The Cross-Linguistic Linked Data project

Robert Forkel

Max Planck Institute for Evolutionary Anthropology, Leipzig

May 27, 2014

Outline

Diversity Linguistics at the MPI EVA

- The department for Linguistics at the Max Planck Institute for Evolutionary Anthropology (MPI EVA) in Leipzig studies the world's language diversity.
- As basis or result of this research, cross-linguistic datasets are collected.
- "Cross-linguistic" in this context means "massively multilingual", i.e. datasets often span hundreds of languages.
- As a corollary, studying under-resourced languages is rather the norm than the exception.

Cross-Linguistic data

These cross-linguistic datasets typically consist of

- lexical or typological data
- on many languages (> 20)
- or on small languages.

Examples

- wordlists (Swadesh, Leipzig-Jakarta, etc.) or dictionaries,
- phoneme inventories,
- typological surveys,
- small collections of glossed text (IGT following the Leipzig Glossing Rules) or bibliographies

The status quo of cross-linguistic data on the Web

A lot of cross-linguistic data has been compiled/collected; many linguists have written a dictionary or a grammar or compiled a typological survey as database for their own research.

- But often it is not (anymore) freely accessible on the web . . .
- . . . but is hidden in books . . .
- . . . or – worse – in drawers.

Why?

- The traditional publication models do not work for this kind of data (databases, dictionaries on small languages, . . .).

Cross-linguistic data on the web

- Provided we can get cross-linguistic data on the web, how do we make sure it will stay?
- What tends to keep web resources accessible is

  usage!
- So bridging the gap between data creation and usage, i.e. publishing data in a usable way will solve our problem.

Use cases for cross-linguistic data

How is cross-linguistic data used?

- Search for universals or the lack of these, i.e. documenting language diversity.
- Areal linguistics – research on areal features of languages, e.g. WALS chapter on *Hand and Arm*
- Historical linguistics – reconstruction of proto-languages, mass comparison of lexical data; e.g. *ASJP*, *Mapping the origin of Indo-European*
- To compute language distances/complexity/... (e.g. Gil, Dahl)

WALS chapter – hand and arm

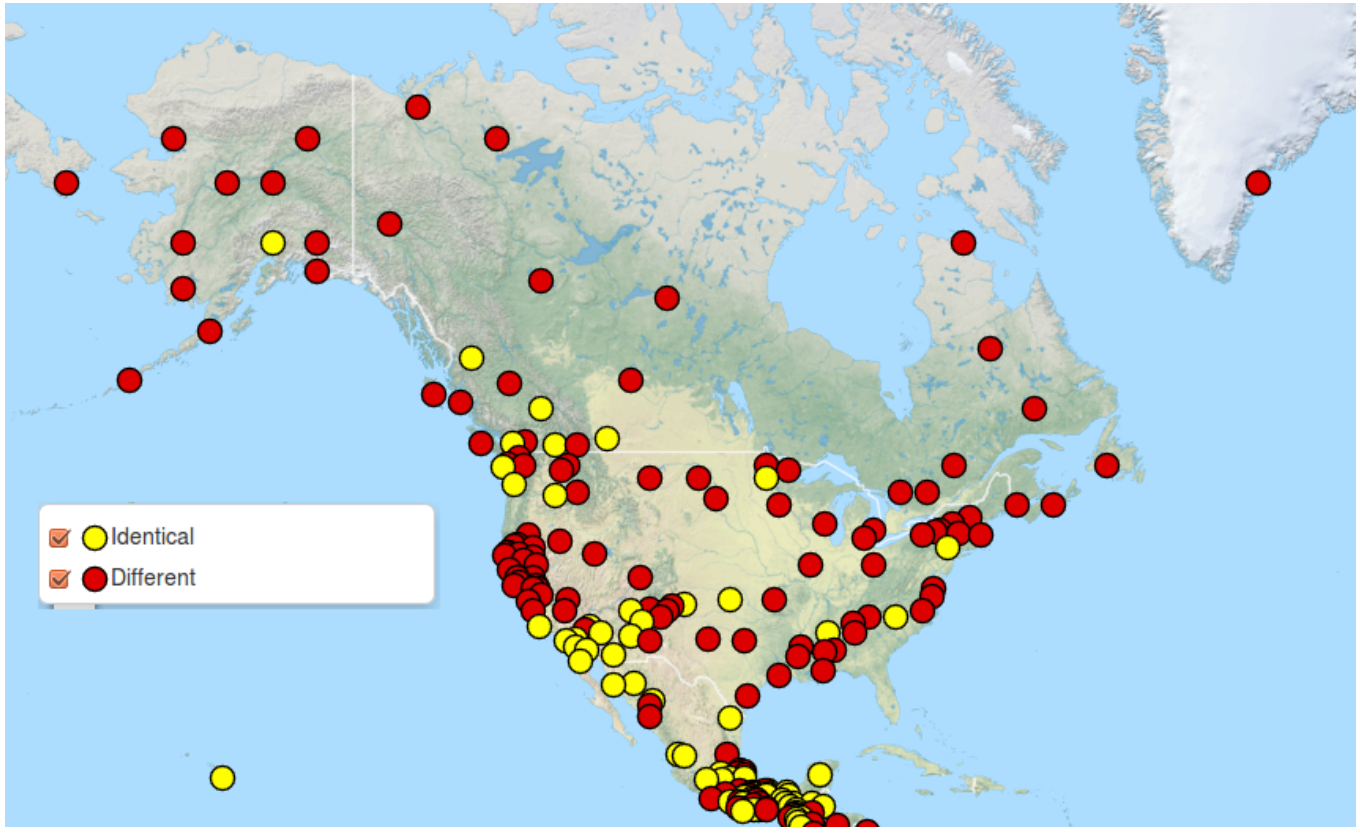

Figure 1: Cecil H. Brown. 2013. Hand and Arm. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online.

