Humanities & Linguistics Data Standards: State of the art & challenges

Arienne M. Dwyer

Co-Director, Institute for Digital Research in the Humanities & Professor of Linguistic Anthropology,
University of Kansas
anthlinguist @ ku.edu

Research powered by:

Gist

- data types & metadata schemas
- metadata schemas
- repositories
- challenges
- focus on linguistics (rel. stdzd/interop)
- commonalities & differences with other domains

Arts & Humanities

- data types - *telling a story via:*
  - common: structured text, images (maps, photos);
  - less common: audio, video, spatial data
- ltd. sense of metadata (but discussed)

Arts & Humanities: metadata schemas

- Humanities: basically none, though
  - TEI (Text Encoding Initiative)
  - MEI (Music Encoding Initiative)
- Arts: Art History - primarily images - Getty, LoC standards.

Archives - largely project-specific or university-specific (D-Space, etc.)

http://www.tei-c.org/index.xml., http://music-encoding.org

Arts: metadata schemas

Controlled vocabularies -

- Getty Categories for the description of works of art
- VRA (visual resources association) core
- Cultural objects name authority

http://www.getty.edu/research/publications/electronic_publications/cdwa/index.html

http://www.loc.gov/standards/vracore/

http://www.getty.edu/research/tools/vocabularies/cona/index.html

Humanities: challenges

- Data rarely considered “data”, instead “sources, materials”
    - quant/stdzn = Trojan horse of positivism?
- metadata is ad hoc (in scope + format)
- existing stds (e.g. TEI), are text/lit focused
- Thus, low interoperability & little incentive to archive… back to this later

HUM Data repositories:

Only 3 repositories that maintain and archive *general humanities* research data:

UK Data Archives (http://www.data-archive.ac.uk)

Cultural Policy & the Arts National Data Archive (CPANDA) (cpanda.org)

Association of Religion Data Archives (ARDA) (thearda.com)

Why not a Tree of Life *or a* Dataverse *for the humanities?*

- We need both general humanities and discipline-specific data repositories.

Dariah - Arts/Hum infrastructure

DARIAH-EU | Digital Research Infrastructure for the Arts and Humanities

Contact | Login | Search: Type something

About | What we do | Contributions | News | Service

What is DARIAH?

DARIAH supports digital research in the arts and humanities. Our members provide digital tools and share data as well as know-how. Learn more about DARIAH

Members and Partners

DARIAH is a network: It connects hundreds of scholars and dozens of research facilities in currently 22 countries. Discover who they are

Latest News

Speaking With One Voice: DARIAH Joins ERIC Forum

To strengthen relations between Research Infrastructures DARIAH's Director Jennifer Edmond signed an agreement with several other ERICs.

16 May 2017

DARIAH Working Group "Community Engagement": "Defining Communities in Research" Survey- First Results Online

The working wroup "Community Engagement" started a campaign to identify research communities relevant to DARIAH's Virtual

Research Infrastructure News

CESSDA: New EU Rules Open the Door to Nordic Cooperation

New data protection rules in the EU may lead to a strengthened Nordic cooperation.

