To appear in:
Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation,
20–22 November 2008, Cebu City, Philippines.

Toward a Global Infrastructure for the Sustainability of Language Resources [*]

Gary F. Simons[a] and Steven Bird[b]

[a]SIL International and Graduate Institute of Applied Linguistics
7500 W. Camp Wisdom Road, Dallas, TX 75236, USA
gary_simons@sil.org

[b]Dept of Computer Science and Software Engineering, University of Melbourne, Victoria 3010, Australia
and Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA 19104, USA
sb@csse.unimelb.edu.au

Abstract. This paper describes work the Open Language Archives Community (OLAC) is doing to contribute to a global infrastructure for the sustainability of language resources. After offering a definition of language resource, it addresses the issue of what makes language resources sustainable by defining six necessary and sufficient conditions for their sustained use, then discusses what it takes to make such sustainability a reality by describing the roles of four key sets of players—creators, archives, aggregators, users. With this background, the paper describes the community infrastructure OLAC has developed for allowing its members to express consensus about best practices for digital archiving, plus the technical infrastructure it has developed to provide aggregation and search for the language resources community. The concluding section probes the broader issue of sustainable development to consider the sustainability of language resources in the context of the sustainability of language development and of languages themselves.

Keywords: Language documentation, metadata, digital archives, sustainable development

1. Introduction

Sustainability has become a byword of our times. In fact, *The Global Language Monitor* recognized *sustainable* as the Top Word of 2006.[1] Behind all the buzz there is an important concept that has significance even for our language resources community. Focus on the sustainability of the planet in the news media is making us increasingly aware that unless we

[*] This work is supported by the NSF Project *OLAC: Accessing the World's Language Resources*, awards 0723357 and 0723864 to the University of Pennsylvania and the Graduate Institute of Applied Linguistics. Components of the infrastructure have been developed with our research assistants Haejoong Lee and Debbie Chang and archiving consultant Joan Spanne. We are grateful to members of the OLAC community for their collaboration in developing and implementing OLAC's standards and recommendations.

[1] http://www.languagemonitor.com/top_word_lists

mend our wasteful ways, we could squander the world's natural resources along with the opportunity of future generations to enjoy the same quality of life that we do.

The language resources community is no stranger to waste. Waste happens when the resources resulting from prior work are no longer available due to the deterioration of the media that store them or the obsolescence of the formats that encode them. Waste also happens when a new project redoes work that has already been done by someone else, because the new project does not know about the prior work or because the prior work was not made available in a form they could use. Even more pervasive is the waste that happens when ordinary users, who would have no ability to create the needed resources, miss out entirely on the opportunity to benefit from resources that already exist because they are not able to discover or access or use them. Rather, we should be seeking to create an environment in which language resources thrive through regular use by all who can benefit from them.

This paper describes the work the Open Language Archives Community (OLAC) is doing to address issues like these. First, the paper defines the scope of OLAC by offering a definition of language resource (§2). Next, it identifies the conditions that are necessary for the sustainable use of language resources (§3) and defines the roles of the four sets of players—creators, archives, aggregators, and users—that are involved in making such sustainability a reality (§4). With this background, the paper is then able to describe what OLAC is doing to contribute to a global infrastructure for supporting the sustainable use of language resources (§5). The concluding section probes the broader issue of sustainable development to consider the sustainability of language resources in the context of the sustainability of language development and of languages themselves (§6).

2. Defining Language Resource

The notion of "language resource" is something our community tends to take for granted and is something that OLAC has taken for granted since its inception. As the founding mission statement says, the primary goal of OLAC is "creating a worldwide virtual library of language resources."[2] OLAC recognizes that the language resources of interest to our community come not only from sources within our community but also from many sources that would not identify themselves as part of our community (e.g., libraries, national archives, book sellers). As OLAC has begun interacting with such institutions in order to bring their resources into a single global infrastructure, we have found it necessary to define exactly what we mean by a language resource. This is the latest version of our definition:

> A language resource is any physical or digital item that is a product of language documentation, description, or development or is a tool that specifically supports the creation and use of such products.

The following paragraphs elaborate the major concepts used in the definition.

Language resources are rooted in the study of language. More specifically, they arise from a series of activities that could be termed the "three D's": language documentation, language description, and language development.

The distinction between language documentation and language description is defined in Himmelmann's (1998) seminal work. Language *documentation* "aims at the record of the linguistic practices and traditions of a speech community." Documentation is concerned with the primary data of language study. It is done by compiling a sample of instances of language in actual use (whether spoken or written), commenting on those instances (such as through situational metadata, transcription, translation, annotation), and then archiving the whole

[2] http://www.language-archives.org/

collection. Language *description,* by contrast, is concerned with the secondary data of language study. It is done by analyzing the primary data and generalizing over it to produce works like a dictionary and grammar that describe how the language works as a system of signs.

