Curating Cross-Linguistic Data with git and GitHub

Robert Forkel

MEaCoM 1, October 23–24, 2017, Alcanena

Max Planck Institute for the Science of Human History

Code is data

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.

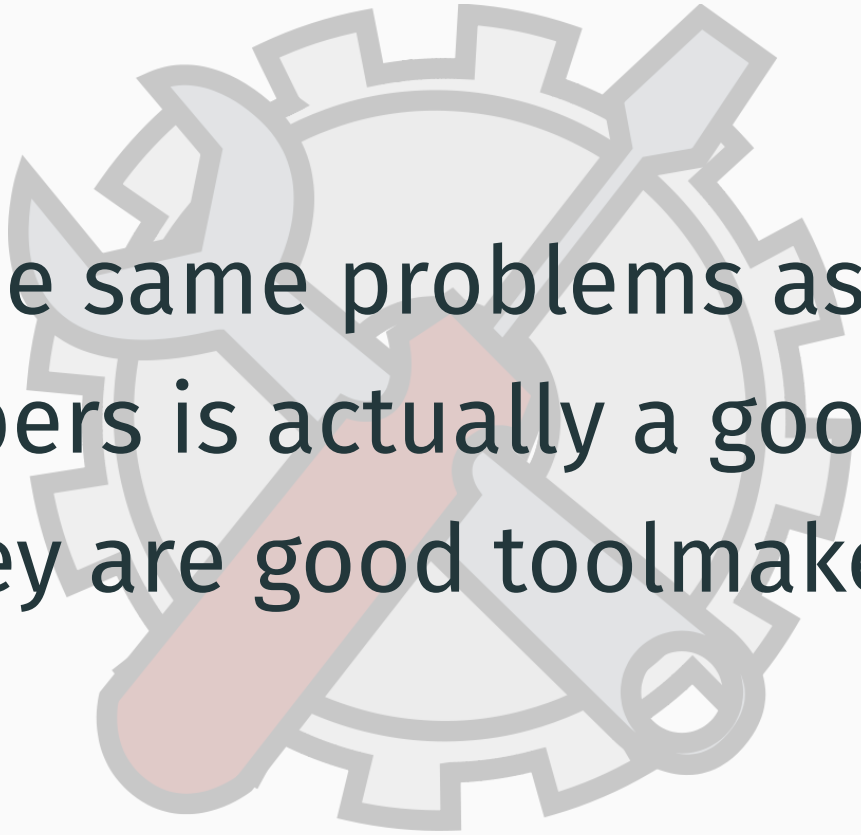
So homoiconicity is a desirable property of a programming language, because it allows for better tooling.

Data is code

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
- not really Big Data
- open (often)



Having the same problems as software developers is actually a good thing!
They are good toolmakers!

Data curation and version control

```
commit 1203f919cb811b7fb57fa350f8dc37c0e756a004
Author: kirbykat <kirbykat@users.noreply.github.com>
Date:   Thu Oct 5 00:21:47 2017 +0200

    correct glottolog id for ancient hebrew

commit 3df866f2965a1a9f7e4f67e8dd92833e28baf489
Author: Hans-Jörg Bibiko <bibiko@shh.mpg.de>
Date:   Wed Sep 13 09:38:49 2017 +0200

    fixed Labor and Lifecycle spelling

commit 8c06135d3fe7cb6f718e48b86e98cdc9e7bb41e9
Merge: 985d3d9 efaf853
Author: xrotwang <xrotwang@googlemail.com>
Date:   Wed Sep 13 09:21:31 2017 +0200

    updated variable definitions
```

Listing 1: Obviously, data curation and version control software are a perfect match!

Data curation with GitHub

Overview

Using line based text files for data and

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: Glottolog

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.

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`
- BibTEXfor 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.

clld/glottolog: BibTeX

```
@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 2: BibTeX is used for reference data.

clld/glottolog: Why BibTeX?

- 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 BibTeX as container format – no LaTeX in field values, but UTF-8 encoded text.

clld/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 3: **INI** files are used for metadata on languoids.

clld/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.

clld/glottolog: File-System Hierarchy

```
$ 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 4: A directory tree is used to model the language classification.

