The Cross-Linguistic Linked Data project

Robert Forkel

Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6, D-04103 Leipzig
robert_forkel@eva.mpg.de

Abstract

The *Cross-Linguistic Linked Data project (CLLD – `http://clld.org`) helps record the world's language diversity heritage by establishing an interoperable data publishing infrastructure. I describe the project and the environment it operates in, with an emphasis on the datasets that are published within the project. The publishing infrastructure is built upon a custom software stack – the `clld` framework – which is described next. I then proceed to explain how Linked Data plays an important role in the strategy regarding interoperability and sustainability. Finally I gauge the impact the project may have on its environment.*

Keywords: *Linked Data, Linguistics, Software, Typology*

1. Cross-Linguistic data – the status quo

For the purposes of this paper I define cross-linguistic data as either data on many languages, or as data about under-resourced languages. I also restrict it to textual data.[1] Thus, this data will mostly come in the form of wordlists, dictionaries, phoneme inventories, typological surveys, small collections of glossed text, grammars, or bibliographies.

This kind of data is the result or forms the basis of much of the work being done at the department of linguistics of the Max Planck Institute for Evolutionary Anthropology (MPI EVA) in Leipzig which is one of the centers of what may be called "language diversity research".

Since data collection via fieldwork is well respected in this community there is not a shortage of data; often this data is not very complex but even more often it is unpublished. And even if this data is published, it may have ended up as a printed grammar or dictionary,[2] which – given the fact that these are reference works – is clearly inferior to a digital medium.[3]

Similar observations can be made for typological databases. While many presentations at ALT 10[4] used data from WALS[5] and complemented it with the author's own data, typically this complementary data is not published.

So there is quite a bit of seemingly low-hanging fruit out there: simple data waiting to get published.

2. The CLLD project

Cross-Linguistic Linked Data (CLLD) is a project funded by the Max Planck Society for four years, setting out to pick this fruit by establishing data publishing infrastructure.

We try to do so by:

- closing the gap between data creation and data publication by making publication easy and attractive,
- overcoming the disconnect between data creators and data consumers[6],
- providing the infrastructure in a sustainable way.

2.1. The datasets

Most datasets under consideration right now have been compiled by or in association with the department of linguistics at the MPI EVA:

WALS The World Atlas of Language Structures is now online in its third implementation.

APiCS The Atlas of Pidgin and Creole Language Structures is a typological database modeled after WALS but focussing on pidgins and creoles.

ASJP (to be published in 2014) The Automated Similarity Judgement Program has collected almost 7000 small wordlists of languages and varieties from all over the world.

IDS (to be published in 2014) The Intercontinental Dictionary Series is a collection of extensive wordlists (ca. 1300 items) collected for a curated set of meanings covering more than 220 languages.

AfBo A world-wide survey of affix borrowing describes 101 cases of affix borrowing from one language into another.

WOLD The World Loanword Database contains extensive vocabularies (similar to IDS) for 41 languages annotated for loanword status and source words (Haspelmath and Tadmor, 2009).

Glottolog Glottolog is a language catalog and bibliographical database, comprising close to 8000 languages and more than 200000 bibliographical records (Nordhoff, 2012).

[1] There does not seem to be much of a Linked Data story for multimedia content anyway.

[2] Publishing printed grammars and dictionaries seems to get more and more difficult, though (Haspelmath, 2014).

[3] Re-publication or aggregation of data from printed media in digital form is fraught with all the license and copyright issues and the interpretations thereof in the community (Austin, 2011).

[4] The 10th Biennial Conference of the Association for Linguistic Typology, Leipzig August 2013

[5] The World Atlas of Language Structures

[6] It is an interesting observation that at ALT 10 the typical presenters of papers working with quantitative methods on linguistic datasets were disjoint from the people creating such databases.

