Graduate Institute of Applied Linguistics

Thesis Approval Sheet

This thesis, entitled

TAPS: Checklist for Responsible Archiving of Digital Language Resources

written by

Debbie Chang

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts

with a major in

Applied Linguistics

has been read and approved

by the undersigned members of the faculty

of the Graduate Institute of Applied Linguistics.

_____

Gary F. Simons (Mentor)

_____

Michael Cahill

_____

Stephen Parker

_____

Date Signed

Chapter 4: Development and Use of the TAPS Checklist

The tool introduced in this chapter, the TAPS (Target, Access, Preservation, and Sustainability) Checklist for Responsible Archiving of Digital Language Resources, is designed to assist linguists in evaluating digital archives. It is an application of the guidelines for digital archives presented in chapter 3, and is tailored to the interests of linguists and language communities. Section 4.1 describes the methodology for developing the checklist, sections 4.2 through 4.5 discuss each section of the TAPS Checklist in detail and serves as a "user's manual," and section 4.6 addresses the limitations of TAPS.

4.1 Methodology

The procedures used for developing and testing the TAPS Checklist are described in Section 4.1.1. The process owes much to the generosity of the individuals and archives involved. Section 4.1.2 describes the uses and scoring of the TAPS Checklist.

4.1.1 The Development of the TAPS Checklist

The TAPS Checklist was formulated by the author of this thesis through a comparison of components common to trustworthy archives as enumerated in three different tools listed in chapter 3: *TRAC: Criteria and Checklist* (OCLC and CRL 2007), *Catalogue of Criteria* (NESTOR 2006), and the *Data Seal of Approval*, version 1-3 (DANS 2008). I went through the items in each of these tools and grouped them in table

format. From that list, four major categories were identified as being most pertinent for choosing a trustworthy and appropriate digital archive from the perspective of a linguist as potential depositor: Target, Access, Preservation, and Sustainability. These categories form the acronym "TAPS." Within this framework, the original list was pared down to include the most essential archival functions that could be readily understood and investigated by a non-expert in digital archiving. These were formulated into four questions for each of the major topic areas to create the sixteen items of the TAPS Checklist.

In all, the TAPS Checklist went through fifteen versions during its development. The help of many individuals was invaluable. Throughout the process, drafts were submitted to my thesis advisor, Gary Simons, who was instrumental in guiding and honing the finished product. In the initial stages of development, the Checklist was reviewed by Wayne Dye and William Reiman, individuals who have prepared digital language documentation for archiving. The Checklist was revised based on their input to reflect the particular interests of linguists and language communities, and their understanding of important issues in digital archiving. An interview with Joan Spanne, a specialist in digital archiving, suggested rearrangement of several items and expanding others to include examples.

I conducted the first site visit in September 2009 at the SIL Language and Culture Archive in Dallas, Texas with the archive director, Jeremy Nordmoe, and archivist, Vurnell Cobbey. This visit identified items that benefitted from re-wording, and highlighted items which required personal input from the linguist. Spanne additionally

verified additions and corrections to this initial evaluation using TAPS. In the next site visit, Heidi Johnson at the Archive of Indigenous Languages of Latin America (AILLA) provided feedback and valuable insights regarding issues concerning access resulting in significant revisions to the Access section of the Checklist. These and other site visits I conducted—with Lydia Motyka of the Florida Digital Archive (FDA), which serves the public university libraries in Florida, and Mary S. Linn of the Division of Native American Languages (NAL) archive within the Sam Noble Oklahoma Museum of Natural History at University of Oklahoma—were crucial to determining metrics for TAPS. Dye also consulted as a linguist in using the Checklist to evaluate the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) at which he deposited digital language data. Bob Conrad generously tested the TAPS Checklist at two special collections in university libraries in which he deposited language materials. I conducted the remaining evaluations using TAPS through the Internet. Information on PARADISEC (in addition to that supplied by Dye), Kaipuleohone, and ELAR were gathered online at the archives' websites, and then real-time interviews were conducted. A Skype interview with Nick Thieberger, a key implementer for both the PARADISEC and Kaipuleohone archives, was instructive in the extent of the state of the art possible given limited resources. Finally, a Skype interview with David Nathan, the director of ELAR, underlined the need for archives to serve language communities well. Nathan additionally shared articles, published and in press, that were illustrative of problems and solutions in digital language archiving.

4.1.2 Uses of the TAPS Checklist

The TAPS Checklist is a consensus of archival best practices with a focus on what is practical for a linguist to investigate. In formulating the Checklist, efforts were made to identify and include key activities common to all trustworthy digital archives and issues important to linguistic data and research. These items are intended to be researched first in information published by the archive (*e.g.*, on its website). Gaps are then filled in by correspondence, phone call, or an onsite visit.

Using TAPS, a linguist should be able to discern an archive that is trustworthy from one that is not and eventually find the best archival home for his or her work. This is done by comparing archives to each other across the four categories of Target, Access, Preservation, and Sustainability, so as to determine the relative quality of each archive. TAPS may also be adapted as archives may also be compared item by item, or by a subset of items in the checklist.