Glottolog and collaboration

The GitHub workflow

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 can be 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
- put an API on your data rather than on your web apps

Glottolog reference search

```
$ glottolog refsearch "author:Holton␣year:2003␣Tobelo"
ID                        Author          Year  Title
------------------------  ------------  ------  -------
mpieva:Holton2003Tobelo   Holton, Gary    2003  Tobelo
wals:2737                 Holton, Gary    2003  Tobelo
hh:g:Holton:Tobelo        Gary Holton     2003  Tobelo
langsci:Holton:03         Holton, Gary    2003  Tobelo
(4 matches)
```

Listing 5: Using the pure-python Whoosh search engine, we can provide full-text search on more than 300,000 Glottolog references from the command line (or any program that can "shell out")

Glottolog languoid search

```
$ glottolog langsearch Deutsch
4 matches
German Sign Language [germ1281] language
./sign1238/deaf1237/dgsi1234/germ1281/md.ini
Deutsche Gebärdensprache

Kaniet-Dempwolff [kani1283] language
./aust1307/nucl1752/.../west2532/anch1239/kani1283/md.ini
hh:hw:Dempwolff:Deutsch-Neuguinea** recorded

Old Saxon [olds1250] language
./indo1319/germ1287/.../alts1234/olds1250/md.ini
Altnieder-deutsch

German [stan1295] language
./indo1319/germ1287/.../high1287/stan1295/md.ini
Deutsch
```

Listing 6: Similar functionality is provided to search language information.

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.

Other useful line based text formats

- bagit: cataloging/packaging hierarchies of files
- csv: tabular data for version control
- csv packages with w3c: multi-table packages with foreign keys
- cldf: cross-linguistic datatypes built on the w3c spec for tabular data

Bagit

```
myfirstbag/
|-- data
|   \-- 27613-h
|       \-- images
|           \-- q172.png
|           \-- q172.txt
|-- manifest-md5.txt
|     49afbd86a1ca9f34b677a3f09655eae9 data/27613-h/images/q172
    .png
|     408ad21d50cef31da4df6d9ed81b01a7 data/27613-h/images/q172
    .txt
\-- bagit.txt
      BagIt-Version: 0.97
      Tag-File-Character-Encoding: UTF-8
```

Listing 7: `https://en.wikipedia.org/wiki/BagIt`

CSV

correct glottolog id for ancient hebrew

master

Browse files

kirbykat committed 6 days ago 1 parent af0c317 commit 1203f919cb811b7fb57fa350f8dc37c0e756a004

Showing **1 changed file** with **1 addition** and **1 deletion**. Unified Split

2 datasets/SCCS/societies.csv View

```
@@ -53,7 +53,7 @@ SCCS57,xd576,Kurd,cent1972,Kurd (SCCS57),,1951,Lang_assignment_change_notes,Kurd
 SCCS55,xd577,Abkhaz,abkh1244,Abkhaz (SCCS55),,1880,Lang_assignment_change_notes,Abkhaz (RI03),http://ehrafworldculture
 SCCS47,xd581,Turks,nucl1301,Turks (SCCS47),,1950,Lang_assignment_change_notes,Turks (MB01),http://ehrafworldcultures.y
 SCCS46,xd588,Rwala,east2690,Rwala Bedouin (SCCS46),Rwala Bedouin,1913,Lang_assignment_change_notes,Rwala (MD04),http:/
-SCCS44,xd589,Hebrews,hebr1245,Hebrews (SCCS44),,-621,Lang_assignment_change_notes,,,31.18,34.92,31.18,34.92,Original
+SCCS44,xd589,Hebrews,anci1244,Hebrews (SCCS44),,-621,Lang_assignment_change_notes,,,31.18,34.92,31.18,34.92,Original
 SCCS45,xd590,Babylonians,akka1240,Babylonians (SCCS45),,-1750,Lang_assignment_change_notes,,,32.58,44.75,32.58,44.75,(
 SCCS59,xd600,Punjabi,panj1256,Punjabi (West) (SCCS59),Punjabi (West) ,1950,Lang_assignment_change_notes,,,32.5,74.0,32
 SCCS58,xd605,Basseri,bass1257,Basseri (SCCS58),,1958,Lang_assignment_change_notes,Basseri (MA10),http://ehrafworldcult
```

Figure 1: CSV plays well with version control and GitHub.

