

CROSS-LINGUISTIC LINKED DATA

Dateninfrastruktur für Diversity Linguistics

Robert Forkel

18.9.2015

Max-Planck-Institut für Menschheitsgeschichte

- Umfasst Teile von Historischer Linguistik, Typologie und Sprachdokumentation
- beschäftigt sich mit **vielen** oder **kleinen** Sprachen
- typische Daten:
 - Wortliste
 - Grammatikskizze
 - Phoneminventar
 - meist graue Literatur
 - traditionelle Publikationskanäle funktionieren kaum noch

CROSS-LINGUISTISCHE DATEN - EIN BEISPIEL

WALS Online - Datapoint Uradhi / Locus of Marking in the Clause - Chromium

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

Home Features Chapters Languages References Authors

Datapoint Uradhi / Locus of Marking in the Clause

Language: Uradhi

Feature: Locus of Marking in the Clause by Johanna Nichols and Balthasar Bickel

Value: Dependent marking

Examples

Sentence igt-470:

wutpu-nku uma-Ø ute-n
wutpu-nku uma-Ø ute-n
old.man-ERG firewood.ABS pick.up- PST
'The old man picked up some firewood.'

References

- Crowley 1983

cross-linguistische Datenbanken:
- "Messwerte" für
- vergleichbare Merkmale
- von Sprachen
- mit Beispielen und
- Referenzen

History

2008-04-21 Dependent marking

Figure: Ein WALS datapoint illustriert typische cross-linguistische Datensammlungen.

- Wir leben also fast im Naturwissenschaftlichen Shangri-La der Forschungsdaten – simple Messreihen
- abgesehen von der Kalibration der "Messgeräte" :), die zwar hochsensibel sind, aber zu Inkonsistenz neigen
- und dem Umstand, dass Datengeber und Datennehmer sich kaum überschneiden

- Von der MPG für 4 Jahre gefördert, anfangs am MPI EVA in Leipzig, jetzt am MPI SHH in Jena.
- Brücke zwischen Datengebern und Datennehmern
- Datenpublikationsplattform für Diversity Linguistics:
 - Referenzdatenbanken: Sprachkatalog und Bibliographie
 - Datenjournals: Dictionaria für Wörterbücher und JCLD für Datenbanken
 - Typologische und lexikalische "standalone" Datenbanken
 - Publikationsformen sind an traditionellen Vorbildern orientiert: Journal, Buchreihe

Wir müssen also in vielen Fällen ein grundlegendes Problem wissenschaftlicher Datenbanken lösen, nämlich Zitierfähigkeit (also Zugang zu älteren Versionen) mit Aktualisierbarkeit in Einklang zu bringen.

RESEARCH DATA MANAGEMENT MIT GITHUB

Linguistische Forschungsdaten haben genug mit code gemein, um tools, workflows und best practices aus der Open Source Software-Entwicklung zu borgen.

Unsere Daten sind

- Text
- oft in zeilenbasierter Form (etwa CSV)
- kleine Datenmenge
- häufig offen zugänglich

- git** · source code management tool
 - Ähnlich wie *CVS* oder *Subversion*, aber *distributed*
 - \implies jeder checkout ist voll funktionsfähiges repository

- GitHub** · Hosting Plattform für git repositories
 - Ergänzt git mit zusätzlichen Kollaborations-tools, insbesondere pull requests
 - webhooks erlauben Integration mit anderen Services, etwa
 - Archivierung via ZENODO
 - continuous integration via Travis-CI

KOLLABORATIVE DATENPFLEGE MIT GITHUB

Best practices der open source Softwareentwicklung übertragen auf Datenpflege:

commit history audit trail für alle Änderungen

merge Prozedur zum Zusammenführen von Änderungen mehrerer Autoren

pull request Prozedur für Einreichung und open peer review neuer Daten

release Publikation

continuous integration Validierung

fork Transparenter Wechsel der Verantwortlichkeiten

DATA IS CODE - BEISPIEL TSAMMALEX

The screenshot shows the Tsammax website interface for the species *Panthera leo* (Lion). The browser address bar shows the URL `tsammax.cld.org/parameters/pantheraleo`. The page title is "Species *Panthera leo* (Linnaeus, 1758) (Lion)".

Navigation tabs include: Home, Names, Languages, Taxa, Ecoregions, References, Images, and Contribute!

Buttons for "Map", "Pictures", and "Names" are visible. The "Map" button is active.