Each of the four questions in the categories was rated on a three-point scale with the following metrics:

Yes = The archive appears to follow best practices. The strongest indicator of this is that the staff is following a written policy or procedure that conforms to best practices. (3 points)

? = The archive is in the planning stages of implementing best practices; or the archive partially follows best practices; or the archive is assuming that another entity to which they outsource follows best

practice in implementing the functions indicated, but cannot document it. (2 points)

No = The item is not in the scope of the archive; or the degree to which the archive follows best practice is unclear. (1 point)

No zero value was given in the metric. The absolute differences between the scores and the statistical significance of the scores remain the same whether the point scale begins at 1 or 0.

Since it does not require the specialized knowledge of a formal auditor, the purpose of the TAPS Checklist is not to establish conclusively that the archive is trustworthy (or not), but to establish whether the depositor feels that the archive can be trusted with respect to all the designated communities involved. If the linguist comes away from investigating an archive with significant doubts about its trustworthiness, then we can conclude that significant problems likely are present since they were discoverable by a non-specialist. In the case where the linguist comes away feeling satisfied after investigating an archive, there is always the possibility that there are detailed technical problems that cannot be uncovered by a non-specialist. Only a full formal audit could ultimately guard against that possibility, but TAPS will at minimum be able to help linguists think through issues concerning the digital archiving of their data and steer clear of the clearly untrustworthy archives. In addition to its use by linguists, it is hoped that TAPS will be useful to the archives themselves in identifying significant shortcomings and will contribute to improvement in their trustworthiness. All the archives that I interviewed received the resulting written evaluations (see appendix B) and were given

an opportunity to correct errors and omissions. Many of the individuals that I interviewed indicated that the TAPS Checklist was helpful for them to identify areas that their archives could improve upon.

Each of the sixteen items in the TAPS Checklist was worded to stand alone as much as possible without extensive explanatory notes. However, the following sections "unpack" the contents of the Checklist: the four sections of TAPS and the 16 individual questions are outlined and discussed below in sections 4.2 through 4.5. At the end of explanations for many of the main questions, more detailed questions are included to aid the linguist in probing for the answer to the main question.

4.2 Target

Target refers to the "fit" of the archive with regard to the data to be deposited and the needs of the identified designated communities. This section of the TAPS Checklist, in items 1 through 4 (section 4.2.1 through 4.2.4), addresses the archive's commitment to maintaining "digital objects" for communities identified by the linguist. The specific questions deal with mission statement, submission criteria, designated communities, and an ongoing relationship to the language community.

4.2.1 Item 1: Mission Statement

Does the archive have a mission statement that reflects a commitment to the long-term preservation of digital information?

A key characteristic of a trustworthy archive is an explicit mission statement that makes clear its intentions to preserve digital information for the long-term (DANS 2010:11, OCLC and CRL 2007:10). Digitization projects and websites may store and disseminate digital materials in the short-term, but they will not have the infrastructure behind them to guarantee long-term preservation of digital materials and cannot make such claims. This is what differentiates archives from non-archives. A linguist should be wary of any institution that offers to "archive" materials but makes no commitment to long-term preservation in its mission statement.

4.2.2 Item 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:_____)?

Submission criteria pertain to the match between the content and types of the materials you want to deposit and the content and formats of materials that an archive accepts. The linguist should specify the content and type of data he or she wishes to deposit. Since archives have different specializations, determining the scope of the archive's collection is important to finding a good fit for your materials.

Trustworthy digital archives accept digital objects from the producers based on defined criteria (NESTOR 2006:18, DSA 2009:11). Well-thought out collection policies and guidelines on accepted formats therefore will indicate a high degree of trustworthiness. Support and procedures for digitizing analog and hardcopies of materials should also be well-defined if that is a need for the depositor.

The fit of the archive in terms of the overall collection and kinds of data that it typically handles should be taken into account since a digital object in an "obscure" subject or an "atypical" format (even if it is an "archival" format) may not be as well preserved, supported for access, or maintained in machine readable formats in the long-run as would materials in areas of specialty for the archive. For such reasons, it is ideal for depositors to dialogue with potential archives regarding submission criteria (including file formats) before embarking on a language documentation project.

2. Submission Criteria In-depth questions: What is the content of material I want to deposit? What formats are they in? Does my content fall within the collection policy of the archive? Do my materials fall within the preferred submission formats of the archive?

4.2.3 Item 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.*)?*

The OAIS reference model defines a designated community as the group or groups to which the archive aims to make content in the archive accessible (CCSDS 2002:1-10). These are the potential consumers who should be able to understand a particular set of information, and may be composed of more than one user community (CCSDS 2002:1-10). For the purposes of our discussion, I reference the intended user community in the plural as designated communities. Thus, the linguistics community,

the larger scholarly community, the language community, and so forth can all be counted as designated communities.

In order to determine the desired designated community or communities, linguists should first consider which audiences would benefit most from the materials deposited in an archive and who should be able to access the data once it is submitted (the language community, the linguistics community, *etc.*). Linguists should then determine whether those groups are a match to those that the archive is committed to serve. As the archive will be responsible for maintaining services to its designated communities over time, it stands to reason that the changing needs of the desired user communities of the linguist are best served by archives that are already well-positioned to serve those communities.