Bells and whistles

Continuous integration

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

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

CI Build history

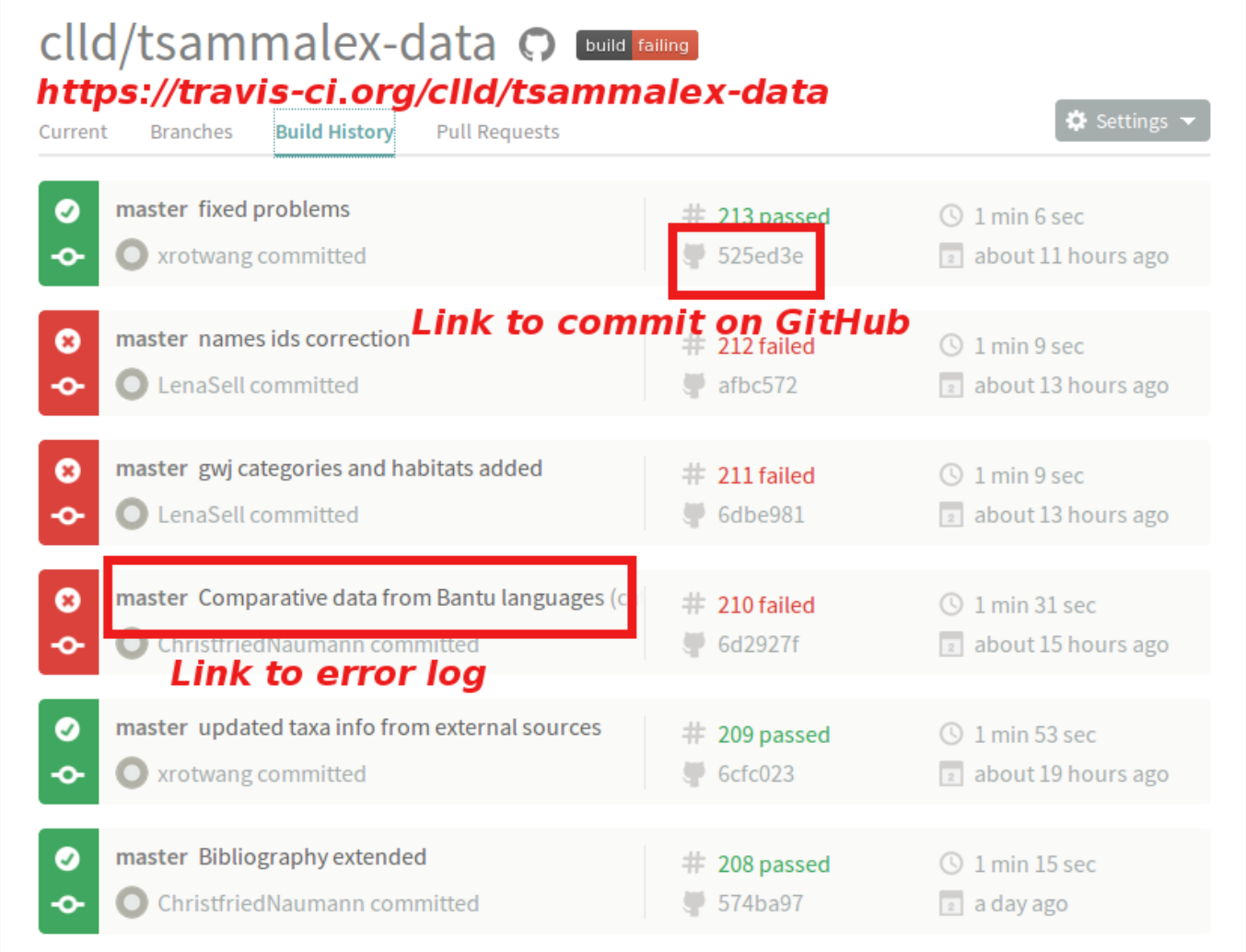


Figure 3: The build history relates builds and repository changes.

CI BUILD LOG

https://travis-ci.org/clld/tsammalex-data/builds/59772936#L4970-L4977

mpg hachgenei

The build log identifies errors and allows line-specific linking.

```
4965     self.test(*self.arg)
4966   File "/home/travis/build/clld/tsammalex-data/tsammalexdata/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 4: Build log for error reporting.

CI: ADDRESSING BUILD ERRORS

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

fixed problems

master

Browse files

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
 ngh-e,Nǁang/Langeberg,nuuu1241,"Nǁang (Nǁng, Nǀuu, Nǀuuki, ǂKhomani) is a Tuu language spoken by probably less thar
 ngh-w,Nǁang/Nǀuu,nuuu1241,"Nǁang (Nǁng, Nǀuu, Nǀuuki, ǂKhomani) is a Tuu language spoken by probably less than 10 r
 nhr,Naro,naro1249,Naro (Nharo) is a Khoe-Kwadi  language spoken by about 10 000 native speakers in Botswana and Nar
-nhr,Naro,naro1249 ,,khoekwadi,-22.3,21.0,Southern Africa,eng;tsn;naq
 nih,Nyiha,nyih1240 ,,bantu,-8.9,33.0,Eastern Africa,eng;swh
 nmn-a,Taa/!Ama,xooo1239,"Taa (!Xoon, !Xóõ) is a Tuu  language spoken by about 3000 native speakers mainly in Botswa
 nmn-e,Taa/East !Xoon,xooo1239,"Taa (!Xoon, !Xóõ) is a Tuu  language spoken by about 3000 native speakers mainly in
```

