Toward a Global Infrastructure for the Sustainability of Language Resources

Gary F. Simons
SIL International and GIAL

Steven Bird
U of Melbourne and U of Pennsylvania

Coordinators, Open Language Archives Community

PACLIC 22, Cebu City, 20-22 Nov 2008

The problem of waste

- Language resources go to waste when
  - Media have deteriorated beyond use or formats have become obsolete
  - Projects reinvent the wheel because existing resources are not accessible
  - Potential users have no idea that relevant resources even exist or cannot access them

Overview of talk

- Foundational definitions
  - What is a *language resource?*
  - What are the necessary conditions for the *sustainable use* of language resources?
  - What are the roles of the *key players* involved in achieving such sustainability?
- OLAC's contribution toward a global infrastructure to support the sustainable use of language resources
- Considering sustainable development more broadly
  - The sustainability of language resources in relation to the sustainability of language development and of languages themselves

What is a language resource?

- From the OLAC mission statement:
  - We are working to create “a worldwide virtual library of language resources”
- Language resources are rooted in the study of language
- They arise from the “Three D’s”
  - Language Documentation
  - Language Description
  - Language Development

Documentation *vs.* description

- The seminal work:
  - Nikolaus Himmelmann, 1998. “Documentary and descriptive linguistics.” *Linguistics* 36:165–191.
- Documentation deals with the primary data
  - Provides “a comprehensive record of the linguistic practices characteristic of a given speech community” by collecting recordings and commenting on them
- Description creates secondary data
  - Aims at “the record of a language … as a system of abstract elements, constructions, and rules” by producing grammars, dictionaries, analyzed texts

Language development

- Resources that focus on acquiring language skills, in two senses:
  - the process by which humans learn language
  - the activities that result from language planning
    - Corpus planning — developing writing systems, terminology, prescriptive dictionary or grammar
    - Acquisition planning — materials for language learning, teaching reading and writing
    - Automation planning — processes that leverage new language technologies to amplify productivity

Tools

- The community that produces language resources is vitally interested in the tools that are used in that work, e.g.
  - A textbook on theory or method
  - A software program that is specifically designed to automate a “Three D” task
  - A document that advises how to do a “Three D” task using generic software

A definition

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

The sustainability problem

- Sustaining language resources =
  - Maintaining the use of language resources over time
- Given the relentless:
  - Entropy that degrades digitally stored information
  - Innovation that obsoletes hardware and software
  - Discovery that provides new ways of doing things
- How do we keep our language resources from
  - Falling into disuse, then
  - Slipping into oblivion

Necessary conditions

- Goal: Sustain the use of language resources
- A resource will be used if it is:
  - **Extant** (i.e., preserved) + Usable + **Relevant**
- A resource is usable if it is :
  - **Discoverable**
  - **Available**
  - **Interpretable**
  - **Portable**
- Thus, to sustain use, we must establish and sustain these six characteristics of language resources

1. Extant

- A language resource cannot be used if a faithful copy of the original resource ceases to exist
- Archiving institution must follow procedures to:
  - Ensure that the resources are preserved against all reasonable contingencies (*e.g.*, offsite backup)
  - Ensure periodic migration to fresh and current media
  - Ensure that all copies are authenticated as matching the original
  - Keep preservation metadata (provenance, fixity)

2. Discoverable

- A language resource cannot be used unless the prospective user is able to find it.
- The key is descriptive metadata:
  - The description of the resource must be published in such a way that the user to whom it is relevant is able to discover its existence when searching.
  - The description of the resource must be done in such a way that the user to whom it is relevant is able to judge it as being relevant without having to first obtain the resource.

3. Available

- A language resource cannot be used unless it is available to the prospective user.
- Availability has two major facets:
  - User must have the right to access and use the resource; the rights must be sorted out when the resource is created and clarified when it is archived
  - User must know the procedure for gaining access
- Open Access fosters the most widespread use
  - Long term access requires persistent URIs

4. Interpretable

- A language resource cannot be used if the user is not able to make sense of the content.
- OAIS standard (ISO 14721) states that:
  - Archives must ensure that resources are “independently understandable” by the designated user community (*i.e.,* no need to consult producer)
- *E.g.,* document the situational context, methodology, terminology, abbreviations, markup conventions, character encodings

