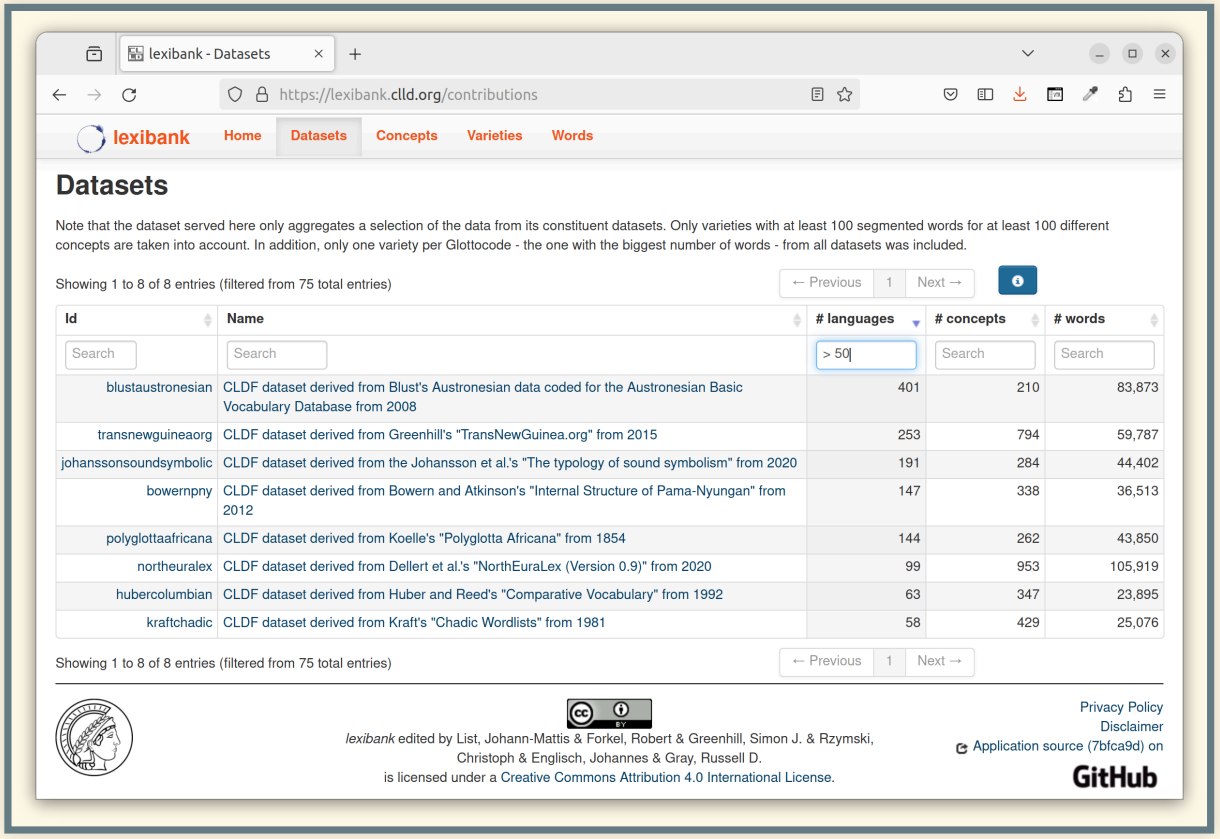
- adding data types (aka tables or components)
- adding properties (aka columns)
- specifying relations between tables/columns and media files


Google
language phylogeny
All
Images
Books
News
Videos
More
Tools
Saved
SafeSearch
english
slavic
indoeuropean
germanic
phylogenetic tree
turkic
evolution
evolution of
turkic lang
Proto-Mayan
Wikipedia
Tree model - Wikipedia
ResearchGate
A dated phylogenetic tree ...
Nature
Language-tree divergence tim...
ResearchGate
A phylogeny of the Indo-European ...
Comparative linguistics
Language evolutionary tre...
j.g. pausas' blog
j.g. pausas' blog » Diversity of languages
Languages Of The World
Malformed Language Tree of Bouckaert ...
Arya McCarthy
Reconstructing Phylogenetic Language ...
ResearchGate
Phylogenetic Trees of the T...
Nature
Phylogenetic evidence for Sino-...

