Introducing the ELAR information system architecture

Robert Munro and David Nathan
Endangered Languages Archive (ELAR),
School of Oriental and African Studies

1 Introduction

The Endangered Languages Archive (ELAR) is part of the Hans Rausing Endangered Languages Project (HRELP), based at the School of Oriental and African Studies (SOAS), University of London.

HRELP was set up in 2003 with a donation of £20 million from the Lisbet Rausing Charitable Fund, with the goal of documenting as many endangered languages as possible. HRELP consists of three programmes, ELAR, and:

- The Endangered Languages Academic Programme (ELAP). ELAP runs PhD and Masters courses in language documentation. There have been about 40 Masters students since 2003, and 10 PhD students. ELAP also runs regular seminars and workshops.

- The Endangered Languages Documentation Programme (ELDP) is primarily a funding body, offering up to £1million in grants each year for the documentation of endangered languages. It also runs training courses for grantees, although most of the teaching for the training is undertaken by ELAP and ELAR staff.

1.1 ELAR's current state

ELAR will archive the materials of all documentation projects funded by ELDP, as well as endangered languages materials from depositors anywhere in the world. ELAR recently opened, accepting deposits since 28th October 2005.[1] As well as archiving, we are active in other tasks including teaching, publishing, and multimedia development for endangered languages.

ELAR has two full-time staff: David Nathan, the archivist; and Robert Munro, the software developer. They both have backgrounds in diverse areas of IT and linguistics. The archive also has two half-time supporting staff, and we are frequently helped by casual staff including student helpers and various specialists.

Currently, we are in the process of designing and implementing key systems:

1. Accession system (ingestion system). This is mostly complete, but some manual systems will be become automated or semi-automated. Some specialist equipment, such as audio-tec's Dobbin, is soon to be installed.

2. Archive information system. This is the internal architecture of the archive information system, including the storage of data and structures supporting the materials, metadata and the relationships between them. The systems analysis for this system is complete; it is currently in the design phase and will be rolled-out from December 2005.

3. Catalogue serving system. This will serve our catalogue for metadata harvesting. The systems analysis and design for this system are complete. We intend to serve our catalogue in OLAC, IMDI and OAI compliant metadata. It will be rolled out following the implementation of the archive information system.

---

[1] Because we are so recently opened and in the process of systems development, this paper is best read as a snap-shot of ELAR in its current state. We will publish more detailed technical specifications at a later date.

4. Archive access system. This will allow users to interact with the archive materials. It will manage access to data, access restrictions, and the processes for adding / updating information. The systems analysis for this system is complete. It is closely tied to the archive information system, so the two are being designed and rolled-out concurrently.

5. Data storage. We have installed a fast storage area network (SAN) with a capacity of 8 terabytes. It is backed up by a dedicated robotic tape library.

6. Long-term backup system. We have reached an agreement with the Oxford Text Archive (OTA) for the additional ingestion and cataloguing of our data. It includes strategies for both disaster recovery and for long-term preservation. We will work with OTA to define the appropriate structures and metadata for data interchange.

The long-term back-up system was negotiated by David Nathan. The infrastructure for the data storage was undertaken by contractors. The other four systems were/are designed through a formal systems analysis and design project with Robert Munro as the analyst, and David Nathan the primary client. The systems analysis and design drew on a wide variety of sources:

1. Literature review. Approximately 60 papers, technical reports and softwares were reviewed. The major points relating to the digital archiving of endangered languages were distilled from each paper, and the (often conflicting) conclusions of the papers were analysed. This allowed us to make a number of useful generalisations about archive structures, design and access patterns, and also highlighted areas where generalisations could not be made. The review document is approximately 100 pages in length. In addition, policies were iteratively refined with the archivist, who defined formal policies, at a fairly fine level of granularity, for all four systems. These policies ranged from preferred formats for accessing, to our designated communities for access to materials. Drawing on the literature review, the analyst commented on how these policies related to other archiving solutions, and made suggestions for changes where the implementation of policies was unnecessarily complicated, not required or conflicted with requirements of the designated communities.

2. Review of exemplar materials. The archivist reviewed examples of documentation materials from other archives. This allowed the analyst to understand the archivist's position on the role of metadata in documentation materials.