3. Designated Communities In-depth questions: What are the desired user communities for the data I want to deposit? Do these fall within the designated communities of the archive? Particularly if the language community is an important user of the deposited materials, does the archive cater to that user community?

4.2.4 Item 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.)

Since it is expected that deposited materials will outlive the depositor who initially acts as a liaison for a language community, it is often desirable for the archive to

interact directly with the language community. If this seems important to the language community as a condition for depositing the materials, then the linguist must ensure that the archive is ready to accept this responsibility. Additionally, certain archives may be able to set up revenue sharing structures for the language community if there are any revenues to be generated from the use of deposited materials. While this may not seem a likely scenario with purely linguistic data, it is not hard to imagine with the licensing of recordings of music and cultural performances.

Several language archives routinely interact with and champion the causes of the language community. The Native American Languages (NAL) archive at the University of Oklahoma is actively involved in language revitalization efforts, grant writing for language communities, and is committed to prosecute those who profit improperly from archived materials. A case at PARADISEC illustrates a solution regarding potential revenues generated by archived materials; the case concerns Dye's work among the Bahinemo of Papua New Guinea. Though the language documentation gathered from the Bahinemo was unlikely to generate revenue, and Dye stressed to the village in which he worked that neither he nor the archive were going to make any money from having their materials archived, the possibility that authors or film producers or some other archive user might materially benefit from the archived recordings remained a point of concern for the community. Thus PARADISEC provided a way to share any potential income as part of its firm commitment to be fair to descendants of language group members. The archiving agreement specifies a local agency to which any royalties resulting from use of

archived materials should be paid. Additionally, Dye liberally shared with the community from funds he had been awarded to do the language documentation work.

4. Ongoing Relationship In-depth questions: What types of interaction do I anticipate needing to take place between the language community and the archive? Will the archive support these? Is potential revenue sharing an issue for my deposit? If so, will the archive offer this service?

4.3 Access

Access refers to the accessibility and usage of the data and corresponding metadata once materials are deposited. The TAPS Checklist addresses Access in questions five through eight concerning discoverability, fixed identifiers, reach, and access restrictions.

4.3.1 Item 5: Discoverability

Are the metadata for materials deposited at the archive searchable online? I.e. *posted on the web 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.*)?*

Once digital material is submitted and archived, it is advantageous for the materials to be discoverable over the Internet to reach the widest possible audience. Generally speaking, an archive's holdings do not have to be available as complete digital files online, but the catalog of what they contain, that is, the descriptive metadata, should

be online so that it is searchable and discoverable with ordinary Internet search engines; see section 2.2.4. If reaching the widest possible audience is deemed undesirable, the essential question is, “Does the archive provide adequate resource discovery opportunities to the designated community?”

Before materials can be cataloged and found on the Internet, sufficient descriptive metadata needs to be provided by the depositor. An archive should have guidelines and standards for this metadata. The quality of the metadata to be searched is another indicator of trustworthiness of the archive.

One way for an archive to ensure that search results for their holdings show up on the “first page” in a routine search on the Internet is to be a member of the Open Language Archives Community (OLAC), which aggregates metadata from all the participating archives into a combined catalog at the OLAC website. This aids the discoverability of materials deposited in both prominent and less prominent archives.

5. Discoverability
In-depth questions: Does the archive have the necessary guidelines and standards to help me, as the depositor, provide quality descriptive metadata? Is the metadata posted on the Internet? If so, are these records easily found on the Internet? Is the metadata aggregated through a service such as OLAC? Does my desired designated community have adequate resource discovery opportunities through the archive’s approach to descriptive metadata?

4.3.2 Item 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?

The archive should have a system to assign externally visible and standardized persistent identifiers to materials in order to enable reliable referencing in academic citations and to ensure that they can be found in the distant future (NESTOR 2006:23, CCSDS 2002:2-6, Bird and Simons 2003:65-66). Each persistent identifier should be assigned permanently and remain unique within the system. The persistent identifier should not be based on any changeable attribute of the material being referenced. For instance, in the Digital Object Identifier (DOI)[1] system used in commercial publishing, a "dumb number" that is not based on any pattern avoids misleading assumptions or loss of meaning over time or across linguistic or cultural barriers. Another example is the Handle System[2] that is used in open-source repository systems like DSpace to assign persistent identifiers.

4.3.3 Item 7: Reach

Will the audience that I wish to reach (specify:_____) be able to access the materials once they are deposited in the archive?

The archive should communicate in advance and in a transparent manner its conditions of access and any costs that may arise. The linguist in turn needs to determine what constitutes reasonable access for the designated community and how that matches

[1] [http://doi.org/]
[2] [http://www.handle.net/]

up with the archive's policies on how materials will be accessed. Accessing the materials may include, but is not limited to, any or all of the following (NESTOR 2006:9):

- Accessing the materials at a given access point (*e.g.*, the access station pictured in figure 4.1)
- Creating or supplying an analog copy (*e.g.*, a print-out or a print-on-demand service)
- Creating or supplying a digital copy (*e.g.*, e-mail delivery or download by the user)
- Creating interfaces to permit online exploration or query of the materials.

