



Glottolog 3.0

A collaborative, versioned catalog of languages and dialects

Robert Forkel

Poznan, 2016-09-15, A new era for cross-linguistic databases

Max Planck Institute for the Science of Human History

1. Glottolog
2. Open Source and Open Data
3. Glottolog and collaboration

Glottolog

What is Glottolog?

Glottolog is a comprehensive online catalog of languages.

- but also of dialects (less comprehensive, though)
- and a bibliography, linked to languages
- and a genealogical classification of languages

Glottolog 2.7 · Polish - Mozilla Firefox

clottolog 2.7 · Polish

glottolog.org/resource/languoid/id/pol11260

Language: Polish

Classification

- Indo-European (583)
 - Albanian (4)
 - Anatolian (10)
 - Armenian (3)
 - Balto-Slavic (22)
 - Eastern Baltic (2)
 - Prussian
 - Slavic (22)
 - East Slavic (5)
 - South Slavic (9)
 - West Slavic (8)
 - Czech-Slovak (3)
 - Czech (2)
 - Kashubian
 - Polish
 - Upper Silesian Polish
 - Sorbian (2)
 - Celtic (16)
 - Declian
 - Germanic (103)
 - Grieco-Phrygian (10)
 - Indo-Iranian (318)
 - Italic (88)
 - Lusitanian
 - Messapic
 - Thracian
 - Tocharian (2)

Subclassification references

- Sussens, Roland and Cubberley, Paul 2006

References

Showing 1 to 100 of 351 entries

| Details* | Name | Title | Year | Pages | Doctype | Provider | da |
|----------|---|--|------|-------|------------------------|------------------|----|
| more | S. A. Tokarev and N. N. Cebokarov 1964 | Narody Zarubečnej Evropy I | 1964 | 1018 | overview, ethnographic | th | |
| more | Dufka-Markowska, Anna 2010 | L'abîmte énonciative dans des textes de presse français et polonais: le conditionnel journalistique et ses traductions en polonais | 2010 | 815 | text | degruyter | |
| more | Barthnick, Barbara and Hansen, Björn and Klemm, | Grammatik des Polnischen | 2005 | 634 | | imperva, selfart | |

Why not Ethnologue or ISO 639-3?

So Glottolog is very much like Ethnologue. Why another one?

- The editorial process of Ethnologue is not fully transparent, the change request process for ISO 639-3 is slow.
- Ethnologue is behind a paywall, ISO 639-3 not fully integrated in the web at large and the semantic web in particular.
- Ethnologue and ISO 639-3 are not really targeted at academia, they have a different business model.

Glottolog wants to provide data like Ethnologue, but curated in a more transparent, collaborative, community owned way.

Open Source and Open Data

So, why are we looking at Open Source software development best practices to improve management of research data like Glottolog?

Open Source collaboration

Open Source software in the age of GitHub is a tremendous success story for worldwide online collaboration.

This is exactly the kind of collaboration we want to enable for data sets like Glottolog, which clearly

- profit from more curators

given enough eyeballs, all bugs are shallow (Linus' Law)

- "belong" to the academic community more than to any one institution, thus – given current funding schemes – will have to be transferred to a different owner at some point.

What spurred this surge in collaboration on Open Source software?

Remember: Licenses grant rights **people wouldn't usually have!**

Open Licenses which allow derived works are the basis of Open Source:

The ability to create derived works means that anyone can also modify the source or data as they see fit. In practice this means forking: creating a new custom version of some software, or a modified (corrected, reformatted) version of a dataset.

(Leigh Dodds)

Infrastructure for Open Source development

The default practice in the open source world is that code will be:

- published in a public repository*
- published with a complete version history [...]*
- published in an environment that supports transparent reporting of issues, bugs and suggestions*
- published in an environment that includes good documentation tools, such as a wiki*
- and, most importantly, published in an environment that allows forks and improvements to be folded back into the original project*

I'd go as far as suggesting that each of these are as important to our modern experience and expectations of open source, as the basic rights granted by open licences.

(Leigh Dodds)

Today, this infrastructure is GitHub.

Research data curation on GitHub: An example

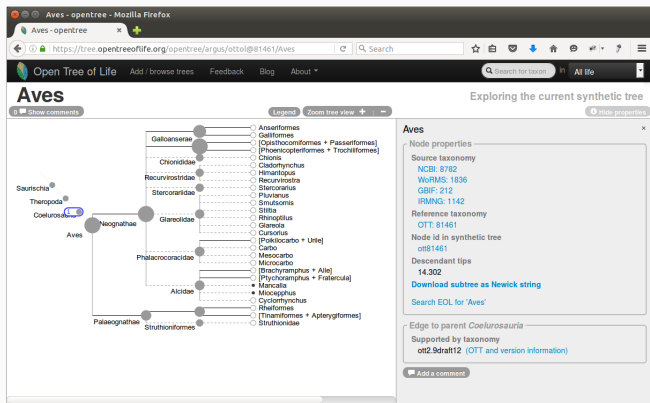


Figure 1: The data behind the Open Tree of Life is curated in a series of GitHub repositories

How do we turn Glottolog into Open Data?

