COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE/if not in response to a program announcement/solicitation enter NSF 04-23

NSF 04-23 **01/15/07**

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

BCS - LINGUISTICS

FOR NSF USE ONLY

NSF PROPOSAL NUMBER

0723357

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
01/26/2007	**18**	**04040000 BCS**	**1311**	**042250712**	**01/29/2007 10:00am**

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)

231352685

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE

University of Pennsylvania

AWARDEE ORGANIZATION CODE (IF KNOWN)

0033787000

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

University of Pennsylvania
Research Services
Philadelphia, PA. 191046205

NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE

PERFORMING ORGANIZATION CODE (IF KNOWN)

ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)
☐ SMALL BUSINESS ☐ MINORITY BUSINESS
☐ FOR-PROFIT ORGANIZATION ☐ WOMAN-OWNED BUSINESS

☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **Collaborative Research: OLAC: Accessing the World's Language Resources**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ **187,264**	**36** months	**07/01/07**	

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

☐ BEGINNING INVESTIGATOR (GPG I.A)
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C)
☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.B, II.C.1.d)
☐ HISTORIC PLACES (GPG II.C.2.j)
☐ SMALL GRANT FOR EXPLOR. RESEARCH (SGER) (GPG II.D.1)
☐ VERTEBRATE ANIMALS (GPG II.D.5) IACUC App. Date _____
☐ HUMAN SUBJECTS (GPG II.D.6)
Exemption Subsection _____ or IRB App. Date _____
☐ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)
_____
☐ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)

PI/PD DEPARTMENT
Department of Linguistics

PI/PD FAX NUMBER
215-573-2175

PI/PD POSTAL ADDRESS
3615 Market Street

Philadelphia, PA 19104
United States

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address
PI/PD NAME **Mark Liberman**	**Ph.D.**	**1975**	**215-573-5490**	**myl@unagi.cis.upenn.edu**
CO-PI/PD **Steven G Bird**	**PhD**	**1991**	**215-898-0464**	**sb@ldc.upenn.edu**
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				

Electronic Signature

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE/if not in response to a program announcement/solicitation enter NSF 04-23

NSF 04-23 **01/15/07**

FOR NSF USE ONLY

NSF PROPOSAL NUMBER

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

BCS - LINGUISTICS

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)

231352685

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE

Graduate Institute of Applied Linguistics

AWARDEE ORGANIZATION CODE (IF KNOWN)

6250012801

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

7500 W Camp Wisdom Road
Dallas, TX 75236-5629

NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE

PERFORMING ORGANIZATION CODE (IF KNOWN)

ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)
☐ SMALL BUSINESS ☐ MINORITY BUSINESS
☐ FOR-PROFIT ORGANIZATION ☐ WOMAN-OWNED BUSINESS

☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **Collaborative Research: OLAC: Accessing the World's Language Resources**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ **107,200**	**36** months	**07/01/07**	

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

☐ BEGINNING INVESTIGATOR (GPG I.A)
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C)
☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.B, II.C.1.d)
☐ HISTORIC PLACES (GPG II.C.2.j)
☐ SMALL GRANT FOR EXPLOR. RESEARCH (SGER) (GPG II.D.1)
☐ VERTEBRATE ANIMALS (GPG II.D.5) IACUC App. Date _____
☐ HUMAN SUBJECTS (GPG II.D.6)
Exemption Subsection _____ or IRB App. Date _____
☐ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.j)
_____
☐ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.G.1)

PI/PD DEPARTMENT
Language Development

PI/PD FAX NUMBER
972-708-7546

PI/PD POSTAL ADDRESS
7500 W Camp Wisdom Road

Dallas, TX 752365629
United States

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address
PI/PD NAME **Gary F Simons**	**PhD**	**1979**	**972-708-7487**	**gary_simons@sil.org**
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				

Project Summary

OLAC: Accessing the World's Language Resources

Intellectual Merit

Language resources are the bread and butter of language documentation and linguistic investigation. They include the primary objects of study such as texts and recordings, the outputs of research such as dictionaries and grammars, and the enabling technologies such as software tools and interchange standards. Increasingly, these resources are maintained and distributed in digital form. Although language resources are beginning to abound on the web, they are often difficult or impossible for interested parties to find and use. Searching on the web for language resources in many languages is a hit-and-miss affair for three reasons: (i) resources are housed in archives that have never put their catalog online, (ii) resources are exposed to online search engines but inadequately described so that searches do not retrieve desired results with precision, or (iii) resources are exposed online but are hidden behind form-based interfaces such that search engines cannot find them. The Open Language Archives Community (OLAC) is addressing these problems by providing a standard set of language resource descriptors and a portal that permits users to query dozens of language archives simultaneously using a single search. However, the current coverage of OLAC is only the tip of the iceberg. New research is needed in order to tap the wealth of new digital library services and web-mining technologies, and to make the discovered language resources maximally accessible to linguists.

The aim of the proposed project is to greatly improve access to language resources for linguists and the broader communities of interest, by achieving an order-of-magnitude increase in the coverage of the OLAC catalog and in the use of OLAC search services. The project seeks to do so through two main areas of activity:

Access to Language Resources in Archives: Develop guidelines and services that encourage best common practices among language archives that will facilitate language resource discovery with precision through OLAC.

Access to Language Resources on the Web: Develop services to bridge the resource catalogs of the repository, library, and web domains, to facilitate language resource discovery with precision through OLAC.

Broader Impacts

The proposed research should have a broad impact across the field of linguistics by developing an online service that gives linguists access to resources for the thousands of languages in the world. But the impact will extend well beyond the linguistics community. Access to these language resources will assist technologists who are endeavoring to make information technologies work with every language, not just a select few. It will also permit educators, students and members of society at large to access a wealth of materials that demonstrate the full range of linguistic diversity in the world. Yet another audience for access to language resources are the actual speakers of all the world's languages. In the case of endangered languages, access to language resources is a critical asset in the process of language revitalization. The project will also serve to advocate the widespread use of ISO 639-3 codes to precisely identify the 7,500 known human languages, past and present. This will encourage reform in the practice of library and archive cataloging, which currently recognizes fewer than 400 languages, and will begin the process of helping the major storehouses of knowledge around the world to appropriately deal with linguistic diversity.