7. Reach

In-depth questions: Will members of the designated communities be expected to have access to the Internet? Will members of the designated communities need to maintain an e-mail address? Will the metadata be available in English only, or will it be available in another language that is more accessible to the designated community? Will the archive charge fees for copies of data on media that are usable by members of the designated communities? Even if the fees are "at cost," will they be affordable for those communities?

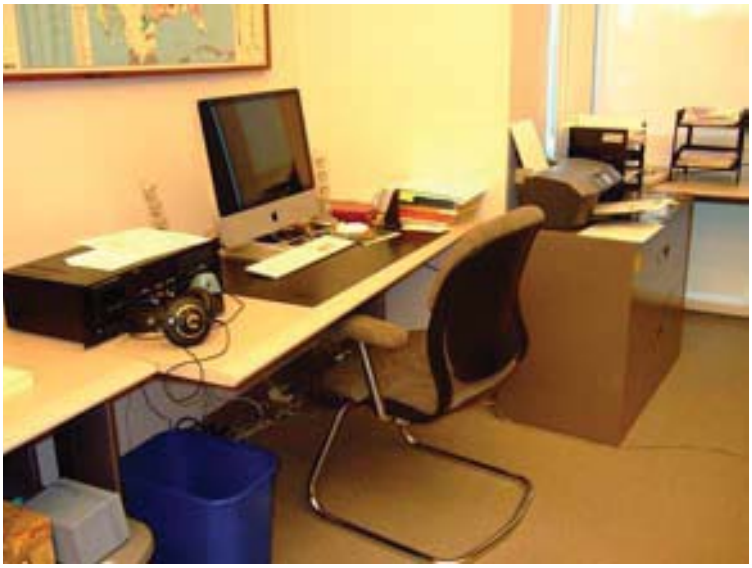

Figure 4.1: An Access Station at NAL. The Division of Native American Languages (NAL) at the Sam Noble Oklahoma Museum of Natural History has designated spaces where persons may access language materials.

4.3.4 Item 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?

The policies and procedures of the archive should articulate usage rights and conditions and their enforcement. The policies should make clear what options are available for open access and for restricted access, and then the linguist needs to ensure that one of those options matches the current and anticipated future needs. The issues surrounding ownership, copyrights, and conditions on the use of deposited materials should also be weighed carefully. The needs of the designated community should be evaluated for compatibility with what the archive will and will not do once the materials are part of the archive's collection.

8. Access and Use Restrictions

In-depth questions: How does the archive deal with copyright? Does the archive require transfer of ownership? Does the archive allow materials to be deposited with restrictions on access? If so, what restrictions are possible and how are requests for access handled? Does the archive allow materials that are closed to access to be deposited? How long will periods of closed access last? What are the archive's conditions of use policies?

When materials are deposited in an archive, a contract that governs the use of the material should be signed by the institution and the depositor. This contract, or deposit agreement, typically takes into account issues of ownership and copyright, access restrictions, and conditions on the use of deposited materials. These issues are described in greater depth in sections 4.3.4.1 through 4.3.4.3.

4.3.4.1 Copyright and Transfer of Ownership

Depositing materials at an archive that makes its holdings freely available on the Internet is essentially publishing them through the archive (*e.g.*, PARADISEC, AILLA). Because of the inherent copyright in performances, one cannot actually do this without the informed consent of the performers to allow this kind of distribution. Once materials are deposited, unless other arrangements are made, the archive decides over questions of access to and use of the materials. These rights are not necessarily exclusive, however

(*e.g.*, AILLA[3]). Depositing in an archive at minimum involves grant of license to reproduce (a necessary condition for many preservation functions) and distribute. At the Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS) archive, copyright and responsibility for archived material may be retained by the depositor or be completely conferred upon the archive (see section 2.2.2).[4] Ultimately, as AIATSIS and PARADISEC make clear in their deposit forms,[5, 6] the archive is responsible to adhere to procedures that "safeguard the interests and sensitivities of relevant indigenous people".[7]

It is debatable whether the traditional knowledge and stories of a people can be copyrighted, or whether such works should be treated as facts or ideas, which cannot be copyrighted.[8] In either case, it is a given person's performance of a work that is protected by copyright. However, under international and U.S. copyright laws, no work is protected in perpetuity though a language community may feel differently about their ownership of their traditional cultural heritage (SEM 2001:16-17). In AIATSIS and PARADISEC's deposit forms, the depositor is required to list "relevant individual(s) and their community(ies) and/or other funding organizations" that may have rights to the material being deposited. Furthermore, in the "explanatory notes" attached to these two Australian archives' deposit forms, similar statements carefully note that "the term ownership refers to ownership of the physical copy of the material being lodged with [the