5. Portable

- A language resource cannot be used if it does not interoperate in user's working environment.
- A resource must work with:
  - User's hardware and operating system
  - Software tools available to the user
  - Best practices of the designated user community
- Maximizing portability means:
  - Formats that are open and transparent (not proprietary)
  - Following best practice markup and terminology

6. Relevant

- A language resource will not be used unless it is relevant to the needs of the prospective user.
- Relevance enters into decisions of what to create, what to fund, what to archive.
  - In the case of endangered languages, the language community itself is a critical user group
  - We have an ethical responsibility to create resources that are relevant to the language community and their aims for their language

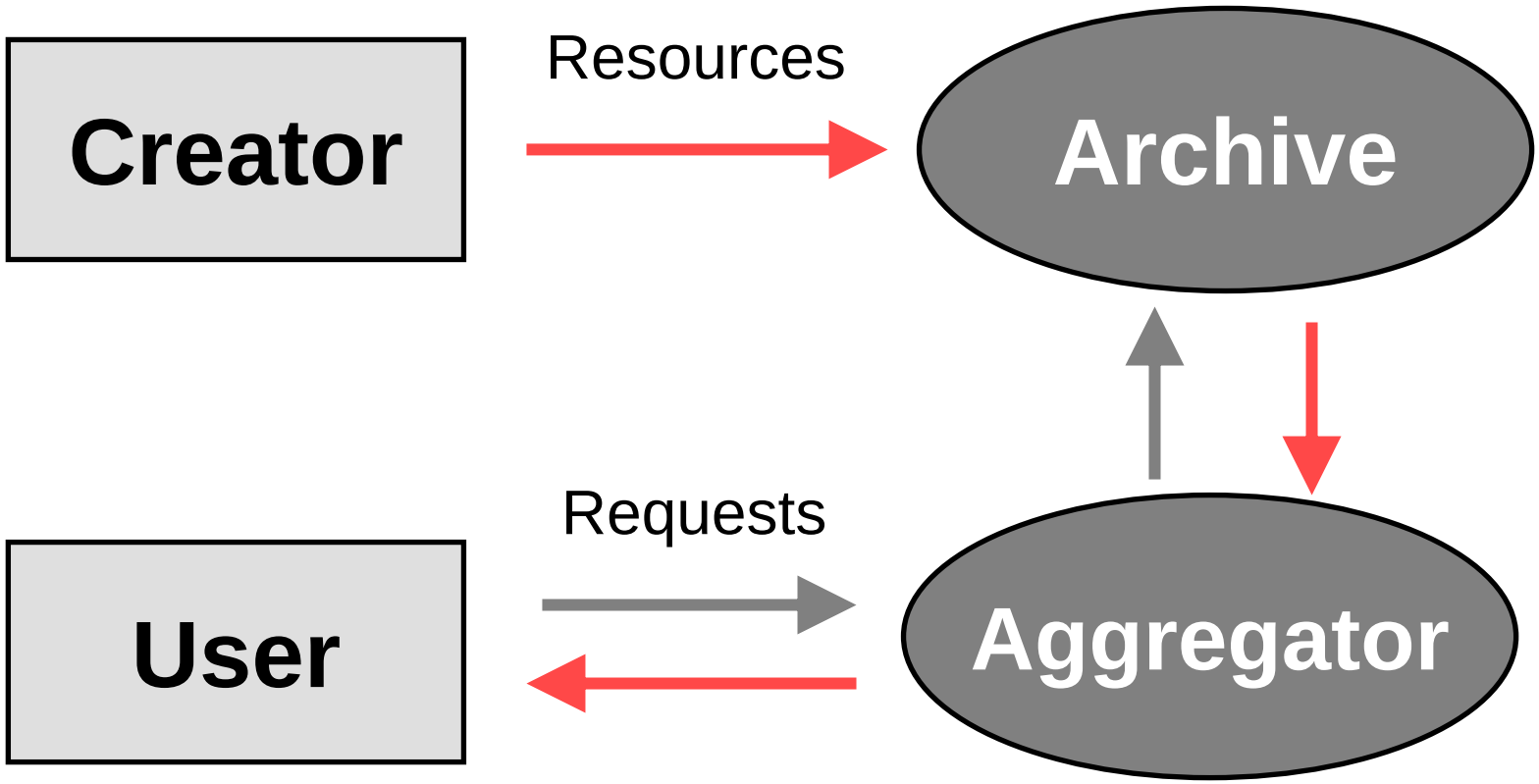
It takes an infrastructure

- ***Linguists*** can create resources that are portable and interpretable.
- They cannot preserve them long term or provide the means of access to all users.
  - That's what ***Archives*** do.
- They cannot make them discoverable.
  - That's what ***Aggregators*** (*e.g.*, Google) do.

