Computational Methods in Linguistics

Bender and Wassink 2012
University of Washington
Week 5: 4/23-4/25

Today

- Metadata: What are your questions about it?
- What's in it for me?
- DCMI Level-1 and OLAC metadata basics
- Publishing data and metadata with papers
- Where do I get metadata for a found resource?
- Introduce lab 5

Our Qs about metadata

- To start: where are you currently storing your metadata?
- Are there best practices regarding where metadata ought to be stored and in what format, so it's pretty easy to look for?
- Where might I find metadata for found corpora that are amalgamations of independent resources?

more Qs

- State of the field regarding what is important wrt metadata categories, labels and valid options for identifiers (e.g., multilingual corpora, genres)
- Cross-fertilization between OLAC/DCMI/LDC. What happened to IMDI?
- What kinds of metadata might be appropriate for attitudinal data? (is DCMI helpful here?)

Metadata

- Definition: structured information about data. Metadata is descriptive information about an object or resource whether it be physical or electronic.
- Dublin Core Metadata Initiative (http://dublincore.org/documents/dces/)
- Linguistic Data Consortium (http://www.ldc.upenn.edu/Creating/documentation.shtml)
- xml, or located in a readme.txt file that travels with resource

http://cyberling.elanguage.net/page/Sociolinguistics+metadata+tags

What's in it for me?

- provide a layer of representation that increases data accessibility and availability for further analysis. One person's annotation (record of analysis) can become another person's primary data.
- Facilitation of collaboration (multiple researchers need to consistently annotate parts of the same dataset; now or 10 years from now)
- Extensibility by self or others

DCMI

- 1995, 4 Levels of elements for "simple and generic" description of electronic resources

  - first on-board: library scientists, computer scientists, text encoding and museum community, and other related fields

- What do they mean by "resource"?

  - content (documents, images, data) published on or delivered via the World Wide Web, particularly linked data

- Includes 15 "core" elements or properties (Level -1), plus several dozen related properties and classes (Levels 2-4)

- Some properties are constrained by a relation to a Vocabulary Encoding Scheme (e.g., Type, Subject)

- Our goal: Informal compliance with Level-1

- Group-Level and Recording-Level Tags only

DCES (Element Set)

1. Identifier
2. Format
3. Title
4. Rights
5. Date
6. Coverage
7. Publisher
8. Subject
9. Relation
10. Language
11. Contributor
12. Creator
13. Description
14. Source
15. Type

Element set - http://dublincore.org/documents/dces/

DCMI Level-1 Tags (detail)

- Identifier - "An unambiguous reference to the resource *within a given context*."

  "Recommended best practice is to identify the resource by means of a string conforming to a formal identification system."

- Format - "The file format, physical medium, or dimensions of the resource."

  "Examples of dimensions include size and duration. Recommended best practice is to use a controlled vocabulary such as the list of Internet Media Types [MIME]."

DCMI Level-1 (detail, 2)

- Title - "A name given to the resource."

  "Recommended best practice is to identify the resource by means of a string conforming to a formal identification system. Typically, a Title will be a name by which the resource is formally known."

  so, what's the difference between a title and an identifier?

DCMI Level-1 (detail, 3)

- Rights - "Information about rights held in and over the resource."

  "Typically, rights information includes a statement about various property rights associated with the resource, including intellectual property rights."

- Date - "A point or period of time associated with an event in the lifecycle of the resource."

  "Date may be used to express temporal information at any level of granularity. Recommended best practice is to use an encoding scheme, such as the W3CDTF profile of ISO 8601."

DCMI Level-1 (detail, 4)

- Coverage - "Spatial topic and spatial applicability may be a named place or a location specified by its geographic coordinates. Temporal topic may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies."

  "Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges."

DCMI Level-1 (detail, 5)

- Publisher - "An entity responsible for making the resource available."

  "Examples of a Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity."

- Subject - "The topic of the resource."

  "Typically, the subject will be represented using keywords, key phrases, or classification codes. Recommended best practice is to use a controlled vocabulary. To describe the spatial or temporal topic of the resource, use the Coverage element."

DCMI Level-1 (detail, 6)

- Relation - "A related resource."

  compare to Sociolab tasks and Elicitation Instruments

- Language - "A language of the resource."

  "from RFC4646:

  Language tags are used to help identify languages, whether spoken, written, signed, or otherwise signaled, for the purpose of communication. This includes constructed and artificial languages, but excludes languages not intended primarily for human communication, such as programming languages."

more on “language”

- allows for “subtags”, which may be used to represent:
  - region
  - variant (registered)

DCMI Level-1 (detail, 7)

- Contributor - "An entity responsible for making contributions to the resource. Examples of a Contributor include a person, an organization, or a service. "

  "Typically, the name of a Contributor should be used to indicate the entity."

- Creator - "An entity primarily responsible for making the resource."

  "Examples of a Creator include a person, an organization, or a service. Typically, the name of a Creator should be used to indicate the entity."

DCMI Level-1 (detail, 8)

- Source - "A related resource from which the described resource is derived."

  "The described resource may be derived from the related resource in whole or in part. Recommended best practice is to identify the related resource by means of a string conforming to a formal identification system."

- Type - "The nature or genre of the resource. (collection, dataset, event, image, interactive resource, sound, text, still image, etc.)."

  "Recommended best practice is to use a controlled vocabulary such as the DCMI Type Vocabulary [DCMITYPE]. To describe the file format, physical medium, or dimensions of the resource, use the Format element.

Gaps in the DCMI

- Linguistics as one "application environment" with particular requirements
- Subject, Type are vague (Topic?)
- No speaker-level tags
- No token-level tags
- Phonetic information is too broad (sampling rate, sampling window size, etc.)