[3] [http://www.ailla.utexas.org/site/ipr.html]
[4] [http://www.aiatsis.gov.au/collections/docs/AVA_deposit.pdf], [http://www.aiatsis.gov.au/collections/docs/AVA_transfer.pdf]
[5] The AIATSIS deposit form is available at: [http://www.aiatsis.gov.au/collections/docs/AVA_deposit.pdf]
[6] The PARADISEC deposit form is available at: [http://www.paradisec.org.au/PDSCdeposit.pdf]
[7] [http://www.aiatsis.gov.au/collections/docs/AVA_deposit.pdf], [http://www.paradisec.org.au/PDSCdeposit.pdf]
[8] [http://www.ailla.utexas.org/site/ipr.html]

archive]. It is not a wider claim to the intellectual property or ownership of any traditional knowledge."[9] PARADISEC's notes go on to say:

> If the material was written, photographed, drawn, recorded or filmed by you, then you are the creator and owner of the physical copy of the material, or if you have collected, found or inherited the material you are the owner of the physical copy of the material and therefore you or your delegate are in a legal position to enter this agreement.

Much has been written about copyright issues, and each country has its own copyright laws for which the linguist is responsible. See sections 2.2.3 through 2.2.5 for a more in-depth discussion. More information may be found at AILLA's webpage on Intellectual Property Rights[10] and the included links, or the Society of Ethnomusicology's *Manual for Documentation and Fieldwork & Preservation, Chapter 2, Ethical and Legal Considerations* (SEM 2000) for details concerning copyright law.

4.3.4.2 Access Restrictions

Varying levels of access to viewing and listening to materials is desirable in some cases to preserve privacy and confidentiality. Though AILLA encourages all depositors to make their resources freely available, its Graded Access System is an example of providing flexibility and leaving it up to depositors to specify restrictions on use.[11] AILLA provides the choice of four access levels to depositors who can choose to assign

[9] [http://www.paradisec.org.au/PDSCdeposit.pdf], [http://www.aiatsis.gov.au/collections/docs/AVA_deposit.pdf]
[10] [http://www.ailla.utexas.org/site/ipr.html]
[11] [http://www.ailla.utexas.org/site/gas.html]

any level to their entire collection or to any part of their collection: Level 1, access is open; Level 2, access is protected by password; Level 3, access is protected by a time limit, and Level 4, the depositor (or someone else) controls access to the resource.[12]

With AIATSIS and PARADISEC, inquiry is made in their deposit forms regarding the depositor's "understanding of the attitude of the [language community towards] … this material being made accessible to other people," and "whether any special conditions should be considered when handling this material, for example, ceremonial or gender restricted material, sensitive genealogical material, photographs or recordings of deceased people" so that the archive may act accordingly. Both archives further offer depositor specified conditions on access, but AIATSIS reserves the right to refuse material which has unreasonable conditions, and PARADISEC will not hold material on permanent closed access.

NAL has a restricted-use policy which allows open access to most materials (*i.e.*, access is not based on tribal membership), but will permanently deny public access to portions of materials containing "injurious gossip." Materials pertaining to formal societies (*e.g.*, Kiowa Black Leggings) and chief societies (*i.e.*, men who are born into a chief family line) can also be restricted, though such sacred or sensitive content is officially "on loan" to the museum according to NAL's *Restrictions on Use Policy*.[13] AILLA recommends that materials of extremely sensitive nature not be deposited.[14]

---

[12] [http://www.ailla.utexas.org/site/forms/ailla_depositor_packet.pdf]
[13] [http://www.snomnh.ou.edu/collections-research/cr-sub/nal/restrictions%20policy.pdf]
[14] [http://www.ailla.utexas.org/site/five_con.html]

4.3.4.3 Conditions of Use

Conditions of use have to do with what users of archived materials are allowed to do with them. NAL's *Restrictions on Use Policy* prohibits commercial or for-profit use of collected materials, and states the archive's commitment to prosecute for the improper use of deposited materials. AIATSIS provides its depositors open-ended choices, which include a choice for the depositor to be contacted each time material is copied. PARADISEC and AILLA outline responsibilities of and limitations on the user in their "Conditions of Access" and "Conditions for Use of Archive Resources" agreements.[15] Users are not authorized to access the archives until they have read and signed these agreements.

4.4 Preservation

Preservation refers to the overall system and technical structures of the archive that ensure materials will be managed in ways that make them available and usable, with their authenticity and integrity intact, far into the future. The TAPS Checklist addresses Preservation in questions 9 through 12 concerning evidence of long-term planning, preservation strategies, integrity, and authenticity.

4.4.1 Item 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*

[15] [http://www.paradisec.org.au/PDSCaccess.pdf], [http://www.ailla.utexas.org/site/use_conditions.html]

implementation and is engaged in formal, periodic review and assessment that responds to technological developments and evolving requirements)?

At the heart of any archive is the plan and implementation of a defined archival process that is sustainable over time (NESTOR 2006:20). At the highest level of planning for the long-term preservation of digital materials, the archive should plan to take into account legal and social changes, the needs and expectations of the designated communities, and technological developments relevant to the preservation and appropriate use of the deposited materials (NESTOR 2006:14). The day-to-day operations of the archive include the definition of digital objects "packaged" in a defined structure for long-term preservation (*i.e.*, content data in a suitable archival format, information needed to interpret the content data, and the relevant metadata). For digital language archives, the structure of complex objects, such as multimedia materials, needs to be adequately described so that they can be reconstructed and used as intended (NESTOR 2006:25). The procedures for creating and maintaining these "archival information packages" should be documented with written policies and procedures (NESTOR 2006:20, CCSDS 2002:1-7). The responsibility for each process may be assigned to particular individuals (NESTOR 2006:18) or to outsourced entities (NESTOR 2006:13). Trustworthy archives have a demonstrable commitment to the archival storage of digital materials to defined specifications, and will regularly review the appropriateness of those specifications over time (DANS 2010:11, NESTOR 2006:12).