Language *development* adds a third dimension involving resources that focus on acquiring language skills. The term "language development" is used in two ways by different subcommunities. It is most widely used to refer to the process by which humans learn language. Documentary corpora of individuals learning language and secondary descriptions based on those primary data are certainly within the scope of language resources. So are works that reflect a second sense of "language development," namely, in reference to the activities that result from language planning (Cooper, 1989). Under the heading of corpus planning, these include terminology development and the production of prescriptive dictionaries and grammars. Under the heading of acquisition planning, these include the development of materials that are designed to help people learn a language or learn language skills like reading and writing. A twenty-first century approach to language planning could also include "automation planning" which involves the development of processes that leverage new language technologies so as to amplify human productivity.

With these background definitions for documentation, description, and development, it is now possible to elaborate the other key terms in the definition. First, a language resource is defined as any resource that is a product of any of the three D's. The intention of *any* in the definition is to place no limit on the form of a resource; for instance, it may be physical or digital, textual or audiovisual, published or unpublished. The intention of *product* is to say that being the output of language documentation, description, or development is what identifies a resource as a language resource, not being the input. The input to description is typically documentation, and the input to development is typically documentation and description, so it is tempting to see the input role as being a defining characteristic. However, it is clear that this approach fails when we consider documentation. If the inputs to documentation were language resources, then every speech event in daily life and every article in a newspaper and every text page on the web would be a language resource. It is true that all of these are potential inputs to the study of language, but they wouldn't become language resources until somebody performed the documentation process of compiling them into a collection, providing metadata (and possibly other commentary) for the collection, and lodging the result in an institutional archive.

Secondly, the community that produces language resources is vitally interested in the tools that are used in that work; in fact, many in the community focus on the development of such tools. Thus, any resource that is a *tool* that supports the creation of language resources is also defined to be a language resource. Such tools can take a number of forms; for instance, a tool might be a textbook on theory or a software program that automates aspects of the work or a blog that gives methodological advice to practitioners. The definition limits the scope to tools that are *specifically* designed to support the creation or use of language resources. For instance, a general word processor or recording device might be used to create a language resource, but it is not itself a language resource. However, a document giving advice on how to use such general tools in creating language resources would be. Similarly, a software tool that automates a specific aspect of language description (like dictionary building) would be a language resource.

This definition of language resource, then, identifies the scope of the worldwide virtual library that OLAC seeks to build. It aims to encompass any resource that is the product of language documentation, description, or development, and any resource that is a tool for supporting these activities.

3. Necessary and Sufficient Conditions for Sustainability

Sustainability, in the general sense, refers to the ability to maintain indefinitely a given process or a desired state. (The richer sense implied by "sustainable development" is discussed below in §6.) In the case of language resources, we want to sustain their use. Thus the problem of sustaining language resources can be understood as the problem of maintaining the use of language resources over time. That problem can be summarized as follows:

> Given the relentless process of entropy that degrades digitally stored information, the relentless process of innovation that makes equipment and methodologies obsolete even while they are still in common use, and the relentless proliferation of information resources of all kinds that makes it ever harder to find language resources of interest, how do we keep our language resources from falling into disuse and wasting away as yet more detritus on a digital scrap heap?

To be sustainable, the results of our work must transcend computer environments, communities of practice, domains of application, and especially the passage of time (Bird and Simons, 2003). Ensuring availability to future generations is particularly crucial for resources that document languages that are themselves in danger of being lost (Simons, 2006).

If our goal is to sustain the use of language resources, then we must begin by asking, "What does it take to sustain the use of language resources?" By identifying the necessary conditions for the use of language resources, we can identify the objectives that an infrastructure for sustainability must meet. If the identified conditions are also sufficient to ensure use, then they would constitute a complete set of objectives. We propose that there are six necessary and sufficient conditions for the use of a language resource. That is, to sustain use, the community's infrastructure must establish and maintain the following characteristics of a language resource:

- The resource must be *extant*.
- The resource must be *discoverable*.
- The resource must be *available*.
- The resource must be *interpretable*.
- The resource must be *portable*.
- The resource must be *relevant*.

The middle four conditions form a group that defines the attributes that make a resource usable. Thus the model can be summarized as follows: A resource will be used if it still exists, if it is usable, and if a user finds it relevant.

The first necessary condition for the use of a language resource is that it be *extant*. Once a resource comes into existence, we cannot assume its ongoing existence, particularly in the case of digital resources which can be lost in an instant through an event like a disk crash or can be lost gradually as storage media degrade over time. Sustainability requires that the custodian of a resource follows procedures to ensure that the resources are preserved against all reasonable contingencies (e.g., via offsite backup), that the resources are periodically migrated to fresh and current media, and that all file copies are authenticated as exactly matching the source file.

Second, the resource must be *discoverable;* it cannot be used unless the prospective user is able to discover its existence and its whereabouts. The key to this is descriptive metadata. Metadata is "data about the data;" it is like the catalog card in a library that describes a book and tells where to find it. But it is not enough that descriptive metadata simply exists. The description of the resource must be published in such a way that the prospective user who knows nothing about the resource is able to discover its existence when searching. Furthermore, the description of the resource must be done in such a way that the prospective user is able to

judge it as being relevant without having to first obtain the resource. If these conditions are met, then a resource is discoverable.