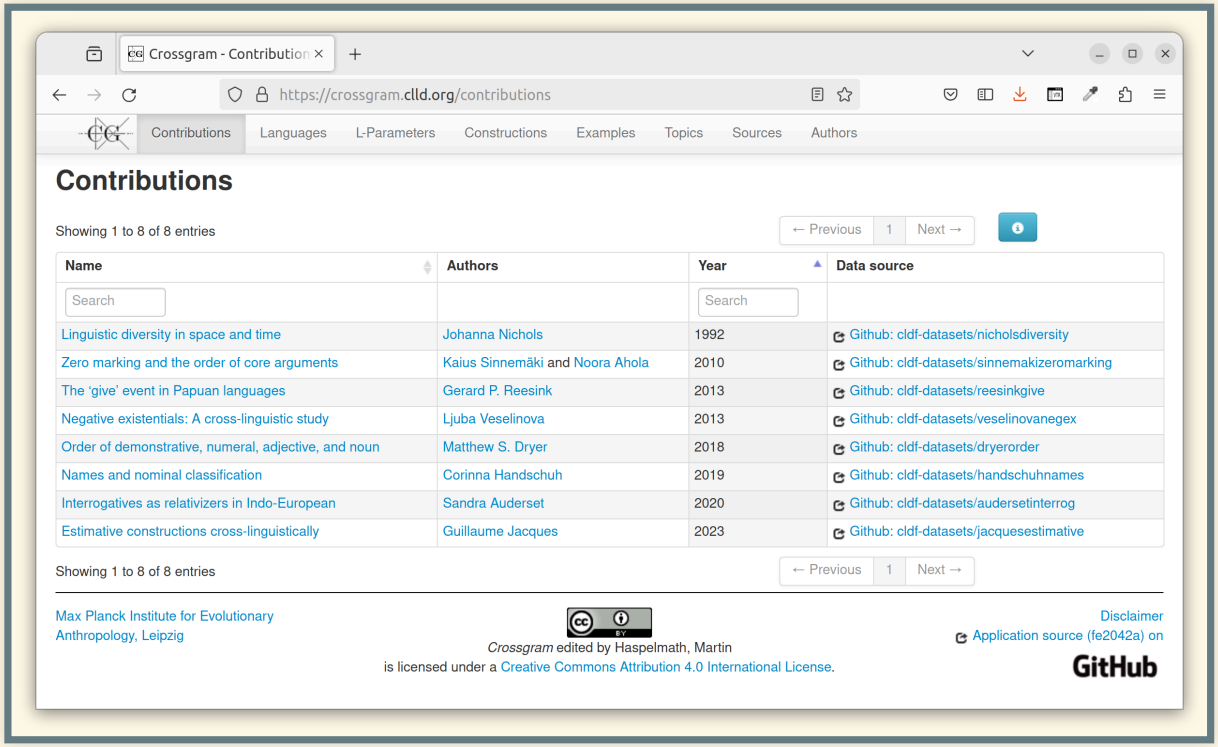
Aggregation

By providing a uniform way to access data across datasets, CLDF is an essential tool for compiling large, aggregated databases.

