# The Cross-Linguistic Linked Data project

Robert Forkel

Max Planck Institute for Evolutionary Anthropology, Leipzig

May 27, 2014

# Outline

# Cross-Linguistic data

Data for cross-linguistic studies is typically

- 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, grammars, 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, . . . ).

# Use cases for cross-linguistic data

What keeps data on the web from vanishing?
Usage!

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

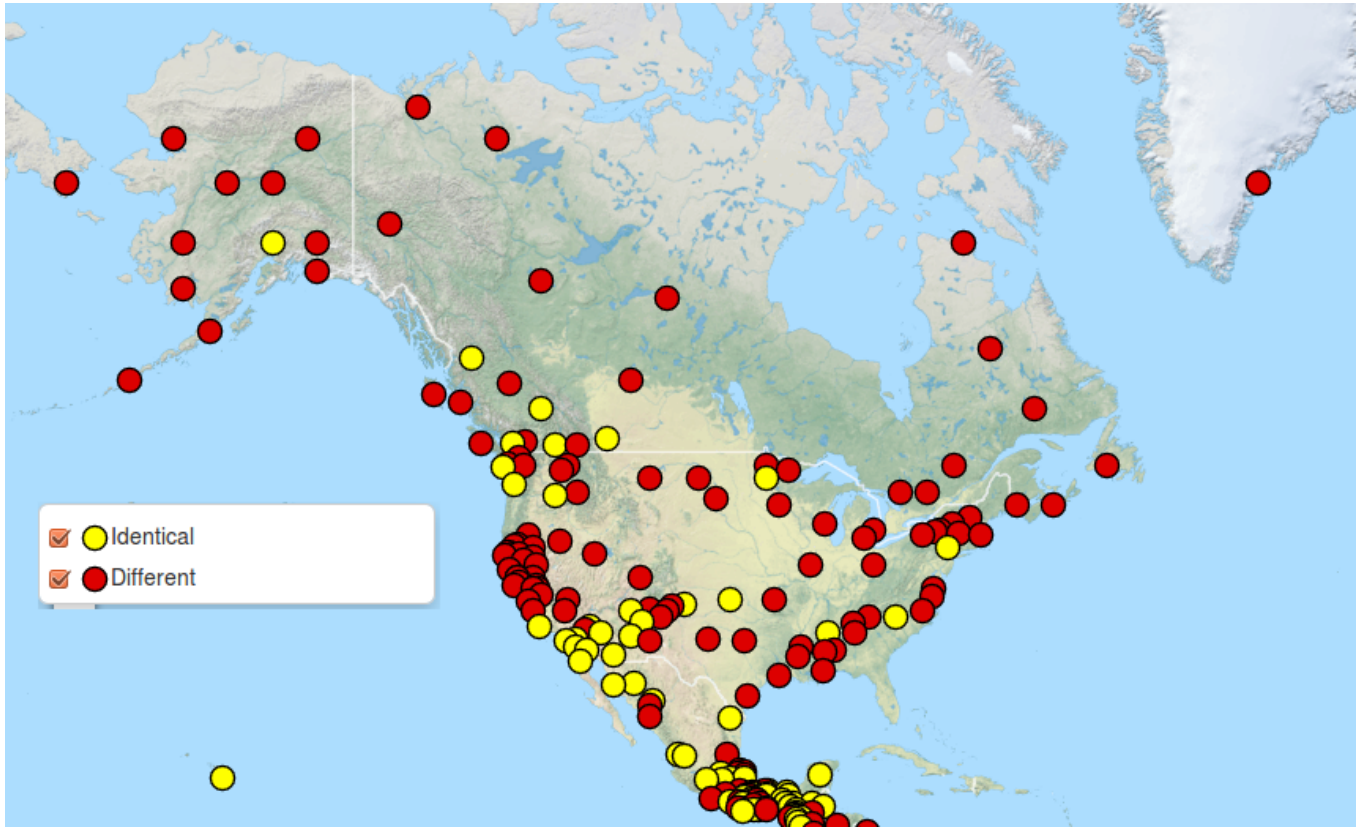


Figure 1: Cecil H. Brown. 2013. Hand and Arm. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) The World Atlas of Language Structures Online.