0723357

OLAC: Accessing the World's Language Resources

1 Background and Results from Prior Support

Today, language technology and the linguistic sciences are confronted with a vast array of *language resources*. These are predominantly resources in or about a human language—texts, recordings, dictionaries, annotations, grammars, and the like. Also included are resources used by language technologists—software, protocols, data models, file formats, newsgroups, web indexes, and the like. The resources are growing in number, in size, in diversity. Multiple *communities* depend on language resources, including linguists, engineers, educators, students, and actual speakers. Language resources are particularly important to language communities who want to revitalize a language that is in danger of extinction. The number of individuals having an immediate, professional interest in language resources is in the tens of thousands. The total set of potential users of language resources must number in the millions.

Today, we have unprecedented opportunities to *connect* these communities to the language resources they need. First, the world wide web is making it possible, like never before, to bring all the members of the language resource community together into a global virtual community. Second, digital publication—both on and off the web—is the most practical and efficient means of sharing language resources. Open encoding standards like the Extensible Markup Language (XML) and the Universal Character Set (Unicode) provide flexible ways to represent structured data and ensure its long-term survival. Finally, the Open Language Archives Community (OLAC) provides a standard set of language resource descriptors (OLAC Metadata), and a portal that permits users to query dozens of language archives simultaneously using a single search. OLAC has the bottom-up, distributed character of the web, while simultaneously having the efficient, structured nature of a centralized database. This combination is well-suited to the language resource community.

OLAC was founded in 2000 with the goals of (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. For the first goal, substantial progress on identifying best practices has now been achieved under the NSF E-MELD project; consideration of best practices is now incorporated into the funding of language documentation projects (DEL, Bird and Simons (2003b)). For the second goal, 35 repositories holding some 30,000 metadata records are participating in OLAC; linguists can search all repositories simultaneously, and the interest level now stands at 100,000 queries per month. Participating archives include: Alaska Native Language Center Archive, Archive of the Indigenous Languages of Latin America, UC Berkeley Audio Archive of Linguistic Fieldwork, Oxford Text Archive, Rosetta Project All Language Archive, SIL Language and Culture Archives, and 29 more.

The results to-date of OLAC demonstrate a high level of participation and interest within the language resources community, but we recognize that the current coverage is only the tip of the iceberg. Google (with billions of items indexed) is still the first search engine of choice for most linguists since it has so much greater coverage. However, using a conventional search engine to find a particular kind of resource for a particular language is not likely to succeed, especially for

an uncommon language. One of three outcomes is more likely: (1) Relevant resources do exist as web pages, but their visibility is *occluded* by fuzziness of search terms and descriptors on the page. (2) Relevant resources are accessible from the web, but they are completely *hidden* from the view of the search engine because they are on the other side of form-based interfaces. (3) Relevant resources are *missing* altogether because the archives that hold them have not published their catalogs on the web.

The aim of the proposed project is to dramatically improve the language resource community's access to resources by achieving an order-of-magnitude increase in the coverage of the OLAC catalog and in the use of OLAC search services. The project seeks to do so through two main areas of activity:

Objective 1: Access to Language Resources in Archives: Develop guidelines and services that encourage best common practices among language archives that will facilitate language resource discovery with precision through OLAC.

Objective 2: Access to Language Resources on the Web: Develop services to bridge the resource catalogs of the repository, library, and web domains, to facilitate language resource discovery with precision through OLAC.

Objective 1 addresses the problem of occluded and missing resources, while objective 2 addresses hidden resources. In the longer term, it is hoped that an OLAC search service with vastly increased coverage will become such a popular tool among linguists that it will provide powerful incentive for them to adopt best practices in digital archiving and deposit their own work in language archives that will make it accessible to others through the same means.

Prior Support: The proposed project represents an outgrowth of prior NSF support to the investigators in the following projects: *Multidimensional Exploration of Linguistic Databases* (#9983258, 02/01/2000–01/31/2003); *ISLE: International Standards for Language Engineering* (#9910603, 03/01/2000–07/31/2003); *TalkBank: A Multimedia Database of Communicative Interactions* (#9978056, 10/01/1999–09/30/2005); *E-MELD: Electronic Metastructure for Endangered Languages Data* (#0094934, 07/01/2001–06/30/2006); *The Rosetta Project: ALL Language Archive* (#0333727, 11/01/2003–02/28/2006); *Querying Linguistic Databases* (#0317826, 08/01/2003–07/31/2007). Key results from these projects were as follows:

(a) OLAC, a community-wide resource discovery framework (Bird and Simons, 2000a,b, 2001, 2002; Simons, 2002; Bird and Simons, 2003a; Simons and Bird, 2003a; Simons, 2003; Simons and Bird, 2003b; Bird and Simons, 2004);

(b) data models for linguistic annotations of text and speech (Bird and Liberman, 2001; Maeda and Bird, 2000; Graff and Bird, 2000; Cotton and Bird, 2002; Cieri and Bird, 2001; Bird et al., 2000b; Bird and Harrington, 2001)

(c) open source linguistic annotation software (Bird et al., 2001b; Maeda et al., 2002; Bird et al., 2002; Ma et al., 2002);

(d) recommendations for digital storage of language data (Bird and Simons, 2003b; Simons, 2006);

(e) methods of interoperating over disparate language resources (Simons et al., 2004a,b);

(f) CD-ROM publications of linguistic field data: (Bird and Bell, 2001; Bird, 2003b,a); and

(g) models and tools for querying linguistic databases (Bird et al., 2001a, 2000a; Lai and Bird, 2004; Bird et al., 2005, 2006)

1.1 Access to Language Resources in Archives

Members of the language resources community have a common need for discovering language resources with high precision and recall. To support this, the community needs to agree on some specialized vocabularies in the metadata for describing resources. Over the past five years, OLAC participants have adopted five such vocabularies: *subject language*, for identifying precisely which language(s) a resource is "about"; *linguistic type*, for classifying the structure of a resource as primary text, lexicon, or language description; *linguistic field*, for specifying relevant subfields of linguistics; *discourse type*, for indicating the linguistic genre of the material; and *role*, for documenting the parts played by specific individuals and institutions in creating a resource. These vocabularies are represented in the OLAC Metadata format, which is an extension of Dublin Core Metadata (Bird and Simons, 2004)—the dominant metadata standard in the digital library and world wide web communities. Participating language archives publish their catalogs in an XML format, and these records are "harvested" twice a day by OLAC services using the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (Simons and Bird, 2003a)—another standard of the digital library community.