Third, once discovered, the resource must be *available;* it cannot be used unless it is truly available to the prospective user. Availability has two major facets. First, the user must have the right to access and use the resource. In order to guarantee sustainable use of a resource, it is essential that the rights of future users be established when the resource is created and clearly stated when it is archived. Where possible, distributing resources under the terms of Open Access[3] (such as under a Creative Commons[4] license) fosters the most widespread use. Second, the user must know the procedure for accessing the resource. In the case of physical resources, this involves knowing how to gain access to the single archived instance of a resource or how to order a copy of a published resource. In the case of digital resources, maximal availability is achieved through dissemination via links on the Internet, but sustainable long term access requires persistent URLs that will not break.

Fourth, once accessed, the resource must be *interpretable;* it cannot be used if the user is not able to make sense of the content. The *Reference Model for an Open Archival Information System* (CCSDS, 2002), the standard adopted as ISO 14721 in 2003, states that one of the fundamental functions of an archive is to ensure that the resources it archives are "independently understandable" by the designated user community. That is, the prospective user should be able to use the resource without needing to consult the creator to clarify any details of content. For a language resource, this means documenting things like the situational context, methodologies, terminologies, abbreviations, markup conventions, and character encodings.

Fifth, once accessed the resource must be *portable;* it cannot be used if it does not operate within the user's working environment. A resource must work with the user's hardware and operating system, and with the software tools that are available to the user. Maximizing portability means using formats that are open and transparent and thus supported by many software vendors (including open source projects), rather than using proprietary formats that force users to buy proprietary software that is not even likely to still be available to future generations. Another practice that promotes use in a wider variety of contexts is following the best practices (such as for markup or terminology) of the target community of practice.

Finally, the resource must be *relevant*. Maintaining the five conditions above makes it possible for a resource to be used well into the future, and unless these conditions are met, it cannot be used. But these conditions are not sufficient to guarantee that a resource will be used. The final condition is relevance; a resource will not be used unless it is relevant to the needs of the prospective user. In the case of endangered languages, the members of the language community themselves form a critical user group. The linguistics community has come to recognize that when we work with endangered languages, we have an ethical responsibility to create resources that are relevant to the language community and their aims for their language (e.g., Nathan, 2006). Funding agencies also play a role when they wield their perceptions of relevance as a factor for deciding which resource development efforts they will fund.

4. The Key Players and their Roles

No single person or institution can achieve the sustainable use of language resources; rather, it takes an infrastructure involving four key sets of players who have distinct roles:

- Creators — Persons who create language resources
- Archives — Institutions that curate language resources for long-term preservation and access

[3] http://www.eprints.org/openaccess/
[4] http://creativecommons.org/

- Aggregators — Institutions that gather language resources from multiple archives and make them interoperate
- Users — Persons who want to use language resources

The basic model, illustrated in Figure 1, is as follows. Individuals (shown as ellipses on the left side of the diagram) create and use language resources, but they depend on institutions (shown as rectangles on the right side of the diagram) to bridge the gap between producer and consumer. Creators place the resources they create under the care of Archives that are committed to preserve the resources and provide access over the long term. It is impossible for an individual Creator to achieve sustainability since life is short; resources cannot be sustainable unless they are under the care of long-lived institutions. However, archival care itself is not enough to guarantee sustained use. This is because there are so many institutions curating language resources that the User who could truly benefit from using the resources cannot possibly know about all the Archives to look in. Aggregators are thus the key to linking the supply (in Archives) with the demand (by Users). Aggregators harvest resources from Archives (either full data or metadata) and provide services where Users can make a single search request to find and retrieve resources from all participating Archives.

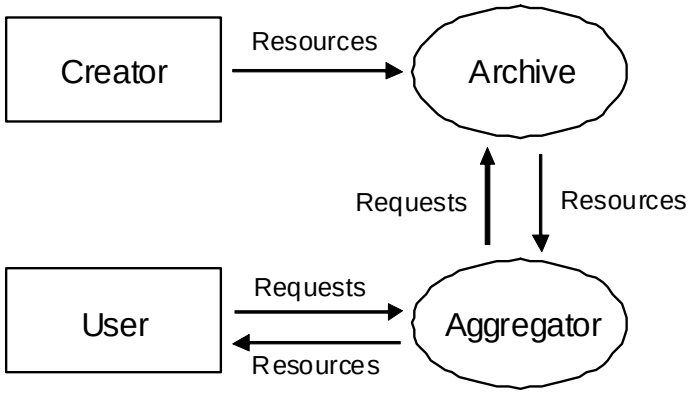

Figure 1: Key players in the infrastructure for sustainable language resource use.

The roles of the players can be further elucidated by considering the function each performs with respect to the six necessary and sufficient conditions introduced above. The results are summarized below in Table 1. The functions listed in the table identify best current practices (which are not always common practices). The following paragraphs explain the roles by considering one condition at a time.