We need to model Glottolog data in a way suitable for distributed version control systems.

- line-based text formats, i.e. text that can be meaningfully handled by **diff**
- BibT_EX for bibliography files
- INI files for languoid metadata.
- A directory tree to model the classification.
- Some tools to simplify manipulation of the language tree.
- An API to access the data in the repository programmatically.

```
@book{94863,  
  address    = {New York},  
  author     = {Sapir, Edward},  
  publisher  = {Harcourt and Brace},  
  title      = {Language},  
  year       = {1949},  
  bibtexkey  = {sapir_language1949},  
  inlg       = {English [eng]},  
  macro_area = {Africa},  
  src        = {wals},  
  srctrickle = {wals#5298}  
}
```

Listing 1: BibT_EX is used for reference data.

- Well supported in many bibliography management tools like
 - Zotero
 - jabref
- Our workflow is already adapted to it
- The (missing) details in the data model – e.g. no splitting of authors – align well with our messy data.
- We only use Bib_T_EX as container format – no \LaTeX in field values, but UTF-8 encoded text.

cldd/glottolog: INI files

```
# -*- coding: utf-8 -*-  
[core]  
name = Abinomn  
glottocode = abin1243  
hid = bsa  
level = language  
iso639-3 = bsa  
latitude = -2.92281  
longitude = 138.891  
macroareas =  
    Papunesia  
countries =  
    Indonesia (ID)  
  
[sources]  
glottolog =  
    Mark Donohue and Simon Musgrave 2007 (89329)
```

Listing 2: **INI** files are used for metadata on languoids.

cldd/glottolog: Why **INI** files?

- Good support (e.g. syntax highlighting) in many text editors.
- The programming language Python supports reading and writing **INI** files out-of-the-box.
- Format is extensible – new sections and options can be added any time without disrupting the processing pipeline.

```
$ tree --charset ASCII languoids/tree/abkh1242/abkh1243/  
    abkh1244/  
languoids/tree/abkh1242/abkh1243/abkh1244/  
|-- abkh1244.ini  
|-- abzh1238  
|   `-- abzh1238.ini  
|-- bzyb1238  
|   `-- bzyb1238.ini  
`-- samu1242  
    `-- samu1242.ini
```

3 directories, 4 files

Listing 3: A directory tree is used to model the language classification.

Glottolog and collaboration

fork Create your own copy of the data repository. The repository you forked from is also called **upstream**.

edit Change the data in your copy.

commit Register meaningful groups of changes in your copy.

pull request Propose merging your changes into upstream, i.e. **clld/glottolog**.

merge Incorporate changes from other forks of the repository.

Use cases: Transfer of ownership

Forks are essential for the open source software development model for another reason as well:

They allow for seamless transfer of ownership of codebases.

For Glottolog this means

- the data repository can be forked - any fork is as good as the original repos
- the code for the web application has an open license, can be run anywhere, and ingest data from any fork
- the only thing bound to an institution that has to be explicitly transferred (with consent of the owner) is the domain name **glottolog.org**

Use cases: Functionality built on the repository

Functionality built on top of the repository – rather than on top of the web application

- reduces traffic at **glottolog.org**
- works off-line
- works for forks, too
- thus, local changes can be incorporated in workflows right away

Use cases: Add "your" language

Working on varieties which are not in Glottolog?

- mint Glottocodes (using functionality built on top of the repository)
- add languoids to your fork of the repository
- use "your" Glottocodes in your data ...
- ...while waiting for "upstream" to incorporate your changes.

What happens when your changes are not accepted and merged into upstream?

- You either discard your changes, revert back to the status before and keep in synch with upstream;
- or you keep your changes,
 - and keep merging changes from upstream, resolving any conflicts resulting from your changes locally
 - or try to convince the community that your fork should become the new upstream repository (the "traditional meaning of fork in Open Source software development").

Example: Changing a language name – fork

The screenshot shows a GitHub repository page for 'clld / glottolog'. The repository name 'clld / glottolog' is highlighted with a red box. The 'Fork' button is also highlighted with a red box. The file path 'glottolog / languoids / tree / sino1245 / kuki1245 / naga1409 / zeme1241 / mara1379 / tkhu1238 / tkhu1238.ini' is highlighted with a red box. The file content is displayed below, showing a configuration file with various settings.