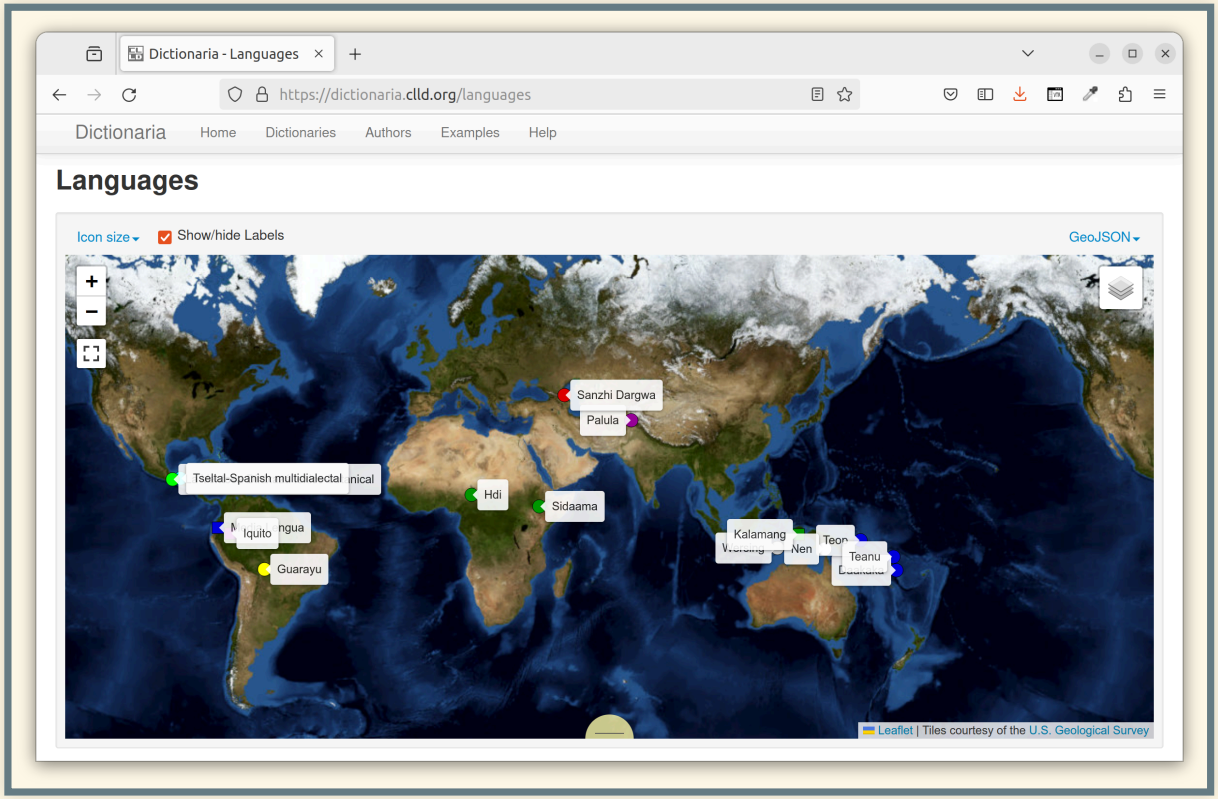
CLDF Wordlist



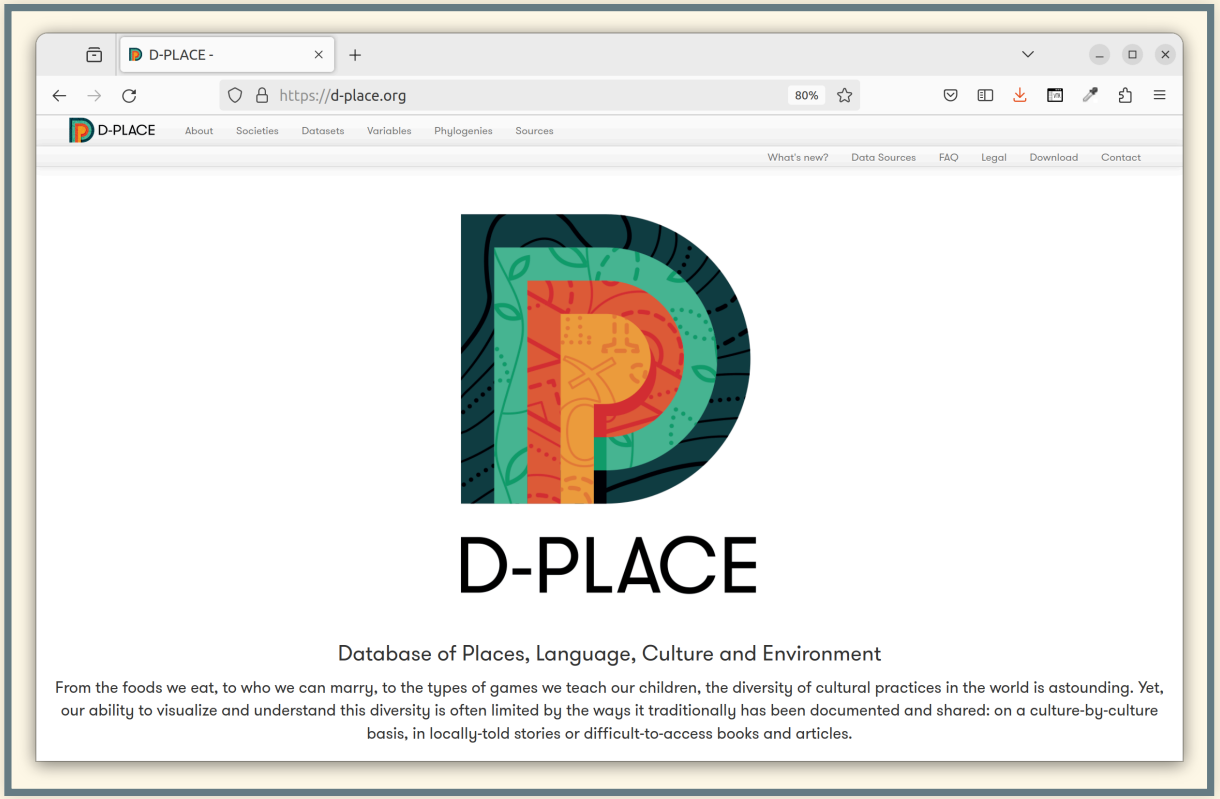
CLDF StructureDataset



CLDF Dictionary



But also


D-PLACE -
https://d-place.org
80%
D-PLACE
About
Societies
Datasets
Variables
Phylogenies
Sources
What's new?
Data Sources
FAQ
Legal
Download
Contact
D-PLACE
Database of Places, Language, Culture and Environment
From the foods we eat, to who we can marry, to the types of games we teach our children, the diversity of cultural practices in the world is astounding. Yet, our ability to visualize and understand this diversity is often limited by the ways it traditionally has been documented and shared: on a culture-by-culture basis, in locally-told stories or difficult-to-access books and articles.

More

CLDF Markdown

```
### Rule 2: Morpheme-by-morpheme correspondence

Segmentable morphemes are separated by hyphens, both in the example and in the
gloss. There must be exactly the same number of hyphens in the example and in the
gloss. E.g.

[example 2](cldf/examples.csv#cldf:2)
```

Rule 2: Morpheme-by-morpheme correspondence

Segmentable morphemes are
gloss. There must be exactly th
the gloss. E.g.

> (2) Lezgian (Haspelmath 199

```
Gila   abur-u-n       ferma   hamišaluǧ   güǧüna   amuq’-da-č.
now    they-OBL-GEN   farm    forever     behind   stay-FUT-NEG
‘Now their farm will not stay behind forever.’
```

Catalog of Linguistic Data

CLDF datasets hold much more “actionable” (meta)data than archive holdings exposed through OLAC.

From

list all holdings characterized as *text* for language X

to

list all segments found in any text for language X

https://cldf.clld.org