Extant. It is the Creator who first brings a language resource into existence. In order to ensure sustained existence, the Creator must supply the complete original resource to an Archive. The Archive confirms that the submission is complete and valid as to formats on acceptance; it then follows "documented policies and procedures which ensure that the information is preserved against all reasonable contingencies, and which enable the information to be disseminated as authenticated copies of the original" (CCSDS, 2002:3-1). The Aggregator harvests information about all extant resources being preserved by Archives, which gives the User the potential to use any extant resource since it will be discoverable through the Aggregator.

Table 1: The roles of the key players in sustaining language resource use.

	Creator	Archive	Aggregator	User
Extant	Supplies complete and valid original	Follows preservation procedures	Harvests data or metadata of extant resources	Has potential to use any extant resource
Discoverable	Supplies descriptive metadata	Publishes interoperable metadata	Provides search service over all resources	Searches for resources of relevance
Available	Secures and documents access rights	Provides access consistent with access rights	Mediates access to resources in archives	Accesses resources that seem relevant
Interpretable	Supplies documentation of content	Ensures independent understandability for target group	Interprets resources to provide aggregation	Uses resources if they are understood
Portable	Supplies a portable original	Migrates resources as formats change	Interoperates over aggregated portable resources	Uses resources if they operate in user's context
Relevant	Prioritizes creation of resources deemed relevant	Prioritizes curation of resources deemed relevant	Provides services relevant to its target community	Determines if resource is actually relevant

Discoverable. The Creator's contribution to making resources discoverable is to provide the descriptive metadata that answers the basic questions of who, what, when, where, why, and how. The User's role is to search for resources that are relevant to the present need. The metadata in a catalog record is what makes it possible to match a User's search query with the resources that are likely to be relevant. The role of the Archive is to ensure that the descriptive metadata follow best practice guidelines and then to publish these descriptions in such a way that Aggregators can harvest them and build search services that interoperate over the resources in all known Archives.

Available. The Creator's contribution to making resources available is to secure the needed permissions for sharing the resources with others and to document any restrictions to access that the rights holders may stipulate. The User's role is to access resources that, upon discovery, seem relevant. The role of the Archive is to ensure that the access rights are clearly known and documented, and then to provide a means by which a User who meets any restrictions on access can obtain a copy of the resource. The role of the Aggregator is to link the User with resources in the Archive, either through a direct URL in the case of an openly accessible digital resource or through information on what to do next in the case of a restricted resource.

Interpretable. It is the role of the Creator to make the language resource understandable to its prospective users. For language documentation, this may involve augmenting a recording with things like transcription, translation, commentary, and a description of the situational context. For language description, this may involve adding definitions of terms, abbreviations, markup conventions, and character encodings. The role of the Archive is to ensure that the resource to be preserved "is Independently Understandable to the Designated Community" (CCSDS, 2002:3-1). In other words, the anticipated User should be able to understand the resource without needing the assistance of the Creator. When the resource is independently understandable, it is possible for an Aggregator to harvest the data itself and provide services that interoperate over aggregated data. It is also possible for a User to interpret and use the

resource. In the case where the Creator has archived a resource that conforms to a community-wide standard for information encoding (such as a markup schema or an ontology), then the Aggregator can harvest such resources to provide services that interoperate over the shared semantics of the standard.

Portable. The Users can use a resource only if it operates in their working environment, including their hardware, operating system, and available software tools. Since different Users have different working environments and the environments of the future will be different yet, the Creator should prepare resources in such a way that they operate across the variety of environments that Users typically have. This means that the Creator must look beyond favorite working and presentation forms to produce archival forms that provide LOTS—lossless, open, transparent, and supported by multiple software suppliers (Simons, 2006). The role of the Archive is to ensure that the form of the resource is adequately portable to be an archival form, and then to provide a service that will migrate the resource to new forms in the future if the archived form ceases to be widely usable. The role of the Aggregator is to harvest portable resources from the Archives and create services (like search and conversion) that exploit their interoperability.

Relevant. The Creator has limited time and resources and thus prioritizes the creation of resources that are deemed relevant. Similarly the Archive cannot afford to preserve every item that has ever been created and thus prioritizes what it will accept for long-term curation to resources that are deemed relevant. An Aggregator has a particular target community and develops services that are relevant for that community; a specialized Aggregator will therefore selectively harvest and interoperate over resources that it deems relevant to its target community. Ultimately, it is the User who determines whether a resource is relevant when deciding whether or not to use it.

5. OLAC's Contribution to Global Infrastructure

As set out in its mission statement, 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." Thus, OLAC's mission has two facets, and these are reflected in two kinds of infrastructure that OLAC is building—a community infrastructure and a technical infrastructure.