But CLLD also provides infrastructure to publish datasets originating outside the MPI EVA:

- Two data journals (one for dictionaries and one for typological databases) will be started in 2014 which are open to submissions. These journals will serve the double purpose of
  - allowing publication of datasets referenced in "traditional" publications (as is increasingly required by funders),
  - applying the traditional model of peer-reviewed publications to data, thereby incentivizing researchers through recognition.
- Bigger datasets can become part of CLLD following an "edited series" publication model. There are already two datasets in this category:

  eWAVE The electronic World Atlas of Varieties of English is a typological database containing information on 76 varieties of English (and highlighting the fact that ISO 639-3 is not sufficient to identify language varieties).

  PHOIBLE (to be published in 2014) The The Phonetics Information Base is a large collection of phoneme inventories for languages from all over the world.
- Last but not least the `clld` framework,[7] upon which all publications are built, is open source software and can be freely reused; i.e. institutions or individuals can build applications based on the `clld` framework to host and publish their own databases.

2.2. The `clld` framework

Recognizing that the field of interoperable linguistic data publication is still in its beginnings[8] adaptability and in general an iterative approach is called for. Thus, we aim to "standardize" on a lower level, namely on the publication platform; in doing so we hope to make published resources – i.e. the interface to the data – more adaptable.[9] So our aim is at the same time more modest than semantic interoperability and more ambitious, because the platform is open to serving non-RDF serializations of resources should these become de-facto standards.

In the first year of the project[10] a cross-linguistic database framework – the `clld` framework[11] – has been developed, which will be the focus of the following sections. Publishing datasets as `clld` applications should be seen as a perfect basis for publishing it as Linked Data while at the same time publishing it in a more traditional way (with respect to Web-publishing). It is also a good way to extend the uniformity of the interface from the machine readable data to the user interface accessed with the browser. While I understand the strength of the Linked Data approach to publishing, being able to also put an attractive human user interface on top of a dataset must not be underestimated when it comes to convincing linguists to open up their data. Thus the `clld` framework provides a

- a common core data model,
- a basic API built on Linked Data principles
- and what could be termed a "reference implementation" of a dataset browser as user-friendly interface for humans.

2.2.1. The data model

The design of the data model was guided by three principles:

1. All the target datasets have to "fit in" without loss.
2. The data model must be as abstract as necessary, as concrete as possible.
3. The data model must be extensible.

Note that these guidelines mirror the requirements set forth in Section 6.1 of Monachesi et al. (2002) for a *linguistic "metalanguage"*, ensuring unified access to typological databases. It turns out that most of the datasets we encountered thus far can be modeled using the following concepts.[12]

Dataset holds metadata about a dataset like license and publisher information.

Language often rather a languoid in the sense of Good and Cysouw (2014).

Parameter a feature that can be coded or determined for a language – e.g. a word meaning, or a typological feature.

ValueSet set of values measured/observed/recorded for one language and one parameter, i.e. the points in the Language-Parameter-matrix.

Value a single measurement (different scales can be modeled using custom attributes).[13]

Unit parts of a language system that are annotated, such as sounds, words or constructions.

UnitParameter a feature that can be determined for a unit.

UnitValue measurement for one unit and one unitparameter.

[7] `https://github.com/clld/clld`

[8] Although this may have been so for almost 10 years.

[9] It should be noted that this is not the first attempt at standardization of a software stack for cross-linguistic databases (Monachesi et al., 2002); but today's options for community driven development of open source software promise to make a real difference.

[10] `http://clld.org/2014/01/03/new-year.html`

[11] Spelled in lowercase conforming to common rules for names of Python software packages

[12] Or as Dimitriadis (2006) put it (further corroborating our experience): "survey databases are all alike".

[13] The only assumption the core data model makes about values is that they have a name, i.e. a textual description; datatypes for values can be implemented as application-specific extensions of this core model.

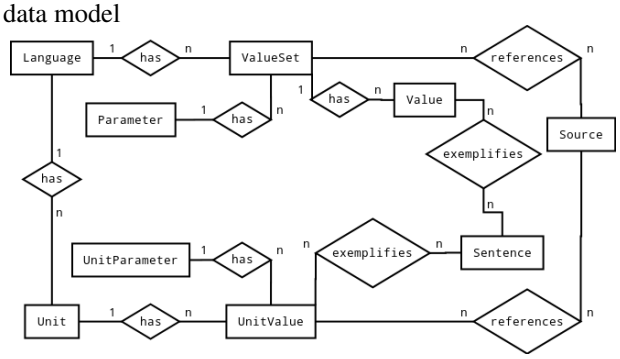
Source pointer to secondary literature - e.g. bibliographical records.

