

Ensuring the sustainability of language resources

Gary F. Simons

SIL International and
Graduate Institute of Applied Linguistics

Workshop on
Processing Text Technological Resources,
Center for Interdisciplinary Research,
University of Bielefeld, 13–15 March 2008

Overview

- Setting the context
 - Understanding language resources in relation to language development
- Addressing the immediate problem
 - Ensuring the sustainability of language resources
- Considering the wider problem
 - Understanding language resources in relation to the sustainability of language development

The starting point

- Language development
 - Actions that are taken to increase a particular community's capacity to use their languages in their efforts to achieve their changing social, cultural, political, economic and spiritual goals.
- A wide range of applications, *e.g.*,
 - Acting to preserve identity and oral use
 - Acting to develop and spread literacy
 - Acting to enhance productivity through text technology

Definitions

■ Sustainability

- The ability to maintain something indefinitely at a desired level over time

■ Language resource

- Any digital resource that is a product of language development or that supports the activities of language development, *e.g.*,
 - Language documentation and description
 - Language teaching materials
 - Text technology tools and 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
- Necessary conditions for use:
 - Discoverable
 - Accessible
 - Authentic
 - Intelligible
 - Interoperable
- Thus, to sustain use, we must establish and sustain these five characteristics of language resources

1. 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.

2. Accessible

- A language resource cannot be used unless the prospective user can gain access to it.
- Access 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

3. Authentic

- A language resource cannot be used if it is not a faithful copy of the originally created resource.
- 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)

4. Intelligible

- 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

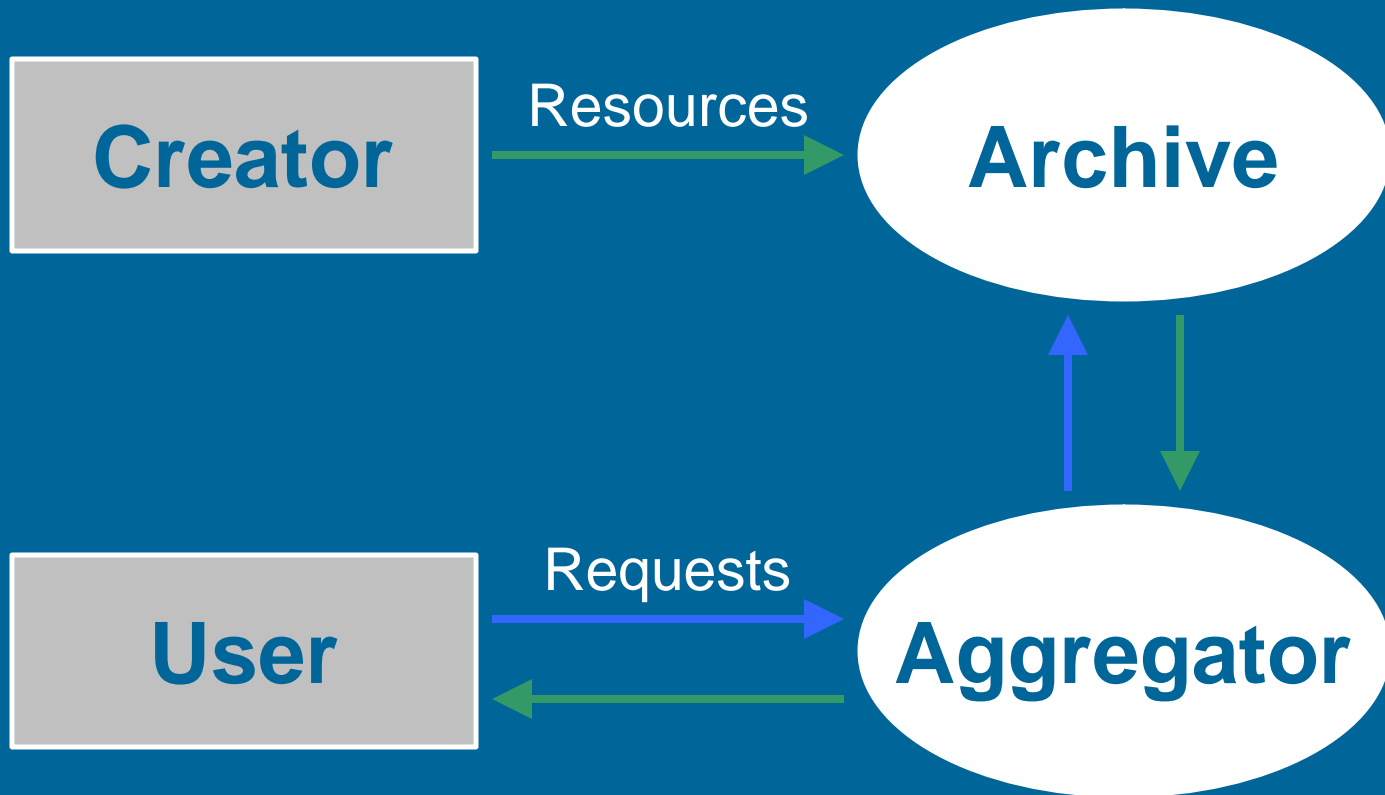
5. Interoperable

- 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 interoperability means:
 - Formats that are open and transparent (not proprietary)
 - Following best practice markup and terminology

