CLDF

When Data meets Analysis

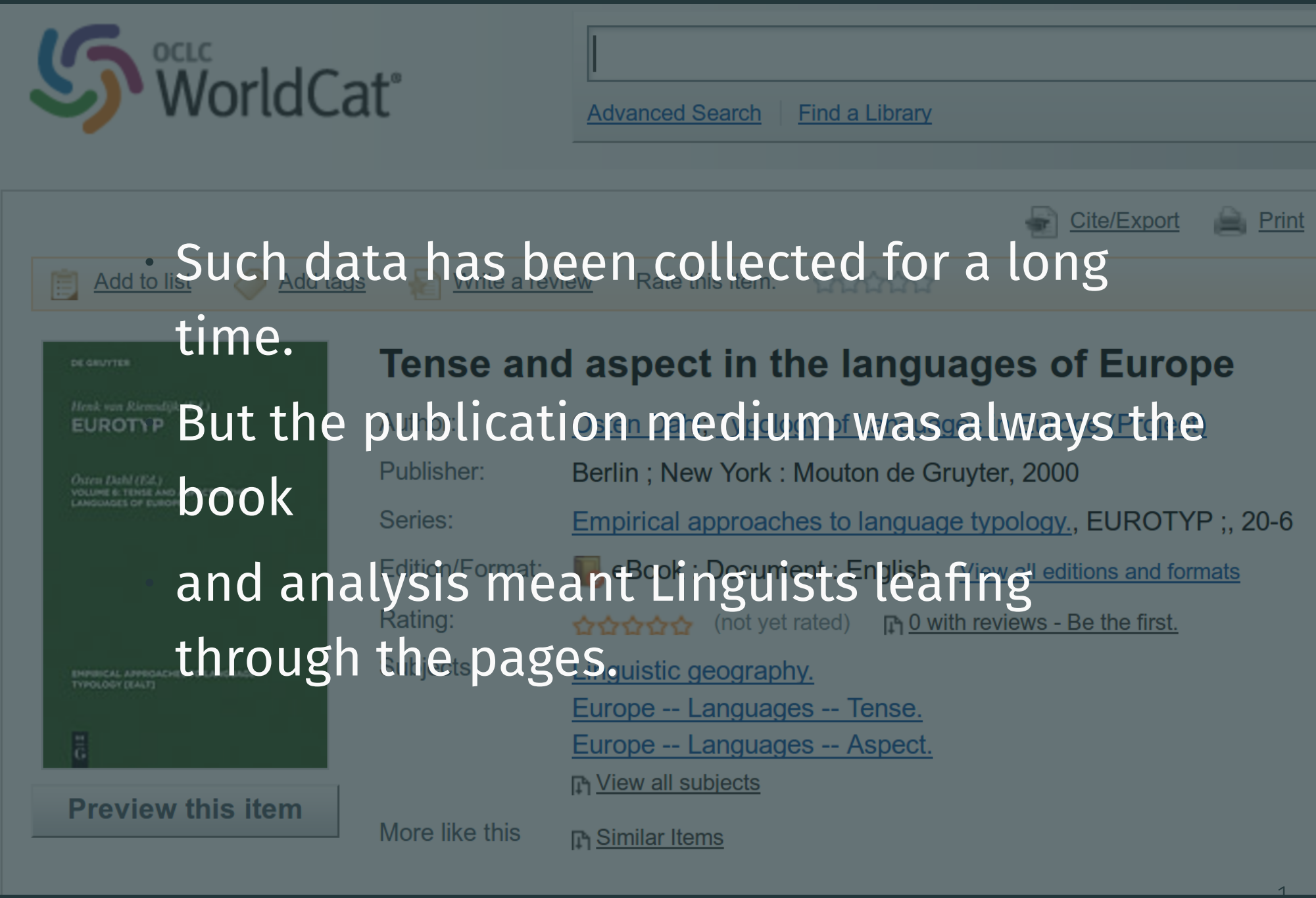
Robert Forkel

Department for Cultural and Linguistic Evolution
Max Planck Institute for the Science of Human History