OLAC has been a developer or early adopter of several other key digital library technologies in service to the language resources community: (i) a simplified method for archives to publish their metadata in "static repositories", thus lowering the barrier to entry—a method now adopted by the OAI as a service to the wider digital archives community; (ii) a Google-like search interface for searching participating archives with specialised support for language identification (Hughes and Kamat, 2005; Hughes, 2006); (iii) a score card system which publishes a metadata quality score for each participating archive to serve as a form of peer review and an incentive for archives to improve the quality of their metadata; (iv) a metadata usage report showing how individual metadata elements and values have been put to use by participating archives; (v) an "OAI crosswalk" permitting users of digital library services outside the language resources community to access all OLAC metadata; and (vi) a web crawler gateway permitting users of search engines such as Google to access all OLAC metadata.

Beyond these technological achievements, OLAC has succeeded in establishing a *community* that operates by means of an open process (Simons and Bird, 2002). In a 2004 publication of the Digital Library Federation, OLAC was singled out for this facet of its work:

> OLAC is exemplary in several ways: the technical and social infrastructure that it has developed to support its community of contributors, based on shared principles and standards; the resources that it provides at its Web site about its purpose, scope, history, tools, news and events; and the efforts of its two leaders ... to articulate the challenges, analyze the options, and recommend possible solutions to their community of contributors in order to improve OLAC (Brogan, 2004).

Despite these successes, OLAC has three significant shortcomings. First, the quality of metadata in the majority of participating archives remains low. According to the OLAC report card system, the average metadata-quality score for an OLAC record is 6.2 out of 10. However, this score is inflated by four of the larger archives that are already following best practice; the average score for the remaining 31 archives is only 3.5 out of 10. Another way in which quality of current metadata suffers is that it is not being kept up-to-date; most archives have not updated their metadata to reflect new acquisitions since joining OLAC. Second, many language archives are not yet participating in OLAC. For example, the American Philosophical Association (Philadelphia) and the National Anthropological Archives (Washington DC) are major national centers having substantial collections of early language documentation for native American languages. The Oriental Institute (Chicago) and the School of Oriental and African Studies (London) have vast collections covering Asian and African languages. Similarly, many smaller archives are yet to participate in OLAC, e.g. the Memorial University of Newfoundland Folklore and Language Archive (St John's); the University of Wisconsin Himalayan Linguistics Archive (Milwaukee); the College of William and Mary Creek Language Archive (Williamsburg). Many linguistics departments have archives holding language documentation collected over many decades by their staff, e.g. UCLA, UC Santa Barbara, University of Chicago. Third, many of the participating "archives" are more accurately described as digitization projects. Once project funding and institutional backing cease, the materials may disappear. Thus, these initiatives are not yet following best practices in digital archiving, e.g. as defined by the Open Archival Information Systems (OAIS) reference model (CCSDS, 2002) which has attained the status of an ISO standard (ISO 14721:2003) and is accepted as best practice within the digital library community.

In order to address these three shortcomings, we believe that the language archives community needs to develop in three ways:

(1) All OLAC repositories should have up-to-date catalogs that contain metadata conforming to best practice.

(2) All major language archives should be participating in OLAC.

(3) All OLAC repositories should conform to current best practices for the long-term curation of their holdings.

1.2 Access to Language Resources on the Web

Language resources of all kinds have been posted on the web and can be discovered with a standard search engine: interlinear texts, dictionaries, linguistic descriptions, and linguistic research papers. Even ordinary web pages may count as primary language documentation when they are in any of the thousands of "low-density" languages, languages with a small presence on the web. Another source of language resources is traditional libraries with their collections of printed dictionaries, grammars, and linguistic monographs. Library automation has resulted in standards like MARC for digital cataloging (US Library of Congress, 2000) and the Z39.50 protocol for searching digital catalogs (NISO, 2003). Today such catalogs typically have a web interface, though their entries

are usually not indexed by search engines. WorldCat[1] is a single service that provides a web-based search engine over the world's major libraries. A more recent trend in the library automation arena has been the development of e-print repositories, often in connection with an academic institution. These use simpler standards, Dublin Core for metadata cataloging and the OAI protocol for metadata harvesting. OAIster is a single service that provides a web-based search engine over all known e-print repositories.

The task of finding language resources on the web is beset with problems. The most obvious one is *scale*: that is, searching for low-density language resources among the billions of pages indexed by Google is like searching for a needle in a haystack. Some linguistic web-mining projects are addressing this problem by probing search engines with linguistic query terms and archiving the found objects (Langendoen et al., 2002; Lewis, 2003; Baldwin et al., 2006). Another problem is *volatility*: for instance, a four-year longitudinal study has shown the half-life of a web page to be approximately two years (Koehler, 2002). Yet another problem is that of the *hidden web*: namely, the content that lies behind search interfaces and is thus opaque to conventional search engines. Within the library automation community, JZKit[2] is an open-source project that addresses this by building an aggregated index of library catalogs that are accessible through about a dozen search interfaces (including the Z39.50 and SRW/SRU[3] standards).

In order for the language resources community to realize the promise of universal access to relevant resources, three major problems still need to be addressed. First, due to the huge scale of the search space and the absence of precise indexing vocabularies, users of web search engines such as Google typically experience low precision and recall when searching for language resources. Searches for scarce resources are often swamped with irrelevant results (low precision). For example searches for "Santa Cruz" will not yield results for the East Papuan language of this name, spoken in the Solomon Islands. Furthermore, many resources are just not returned at all because search terms do not match the synonymous terms used in the desired documents (low recall). For example, searching for "Dschang lexicon" will not return the Dschang lexicon developed by one of the PIs (Bird and Tadadjeu, 1997); it was published as "Lexique Yémba," since the language of wider communication is French, and the language autonym for *Dschang* is *Yémba*. Such materials are straightforward to find, however, using linguistically-aware web mining programs that probe Google's index by querying for all synonyms and translations of linguistic terminology and by searching for pages that contain IPA characters. A harder problem is to reliably identify the subject language and the linguistic type of the found resource (Hughes et al., 2006).