ASJP language tree

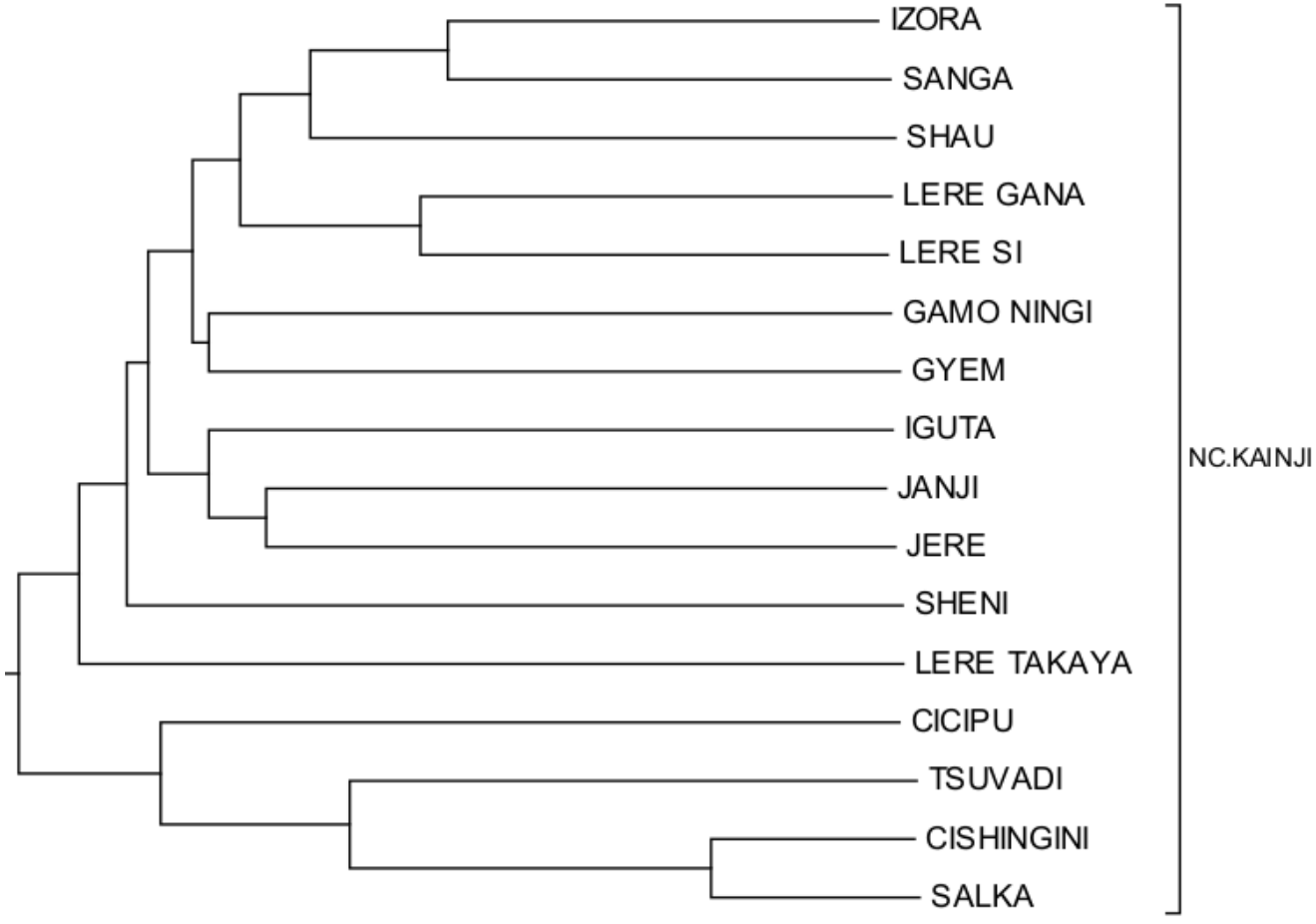

Figure 2: The ASJP Consortium. 2013. ASJP World Language Trees of Lexical Similarity: Version 4 (October 2013).

Mapping the origin of Indo-European

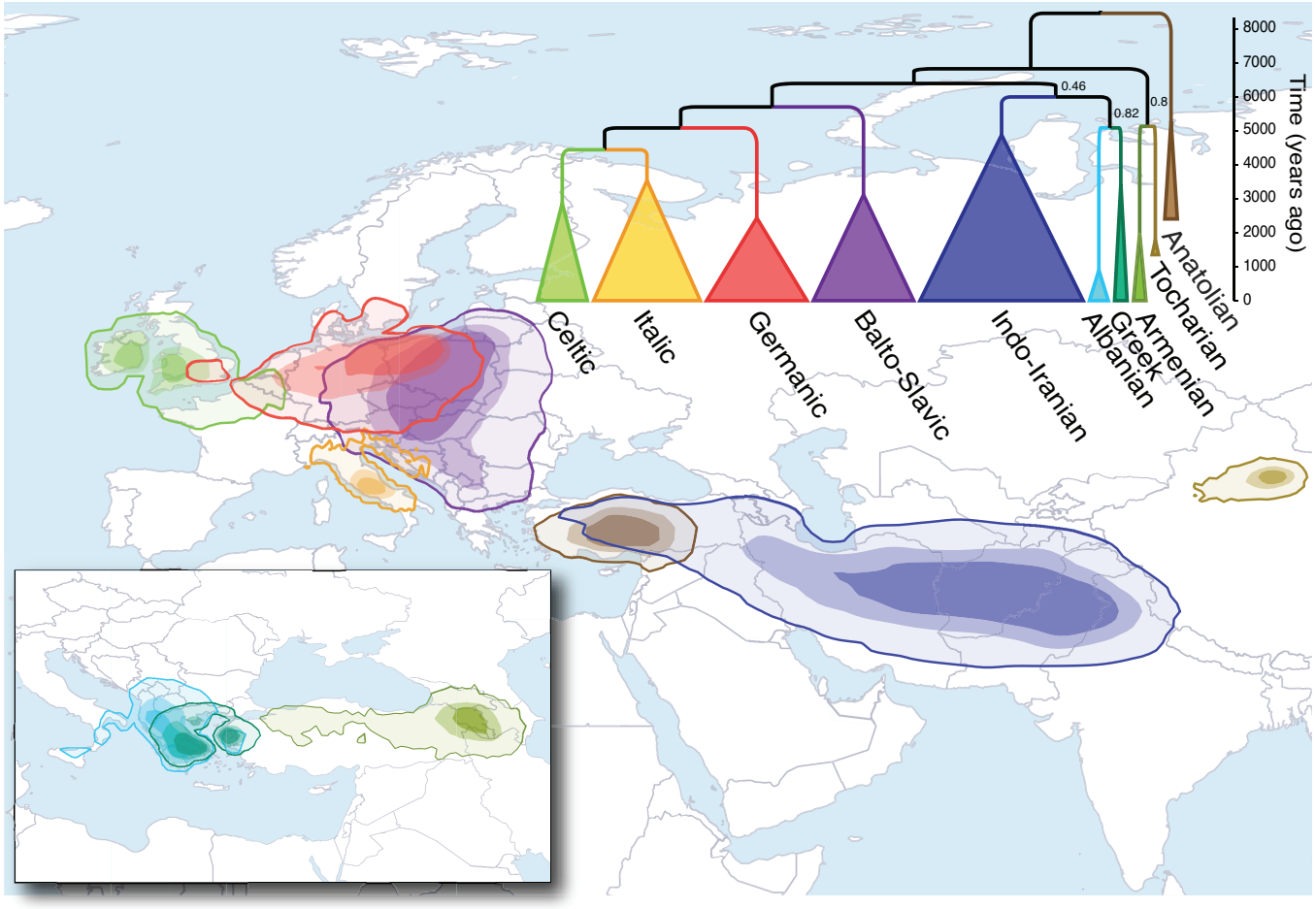

Figure 3: Figure 2 from Bouckaert, R. et al. (2012). Mapping the origins and expansion of the Indo-European language family. Science, 337:957–960.