3. Interaction with associated archives. Our involvement in DELAMAN and DAM-LR has allowed us to discuss archive design and archive policy with a range of archivists and archive technical staff. We are especially thankful to Peter Wittenburg and Daan Broeder of the DOBES archive.

4. Interaction with ELDP grantees. The archivist and analyst both took part in the training of ELDP grantees. This allowed us to workshop some of our early ideas, and to get feedback from documenters on their interaction with other archives. For those that were members of endangered languages communities or worked with language development organisations, we were able to obtain information about how the archive could ideally serve these potential users. The ELDP grantees will also become depositors, so we were able to get an understanding of the nature of the materials that will be submitted, and to work with the grantees in order to make their working formats easily convertible to appropriate archive accession formats.

5. Interaction with members of ELAP. Our close relationship with ELAP staff and students gave us examples of requirements that documenters and EL community members will have of the archive, and of the materials they might wish to deposit.

6. Departmental seminars on language documentation. By taking part in the regular seminars on language documentation run by ELAP, we were able to get a clear idea of the variety of phenomena that documenters will record and wish to archive, and to make focussed queries with a view to understanding the architecture required to represent such phenomena in an archive.

7. Seminars focused on archiving. We ran a department seminar focussing on accessioning data. With language documenters, we workshopped our deposit forms, deposit requirements, deposit metadata requirements and accession policies. The acceptance of the outcomes of the seminar by both the documenters and the archive allowed us to launch our accession system.

2   The ELAR architecture

The ELAR architecture is strongly informed by the Open Archive Information System (OAIS) Reference Model (CCSDS, 2002). Figure 1 gives a top level model of the flow of information – note that producers and designated communities can be the same people in some instances. The model makes a clear distinction between the formats and structures appropriate for each of Ingestion (Accessioning), Archiving, and Dissemination. The Ingestion Packages are the formats and structures the archive will accept for deposit. These are designed to minimise the cost of conversion and allow incremental assembly of the deposit. There should be multiple Dissemination Packages, each designed to meet the needs of a Designated Community. The Archive architecture should be designed to support these Dissemination Packages, keeping in mind that their nature will evolve with changing user requirements. We believe that OAIS provides the most well-developed model for digital archives, and recommend others to also adopt its architecture and internal terminology, to make discussions and operations between archives as transparent as possible. The influential DSpace system (Bass et al, 2002a; Bass et al, 2002b) also draws on the OAIS model.
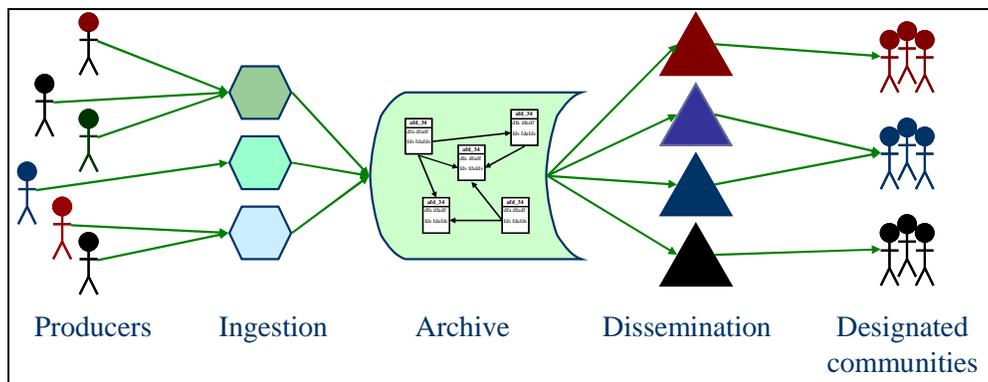

Figure 1 The OAIS reference model

3   User Requirements

JISC (2001) explored the different types of digital archives. ELAR will span the functions of a number of these archives, and so many of the user requirements we identify draw strongly on this study. However, there was no archive in JISC (2001) with a designated community that corresponded to that of 'member of endangered language community'.

We identify the following designated communities:

- Depositor
- Researcher
- Endangered language speaker
- Publisher - repurposing
- Long-term preserver
- Publisher –public heritage
- General Public

See the Appendix for a list of user requirements, examples and their implications for the data architecture of ELAR.

