A Gentle Introduction to Metadata

Jeff Good
University of California, Berkeley

Source: http://www.language-archives.org/documents/gentle-intro.html

1. Introduction

Metadata is a new word based on an old concept. Any summary of the contents of a library or archive, like a card catalog, contains metadata. It is the preferred term of the technical community to refer to ``card-catalog" data, and it will, therefore, become increasingly used as more technical tools are developed for linguistic research. The purpose of this document is to provide a non-technical introduction describing what metadata is, what the general linguist should know about it, and also to describe some aspects of the metadata standard used by the Open Language Archives Community (OLAC).

2. What metadata is

There is little conceptually new about metadata. The shortest definition for the term is ``data about data," and, while many of us may not be accustomed to thinking about metadata very much, we create it and make use of it all the time. A citation like the one below, is a form of metadata:

Bloomfield, Leonard. 1933. *Language*. New York: Holt, Rinehart & Winston.

The reference above is information about a book--that is, data about data.

Something important to note about a reference like the one above is that we understand the information it is trying to convey by convention. The first element is the author's name, then the year, then the title, then the city and publisher. Also, our knowledge of the basic structure of references allows us to be fairly sure this citation refers to a book.

Another way of representing the metadata in the above reference would be as follows:

Document type	Book
Last name of author	Bloomfield
First name of auther	Leonard
Year of publication	1933
Title	Language
City	New York
Publisher	Holt, Rinehart & Winston

The standard citation and the above table are both different representations of the same metadata. The table has the advantage of explicitly marking the category of information encoded in the various pieces of the citation.

A book citation is a fairly well-known kind of metadata, but the general idea of ``data about data'' is far more inclusive. An annotated bibliography, for example, also constitutes metadata which is very much like a list of references except that it also includes an extra level of description in addition to the basic metadata for the document.

Though they do not match the traditional notion of ``data type'', it's also easy to find metadata about linguistic software, which can often be quite valuable. Below is metadata about the SIL Shoebox field linguist's tool.

Name	The Linguist's Shoebox
Platform(s)	Windows (3.1, 95/98, NT), Macintosh (OS 6, OS 7, 68K, PPC)
Categories	data management, parser, text analysis
Domains	morphology, syntax, discourse, lexicon/dictionary
Keywords	field notes
Version	5
Date	September 2000
Description	Shoebox is a computer program that helps field linguists and anthropologists integrate various kinds of text data: lexical, cultural, grammatical, etc. It has flexible options for sorting, selecting, and displaying data. It is especially useful for helping researchers build a dictionary as they use it to analyze and interlinearize text.
More information	Shoebox Home
Contact	shoebox_support@sil.org
Licensing	Shoebox is licensed for $45.00

One can make metadata for any linguistic resource--there could even be metadata about metadata. For example, a description of a collection of bibliographies would be data about data about data (a.k.a. meta-metadata).

It's also fairly common to include the metadata about a linguistic resource within the resource itself. The title page of a book is one example. The subject and language indices often found at the end of a linguistics book are also good examples of metadata. Both types of indices tell the user about the internal content of that book.

While the basic idea behind metadata may be old, the rise of the internet will allow it to become more useful than it ever has been since metadata will be much more accessible than before--and if metadata is more accessible, then linguistic resources of all kinds will also be more accessible. In order for this to happen, however, researchers must become aware of how to produce good metadata and how to use metadata produced by others.

3. What can be done with metadata

One of the most important uses for metadata is to locate a resource. Thus, a book reference is designed to give enough information to allow someone to find that book.

The other primary use of metadata is resource discovery--that is, finding resources relevant to one's research but which one is unaware of. The subject index of a card catalog is a metadata collection which is good for such a purpose. With the advent of new technologies, there are many more possibilities for the discovery of resources than were previously available. Online indexes of abstracts, like LexisNexis, are well-known tools which makes use of metadata and new technology to greatly enhance the researcher's ability to find relevant resources. These tools are not yet perfect, by any means, but they do represent a significant step forward in resource discovery.

One of the primary goals of the Open Language Archives Community (OLAC) is to help linguists create metadata and distribute it in order to assist in the discovery and location specifically of linguistic resources. The basic reason linguists should consider creating metadata for their resources is to allow other researchers to locate those resources. Of course, this might not confer direct benefit to the owner of the resource. However, as more metadata becomes available, all linguists will ultimately benefit since it will allow them to the find resources the need more easily.

4. What can't be done with metadata

It's important to realize that metadata is simply data about data and not the data in and of itself. Thus, making metadata public is fairly safe--metadata alone does not give people access to the data. A book reference is not a book, and metadata about online data is also not the online data itself.