Sentence a small piece of text, preferably in the form of interlinear glossed text[14] according to the Leipzig Glossing Rules.[15]

Contribution a collection of ValueSets that shares provenance information, e.g. authorship.[16]

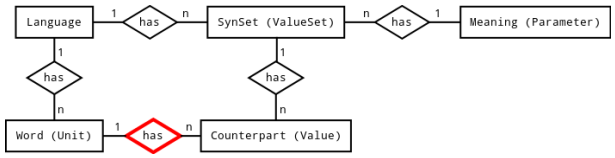
The relations between these entities are shown in Figure 1. Note that Dimitriadis' *Construction* maps to our *Unit* and *Example* to *Sentence* (Dimitriadis, 2006, p. 15).

Figure 1: Entity-relationship diagram of the CLLD core data model



In a concrete incarnation this core data model can be interpreted as shown in Figure 2. Note the additional relation between *Word* and *Counterpart* which is not present in the core model. The `clld` framework uses the *joined table inheritance* feature of the `SQLAlchemy` package to transparently add attributes to entities of the core data model. (see section 2.2.2.).[17]

Figure 2: Entity-relationship diagram of the WOLD data model; *SynSet*s are sets of synonyms, a *Counterpart* is an instance of the many-to-many relation between Words and Meanings in the sense of Haspelmath and Tadmor (2009).



2.2.2. The implementation

CLLD applications are implemented using the `clld` framework.[18] This framework in turn is based on the python packages `pyramid` and `SQLAlchemy` and allows building web applications accessing a relational database. Using an RDF graph database as main storage was out of the question because of its non-standard requirements in terms of deployment, administration and maintenance, which would conflict with the strategy for sustainability described in section 2.3.

These technology choices offer the following two essential mechanisms for extensibility:

1. The joined table inheritance[19] model provided with `SQLAlchemy` allows for transparent extension of core database entities. For each core entity a CLLD application may define an extended entitity, adding attributes and relations. Accessing the database using `SQLAlchemy` makes sure that whenever the core entitity is queried an instance of the extended entity is returned.

2. The Zope component architecture[20] within pyramid[21] provides an implementation of concepts like interface and adapter, which in turn make it possible to provide default behavior for entities which works with extended entities as well and can easily be overridden by registering custom behavior.

Using these mechanisms deviations in terms of data model or user interface are possible, but the default behavior[22] should be good enough most of the time (at least for the data consumed by machines only).

2.3. Sustainability: The idea of graceful degradation of service

Lacking longterm institutional/financial support, a project may employ several methods to gain sustainability:

1. Make fundraising part of the project activities.

2. Make transfer of ownership easy.

With respect to the CLLD databases the latter means that we always have to take into consideration what running such a service entails. From our own experience it seems clear that running interactive web applications without the ability to further develop the software will lead to dysfunctional applications quickly (within a few years).

But since this scenario is not unlikely in the event of a transfer of ownership, we would like to define a more stable, and more easily maintainable level of service for our applications. Thus, we use the Linked Data principles to define a lowest level of service we want to guarantee for CLLD applications. Doing so means that running CLLD applications can be as cheap as hosting static files on a web server (and possibly keeping domain registrations valid).

[14] `http://en.wikipedia.org/wiki/Interlinear_gloss`

[15] `http://www.eva.mpg.de/lingua/resources/glossing-rules.php`

[16] Thus, a CLLD dataset is similar to an edited volume in that it may aggregate data from multiple contributors.

[17] It should be noted that this data model provides sufficient structure to allow conversion to the RDF model for wordlists proposed by Poornima and Good (2010).