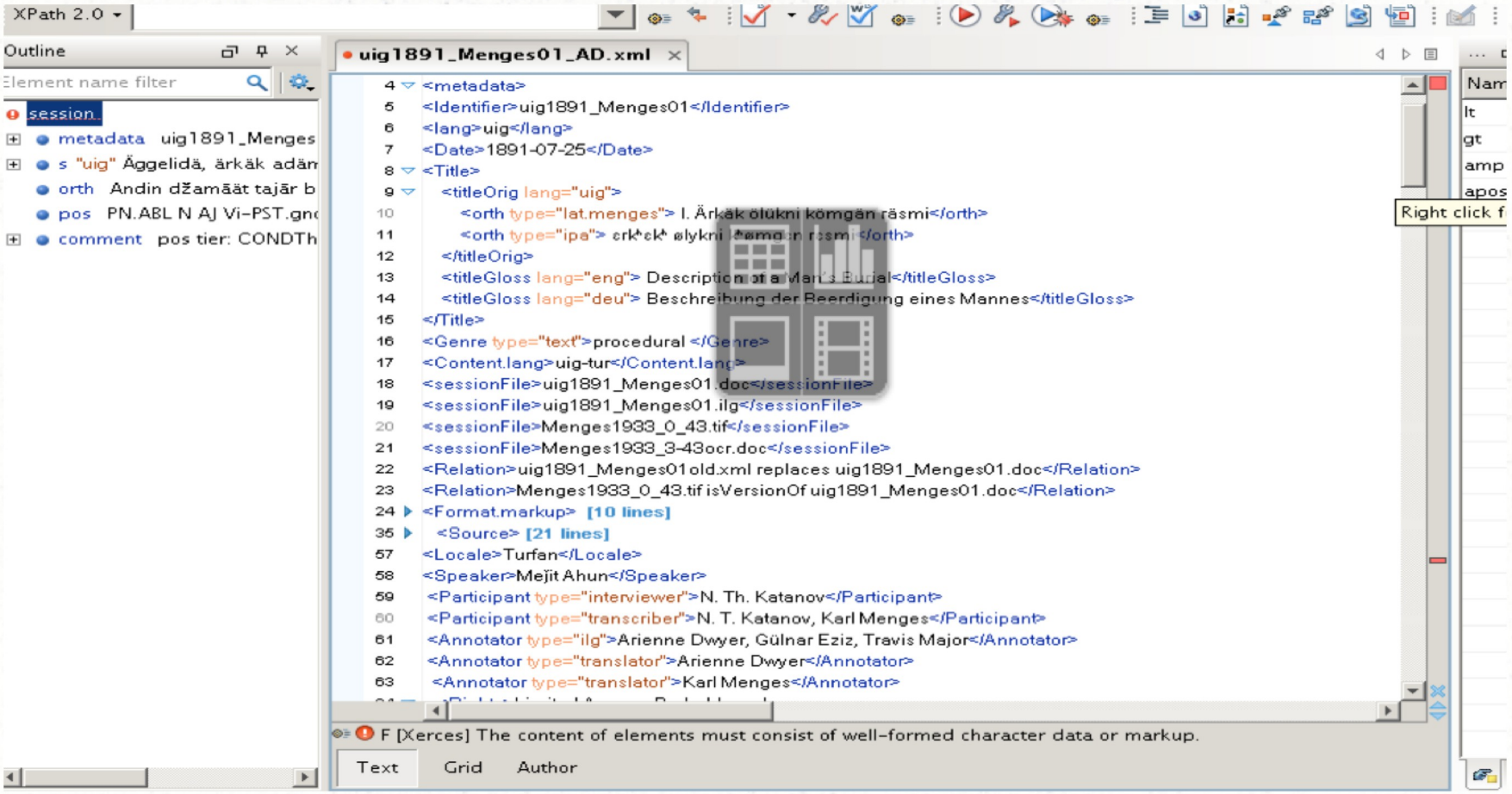
08 May 2017

CESSDA: Useful Tools for Funding Your Social Science Data Archive

The toolkit aims to support organisations wishing to

Documentary Linguistics: ~6700 languages

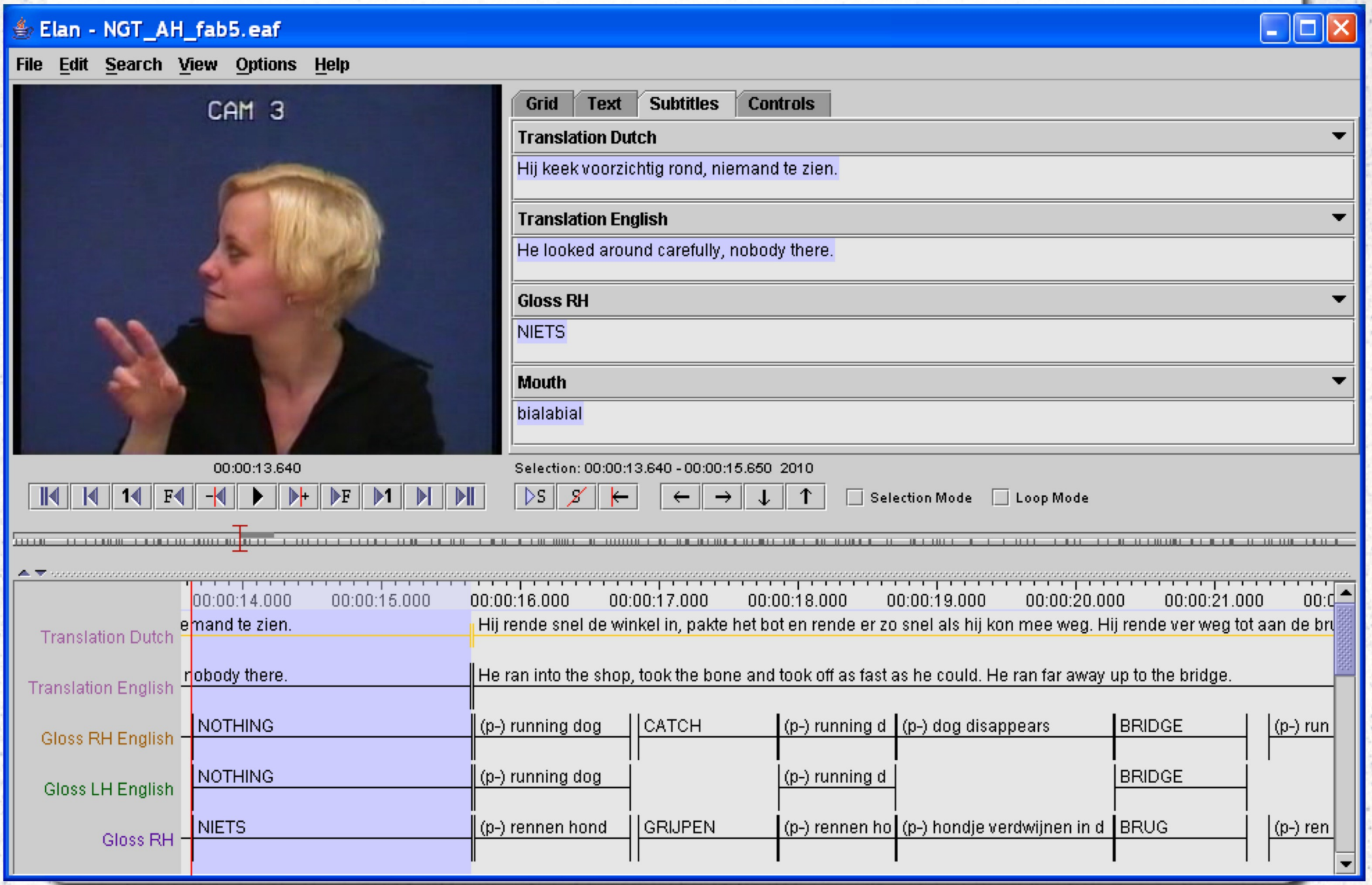
Transcription - structured XML



Data: A/V, situation, participants, genre