# ASJP language tree

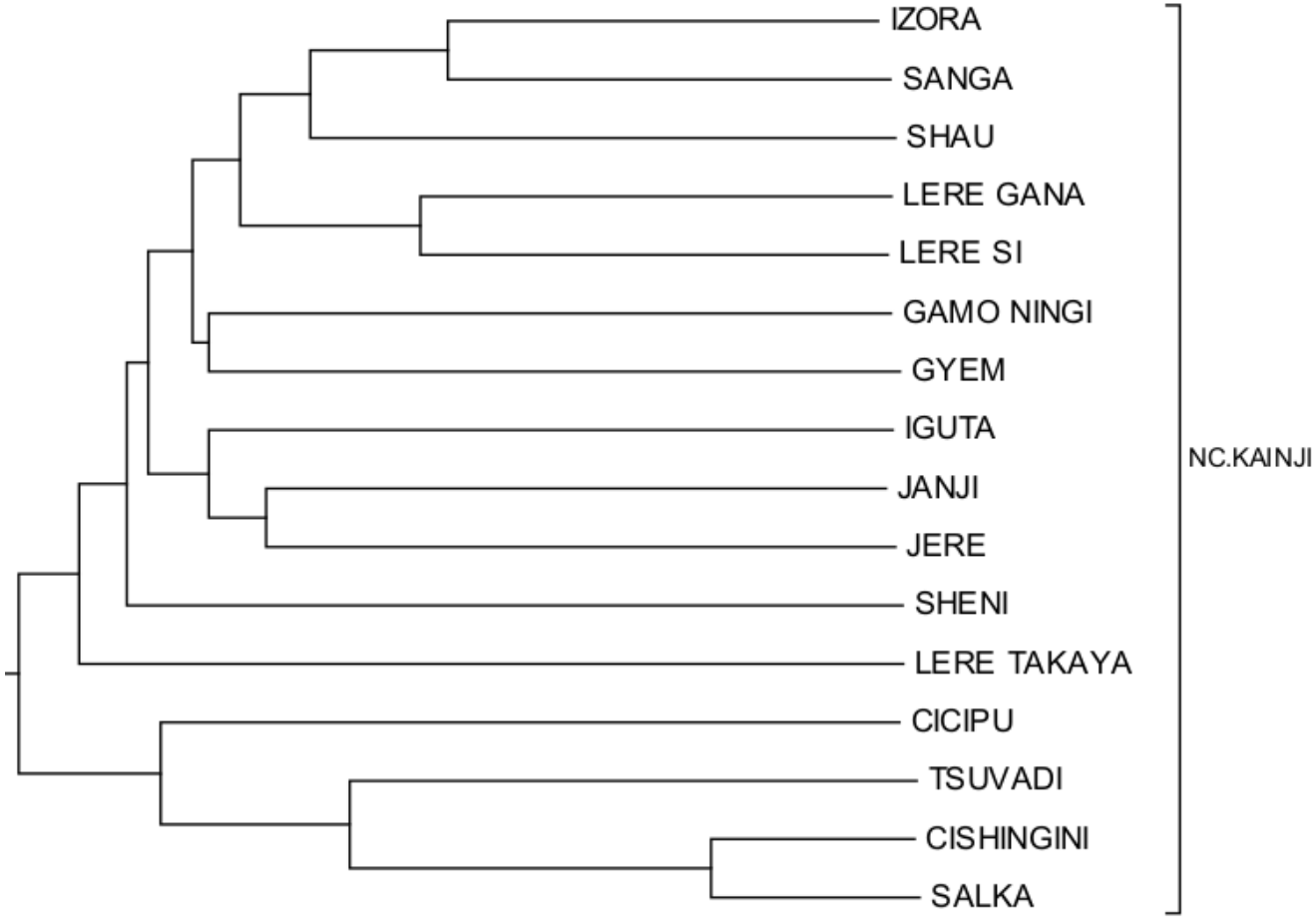


Figure 2: The ASJP Consortium. 2013. ASJP World Language Trees of Lexical Similarity: Version 4 (October 2013).