2 tsammalexdata/data/lineages.csv View

```
@@ -5,3 +5,5 @@ tuu,Tuu,,tuuu1241,,,66ff33
 kxa,Kx'a,,kxaa1236,,,990099
 germanic,Germanic,,germ1287,Indo-European,indo1319,dd0000
 dogon,Dogon,,dogo1299,,,ffffff
+cushitic,Cushitic,,cush1243,,,000000
+isolated,Isolated,,,,,00ff00
```

Figure 5: 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

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

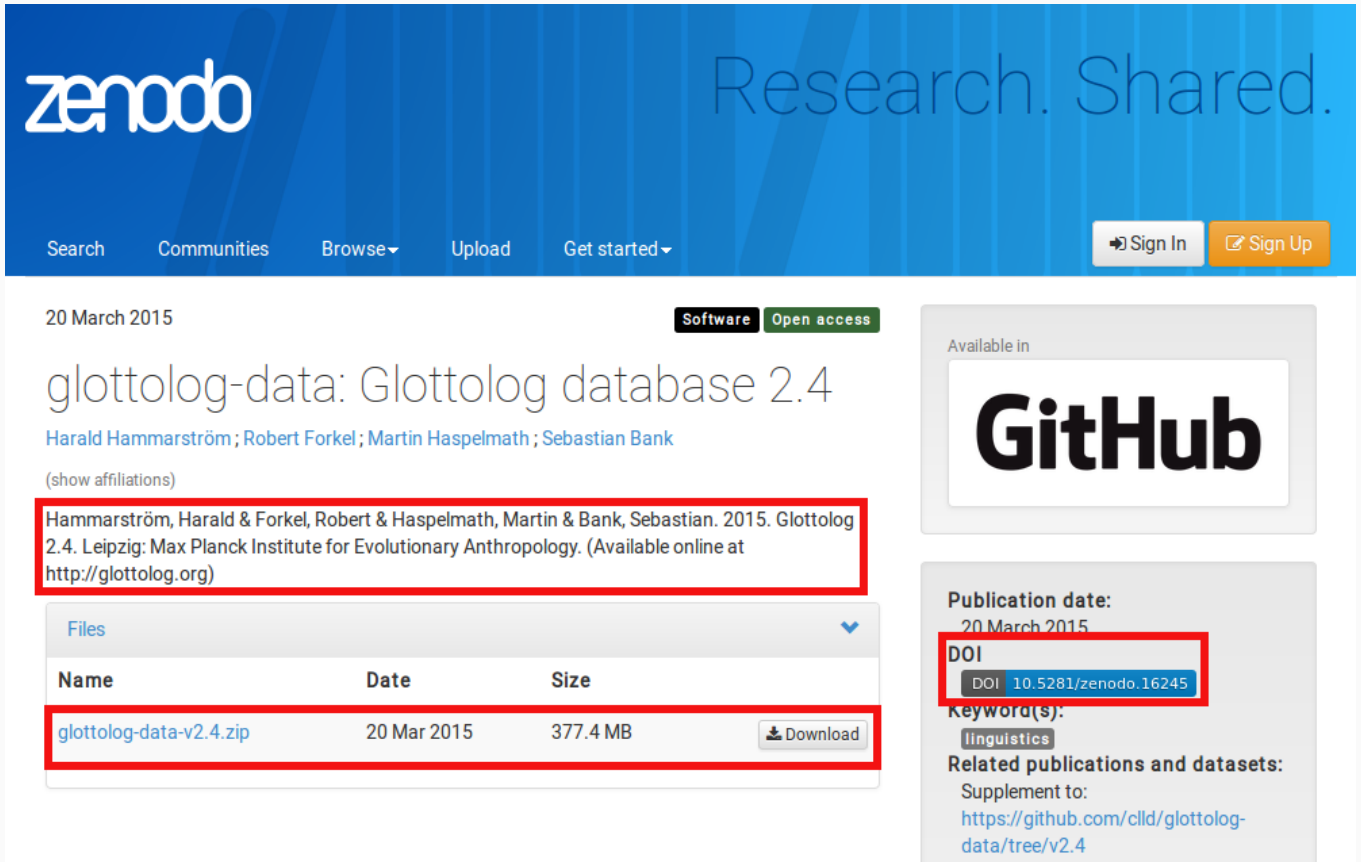


Figure 6: `http://dx.doi.org/10.5281/zenodo.16245`

Summary

If your data is code, treat it as such.

And yes, GitHub is the missing editorial backend of your system.