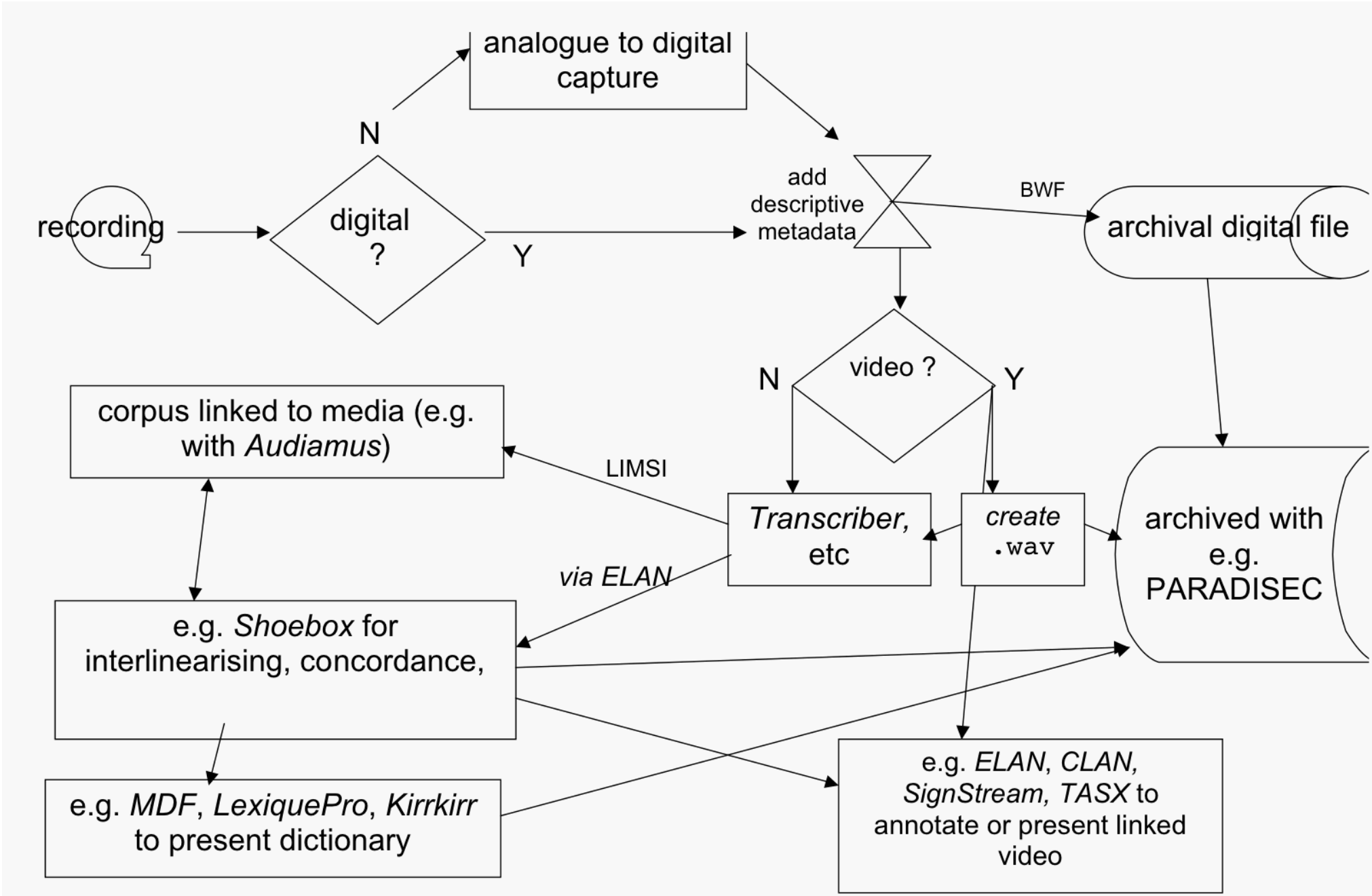
photo © Arienne M. Dwyer

+gesture - also XML, interoperable



https://tla.mpi.nl/tools/tla-tools/elan/attachment/elan-screenshot/

An iterative workflow:



https://www.anu.edu.au/linguistics/nash/fm/flow.html

Documentary Linguistics

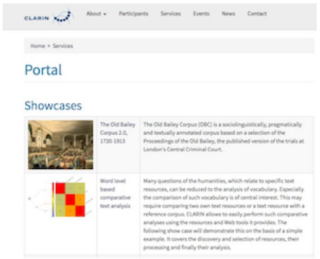
- data types: audio-video, text trans-cription, published texts spatial, etc.
- MD standards: OLAC, IMDI

Infrastructure - CLARIN

Archives - mostly regional, e.g. AILLA, ALNA (US), PARADISEC, (Australia)

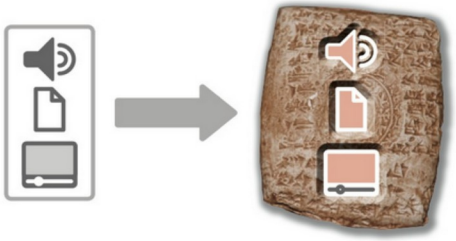
global (but only grantees): The Language Archive (EU/NL); ELAR (UK)