# Mapping the origin of Indo-European

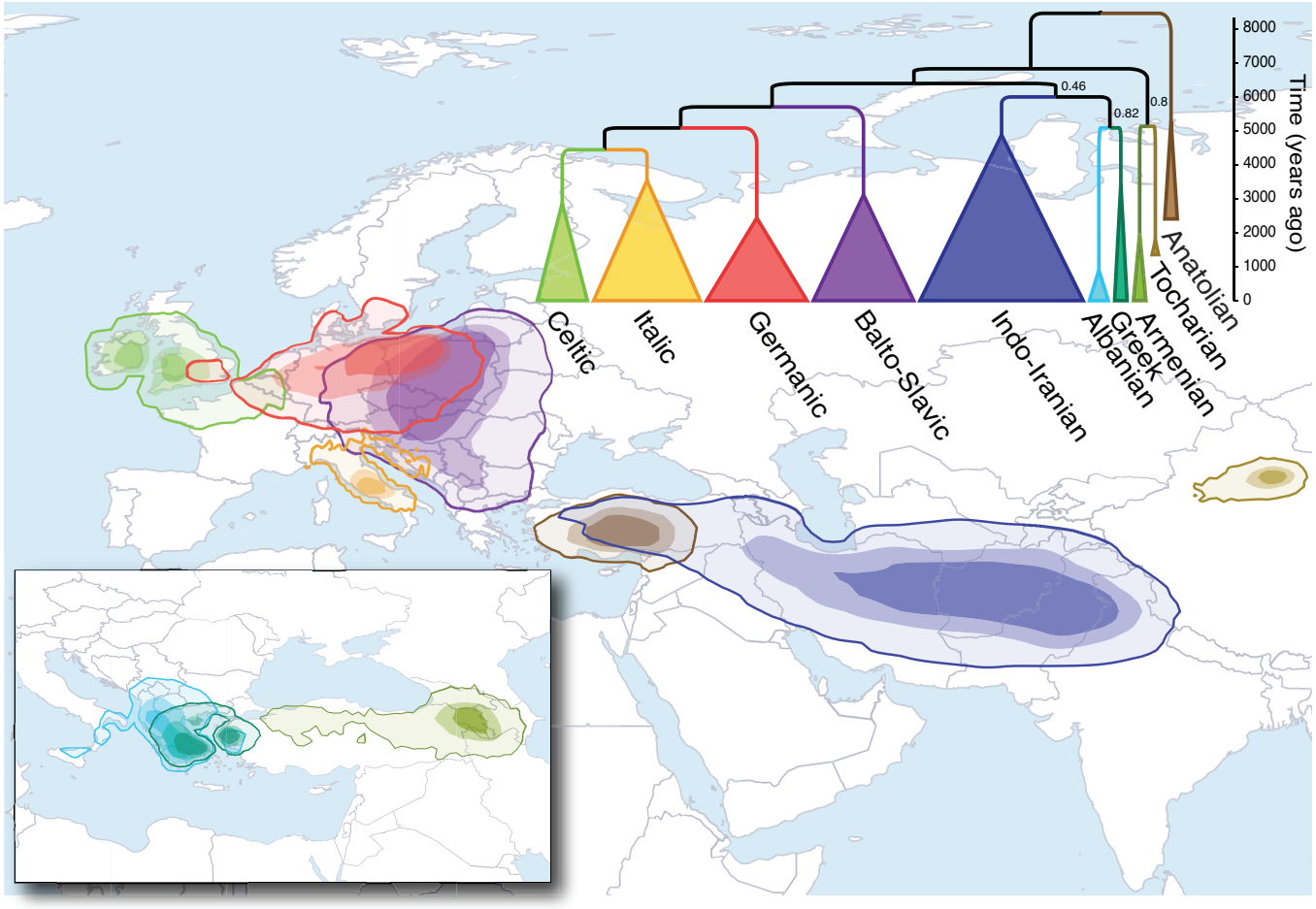


Figure 3: Figure 2 from Bouckaert, R. et al. (2012). Mapping the origins and expansion of the Indo-European language family. Science, 337:957–960.