9. Evidence of Long-term Planning

In-depth questions: Does the archive have written procedures for the tasks involved in implementing their defined archival process? Do these procedures specify deadlines for completing upcoming tasks as they pertain to the creation and maintenance of archival information packages? Is responsibility for each process clearly assigned to specific individuals or outsourced entities? Is the archive explicitly monitoring substantial changes, whether technical, organizational, or community-based? Will the archive change its procedures as needed?

4.4.2 Item 10: Preservation Strategies

Will the archive refresh and update digital materials as needed to counter obsolescence of hardware and software over time?

A trustworthy archive has an overall strategy for preserving digital materials within their collection. The monitoring of technical developments noted above in section 4.4.1 includes the development and standardization of new file formats and new storage techniques and the phasing out of existing technologies as needed. The archive should keep pace with ongoing technical developments (such as changes to data carriers, data formats, and user demands), but even in the absence of such changes, the archive must have a plan to deal with the deterioration of the media on which the data is recorded, sometimes called “bit rot” or “digital decay,” in which the bits of data themselves are subject to corruption over time.

In order to carry out such responsibilities, the digital repository must identify which characteristics of the digital objects are significant for information preservation (NESTOR 2006:19). The process should be defined to determine for each item archived whether a maintenance measure must be undertaken to ensure long-term preservation, and when needed, the corresponding measure should be carried out and any changes to the digital object documented (see section 4.4.4 below on Authenticity) (NESTOR 2006:21, 25). Two long-term preservation measures are:

- Refreshing: the transfer of data from one medium to the same type of medium without any alteration to the data at the bit-level. Refreshing guards against the deterioration of physical media, but not obsolescence.
- Data migration: transferring files to a newer format (for example in 2001, from JPEG to JPEG 2000), when software or hardware required to read the data is no longer available. Data migration can be a time-consuming process, involving alterations to the data, and often sacrificing an element of the 'look and feel' of the original material.

Even when the migration strategy involves changing format, bit-level preservation of originals seems to be emerging as a best practice in the digital archiving community (Caplan 2004:6). Keeping the original file as it was submitted aids in demonstrating the integrity (see section 4.4.3 below) and authenticity (see section 4.4.4 below) of the digital objects. Additionally, there is always the possibility that a better migration algorithm may arise; if the original file are always retained, a "do-over" of the migration is always possible (Caplan 2004:6). No archive that I interviewed had conducted a full-scale

migration as the need had not arisen. The FDA had conducted proof-of-concept migrations, however.

10. Preservation Strategies In-depth questions: On what medium will the archive store the materials I submit? What is their schedule for refreshing data on that medium? What will the archive do with the data as the medium approaches obsolescence? What will the archive do if the format in which the data are stored becomes obsolete? When was the last time the archive completed a migration from an obsolete format to a newer format?

4.4.3 Item 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.*)?*

The archive should ensure the integrity of the digital materials and metadata throughout their lifecycle within the archive (as they are processed, stored, copied, and used). Here, integrity refers to (1) the completeness of the digital object, including metadata, and (2) the exclusion of unintended modifications as defined in the preservation rules. Integrity is measured in terms of the characteristics of the particular digital material being preserved (NESTOR 2006:41). Inappropriate modifications may be caused by human error (deliberate or accidental), imperfections in media, or damage to the technical infrastructure. The archive should take both organizational and technical precautions to secure the integrity of objects within their custody; that is, the archive

should operate a data management system that is able to ensure integrity of digital materials (NESTOR 2006:15). Best practice is to use fixity metadata, like checksums and digital signatures, to ensure the integrity of copies. The use of fixity metadata reflects an archive's institutional commitment to the integrity of digital materials, and is also an indicator of the quality of its archival implementation.

A checksum is created using an algorithm that adds all the bytes or words in an arbitrary block of data to create a value that is stored as part of the fixity metadata of the digital object. When data is transmitted or copied, the checksum is recomputed and compared to the checksum value stored in the metadata in order to detect an error. If the checksums match, it is unlikely that there was an error in transmission or copying, though it is possible that some pattern of altered bits in a message can result in an erroneously matching checksum value (Maxino 2006:1). A good checksum algorithm will yield a different result with high probability when the data is accidentally corrupted; thus, if the checksums match, the data is very likely to be free of accidental errors.[16] There are tradeoffs, however, between the computing power used on the checksum calculation, the size of the block of data checked, and the probability of such undetected errors (Maxino 2006:1).