4 ELAR – ingestion (accession) formats

Suitable ingestion[2] formats have been well-discussed. Generally, those meeting the seven plus one dimensions of portability are acceptable (Simons and Bird, 2003; Johnson 2004).

ELAR will accept a range of formats, with a preference for the following:

- text - plain text, with or without markup
- documents - plain text, PDF or postscript
- structured text - XML, other markup (with description of markup system)
- structured data in commonly available Office formats - ELAR will convert them to archive-suitable formats
- character encoding
    - preferred encoding is ASCII or Unicode
    - any other encodings to be clearly documented
    - font substitutions to handle non-Roman characters may need to be converted
- sound - WAV
- image - BMP, TIFF, JPEG
- video - MPEG2

ELAR will recommend and support the incremental development of complete metadata descriptions. However, the following metadata elements are the minimum required for accession (drawing on OLAC terminology):

- Identifier: A means to uniquely identify each item in the submission information package. This might be either:
    - a unique name for each item listed together with the full filename (and media carrier label if relevant) or
    - a unique filename for each item
- Format: Describe formats
    - file format
    - mark-up format
    - character encoding format
- Creator: Entity primarily responsible for making the content
- Subject.language: The language(s) which is described or documented
- Language: The language in which the content is expressed or introduced.
- Rights: Information about rights held in and over the resource (ELAR will apply default values if required)

ELAR also recommends that the following metadata fields are provided at the time of submission:

- Title: A short name given to the resource
- Description: An account of the content of the resource
- Type: The nature or genre of the content of the resource

---

[2] In ELAR, we prefer the term 'accession' over 'ingestion' but use the OAIS reference model terms here. We also prefer 'deposit' over the OAIS's 'submission', and 'depositor' over 'producer', but have used all four terms here.

Additional metadata fields improve the description of the resource for various communities that may have an interest in it (such as linguists, language community members, educationalists etc). These can be provided at or after the time of deposit. Since language documentation and description is an ongoing process, some metadata will naturally be created, added, or updated after a resource has been deposited. ELAR recognises that changing depositor and user needs will require us to evolve and augment our metadata systems over time.

ELAR accepts metadata structured in several ways:

1. TABLE. As a table where each row represents one item (file), and each column (field) a metadata field. Each field must be documented or mapped to a OLAC or IMDI field;

2. XML format conforming to the OLAC metadata schema.

3. XML format conforming to the IMDI metadata schema.

4. RELATIONAL. A relational format, such as a relational database. The database structure should be documented, and each field documented or mapped to an OLAC or IMDI field.

Generally, a submission package should only be used as an archive package if the user community and the depositor community are one and the same. This is the classic view of an archive as secure backup, that is, a data bank (JISC, 2001). While this model is often used in linguistics for digitisation projects (JISC, 2001), these 'corpus repositories' are typically only interpretable by the data creators, and are therefore not portable (Bird and Simons, 2003).

With increasing awareness of the importance of portability and archiving, and the availability of linguistic tools supporting portable formats and structures (e.g. ELAN), submission packages will sometimes require very little effort in order to convert them to archive format. In the best case scenario, the archive can simply import the materials and verify that the files are uncorrupted. However, this scenario is rarely seen, and much of the depositor's 'working format' might be present in the submission, or information expressed by file naming or distribution in directories. ELAR does not store data using the filenames and directory structure of the submission packages, as these may be platform-dependent. However, naming and distribution of deposited files may provide meaningful information not formally captured in metadata, so we do record the submission filenames, directory paths and delivery media in the archive information system. This also allows the depositor to navigate their collection via the names and directories that they are familiar with, while the archive's underlying architecture can take advantage of the much richer structures that a database management system offers.

5 ELAR – archive and dissemination formats

ELAR's catalogue serving system is part of the dissemination strategy, serving our metadata in OLAC, IMDI, OAI and TEI formats.

