

CROSS-LINGUISTIC DATA ... IS CODE

Robert Forkel

April 30, 2015

Max Planck Institute for evolutionary Anthropology

Homoiconicity in programming languages means:

code can be treated as a basic data structure that the programming language knows how to access.

<http://blogs.mulesoft.org/code-is-data-data-is-code/>

Thus the programming language can manipulate code more reliably (no syntax errors!) than your text editor.

So homoiconicity is a desirable property of a programming language, because it allows for better tooling.

Turns out – in a rather mundane way – data is code, too.

Or less catchy: cross-linguistic data is similar enough to (open source) code to share its tools:

- textual data

Turns out – in a rather mundane way – data is code, too.

Or less catchy: cross-linguistic data is similar enough to (open source) code to share its tools:

- textual data
- line based (thanks to Harald and cldf)

Turns out – in a rather mundane way – data is code, too.

Or less catchy: cross-linguistic data is similar enough to (open source) code to share its tools:

- textual data
- line based (thanks to Harald and cldf)
- not really Big Data

Turns out – in a rather mundane way – data is code, too.

Or less catchy: cross-linguistic data is similar enough to (open source) code to share its tools:

- textual data
- line based (thanks to Harald and cldf)
- not really Big Data
- open (often)

DATA CURATION WITH GITHUB

Using

`cldf` as format for our data,

`git` (a tool for distributed source code management) and

GitHub (a platform hosting git repositories).

we get a platform for collaboratively curating cross-linguistic data.

EXAMPLE: TSAMMALEX

GitHub, Inc. (US) | https://github.com/clld/tsammalex-data | mpg hachgenei

This repository Search Explore Gist Blog Help xrotwang +

clld / **tsammalex-data** Unwatch 4 Star 0 Fork 2

Tsammalex is a multilingual lexical database on plants and animals. <http://tsammalex.clld.org> — Edit

250 commits **History** 1 branch 0 releases 6 contributors **Who?**

branch: master tsammalex-data / +

File	Commit Message	Time Ago
fixed problems		
xrotwang authored		12 hours ago
tsammalexdata	fixed problems	12 hours ago
.gitattributes	fixed suffix	5 months ago
.gitignore	updated data	5 months ago
.travis.yml	added email notification config for travis	4 months ago
MANIFEST.in	implemented functionality to harvest external data	5 months ago
README.md	work on integration of dogon data	4 months ago
setup.py	updated taxa info from external sources	19 hours ago
species.json	added data parsed from medlawiki	a year ago

SSH clone URL
git@github.com:clld/tsammalex-data.git

You can clone with HTTPS, SSH, or Subversion

Download ZIP

Can I add my data?

Can I get your data?

Figure: The data for Tsammalex is curated at clld/tsammalex-data.

A DATABASE JOURNAL

This can be easily extended to a platform for data journals:

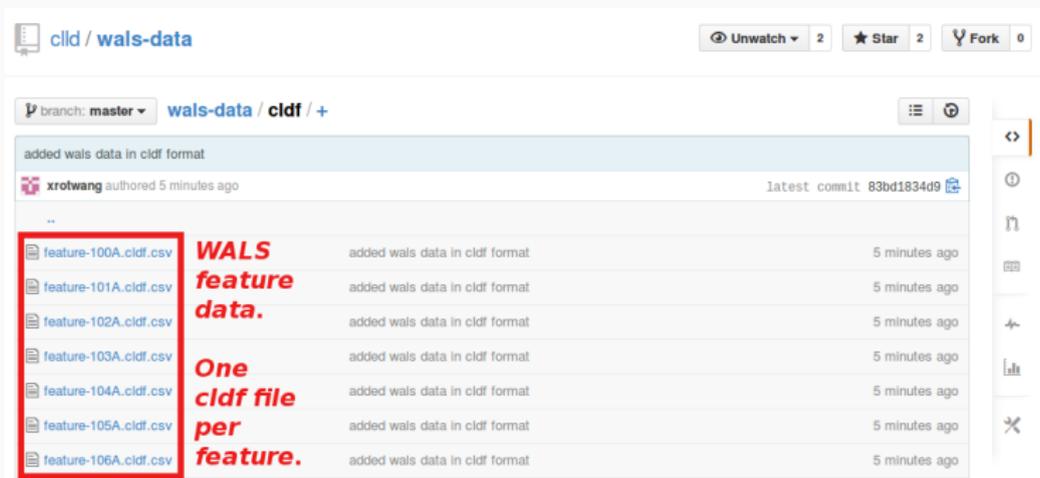
cldf the submission guidelines

pull request submission with open review

merge accepting a submission

release publication

EXAMPLE: WALS AS DATABASE JOURNAL



The screenshot shows a GitHub commit page for the repository 'cldf / wals-data'. The commit message is 'added wals data in cldf format' and was authored by 'xrotwang' 5 minutes ago. The commit hash is '83bd1834d9'. The commit includes seven files, each named 'feature-100A.cldf.csv' through 'feature-106A.cldf.csv'. A red box highlights the file names, and red text annotations emphasize 'WALS feature data.' and 'One cldf file per feature.'

File Name	Commit Message	Time
feature-100A.cldf.csv	added wals data in cldf format	5 minutes ago
feature-101A.cldf.csv	added wals data in cldf format	5 minutes ago
feature-102A.cldf.csv	added wals data in cldf format	5 minutes ago
feature-103A.cldf.csv	added wals data in cldf format	5 minutes ago
feature-104A.cldf.csv	added wals data in cldf format	5 minutes ago
feature-105A.cldf.csv	added wals data in cldf format	5 minutes ago
feature-106A.cldf.csv	added wals data in cldf format	5 minutes ago

Figure: WALS as database journal.

BROWSING CLDF ON GITHUB



The screenshot shows a GitHub commit interface. At the top, it indicates the branch is 'master' and the file path is 'wals-data / cldf / feature-1A.cldf.csv'. A commit by user 'xrotwang' is shown, with a message '8 minutes ago added wals data in cldf format'. Below this, it says '1 contributor' and features a red banner that reads 'csv and GitHub go well together'. The file statistics show '565 lines (564 stoc) | 129.153 kb'. Navigation buttons for 'Raw', 'Blame', and 'History' are visible. A search bar is present with the text 'Search this file...'. The main content is a table with three columns: 'LANGUAGE_NAME', 'VALUE', and 'SOURCE_NAME'. The table contains 14 rows of data, including languages like Abipón, Abkhaz, Aché, Achumawi, Acoma, Adzera, Aghem, Ahtha, Alkana, Ainu, Aizi, and Akan.

	LANGUAGE_NAME	VALUE	SOURCE_NAME
1			
2	Abipón	Moderately small	Najlis 1966
3	Abkhaz	Large	Hewitt 1979
4	Aché	Small	Susnik 1974
5	Achumawi	Moderately small	Climsted 1964;Climsted 1966
6	Acoma	Large	Miller 1966
7	Adzera	Moderately small	Holzknicht 1973
8	Aghem	Average	Hyman 1979
9	Ahtha	Moderately large	Kari 1990
10	Alkana	Average	Hanke 1956;Alkhenvald and Dixon 1999[362-363]
11	Ainu	Small	Patrie 1982;Simeon 1969
12	Aizi	Average	Herault 1971
13	Akan	Average	Dolphyne 1968;Ladefoged 1964;Schachter and Fromkin 1968;Stewart 1967;Weimers 1946

Figure: GitHub supports tabular data well.

PULL REQUESTS AS SUBMISSIONS

Comparing changes

Choose two branches to see what's changed or to start a new pull request. If you need to, you can also [compare across forks](#).

The "journal" we want to submit to.

base fork: **cldf/wals-data** base: **master** head fork: **xrotwang/wals-data** compare: **master**

✓ **Able to merge.** These branches can be automatically merged.

Create pull request Discuss and review the changes in this comparison with others. **A "pull request" is a submission.**

2 commits 2 files changed 0 commit comments 1 contributor

Commits on Apr 24, 2015

- xrotwang Added a new (fake) feature: 145A
- xrotwang Added feature metadata

And here's what we want to submit.

