

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/tsammaxlex-data> | mpg hachgenei

This repository | Search | Explore | Gist | Blog | Help | xrotwang | + | Settings | Fork

clld / **tsammaxlex-data** | Unwatch | 4 | Star | 0 | Fork | 2

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

256 commits | **History** | 0 releases | 6 contributors | **Who?**

branch: master | **tsammaxlex-data** / +

fixed problems		
xrotwang authored 12 hours ago		latest commit 525ed3e597
tsammaxlexdata	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 mediawiki	a year ago

Can I add my data?

Can I get your data?

SSH clone URL
git@github.com:clld

You can clone with HTTPS, SSH, or Subversion

Download ZIP

Figure: The data for Tsammaxlex is curated at clld/tsammaxlex-data.

A DATABASE JOURNAL

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

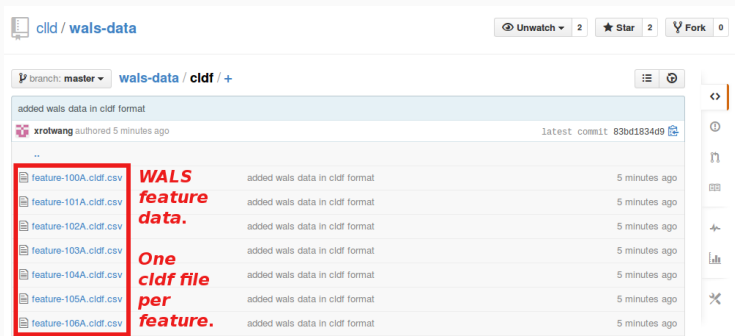
cldf the submission guidelines

pull request submission with open review

merge accepting a submission

release publication

EXAMPLE: WALS AS DATABASE JOURNAL



cldf / wals-data

Unwatch 2 Star 2 Fork 0

branch: master wals-data / cldf / +

added wals data in cldf format

xrotwang authored 5 minutes ago latest commit 83bd1834d9

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

WALS feature data.

One cldf file per feature.

Figure: WALS as database journal.

BROWSING CLDF ON GITHUB



branch: master wals-data / cldf / feature-1A.cldf.csv

xrotwang 8 minutes ago added wals data in cldf format

1 contributor *csv and GitHub go well together*

565 lines (564 sloc) | 129.153 kb

Raw Blame History

Search this file...

	LANGUAGE_NAME	VALUE	SOURCE_NAME
1	Abipón	Moderately small	Najlis 1966
2	Abkhaz	Large	Hewitt 1979
3	Aché	Small	Susnik 1974
4	Achumawi	Moderately small	Olmsted 1964;Olmsted 1966
5	Acoma	Large	Miller 1966
6	Adzera	Moderately small	Holzkecht 1973
7	Aghem	Average	Hyman 1979
8	Ahtna	Moderately large	Kari 1990
9	Alkana	Average	Hanke 1956;Alkhenvald and Dixon 1999[362-363]
10	Ainu	Small	Patrie 1982;Simeon 1969
11	Aizi	Average	Herault 1971
12	Akan	Average	Dolphyne 1988;Ladefoged 1964;Schachler and Fromkin 1968;Stewart 1967;Welmers 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

```
... 00 -0,0 +1,6 00
1 +LANGUAGE_NAME,VALUE,SOURCE_NAME,ID,LANGUAGE_ID,FEATURE_ID,SOURCE
2 +Abipón,Moderately small,Najlis 1966,http://wals.info/valuesets/145A-abl,http://wals.info/liguoid/lect/wals_code_al
3 +Abkhaz,Large,Hewitt 1979,http://wals.info/valuesets/145A-abk,http://wals.info/liguoid/lect/wals_code_abk,http://w
4 +Aché,Small,Susnik 1974,http://wals.info/valuesets/145A-ach,http://wals.info/liguoid/lect/wals_code_ach,http://wal
5 +Achumawi,Moderately small,Olmsted 1964;Olmsted 1966,http://wals.info/valuesets/145A-acm,http://wals.info/liguoid/
6 +Acoma,Large,Miller 1966,http://wals.info/valuesets/145A-aco,http://wals.info/liguoid/lect/wals_code_aco,http://wa
```

Figure: pull requests as submission mechanism.

LIST OF SUBMISSIONS

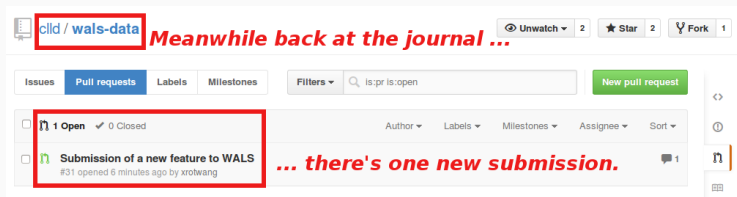


Figure: The list of open pull requests serves as the backlog for the editors.