For our archive access system, we are addressing these key areas:

- Granularity and relationships:
    o archive objects might consist of bundles - i.e. two or more objects – this requires sophisticated modelling of relationships
    o archive objects might consist of a fragment of material within a file – this requires modelling of various types of subsections / subsequence
    o the types of related materials and their relationships should play a part in the search options
- Version control:
    o modelling versions of materials is required
    o multiple types of versioning might be required (migration / dissemination / content update)
    o versions will be 'invisible' to most dissemination packages

- Adding materials and metadata:
  - users can add comments to data
  - users can add metadata values not provided by a depositor
  - users can make relationships between items, including mapping terms used in deposit materials to well-known schemas and ontologies
  - users can supplement the kinds of metadata and relationships in the archive. For example, if the archive did not have a metadata element representing 'Genre', a user could add it
  - **note:** all the above require moderation, and an architecture supporting both the process and the moderation
- Language support:
  - users should be able to add comments / metadata in any language – this is problematic for languages without a stable writing system
  - users should be able to navigate the archive access system via the language preference(s) of their choice
  - the archive architecture needs to support translations of metadata and comments – this requires unambiguous reference to values in metadata fields (not currently possible in OLAC or IMDI)
- Archive services
  - the archive will provide advice and conversion services to depositors
  - the archive will respond to users' requests for information. This could range from locating materials, to multimedia development and other one-off value-adding functions
  - the archive will needs to provide the means for individuals associated with the archive to communicate with each other, especially regarding access to materials

Note that while the archive architecture needs to be flexible and atomic, dissemination will typically involve a limited set of formats and integration of materials into suitable presentation formats. Therefore, while embedding metadata within files is not a good idea for archive format, as that metadata will change (Cedars, 2002), it is a good idea for dissemination formats as it can carry useful information (Barwick, 2003b; Thieberger, 2005; Adobe Systems, n.d.).

5.1 ELAR – database technologies

ELAR uses a MySQL database to support its archive information system. MySQL is the most widely used open source relational database, and is the most popular choice of database management system for the archives surveyed by the OAI (OAI, 2003). However, its previous lack of UNICODE support made it less attractive for endangered language archives. At the 2004 EMELD workshop 'Linguistic Databases and Best Practice' most authors found XML a more appropriate format for an archive information system than a standard relational database. A common criticism of relational databases was that there was no established open-source relational database with UNICODE support or XML export (Ferrara and Moran, 2004; Gibbon, Trippel and Hell, 2004; Wittenburg, Broeder, Piepenbrock and vd Veer, 2004). MySQL only had UNICODE compliance in the beta release of an upcoming version at the time and only Ferrara and Moran (2004) looked at the new UNICODE functionality for MySQL, but they did not have a UNICODE compliant interface for the database so it remained largely untested and unreported. Similarly, EMELD chose the proprietary ORACLE software for its UNICODE support (Hooyenga, 2003).

We report that MySQL now supports UNICODE. We have not experienced any problem adding UNICODE characters to the database via online forms and PHP connectivity. Also, the XML export functions as reported, addressing some of the problems found by Wittenburg, Broeder, Piepenbrock and vd Veer (2004). We still see XML as the better interchange format for most dissemination packages and for representing structures within files.

5.2   Modelling relationships in the archive

The modelling of relationships between archive objects is a crucial part of the archive information system. However, there is still no widely accepted schema for modelling relationships, with many research communities favouring their own tailored models (OAI, 2003). There has been some work on this in the EL archive community (Wittenburg, Broeder, and Buitelaar 2004; Johnson and Dwyer, 2002), but it is a largely unresolved issue. The DSpace project (Bass et al, 2002a; Bass et al, 2002b) is informed by METS, but does not seem to be explicitly supporting the METS schema for their archive system.

One of the generalisations that we were able to make about the access requirements of endangered languages communities is that relationships between items are often more important than keywords from controlled vocabularies (McConvell, 2002b; NMAI, 2005). The National Museum of the American Indian also states that context is often more important than keywords (NMAI, 2005), and should be a factor in searching (although some types of context *can* be represented by keywords).

With good data modelling, any relationship between items in a database can be represented and documented clearly. For the archive information system, it may be enough to model the relationships between items in this way without a dedicated schema for describing the internal format, allowing flexibility and extensibility. For dissemination packages, this can be mapped to the relationships found in bundling schemas such as METS and MPEG-21 (NDIIPP. n.d.). We are currently following this strategy.

