Do you have valuable linguistic data that needs an archival home?

In the very generation when the rate of language death is at its peak, we have chosen to use moribund technologies and to create endangered data.

—Bird and Simons

Only some cultures erect grand built monuments by which we can remember their achievements. But all cultures encode their genius in their languages, stories, and lexicons.

—K. David Harrison

TAPS: Checklist for responsible archiving of digital language resources

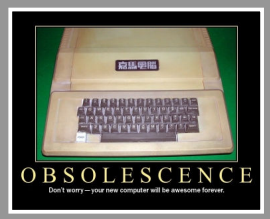
Debbie Chang
Graduate Institute of Applied Linguistics, Dallas, Texas

Top left: Magufa Minal (left) of the Bahinemo of the East Sepik Province, Papua New Guinea gives a story as he is recorded by Ian Tupper and Rachel Gray. Yashia Kwiba and village children are present. (Photo courtesy of Wayne and Sally Dye)

Top right: A few examples of the diverse formats of older, "legacy" materials await digitization at the Archive of Indigenous Languages of Latin America at the University of Texas, Austin.

Bottom left: Regarding the CD-eating fungus, Marc Valls, a biologist at Spain's National Centre for Biotechnology says, "Nature is very clever and, for all the materials that we design, sooner or later they will be degraded by some organism." (Screen capture from http://news.bbc.co.uk/2/hi/science/nature/1402533.stm)

Bottom right: All hardware and software are headed for eventual obsolescence. (From http://news.c-scribe.com/280)



TAPS (Target, Access, Preservation, and Sustainability): Checklist for Responsible Archiving of Digital Language Resources

Archive: Date: Representative:

	TARGET	Yes	No	Comments
1.	**Mission Statement**: Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?	☐	☐	
2.	**Submission Criteria**: Does the material that I want to submit fall within the scope of the archive's collection policy in terms of content and type (specify _____)?	☐	☐	
3.	**Designated Communities**: Is my desired audience (specify _____) a good match for the groups of users the archive targets (e.g., language community, academic community, etc.)?	☐	☐	
4.	**Ongoing Relationship**: Does the archive accept the responsibility to interface with the language community as a provider community? (This could involve revenue sharing and interaction with the language community as owners of their own language development efforts.)	☐	☐	

	ACCESS	Yes	No	Comments
5.	**Discoverability**: Are the descriptive metadata for materials deposited at the archive searchable online? That is, the metadata is posted on the web and/or aggregated through participation in a service such as OLAC so that they are discoverable through Internet search engines (e.g., Google, Yahoo!, Bing, etc.)?	☐	☐	
6.	**Fixed Identifiers**: Does the archive assign a persistent identifier to each item among its digital holdings so that it can be referenced and located in perpetuity?	☐	☐	
7.	**Reach**: Will the audience that I wish to reach (specify _____) be able to access the materials once they are deposited in the archive?	☐	☐	
8.	**Access and Use Restrictions**: Does the archive have policies and procedures to ensure that any restrictions I or the provider community place on access to the materials will be honored?	☐	☐	

	PRESERVATION	Yes	No	Comments
9.	**Evidence of Long-Term Planning**: Does the archive adhere to written policies and procedures for the long-term preservation of digital materials (e.g., the archive has written standards for implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?	☐	☐	
10.	**Preservation Strategies**: Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?	☐	☐	
11.	**Integrity**: Does the archive use fixity metadata to ensure that copies of digital materials will be complete and unchanged (e.g., a checksum, or digital signature, etc.)?	☐	☐	
12.	**Authenticity**: Does the archive ensure that digital materials contain what they claim to contain (e.g., by verifying that digital materials are what the metadata say they are, by permanently associating adequate metadata, and by faithfully maintaining provenance metadata to document any changes to the digital holdings)?	☐	☐	

	SUSTAINABILITY	Yes	No
13.	**Adequate Infrastructure**: Does the archive appear to be adequately staffed (in terms of numbers of staff and skill sets of the staff) and have the technical infrastructure to ensure continuing maintenance and security of materials (e.g., quality media, environmentally-controlled storage, access-controlled storage area)?	☐	☐
14.	**Financial Sustainability**: Does the archive appear to have secured sources of long-term funding?	☐	☐
15.	**Disaster Preparedness**: Is the archive engaged in responsible backup practices and prepared to recover its digital holdings in case of disaster (e.g., disaster recovery plan, offsite storage of backups)?	☐	☐
16.	**Succession Plan**: Does the archive have a reasonable succession plan to ensure that materials will be accessible and preserved elsewhere if the archive ceases to exist?	☐	☐

Yes = best practice
No = not in scope of archive / unclear

The state of the art in language archiving

NAME OF ARCHIVE	TARGET (out of 4)	ACCESS (out of 4)	PRESERVATION (out of 4)	SUSTAINABILITY (out of 4)	TOTAL SCORE (out of 16)
Archive of Indigenous Languages of Latin America (AILLA) University of Texas at Austin	3	3	1	2	9
American Philosophical Society: A Digital Archive for Recordings of Endangered Native American Languages Philadelphia, Pennsylvania	4	4	3	3	14
Endangered Languages Archive (ELAR) School of Oriental and African Studies, University of London, UK	3	3	2	1	9
Kaipuleohone University of Hawai'i at Mānoa	2	3	1	0	6
The Language Archive (TLA)/ Dokumentation Bedrohter Sprachen (DoDeS) collection Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands	4	4	4	4	16
Leipzig Endangered Languages Archive (LELA) Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany	3	1	1	2	7
Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) University of Sydney, University of Melbourne, and Australian National University, Canberra, Australia	4	3	2	1	10
Sam Noble Oklahoma Museum of Natural History (SNOMNH): Department of Native American Languages (NAL) Oklahoma University, Norman, Oklahoma	4	2	3	2	11
SIL International Language and Culture Archives SIL International, Dallas, Texas	3	3	2	2	10
The Smithsonian National Anthropological Archives (NAA) Washington, DC	3	3	4	4	14
AVERAGE SCORE	3.3	2.9	2.3	2.1	10.6