# The CLLD project

The CLLD project sets out to pick the low-hanging fruit – to bring existing but unpublished cross-linguistic data to the web by establishing sustainable publication infrastructure.

# CLLD – datasets

CLLD was motivated by datasets collected by the department of linguistics at MPI EVA.

| | |
|---|---|
| WALS | The World Atlas of Language Structures, |
| APiCS | The Atlas of Pidgin and Creole Language Structures, |
| WOLD | The World Loanword Database, |
| 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.

# CLLD – the software

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

- a CMS tailored towards cross-linguistic data following the idea of Dimitriades,
- 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 data model I

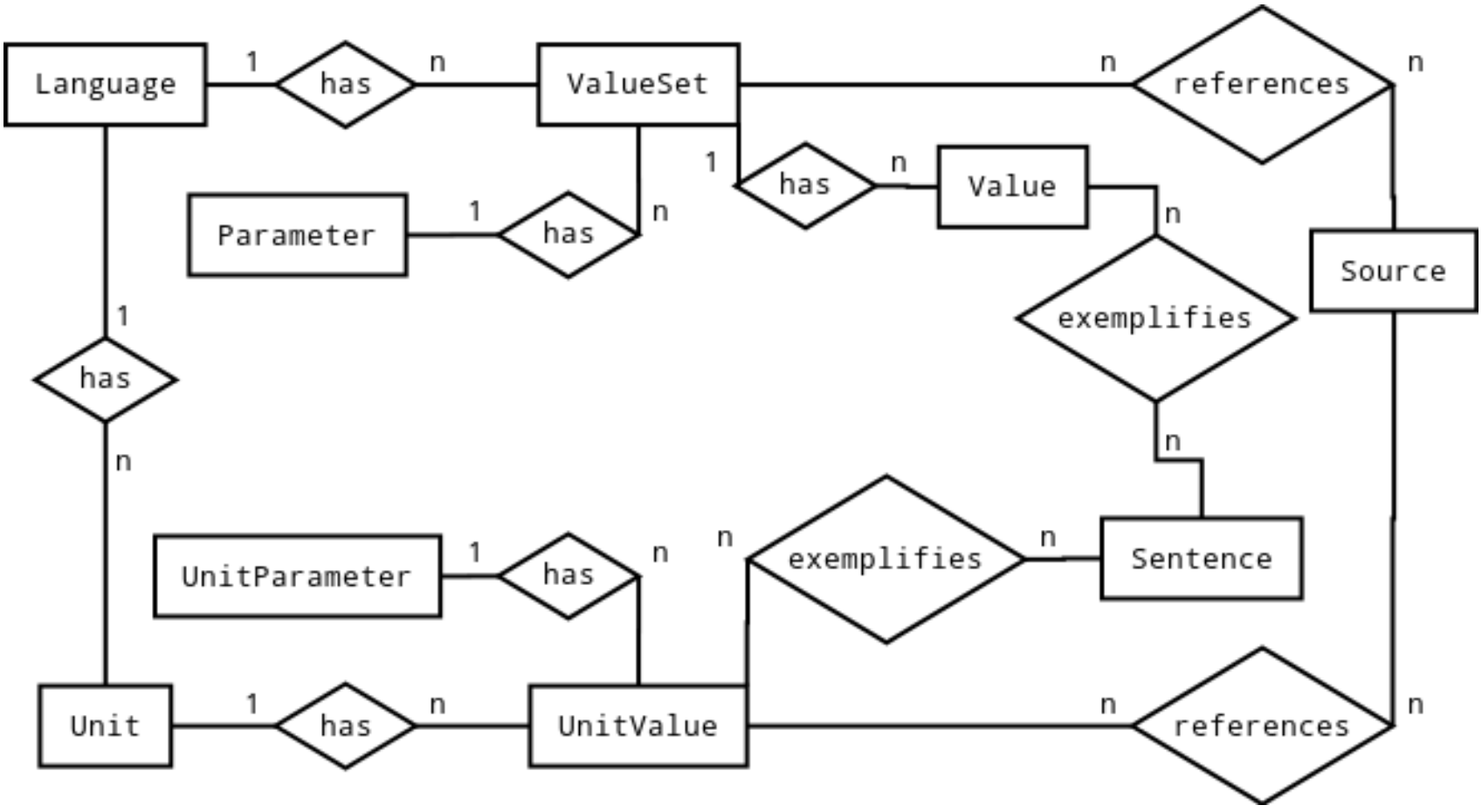


Figure 4: The default clld data model.

# clld data model II

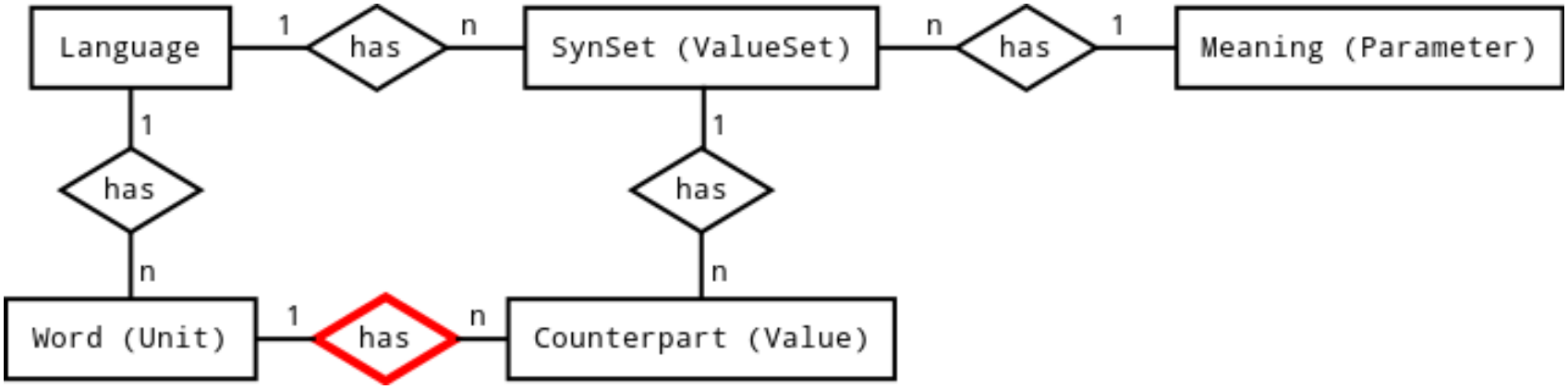


Figure 5: The WOLD instantiation of the data model.

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

# Linked Data – 3-out-of-5 stars

Generally, I want to stress the usefulness of "3-out-of-5 stars" Linked Data:

- 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 (Google, etc.),
- allows easy hosting (thus helps with sustainability, and is attractive for developers/administrators as well).

# Linked Data – the 4th star

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

- 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

- 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*.

# 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*.
- PHOIBLE may play such a role for phonological segments.

and vocabularies:

- stick with rather generic vocabularies by default: dcterms, skos, foaf.
- semantic interoperability by default only for stable interpretations across apps: bibliographical data, provenance data: void, bibo, ...