6   Conclusions

This paper has introduced the ELAR information system architecture. Significant features include: modelling archive objects at different levels of granularity; modelling relationships between objects; allowing users to enter their own metadata and relationships; allowing users to translate information into the language of their choice; allowing users to navigate via the language(s) of their choice.

7   References

Adobe Systems. n.d. "PDF as a standard for archiving". http://www.adobe.com/enterprise/pdfs/pdfarchiving.pdf. Viewed 26/7/05

Barwick, L. 2003a "Turning it all upside down - Imagining a distributed digital archive." Presentation to DRH Conference, Cheltenham, UK, 2 September 2003

Barwick, Linda. 2003b. Planning for PARADISEC: The Pacific And Regional Archive for Digital Sources in Endangered Cultures. Presentation to the Ozeculture conference, Brisbane Powerhouse.

Bass, M.J., Branschofsky, M., Stuve, D., Breton, P., Tansley, R., Carmichael, P., Cattey, B., Chudnov, D., Ng, J. (2002a), DSpace – Functionality, MIT

Bass, M.J., Branschofsky, M., Stuve, D., Breton, P., Tansley, R., Carmichael, P., Cattey, B., Chudnov, D., Ng, J. (2002b), DSpace – Technology & Architecture, MIT

Bird, S & G Simons. 2003. Seven Dimensions of Portability for Language Documentation and Description, Language 79/3: 557-582.

CCSDS 650.0-B-1: Reference Model for an Open Archival Information System (OAIS). Blue Book. Issue 1. National Aeronautics and Space Administration, Washington January 2002

Cedars Guide to: Digital Preservation Strategies. 2002 http://www.leeds.ac.uk/cedars/guideto/dpstrategies/dpstrategies.html, viewed 26/7/05

Csató, Eva and David Nathan. 2003. Multimedia and documentation of endangered languages In Peter K Austin (ed.) Language Documentation and Description, Volume 1

Farrar, S. 2003a Markup and the GOLD ontology. Proceedings of EMELD 2003.

Ferrara, Marisa and Steve Moran (Eastern Michigan U.): Review of DBMS for Linguistic Purposes. Proceedings of EMELD 2004.

Gibbon, Dafydd, Thorsten Trippel and Ben Hell. 2004. Archiving Language Resource Objects in XML. Proceedings of EMELD 2004

Grinevald, C 2003 Speakers and documentation of endangered languages. In Peter K Austin (ed.) Language Documentation and Description, Volume 1

Holton, G 2003 Approaches to digitization and annotation: A survey of language documentation materials in the Alaska Native Language Center Archive. Proceedings of the EMELD 2003

Hooyenga, Susan. 2003. The Challenges of Archiving Endangered Languages. Online: http://cf.linguistlist.org/cfdocs/emeld/school/readingroom/e-meld_paper.doc

JISC. 2001. A Strategic Policy Framework for Creating and Preserving Digital Collections. http://ahds.ac.uk/strategic.pdf, viewed 26/7/05

Johnson, Heidi and Arienne Dwyer 2002. Adapting the IMDI schema. Proceedings of the Language Resources and Engineering Conference (LREC), Las Palmas, Spain

Johnson, Heidi 2002 The Archive of the Indigenous Languages of Latin America: Goals and visions for the future. Proceedings of the Language Resources and Engineering Conference, Las Palmas

Johnson, H 2004 Language documentation and archiving, or how to build a better corpus, in Peter K Austin (ed.) Language Documentation and Description, Volume 2. London: SOAS.