Showing 2 changed files with 20 additions and 0 deletions. Unified Split

```
6 cldf/feature-145A.cldf.csv View
... .. @@ -0,0 +1,6 @@
1 +LANGUAGE_NAME, VALUE, SOURCE_NAME_ID, LANGUAGE_ID, FEATURE_ID, SOURCE
2 +Abipón, Moderately small, Najlis 1966, http://wals.info/valuesets/145A-ab1, http://wals.info/languoid/lect/wals_code_ab
3 +Abkhaz, Large, Hewitt 1979, http://wals.info/valuesets/145A-abk, http://wals.info/languoid/lect/wals_code_abk, http://w
4 +Aché, Small, Susnik 1974, http://wals.info/valuesets/145A-ach, http://wals.info/languoid/lect/wals_code_ach, http://w
5 +Achumawi, Moderately small, Olmsted 1964;Olmsted 1966, http://wals.info/valuesets/145A-acm, http://wals.info/languoid/
6 +Acoma, Large, Miller 1966, http://wals.info/valuesets/145A-aco, http://wals.info/languoid/lect/wals_code_aco, http://wa
```

Figure: pull requests as submission mechanism.

LIST OF SUBMISSIONS

cclid / wals-data *Meanwhile back at the journal ...*

Unwatch 2 Star 2 Fork 1

Issues Pull requests Labels Milestones Filters New pull request

1 Open ✓ 0 Closed

Submission of a new feature to WALS ... *there's one new submission.*
#31 opened 6 minutes ago by xrotwang

Figure: The list of open pull requests serves as the backlog for the editors.

OPEN REVIEW

Submission of a new feature to WALS #31

 **xrotwang** wants to merge 2 commits into `clld:master` from `xrotwang:master`

 Conversation **0**  Commits **2**  Files changed **2**

 **xrotwang** commented a minute ago Owner 

... description of the feature ...

 **xrotwang** added some commits 9 minutes ago

-   Added a new (fake) feature: 145A ... fea633d
-   Added feature metadata 069899a

 **reviewer** commented just now Owner  

Review comment: New features should cover at least x languages. ***A review comment.***

Figure: Reviewers can comment on pull requests.

ADDRESSING REVIEW COMMENTS

Submission of a new feature to WALS #31

 **Open** xrotwang wants to merge 3 commits into `c11d:master` from `xrotwang:master`

 Conversation **1**  **Commits 3**  Files changed **2**

 xrotwang commented 4 minutes ago Owner 

... description of the feature ...

 xrotwang added some commits 12 minutes ago

-  Added a new (fake) feature: 145A ... fea633d
-  Added feature metadata 069899a

 xrotwang commented 3 minutes ago Owner  

Review comment: New features should cover at least x languages.

-   Update feature-145A.c1df.csv 00b4c0d

A modification addressing the review comment.

Figure: Review comments can be addressed by adding commits to the pull request, i.e. amending the submission.

SUBMITTING LANGUAGE SURVEYS

The screenshot shows a GitHub pull request titled "added values for features 1A-3A for Language Aand Another Language #32". The pull request is from user "xrotwang" and is ready to be merged. It shows 3 files changed: cldf/feature-1A.cldf.csv, cldf/feature-2A.cldf.csv, and cldf/feature-3A.cldf.csv. A red text overlay reads: "A different kind of submission: adding values for existing features." Below this, a diff view for cldf/feature-1A.cldf.csv is shown, with a table of changes:

...	...	@@ -1,4 +1,5 @@
1	1	LANGUAGE_NAME, VALUE, SOURCE_NAME, ID, LANGUAGE_ID, FEATURE_ID, SOURCE
2	2	+Aand Another Language, Moderately small, Najlis 1966, http://wals.info/valuesets/1A-abi, http://wals.info/languoid/lect
3	3	Abipón, Moderately small, Najlis 1966, http://wals.info/valuesets/1A-abi, http://wals.info/languoid/lect/wals_code_abi,
4	4	Abkhaz, Large, Hewitt 1979, http://wals.info/valuesets/1A-abk, http://wals.info/languoid/lect/wals_code_abk, http://wal
5	5	Aché, Small, Susnik 1974, http://wals.info/valuesets/1A-ach, http://wals.info/languoid/lect/wals_code_ach, http://wals.:

Figure: Since pull requests can bundle changes to many parts of the repository, allowing language surveys in addition to new features as submissions is not a problem.