The CLLD project

The CLLD project sets out to pick the (seemingly) low-hanging fruit – to bring existing but unpublished cross-linguistic data to the web by establishing sustainable publication infrastructure.

CLLD – datasets

WALS The World Atlas of Language Structures a database of structural properties of more than 2600 languages,

APiCS The Atlas of Pidgin and Creole Language Structures,

WOLD The World Loanword Database contains vocabularies of 41 languages from around the world annotated for loanword status,

IDS The Intercontinental Dictionary Series (to be published in CLLD in 2014),

ASJP The Automated Similarity Judgement Project (to be published in 2014),

Glottolog A language catalog and comprehensive bibliography.

But CLLD can publish non-MPI EVA datasets as well and has done so: eWAVE, SAILS, PHOIBLE.

CLLD – publication models

CLLD provides three publication models for cross-linguistic datasets:

- Standalone databases following an "edited series" model, like WALS, WOLD, ...
- Two journals for cross-linguistic datasets,
  - *Dictionaria* a journal for dictionaries,
  - *The Journal for Cross-linguistic Datasets* for typological surveys and simila datasets.
- Self-hosting using the `clld` software.

These models are explicitely based on examples from the traditional publishing world.

CLLD – publication models

- Incentivize data publication through recognition – following traditional models.
- Favor expert submission and editorial control over crowdsourcing
  - for easier quality control,
  - easier provenance assessment,
  - and because we have the experts at hand.

CLLD – the software

The datasets are hosted as web applications built on the `clld` python package,

- a CMS tailored towards cross-linguistic data,
- provides a common data model which
  - includes the generally accepted practice of basing all "measurements" on sources,
  - fits typological and lexical data,
  - is customizable per application.
- Each `clld` app has full control over its output.
- So we have a lot of datasets served by software fully under our control – time to think about standards!

CLLD – the software

- Trying to "standardize" on the software – i.e. making things interoperable by putting all on the same technology stack – has been tried before.
- Taking the second step – i.e. using common specifications, protocols is essential.
- Still, being able to put a nice interface browsable by humans on your database goes a long way in attracting data submissions.

clld data model

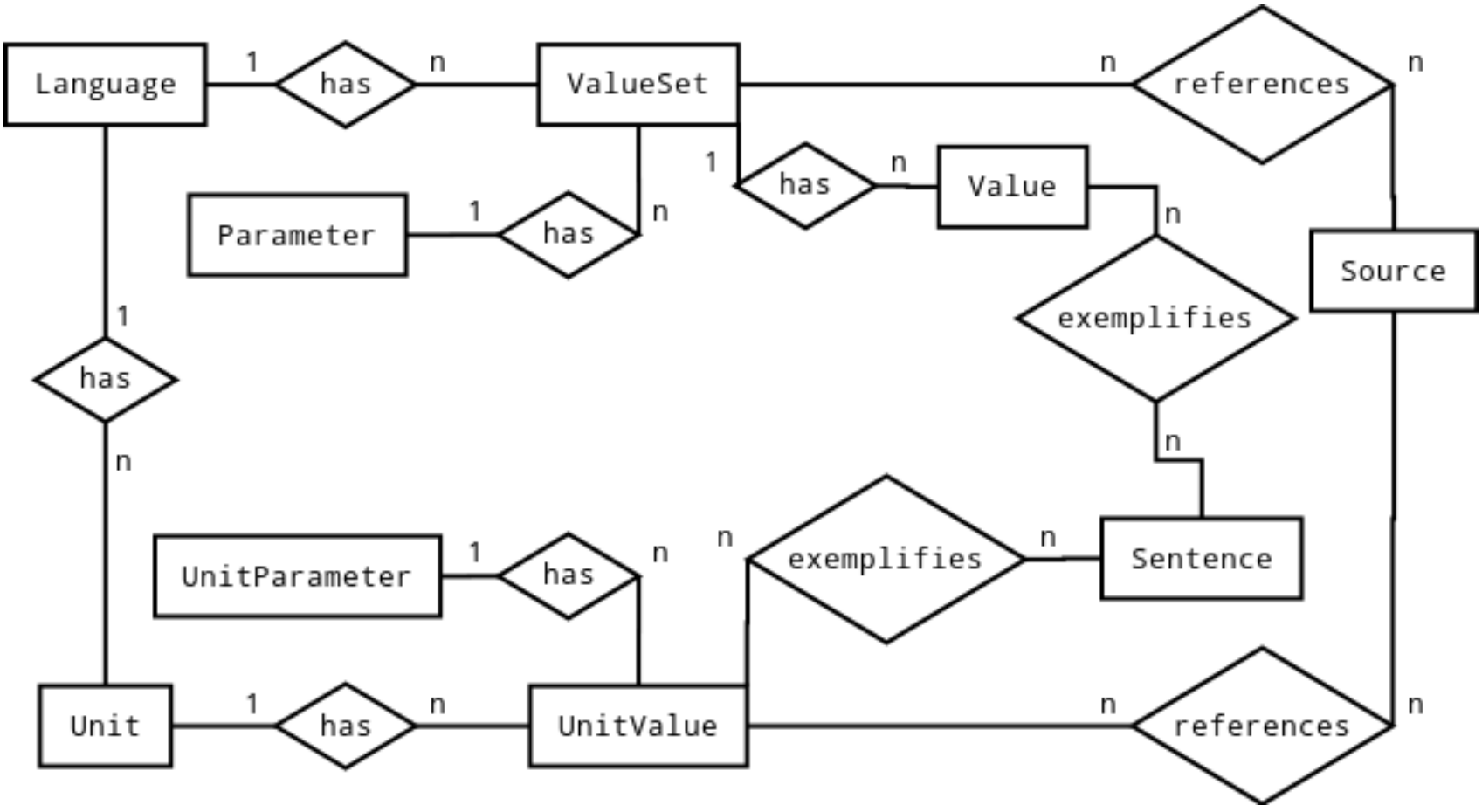

Figure 4: The default clld data model.