McConvell, Pat. 2002a. Archive Initiatives in Australia and their Focuses. Proceedings of the Language Resources and Engineering Conference (LREC), Las Palmas, Spain

McConvell, Pat. 2002b. Linking Resources, Linking Communities. An Australian Indigenous Languages database, multimedia projects, and the role of metadata. Proceedings of the Language Resources and Engineering Conference (LREC), Las Palmas, Spain

Michailovsky, Boyd and Michel Jacobson. 2002. The LACITO Archive - its purpose and implementation. Proceedings of the Language Resources and Engineering Conference (LREC), Las Palmas, Spain

National Museum of the American Indian. nd. "Executive Summary and recommendations." http://www.u.arizona.edu/~aildi/Executive%20Summary%20and%20Recommendations.pdf. Viewed 26/7/05

NDIIPP. n.d. Sustainability of Digital Formats Planning for Library of Congress Collections: Formats, Evaluation Factors, and Relationships. National Digital Information Infrastructure And Preservation Program. http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml Visited 07/11/05

OAI. 2003. Open archives Forum: Project deliverable (D2.3). http://www.oaforum.org/otherfiles/oaf_d23_technical2.pdf. Viewed 26/7/05

Senft, Gunter. 2002. What should the ideal online-archive documenting linguistic data of various (endangered) languages and cultures offer to interested parties? Some ideas of a technically naive linguistic field researcher and potential user. Proceedings of the Language Resources and Engineering Conference (LREC), Las Palmas, Spain

Simons, Gary and Steven Bird 2003a. Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. Computing and the Humanities 37, 375-388.

Simons, Gary and Steven Bird. 2003b. The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. Literary and Linguistic Computing 18: 117-128.

Thieberger, Nick. 2005. PARADISEC Metadata listing Draft 4. [HTTP http://www.paradisec.org.au/downloads.htm] last checked 21/07/2005

Wittenburg, P. 2003a. "Archiving Strategies for Multimedia Language Documentation". Paradisec Workshop, Sydney

Wittenburg, Peter. 2003b. The DOBES model of language documentation. In Peter K Austin (ed.) Language Documentation and Description, Volume 1

Wittenburg, Peter, Daan Broeder and P. Buitelaar. 2004. Towards Metadata Interoperability. Proceedings of ACL 2004

Wittenburg, Peter, Daan Broeder, Richard Piepenbrock and Kees vd Veer. 2004. Databases for Linguistic Purposes: a case study of being always too early and too late. Proceedings of EMELD 2004

Appendix: User requirements

User	Requirement	Example	Example sources of information about requirement	Implications for the data architecture of ELAR
Depositor	Preservation of deposit structure (filenames, directories etc)	A recording can be retrieved as deposited after a period of time has elapsed, via the name & directory structure it had at deposit.	ELDP grantee training workshops. Linguists who had previously archived data elsewhere complained their data was 'transformed'.	Depositors should be able to browse the archive by the filenames / directories of the materials as at the time of deposit. Ideally, the archive should maintain the information required to 'reconstitute' a deposit as a dissemination package.
Depositor	Materials persistently attributable to them	A recording is accompanied by an explicit statement identifying the creator(s)	ELDP grantee training workshop. Many people expressed concern that archived materials might be distributed and used without acknowledgement. The archive information system should explicitly model the ownership of materials.	The metadata used in dissemination packages should have fields representing owner/creator etc. The archive needs to manage the access conditions, and protocols for access, which should include conditions for use of the materials in the dissemination packages.
Depositor	Update materials	A transcription that was only half complete at the time of deposit has been completed and needs to replace the existing one	Holton, 2003; Woodbury, 2003; CCSDS, 2002; Johnson, 2002; JISC, 2001	Superseded material should not be deleted without careful consideration. Version control architectures are required. The new material should replace the old for dissemination packages, unless a user explicitly requests the previous version. Metadata should be able to be entered after materials have been deposited.
Researcher	Broad search	Search for the existence of materials in a given language	EMELD, 2003; Simons and Bird, 2003a; Simons and Bird, 2003b	ELAR should map fields to OLAC, and become an OLAC provider.