The key players

Creator	A person who creates language resources
Archive	An institution that curates language resources for long-term preservation
Aggregator	An institution that makes resources from many archives interoperate
User	A person who wants to use language resources

The big picture



Overview

- Foundational definitions
  - language resource
  - conditions for sustainable use
  - key players — creator, archive, aggregator, user
- OLAC's contribution toward a global infrastructure to support the sustainable use of language resources
- Considering sustainable development more broadly
  - The sustainability of language resources in relation to the sustainability of language development and of languages themselves

Open Language Archives Community

www.language-archives.org

- OLAC is an international partnership of institutions and individuals who are creating a world-wide virtual library of language resources by:
  - Developing consensus on best current practice for the digital archiving of language resources
  - Developing a network of interoperating repositories & services for housing and accessing such resources
- Founded in December 2000
  - Now has 34 participating archives

Who's involved?

- Aboriginal Studies Electronic Data Archive
- Academia Sinica
- Alaska Native Language Center
- Archive of Indigenous Languages of Latin America
- ATILF Resources
- Berkeley Language Center
- Centre de Ressources pour la Description de l'Oral
- CHILDES Data Repository
- Comparative Corpus of Spoken Portuguese
- Cornell Language Acquisition Laboratory
- Dictionnaire Universel Boiste 1812
- DOBES catalogue (MPI, Nijmegen)
- Ethnologue: Languages of the World
- European Language Resources Association
- Laboratoire Parole et Langage
- Linguistic Data Consortium Corpus Catalog
- LINGUIST List Language Resources
- Natural Language Software Registry
- Online Database of Interlinear Text (ODIN)
- Oxford Text Archive
- PARADISEC
- Perseus Digital Library
- Research Papers in Computational Linguistics
- Rosetta Project 1000 Language Archive
- SIL Language and Culture Archives
- Surrey Morphology Group Databases
- Survey for California and Other Indian Languages
- TalkBank
- Tibetan and Himalayan Digital Library
- TRACTOR
- Typological Database Project
- University of Bielefeld Language Archive
- University of Queensland Flint Archive
- Virtual Kayardild Archive (Melbourne)

Community infrastructure

- How can the players of the language resources community organize to make decisions?
- OLAC plays a role by providing an agreed upon process for the community to develop and document its consensus on best practices in archiving of digital language resources.
- Defined in the *OLAC Process* standard
  - http://www.language-archives.org/OLAC/process.html