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 language resources interoperate
User	A person who wants to use language resources

The big picture



The role of the creator

Discoverable	Supplies descriptive metadata
Accessible	Secures & documents access rights
Authentic	Supplies a complete & valid original
Intelligible	Supplies documentation of content
Interoperable	Supplies an interoperable original

The role of the archive

Discoverable	Ensures adequacy of metadata Publishes metadata in interoperable form
Accessible	Ensures clarity of access rights Provides a means of access
Authentic	Follows preservation practices that ensure dissemination of authenticated copies
Intelligible	Ensures that content is independently understandable to target users
Interoperable	Ensures that format is adequately interoperable in current environments

The role of the aggregator

Discoverable	Provides a search facility that interoperates across metadata or content from all archives
Accessible	Provides links to what the archives have made accessible
Authentic	Harvests only authenticated information
Intelligible	Makes the description of search and results understandable to target users
Interoperable	Provides interoperability with other aggregators to widen the circle of discovery

Two relevant aggregators



- DRIVER: Digital Repository Infrastructure Vision for European Research
 - Aggregates Open Access research results from hundreds of European institutions
 - <http://www.driver-repository.eu/>



- OLAC: Open Language Archives Community
 - Aggregates language resources from 34 institutions worldwide
 - <http://www.language-archives.org/>

What the user adds: Relevance

- The five factors are necessary for sustained use, but are not sufficient to guarantee it.
 - A sixth factor is involved: Relevance
 - This is where the role of the user enters in
- A language resource will not be used unless it is relevant to the needs of the prospective user.
 - David Nathan, 2006. “Proficient, permanent, or pertinent: aiming for sustainability.” In *Sustainable data from digital fieldwork*. University of Sydney.

Widening the view

- Given my definition of language resource as:
 - Any digital resource that is a product of language development or that supports the activities of language development
- We should also consider the sustainability of language resources as it relates to the sustainability of language development.
- Language development can be viewed as action to achieve an ever increasing series of goals.

Hierarchy of goals for sustainable language development

Sustainable Identity	Development that will sustain identity of a people even with language loss
Sustainable Orality	Development that will sustain the oral use of a language
Sustainable Literacy	Development that will sustain reading, writing, and literature in a language
Sustainable Automation	Development that will sustain the benefits of language automation

The cycle of sustainability

- Existing resources feed development
 - which in turn produces new resources
 - which in turn feeds more development, which ...
- Since language resources are the fuel of language development,
 - Sustainable language development depends on sustaining resources
 - Lest each generation must start over again
 - Or worse, lest in the absence of adequate development the language itself is not sustained

Biocultural diversity at risk

- Emphasis on sustainability arises from the global concern over the deteriorating natural and social environment in many parts of the world
- *The extinction crisis:* If non-sustainability continues at current rates, by end of century we will lose:
 - 50% of plant and animal species
 - 50% to 90% of languages
- Serious problem because diversity of life forms is what allows ecosystems to adapt and diversity of knowledge is what allows humankind to adapt



OLAC coverage in relation to language size

Where language identification is based on ISO 639-3 codes in descriptive metadata

<i>Population range</i>	<i>Langs.</i>	<i>In OLAC</i>		<i>Items</i>	<i>Per lg.</i>
10,000,000 or more	83	81	98%	2,533	30.6
1,000,000 to 9,999,999	264	209	79%	1,281	4.9
100,000 to 999,999	892	507	57%	2,097	2.4
1,000 to 99,999	3,746	1,593	43%	8,131	2.2
100 to 999	1,071	356	33%	1,866	1.7
1 to 99	548	237	43%	704	1.3
Unknown	308	56	18%	167	
<i>All living languages</i>	6,912	3,039	44%	16,779	
<i>Extinct languages</i>		95		231	

Sustainable development

- Most cited definition comes from the final report of the World Commission on Environment and Development convened by the UN in 1983.
 - "Sustainable development is development that meets the needs of the present without compromising the ability of future generations to meet their own needs."
- Language development does not occur in isolation
 - *E.g.*, economic development of the present may compromise a people's language, identity, and future well-being as they lose language and identity

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 five 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