Planning Information Infrastructure through a New Library Research Partnership

Interim Report
July 2005

Janet McCue, Barbara Lust
Brian Lowe, Elaine Westbrooks, Jonathan Corson-Rikert, Frances Webb, Joy Paulson

---

Table of Contents

Executive Summary

The Cornell Language Acquisition Laboratory and Albert R. Mann Library are in the midst of developing an innovative collaboration between a research laboratory and an academic library to plan for the data preservation and discovery needs of the twenty-first century. Digital technology and internet communication now provide the opportunity to revolutionize the research process, through the ability to store, preserve, share, discover, and reanalyze vast amounts of data. While some disciplines, such as genomics or astronomy, have already developed sophisticated information technology infrastructure for these tasks, others are only beginning such work. In many, if not most research fields, it is especially difficult for the uninitiated to discover where data are located, what they describe, and how they may be used.

This project has begun to tackle these issues by taking advantage of the library's existing expertise in preservation, archiving, and metadata creation, building on the existing ontology-software tools the library has developed, and introducing a new conceptual framework that divides the tasks of data sharing into discrete levels that may be managed and presented in different ways not only for different audiences but respecting political divisions and control issues that will always be present throughout the laboratories and institutions of academia.

The work accomplished so far and the challenges uncovered in considering language acquisition data and conceptual modeling issues for the domain of linguistics will be highly informative as we continue our efforts to quantify the work involved, implement a concrete prototype, and generalize our model for data sharing across academic domains.

Introduction

Background

In this project we have begun to explore the possibility of extending the role of the university library in providing services related to research data. We are assessing the viability of this new role conceptually (i.e. in what ways does it make sense), structurally (i.e. where is the line between the library and the lab with regard to responsibility for research data), and financially (i.e. what are its costs now and what are the issues related to sustainability).

We are exploring this possibility through a unique collaboration between Cornell University's Albert R. Mann Library and the Cornell Language Acquisition Lab, which houses a large and unique amount of cross-linguistic language data and research materials in the interdisciplinary areas relevant to cognitive science, as well as other related areas such as neuroscience. At the same time, we use the University Library-CLAL collaboration as an exemplar of a more general model, which can be applied across different disciplines.

Under the direction of Professor Barbara C. Lust, the Cornell Language Acquisition Lab (CLAL) constitutes one of the world's leading centers for research on human language acquisition.[1] For over twenty years, interdisciplinary teams in the CLAL have collected data in more than twenty languages for the scientific study of language acquisition by children and adults.

The CLAL is a core component of the Virtual Center for the Study of Language Acquisition (VCLA),[2] which includes eight domestic and several foreign institutions poised to:

a. Share data
b. Share materials for best practices for scientific study of language acquisition
c. Collaborate on cross-linguistic research.

The academic library's position as expert in information management and dissemination and the Cornell library's position as a leader in digital resource development and management situate it as a natural potential collaborator with the CLAL and VCLA. Cornell's Albert R. Mann Library offers strengths in metadata, preservation, and archiving that make this library and other academic libraries natural partners in solving the data and metadata challenges of the magnitude and complexity facing the CLAL and VCLA.

Problem Statement

The field of language acquisition is central to developmental psychology, linguistics, and cognitive science. This interdisciplinary field requires collection of language data, and its scientific analysis (linguistics), in conjunction with scientific study of the learner (developmental psychology). Language acquisition research has much in common with other fields in its use of primary data (audiovisual recordings) which can and must be transcribed, analyzed, and interpreted in different ways by different researchers and at different points in time as theories and methods evolve. Such data, however, are accumulating from various sources without an infrastructure for their long-term storage and preservation, widespread access and dissemination, or collaborative research. A system to manage the resulting proliferation of metadata and facilitate easy access to both primary data and their related analyses should have wide application throughout research communities.

[1] Cornell Language Acquisition Lab, Homepage, http://www.clal.cornell.edu/index.php.
[2] Virtual Center for Language Acquisition, Homepage, http://www.clal.cornell.edu/vcla.