Community Infrastructure. Language resource creators, archives, and users comprise a loose network involving scholars, language learners, archivists, and technologists, along with their associated institutions. As we have already discussed, the key players engage with language resources in various ways. There is no centralized coordination of activities among all these players, and innovations tend to spread virally as new standards and practices are supported by tools. OLAC has a special role in this community, namely, that of providing an agreed upon process for the community to develop and document its consensus on best practices in digital archiving. This is done by means of a standards process defined in *OLAC Process,*[5] the founding standard adopted by OLAC to define its governing ideas (i.e. the purpose, vision, and core values) and to describe how it is organized and how it operates.

The initial focus has been on discoverability as a necessary condition for language resource use. To this end OLAC has established a consensus standard on metadata for describing language resources (Bird and Simons, 2004). OLAC has extended Dublin Core metadata,[6] the

[5] http://www.language-archives.org/OLAC/process.html
[6] http://dublincore.org/documents/dcmi-terms/

dominant metadata standard in the digital library and World Wide Web communities, by providing the following additional descriptors that are tailored to language resources:

- Subject language: for identifying precisely (with a code from the ISO 639 standard[7]) 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 a relevant subfield 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.

All of these vocabularies are formally defined on the OLAC site, along with best practice recommendations[8] and comprehensive usage guidelines.[9] This metadata infrastructure slots into the discoverability row of Table 1. The OLAC process has the potential for being used to develop infrastructure in other rows of the table, for example, defining consensus on best practices for preservation, for systematic description of access rights, for encoding specific data types, and more.

Technical Infrastructure. Aggregators are a key part of the global infrastructure set out in Table 1, since they permit users to discover and access relevant language resources without needing to know about all the individual archives and resource creators. In the early days of the web, manually constructed topical indexes played this role. However, such indexes go out of date quickly and do not scale up as the number of resources to index increases exponentially. Here, too, OLAC has a special role within the language resources community, namely, that of providing the primary aggregator dedicated to language resources. 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,[10] another standard of the digital library community (Simons and Bird, 2003). OLAC's technical infrastructure takes care of archive registration, metadata validation, crosswalks to Dublin Core and HTML for discovery in broader communities, together with search services and usage tracking.

The technical infrastructure for the OLAC aggregator is defined in two standards that have been adopted by the community. *OLAC Metadata*[11] defines the XML format used for the interchange of metadata records within the framework of the Open Archives Initiative OAI. *OLAC Repositories*[12] defines the standards OLAC archives must follow to implement a metadata repository that can be harvested by an aggregator using the OAI Protocol for Metadata Harvesting. The list of participating "archives" includes more than just archives. The metadata infrastructure works well to describe online services; thus, many participants are services that publish catalogs (indexed by language) of the language resources they provide.

Examples. The following listing shows a sample metadata record in the XML format prescribed by the *OLAC Metadata* standard. It is the description of a Shoebox[13]-format lexicon of the Ega language (Côte d'Ivoire) that is housed in the University of Bielefeld Language Archive.

---

[7] http://www.sil.org/iso639-3/
[8] http://www.language-archives.org/REC/bpr.html
[9] http://www.language-archives.org/NOTE/usage.html
[10] http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
[11] http://www.language-archives.org/OLAC/metadata.html
[12] http://www.language-archives.org/OLAC/repositories.html
[13] http://www.sil.org/computing/shoebox/

```
<olac:olac xmlns:olac="http://www.language-archives.org/OLAC/1.0/"
    xmlns:dc="http://purl.org/dc/elements/1.1/">
   <dc:title>Ega lexicon (Gbery)</dc:title>
   <dc:creator>Gbery, Eddy Aime</dc:creator>
   <dc:creator>Baze, Lucien</dc:creator>
   <dc:contributor>Lindenlaub, Juliane</dc:contributor>
   <dc:description>Ega lexicon in Shoebox format</dc:description>
   <dc:date>2003-03</dc:date>
   <dc:type xsi:type="olac:linguistic-type" olac:code="lexicon"/>
   <dc:format>shoebox</dc:format>
   <dc:subject xsi:type="olac:language" olac:code="ega"/>
   <dc:language xsi:type="olac:language" olac:code="fra"/>
   <dc:language xsi:type="olac:language" olac:code="eng"/>
   <dc:language xsi:type="olac:language" olac:code="ega"/>
   <dc:language xsi:type="olac:language" olac:code="deu"/>
   <dc:coverage>Cote d'Ivoire</dc:coverage>
</olac:olac>
```

Note the use of domain-specific code values on the Type, Subject, and Language elements. These make it possible to support precise searching for languages and resource types within the aggregated catalog. The entire metadata repository in which this record can be found is openly available (in the format prescribed by the *OLAC Repositories* standard) at:

```
http://www.spectrum.uni-bielefeld.de/langdoc/olac.xml
```

In accordance with the OAI Protocol for Metadata Harvesting, the archive has assigned the above record a unique identifier of *oai:langdoc.uni-bielefeld.de:UBI-EGA-010*. The following URL is thus the *GetRecord* command (following the OAI protocol) for harvesting that single record from the repository at the University of Bielefeld in OLAC format:

```
http://www.language-archives.org/sr/
   www.spectrum.uni-bielefeld.de/langdoc/olac.xml?
   verb=GetRecord&
   identifier=oai:langdoc.uni-bielefeld.de:UBI-EGA-010&
   metadataPrefix=olac
```

Current Status. At present, OLAC has some 35 participating archives, and the OLAC search engine indexes a combined total of approximately 36,000 records. The larger participating archives include the Alaska Native Language Center (U Alaska), Archive of the Indigenous Languages of Latin America (U Texas), Audio Archive of Linguistic Fieldwork (UC Berkeley), Oxford Text Archive, PARADISEC (Pacific And Regional Archive for Digital Sources in Endangered Cultures), SIL Language and Culture Archives, and the Linguistic Data Consortium (U Penn).