BELLS AND WHISTLES

BLAME

tsammax-data / tsammaxdata / data / uses.csv 

Newer  Older

100644 | 29 lines (28 sloc) | 0.711 kb

Raw Normal view History

Commit Hash	Author	Date	Line	Content
7fd32c2	xrotwang	Dec 5, 2014	1	id,name,description
3f79586	LenaSell	Mar 12	2	animalfeeding,animal feeding,
926de1e	LenaSell	Jan 15	3	artscrafts,arts & crafts,
926de1e	LenaSell	Jan 15	4	boats,boats,
926de1e	LenaSell	Jan 15	5	clothingtextiles,clothing & textiles,
7fd32c2	xrotwang	Dec 5, 2014	6	construction,construction,
926de1e	LenaSell	Jan 15	7	cosmeticshygiene,cosmetics & hygiene,
05dfd15	LenaSell	Feb 10	8	dyetanning,dye & tanning,

**"Blame":
per-line
annotation
of last
modification.**

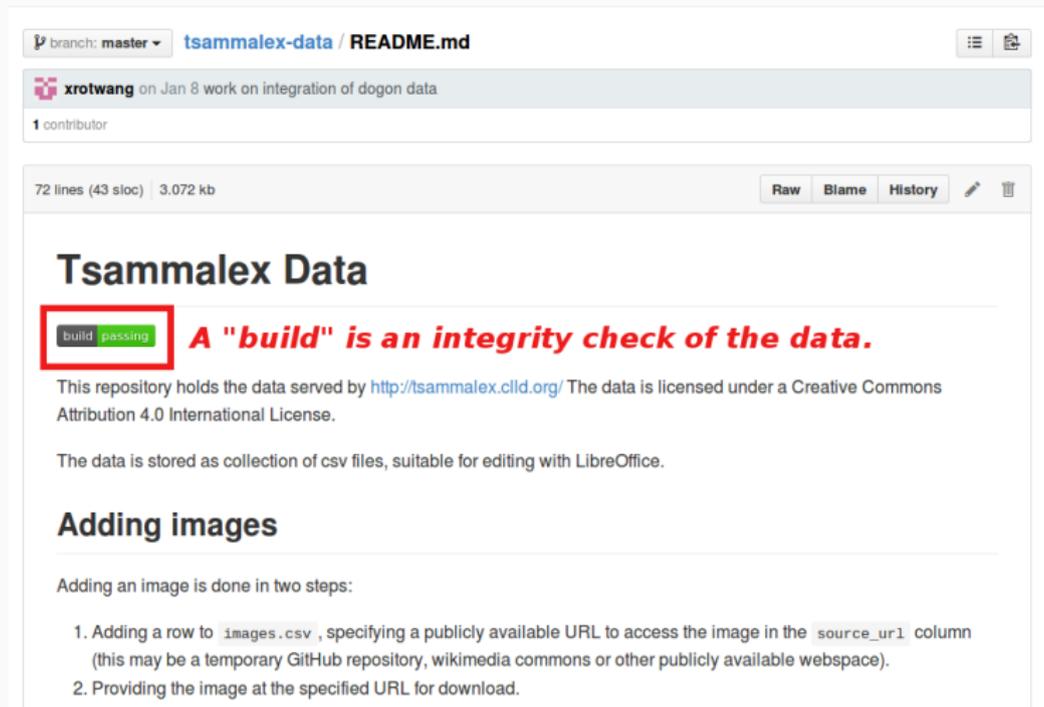
Figure: The Git *blame* functionality provides per-line annotation of last modification - thus allowing provenance tracking not only on file level.

Let's go further borrowing best practices in software development.

Continuous integration

In addition to automated [...] tests, organisations using CI typically use a build server to implement continuous processes of applying quality control in general – small pieces of effort, applied frequently.

- http://en.wikipedia.org/wiki/Continuous_integration



branch: master | tsammalex-data / README.md

xrotwang on Jan 8 work on integration of dogon data

1 contributor

72 lines (43 sloc) | 3.072 kb

Raw Blame History

Tsammalex Data

build passing *A "build" is an integrity check of the data.*

This repository holds the data served by <http://tsammalex.clld.org/> The data is licensed under a Creative Commons Attribution 4.0 International License.

The data is stored as collection of csv files, suitable for editing with LibreOffice.

Adding images

Adding an image is done in two steps:

1. Adding a row to `images.csv` , specifying a publicly available URL to access the image in the `source_url` column (this may be a temporary GitHub repository, wikimedia commons or other publicly available webspace).
2. Providing the image at the specified URL for download.

Figure: GitHub repositories can be registered with CI service provider Travis-CI.