clld default data model – with a twist

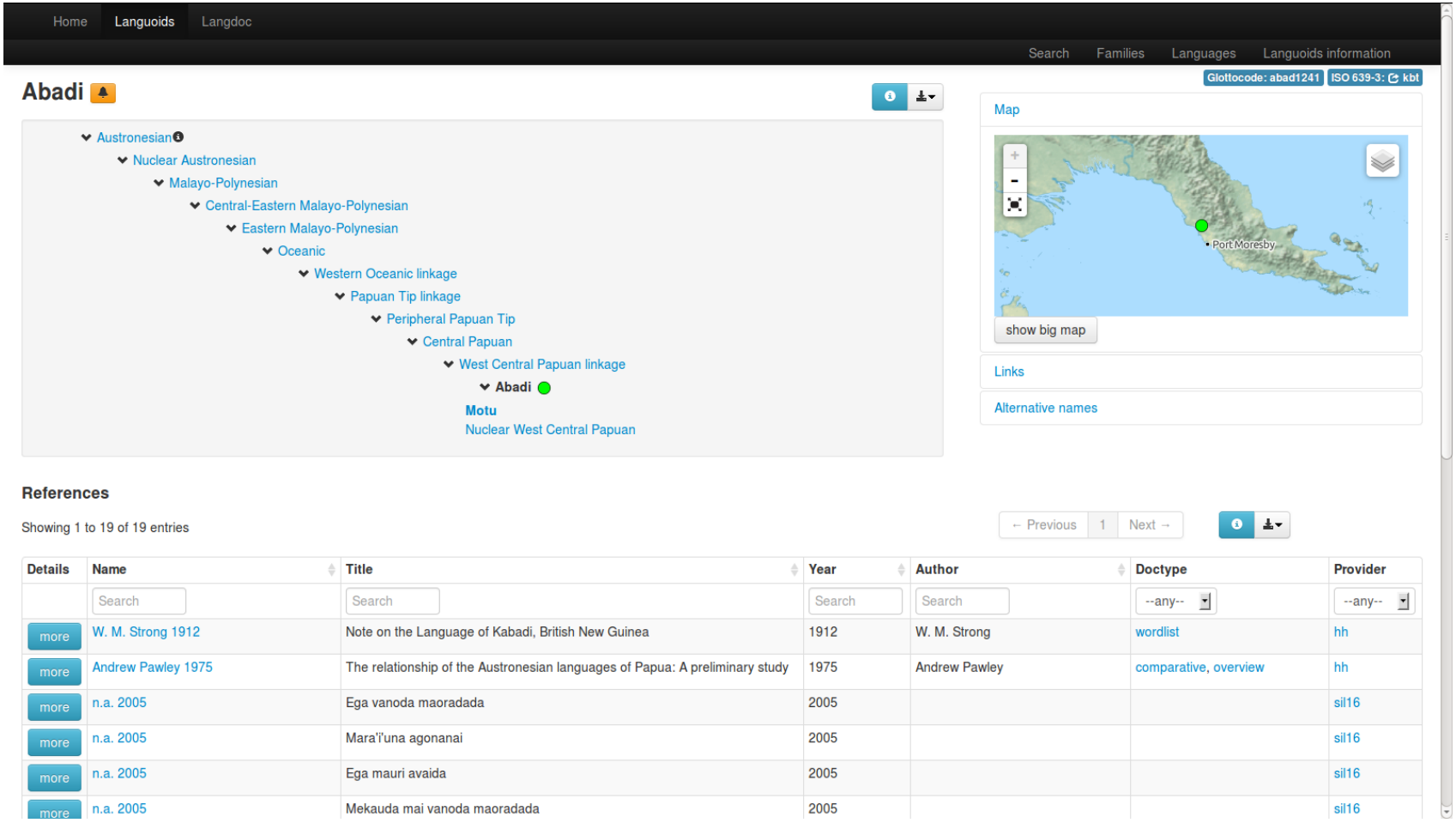

Figure 5: Glottolog page for Icelandic, listing genealogy and sources.