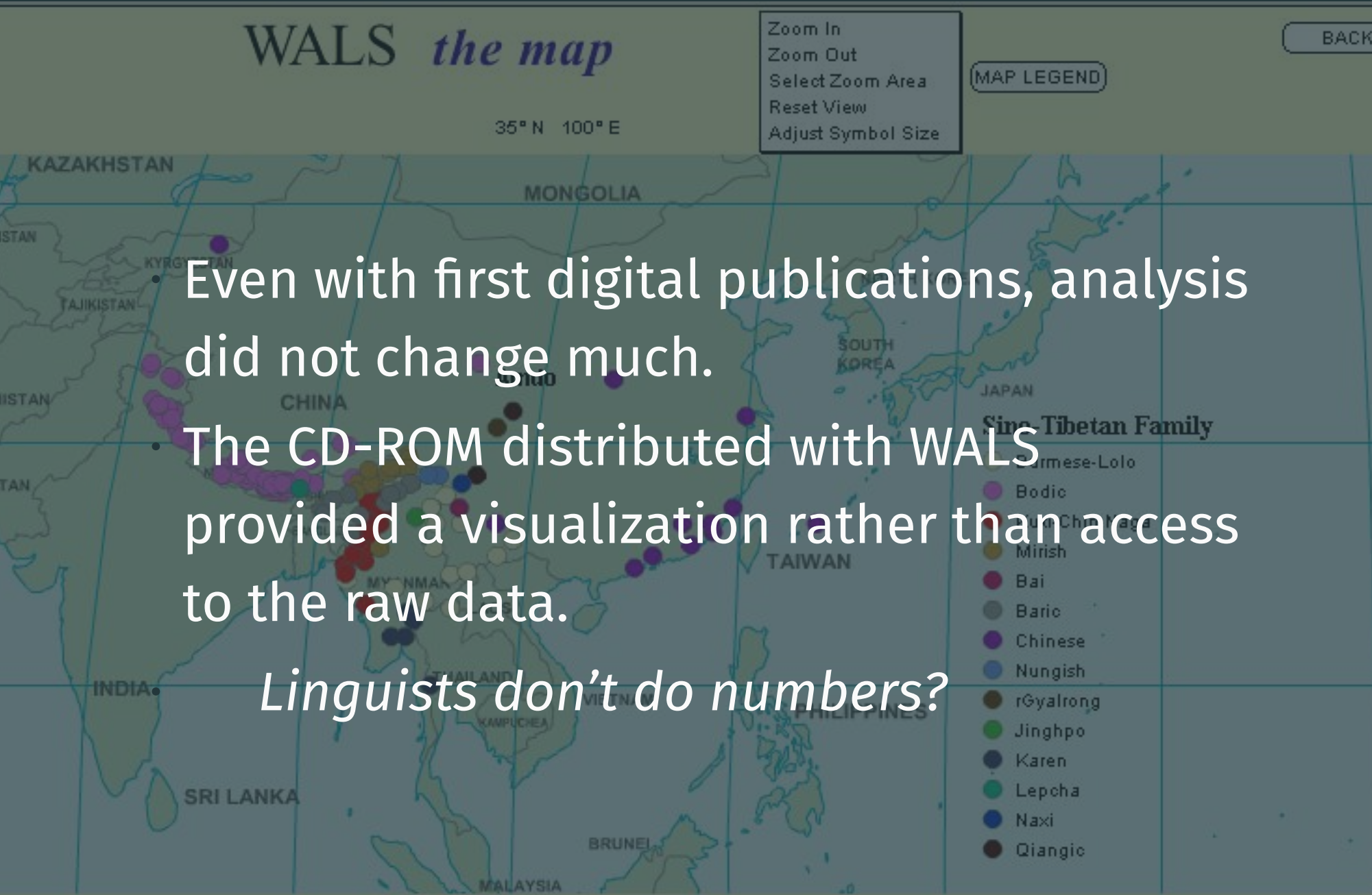
Cross-Linguistic Data

Cross-Linguistic data is data which is available for **hundreds or thousands** of languages. This limits the relevant data types to mostly

- wordlists, i.e. lists of meaning-word pairs
- dictionaries
- typological surveys
- grammars



- Such data has been collected for a long time.
- But the publication medium was always the book
- and analysis meant Linguists leafing through the pages.



- Even with first digital publications, analysis did not change much.
- The CD-ROM distributed with WALS provided a visualization rather than access to the raw data.

Linguists don't do numbers?