CI BUILD HISTORY

clld/tsammalex-data build failing
<https://travis-ci.org/clld/tsammalex-data>
Current Branches **Build History** Pull Requests Settings

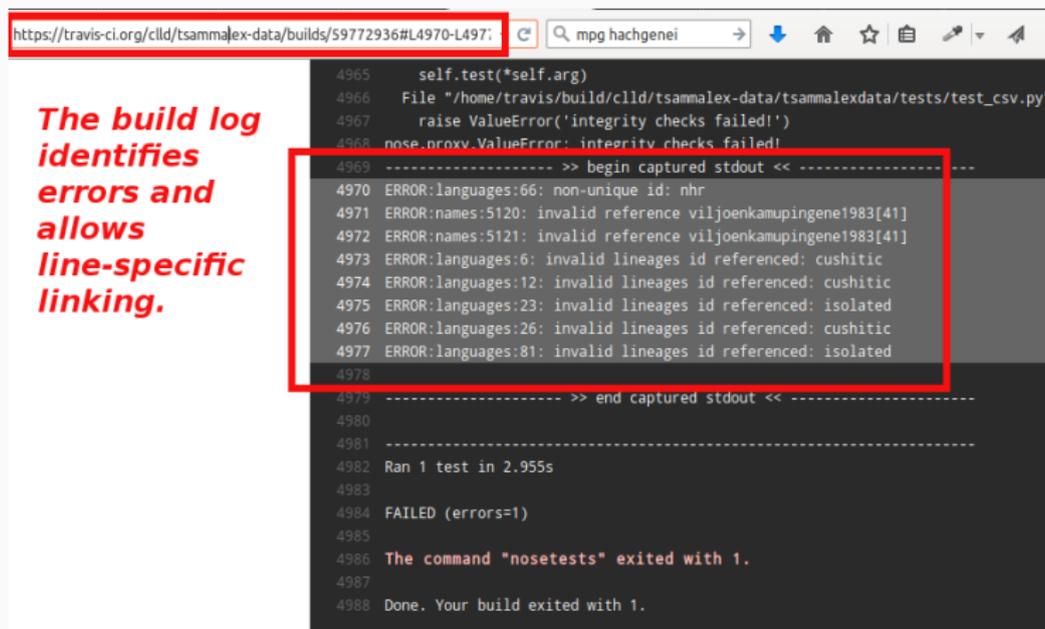
✓	master fixed problems xrotwang committed	# 213 passed 525ed3e	1 min 6 sec about 11 hours ago
✗	master names ids correction LenaSell committed	# 212 failed afbc572	1 min 9 sec about 13 hours ago
✗	master gwj categories and habitats added LenaSell committed	# 211 failed 6dbe981	1 min 9 sec about 13 hours ago
✗	master Comparative data from Bantu languages (c ChristfriedNaumann committed	# 210 failed 6d2927f	1 min 31 sec about 15 hours ago
✓	master updated taxa info from external sources xrotwang committed	# 209 passed 6cfc023	1 min 53 sec about 19 hours ago
✓	master Bibliography extended ChristfriedNaumann committed	# 208 passed 574ba97	1 min 15 sec a day ago

Link to commit on GitHub

Link to error log

Figure: The build history relates builds and repository changes.

CI BUILD LOG



The screenshot shows a web browser window displaying a Travis CI build log. The URL in the address bar is `https://travis-ci.org/clld/tsamalex-data/builds/59772936#L4970-L4971`. The browser's search bar contains the text "mpg hachgenei". The build log content is as follows:

```
4965 self.test(*self.args)
4966 File "/home/travis/build/clld/tsamalex-data/tsamalexdata/tests/test_csv.py"
4967 raise ValueError('integrity checks failed!')
4968 nose_nroxy.ValueError: integrity checks failed!
4969 ----- >> begin captured stdout << -----
4970 ERROR:languages:66: non-unique id: nhr
4971 ERROR:names:5120: invalid reference viljoenkamupingene1983[41]
4972 ERROR:names:5121: invalid reference viljoenkamupingene1983[41]
4973 ERROR:languages:6: invalid lineages id referenced: cushitic
4974 ERROR:languages:12: invalid lineages id referenced: cushitic
4975 ERROR:languages:23: invalid lineages id referenced: isolated
4976 ERROR:languages:26: invalid lineages id referenced: cushitic
4977 ERROR:languages:81: invalid lineages id referenced: isolated
4978
4979 ----- >> end captured stdout << -----
4980
4981 -----
4982 Ran 1 test in 2.955s
4983
4984 FAILED (errors=1)
4985
4986 The command "nosetests" exited with 1.
4987
4988 Done. Your build exited with 1.
```

The error messages from line 4970 to 4977 are enclosed in a red rectangular box. To the left of the log, there is a red text annotation: *The build log identifies errors and allows line-specific linking.*

Figure: Build log for error reporting.