Project Summary

A. Project Goals

The project set out to do the following:

1) Develop a conceptual model for library-laboratory collaboration to address the problems of data preservation, discovery, repurposing, and reanalysis with an interest in interdisciplinary research
2) Prepare documentation of principles of data preservation, access, and dissemination including the creation of a functional prototype search tool
3) Develop the basis for inter-library and inter-lab relations to foster inter-institutional cooperation
4) Consult with industrial and academic support organizations with regard to outsourcing of data processing and storage; and
5) Assess and report both initial and ongoing costs required for the planned infrastructure and research data management.

B. Project Team

The team, led by principal investigators Janet McCue (Assistant University Librarian for Life Sciences and Director of Albert R. Mann Library) and Professor Barbara Lust, consists of a metadata librarian, a preservation librarian, two programmers, and an ontologist with domain knowledge in linguistics. Mann also provided the services of a sound technician to inventory CLAL's collection and to recommend a plan to reformat language data from audio cassette tapes.

C. Project Progress Report

The project team began work in August 2004 to further collaboration between the research lab and the library. Mann Library staff provided expert advice and consultation regarding metadata, organization of information, reformatting data, information technology, and preservation to the CLAL staff; the CLAL shared their extensive knowledge of the disciplinary framework and their experience with data transcription and analysis tools. Within this collaborative framework, the team accomplished the following:

- A clarified functional requirement for administrative metadata
- Advancement of the CLAL's metadata schemes to conform to metadata standards required for Open Archiving, including those set forth by the Linguistic Discipline's Open Language Archives Community (OLAC)[3]
- A detailed inventory of the CLAL's holdings
- A best practices document for audio reformatting
- Recommendations related to CLAL-specific data reformatting
- Initial estimates of the costs of reformatting language assets and storing them long-term
- A unique, hierarchical conceptual framework for linking specific research data ontologies to general University Library higher level ontologies (See Appendix)
- Development of a prototype metadata indexing tool to demonstrate this framework

[3]Open Language Archives Community, Homepage, http://www.language-archives.org/

The CLAL Research Lab meanwhile developed:

- Data Management – i.e., location and organization of physical holdings (e.g., more than 4,000 audio tapes and related written and electronic records)
- Data Identification and Description – i.e., developing a metadata system which integrates University Library upper level description with field and discipline specific data description required in linguistic research.
- Best Practices Manuals and Materials which incorporate these new management and description procedures.
- Software Development (e.g., a Data Transcription and Analysis tool which structures the researcher's data entry and integrates both domain specific and upper level metadata analyses.)
- Web development for interfacing Lab, Virtual Center and Library information
- Preliminary Audio-Video Digitizing Practices and examples for Data reformatting
- Server administration for structured related storage and dissemination of all materials and data under construction.

In addition, the project team engaged in the following collaborative activities:

- Consulted with Cornell's Department of Information Science faculty and staff (Carl Lagoze, William Arms, Claire Cardie et. al).
- Consulted with staff from the Cornell Laboratory of Ornithology.
- Met with MIT collaborators on two occasions (December 2004 and July 2005) The MIT collaborators involved MIT Librarian (Theresa Tobin) and MIT research lab faculty member (Prof. Suzanne Flynn). Discussions involved the foundations for inter-library exchange of research data and materials, as well as for generalizing methods developed in this project. Topics included the D-Space federation, new data grid development, and the necessity to calibrate metadata standards across both libraries and labs. These meetings initiated research lab-University Library collaboration which had not previously existed at MIT.
- Presented the poster, "Developing Adequate Documentation for Multi-faceted Cross Linguistic Language Acquisition Data" at the Conference on Language Documentation: Theory, Practice, and Values, LSA Linguistic Institute, MIT/Harvard, July 9 - 10, 2005.
- Presented the paper, "Searching Interoperability between Linguistic Coding and Ontologies for Language Description: Language Acquisition Data" at the E-MELD (Electronic Metastructure for Endangered Languages Data) workshop on digital language documentation, sponsored by NSF, July 2005, Harvard University.