Repository: **clld / glottolog** (Unwatch 6, Star 1, Fork 5)

Branch: master

Find file Copy path

glottolog / languoids / tree / sino1245 / kuki1245 / naga1409 / zeme1241 / mara1379 / tkhu1238 / tkhu1238.ini

xrotwang Revert "julyupdates" ad92ae7 on 3 Aug

2 contributors

14 lines (11 sloc) 161 Bytes

```
1 # -*- coding: utf-8 -*-
2 [core]
3 name = T. Khullen
4 glottocode = tkhu1238
5 level = dialect
6 macroareas =
7     Eurasia
8 countries =
9
10 [altnames]
11 multitree =
12     T. Khullen
13
```

Example: Changing a language name – edit

The screenshot shows a web browser displaying the GitHub repository page for `shh-dlce / glottolog`. The repository is forked from `cidlglottolog`. The file `tkhu1238.ini` is selected, and its content is displayed in a code editor. The file path `glottolog / languoids / tree / sino1245 / kuki1245 / naga1409 / zeme1241 / mara1379 / tkhu1238 / tkhu1238.ini` is highlighted in the breadcrumb navigation. The code editor shows 14 lines of text, including comments and variable assignments. The 'Edit' button (pencil icon) is highlighted in the top right corner of the code editor. The commit history shows a revert by `xrotwang` on 3 Aug.

glottolog/tkhu1238.ini at master · shh-dlce/glottolog · Mozilla Firefox

glottolog/tkhu1238... x

GitHub, Inc. (US) | https://github.com/shh-dlce/glottolog/blob/master/languoids/tree/sino1245/l

This repository Search Pull requests Issues Gist

shh-dlce / glottolog
forked from cidlglottolog

Unwatch 1 Star 0 Fork 5

Code Pull requests 0 Wiki Pulse Graphs Settings

Branch: master Find file Copy path

glottolog / languoids / tree / sino1245 / kuki1245 / naga1409 / zeme1241 / mara1379 / tkhu1238 / tkhu1238.ini