The Effect of Language on Economic Behavior: Evidence from Savings Rates, Health Behaviors, and Retirement Assets

M. Keith Chen

AMERICAN ECONOMIC REVIEW
VOL. 103, NO. 2, APRIL 2013
(pp. 690-731)

Economists do!

And so more Linguists started to think about how their data may be analysed quantitatively.

Download Full Text PDF

Article Information

Abstract

Languages differ widely in the ways they encode time. I test the hypothesis that the languages that grammatically associate the future and the present, foster future-oriented behavior. This prediction arises naturally when well-documented effects of language structure are merged with models of intertemporal choice. Empirically, I find that speakers of such languages: save more, retire with more wealth, smoke less, practice safer sex, and are less obese. This holds both across countries and within countries when comparing demographically similar native households. The evidence does not support the most obvious forms of common causation. I discuss implications for theories of intertemporal choice. (JEL D14, D83, E21, I12, J26, Z13)

DATA

Bad Data Costs the U.S. $3 Trillion Per Year

by Thomas C. Redman

SEPTEMBER 22, 2016

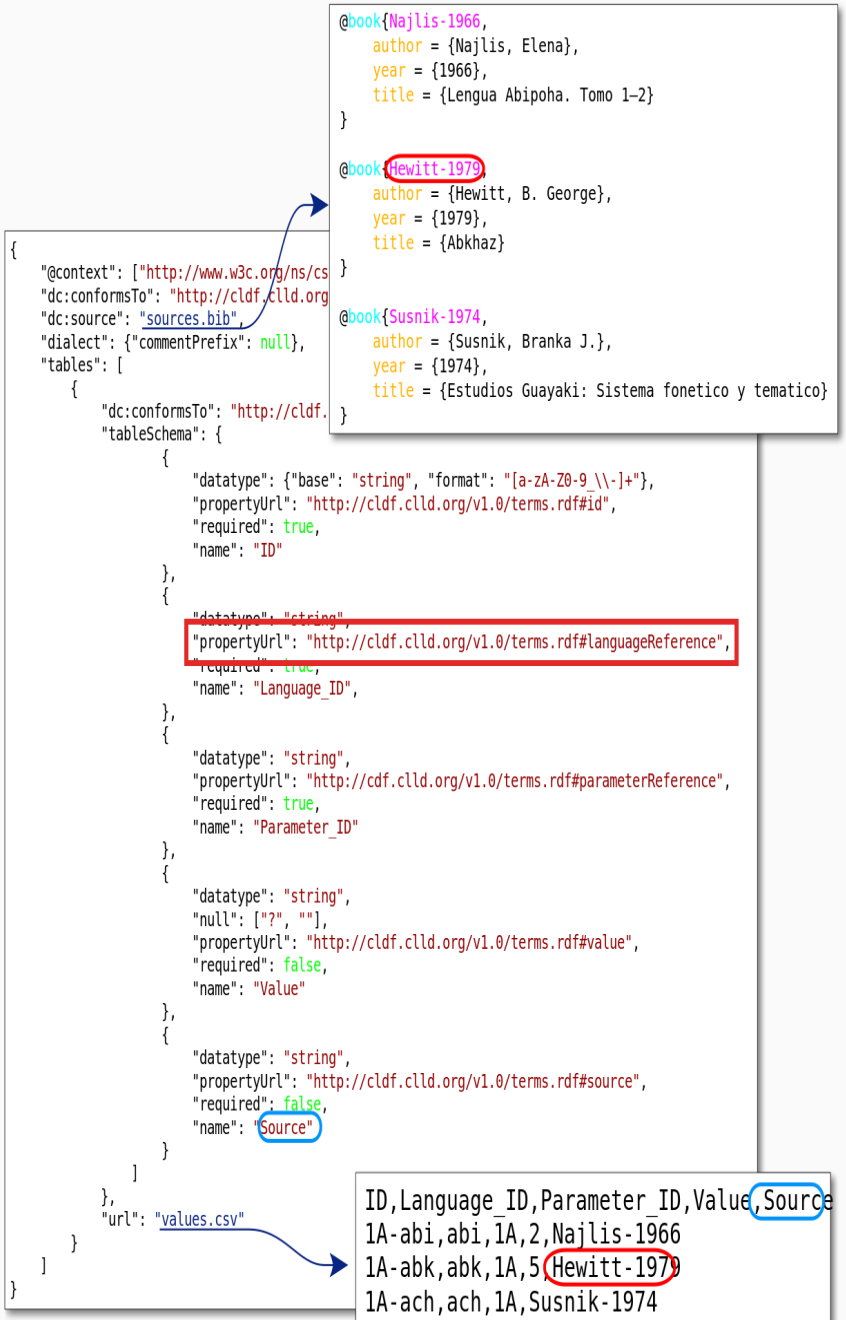
Just to find out that this was harder than expected:

- WALS had the “sparse matrix” problem, and no simple way to merge in other data,
- WALS also has inter-dependent variables.
- IDS has no specification of the transcriptions it uses.
- Language were often identified by name.