This is an excellent beginning, but in surveying the current status we have identified three significant shortcomings. First, the quality of metadata in the majority of participating archives does not meet the level of the best practice recommendations and this limits the quality of search. Second, many significant language archives are not yet participating in OLAC. Third, many of the participating “archives” are more accurately described as digitization projects and are not yet following best practices in digital archiving that will ensure long-term preservation (e.g., CCSDS, 2002).

Future Directions. In addition to these shortcomings, OLAC—and the language resources community more broadly—is confronted with three major challenges that need to be addressed before the promise of universal access to relevant resources is realized. 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). 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). A second challenge is that library automation solutions are part of the deep web and remain hidden to the language resources community at large. This means that users searching for language resources need to visit other services like WorldCat and OAIster; it would be better for the language-resource content of these services to be fully integrated with OLAC. A third challenge 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 shortcomings and challenges, OLAC has been awarded NSF sponsorship for a new project named "OLAC: Accessing the World's Language Resources"[14] which aims to greatly improve access to language resources by achieving an order-of-magnitude increase in the coverage of the OLAC catalog and in the use of OLAC search services. This involves improving access to language resources on two levels. First, to address the above-listed shortcomings and improve access to resources in language archives, the project includes activities aimed at achieving the following outcomes:

- All OLAC repositories should have up-to-date catalogs that contain metadata conforming to best practice.
- All major language archives should be participating in OLAC.
- All OLAC repositories should conform to current best practices for the long-term curation of their holdings.

Second, in order to address the above-listed challenges and improve access to language resources on the web, the project includes activities aimed at achieving the following outcomes:

- Low-density language materials identified by linguistic web mining should be reliably categorized with OLAC vocabularies.
- Language resources held in libraries and digital repositories should be indexed in OLAC through services that crosswalk and enrich existing catalog records.
- 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.

The first phase of the project is drawing to a close. It has focused on developing documentation and services to improve the quality of metadata and search. The improved technology infrastructure is now ready to accept registrations from new participants; all interested projects or institutions which archive language resources or which offer language resources through online services are invited to contact the authors.

6. Toward Sustainable Language Development

The recent emphasis on sustainability in public discourse arises from the global concern over the deteriorating natural environment in many parts of the world. Damage to the environment is leading to what many refer to as "the extinction crisis."[15] For instance, noted biologist Edward O. Wilson (2002) warns that human activities, if left unchecked, could result in the extinction of half the world's plant and animal species by the end of this century.

The pressures of globalization are having a similar effect on the world's minority languages. Early in the last decade, linguist Michael Krauss extrapolated from what had already taken place in Australia and North America to warn that the twenty-first century "will see either the

[14] http://olac.wiki.sourceforge.net/
[15] http://www.well.com/~davidu/extinction.html

death or the doom of 90% of mankind's languages" (1992:7). His essay closes with a sobering challenge: "Obviously we must do some serious rethinking of our priorities, lest linguistics go down in history as the only science that presided obliviously over the disappearance of 90% of the very field to which it is dedicated" (1992:10). He advocates going beyond the scientific work of documenting and describing languages to also working with members of the language community to participate in language development and even working politically beyond the community to increase the language's chance of survival.

We are thus confronted with a challenge that is even greater than the sustainability of language resources, namely, the sustainability of languages themselves. Where languages are threatened because children are no longer learning them, acquisition planning becomes a priority and the sustained products of language documentation and description are key inputs to the language development activities that are needed. Thus, the sustainability of language depends in part on the sustainability of the language resources that contribute to language development, which will lead to the production of new language resources that can in turn enable further development, and so the cycle of sustainability continues.

In 1983, global concern over the world's deteriorating natural and social environment prompted the UN General Assembly to establish the World Commission on Environment and Development. The commission's final report (Brundtland, 1987) is what brought the term *sustainable development* to the world's attention. To this day, their definition of sustainable development is most often cited definition, namely, "development that meets the needs of the present without compromising the ability of future generations to meet their own needs." The commission recognized that the solution must simultaneously address interrelated environmental, economic, and social dimensions—the so-called "three pillars of sustainability." Elkington (1994) picked up this basic model and applied it to doing business in terms of the "triple bottom line"—later popularized as "People, Planet, Profit." The idea is that sustainability is achieved not by maximizing shareholder profit but by coordinating the interests in all three areas of all stakeholders (that is, of everyone affected by the business activities, whether directly or indirectly). Those interests are to simultaneously pursue the three bottom lines of economic prosperity, environmental quality, and social equity.