A second problem is that the library automation solutions are part of the hidden web and remain hidden to the language resources community at large. Users searching for language resources also need to visit other services like WorldCat and OAIster;[4] it would be better for the language-resource content of these services to be fully integrated with OLAC. For example, OAI e-print repositories include many articles relevant to languages and linguistics,[5] for which results are not

[1] `http://www.oclc.org/worldcat/`
[2] `http://developer.k-int.com/jzkit2/`
[3] `http://www.loc.gov/standards/sru/`
[4] `http://oaister.umdl.umich.edu/`
[5] e.g. `http://romeo.eprints.org/search.php?t=Linguistics`

currently returned in OLAC searches. Similarly, searching the world's libraries could be done with a Google-mediated search of the WorldCat site, but this would not permit systematic aggregation and integration, and would still suffer from the problems of low recall and precision. A third shortcoming is that users who try to find language resources using any of these non-OLAC services are unlikely to discover that OLAC can provide additional value, such as richer metadata and more focused result sets.

In order to address these three shortcomings, we believe that the language archives community needs to develop in three ways:

(1) All repository and library holdings relevant to language documentation should be indexed in OLAC, by crosswalking and enriching existing catalog records.

(2) Low-density language materials identified by linguistic web mining should be reliably categorized with OLAC vocabularies.

(3) Web search engines should index all OLAC records, so that users who discover language resources using a conventional web search quickly find OLAC records and are drawn to the OLAC site for more precise searching.

2 Methods and Workplan

The discussion of methods and workplan is organized according to the two major objectives listed in the introduction. Each objective is in turn elucidated in terms of three outcomes (as listed at the conclusion of each of the preceding subsections). The basic plan is that outcomes 1.1 and 2.1 will be pursued during year 1; outcomes 1.2 and 2.2 during year 2; and outcomes 1.3 and 2.3 during year 3. The following discussion fleshes out the six outcomes in terms of the tasks that must be performed in order to achieve them. The allocation of activities to project personnel is discussed in the budget justification.

Objective 1: Access to Language Resources in Archives

Develop guidelines and services that encourage best common practices among language archives that will facilitate language resource discovery with precision through OLAC.

Outcome 1.1: All OLAC repositories should have up-to-date catalogs that contain metadata conforming to best practice.

1.1a Best practice document: Prepare and adopt an OLAC recommendation on metadata best practice, adapting the DC Usage Guidelines and fleshing out published recommendations (Bird and Simons, 2003b; Simons, 2006).

1.1b Score cards: Update the automated score card system that reports on the quality of metadata records to conform to the adopted recommendations.

1.1c *Metadata quality:* Identify low-scoring archives and work with them to improve the quality of their metadata, particularly the use of OLAC vocabularies. Develop an OLAC metadata usage note based on this experience.

1.1d *Updating archives:* Identify OLAC archives whose metadata records have gone stale, and work with them to bring their catalogs up-to-date and to set up automatic processes to expose their live catalogs.

1.1e *Quarterly reporting:* Develop an automated quarterly report to be emailed to curators to remind them about the quality and currency of their metadata, and to inform them of statistics concerning queries on their archives.

1.1f *Metrics:* Develop metrics for monitoring the size, coverage, and use of the complete OLAC metadata catalog, and implement such reporting on the web site.

Outcome 1.2: All major language archives should be participating in OLAC.

1.2a *Non-participating archives:* Compile a list of all known non-participating language archives.

1.2b *Collaboration:* Contact all identified archives and consult with them to develop best strategy for generating their metadata.

1.2c *Scraping:* Do automatic scraping of the HTML metadata already published on archive sites and perform a quick low-accuracy conversion to OLAC metadata in order to show archivists their search results coming up in user queries; use this to show archivists the value of mapping their terms to OLAC terms and publishing OLAC metadata in order to improve precision and recall.

1.2d *OLAC export:* Work with these archives to expose their catalog in OLAC format.

1.2e *Repository editor:* Configure an open source XML editor (like XMLmind) to create a static repository editor suitable for use by small archives.

Outcome 1.3: All OLAC repositories should conform to current best practices for the long-term curation of their holdings.

1.3a *Recommendation:* Prepare and adopt an OLAC recommendation document on best practices for the long-term curation of language archive holdings.

1.3b *Archive categories:* Refine the `<olac-archive>` description to include more fine-grained categorization of participating archives. Categories should be based on published criteria (in the OAIS reference model) and assignment of categories should be vetted by OLAC.

1.3c *Score cards:* Incorporate these criteria into the automated score cards. Generalize the score card mechanism to facilitate the addition of new measures over time (as best practices evolve).

Objective 2: Access to Language Resources on the Web

Develop services to bridge the resource catalogs of the repository, library, and web domains (OAI, MARC, Google) to facilitate language resource discovery with precision through OLAC.

Outcome 2.1: All repository and library holdings relevant to language documentation should be indexed in OLAC, by crosswalking and enriching existing catalog records.

2.1a Language identification: Develop an automated procedure for finding language identification in a catalog record and translating it to an appropriate ISO 639-3 code.

2.1b Data type identification: Develop an automated procedure for finding linguistic data type information in a catalog record and translating it to an appropriate OLAC linguistic data type code.

2.1c OAI Crosswalk: Develop an OLAC data provider that harvests from the leading OAI aggregator and serves up all records that are determined to be language resources with the metadata enriched for precise identification of language and linguistic data type (primary text vs lexicon vs linguistic description).

2.1d Z39.50 Crosswalk: Develop an OLAC data provider that harvests from one or more Z39.50 gateways (e.g. using JZKit), eliminates duplicates, applies the MARC to DC crosswalk[6], and serves up all records that are determined to be language resources with the metadata enriched for precise identification of language and linguistic data type.

Outcome 2.2: Low density language materials identified by linguistic web mining should be reliably categorized with OLAC vocabularies.

2.2a Language identification: Develop an automated procedure for finding low-density language identification from an arbitrary web document, (e.g. using character encodings, keywords, n-gram signatures, anchor text, and place-name references) and translating this to an appropriate ISO 639-3 code.

2.2b Data type identification: Develop an automated procedure for identifying the linguistic data type of an arbitrary web document (e.g. using Bayesian classifiers trained on existing metadata records) and translating this to an appropriate OLAC linguistic data type code.

2.2c Web mining: Generate OLAC metadata records for low-density language resources discovered by linguistic web-mining projects, such as Hughes' collection of 600,000 low-density language URLs.

Outcome 2.3: Web search engines should index all OLAC records, so that users who discover language resources using a web search quickly find OLAC records and are drawn to the OLAC site for more precise searching.