The most useful metadata makes it clear how a resource, once discovered, can be located and accessed. The owner of the resource gets to determine how easy or difficult access should be. Obviously, it's not very useful to let people access metadata for a document that no one will ever be allowed to access. However, it's important to realize that making your metadata publicly available in no way implies that the resource the metadata describes is also publicly available.

Furthermore, the creator of the metadata gets to determine what information will be included in publicly-accessible metadata. While organizations like OLAC will typically have recommendations for what needs to be included in the metadata for a resource, these recommendations can be overridden if there are any concerns about making any information publicly available.

5. Why should linguists create and distribute metadata

There are a range of reasons to create and distribute metadata. Some of them are listed below.

Management of an archive: Good metadata makes it much easier for a researcher to manage and keep track of the resources in an archive. Large archives almost always have internal metadata standards which are important to their maintenance. Metadata is crucial for the maintenance of any archive in the long term, since it is the best way to ensure that a resource in an archive can be tracked down when someone other than the original archivist needs to locate it.

Increasing awareness of your resources: Though not always desirable, many researchers actively want academics and the general public to be aware of the resources they maintain. Good metadata helps people find resources since it creates an easy way for them to get basic information about them. Using a standard metadata format (like the OLAC standard) is especially useful in this area since search tools are being developed which make direct use of these standards.

Increasing the value of community-wide metadata resources: Metadata is one of those things that becomes more valuable as more of it is made available. If every linguist creates metadata for their resources, then we all benefit by being able to find the data we need for our research. So, even if one of your priorities is not to ``advertise'' your resources, unless you have a good reason to keep them unavailable, you should consider making and distributing metadata about them so that other people will know they are out there. While this may not confer a direct benefit to you, the indirect benefits of being part of a community of researchers where there is an expectation that metadata will be made available are potentially enormous. Consider, for example, the possibilities of having a linguist's ``Google'' where you could type in a language name and be reasonably sure you've found the bulk of documentary work done on that language.

6. Steps required for metadata creation

Most researchers are already fairly skilled at making metadata, even if they are not skilled at making structured, machine-readable metadata. This simply requires understanding the important parameters needed to describe a resource and then recording those parameters in a standardized way. Creating a list of references at the end of a book or article requires a basic knowledge of metadata. Also, most well-managed archives have collected their metadata and stored it in a structured way, most typically in a database.

One of the primary goals of OLAC is to create a standard way to document linguistic resources to assist in their location and discovery. Importantly, OLAC does not seek to dictate how linguists must design their metadata. Rather, it seeks to build a community-driven consensus about linguistic metadata. (An article giving an overview of OLAC is available on the OLAC web site at http://www.language-archives.org as is an article describing the proposed OLAC metadata set.)

Also, even an OLAC-compliant archive can store its metadata in a non-OLAC format. Such a resource would simply ``translate'' its metadata into OLAC metadata in order to make it more widely available.

A more technical term for translating metadata from one format to another is *mapping* and sometimes another term *crosswalking* is also used for this. Many archives have either created non-OLAC metadata or will have good reasons for wanting to do so. However, for the most part, mapping their metadata to OLAC metadata will be fairly straightforward since OLAC metadata uses very general descriptors like ``creator", ``subject", and ``format". Since OLAC metadata is specifically designed for linguistic resources, it also includes a few descriptors like ``subject language" which are of particular value to language researchers.

An example will help to illustrate how mapping from a particular metadata format to the OLAC standard could work. The record below is taken from the internal metadata database for the Comparative Bantu Online Dictionary.

Folder Name	Gevove.vdVeen1994
Language	Gevove
Guthrie	B.30 (A.30?)
Contributor	Van Der Veen, Lolke
Institution	DDL
Author	Van Der Veen, Lolke
Citation	Van Der Veen 1994
Count	1450
Formats	Word 5, Text
Gloss Language	French
Condition	1
Web Search	Y
MapMaker Status	None
Recon Status	None
Remarks	The Word version has a nice introduction which I made into an info file. --JG
Record date	9/23/2001
Set	CBOLD:dictionaries

This record contains a range of information, not all of which is likely to be of equal importance to linguists for resource location and discovery. The language, Gevove, is of obvious interest, and the name of the creator of the data is also fairly important. Other parameters like ``Recon Status" (meaning how much lexical reconstruction had been done on the resource) are likely to be less important for the general researcher (though they may be of importance to a minority of them).