By analogy this threefold purpose can inform the broader agenda of the language resources community. (1) As for the economic agenda, doing linguistics can be likened to a quest for riches—the riches of knowledge about language in general and about thousands of languages in particular (Simons, 2007). Developing a central aggregator gives the language resources community a means of amassing its treasures into a single virtual storehouse and of being able to measure the size and scope of that treasury. Such an initiative is already underway through the efforts of OLAC. (2) As for the environmental agenda, the analog for the language resources community is improving the quality of the linguistic ecosphere. One piece of our global infrastructure is the *Ethnologue* (Gordon, 2005) which monitors factors like the population and vitality of all known languages. Where it is clear that a language is endangered, one goal of the language resources community should be to at least ensure that good documentation and description of the language are preserved so that future generations (especially of the ethnic community) will still have access to the language in some form. (3) As for the social agenda, the language resources community should be looking to attain a form of social equity in which minority languages are not overlooked in the efforts of language resource development and in which the products that result include ones that are relevant to the needs and aspirations of the language communities themselves and not just ones that are relevant to outsiders.

The technical infrastructure of OLAC could be exploited to help the language resources community track the availability of documentation, description, and development resources for

all the languages of the world. The community could thereby monitor the world situation with respect to the triple bottom line of sustainable development. Table 2 gives a taste of what is possible. It shows a breakdown of language resources currently known to the OLAC aggregator in terms of the size of the associated language. (Resources that are not cataloged with a specific ISO 639 language code are not included in the tabulation; nor are the 7,296 records for the language descriptions in *Ethnologue*.) The *Languages* column gives the number of known living languages in the world in the given population range as reported in the *Ethnologue* (Gordon, 2005). The *In OLAC* column gives the number of languages for which resources are cataloged in OLAC, first as an absolute number and then as a percentage of the known languages in that population range. These two columns give some indication of the quality of the linguistic environment, first in terms of the languages themselves and then in terms of the response of the language resources community. For instance, 99% of the languages with more than 10 million speakers have resources that are discoverable through OLAC, but only 37% of the languages with 100 to 999 speakers do.

Table 2: OLAC coverage in relation to language size. (as of 15 Sept 2008)

Population Range	Languages	In OLAC		Resources
10,000,000 or more	83	82	99%	3,341
1,000,000 to 9,999,999	264	223	84%	1,431
100,000 to 999,999	892	575	64%	2,607
1,000 to 99,999	3,746	1,797	48%	9,012
100 to 999	1,071	392	37%	2,305
1 to 99	548	271	49%	832
Unknown population	308	86	28%	307
Total living languages	*6,912*	*3,426*	*49%*	*19,835*
Extinct languages	602	130	22%	315

The *Resources* column in Table 2 is a count of the total number of OLAC resources for all languages in the given population range. Dividing *Resources* by *Languages* gives the average number of OLAC resources per language in the population range. That number is plotted for each population range as a bar graph in Figure 2. The graph gives some indication of how the language resources community is performing with respect to a goal of "social equity." We see that the largest languages have more resources by more than an order of magnitude and that the number of resources available declines steadily as language groups get smaller.

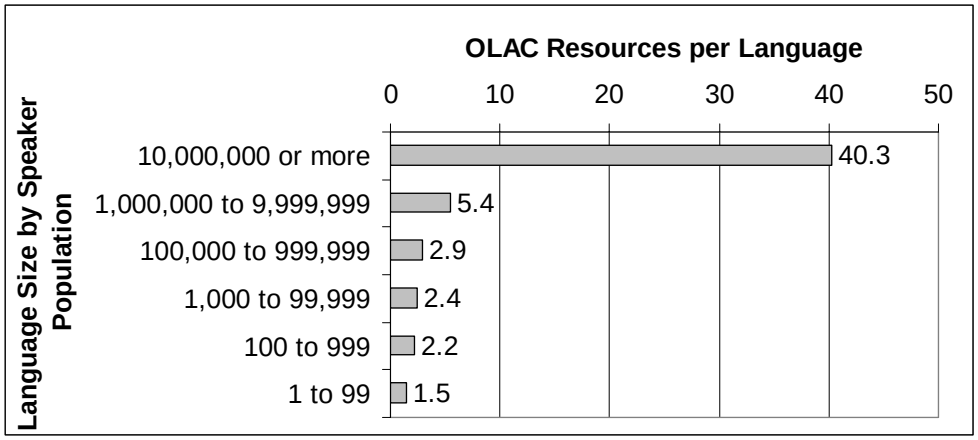

Figure 2: Average OLAC resources per language by language size.