[6] `http://www.loc.gov/marc/marc2dc.html`

2.3a Static HTML: Generate a static HTML page for each OLAC metadata record, by adding a new module to the OLAC harvester.

2.3b Synonyms: Add synonyms for all identified linguistic terminology to static HTML pages, so that these pages are more likely to be found in conventional web searches.

2.3c Embedded queries: Enrich static pages with OLAC query links, so that users who discover OLAC records via a Google search result are encouraged to remain on the OLAC site for more precise searching.

3 Dissemination Plans

We will present the results of our research at major international conferences in linguistics, computational linguistics, and digital libraries, and submit extended discussions of the research to leading journals. The results of our research will also be deployed on the OLAC web site as they come into existence. In this way the results will be put into action immediately by the institutions who are contributing their resources to the OLAC catalog and by individuals who are using the OLAC search services to find them. All tools and data created by the project will be disseminated with open source and open content licenses. We will make occasional use of LanguageLog, when appropriate, to disseminate information of broader interest that is discovered using OLAC search services.

References

Baldwin, T., Bird, S., and Hughes, B. (2006). Collecting low-density language materials on the web. In *Proceedings of the 12th Australasian Web Conference*. Southern Cross University.

Bird, S., editor (2003a). *Grassfields Bantu Fieldwork: Dschang Tone Paradigms*. Linguistic Data Consortium. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2003S02`, ISBN 1-58563-254-6.

Bird, S., editor (2003b). *Grassfields Bantu Fieldwork: Ngomba Tone Paradigms*. Linguistic Data Consortium. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001S16`, ISBN 1-58563-216-3.

Bird, S. and Bell, J., editors (2001). *Grassfields Bantu Fieldwork: Ngomba Tone Paradigms*. Linguistic Data Consortium. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2001S16`, ISBN 1-58563-216-3.

Bird, S., Buneman, P., and Liberman, M., editors (2001a). *Proceedings of the IRCS Workshop on Linguistic Databases*. `http://www.ldc.upenn.edu/annotation/database/`.

Bird, S., Buneman, P., and Tan, W.-C. (2000a). Towards a query language for annotation graphs. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 807–814. Paris: ELRA. `http://arXiv.org/abs/cs/0007023`.

Bird, S., Chen, Y., Davidson, S. B., Lee, H., and Zheng, Y. (2005). Extending XPath to support linguistic queries. In *Programming Language Technologies for XML (PLANX)*, pages 35–46. `http://eprints.unimelb.edu.au/archive/00001451/`.

Bird, S., Chen, Y., Davidson, S. B., Lee, H., and Zheng, Y. (2006). Designing and evaluating an XPath dialect for linguistic queries. In *22nd International Conference on Data Engineering*, pages 52–61. `http://eprints.unimelb.edu.au/archive/00001455/`.

Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C., and Liberman, M. (2000b). ATLAS: A flexible and extensible architecture for linguistic annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. Paris: ELRA. `http://arXiv.org/abs/cs/0007022`.

Bird, S. and Harrington, J., editors (2001). *Speech Communication: Special Issue on Speech Annotation and Corpus Tools*, volume 33 (1–2). Elsevier.

Bird, S. and Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Communication*, 33:23–60. `http://arxiv.org/abs/cs/0010033`.

Bird, S., Maeda, K., Ma, X., and Lee, H. (2001b). Annotation tools based on the annotation graph api. In *Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*.

Bird, S., Maeda, K., Ma, X., Lee, H., Randall, B., and Zayat, S. (2002). TableTrans, MultiTrans, InterTrans and TreeTrans: Diverse tools built on the annotation graph toolkit. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 364–370. Paris: ELRA. http://arXiv.org/abs/cs/0204006.

Bird, S. and Simons, G., editors (2000a). *Proceedings of the Workshop on Web-Based Language Documentation and Description*. http://www.ldc.upenn.edu/exploration/expl2000/.

Bird, S. and Simons, G. (2000b). A survey of the state of the art in digital language documentation and description. http://www.language-archives.org/docs/survey.html.

Bird, S. and Simons, G. (2001). The OLAC metadata set and controlled vocabularies. In *Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*. http://arXiv.org/abs/cs/0105030.

Bird, S. and Simons, G., editors (2002). *Proceedings of the IRCS Workshop on Open Language Archives*. http://www.language-archives.org/events/olac02/.

Bird, S. and Simons, G. (2003a). Extending Dublin Core metadata to support the description and discovery of language resources. *Computers and the Humanities*, 37:375–388. http://arxiv.org/abs/cs.CL/0308022.

Bird, S. and Simons, G. (2003b). Seven dimensions of portability for language documentation and description. *Language*, 79:557–82.

Bird, S. and Simons, G. (2004). Building an Open Language Archives Community on the DC foundation. In Hillmann, D. and Westbrooks, E., editors, *Metadata in Practice: a work in progress*. Chicago: ALA Editions.

Bird, S. and Tadadjeu, M. (1997). *Petit Dictionnaire Yémba-Français (Dschang-French Dictionary)*. Cameroon: ANACLAC.

Brogan, M. L. (2004). A survey of digital library aggregation services. http://www.diglib.org/pubs/brogan/.

CCSDS (2002). *Reference Model for an Open Archival Information System (OAIS): Blue Book*. NASA Consultative Committee for Space Data Systems. http://public.ccsds.org/publications/archive/650x0b1.pdf.

Cieri, C. and Bird, S. (2001). Annotation graphs and servers and multi-modal resources: Infrastructure for interdisciplinary education, research and development. In *Proceedings of ACL/EACL Workshop on Sharing Tools and Resources*, pages 23–30. Somerset, NJ: Association for Computational Linguistics.

Cotton, S. and Bird, S. (2002). An integrated framework for treebanks and multilayer annotations. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1670–1677. Paris: ELRA. http://arXiv.org/abs/cs/0204007.

Graff, D. and Bird, S. (2000). Many uses, many annotations for large speech corpora: Switchboard and TDT as case studies. In *Proceedings of the Second International Conference on Language Resources and Evaluation*. Paris: ELRA. http://arXiv.org/abs/cs/0007024.

Hughes, B. (2006). Searching for language resources on the web: User behaviour in the Open Language Archives Community. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 601–604. Paris: European Language Resources Association.

Hughes, B., Baldwin, T., Bird, S., Nicholson, J., and MacKinlay, A. (2006). Reconsidering language identification for written language resources. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 485–488. Paris: European Language Resources Association.