Services



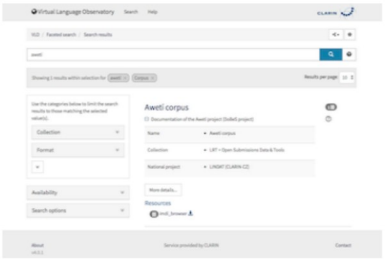
CLARIN portal

Get an example-based impression of what's currently available



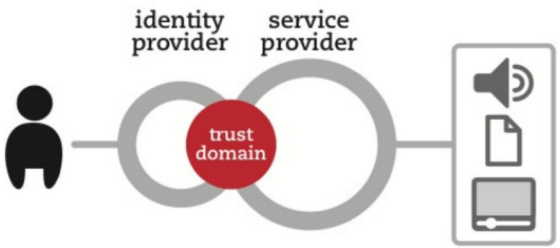
Depositing services

Store language resources in a sustainable repository at a CLARIN centre



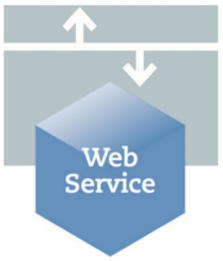
Virtual Language Observatory

Discover language resources using a faceted browser or a map



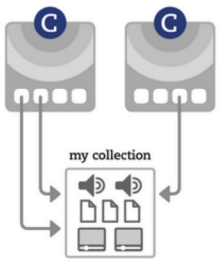
Easy access to protected resources

Get easy access to protected resources, with your institutional



Web services and applications

Explore and analyze language data with a wide variety of tools



Virtual Collections

Create your own digital bookmarks, ideal for citing data sets.

OLAC: Open Language Archives Community

OLAC Mission

OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.

News

OLAC Joins the Linguistic Linked Open Data Cloud: The OLAC system has now been integrated with LLOD... More...

New OLAC Page Listing Archive Submission Policies: As a service to linguists who are in search of an archive that could receive a deposit... More...

New OLAC Search Service: In December 2010, to mark our 10th anniversary, OLAC announces a new search service... More...

Language Archives Tutorial at LSA: In January 2009, an OLAC-endorsed tutorial on language archives was held at the winter meeting of the Linguistic Society of America... More...

OLAC Presented at three Pacific Conferences: In 2008 and 2009, OLAC is being presented at conferences in Australia, the Philippines, and Hawaii... More...

More news ...

Documents

OLAC Standards - specify how OLAC operates

Find Language Resources

Try OLAC's new search engine at: http://search.language-archives.org/

Search

Archive: -- all archives --

Region: Africa Americas Asia Europe Pacific

OLAC Coverage

OLAC Archives contain over 100,000 records, covering resources in half of the world's living languages. More statistics on coverage.