clld data model for structural data

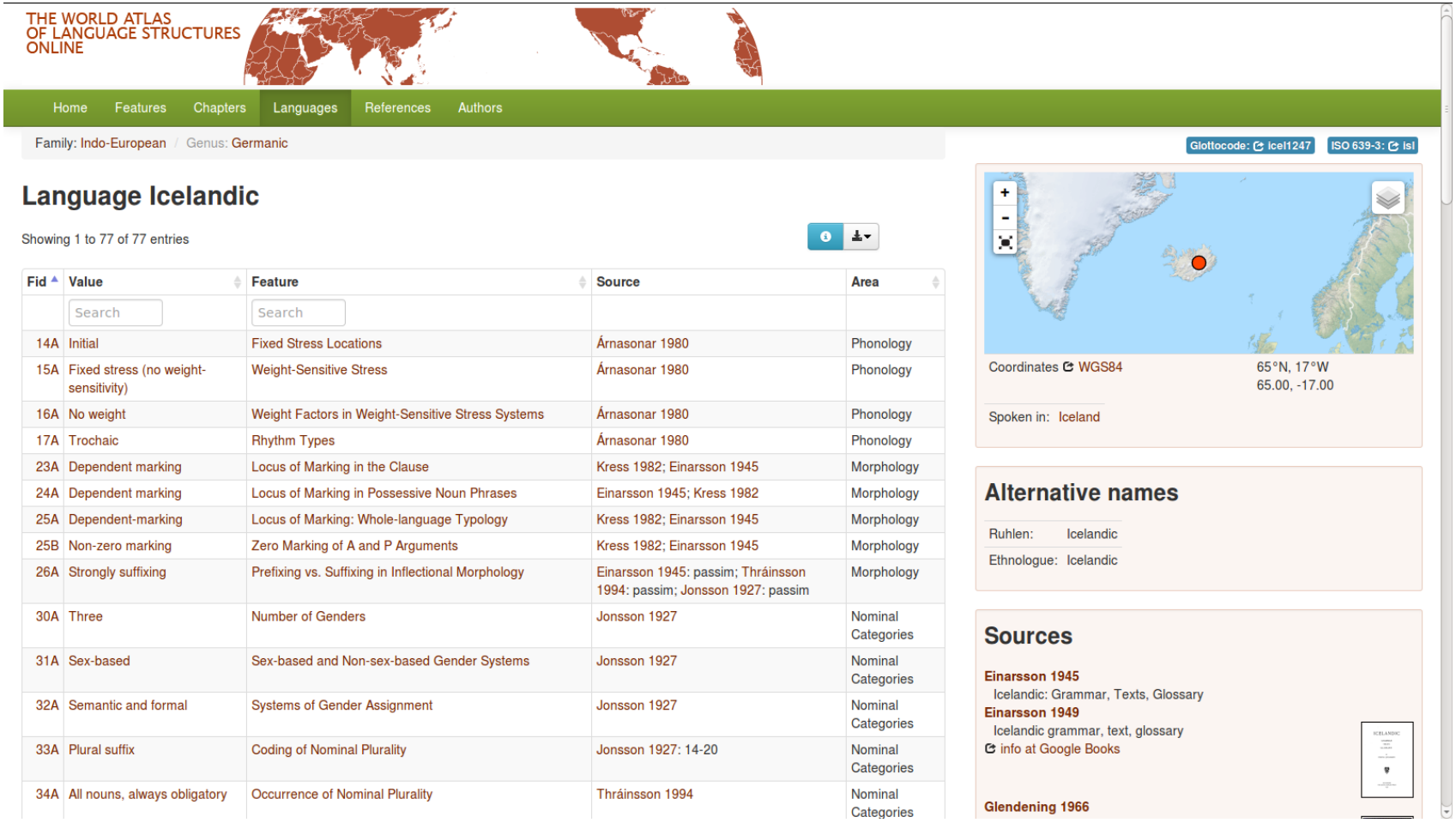

Figure 6: WALS page for Icelandic, listing properties and sources.

clld data model for lexical data

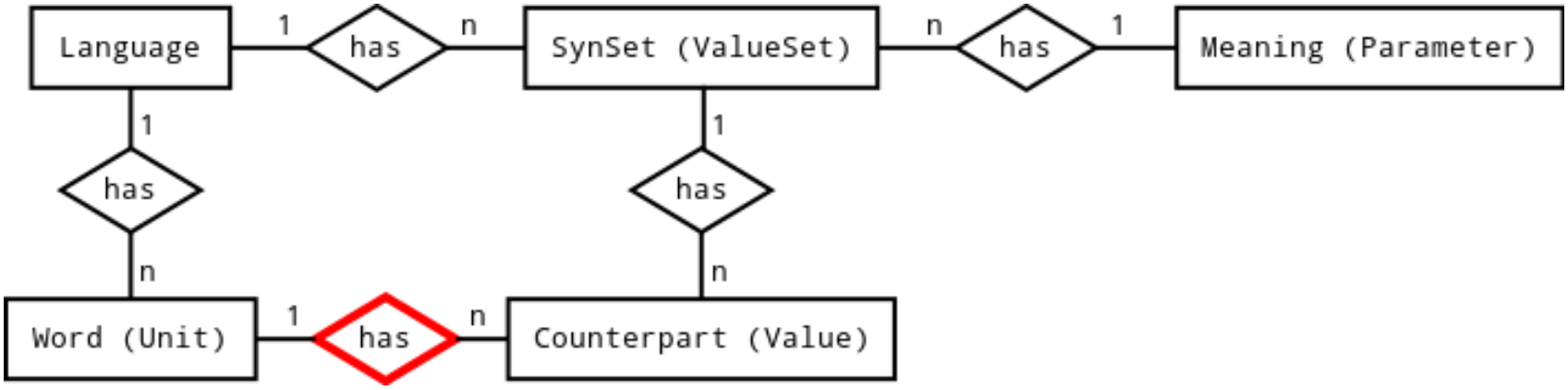

Figure 7: The WOLD instantiation of the data model.

- Additional relation in custom data model,
- lexical data model can be mapped to *lemon* (*Counterpart* maps to *LexicalSense*).

clld data model for lexical data

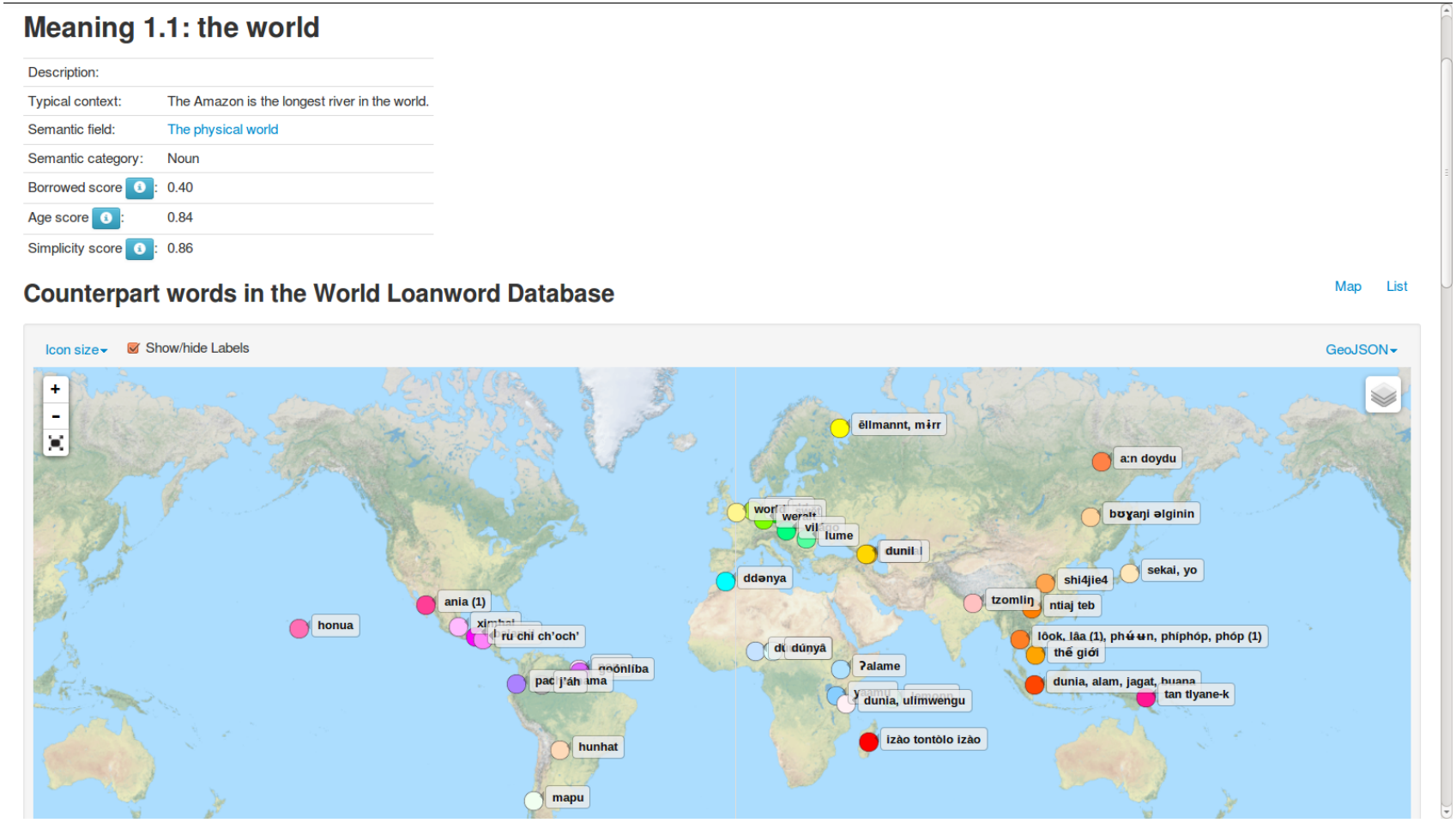

Figure 8: WOLD page for comparison meaning "World".