OPEN REVIEW

Submission of a new feature to WALS #31

 **xrotwang** wants to merge 2 commits into `c11d: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 `clld: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.cldf.csv 00b4c08

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

added values for features 1A-3A for Language Aand Another Language #32 Edit

Open xrotwang wants to merge 1 commit into cldf:master from xrotwang:master

Conversation 0 Commits 1 Files changed 3 +3 -0

Showing 3 changed files with 3 additions and 0 deletions. ***A different kind of submission: adding values for existing features.*** Unified Split


- cldf/feature-1A.cldf.csv +1 -0
- cldf/feature-2A.cldf.csv +1 -0
- cldf/feature-3A.cldf.csv +1 -0

1 cldf/feature-1A.cldf.csv View

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

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

tsammax-data / tsammaxdata / data / uses.csv 

Newer Older







100644 29 lines (28 sloc) 0.711 kb			Raw	Normal view	History
	updated data xrotpang authored on Dec 5, 2014	7fd32c2	1	id,name,description	"Blame": per-line annotation of last modification.
	gwj names added, taxa added, min... LenaSell authored on Mar 12	3f79586	2	animalfeeding,animal feeding,	
	correction last submit; uses added LenaSell authored on Jan 15	926de1e	3	arts crafts,arts & crafts,	
			4	boats,boats,	
			5	clothingtextiles,clothing & textiles,	
			6	construction,construction,	
	updated data xrotpang authored on Dec 5, 2014	7fd32c2	7	cosmeticshygiene,cosmetics & hygiene,	
	correction last submit; uses added LenaSell authored on Jan 15	926de1e	8	dyetanning,dye & tanning,	
	gwj names, taxa lock deleted LenaSell authored on Feb 10	95dfd15			

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

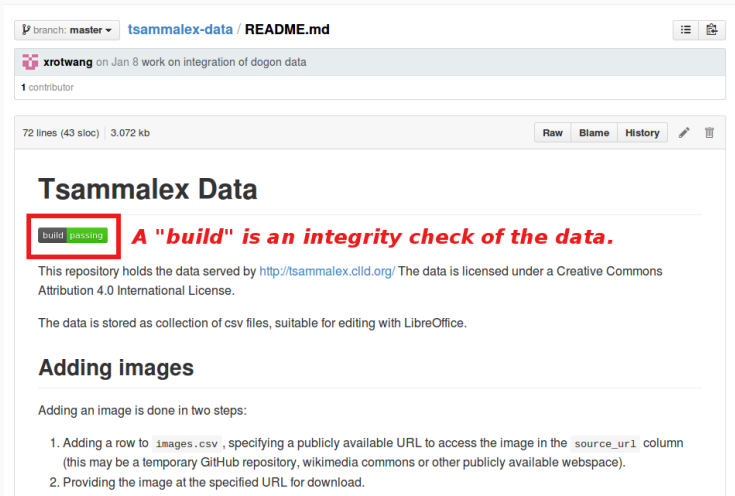
Let's go further borrowing best practices in software development.

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

CI FOR GITHUB



The screenshot shows the GitHub interface for the repository 'tsammalex-data' on the 'master' branch, specifically the 'README.md' file. At the top, it indicates the file was last updated by 'xrotwang' on Jan 8. Below this, it shows '1 contributor' and file statistics: '72 lines (43 sloc) | 3.072 kb'. Action buttons for 'Raw', 'Blame', and 'History' are visible. The main content of the README starts with the title 'Tsammalex Data'. Below the title, there is a green 'build passing' badge, which is highlighted by a red rectangle. To the right of the badge, a red text overlay reads: ***A "build" is an integrity check of the data.*** The text continues: 'This repository holds the data served by <http://tsammalex.cld.org/>. The data is licensed under a Creative Commons Attribution 4.0 International License.' It then states: 'The data is stored as collection of csv files, suitable for editing with LibreOffice.' Below this is a section titled 'Adding images' with the text: 'Adding an image is done in two steps:'. A numbered list follows: 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 webpace). 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

cld/tsammaxlex-data build failing

<https://travis-ci.org/cld/tsammaxlex-data>

Current Branches **Build History** Pull Requests Settings

Status	Commit Message	Commit Hash	Duration	Time Ago
✓	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 build log identifies errors and allows line-specific linking.

```
4965     self.test(*self.arg)
4966     File "/home/travis/build/clld/tsamalex-data/tsamalexdata/tests/test_csv.py"
4967     raise ValueError('integrity checks failed!')
4968 nose.proxy.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.
```

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 tsamalexdata/data/languages.csv View

63	63	ngn-e,Nläng/Langeberg,nuuu1241,"Nläng (Nläng, Nluu, Nluuki, #Khomani) is a Tuu language spoken by probably less than 10 000 native speakers in Botswana and Namibia"
64	64	ngn-w,Nläng/Nluu,nuuu1241,"Nläng (Nläng, Nluu, Nluuki, #Khomani) is a Tuu language spoken by probably less than 10 000 native speakers in Botswana and Namibia"
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,,bantuu,-8.9,33.0,Eastern Africa,eng;sw
68	67	nmn-a,Taa/I Ama,xooo1239,"Taa (I Xoon, I Xoó) is a Tuu language spoken by about 3000 native speakers mainly in Botswana and Namibia"
69	68	nmn-e,Taa/East I Xoon,xooo1239,"Taa (I Xoon, I Xoó) is a Tuu language spoken by about 3000 native speakers mainly in Botswana and Namibia"

2 tsamalexdata/data/lineages.csv View

5	5	kxa,Kx'a,,kxaa1236,,990000
6	6	germanic,Germanic,,germ1287,Indo-European,indo1319,dd0000
7	7	dogon,Dogon,,dого1290,,ffffff
8	8	+cushitic,Cushitic,,cush1243,,000000
9	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
[DOI: 10.5281/zenodo.16245](https://doi.org/10.5281/zenodo.16245)

Keyword(s):
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`