The OLAC document process

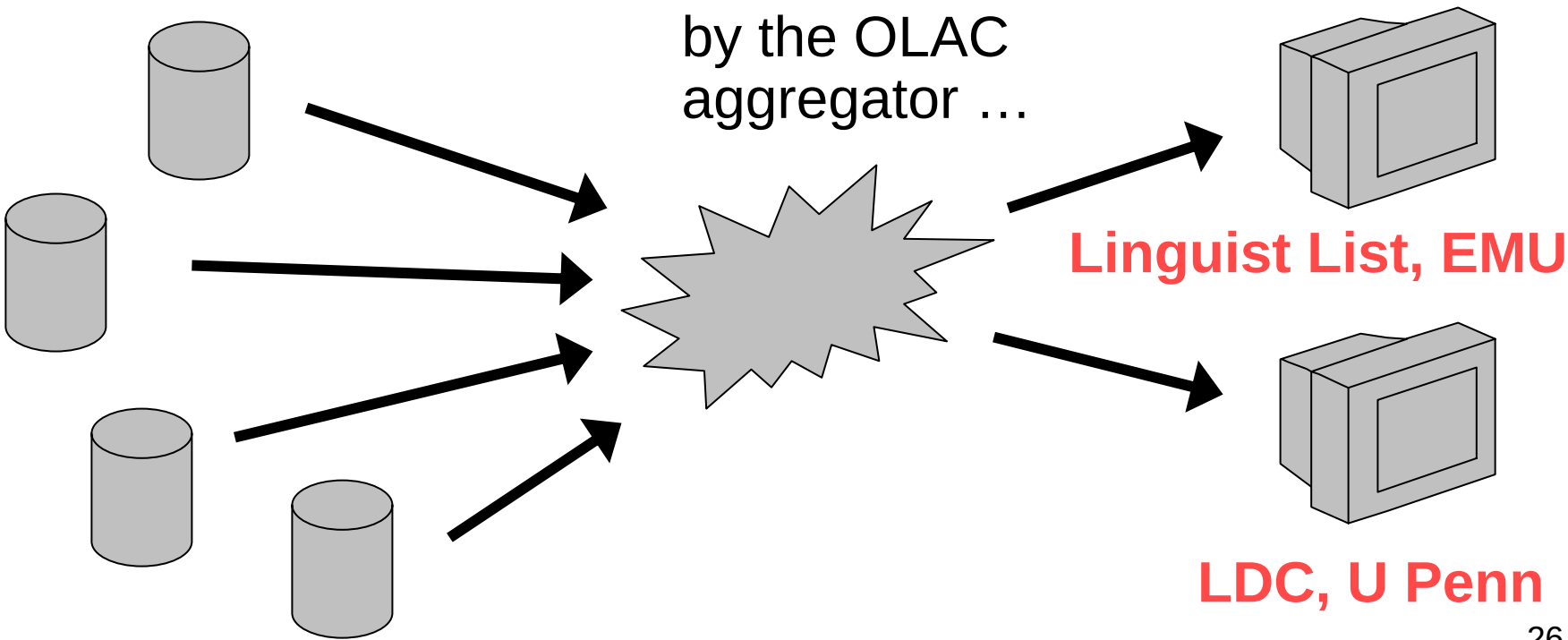
- Three types of document:
  - Standards, Recommendations, Notes
- Six status levels:
  - Draft, Proposed, Candidate, Adopted, Retired, Withdrawn
- The process defines the phases of the life cycle in terms of how a document moves from one status to another.

OLAC metadata standard

- Dublin Core metadata standard has:
  - Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type
- OLAC adds extensions (with controlled vocabularies) specific to our community:
  - Language Identification (ISO 639-3), Linguistic Data Type, Linguistic Field, Participant Role, Discourse Type

Technical infrastructure

- The 34 archives publish catalogs in a standard XML form …
- to be harvested by the OLAC aggregator …
- which supplies information to search services.



Record as published

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

OAI Protocol for Metadata Harvesting

- There are six verbs:
  - GetRecord, Identify, ListIdentifiers, ListMetadataFormats, ListRecords, ListSets
- Requests expressed as URLs:
  - *baseURL*?verb=*value&parameters*
- For instance:
  - http://www.ethnologue.com/oai_server.asp?verb=Identify
- Answer returned as an XML document

Search
OLAC: potawatomi Find -- All archives --

Search results for "potawatomi" in all OLAC archives 9 results from 4 archive(s)

Results from "ethnologue.com"

1. ★★★★★ oai:ethnologue.com:POT Similar records by: score date
title: *POTAWATOMI*: a language of USA
description: A page from the Web edition of Ethnologue: Languages of the World (14th edition) giving basic facts about the language and where it is spoken.

Results from "linguistlist.org"

1. ★★★★★ oai:linguistlist.org:lang_POT Similar records by: score date
title: LINGUIST List Resources for *Potawatomi*
description: A page listing all resources ...

Results from "sil.org" List all results from this archive (2 matches)

1. ★★ oai:sil.org:11119 Similar records by: score date subject
title: Patterns of person-number reference in *Potawatomi*
description: http://www.ethnologue.com/show_work.asp?id=11119
subject: Reference

Results from "perseus.tufts.edu" List all results from this archive (5 matches)

1. ★ oai:perseus.tufts.edu:Perseus:text:2000.03.0068 Similar records by: score language type
description: Descriptions of the *Potawatomi*, Miami, Sauk, Menomone [Menominee], Winnebago, and Dacota [Sioux] provide insights about the observers as well as the peoples observed.
title: Narrative of an expedition to the source of St. Peter's River, Lake Winnepeck, Lake of the Woods, &c. &c. performed in the year 1823, by order of the Hon. J.C. Calhoun, Secretary of war, under the command of Stephen H.