clld data model for lexical data

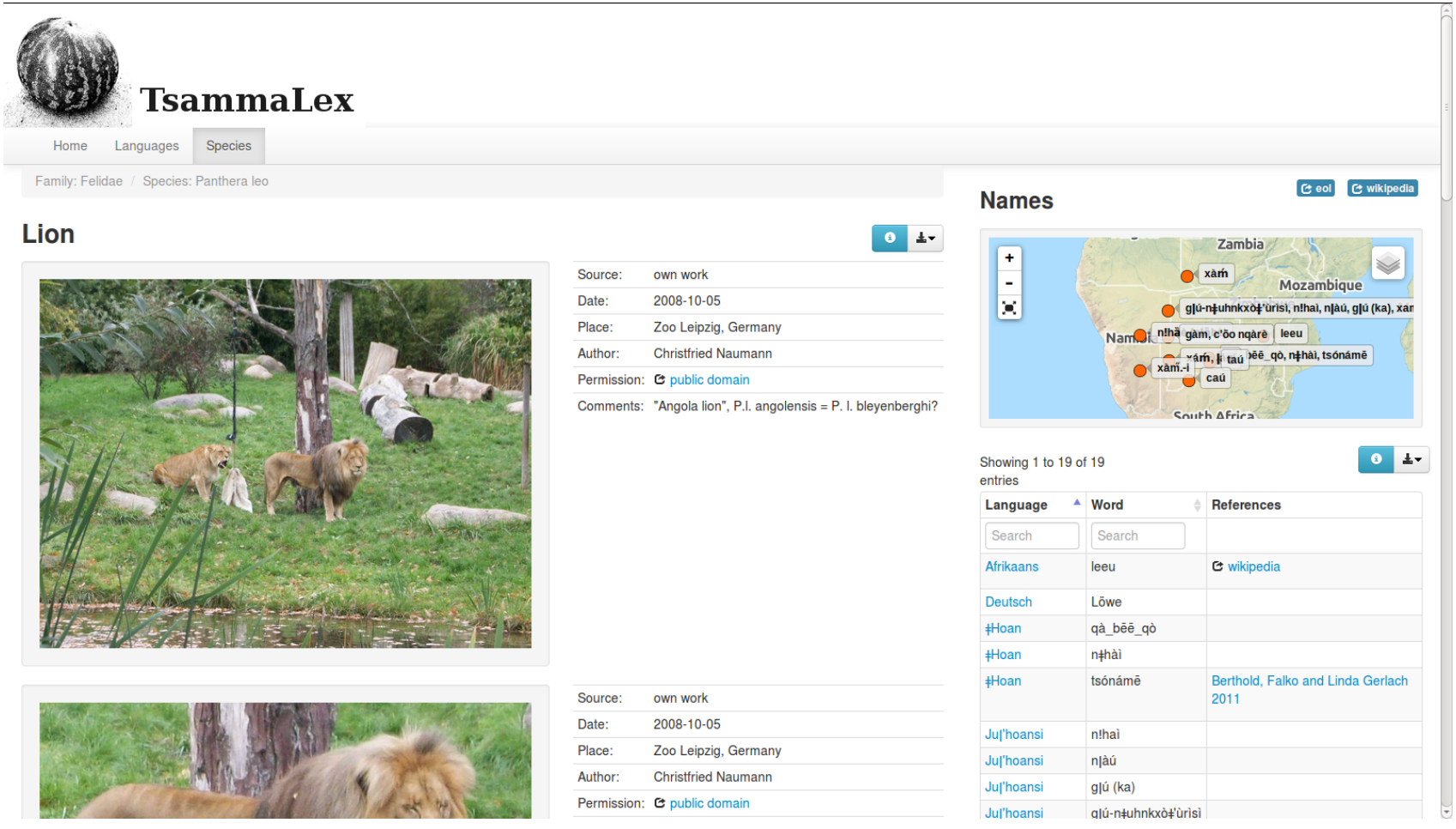

Figure 9: Tsammalex page for species "Lion".

CLLD and Linked Data

So, with this publication platform and lots of datasets at our hands, what to do?

- We regard Linked Data principles as rules of best practice for publishing data on the web.

Linked Data – the first star

Publish data on the web with an open license!

- Most CLLD datasets are published under CC-BY licenses.
- Implies using HTTP URIs as names,
- which forces you to think about the things you want to describe and at which level of granularity,
- enables distributed development of data and the basis for merging via globally unique identifiers,
- puts coarse provenance information in each identifier.

Linked Data – 3-out-of-5 stars

Generally, the usefulness of "3-out-of-5 stars" Linked Data has to be stressed:

- Linked Data as uniform data access API (following the "crawler" paradigm)
- enables distributed databases,
- allows follow-your-nose API discovery,
- plays well with the web at large (Internet archive, bookmarking, etc.),
- allows easy hosting (thus helps with sustainability, and is attractive for developers/administrators as well) – which cannot be said about SPARQL endpoints.

Linked Data – the 4th star

That being said, for common domains RDF models are useful, e.g. to describe provenance and links between resources.

- All CLLD datasets have editors (are)edited.
- VoID is used to convey basic provenance and license information.
- Typically all statements of linguistic interest (i.e. value assignments) are linked to sources.

Linked Data – the 4th star

- Our publication platform does spit out RDF.
- The RDF model for a particular `clld` app can be completely customized.
- But should it?
- Balance between
  - uniform access across CLLD apps and
  - semantic interoperability with existing infrastructure.
  - Is it more useful to model resources as having multiple types or provide mappings?
- Example: Model lexical data using *lemon*?
- Generally, in terms of user-friendliness, the problem is not a choice of RDF models but csv, Newick, ...

Linked Data – the 5th star

Linking with other resources . . .

- Glottolog as hub in the CLLD LOD cloud:
  - language catalog (linking in turn to lexvo, dbpedia, etc.), iso639-3 is often not sufficient.
  - shared bibliography
- WOLD as catalog for comparison meanings (cf Leipzig-Jakarta list) – a *concepticon*, or an *ontology*.
- PHOIBLE may play such a role for phonological segments, e.g. as reference for transcriptions.