Hughes, B. and Kamat, A. (2005). A metadata search engine for digital language archives. *DLib Magazine*, 11(2).

Koehler, W. (2002). Web page change and persistence—a four-year longitudinal study. *Journal of the American Society for Information Science and Technology*, 53:162–171. http://portal.acm.org/citation.cfm?id=506072.506080&coll=ACM&dl=ACM&CFID=77507462&CFTOKEN=24608581.

Lai, C. and Bird, S. (2004). Querying and updating treebanks: A critical survey and requirements analysis. In *Proceedings of the Australasian Language Technology Workshop*, pages 139–146. http://eprints.unimelb.edu.au/archive/00000774/.

Langendoen, D. T., Farrar, S., and Lewis, W. D. (2002). Bridging the markup gap: smart search engines for language researchers. In *Proceedings of the International Workshop on Resources and Tools in Field Linguistics*. Paris: ELRA. http://faculty.washington.edu/wlewis2/papers/LangFarLew02.pdf.

Lewis, W. D. (2003). Mining and migrating interlinear glossed text. In *Proceedings of the EMELD Workshop on Digitizing and Annotating Texts and Field Recordings*. http://emeld.org/workshop/2003/Lewis-paper.pdf.

Ma, X., Lee, H., Bird, S., and Maeda, K. (2002). Models and tools for collaborative annotation. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 2066–2073. Paris: ELRA. http://arXiv.org/abs/cs/0204004.

Maeda, K. and Bird, S. (2000). A formal framework for interlinear text. In Bird, S. and Simons, G., editors, *Proceedings of the Workshop on Web-Based Language Documentation and Description*. http://www.ldc.upenn.edu/exploration/expl2000/papers/.

Maeda, K., Bird, S., Ma, X., and Lee, H. (2002). Creating annotation tools with the annotation graph toolkit. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 1914–1921. Paris: ELRA. http://arXiv.org/abs/cs/0204005.

NISO (2003). Information retrieval (z39.50): Application service definition and protocol specification. http://www.loc.gov/z3950/agency/Z39-50-2003.pdf.

Simons, G. (2002). A query facility for selective harvesting of OLAC metadata. http://www.language-archives.org/NOTE/query.html.

Simons, G. (2003). Specifications for an OLAC metadata display format and an OLAC-to-OAI_DC crosswalk. http://www.language-archives.org/NOTE/olac_display.html.

Simons, G. (2006). Ensuring that digital data last: The priority of archival form over working form and presentation form. SIL Electronic Working Papers 2006-003, SIL International. An expanded version of a paper originally presented at the: EMELD Symposium on Endangered Data vs. Enduring Practice, Linguistic Society of America annual meeting, 8–11 January 2004, Boston, MA, http://www.sil.org/silewp/abstract.asp?ref=2006-003.

Simons, G. and Bird, S. (2002). OLAC process. http://www.language-archives.org/OLAC/process.html.

Simons, G. and Bird, S. (2003a). Building an Open Language Archives Community on the OAI foundation. *Library Hi Tech*, 21:210–218. http://www.arxiv.org/abs/cs.CL/0302021.

Simons, G. and Bird, S. (2003b). The Open Language Archives Community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18:117–128.

Simons, G., Fitzsimons, B., Langendoen, D. T., Lewis, W., Farrar, S., Lanham, A., Basham, R., and Gonzalez, H. (2004a). A model for interoperability: Xml documents as an rdf database. In *Proceedings of the EMELD Workshop on Linguistic Databases and Best Practice, 15-18 July 2004, Detroit, MI*. http://emeld.org/workshop/2004/simons-paper.pdf.

Simons, G., Lewis, W., Farrar, S., Langendoen, D. T., Fitzsimons, B., and Gonzalez, H. (2004b). The semantics of markup: Mapping legacy markup schemas to a common semantics. In Wilcock, G., Ide, N., and Romary, L., editors, *Proceedings of the 4th workshop on NLP and XML (NLPXML-2004)*, pages 25–32. Association for Computational Linguistics. http://emeld.org/documents/SOMFinal1col.pdf.

US Library of Congress (2000). MARC 21: Specifications for record structure, character sets, and exchange media. http://lcweb.loc.gov/marc/specifications/.

Mark Y. Liberman

Director, Institute for Research in Cognitive Science
Director, Linguistic Data Consortium
Trustee Professor of Phonetics in Linguistics
Professor, Computer and Information Science Department

Degrees:
MSc 1972 Linguistics, Massachusetts Institute of Technology
PhD 1975 Linguistics, Massachusetts Institute of Technology

Professional Employment:
AT&T Bell Laboratories: Member of Technical Staff, 1975-1987;
Head, Linguistics Research Department, 1987-1990.
University of Pennsylvania: Professor, 1990–present.

Brief Research Summary:
The phonology and phonetics of tone and intonation; gestural, prosodic morphological and syntactic ways of marking focus,and their use in discourse; formal models for linguistic annotation; multi-language information retrieval; corpus linguistics.

Most Relevant Publications:

- S. Bird and M. Liberman, "A Formal Framework for Linguistic Annotation", Speech Communication 2001 33(1,2) pp. 23-60. (ftp://ftp.cis.upenn.edu/pub/sb/papers/specom00/specom00.pdf)
- S. Bird and M. Liberman, "Annotation Graphs as a Framework for Multidimensional Linguistic Data Analysis", in Proceedings, Workshop on Standards and Tools for Discourse Tagging. Association for Computational Linguistics, 1999. (http://www.ldc.upenn.edu/Papers/DTAG1999/dtag.pdf)
- M. Liberman and C. Cieri "The Creation, Distribution and Use of Linguistic Data," *Proceedings, First International Conference on Language Resources and Evaluatio*n, Granada, 1998. (http://www.ldc.upenn.edu/Papers/LREC1998/LREC1998.pdf)
- M. Liberman, J.M. Schultz, S. Hong and V. Okeke, "The Phonetic Interpretation of Tone in Igbo," Phonetica 1993 50(3) pp. 147–160.
- M. Liberman and J. Pierrehumbert, "Intonational invariance under changes in pitch range and length", in M. Aronoff and R. Oehrle, Eds,, *Language Sound Structur*e, MIT Press, 1984, pp. 157–223.