[18] `https://github.com/clld/clld`

[19] `http://docs.sqlalchemy.org/en/latest/orm/inheritance.html`

[20] `http://www.muthukadan.net/docs/zca.html`

[21] `http://docs.pylonsproject.org/projects/pyramid/en/latest/narr/zca.html`

[22] By default, each resource class comes with a list view and a detailed view for each instance, which in turn can be serialized in various formats like JSON, RDF+XML, etc.

Essentially we are following the Linked Data principles to provide a REST API for our datasets that is easy to maintain. Notably, this API is already used today by search engines, so this aspect of the service will survive also in the lowest level. This also means that we hope *sustainable operability* as defined by Windhouwer and Dimitriadis (2008) can be provided on top of the Linked Data stack, in particular on top of public sparql endpoints. Thus, we propose Linked Data to serve as the *Integrated Data and Documentation Format* described in Windhouwer and Dimitriadis (2008) with the additional benefit of a well-defined access protocol.

The `clld` framework will provide an "emergency exit" feature, which will create a set of files (corresponding to the list and detailed views in various formats as described above) in an appropriate directory structure to be put on a vanilla webserver. This can be done by enumerating the resource types, instances and available representations.

So while Linked Data is still not the way many researchers interested in our datasets[23] actually do or want to access data (at least if they can get away with csv instead), there is something to be gained for the developer: A stable API across phases of deployment which can be used by any additional services built on top of the data.

2.4. Linked Data

As described above, Linked Data plays an important role in the strategy of the CLLD project. In the following sections I describe our design choices regarding the implementation of Linked Data principles for the publication of CLLD datasets.

2.4.1. URLs and resources

We do not distinguish *things* from *Web documents* as recommended by Sauermann and Cyganiak (2008), because the solutions to achieve this conflict with our requirements for easy hosting of the lowest level of service outlined in section 2.3. These conflicts are also echoed in the list of practical limitations given in Tennison 2011 (Tennison, 2011). Arguably, using a concept of languages as sets of doculects (following Good and Cysouw (2014)), the *thing* can to some extent be identified with the web document describing it anyway.

While each RDF resource in CLLD datasets links to its originating dataset, and this dataset is described by a VoID description (see below), perhaps the most important bit of provenance information is the domain part of a resource's identifying URL.[24] Each dataset can employ additional schemes of conveying additional provenance information, though, like adding a version history. It is an explicit goal of the project to keep the resource URLs stable and resolvable for as long as possible, thus, we intend our URLs to be "cool" in the old sense, too, and more generally to fulfill the "social contract" between publisher and user outlined in Hyland et al. (2014).

All entities in the clld data model (see section 2.2.1.) map to resource types in the RDF formulation of this model. Since all entities have a local identifier, a name and a description, and each CLLD dataset is served from a distinct domain, we already have all the pieces necessary to fulfill basic requirements for RDF descriptions.[25]

2.4.2. VoID

The `clld` framework provides a VoID dataset description for each dataset. This description is populated from the metadata specified for the dataset, but also using the knowledge the framework has about its entities and capabilities. E.g. the VoID description for Glottolog[27] describes partitions of the dataset into entitity-type specific subsets (`void:Dataset`), and points to data dumps for these, because the number of resources would make accessing the data via crawling (while still possible) time consuming.

The VoID description and the backlinks of each resource to the dataset are used to provide provenance and license information for each resource. Following the recommendations for deploying VoID descriptions in (Alexander et al., 2011), the description is available at the path `/void.ttl` of CLLD applications as well as via content negotiation at the base URL.

2.4.3. HTTP

The `clld` framework uses content negotiation to make sure that RDF resources can be accessed right now just as they would in the "plain file on webserver" scenario.

HTTP link headers are used to identify canonical URLs and alternative representations.

While this feature might not survive in the lowest level of service (unless some custom webserver configuration is added), it shows the capability of the framework to enumerate the URL space of its resource graph.

2.4.4. Linking with other resources and vocabularies

Linking to resources outside the CLLD universe is clearly in need of further investigation. Linking to dbpedia and lexvo based on ISO 639-3 codes of languages is possible. Linking sources to bibliographical records e.g. in WorldCat is hampered by the fact that identification of matching records is error prone and not doable "by hand" for our typical source collections with more than 1000 records.