Researcher	Domain specific search	Search for materials with specific conditions	Farrar, 2003a; Wittenburg, 2003b; Senft, 2002	For some broader searching by linguistic researchers, an archive should map to IMDI and become an IMDI provider.
				For domains other than linguistics, if there is a potentially significant usage, the archive should investigate and consider mapping to schemas designed for that research community.
				Archive staff will not have the expertise to make all appropriate mappings. However, an archive should allow the phenomena of a specific field to be represented in the archive and mapped to the resources in the archive. Therefore, a range of users should be able to add fields from established ontologies and to map archive resources to those fields. Some mediation of this process by the archive will be necessary.
Researcher	Search for specific content	Search for a recorded utterance of a particular word.	Bird and Simons, 2003	It is likely that such materials are only found within files. The archive information system should support an unambiguous way to reference the passage / sequence in which some material occurs, and allow that fragment to be assigned its own metadata. In short, the granularity of the representation of materials will need to extend below the granularity of the files at submission
Researcher	Group materials according to some criteria	Formally define the relationships between all materials in a deposit, or that are of a certain language, contain a particular phenomenon etc	OAI, 2003; Johnson and Dwyer, 2002; Johnson, 2002; NDIIPP. n.d.	The archive information system should support an unambiguous way to reference phenomena that might only be meaningful as the collection or aggregate of several materials, ie. a bundle. In short, the granularity of the representation of materials will need to extend above the granularity of the files at submission. The system should also support the modelling of relationships between items.

Researcher	Add results of research	Add the results of some analysis to the archive and relate it to the relevant materials.	ELDP grantee training: linguists expressed an interest in being able to comment on materials in archives, and to add the results of analysis. Also, discussion with Peter W and Daan B (DOBES) about ongoing issues in archiving.	The archive should allow persons other than the original depositor to add metadata about deposited materials. This extends to adding relationships between items. This will require moderation by the archive and owners of materials. Note the owner is not necessarily the depositor, and we may allow people to act as moderators who are not the owners of materials.
Endangered language speaker	Continuation of ownership[3] of language and materials	An organisation representing an EL community	McConvell, 2002b; NMAI, 2005; discussions with language documenters who are speakers of endangered languages (ELDP grantees and ELAP students); discussions with people working with EL communities.	EL speakers should be able to define their own metadata fields and values for materials, even if these are 'local'. This extends to making relationships between materials, and defining the types of possible relationships. The language through which materials are accessed should be the most appropriate ones for the EL speaker. That is, the language(s) with writing systems that the speakers are most likely to know. The archive staff can not know all these, so the archive should enable language speakers to translate metadata / comments etc into a language of their choice. Equally important, the ability to describe materials in their own language is one way of expressing their ownership of them. The archive needs to manage the access conditions, and protocols for access to these materials.
Publisher - repurposing	Obtain high quality data for repurposing	A language revitalisation project wants to create multimedia language learning materials	FATSIL, 2004; Csató and Nathan, 2003; Cedars, 2002; NDIIPP. n.d	Materials should be made available in dissemination packages that contain formats and structured tailored to the specific needs of a publisher. The archive structures need to be flexible, ie. not tied to formats or standards that limit restructuring.

---

[3] Note that 'ownership', as used here, does not necessarily mean ownership of IP, copyright, or property in a formal sense. Someone may feel ownership over some item through their use and ability to interact with it. In project management, the term 'owner' is sometimes used synonymously with 'stakeholder'. If an institution wants some individual to commit time and resources to a certain process or item, it is an important motivating factor that the individual feels that they are a stakeholder in that process or item. This is especially true in the absence of more material motivations, such as payment for services. If an individual feels that a resource is their property (even if not recognised by law) then it is important that the curators of an item or process do not seem to restricting the access to that item or process. For example, an EL archive would need to allow speakers of endangered languages to make comments on materials as easily as researchers can, taking into account the nature of comments and ideal access methods will most likely be different.

Long-term preserver	Obtain clearly structured data	A digital archive providing long-term preservation wishes to ingest the entire archive.	CCSDS, 2002	An archive should make arrangements with an archive that has long-term funding. They will need to negotiate the appropriate formats, structures and protocols for data interchange.
Publisher – public heritage	Exhibit materials	A museum wishes to exhibit a video of a ceremony	JISC, 2001	It is likely that the archive will act as a mediator between the museum and the owner(s) of the materials. Such ad hoc users are unlikely to be familiar with controlled vocabularies.
General Public	Browse materials	A person with an interest in endangered languages and/or cultures is browsing the archive		The archive should support dissemination packages that will inform and appeal to people with limited time and interest. Interaction is likely to be infrequent, so becoming familiar with controlled vocabularies is not an option. Archives should provide publicly-consumable 'exemplars', with an emphasis on multimedia and accessibility.