ISO 639-3

- OLAC achieves interoperation across archives by using this standard to precisely identify languages
- *ISO 639-3: Alpha-3 code for comprehensive coverage of languages* (published 2007-02-05)
  - Three-letter codes for ~6,900 living languages
  - Three-letter codes for ~600 extinct, historical, ancient, and constructed languages
  - RA site: **http://www.sil.org/iso639-3/**
- E.g., <dc:subject xsi:type="olac:language" olac:code="ega"/>

What is the current coverage?

	All archives	Excluding *Ethnologue*
Items in catalog	36,161	28,892
Items that are online	21,579 (60%)	14,310 (50%)
ISO 639-3 languages included	7,334	3,556

Current development

- We have begun the second year of a three year grant from the National Science Foundation to GIAL and U Penn with the aim of achieving
  - An order of magnitude increase in the coverage of the OLAC catalog through recruiting more archives to join and building gateways to library catalogs and institutional repositories
  - An order of magnitude increase in the use of the OLAC catalog through improved search services over the enlarged collection

Call for participation

- All institutions and projects with language resources to share are enthusiastically invited to participate.
- Visit www.language-archives.org to:
  - Try our two search services
  - Read workpapers and published articles
  - Subscribe to the OLAC-General mailing list
  - Learn how to publish your metadata catalog

Overview

- Foundational definitions
  - language resource
  - conditions for sustainable use
  - key players — creator, archive, aggregator, user
- OLAC's contribution toward a global infrastructure to support the sustainable use of language resources
- Considering sustainable development more broadly
  - The sustainability of language resources in relation to the sustainability of language development and of languages themselves

The extinction crisis

- Public discourse on sustainability arises from the global concern over the deteriorating natural environment.
  - Damage to the environment is leading to what many refer to as “the extinction crisis.”
- E.g., Biologist Edward O. Wilson in *The Future of Life* (2002) warns:
  - 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.

Language endangerment

- Krauss, Michael. 1992. “The world’s languages in crisis,” *Language 68(1):*4-10.
  - USA/Canada: 149 of 187 languages (or 80%) were NO longer learned by children; Australia: 90% were already moribund.
  - Unless we do something: “The coming century will see the death or the doom of 90% of mankind’s languages.”
  - He suggests we should go beyond the scientific work of documenting and describing languages to also working with members of the language community to participate in language development.

The cycle of sustainability

- We face an even greater challenge
  - The sustainability of languages themselves
- Revitalization is a kind of language development
  - The sustained products of language documentation and description are key inputs to the needed development activities
- Thus, the sustainability of language depends in part on the sustainability of the language resources
  - that contribute to development activities
  - which in turn produce new resources
  - which in turn feeds more development, and so on …

Pillars of sustainability

- World Commission on Environment and Development, 1987:
  - To be sustainable, development must simultaneously address the “three pillars of sustainability” — environmental, economic, and social issues
- Parts of business community have embraced this as the “triple bottom line” — Planet, Profit, People
  - The aim is not to maximize shareholder profit but to coordinate the interests of all stakeholders in all three areas, i.e., simultaneously pursue three bottom lines of
  - environmental quality, economic prosperity, social equity