Linked Data – the 5th star

. . . and vocabularies

- we stick with rather generic or focused vocabularies by default: dcterms, skos, foaf;
- and aim at semantic interoperability by default only for stable interpretations across apps: bibliographical data, provenance data, metadata: VoID, bibo, . . .
- Creation of a CLLD ontology might be warranted – but this may be putting the cart before the horse.

Glottolog – linking languages to descriptions

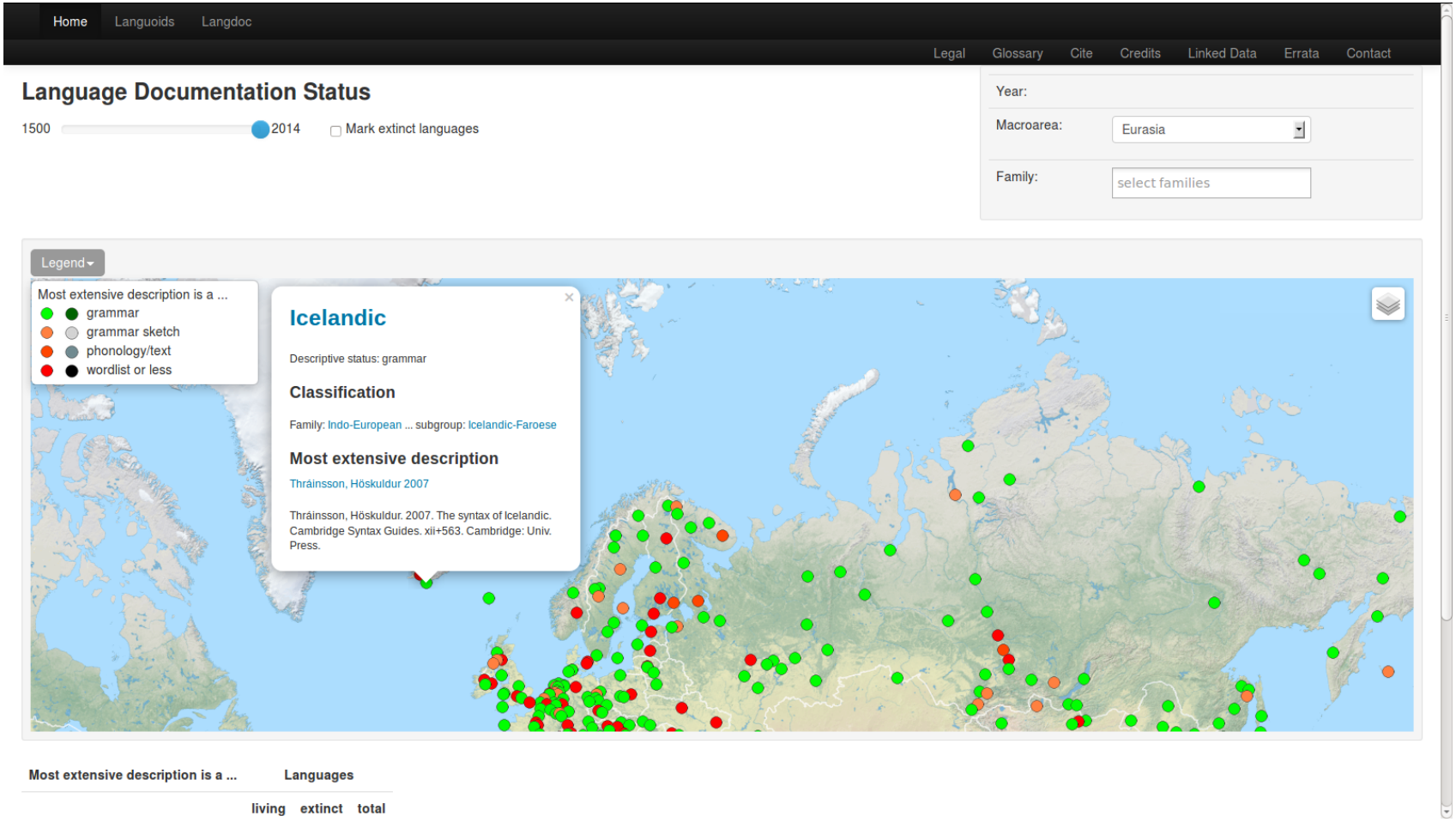

Figure 10: Glottolog description status browser – the LRE map for language descriptions.

Use cases – revisited

How does CLLD help with our use cases and how does Linked Data benefit cross-linguistic research?

- Clearly all use cases will benefit from more and better data, e.g.
  - constructing language phylogenies with a time-depth of more than a couple thousands of years is thought to be more accurate with structural than with lexical data,
  - the WALS language–feature matrix is still pretty sparse.
- Many of the listed datasets have been available in some digital form before, being able to access them in a unified way could help grow a unified toolset.

A workflow for research based on CLLD data

1. Identify suitable datasets.
2. Aggregate the data in a triple store (crawling/importing dumps).
3. Filter data in the triple store (using provenance information, etc.).
4. Export data to suitable format for analysis.

- CLLD and Linked Data will mainly play a role during aggregation of raw data.

Semantic interoperability I

- Being able to evaluate provenance data during the aggregation of a dataset is useful (e.g. in the ASJP project, some sources of wordlists are regarded as less trustworthy than others).
- Unambiguous identification of languages is required; Glottolog will help with that.
  - Being able to answer the question "which data do we have on a selected sample of languages?" as well as
  - "what sample of languages can we investigate given we need a certain selection of data (lexical, structural, etc.)?"
- For lexical data *lemon* can help to interpret the raw data, i.e. matching senses across languages (cf. Moran and Brümmer 2013).
- The requirements of statistical methods may lead to a standardisation of structural language parameters (features in the WALS sense), but we are not there yet.

