Dative: from collaborative database to Archival Information Package

ICLDC 5
University of Hawai‘i at Mānoa
March 3, 2017

Joel Dunham (Concordia University & Artefactual Systems)
Open Source Linguistic Data Management
Dative: GUI
Online Linguistic Database
Archivematica: preservation pipeline
OLD LOD Export

- Linked Open Data (LOD) export
- JSON-LD, Turtle, N-triples, RDF/XML
- Uses existing ontologies: GOLD, Dublin Core, PROV, FOAF, Schema.org, SKOS, etc.
- Static, citeable, reusable, web-available, semantically specified data sets
Towards 5 ★ Open Data
**Existing Ontologies**

**gold**: General Ontology for Linguistic Description

**lexvo**: SIL’s ISO 639-3 language codes

**prov**: provenance information: elicitation events, data entry events, etc.

**sioc**: Semantically Interlinked Communities: OLD accounts, users, roles, etc.
BagIt (LoC) transmission format: structure, metadata & checksums

Dublin Core descriptive metadata about data set

OLD database as linked data (triples, JSON-LD, Turtle, etc.)

OLD binaries: audio/video, images, parsers, treebanks, etc.
Why LOD Export?

- Data sets that make sense
- even without Dative/OLD
- even centuries from now
- Exit strategy: get your data out of Dative/OLD
- Reuse: others can query (SPARQL), harvest, link to, cite/attribute your public data sets
“This term is often used in linguistics and phonetics to refer to any entity which constitutes the focus of an enquiry. The unit is the stretch of language that carries grammatical patterns, and within which grammatical choices are made. For example, the unit sentence consists of one or more instances of the unit clause, and so on. [Crystal 2008: 503]”
Interlinear glossed text (IGT) is a linguistic data structure meant to display morphosyntactic structure: morphemes, morpheme boundaries, morpheme types (clitics, prefixes, reduplicated forms), morphosyntactic features/values and part of speech information. At a minimum, an instance of IGT includes a single line of source language followed by a translation line. [...] Standards for IGT include the Leipzig Glossing Rules.”
The relation between some linguistic unit and its corresponding written expression.

The written physical form of language as distinct from either signed or spoken expressions. A written expression is the physical product of the writing process.
OLD-independent linguistic description
Provenance metadata

“This linguistic unit was created by means of an elicitation activity involving a speaker and an elicitor and an entry activity into an OLD application.”
Descriptive metadata (Dublin Core)
Descriptive metadata about entire dataset
OLD LOD/BagIt Exports

• public or private
• public accessible at stable URLs
• queryable via SPARQL / RDF libraries
Archivematica is a preservation pipeline

- standard packaging
- metadata (METS, PREMIS)
- fixity
- file identification & characterization
- file normalization
METS file: everything that happened to your data in the preservation process.
Archivematica METS / PREMIS

• Archivematica uses archival standards—PREMIS in METS—to encode technical & descriptive metadata about the OLD data set and the preservation actions taken on it.

• METS = Metadata Encoding and Transmission Standard

• PREMIS = Preservation Metadata: Implementation Strategies
Descriptive Dublin Core metadata

from our OLD export
Technical metadata

checksum
PRONOM format id
MIME type
Digital provenance metadata

Archivematica created a .tif preservation derivative (normalization event) from an OLD .jpg file.
Archival Information Package

- linked open data
- citable
- web-accessible
- reusable
- preservation-ready
Open Source

https://github.com/jrwdunham/dative

https://github.com/jrwdunham/old-pyramid

https://github.com/jrwdunham/deploy-dative-old

https://github.com/artefactual/archivematica
How to use Dative/OLD

A. Concordia University community server (free, up to 1GB storage space)

B. Install it yourself—it’s open source and free (and has documentation and Ansible install scripts)

C. Contact LambdaBar (joel@lambdabar.com) for custom installations and feature development
Contribute to OLD’s LOD semantics

- https://github.com/jrwdunham/old-pyramid/tree/dev/issue-jsonld-export