CI: ADDRESSING BUILD ERRORS

fixed problems

master

xrotwang authored 12 hours ago 1 parent 921fb66 commit 525ed3e597e9004ea2bc1a387ddea07d44f5bd3c

Showing 4 changed files with 4 additions and 20 deletions. Unified Split

1 tsammalexdata/data/languages.csv View

```
@@ -63,7 +63,6 @@ ndo,Wambo/Ndonga,ndon1254 ,, bantu, -18,2,17.1,Southern Africa,eng;kua;umb
63 63 ngh-e,NiAng/Langeberg,nuuu1241,*NiAng (NiAng, Nluu, Nluuki, #Khomani) is a Tuu language spoken by probably less than 10
64 64 ngh-w,NiAng/Nluu,nuuu1241,*NiAng (NiAng, Nluu, Nluuki, #Khomani) is a Tuu language spoken by probably less than 10
65 65 nhr,Naro,naro1249,Naro (Nharo) is a Khoe-Kwadi language spoken by about 10 000 native speakers in Botswana and Namibia
66 66 -nhr,Naro,naro1249 ,,khoekwadi, -22.3,21.0,Southern Africa,eng;tsn;naq
67 66 nih,Nyiha,nyih1240 ,,bantu, -8.9,33.0,Eastern Africa,eng;smh
68 67 mnm-a,Taa/!Ama,xooo1239,*Taa (!Xoon, !Xoó) is a Tuu language spoken by about 3000 native speakers mainly in Botswana and Namibia
69 68 mnm-e,Taa/East !Xoon,xooo1239,*Taa (!Xoon, !Xoó) is a Tuu language spoken by about 3000 native speakers mainly in Botswana and Namibia
```

2 tsammalexdata/data/lineages.csv View

```
@@ -5,3 +5,5 @@ tuu,Tuu,,tuuu1241,,66ff33
5 5 kxa,Kx'a,,kxaa1236,,990099
6 6 germanic,Germanic,,germ1287,Indo-European,indo1319,dd0000
7 7 dogon,Dogon,,dogo1290,,ffffff
8 +cushitic,Cushitic,,cush1243,,000000
9 +isolated,Isolated,,,,00ff00
```

Figure: The URL to the build log could be used in a commit log to link changes back to the error report.

BUT WHAT IF GITHUB ...?

Does this introduce too much dependence on GitHub.com?

There are some mitigating factors:

- git is a *distributed scm*, thus each clone contains all the data!
- There are alternative git hosting platforms like BitBucket.
- and then there's ZENODO

ZENODO solves the longterm preservation and citability issue for GitHub repositories by

- archiving releases ("issues") of GitHub repositories
- assigning a DOI to each release

GLOTTOLOG 2.4 AT ZENODO

zenodo Research. Shared.

Search Communities Browse Upload Get started Sign In Sign Up

20 March 2015 **Software** **Open access**

glottolog-data: Glottolog database 2.4

Harald Hammarström; Robert Forkel; Martin Haspelmath; Sebastian Bank
[\(show affiliations\)](#)

Hammarström, Harald & Forkel, Robert & Haspelmath, Martin & Bank, Sebastian. 2015. Glottolog 2.4. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://glottolog.org>)

Name	Date	Size	
glottolog-data-v2.4.zip	20 Mar 2015	377.4 MB	Download

Available in **GitHub**

Publication date: 20 March 2015
DOI: [10.5281/zenodo.16245](https://doi.org/10.5281/zenodo.16245)
Keywords: [Linguistics](#)

Related publications and datasets:
Supplement to:
<https://github.com/clld/glottolog-data/tree/v2.4>

Figure: <http://dx.doi.org/10.5281/zenodo.16245>

If your data is code, treat it as such.

And yes, GitHub is the missing editorial backend of `clld` :).

`clld.org`