Current Research Funding:

- Talkbank (NSF KDI and SBE): co-P.I., 9/99–9/2002
- Multilingual Information Access and Management (NSF): P.I., 3/2000–3/2002
- Resources for Multilingual Information Technology (DARPA): P.I., 1/2000–1/2003
- Empirical Multilingual Processing (DARPA): co-P.I., 1/2000–1/2005
- Multidimensional Exploration of Linguistic Databases (NSF): co-P.I., 2/2000 to 2/2002

Synergistic Activities:

- Founded and directs the Linguistic Data Consortium, which has produced more than 160 digital publications since 1992. These speech and text corpora and lexicons have been used by nearly 1000 organizations around the world, as a basis for research and

development in speech- and language-related science and engineering. (http://www.ldc.upenn.edu)

- Director of the Institute for Research in Cognitive Science at the University of Pennsylvania, an NSF Science and Technology Center that fosters interdisciplinary research through the interaction of investigators from the disciplines of Computer Science, Linguistics, Mathematical Logic, Neuroscience, Philosophy and Psychology.
- Resident Faculty Master, Ware College House, University of Pennsylvania. Ware is an undergraduate dorm that is home to about 500 students.
- Current or recent editorial advisory boards: Cognition; Computer Speech and Lang uage; Speech Communication; International Journal of Corpus Linguistics.

Collaborators and Other Affiliations:

- *Collaborators within past 48 months (other than at Penn):* Akin Akinlabi (Rutgers University), Claude Barras (LIMSI, Paris), Edouard Geoffrois (ETCA, Paris), Brian MacWhinney (CMU).
- *Graduate Advisors:* Morris Halle (MIT), Noam Chomsky (MIT). Paul Kiparsky (Stanford), Ken Stevens (MIT).
- *Recent advisees or thesis committee connections:* Michael Collins (AT&T Labs), Jason Eisner (Rochester University), Carmen Fought (Pitzer), Beth Ann Hockey (NASA Ames), Dan Melamed (West), Corey Miller (Nuance), Adwait Ratnaparki (IBM), Jeff Reynar (Microsoft), Michael Schultz (amazon.com).
- Total PhD students (current and past): 20. Total Postdocs sponsored: 3.

Biographical Sketch

Steven Bird

A Professional Preparation

University of Melbourne, Australia	Computer Science and Mathematics	B.Sc. (hons) 1985
University of Melbourne, Australia	Computer Science	M.Sc. 1987
University of Edinburgh, UK	Cognitive Science	Ph.D. 1991

B Appointments

1. Associate Professor, Department of Computer Science and Software Engineering, University of Melbourne (2002–present);
2. Senior Research Associate, Linguistic Data Consortium, University of Pennsylvania (2002–present);
3. Associate Director, Linguistic Data Consortium, Adjunct Associate Professor, Computer and Information Science, and Adjunct Associate Professor, Linguistics, University of Pennsylvania (1998–2002);
4. Research Fellow, Centre for Cognitive Science, University of Edinburgh (1990–98).

C Publications

C.1 Five Relevant Publications

Bird, S, Y Chen, S Davidson, H Lee, & Y Zheng (2006). Designing and Evaluating an XPath Dialect for Linguistic Queries. *Proceedings of the 22nd International Conference on Data Engineering (ICDE)* pp 52–61.

Goldman, J, S Renals, S Bird, F de Jong, M Federico, C Fleischhauer, M Kornbluh, L Lamel, D Oard, C Stewart & R Wright (2005). Accessing the Spoken Word, to appear in *International Journal on Digital Libraries* 5.

Bird, S & G Simons (2004). Building an Open Language Archives Community on the DC Foundation. In Hillmann and Westbrooks (editors), *Metadata in Practice: A Work in Progress*, ALA Editions, pp 203–222.

Simons, G & S Bird (2003). Building an Open Language Archives Community on the OAI Foundation, to appear in *Library Hi Tech* 21, 210–218. Special Issue on the Open Archives Initiative.

Bird, S & G Simons (2003). Seven Dimensions of Portability for Language Documentation and Description, *Language* 79, 557–582.

C.2 Five Other Publications

Simons, G & S Bird (2003). The Open Language Archives Community: An infrastructure for distributed archiving of language resources *Literary and Linguistic Computing* 18: 117-128.

Bird, S & G Simons (2003). Extending Dublin Core Metadata to support the description and discovery of language resources, *Computing and the Humanities* 37, 375–388.

Bird, S & M Liberman (2001). A formal framework for linguistic annotation. *Speech Communication*, 33, 23–60.

Bird, S & G Simons (2000). *Web-Based Language Documentation and Description*. http://www.ldc.upenn.edu/exploration/expl2000/

Bird, S (1999). Multidimensional exploration of online linguistic field data. *Proceedings of the 29th Annual Meeting of the Northeast Linguistics Society*, pp 33–47.

D Synergistic Activities

Coordinator of the Open Language Archives Community, a worldwide virtual library of language resources [www.language-archives.org]

Editor of ACL Digital Anthology [acl.ldc.upenn.edu]

Guest editor of a special issue of *Speech Communication* on speech annotation and corpus tools.

E Collaborators and Other Affiliations

1. *Collaborators/co-editors:* Catherine Bow (U Melbourne), Peter Buneman (U Edinburgh), Christopher Cieri (U Penn), Susan Davidson (U Penn), Franciska de Jong (U Twente), Marcello Federico (U Trento), Carl Fleischhauer (Library of Congress), Jerry Goldman (U Chicago), Jonathan Harrington (U Kiel, Germany), Chu-Ren Huang (Academia Sinica, Taiwan), Baden Hughes (U Melbourne), Ewan Klein (U Edinburgh), Mark Kornbluh (Michigan State), Lori Lamel (LIMSI), Haejoong Lee (U Penn), Mark Liberman (U Penn), Edward Loper (U Penn), Xiaoyi Ma (U Penn), Brian MacWhinney (CMU), Kazuaki Maeda (U Penn), Craig Martell (U Penn), Douglas Oard (U Maryland), Steve Renals (U Edinburgh), Gary Simons (SIL), Claire Stewart (Northwestern U), Richard Wright (BBC).

2. *Graduate Advisors:* Roland Sussex, University of Queensland, Australia; Ewan Klein, University of Edinburgh, UK; Robin Cooper, Göteborg University, Sweden.