Digital signatures are typically used to verify authenticity, which is the process of determining if a user or entity is who he, she, or it claims to be (OWASP 2002), but they also serve the purpose of simultaneously providing integrity over the signed data. This is a consequence of a necessary property of cryptographic hash algorithms and signature

[16] [http://en.wikipedia.org/wiki/Checksum]

algorithms as any change in the input data leads to a large, unpredictable change in the output with very high probability. In other words, if the data has changed, the signature will fail to verify, and the loss of integrity will be obvious. If, on the other hand, the signature verifies, the digital object is likely unaltered (Adams and Lloyd 1999).

4.4.4 Item 12: Authenticity

*Does the archive ensure that digital materials contain what they claim to contain (*e.g.*, by verifying that digital objects 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 objects that occur while they are in the care of the archive)?*

The archive should ensure the authenticity of digital materials and metadata throughout their lifecycle within the archive (as they are processed, stored, and used). Authentic here means that a digital object actually contains what the metadata claims that it contains. When authenticity cannot be demonstrated for a particular holding, the archive should document this fact in the metadata.

Figure 4.2: Materials Being Checked for Authenticity at NAL. OU graduate student, Amber Neely, listens to Kiowa language materials at NAL, checking the authenticity and noting any discrepancies.

After authenticity is verified in the initial deposit, it can be preserved through permanently associating adequate metadata so that the match between deposited materials and associated metadata can be verified, and using provenance metadata to document the origins and all changes to the materials and metadata (NESTOR 2006:17, 25). In language archives, the depositor may be solely responsible for vouching for the authenticity of deposited materials.

Provenance metadata should contain information about how the digital objects came about, and careful records of the outcome of preservation processes. In cases where material is migrated to new formats, users must understand which versions of a particular digital resource are available for access, and how the resources have been changed as a consequence of preservation (Lavoie and Dempsey 2004).

4.5 Sustainability

Sustainability refers to the demonstrated organizational robustness of the archive, lending long-range viability to the functions that it performs. The TAPS Checklist addresses sustainability in questions 13 through 16 concerning adequate infrastructure, financial sustainability, disaster preparedness, and succession planning.

4.5.1 Item 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)?*

Adequate infrastructure addresses two aspects of the archive: the staff and the technical infrastructure.

Staff: The qualifications and training of the staff should be adequate to the defined processes and mission of the archive (NESTOR 2006:12). Staff numbers should be sufficient to fully complete the tasks of the archive (OCLC and CRL 2007:11). Additionally, there should be programs to ensure adequate professional development of staff over the long-term.

Technical Infrastructure: The technical infrastructure of the archive should ensure the continuing maintenance and security of its digital objects. This infrastructure includes good overall computing practices described by international management standards, for example, ISO 27002, formerly ISO 17799 (OCLC and CRL 2007:43). It is

recognized that "without a secure and trusted infrastructure, the functions carried out on [archived materials] cannot be trusted," and such an archive would be "built on a house of cards" (TRAC 2007:43).[17] Responsible back-up procedures are included in section 4.5.3, Disaster Preparedness, but since it is likely beyond the scope of an informal interview to check the computing practices of a given archive, the TAPS question lists more tangible indicators of the quality of technical infrastructure as possible items to check. "Quality media" refers to the physical media on which data is stored; for example, hard disks are more durable and less prone to failure than CDs or DVDs. "Environmentally-controlled storage" refers to the physical environment in which physical copies of materials are stored, including temperature, humidity and pest controls. And "access-controlled storage area" indicates that digital and physical copies of materials are protected from misuse or theft by virtue of the security in the facilities in which they are kept.

4.5.2 Item 14: Financial Sustainability

Does the archive appear to have secured sources of long-term funding?

The archive should be able to demonstrate its financial sustainability. Though an archive may not be a for-profit business, it should adhere to good business practices and should have a plan for how it will "stay in business." The business plan comprises a set of documents that lays out the past, present, and future of the repository and its activities, and which takes into account the financial implications related to development and

[17] The requirements for an adequate technical infrastructure as it applies to digital archives are laid out in Section C of the *TRAC Criteria and Checklist.*

normal production activities, and may note factors that would affect operations. It is recommended that the business plan and financial fitness be reviewed at least annually (TRAC 2007:16). The digital repository should be able to demonstrate that the proposed services can be financed, both in the short and long term, whether it is on the basis of guaranteed funding or on the basis of charging for use of its services (NESTOR 2006:11).

4.5.3 Item 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)?*

The archive should ensure that it has adequate hardware and software support for backup functionality that is sufficient for the services it provides and for its digital holdings. The following can demonstrate the adequacy of the processes, hardware, and software of an archive's backup systems: documentation of what is being backed up and how often; audit log of backups; validation of completed backups; "firedrills"—testing of backups; support contracts for hardware and software for backup mechanisms (TRAC 2007:44-45). Another important requirement is that backups be stored in a different physical location than the archive itself in order to mitigate the risk of fire, flood, tornado, and other disasters that could befall the building that houses an archive. The existence of (and long distances between) "mirror" sites also lend confidence to an archive's disaster preparedness.