Developments

To address the problem of data discovery and data preservation, the team developed a conceptual model that incorporated the following principles:

1. Local control over implementation of data storage and access coupled with sophisticated use of metadata to enable widespread discovery
2. A hierarchical ontology model to enable various audiences to easily search and browse for relevant information
3. A prototype search tool based on a successful ontology-driven system developed at Mann

4. Interfacing with existing ontologies to enhance searching, even when such ontologies may not be fully developed
5. Recommendations for digital rights management, privacy, and confidentiality

The Hierarchical Ontology Model

The hierarchical ontology model (Appendix) enables the creation of a single point of high-level discovery for disparate data from multiple disciplines, institutions, and laboratories. This requires the acknowledgement that disciplines do not necessarily organize or understand their data in mutually compatible frameworks, and that the ongoing modeling of new concepts is a vital aspect of research. The hierarchical model attempts to work within these restrictions by creating vertical bridges between levels, each of which could be implemented within a single repository or as distributed repositories providing metadata to the next higher level and access to component datasets. The interlocking triangles show how each node incorporates only the top, most general information contained in the nodes below.

A. Testing the Model – the Prototype Indexing Tool

Mann Library has created and partially populated an indexing tool based on the hierarchical ontology model described in the previous section. Our prototype uses the software framework created for the Cornell Virtual Life Science Library, VIVO.[4] VIVO indexes the people, events, publications, and research activities in the life sciences on the several campuses of Cornell University and provides a seamless navigational experience to users who may have no prior knowledge of the organizational structures involved. Patterned on the Protégé project at Stanford,[5] VIVO uses an underlying ontology structure to encode relationships among entries as well as detailed metadata about each entry itself. The full contents of the ontology-structured index are searchable, effectively combining the advantages of searching and browsing.

For this planning grant, we have extended the VIVO software framework to support the layering of successively more detailed information in related but independent portals. Our ongoing work will involve further adaptation of the VIVO codebase to permit importing and exporting of metadata through more standardized Semantic Web languages, such as RDF/OWL. Being able to incorporate external ontologies and easily forge general links between low-level markup, higher-level search categories, and administrative metadata will enhance the practicality of data sharing.

B. Management of Digital Rights

The project has involved the consideration of a number of concerns regarding digital rights confidentiality, and privacy. Our model allows for individual institutions to retain local control over these issues while still permitting data discovery. We have also formulated a number of specific recommendations for participating laboratories, and emphasize the need to thoroughly examine rights and privacy issues before engaging in preservation or archiving activities. Time and money should be invested into open resources or those which are permitted to be shared.

[4] Cornell Virtual Life Science Library, homepage, http://vivo.library.cornell.edu
[5] Protégé project at Stanford, homepage, http://protege.stanford.edu

Cost Framework

It is our intention to produce a set of statistics, in accord with the structure in Table 1. In these estimates we intend to separate those involved with general academic library infrastructure, (e.g., those that illustrate the costs associated with ontology building, metadata creation, development, and production, as well as digital preservation) from those that are discipline and lab specific and which arise from using the CLAL body of language acquisition data as a metric. For example, in the case of the CLAL, language data is inherently multi-media, involving both audio and visual technologies, and raises particular issues of portability.[6] In addition, much of the data in the lab is not digitized; different media have been used over time. Specific costs will vary from field to field and depend on the types of data and media involved. In addition, metadata and ontology developments will reflect the availability of existing ontologies in a specific discipline, (e.g. various developing ontologies in linguistics, such as GOLD).