Cross-Linguistic Data Formats provides a framework to address these data quality problems.

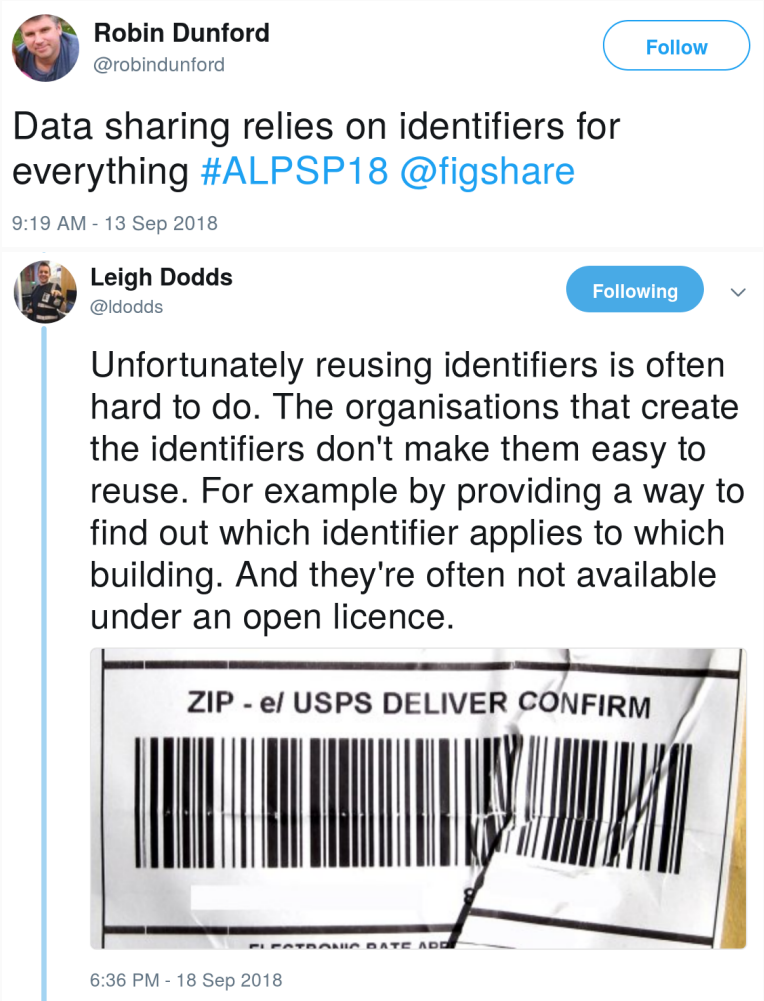
- A CSV package format
- adding semantics via
  - an ontology,
  - modularization and
  - reference catalogs

Package format



based on w3c's spec for "Tabular Data and Metadata on the Web" (CSVW) using JSON-LD (Linked Data) to tie data to the CLDF ontology and reference catalogs.

Reference Catalogs



We provide reusable identifiers via reference catalogs for

- languages and varieties: **Glottolog**
- semantic concepts: **Concepticon**
- transcription systems: **CLTS**
- structural features of languages: **Grammaticon**

Tooling

- `pycldf` – a Python package
  - to read and write CLDF (although CLDF can be edited “by hand”, too)
  - to convert CLDF to SQLite for scalable analysis
- CLDF decouples tool development from particular datasets

  Code against an interface not an implementation!

  e.g. BEAST analyses can use CLDF data via BEASTling
- Environments without full CLDF support (e.g. R) can still access the CSV data in CLDF packages (or the SQLite databases created by `pycldf`)

Summary

Getting cross-linguistic data ready for quantitative, computational analysis is a lot of work. CLDF allows us to automate some of it.

`https://cldf.clld.org`