These results are only suggestive since the coverage of OLAC is far from complete. However, they do give a glimpse of how the technical infrastructure offered by OLAC could help the language resources community to monitor the availability of documentation, description, and development resources for all the languages of the world. The simple counting of resources as employed in Table 2 is a rather crude metric since resources vary so widely in type and in their extent. The work of McConvell and Thieberger (2001, especially pp. 69–70) is instructive in pointing a way forward. They developed a 20-point index for assessing the level of documentation and description for the endangered languages of Australia. If OLAC were to adopt a standardized vocabulary for identifying the complete range of language resource types, as well as for quantifying their extent, it would be possible to use the aggregated catalog to automatically generate indices for the level of documentation, description, and development of languages as reflected in the total set of resources that are known to exist for each.

7. Conclusion

The development community has long recognized that achieving sustainable development requires coordinated efforts of many actors. This is equally true of sustainable language development. A recent World Development Report (World Bank, 2003:xiv) observes that sustainable development efforts fail when:

- The actors fail to take the long view. That is, they opt for the short-term solution which ends up creating a bigger problem in the long term.
- The actors fail to represent dispersed interests. That is, powerful actors are driven by self interest with the result that they benefit at the expense of the less powerful who are adversely affected.
- The actors fail to commit to allowing assets to thrive. That is, over consumption or hoarding of resources leads to their ultimate loss.

We must not make those same mistakes as a language resources community. Let us not fail to take the long view; rather, by embracing the six factors for sustainability of language resources, we should strive to ensure their long-term use. Let us not fail to represent dispersed interests; rather, by giving attention to disempowered minority languages that are under threat and by pursuing language development efforts that are relevant to their needs and aspirations, we can encourage their survival. And finally, let us not fail to commit to allow our assets to thrive; rather, by committing to both of the above we will help the language resources and the languages themselves to thrive through sustained use.

References

Bird, S. and G. Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description. *Language*, 79, 557–582. <http://www.language-archives.org/documents/portability.pdf>

Bird, S. and G. Simons. 2004. Building an Open Language Archives Community on the DC Foundation. In D. I. Hillmann and E. L. Westbrooks, eds., *Metadata in Practice*, pp. 203–222. Chicago: American Library Association. <http://www.ldc.upenn.edu/sb/home/papers/mip.pdf>

Bruntland, G., ed. 1987. *Our Common Future: The World Commission on Environment and Development.* Oxford: Oxford University Press.

CCSDS. 2002. *Reference Model for an Open Archival Information System (OAIS).* CCSDS 650.0-B-1. Blue book, Issue 1. Consultative Committee for Space Data Systems. <http://public.ccsds.org/publications/archive/650x0b1.pdf>

Cooper, R. L. 1989. *Language Planning and Social Change.* Cambridge: Cambridge University Press.

Elkington, J. 1994. Towards the Sustainable Corporation: Win-win-win business strategies for sustainable development. *California Management Review*, 36(2), 90–100.

Gordon, R. G., Jr., ed. 2005. *Ethnologue: Languages of the World, Fifteenth Edition.* Dallas: SIL International. Online version: <http://www.ethnologue.com/>

Himmelmann, N. 1998. Documentary and Descriptive Linguistics. *Linguistics*, 36, 165–191.

Krauss, M. 1992. The World's Languages in Crisis. *Language*, 68(1), 4–10.

McConvell, P. and N. Thieberger. 2001. *State of Indigenous languages in Australia — 2001.* Australia State of the Environment Second Technical Paper Series (Natural and Cultural Heritage), Canberra: Department of the Environment and Heritage, <http://www.environment.gov.au/soe/2001/publications/technical/pubs/indigenous-languages.pdf>

Nathan, D. 2006. Proficient, Permanent, or Pertinent: Aiming for sustainability. In L. Barwick and N. Thieberger, eds., *Sustainable Data from Digital Fieldwork: From Creation to Archive and Back*, pp. 57–68. Sydney: Sydney Univ. Press. <http://ses.library.usyd.edu.au/bitstream/2123/1618/4/sddftoc.pdf>

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. <http://www.sil.org/silewp/abstract.asp?ref=2006-003>

Simons, G. 2007. Doing Linguistics in the 21st Century: Interoperation and the quest for the global riches of knowledge. *Proceedings of the E-MELD/DTS-L Workshop: Toward the Interoperability of Language Resources*, 13–15 July 2007, Palo Alto, CA. <http://linguistlist.org/tilr/papers/TILR%20Plenary.pdf>

Simons, G. and S. Bird. 2003. Building an Open Language Archives Community on the OAI Foundation. *Library Hi Tech*, 21(2), 210–218. <http://arxiv.org/abs/cs.CL/0302021>

Wilson, E. O. 2002. *The Future of Life.* New York: Alfred A. Knopf.

World Bank. 2003. *World Development Report 2003: Sustainable Development in a Dynamic World.* New York: World Bank and Oxford University Press.

Author/s:
Simons, G. F.; Bird, S.

Title:
Toward a global infrastructure for the sustainability of language resources

Date:
2008

Citation:
Simons, G. F. & Bird, S. (2008). Toward a global infrastructure for the sustainability of language resources . In Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation, Cebu City, Philippines.

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/25076

File Description:
Toward a global infrastructure for the sustainability of language resources