Personnel			
	Position / Title	**%FTE Short-Term Costs**	**#years as Long-Term Costs**
1.	Domain Expert/Ontologist	50	TBD
2.	Metadata Specialist	20	TBD
3.	Preservationist	15	TBD
4.	Curator	15	TBD
5.	Programmer	20	TBD
6.	Server Administrator	15	TBD
Storage of Analog Assets			
1.	87 Linear Feet (4200 Hours)	**$260.00/month +**	TBD
Digitizing Analog Assets			
1.	4200 Hrs digitized at 44.1kHz / 24 bit	**$200,000.00-$400,000.00**	TBD
Storage of Digital Assets			
1.	2.0 TB	**$800.00/month +**	TBD

Table I: Cost structure (supplied values are CLAL examples)

We intend to investigate the ongoing costs of the following:

1) Middleware level required to create metadata and other data documentation and propagate it to higher levels
2) Data in digital and analog form
   - Long term digital data storage
   - Long term physical data storage, if applicable
3) Research lab maintenance and sustainability

[6] Bird, Steven and Simons, Gary. (2003) "Seven Dimensions of Portability for Language Documentation and Description," *Language, 79*. Available at: http://www.ldc.upenn.edu/sb/home/papers/0204020/0204020-revised.pdf. and Breen, M., G. Flam, I. Giannattasio, P. Holst, P. Pellizzari, D. Schuller (October 2003). Task Force to establish selection criteria of analogue and digital audio contents for transfer to data formats for preservation purposes. International ASsociation of Sound and Audiovisual Archives (IASA).2004. Available at: http://www.iasa-web.org/taskforce.pdf.

Growth Issues

The project has unearthed several growth issues. These include among others:

1) Division of budget between generic Library infrastructure costs and Domain and Data specific costs.
2) Division of budget to separate costs of "Long-Lived Digital Data Collections" from immediate costs of data archiving and preservation.
3) Division of contributions between University Library and Research Lab. Although the Mann Library and CLAL relationship has been extremely close and collaborative, this close degree of collaboration could not be assumed to be available or feasible across all research labs and fields. Interactive lab-library structures must be standardized and generalized.
4) General Issues of sustainability. For the particular type of data represented in the area of language we intend to consult the Laboratory of Ornithology on this issue.

Next Steps

Our intention now includes the following major areas: (i) We would like to expand our work to different types of research data from different fields/disciplines in order to evaluate the generalizability of the methods and infrastructures we have developed, and in order to provide more generalizable cost structures; (ii) We would like to refine the prototype we have developed for the current CLAL collaboration to support multi-level browsing and searching, and to instantiate examples of true inter-institutional data and materials exchange.

We intend to identify likely new sources of data to integrate into the model in order to provide a more concrete demonstration of a multilevel discovery structure and interdisciplinary linkages. By using these new sources of data, we also expect to provide examples of the costs involved in preserving, storing, and describing different types of research data. We also wish to outline plausible methods of dividing the financial and administrative burdens associated with these activities between research labs and the library.

Publications, Proposals and Presentations

Brian Lowe, Barbara Lust, Jonathan Corson-Rikert, Suzanne Flynn and Maria Blume, "Searching Interoperability between Linguistic Coding and Ontologies for Language Description: Language Acquisition Data" at the E-MELD (Electronic Metastructure for Endangered Languages Data), Harvard University, July 1-3, 2005. Available at: http://emeld.org/workshop/2005/papers/lust-paper.doc

Brian Lowe, Steven Pantle, and Elaine L. Westbrooks, (In Prep). Audio Archiving Best Practices.

