The Open Language Archives Community

Steven Bird*, Hans Uszkoreit†, and Gary Simons‡

*Linguistic Data Consortium, University of Pennsylvania
3615 Market Street, Suite 200, Philadelphia, PA 19104-2608, USA
sb@ldc.upenn.edu

†Deutsches Forschungszentrum für Künstliche Intelligenz
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, GERMANY
uszkoreit@dfki.de

‡SIL International 7500 W. Camp Wisdom Road, Dallas, TX 75236, USA
Gary_Simons@sil.org

Abstract

The goal of this symposium is to disseminate the OLAC vision to the language resources community, and to the European research community more broadly. We hope to encourage the community to archive and publish their resources using archival formats, and to document them using standard metadata. Presentations will address the following questions: What is the Open Language Archives Community? Why is language archiving important? What does it take to participate in OLAC? Discussion time will be used to clarify the OLAC model and to identify and address any concerns raised by the audience. Substantive feedback will help to guide the future evolution of OLAC. This symposium will mark the official launch of OLAC in Europe.

1. Introduction

Language technologists depend on a vast array of language resources, including texts, recordings, lexicons and annotations. The needs extend to software resources, such as protocols, interchange formats, data models and character encodings. Beyond data and tools, advice is another kind of language resource which exists alongside the data and tools. As resources proliferate we need a systematic way to discover them. This calls for language archives linked by community-specific metadata (i.e. catalog information) and a centralized union catalog.

OLAC, the Open Language Archives Community, is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources.

OLAC was founded at the NSF-funded workshop on Web-Based Language Documentation and Description, held in Philadelphia in December 2000 [http://www.ldc.upenn.edu/exploration/expl2000/]. The OLAC infrastructure is based on the Open Archives Initiative [www.openarchives.org] and the Dublin Core Metadata Initiative [dublincore.org].

The operation of OLAC is governed by three standards which participating archives agree to follow. During calendar year 2002 these standards are being frozen in candidate status so that new participants can implement a data provider without worrying about the foundations changing beneath them. In 2003 the standards will be revised by the community and formally adopted.

1. **OLAC Process**: summarizes the governing ideas of OLAC (i.e. the purpose, vision, and core values) and then describes how OLAC is organized and how it operates.
2. **OLAC Protocol for Metadata Harvesting**: defines the protocol OLAC service providers use to harvest metadata from OLAC data providers. It defines the responses that OLAC data providers must make to the requests of the protocol.
3. **OLAC Metadata Set**: specifies the metadata set used by OLAC for the interchange of metadata within the framework of the Open Archives Initiative.

These candidate standards are posted on the OLAC website [www.language-archives.org]. Also on this site is the list of 20 participating archives and a link to the cross-archive searching service hosted by the Linguist List.

2. Program: Wednesday 29 May Room Atlantico 14:40-16:40

The symposium will consist of an opening statement, six presentations, open discussion, and a closing statement.

14:40 Opening Statement
Antonio Zampolli

14:45 The Seven Pillars of Open Language Archiving
Gary Simons

15:00 OLAC, EMELD and "Us"
Helen Aristar-Dry and Anthony Aristar

15:20 Ontologies for Language Resource Description
Hans Uszkoreit

15:35 Opening the Archives: OLAC, TRACTOR and the OTA
Martin Wynne

15:50 Experience with OLAC for the ATILF archives
Laurent Romary and Zina Tucsnak

16:05 Getting involved in OLAC
Steven Bird

16:15 General Discussion

16:25 Closing Statement
Nicholas Ostler

3. Short Abstracts

The Seven Pillars of Open Language Archiving

Gary Simons
SIL International

Digital archiving of language documentation and description on the World-Wide Web holds the promise of unparalleled access to language information. But if it is not done well, it also offers the specter of frustration and chaos on an unparalleled scale. This talk presents an executive summary of our vision for the kind of infrastructure that could unlock the promise. Special focus is given to the seven pillars on which the OLAC infrastructure has been erected: DATA, TOOLS, ADVICE, GATEWAY, METADATA, REVIEW, and STANDARDS.

OLAC, EMELD and "Us"

Helen Aristar-Dry and Anthony Aristar
Linguist List, Eastern Michigan University, and Wayne State University.

Over the past 11 years, the Linguist List has become the primary source of information for the linguistics community, reaching out to 15,500 subscribers worldwide, and having four complete mirror sites. The Linguist List will be augmenting its service by providing the primary entry point for OLAC, and permitting linguists to browse distributed language resources at a single place. This talk will include a demonstration of a new Linguist List "service provider", and also report progress on a new NSF-sponsored project to create a "Showroom of Best Practice" for language documentation.

Ontologies for Language Resource Description

Hans Uszkoreit
Deutsches Forschungszentrum für Künstliche Intelligenz

The success of global digital archiving initiatives depends on usable and widely accepted schemes for describing informational resources. In order to arrive at useful metadata sets one has to combine accepted generic resource ontologies with transparent specialized ontologies for one or more relevant subjects areas. A useful ontology for the transdisciplinary area of language technology has to unify ontologies of language research and resources with ontologies of several areas of information technology. We will demonstrate how such an ontology can be developed and employed for creating interfaces between the OLAC metadata set and other resources such as the portal LT-World, the ACL/DFKI Software Registry, and the Survey of the State of the Art in Language Technology.

Opening the Archives: OLAC, TRACTOR and the OTA

Martin Wynne
Oxford Text Archive

This talk will reflect on the experience of delivering resource descriptions from two archives: the TELRI Research Archive of Computational Tools and Resources (TRACTOR) and the Oxford Text Archive. The process of migration and harvesting of the metadata records from both archives is examined. The merits and drawbacks of the OLAC metadata set, and its suitability to multilingual language tools and resources are appraised, and some alternatives are considered. Some preliminary thoughts are offered on the experience of participating in OLAC with two significant archives, which have vastly different holdings, archiving and distribution policies and metadata standards.

Experience with OLAC for the ATILF archives

Laurent Romary and Zina Tucsnak
Analyse et Traitements Informatisés de la Langue Française

We will provide some feedback concerning the experience of implementing OLAC in the context of ATILF, an institution which acts as a central point in France for online delivery of textual and lexical data. We will show that, despite the simplicity of deployment offered by the OLAC principles and format, the main difficulty of implementing OLAC in a large institution like ATILF is to provide coherent metadata covering a wide and heterogeneous set of information sources, both from a technical and a conceptual point of view. This requires that we implement various methods of recovery from those information sources.

Getting involved in OLAC

Steven Bird
University of Pennsylvania

This talk will describe the OLAC "starter kit", a low entry-cost method for resource creators to document their work. Various routes for exporting existing archive catalogs to OLAC format will be described. The talk will conclude with a call for widespread participation in OLAC.