Biological classification:

- kingdom: Animalia
- phylum: Chordata
- class: Mammalia
- order: Carnivora
- family: Felidae
- species: *Panthera leo*

Characteristics: carnivore; height: 110-120 cm; the Lion is the largest cat in Africa and the only one where the males have manes, the colour of the mane ranges from blonde to black, the back of the round ears is black; the males are much larger than the females and have manes;

Countries: [Map icon]

Ecoregions: [Map icon]

Links: [col](#), [wikipedia](#), [GBIF](#), [CatalogueOfLife](#), [BHL](#)

References: [Cillie 1997: 110](#)

Map: A map of southern Africa showing the distribution of *Panthera leo*. The map includes labels for various locations: xam, gám, c'óo nqàré, xám, qà_bèè_qò, n'nhà, tsón, Gaborone, taú, and k'ò:i. The map also shows the Ghanzi region and the Gaborone region. The map is interactive, with a "Show/hide Labels" checkbox checked and a "GeoJSON" button.

Figure: Tsammax Applikation

DATA IS CODE - TSAMMALEX DATA REPOSITORY

The screenshot displays the GitHub repository page for `cclid/tsammalex-data`. The repository is described as a multilingual lexical database on plants and animals. Key statistics shown include 316 commits, 1 branch, 3 releases, and 6 contributors. The file list includes `fixed tests`, `tsammalexdata`, `.gitattributes`, `.gitignore`, `.travis.yml`, `CONTRIBUTING.md`, `MANIFEST.in`, `README.md`, `RELEASE.md`, `setup.py`, and `species.json`. The sidebar on the right contains navigation options: Code, Issues (6), Pull requests (0), Pulse, and Graphs. The `HTTPS clone URL` is `https://github.com/cclid`, and there is a `Download ZIP` button.

File Name	Description	Last Commit
fixed tests	fixed tests	2 months ago
.gitattributes	fixed suffix	9 months ago
.gitignore	updated data	9 months ago
.travis.yml	added email notification config for travis	8 months ago
CONTRIBUTING.md	Create CONTRIBUTING.md	5 months ago
MANIFEST.in	implemented functionality to harvest external data	9 months ago
README.md	work on integration of dogon data	8 months ago
RELEASE.md	Update RELEASE.md	5 months ago
setup.py	updated taxa info from external sources	5 months ago
species.json	added data parsed from mediawiki	2 years ago

Figure: Tsammalex data repository

DATA IS CODE - UPDATES

Submission of a new feature to WALS by xrotwang · Pull Request #31 · cld/wals-data · GitHub · Chromium

Submission of a new feature to WALS #31

Closed xrotwang wants to merge 3 commits into `cld:master` from `unknown repository`

Conversation 2 Commits 3 Files changed 2

pull request bündelt vorgeschlagene Änderungen und open review

xrotwang commented on Apr 24 Owner

... description of the feature ...

xrotwang added some commits on Apr 24

- Added a new (fake) feature: 145A ... fea633d
- Added Feature metadata 069899a

xrotwang commented on Apr 24 Owner

Review comment: New features should cover at least x languages.

- Update feature-145A.cldf.csv 00b4fa8

pull request bündelt review Kommentare und daraus resultierende Modifikationen

Figure: Pull request: Transparentes update mit review

DATENBANK ON-DEMAND

Oft ist eine Datenbank mehr als nur ein dump der Daten: Die Applikation

- implementiert die Standardinterpretation der Daten und
- vermittelt Standardzugang zu den Daten (die API)

Wie können wir Zugänglichkeit der API von früheren Bearbeitungsständen der Datenbank ermöglichen?

The screenshot shows the Tsammalex website in a Chromium browser. The website has a navigation menu with tabs: Home, Names, Languages, Taxa, Ecoregions, References, Images, and Contribute!. Below the navigation is a secondary menu: Legal, Download, Contact, Help, and Contributors. The main content area is titled "Welcome to Tsammalex" and contains introductory text. A terminal window is overlaid on the page, showing a Python script that uses the Tsammalex API to retrieve dataset information. The terminal output shows the successful execution of the script, including the dataset name, license, and citation information. A "Cite" section on the website provides the citation details for the dataset, including the authors, title, and DOI. A "Version" section indicates the latest released version of the data. A red box highlights the "Version" section.

```
Python 2.7.3 (default, Jun 22 2015, 19:33:41)
[GCC 4.6.3] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> from clldclient.database import Database
INFO:rdflib:RDFLib Version: 4.2.0
>>> tsammalex = Database('tsammalex.clld.org')
>>> tsammalex.dataset.license
INFO:clldclient.cache:cache hit http://tsammalex.clld.org/
http://creativecommons.org/licenses/by/4.0/'
>>> print tsammalex.dataset.citation
Christfried Naumann & Steven Moran & Guillaume Segeer & Robert Forkel (eds.) 2015.
Tsammalex: A lexical database on plants and animals.
Leipzig: Max Planck Institute for Evolutionary Anthropology.
(Available online at http://tsammalex.clld.org, Accessed on 2015-09-15.)
```