In conjunction with responsible backup practices, the archive should have a written plan regarding what happens in specific types of disaster (fire, flood, earthquake, explosion, system compromise, *etc.*), and who has responsibility for which actions (TRAC 2007:49). Disaster with respect to digital archives is defined as "any event that threatens or interrupts the operation of the repository and that, without corrective action, threatens the long-term preservation of its holdings" (TRAC 2007:81). The level of detail in a disaster plan, and the specific risks addressed, are determined by the location and expected services of the archive. The disaster plan should also deal with specific consequences arising from unspecified situations, such as lack of access to a building or prolonged network outages. The archive should keep written disaster preparedness and recovery plans, including at least one off-site backup of all preserved information together with an off-site copy of the recovery plans (TRAC 2007:49).

15. Disaster Preparedness

In-depth questions: Is there a formally documented procedure for regular backups? Is compliance with the procedures audited? Where are the backups of archived material kept? If the building or location housing the archive is destroyed, how will materials be recovered? Is there a written disaster recovery plan, located off-site, that makes explicit what to do if a disaster occurs?

4.5.4 Item 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?

The archive should ensure the continuation of the preservation tasks if the archive itself ceases to exist. In order to avoid irreparable loss, consideration needs to be given to this responsibility while the archive and its holdings are viable, not when a crisis is occurring (TRAC 2007:10). To this end, the archive ideally should have a formal succession plan that includes trusted inheritors (TRAC 2007:10). Succession plans should describe processes that will enable the preservation work to continue within an alternative organizational framework, thereby ensuring that the requirements can continue to be completed; where this is not possible, any restrictions should be documented (NESTOR 2006:12).

If a formal succession plan is not in place, the archive should at minimum be able to identify the basis of a plan, for example, partners, commitment statements, likely heirs, and so forth. Succession plans do not need to transfer the entire collection to a single organization if this is not feasible. Multiple inheritors are acceptable as long as the data remains accessible (TRAC 2007:10).

It should be noted that, organizationally, the materials in an archive can be at risk whether the archive is run by a commercial organization or a government entity (*e.g.*, national library or archives):

> At government-managed repositories and archives, a change in government that significantly alters the funding, mission, collecting scope, or staffing of the institution may put the data at risk. These risks are similar to those faced by commercial and research based repositories and

> should minimally be addressed by succession plans for significant collections within the greater repository (TRAC 2007:10).

4.6 Limitations of the TAPS Checklist

This checklist is not a comprehensive tool and is not intended to be used to perform an outside audit of a given archive. Instead, a high degree of trust is placed on the self-reporting of the archives on their practices with regard to the items pinpointed in the checklist. The criteria contained in the TAPS Checklist are not exhaustive at sixteen items,[18] but they are essential to the trustworthiness of digital language archives and concerns of linguists and language communities. Table 4.1 shows how the sixteen items of the TAPS Checklist align with "ten basic characteristics of digital preservation repositories" [19] identified by four preservation organizations that convened in 2007 in Chicago under the auspices of the Center for Research Libraries. Note that some TAPS items are listed more than once in the table. Item 4, ongoing relationship, is the only item that does not appear as it pertains to the rights of language communities, which are not addressed by general digital archiving standards. The preservation organizations were: the Digital Curation Center (U.K.) and Digital Preservation Europe which created DRAMBORA, NESTOR (Germany) which created the *Catalogue of Criteria,* and the CRL (international consortium based in North America) which created TRAC.

---

[18] TRAC, the most extensive of the auditing tools described in chapter 3 with 114 checklist items, refers to itself as a "starting point" and not an all-inclusive checklist.

[19] [http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re]

Table 4.1: Distribution of TAPS Checklist Items among Ten Basic Characteristics of Digital Preservation Repositories (material from [http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re] and appendix A)

Ten Basic Characteristics of Digital Preservation Repositories	TAPS Checklist
1. The repository commits to continuing maintenance of digital objects for identified community/communities.	1. Mission Statement 3. Designated Communities
2. Demonstrates organizational fitness (including financial, staffing structure, and processes) to fulfill its commitment.	14. Financial Sustainability 13. Adequate Infrastructure 9. Evidence of Long-Term Planning 10. Preservation Strategies
3. Acquires and maintains requisite contractual and legal rights and fulfills responsibilities.	8. Access and Use Restrictions
4. Has an effective and efficient policy framework.	9. Evidence of Long-Term Planning 15. Disaster Preparedness 16. Succession Plan
5. Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.	2. Submission Criteria 9. Evidence of Long-Term Planning
6. Maintains/ensures the (a) integrity, (b) authenticity and (c) usability of digital objects it holds over time.	11. Integrity 12. Authenticity 10. Preservation Strategies
7. Creates and maintains requisite metadata about (a) actions taken on digital objects during preservation as well as about (b) the relevant production, access support, and usage process contexts before preservation.	9. Evidence of Long-Term Planning 11. Integrity
8. Fulfills requisite dissemination requirements.	5. Discoverability 6. Fixed Identifiers 7. Reach 8. Access Restrictions
9. Has a strategic program for preservation planning and action.	9. Evidence of Long-Term Planning 10. Preservation Strategies
10. Has technical infrastructure adequate to continuing maintenance and security of its digital objects.	13. Adequate Infrastructure