Glottolog 3.0: A collaborative, versioned catalog of languages and dialects

Robert Forkel, Max Planck Institute for the Science of Human History, forkel@shh.mpg.de

Glottolog is a catalog of languages and a comprehensive bibliography for Diversity Linguistics. While Glottolog was always edited by scholars, release 3.0 will introduce a much more powerful and convenient collaboration model to allow the community of linguists to get involved in the curation of the Glottolog data.

The enhanced collaborative features are made possible by leveraging the distributed version control software git and the public git-hosting platform GitHub. Borrowing tools typically used for software development allows us to build on well-established and well-documented best practices which make collaborative curation of research data more transparent.

Using version control software means that tracking the history and provenance of the data is easy. The paradigm of "distributed version control" - i.e. each instance or clone of the data repository (called clld/glottolog following the GitHub naming scheme <organization>/<repository>) functions as a full-fledged repository itself - allows a well-understood procedure to integrate new data, following the fork-and-pull-request workflow which is now standard for open source software projects.

In particular, this allows a new way to incorporate new Glottocodes:

- Anyone can add languages (in their own fork of the repository), and even assign Glottocodes for these.
- These changes can then be submitted back to clld/glottolog via pull requests.
- Even if these requests are not accepted, others could merge them into their forks, exploiting the fact that the system is distributed.

This is in clear contrast to the ISO 639-3 model, where researchers can only submit change requests and then wait. With the Glottolog system, they have a fully functional repository including their changes right away. Only if changes are rejected by clld/glottolog will they have to decide whether to discard their changes to stay compatible with clld/glottolog or to keep the changes, thus offering an alternative catalog of the world's languages (which will still be technically compatible with software built on top of the data repository, e.g. the web application serving http://glottolog.org).

Thus, the possibility of "forking" a repository also goes a long way towards solving one of the biggest problems in longterm research data curation: What if funding runs out? Any successor willing to take over can simply fork a repository and if the cmmunity follows, a perfectly transparent transfer of ownership can be implemented.