xrotwang Revert "julyupdates" ad92ae7 on 3 Aug

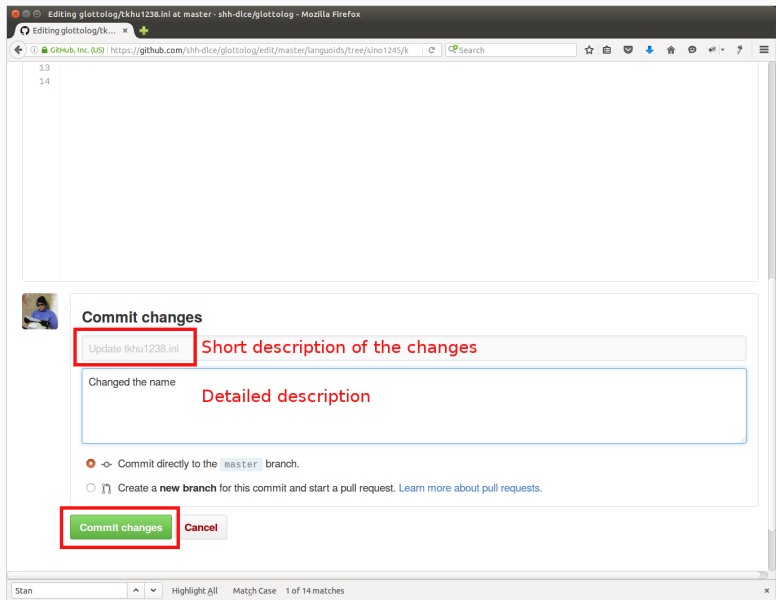
2 contributors

14 lines (11 sloc) 161 Bytes Raw Blame History Edit

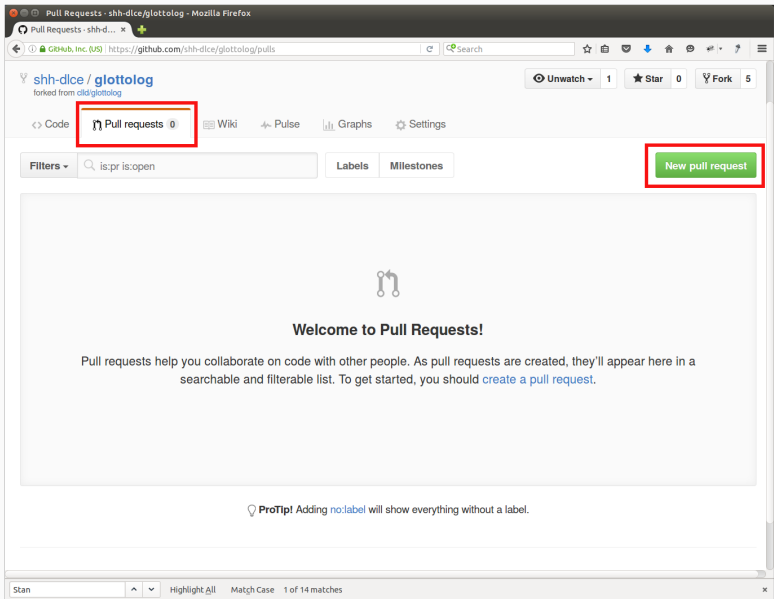
```
1 # -*- coding: utf-8 -*-
2 [core]
3 name = T. Khullen
4 glottocode = tkhu1238
5 level = dialect
6 macroareas =
7     Eurasia
8 countries =
9
10 [altnames]
11 multitree =
12     T. Khullen
13
```

Stan Highlight All Match Case 1 of 14 matches

Example: Changing a language name – commit



Example: Changing a language name – pull request



Example: Changing a language name – pull request

The screenshot shows a GitHub web interface for a pull request. At the top, the browser address bar shows the URL `https://github.com/clld/glottolog/compare/master...shh-dice:master`. The page title is "Comparing changes". Below the title, a message says "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](#)."

A red box highlights the branch selection area, which includes dropdowns for "base fork: cllid/glottolog", "base: master", "head fork: shh-dice/glottolog", and "compare: master". Below this, a green checkmark indicates "Able to merge. These branches can be automatically merged."

Another red box highlights the "Create pull request" button, with the text "Discuss and review the changes in this comparison with others." next to it.

Below the buttons, statistics show "1 commit", "1 file changed", "0 commit comments", and "1 contributor".

The "Commits on Sep 09, 2016" section shows a commit by "xrotwang" titled "Update tkhu1238.ini" with commit hash "21fde7e".

A message states "Showing 1 changed file with 1 addition and 1 deletion." To the right are "Unified" and "Split" view options.

The file diff for `languoids/tree/sino1245/kuki1245/naga1409/zeme1241/mara1379/tkhu1238/tkhu1238.ini` is shown. A red box highlights the change on line 3, where the language name is updated from `-name = T. Khullen` to `+name = T Khullen`. Other lines show `glottocode = tkhu1238` and `level = dialect`.

At the bottom, a search bar contains the text "Stan" and shows "1 of 14 matches".

Example: Changing a language name – pull request

The screenshot shows a web browser window with the URL `https://github.com/clld/glottolog/compare/master...shh-dice:master`. The page title is "Open a pull request". Below the title, it says "Create a new pull request by comparing changes across two branches. If you need to, you can also [compare across forks](#)."

The comparison section shows "base fork: clld/glottolog", "base: master", "head fork: shh-dice/glottolog", and "compare: master". A green checkmark indicates "✓ Able to merge. These branches can be automatically merged."

The pull request title is "Update tkhu1238.ini". The description is "Changed the name". The "Write" tab is active, showing a rich text editor with a toolbar. Below the editor, it says "Attach files by dragging & dropping or [selecting them](#)."

On the right side, there are sections for "Labels" (None yet), "Milestone" (No milestone), and "Assignees" (No one—assign yourself).

At the bottom of the pull request form, there is a checkbox "Allow edits from maintainers." with a link "Learn more". A red rectangle highlights the "Create pull request" button.

At the bottom of the page, there is a summary bar showing "1 commit", "1 file changed", "0 commit comments", and "1 contributor".

Example: Changing a language name – merge

Update tkhu1238.ini by xrotwang · Pull Request #24 · cld/glottolog - Mozilla Firefox

Update tkhu1238.ini... x

cld / glottolog Now we are in the "upstream" repository!

Unwatch 6 Star 1 Fork 5

Code Issues 12 Pull requests 2 Wiki Pulse Graphs Settings

Update tkhu1238.ini #24

Open xrotwang wants to merge 1 commit into cld:master from shh-dice:master

Conversation 0 Commits 1 Files changed 1 +1 -1

xrotwang commented just now Cross-Linguistic Linked Data member

Changed the name

Update tkhu1238.ini ... 21fde7e

Add more commits by pushing to the **master** branch on **shh-dice/glottolog**.

This branch has no conflicts with the base branch
Merging can be performed automatically.

Merge pull request or view command line instructions.

