

## Community Involvement in the Infrastructure Initiative: The LINGUIST List

Helen Aristar-Dry, Eastern Michigan University  
hdry@linguistlist.org

Paper presented at the workshop on  
[Web-Based Language Documentation and Description](#)  
12-15 December 2000, Philadelphia, USA.

### PANEL 2: Community Infrastructure

If we use the term “community” loosely, The LINGUIST List might be said to represent quite a large linguistic community. The list now has 13,500 subscribers in 105 different countries, and the main web site at Eastern Michigan University receives 300,000 page views a week. There are full mirror sites at the University of Tübingen, University of Stockholm, University of Edinburgh, and Moscow State University; and together all five sites receive almost 3,000,000 page views a week. Furthermore, LINGUIST has begun to host and/or archive other linguistics related mailing lists; and over 70 other language oriented mailing lists are currently archived on our site.

This is relevant to the topic of this panel in that LINGUIST’s size and international reach would seem to make it suitable to develop parts of the infrastructure that require either centralization of resources or outreach to the linguistics community. And we are proposing to various funding agencies a project which exploits LINGUIST’s potential in these two areas.

The planned outreach efforts include workshops, digital institutes, and a Showroom of Best Practice, which will offer data from 10 endangered languages illustrating best practice in markup and corpus design. The Showroom is important, because we firmly believe that standards cannot be promulgated in the abstract; rather, they have to be embodied in data that demonstrates their usefulness. However, it is the centralization of information that I will focus on here, since this workshop gives us an opportunity to solicit feedback on LINGUIST’s plan to establish a central metadata server.

The need for such a server was clearly articulated in the Workshop Overview and in the RFC listing the infrastructure requirements of the different stakeholders. As indicated on the Overview chart (a few lines of which are reproduced below), a central metadata server would store metadata about language resources, support user query of the database, and display the results.

	Needs				
	Data Models & Archives	Data Creation		Data Access	
Data Type	Store	Create	Convert	Display	Query
Metadata	Linguist List	LL?	LL?	LL	LL

Word Lists					
------------	--	--	--	--	--

Primary responsibility for creation of metadata would lie with the data provider, and other organizations, e.g., the LDC, would be the primary developers and purveyors of conversion software. However, as I will explain below, a central metadata server might also usefully become involved in these enterprises.

Such a server, and the associated metadata standards, would go far toward fulfilling User Requirements 1-5 (reproduced below), as listed in the RFC.

**User Requirements:**

1. There is a single site on the Web where any user can go to discover what language information resources are available, regardless of where they may be archived.
2. All language resources (regardless of where they may be archived) are catalogued with a consistent set of metadata descriptions, so that the user can ascertain all the basic facts about a resource without having to download it.
3. Uniform metadata descriptions can be used to perform focused searching of language resources by metadata categories regardless of where on the Web they may actually be archived.
4. All language resources (regardless of where they may be archived) are tagged in a consistent way to identify the languages they relate to, so that a single search for a particular language will retrieve all relevant resources on the Web.
5. When a user discovers the existence of a resource, full information is available on how to obtain the resource, and on any restrictions concerning the use of the resource. The resource can be obtained in a timely fashion.

Requirements 2 and 3 will be met if consensus is reached about a metadata standard and that standard is widely implemented. Requirements 4 and 5 treat the nature of the standard, suggesting that a uniform scheme for language-naming and language classification be adopted, and that metadata should include information about resource availability. Although establishing a metadata server is logically separate from defining a metadata standard, the server can be instrumental in publicizing the recommendations. And indeed the existence of a centralized repository, with flexible search and retrieval capabilities, may even be crucial to the standard's wide adoption, since what is true of language markup is also true of metadata: standards can not be promulgated in the abstract; rather they must be embodied in something which demonstrates their usefulness.

The greatest difficulty in creating such a repository may lie not in establishing a central site but in maintaining it--in finding language resources, inducing their owners to provide appropriate metadata, and keeping track of the resources afterward. In our admittedly biased view, an appropriate central site already exists; and it is equipped with 5 powerful Unix machines and an installation of Oracle database software. But, unless a much better-funded organization than LINGUIST takes it over, the repository of language-

related metadata will be a monitored but partly user-maintained facility. And, like all user-maintained databases, it will have to confront the very human obstacles of ignorance and sloth. As yet, most linguists have never heard of metadata. And, even if they learn of a metadata collection effort, many may not wish to participate—perhaps because they don't recognize the value of their resources, perhaps because they feel intimidated by digital enterprises, perhaps because they simply don't want to go to the trouble.

To overcome these obstacles, we may need more than publicity efforts, user-friendly input forms, and field software which facilitates appropriate documentation. We may need a repository which can find relevant sites by itself and instigate metadata creation and collection. Although this may seem a distant goal, we believe that the foundations have already been laid.

Running on the LINGUIST site is a piece of software which extracts URL's from LINGUIST issues and appends them to a file. This software has been running since 1997, and in that time virtually every URL relevant to the discipline of linguistics has been mentioned in some LINGUIST posting or other; moreover, we will shortly extend this URL-grabber to the other lists we archive. We have now implemented a spider which searches this list of URLs and indexes all the words on the listed sites. A primitive search facility based on this index is available at

<http://linguistlist.org/~zheng/sitesearch2.html>

This facility could be extended so that it recognizes language documentation using a set of keywords, categorizes material within <meta> tags separately from the rest of the text, and stores the metadata in a database. Moreover, if a relevant site lacks metadata, the system could send a message to any email addresses on the pages and request that the recipient complete a form at our site. This form could not only generate standardized metadata and save it to our database, but it could also return a formatted copy to the information provider, with the request that it be included in the web pages.

And this, in turn, could help to track perambulating sites. As noted in Archivist Requirement 4, some type of unique identifier—a kind of ISBN—must be assigned to resources, so that they can be recognized even if their URL changes. Assignment of the identifier might reasonably take place when the resource's metadata is entered into the central database. The identifier could then be included in the standardized metadata returned to the resource-holder, and if the metadata is put into the web pages as requested, the identifier will allow search engines to find the resource, whatever its URL.

The need for such an identifier, then, may constitute an additional argument for establishing a central repository of metadata. Whether or not that repository is established on The LINGUIST List site, we are eager to support its development.