Lexical and biological data can be accessed directly (tabs "Names" and "Taxa", respectively) or filtered for specific languages ("Languages") or geographical regions ("Ecoregions"), with varying details. The tabs "References" and "Images" include lists of sources and individual images, while "Contribute!" provides more information, especially for potential contributors.

Cite

Christfried Naumann & Steven Moran & Guillaume Segeer & Robert Forkel (eds.) 2015. Tsammalex: A lexical database on plants and animals. Leipzig: Max Planck Institute for Evolutionary Anthropology. (Available online at <http://tsammalex.clld.org>, Accessed on 2015-09-15.) DOI: [10.5281/zenodo.17571](https://doi.org/10.5281/zenodo.17571)

Version

tsammalex.clld.org serves the latest released version of data curated at [clld/tsammalex-data](https://github.com/clld/tsammalex-data) - currently v0.3

Figure: Applikation stellt API für Daten zur Verfügung

DATA AND CODE

- Daten müssen zusammen mit code publiziert und archiviert werden.
- Code muss “bootstrapping” unterstützen, d.h. die Initialisierung einer lokalen Datenbank und Applikationsinstanz.

```
robert@astroman: ~/venvs/clldclient
Python 2.7.3 (default, Jun 22 2015, 19:33:41)
[GCC 4.6.3] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>> from clldclient.database import Database
INFO:clldclient.database:12.3
>>> tsammalex = Database('localhost:6543')
INFO:clldclient.cache:cache miss http://localhost:6543/
INFO:requests.packages.urllib3.connectionpool:Starting new HTTP connection (1): localhost

Christfried Naumann & Steven Moran & Guillaume Segerer & Robert Forkel (eds.) 2015.
Tsammalex: A lexical database on plants and animals.
Leipzig: Max Planck Institute for Evolutionary Anthropology.
(Available online at http://tsammalex.clld.org, Accessed on 2015-09-15.)

>>>
```

⇒ Datenbank on-demand!

DATENBANK ON-DEMAND

DOI 10.5281/zenodo.11040

25 July 2014

Dataset Open access

WALS Online, July 2014, with unfreeze support.

Matthew J. Haspelmath, Martin J. Forkel, Robert

Matthew J. Haspelmath, Martin J. Forkel, Robert

Institute for Linguistic Theory and Applied Linguistics

Files

Name	Size	Type
solite.ini	1.2 kB	Configuration file
setup.py	1.2 kB	Python script
setup.cfg		
requirements.txt		
README.txt		
README.md		
MANIFEST.in		
data.zip		
CHANGES.txt		
alembic.ini		
.gitignore		
wals3		
migrations		
data		

```
virtualenv --no-site-packages wals
cd wals/
. bin/activate
curl -O http://zenodo.org/record/11040/files/wals3-v2014.2.zip
unzip wals3-v2014.2.zip
python c11d-wal
cd wals3/
pip install -r
python setup.py
```

WALS Online -- Chromium

localhost:6543

THE WORLD ATLAS OF LANGUAGE STRUCTURES ONLINE

Home Features Chapters Languages References Authors

Welcome to WALS Online

Figure: DOI -> DB

Der Teufel steckt im Detail (und externe Abhängigkeiten überall):

- Python-Pakete von PyPI
- Javascript vom CDN
- Base-layer für maps

Dennoch

- moderne Container- oder Virtualisierungslösungen (docker, ec2) ...
- mit standardisierten Applikationstemplates ...
- angeboten von Rechenzentren

könnte DOI -> DB als one-click Angebot für Wissenschaftler möglich machen.

ZUSAMMENFASSUNG

Für Datenerzeuger

Für code-ähnliche Daten lohnt ein Blick über den Zaun zur Software-Entwicklung – entsprechend John Nerbonne's Digital Humanities Motto:

Beg, buy, steal or borrow! – <https://twitter.com/TomKnieper/status/448820937446932480>

Für Humanities Data Center

"DOI->DB" könnte eine Antwort auf die "Mehr als wegspeichern?" Frage sein.

clld.org