# 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.

Notes:

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

# 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.

# 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

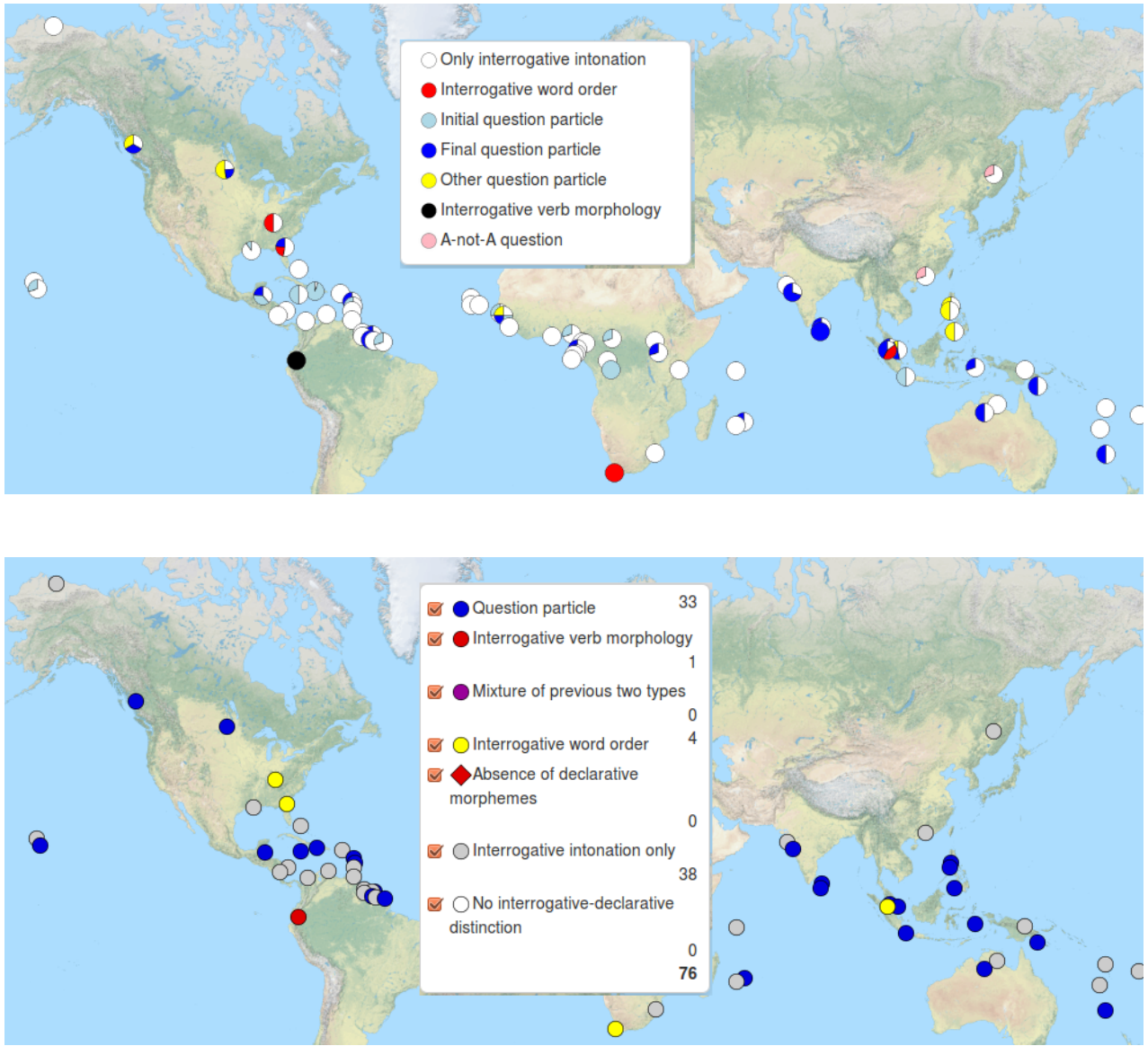


Figure 6: APiCS feature *Polar questions* – original and WALSified.

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!