Relative strengths and weaknesses of digital language archives

TAPS CATEGORY	ITEM	ARCHIVES FOLLOWING BEST PRACTICE
Target	1. Mission Statement	100%
Target	3. Designated Communities	100%
Access	8. Access and Use Restrictions	100%
Target	2. Submission Criteria	90%
Access	6. Fixed Identifiers	80%
Preservation	12. Authenticity	80%
Sustainability	15. Disaster Preparedness	70%
Access	5. Discoverability	60%
Preservation	11. Integrity	60%
Sustainability	13. Adequate Infrastructure	60%
Sustainability	14. Financial Sustainability	60%
Access	7. Reach	50%
Preservation	10. Preservation Strategies	50%
Target	4. Ongoing Relationship	40%
Preservation	9. Evidence of Long-term Planning	40%
Sustainability	16. Succession Plan	20%

Implications of the Digital Age

Technological innovation has many advantages, and is in many ways inescapable. However, digital data is fragile and subject to two main threats:

- Degradation of physical media
- Obsolescence of hardware and software

While digital data holds the promise of copying and sharing without loss, there is potential for "catastrophic" loss when they are not adequately preserved.

- According to a UNESCO study, about 80% of our recordings about languages and cultures are endangered and urgent measures are called for to safeguard this valuable data.
- Digital materials need curation from the beginning to be viable long-term.

Thesis at: http://www.gial.edu/theses/index.htm

Digital archives: best practices

Simply digitizing materials and making them available on a website does not guarantee long-term preservation and access. TAPS is based on tools and criteria developed by the professional archiving community to describe the required operations and implementation of trustworthy digital archives (see list of standards consulted). Work is proceeding nationally and internationally towards the goal of digital archive certification, but to date, few digital archives, and no digital language archives, have achieved formal certification.

The purpose of TAPS

The TAPS Checklist was developed to help linguists and other depositors choose an archival home for their language resources.

- Represents a distillation of best practices in digital archiving
- Addresses issues important to linguists and language communities regarding archiving language materials

TAPS is an acronym standing for the four parts of the checklist (Target, Access, Preservation, and Sustainability).

- Not intended to be an exhaustive inventory or for an outside "audit" of an archive
- Includes things a linguist can reasonably check out

Survey methodology

To survey the state of the art in digital language archiving, I used the TAPS checklist to conduct interviews with a sample of digital language archives. Between September 2009 and January 2011, I interviewed representatives from ten digital language archives located in Australia, Germany, the Netherlands, UK, and US. Half the interviews were conducted in person during site visits, and half were conducted through Skype calls. These archives represented both public and private institutions with a global or regional collections focus, restricted or unrestricted submitter policies, and purely digital or mixed holdings.

Each of the ten archives interviewed were scored on a binary scale for each item on the TAPS checklist (yes = 1, no = 0), for a maximum score of four on each section, and a total maximum score of sixteen.

Archiving standards consulted in developing TAPS

DANS. 2010. *Data Seal of Approval - Quality guidelines for digital research data in the Netherlands*. Second edition, ed. by Laurents Sesink, Rene van Horik, and Henk Harmsen. The Hague: Data Archiving and Networked Services. [http://www.datasealofapproval.org/sites/default/files/Data_Seal_of_Approval_1_4_3.pdf]

DCC and DPE. 2009. *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) Interactive: User Guide*. Edinburgh: Digital Curation Centre and Glasgow: Digital Preservation Europe. [http://www.dcc.ac.uk/docs/tools/DRAMBORA_Interactive_Manual.pdf]

ISO. 2003. *Space Data and Information Transfer Systems—Open Archival Information System—Reference Model* ISO 14721:2003. Geneva: International Organization for Standards. [http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683]

NESTOR. 2006. *Catalogue of Criteria for Trusted Digital Repositories Version 1 (draft for public comment)*. Frankfurt am Main: Network of Expertise in long-term STOrage and long-term availability of digital Resources. [http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf]

OCLC and CRL. 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist, Version 1.0*. Dublin, Ohio: Online Computer Library Center, Inc. and Chicago: The Center for Research Libraries. [http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf]

Conclusions

The digital language archives evaluated show that they are strongest in issues that linguists are most concerned about: addressing a specialized audience and concerns surrounding access. The archives are weakest, however, in technical and organizational issues that may impact the overall longevity of the archived material.

Acknowledgements

The work represented in this poster was funded by a grant from the National Science Foundation (BCS-0723864, *OLAC: Accessing the World's Language Resources*) awarded to the Graduate Institute of Applied Linguistics (GIAL) in Dallas, Texas. I am grateful for Gary Simons, my thesis advisor, and many individuals who gave generously of their time.