The "merge" button is only visible to editors of this repository

Labels: None yet

Milestone: No milestone

Assignees: No one—assign yourself

1 participant

Notifications: You're receiving notifications because you authored the thread.

Write Preview AA B i “ < > ☰ ☷ ☶ ↶ @

Stan Highlight All Match Case 1 of 14 matches

Example: Reviewing pull requests

The screenshot shows a GitHub pull request page for the repository 'cild / glottolog'. The pull request is titled 'Update kxoe1243.ini #12' and is in the 'Merged' state. It was merged by 'd97hah' on 15 Jun. The pull request description, highlighted with a red box, states: 'Changed lat/long to those of Rundu in the West Caprivi, which is nowadays the main settlement area of Kxwe-speakers; previous lat/long implied that Kxwe is a language of Zambia, which is not actually the case'. The pull request was merged into the 'cild:master' branch from the 'afehn:afehn-patch-1' branch. The 'Revert' button is also highlighted with a red box. The 'Write' and 'Preview' buttons in the comment section are also highlighted with a red box. The comment section includes a text input field with the placeholder 'Leave a comment' and a file upload area with the text 'Attach files by dragging & dropping or selecting them.'

Update kxoe1243.ini by afehn · Pull Request #12 · cild/glottolog - Mozilla Firefox

Update kxoe1243.in... x +

GitHub, Inc. (US) | https://github.com/cild/glottolog/pull/12

cild / glottolog

Unwatch 6 Star 1 Fork 5

Code Issues 12 Pull requests 1 Wiki Pulse Graphs Settings

Update kxoe1243.ini #12

Merged d97hah merged 1 commit into cild:master from afehn:afehn-patch-1 on 15 Jun

Conversation 0 Commits 1 Files changed 1 +2 -2

afehn commented on 16 May

Changed lat/long to those of Rundu in the West Caprivi, which is nowadays the main settlement area of Kxwe-speakers; previous lat/long implied that Kxwe is a language of Zambia, which is not actually the case

Update kxoe1243.ini 59f28ec

d97hah merged commit c5be6f6 into cild:master on 15 Jun

Revert

2 participants

Write Preview

Leave a comment

Attach files by dragging & dropping or selecting them.

Labels: None yet

Milestone: No milestone

Assignees: No one—assign yourself

Notifications: You're receiving notifications because you're subscribed to this repository. Unsubscribe

Example: Reclassifying Dogon

More complex changes – such as re-arranging the classification of a subgroup or whole language family – typically

- start out as **issues**
- which can be discussed
- and eventually may lead to pull requests
- issues can easily be referenced in
 - commit messages
 - pull request descriptions

Example: Reclassifying Dogon

The screenshot shows a GitHub issue page for the repository 'cld / glottolog'. The issue title is 'Re-classifying Dogon #20', which is highlighted with a red box. Below the title, it says 'Open' and 'xrotwang opened this issue on 27 Jul · 7 comments'. To the right of the title, there are buttons for 'Edit' and 'New issue', with the latter also highlighted by a red box. The issue content, also highlighted with a red box, is a comment from 'xrotwang' dated '27 Jul'. The comment text is: 'Classification of Dogon based on Heath's current analysis (language code [xxxxxxx] indicates missing language or branch; for the former there may already be a Glottolog code with different canonical label, except Tommo So, this language should be added to Glottolog; see [McPherson 2013](#):'.

```
'Dogon [dogo1299]'
-- 'East Dogon [xxxxxxx]'
---- 'Toro Tegu [toro1253][dtt]'
---- 'Jamsay [jams1239][djm]'
---- 'Toro So [toro1252][dts]'
---- 'Tomo Kan [tomo1243][dtm]'
---- 'Southeast Dogon [xxxxxxx]'
----- 'Ben Tey [bent1238][dbt]'
----- 'Bankan Tey [bank1259][dbw]'
----- 'Nanga [nang1261][nzz]'
----- 'Togo Kan [togo1254]'
----- 'Tengou Kan [xxxxxxx]'
----- 'Donno So [donn1239]'
----- 'Tommo So [xxxxxxx]'
-- 'West Dogon [west2779]'
---- 'Yanda Dom [yand1257][dym]'
```

On the right side of the issue, there are sections for 'Labels' (None yet), 'Milestone' (No milestone), 'Assignees' (No one—assign yourself), and '4 participants' (with three user avatars). Below these is a 'Notifications' section with an 'Unsubscribe' button, which is also highlighted with a red box. The text below the button says: 'You're receiving notifications because you authored the thread.'

Anyone with a GitHub account can open issues

<https://github.com/clld/glottolog>