It should be noted, though, that some of our datasets carry the potential to serve as hubs in the Linked Data cloud themselves, and do so within the CLLD sub-cloud:

- Glottolog as language catalog and comprehensive bibliography. The desirability of alternative language catalogs (in addition to Ethnologue or ISO 639-3) is described in Haspelmath (2013) and can easily be seen looking at a dataset like eWAVE or APiCS, where many of the languages under investigation are not included in either Ethnologue or ISO-639-3.

[23] Most of the datasets under consideration here are more interesting for typologists than for computational linguists.

[24] Since CLLD datasets can be aggregations of multiple contributions, additional – more fine grained – provenance information is typically available, but for purposes of quality assessment the overriding factor will often be the fact that a ValueSet is part of an aggregation compiled under editorial control.

[25] e.g. as specified for bio2rdf in its RDFization-Guide[26]

[27] http://glottolog.org/void.ttl

- IDS and WOLD as providers of semi-standard comparison meanings for the creation of wordlists, i.e. as concepticons in the sense of Poornima and Good (2010).[28]

While the comprehensive ambition of the CLLD project might warrant the creation of a CLLD ontology, we have refrained from creating one. This is in large part due to the lack of use cases (evidenced by lack of usage) for the RDF data.
In an environment where the preferred data exchange format is still csv, I did not want to venture on an undertaking that might leave us with a bunch of vocabulary URLs to maintain which no one uses. Thus, CLLD's current RDF model reuses fairly generic terms from the most common vocabularies: `dcterms`, `skos`, `foaf`, `wgs84`. Due to the extensible architecture provided by the `clld` framework described in section 2.2.2. each CLLD application is in complete control of the RDF graphs of its entities, though.[29]

3. Where does this get us?

With WALS, APiCS, WOLD, Glottolog and some more datasets[30] published as CLLD applications – i.e. with their data available as Linked Data described via VoID – I will try to gauge the impact of the CLLD project looking at some use cases:

- At the most basic level, fulfilling the request "give me all you have on language x" (where x is chosen from one of the supported language catalogs) should be possible – using a local triplestore filled by harvesting the CLLD apps individually or later this year using the CLLD portal.[31]

- Testing the conjecture mentioned in WALS chapter "Hand and Arm" (Brown, 2013)

  > The presence of tailored clothing covering the arms greatly increases the distinctiveness of arm parts and renders more likely their labeling by separate terms [...]. Another potentially fruitful investigatory strategy would be to cross-tabulate values against the tailoring technologies of peoples who speak each of the 620 languages of the sample - an enormous research effort this author must leave to future investigators.

  can still not be done fully automatically, but it should be possible to connect languages to descriptions about their speakers via dbpedia and start from there.[32]

- Seeding/expanding datasets like "Tea" (Dahl, 2013)[33] with data from lexical databases like WOLD[34] is already possible.

Arguably, in particular for the case of typological datasets, completely automated use is still far off.[35] The typical process for analysis of typological datasets will remain a sequence of data download, manual inspection, massaging the data, then running analysis software; for this workflow, the uniform access aspect of Linked Data is the most important.
Thus, the future plans for the project focus on

- aggregators or portals:

  Lexicalia the portal to lexical data in CLLD datasets[36] and

  CrossGram the portal to typological data in CLLD datasets[37]

  are already under way. In general we would like to follow the example of bio2rdf in providing a standard, versioned, curated aggregation of linguistic data around which a community can grow and share analyses, methods and data.

- "curated triplestores" and "on-demand triplestores" in the sense of Tennison (2010).[38] On-demand triplestores also look like a good way to go forward with reproducible research[39] without putting the burden of versioning on each database: one will not be able to simply publish a sparql query and the URL of the endpoint, but would have to deposit the triples as well.

[28] It would probably make sense to link these meanings to wordnet wordsenses but it seems difficult to determine authoritative URLs for these.