The information in the above metadata record was mapped to OLAC metadata by the author of this article in the following way:

Title	Gevove.vdVeen1994
Subject language	Gevove
Contributor	Van Der Veen, Lolke
Date	2001-09-23
Type	dataset
Formats	Word 5, Text
Identifier	http://bantu.berkeley.edu/CBOLDFTP/CBOLD_Data/Gevove.vdVeen1994
Language	French
Rights	See: http://linguistics.berkeley.edu/CBOLD/Data/CBOLD.data.html

The first thing to note about the mapping from CBOLD internal metadata to OLAC metadata is that information is lost--for example, there is no mention of ``Recon status'' in the OLAC version. OLAC standard metadata does not contain a ``Recon status'' field since this is too specific for general language resources. Translating from a more specific metadata format to a more general one often involves losing some information. This isn't generally a problem unless it substantially hinders resource location and discovery.

It would have been possible to include information like ``Recon status'' in the OLAC metadata because the OLAC standard is flexible. For example, though it does not define a ``Recon status'' field, it does define a *Description* field, where any information not belonging to another field can be put.

If it is possible to put information like ``Recon status'' into OLAC metadata via the *Description* field, why wasn't it done? As the creator of the above metadata records, I can say that this was because it made the task of mapping simpler for me. I made sure the most important aspects of the metadata (like subject language) appeared in OLAC metadata, of course, but also chose to leave some information out.

This ``incomplete'' mapping of CBOLD-internal metadata to OLAC metadata illustrates an important point: Ideal metadata would contain all information possible about a resource. However, OLAC does not enforce this. Rather, it leaves it up to the creator to decide what information to put in the metadata. OLAC standard metadata must have a particular structure and make use of OLAC-defined descriptive parameters, but the creator of the metadata otherwise has a lot of flexibility.

If there's one general piece of advice to give for making metadata, it is to think of the needs of the potential end user of it. This can be difficult, because it is hard to exactly who will need to use your resource. From the OLAC perspective, though, it will usually be worthwhile to think about what kinds of linguists might find yourself valuable and what sort of search criteria would they be likely to use to locate it.

7. How to distribute metadata

Many archives have already designed ways to distribute their metadata, most typically through some sort of online access to their database. The Linguistic Data Consortium, for example, has on online catalog to allow access to its metadata.

Online archive access managed by the archive is certainly very valuable, but it has the problem that the user must know about the archive in the first place in order to locate resources in that archive. Ideally, metadata from all linguistic archives could be accessed and searched in a centralized way so that a linguist need not know about a relevant archive in order to locate useful resources--this requires a generalized system of metadata distribution.

The OLAC metadata standard is specifically designed to make general distribution of linguistic metadata possible. This metadata standard is the first step to the creation of search engines designed specifically for the needs of language researchers. There are various ways to distribute OLAC metadata, some of which are simpler than others. These ways are described in the Implementers FAQ document.

Search engines making use of OLAC metadata already exist, the most important being the one hosted by the LINGUIST List at http://www.linguistlist.org/olac/. The value of these search engines is largely contingent on the metadata made available to them. The more researchers who create OLAC standard metadata, the easier it will be for all researchers to find the data they need.

8. Next steps and further resources

There are a number of resources on the web for anyone interested in learning more about metadata. For linguists, the OLAC implementers site is very useful.

http://www.language-archives.org/docs/implement.html

This site tells you how to become and OLAC data provider and gives various menthods to do so.

Another important set of documents can be found at:

http://www.language-archives.org/documents.html

There, you can find documentation for the latest OLAC metadata standards. These are community-driven standards. So, comments and suggestions for improving them can be made by anyone who intends on using the OLAC standards.

The OLAC metadata standard is based on the metadata standards developed by the Dublin Core Metadata Initiative. The Dublin Core standard is meant to describe all types of resources, and the OLAC standard is an elaboration of the Dublin Core tailored for language resources. The Dublin Core Usage Guide can be found at

http://dublincore.org/documents/usageguide/

This guide is a more technical introduction to metadata than the present document but is still fairly accessible. In addition to basic information on metadata, it outlines the basic rationale behind the Dublin Core and also gives some examples about how metadata can actually be represented in digital document--something which has not been discussed here.

The standard language for encoding OLAC metadata is XML (Extensible Markup Language). XML can be understood as a more generalized form of HTML, the standard language used for web documents. The Text Encoding Initiative Guidelines contain a useful chapter introducing some of the basic concepts of XML:

http://www.tei-c.org/P4X/SG.html

Another very useful resource is the home page for the Cornell University Library Metadata Working Group, which contains links to a wide range of online metadata resources:

http://metadata-wg.mannlib.cornell.edu/readings/index.htm

Finally, if you have any questions that weren't answered by this document, please feel free to contact the OLAC coordinators at olac-admin@language-archives.org.