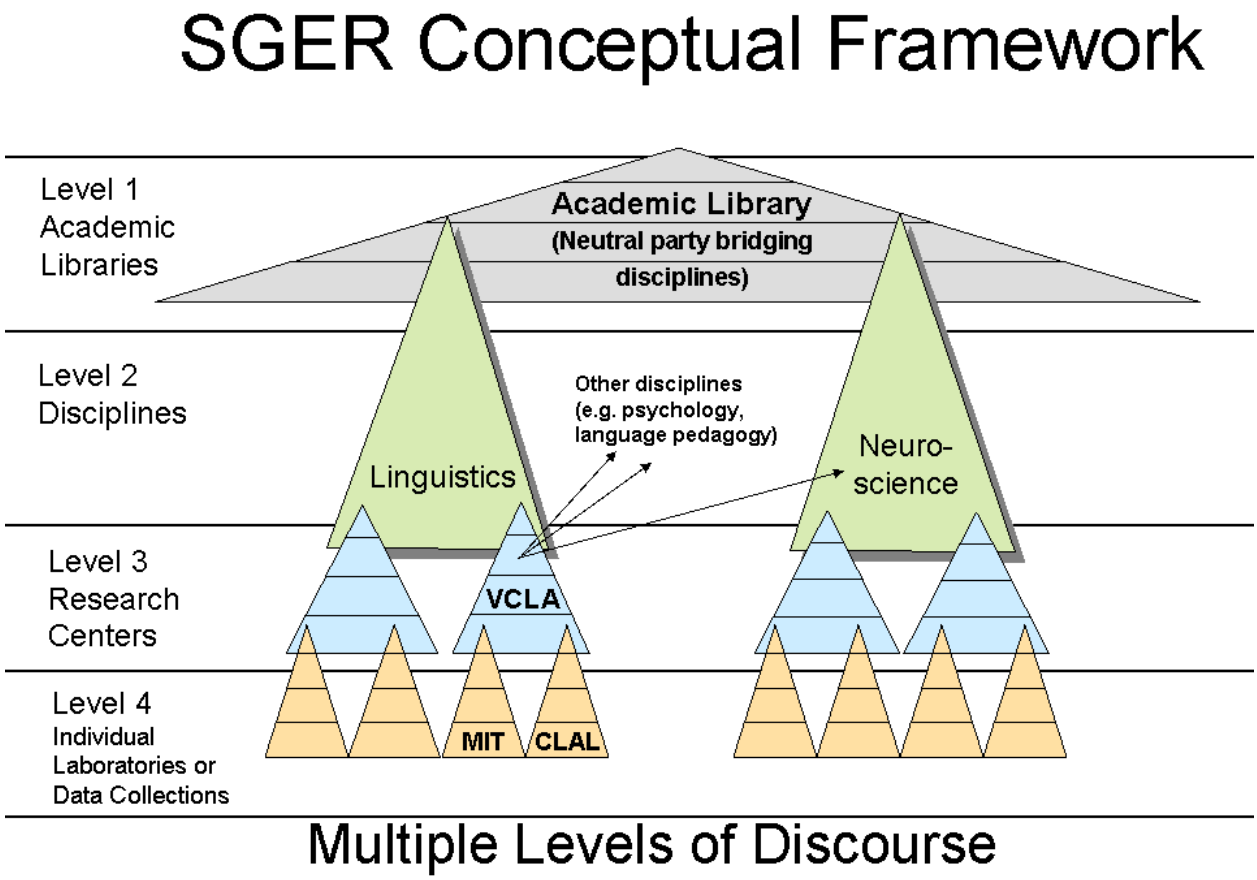
Barbara Lust. 2005. NSF CRI (Children's Research Initiative) Planning Grant: A Virtual center for Child Language Acquisition Research. Supplement Request: Integrating the Virtual Center for Child Language Acquisition Research in Cyberinfrastructure.

Barbara Lust, 2005. Transforming the Primary Research Process through Cybertool Development: A Model from Language Acquisition Data Management. Proposal submitted to NSF CI-Team program.

Joy Paulson, Barbara Lust, Elaine L. Westbrooks, 2004. Language Acquisition Digital audio Archiving: The South Asian Component. A proposal submitted to the National Endowment for the Humanities, Division of Preservation and Access.

Elaine L. Westbrooks, Barbara Lust, Suzanne Flynn, Theresa Tobin, 2005. "Developing Adequate Documentation for Multi-faceted Cross Linguistic Language Acquisition Data," Conference on Language Documentation: Theory, Practice, and Values, MIT/Harvard, July 9-10, 2005. http://www.lsadc.org/languagedocumentation/program.html

Appendix: Diagram of Hierarchical Model



Level one of the model represents the cross-discipline discovery tool that is the end result of all of these vertical bridges. It uses a general, non-discipline-specific ontology to provide information about types of research data, their languages, formats and availability. Interested searchers may follow links to more detailed information in the second level: the discipline node.