3. *recent advisees or thesis committee connections:* John Bell, Phil Blunsom, Trevor Cohn, Rod Farmer, Catherine Lai, Patrick Ye.

4. *Total number of Ph.D. students:* 10

Biographical Sketch

Gary F. Simons

A Professional Preparation

Seattle Pacific University	Interlanguage	B.A. (Summa Cum Laude) 1974
Cornell University	Linguistics	M.A. 1976
Cornell University	Linguistics	Ph.D. 1979

B Appointments

Associate VP for Academic Affairs, SIL International (1999–)
Adjunct Associate Professor of Language Development, Graduate Institute of Applied Linguistics (1999–)
Adjunct Assistant Professor of Linguistics, University of Texas at Arlington (1985–)
Director, Academic Computing, SIL (1986–1999)
Manager, Language Data Processing, SIL (1984–1985)
Translation Advisor for North Malaita, Solomon Islands (with SIL, 1979–1983)
Graduate Research Assistant, Cornell University (1976–1978)

C Publications

C.1 Five Relevant Publications

Bird, Steven and Gary Simons (2004). Building an Open Language Archives Community on the DC Foundation. In Diane Hillmann and Elaine Westbrooks (eds.), *Metadata in Practice*, pp. 203–222. Chicago: American Library Association. Preprint: `http://www.sil.org/~simonsg/preprint/Metadata%20in%20Practice.pdf`

Bird, Steven and Gary Simons (2003). Seven Dimensions of Portability for Language Documentation and Description, *Language* 79(3):557–582. Preprint: `http://www.sil.org/~simonsg/preprint/Seven%20dimensions.pdf`

Simons, Gary and Steven Bird (2003). Building an Open Language Archives Community on the OAI Foundation, *Library Hi Tech* 21(2):210–218. Special Issue on the Open Archives Initiative. Preprint: `http://arxiv.org/abs/cs.CL/0302021`

Simons, Gary and Steven Bird (2003). The Open Language Archives Community: An infrastructure for distributed archiving of language resources *Literary and Linguistic Computing* 18(2):117-128. Preprint: `http://arxiv.org/abs/cs.CL/0306040`

Bird, Steven and Gary Simons (2003). Extending Dublin Core Metadata to support the description and discovery of language resources, *Computers and the Humanities* 37(4):375–388. Preprint: `http://arxiv.org/abs/cs.CL/0308022`

C.2 Five Other Publications

Simons, Gary F. (2006). Ensuring that digital data last: The priority of archival form over working form and presentation form. *SIL Electronic Working Papers* 2006-003. Dallas: SIL International. Online: `http://www.sil.org/silewp/abstract.asp?ref=2006-003`

Simons, Gary F., William D. Lewis, Scott O. Farrar, D. Terence Langendoen, Brian Fitzsimons, and Hector Gonzalez (2004). The semantics of markup: Mapping legacy markup schemas to a common semantics. In

Graham Wilcock, Nancy Ide, and Laurent Romary (eds.), *Proceedings of the 4th workshop on NLP and XML (NLPXML-2004)*, pp. 25–32. Association for Computational Linguistics.
Preprint: `http://linguistlist.org/emeld/emeld/documents/index.cfm`

Simons, Gary F. (1998). The nature of linguistic data and the requirements of a computing environment for linguistic research. In John Lawler and Helen Aristar Dry (eds.), *Using Computers in Linguistics: A Practical Guide*, pp. 10-25. London and New York: Routledge.

Pike, Kenneth L. and Gary F. Simons (1996). Toward the historical reconstruction of matrix patterns in morphology. In Kenneth L. Pike, Gary F. Simons, Carol V. McKinney, and Donald Burquest, *The Mystery of Cultural Contacts, Historical Reconstruction, and Text Analysis*, pp. 1-37. Georgetown University Press.

Langendoen, D. Terence and Gary F. Simons (1995). A rationale for the TEI recommendations for feature-structure markup, *Computers and the Humanities* 29(3):191-209. [Reprinted in Nancy Ide and Jean Veronis (eds.), *The Text Encoding Initiative: Background and Context*, pp. 191-209. Dordrecht: Kluwer Academic Publishers.]

D Synergistic Activities

Coordinator (with Steven Bird) of the Open Language Archives Community, a worldwide virtual library of language resources [`www.language-archives.org`] (2000–)

Developer (with Peter Constable) of *ISO 639-3: Codes for the representation of names of languages* (providing unique three-letter identifiers for approximately 7,500 human languages, past and present) and developer of the Registration Authority web site [`www.sil.org/iso639-3/`] (2002–)

Executive editor of *Ethnologue: Languages of the World*, print and web versions, fifteenth edition (published 2005) and sixteenth edition (in preparation)

External consultant on past NSF-sponsored projects for Linguist List (Eastern Michigan U)—Software Development for the LINGUIST Network, The LINGUIST Multi-List Support Project, Database Design for Endangered Languages Data, E-NELD: Electronic Metastructures for Endangered Language Data—and for the Linguisitic Data Consortium (U Pennsylvania)—Linguistic Exploration, International Standards for Language Engineering, TalkBank. Consultant to three current projects—LL-Map: Language and Location (Eastern Michigan U), Multi-Tree: A Digital Library of Language Relationships (Wayne State U), and Five Languages of Eurasia: Field Work, Analysis and Digital Archiving (Colgate U)—with combined commitment of 12 days per year.

Committee on Endangered Languages and their Preservation, Linguistic Society of America (2006–)

E Collaborators and Other Affiliations

1. *Collaborators/co-eds.:* Anthony Aristar (Wayne State U), Steven Bird (U Melbourne), Peter Constable (Microsoft), Helen Aristar Dry (Eastern Michigan U), Scott Farrar (U Washington), Hector Gonzalez (California State U, Fresno), Larry Hayashi (SIL International), Chu-Ren Huang (Academia Sinica, Taiwan), Baden Hughes (U Melbourne), D. Terence Langendoen (NSF), William Lewis (U Washington), Mike Maxwell (CASL), Alexander Nakhimovsky (Colgate U)

2. *Graduate Advisor:* Joseph Grimes (SIL and U Hawaii at Manoa)

3. *Thesis advisees:* Eric Albright (SIL), Eugene Gruber (?), Linda Humnick (SIL), Lars Huttar (SIL) Taeho Jang (SIL), Ken Prettol (SIL), John Wimbish (SIL). *Total: 7*