Join the OLAC Community

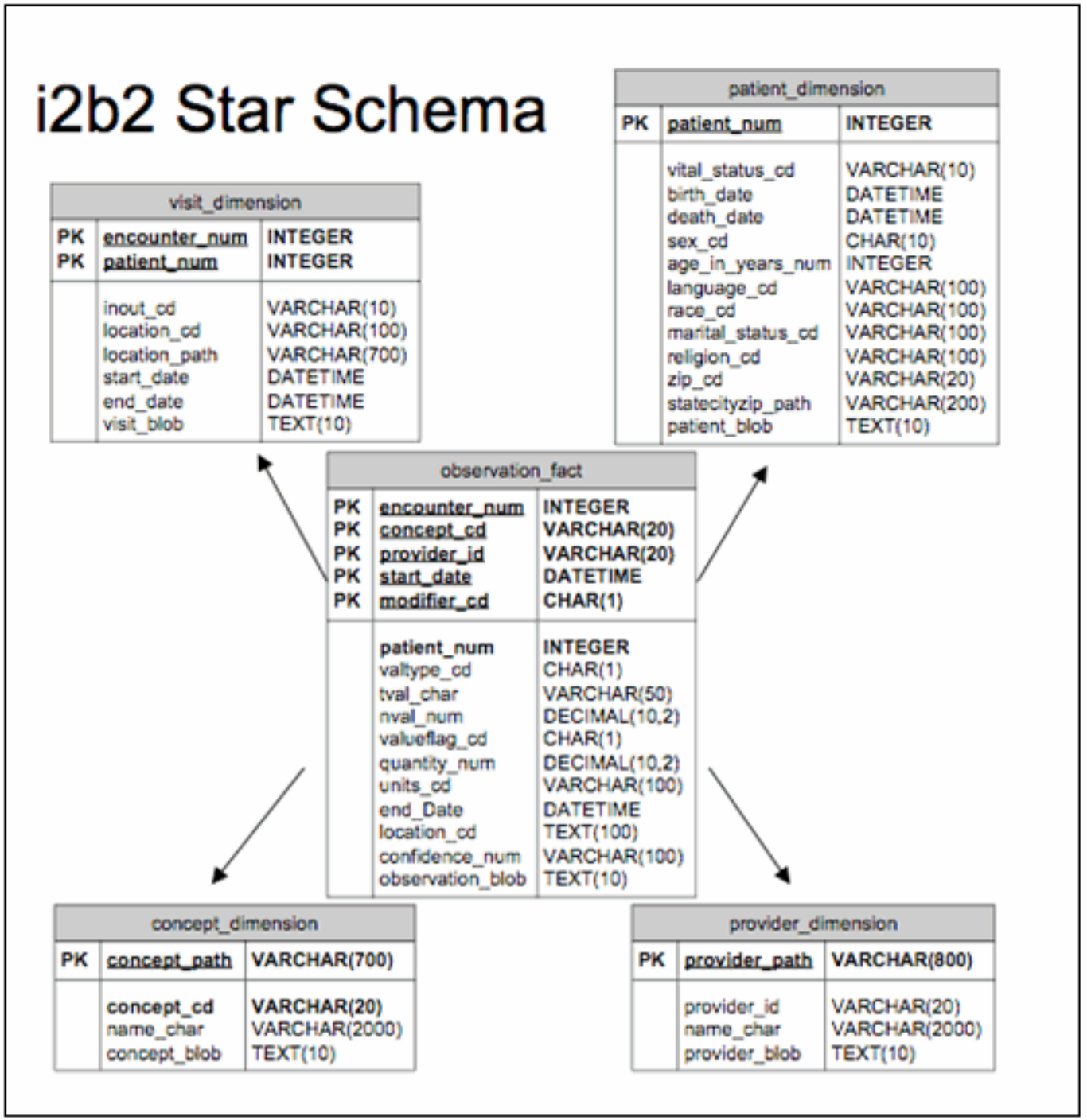
Sign-up for the OLAC mailing list and stay current with standards and best practices for language resource archiving (Archives).

OLAC = DC + extensions

controlled vocabularies for:

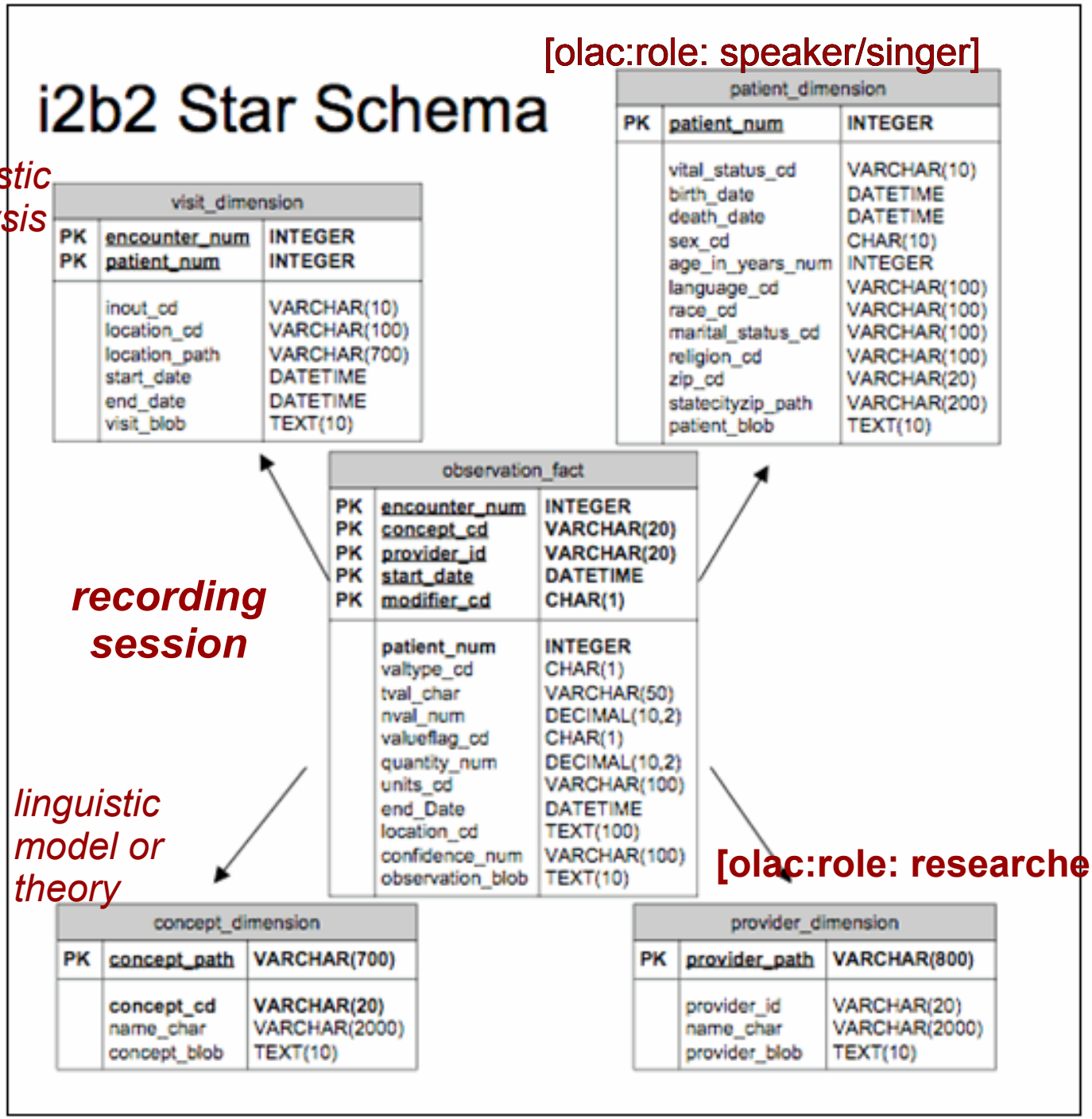
- [olac:discourse-type] - oratory, formula, ludic...
- [olac:language] - 639-3 and beyond
- [olac:linguistic-field] - syntax, semantics...
- [olac:linguistic-type] - grammar, lexicon, text..
- (Participant) [olac:role] - singer, annotator...

i2b2 Star Schema



https://i2b2.cchmc.org/faq

re-imagining the i2b2 star schema for linguistics



https://i2b2.cchmc.org/faq

Challenges

- Multilingual (and incompletely understood) - gathering; code-switching.
- Multi-modal - sound, gesture, social setting.
- Researchers serve Indigenous community
  - Informed consent and possible embargo
    - Difficult to anticipate all future uses of data
    - Avoid embarrassment or taboo

Archival products useful to community & academics

Participatory archiving

- Barriers to Indigenous archive access
- Alternative discovery tools
- Different products (maps, photos, texts vs. linguistic analysis)
- Embargoing
    - (for academics)
    - for community members
    - challenges of changing permissions

Summary

- Art history: metadata stds & repositories well-developed
- Humanities: at least has TEI XML
- Linguistics & heritage language domains
  - OLAC (DC + language info) metadata
  - i2b2-type interactional model possible
  - repositories, but area- or funder-specific
  - participatory: lg. speakers' agency in data collection and access