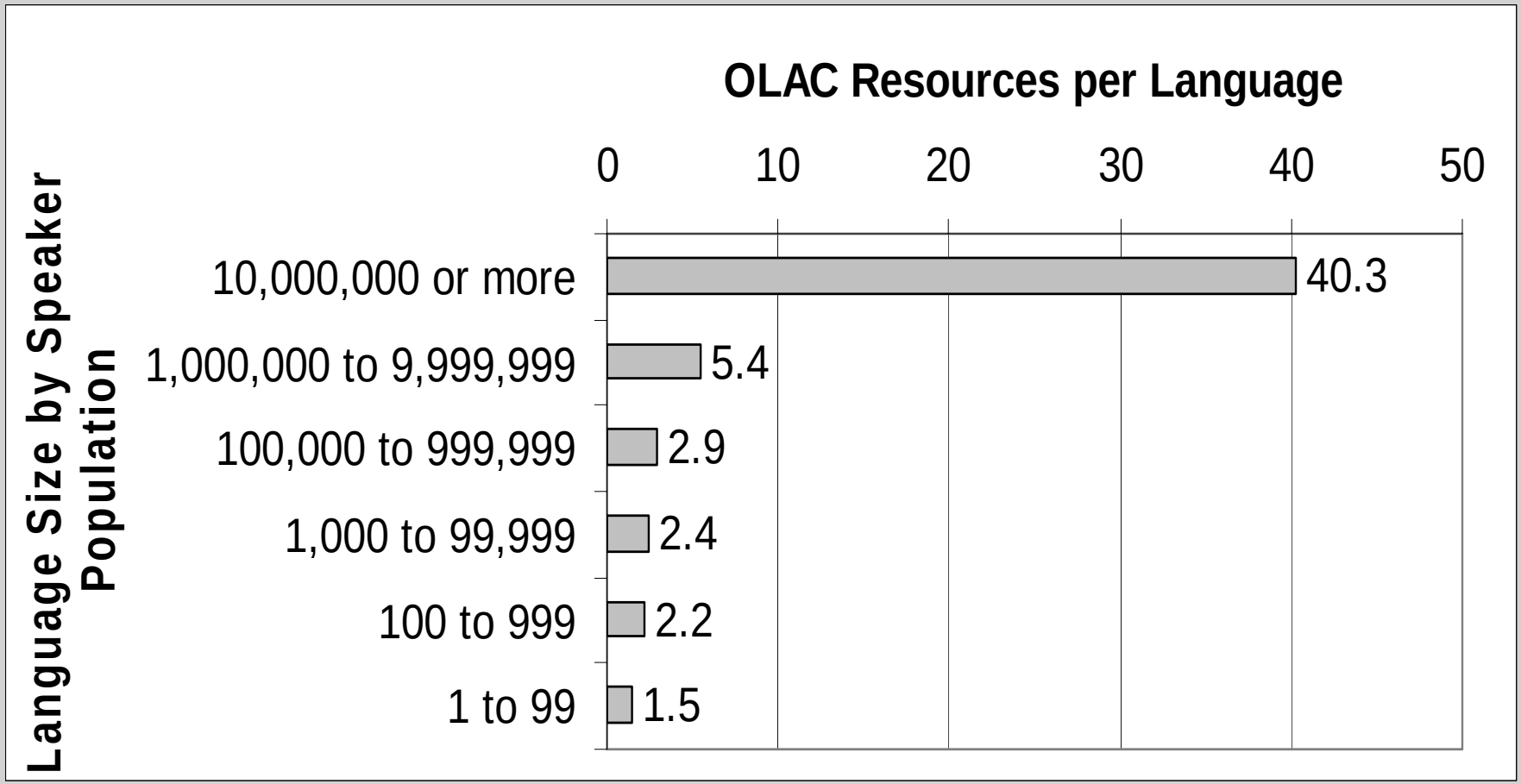
The triple bottom line

- Economic agenda
  - Developing a central aggregator as a virtual storehouse for the treasures of knowledge about language in general and about thousands of languages in particular
- Environmental agenda
  - Improving the quality of the linguistic ecosphere, including ensuring preservation of resources for seriously endangered languages so that future generations of the ethnic community will still have access to their language
- Social agenda
  - Seeking a form of social equity in which minority languages are not overlooked in the efforts of language resource development

OLAC coverage by language size

Population range	*Languages*	*In OLAC*		*Items*
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	308	86	28%	307
All living languages	*6,912*	*3,426*	*50%*	*19,835*
Extinct languages	602	130	22%	315

Inequity of resources in relation to language size



Common failings

- Achieving sustainable development requires coordinated efforts of many actors, which fail when:
  - The actors fail to take the long view
    - The short-term fix creates a bigger long-term problem
  - The actors fail to represent dispersed interests
    - Powerful minorities benefit at expense of everyone else
  - The actors fail to commit to allowing assets to thrive
    - Over consumption or hoarding leads to ultimate loss

— *World Development Report 2003*, World Bank, p. xiv

Conclusion: Toward sustainability

- Let us not fail to take the long view
  - Embracing the six factors for sustainability of language resources will ensure their long-term use
- Let us not fail to represent dispersed interests
  - Attending to sustainability of language development for all languages will encourage their survival
- Let us not fail to commit to allow assets to thrive
  - By committing to both of the above we will help the resources and the languages themselves to thrive