Open Language Archives Community (OLAC)

- A metadata set for the "language resource community" application environment
- based upon the Dublin Core metadata set, using all 15 elements (Level -1)
- ...But it is meant to be more precise, allowing for greater precision in resource description.
- Guidelines for xml implementation

http://www.language-archives.org/OLAC/metadata.html

olac:discourse-type

- [dc:type, dc:subject] Provides a controlled vocabulary for identifying approximately ten discourse types. It is used with Type to identify the genre of a language resource (particularly a primary text). It may also be used with Subject to identify a work as being about a particular genre.
- discourse-types:

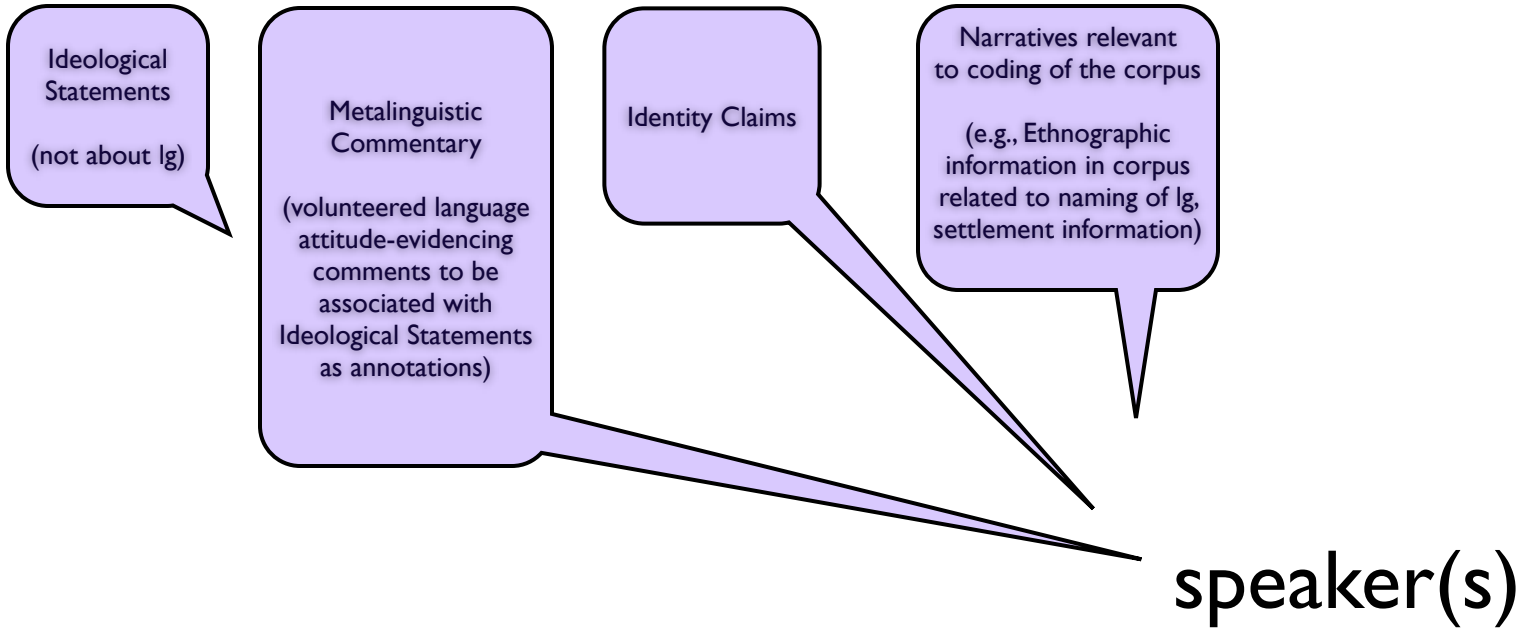
dialogue	oratory	singing
drama	narrative	unintelligible*
formulaic	procedural	
ludic	report	

other OLAC extensions

- olac:language - uses ISO639 codes
- olac:field - allows specification of lx subfield or subject for which a resource was created
- olac:data-type - specifies nature of resource (lexicon, a primary text, or a language description.
- olac:participant - specifies role of participant using 24-item controlled vocabulary)

Still missing

Attitudes



Publishing data with papers

- Linguistic Society of America Technical Advisory Committee
- Journal publishers currently requiring data and analyses, e.g.:
  - Journal of Experimental Linguistics
- ...or encouraging submission of data and analyses:
  - Routledge (Taylor & Francis): "More extensive supplementary material (analyses rather than data) ideally should be subject to peer review."
  - Elsevier (Science Direct)
  - Proceedings of the ACL and related conferences

LSA Technical Advisory Committee Recommendations

Minimally:
Data collection methodology should be transparently and truthfully described:
-- when and where the data was collected,
-- information about informants/consultants
--method of collection (introspection, fieldwork, gleaned from a database, text annotation, …)
-for secondary data, the author should provide clear citations and state
how the correctness of this data has been evaluated

The reviewers should check whether the paper follows these recommendations
and also consider whether the analysis is based on sufficient quantities of data to be deemed reliable.

Data availability should be encouraged (what, where, under which conditions. At the very least all data that the reviewers might be interested in should be made available to them.

Recommendations:
The standard procedure should be that the data are made available to the scientific community from a website that has a reasonable degree of permanency and openness.

Each journal asks submitting authors to fill out an "availability of data" form

Further study to clarify the following two technical points:

1. Hosting data: Should journals be involved or special purpose repositories be created?
2. what type of data should be stored different subfields of linguistics?