[29] E.g. in WALS chapters carry information on the corresponding linguistic field which often can be linked to dbpedia; WALS languages can be linked via `dcterms:spatial` relations to geonames.org countries.

[30] `http://clld.org/datasets.html`

[31] Cf. `http://portal.clld.org/?q=english`

[32] If a large, curated database like eHRAF were part of Linked Open Data this could be possible, though. It should also be noted that cultures, thus anthropological data, are often identified/distinguished by their language, so that this kind of data would also fit into the CLLD framework.

[33] This dataset lists for many languages whether the corresponding word for "tea" is derived from Sinitic "cha" or Min Nan Chinese "te".

[34] `http://wold.livingsources.org/meaning/23-9` lists counterparts for "tea" in various languages including their loanword status.

[35] Judging by our experience with making APiCS and WALS structure sets comparable (where APiCS was created with comparability to WALS as an explicit goal), and evidence provided by Round (2013) at ALT 10 for the difficulty of designing comparable datasets, it seems clear that "know your data" will remain an obligation of the researcher that cannot be shifted to the machine.

[36] `http://portal.clld.org/lexicalia`

[37] `http://portal.clld.org/crossgram`

[38] Imagine a service that would allow one to: 1. collect a custom dataset from selected features from WALS, APiCS and eWAVE; 2. post-process it with software from CLLD's community repository; 3. dump it in virtuoso and package the appliance as Amazon EC2 AMI ...

[39] `http://languagelog.ldc.upenn.edu/nll/?p=830`

4. References

Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. 2011. Describing linked datasets with the void vocabulary. http://www.w3.org/TR/void/.

Peter Austin. 2011. They're out to get you (or your data at least). http://www.paradisec.org.au/blog/2011/04/theyre-out-to-get-you-or-your-data-at-least/.

Cecil H. Brown. 2013. Hand and arm. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Östen Dahl. 2013. Tea. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Alexis Dimitriadis. 2006. An extensible database design for cross-linguistic research. http://languagelink.let.uu.nl/burs/docs/burs-design.pdf.

Jeff Good and Michael Cysouw. 2014. Languoid, doculect and glossonym: Formalizing the notion 'language'. *Language Documentation & Conservation*, 07.

Martin Haspelmath and Uri Tadmor, 2009. *The Loanword Typology Project and the World Loanword Database*, page 1–33. De Gruyter.

Martin Haspelmath. 2013. Can language identity be standardized? on morey et al.'s critique of iso 639-3. http://dlc.hypotheses.org/610.

Martin Haspelmath. 2014. A proposal for radical publication reform in linguistics: Language science press, and the next steps. http://dlc.hypotheses.org/631.

Bernadette Hyland, Ghislain Atemezing, and Boris Villazón-Terrazas. 2014. Best Practices for Publishing Linked Data. http://www.w3.org/TR/ld-bp.

Paola Monachesi, Alexis Dimitriadis, Rob Goedemans, Anne-Marie Mineur, and Manuela Pinto. 2002. A unified system for accessing typological databases. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 3)*.

Sebastian Nordhoff. 2012. Linked data for linguistic diversity research: Glottolog/langdoc and asjp online. In Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors, *Linked Data in Linguistics. Representing and Connecting Language Data and Language Metadata*. Springer, Heidelberg.

Shakthi Poornima and Jeff Good. 2010. Modeling and encoding traditional wordlists for machine applications. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, NLPLING '10, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics.

Erich Round. 2013. How to design a dataset which doesn't undermine automated analysis. Talk given at ALT 10.

Leo Sauermann and Richard Cyganiak. 2008. Cool URIs for the Semantic Web. http://www.w3.org/TR/cooluris/.

Jeni Tennison. 2010. Distributed publication and querying. http://www.jenitennison.com/blog/node/143.

Jeni Tennison. 2011. What do uris mean anyway? http://www.jenitennison.com/blog/node/159.

Menzo Windhouwer and Alexis Dimitriadis. 2008. Sustainable operability: Keeping complex resources alive. In *Proceedings of the LREC Workshop on the Sustainability of Language Resources and Tools for Natural Language Processing*.