Language identification

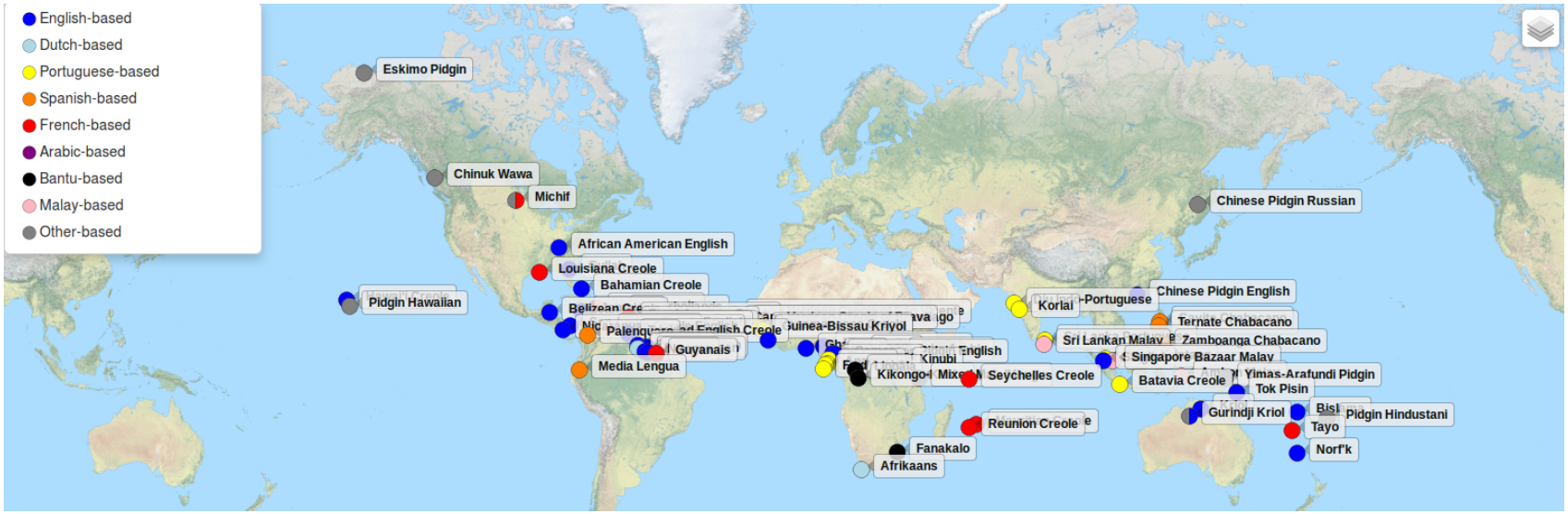

Figure 11: APiCS languages.

Semantic interoperability II

- Often cultures are identified with language codes, e.g. iso639-3; being able to link to anthropological data about these would be very valuable. Quoting from the WALS chapter on *Hand and Arm*:

  > *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.*

Semantic interoperability III – limits

- Generally, useful data formats will be dictated by the needs of the analysis tools (e.g. phylogenetic software),
- so doing analyses directly on the RDF model can not be expected.
- Computing language phylogenies: Construction of the dataset on which to base analyses is part of the intellectual research effort.
- Example APiCS: Interoperability of typological resources is hampered by the difficulty of cross-linguistic categories.

Semantic interoperability – APiCS and WALS

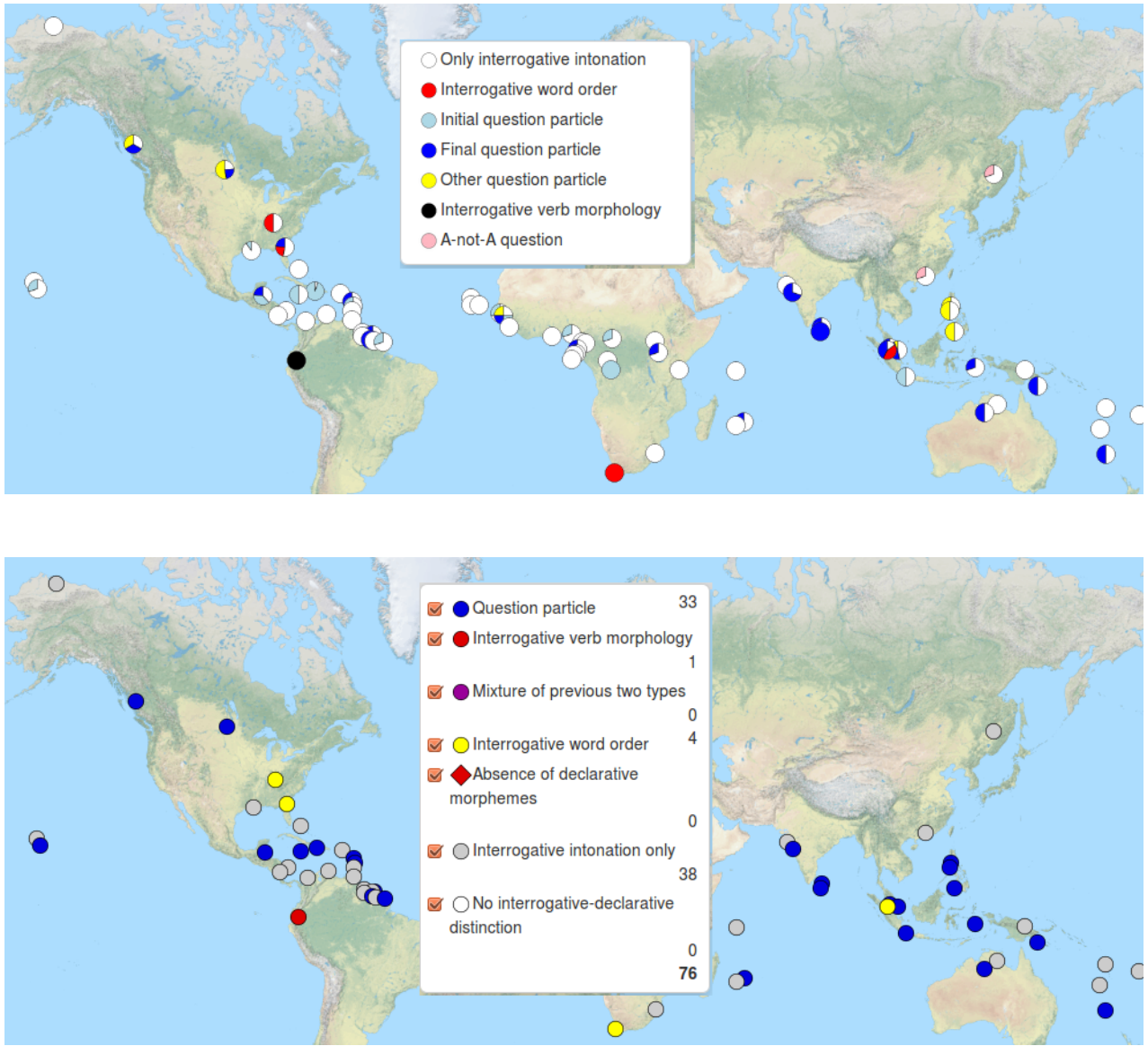

Figure 13: APiCS feature *Polar questions* – original and WALSified.

Final remarks:

- If you are a linguist and have unpublished cross-linguistic datasets, get in touch!
- If you have a research question that might be possible to answer using the kind of data we have, get in touch!
- If you are a Linked Data specialist with ideas how to model cross-linguistic data in RDF, let us know!

http://clld.org

Thank you!