Publish or Cherish?
Marshalling Documentation at the Endangered Languages Archive

David Nathan
Endangered Languages Archive
Department of Linguistics
School of Oriental and African Studies
University of London

Research Centre for Linguistic Typology, La Trobe University
Oct 15, 2010
Summary

- ELAR @ HRELP @ SOAS
- redefining the endangered languages archive
- why? – diversity and protocol
- demonstration
- topics
  - metadata
  - the form of documentation
  - trends
Hans Rausing Endangered Languages Project

- established 2002
- based at the School of Oriental and African Studies, University of London
- supported by the Arcadia Fund
Three programs

- ELDP – documentation funding, ~ $1.5 million per year, currently ~ 200 teams worldwide
- ELAR – archiving & publishing of documentation, training, technical services
- ELAP – postgraduate programs in documentary and field linguistics MA, PhD, post-docs
  - ELDP and ELAR funded until 2016
ELDP Funded projects 2003 - 2007
Endangered Languages ARchive (ELAR)

- ELAR has a tiny staff; archivist, software developer, technician (0.5)
  - archive policies
  - develop archive infrastructure
  - curation, cataloguing and dissemination
  - support facilities
  - training and advice
  - materials development and publishing
  - participate in HRELP-wide activities
ELAR’s holdings

- ELAR currently holds about 90 deposits with a total volume of approx 10 TB.
- the average deposit is about 60 GB
- sizes vary widely, with a small number of huge deposits. The median size is around 15GB
- we expect volume to nearly double over the next 18 months
Other activities ...

- train grantees and others in London, Tokyo, Winneba (Ghana)
Other activities ...

- 3L summer schools (with Leiden and Lyon)
- LDLT (language documentation and linguistic theory) conferences
- “Endangered Languages Week”
- publish journal, CDs
- websites:
  - www.hrelp.org
  - elar.soas.ac.uk
Redefining the endangered language archive
What is a digital language archive?

- a collection of language materials
- a trusted repository created and maintained by an institution with a commitment to the long-term preservation of archived material
- a forum/platform for building and conducting relationships between data providers and data users
OAIS model

- OAIS archives define three types of ‘packages’

  ingestion, archive, dissemination:
Boundary between depositors, users and archive:
- users add, update content; negotiate access
Why redefine the endangered languages archive?
Documentation, archives, and languages

- Susanne Romaine: revitalisation efforts mean that *institutions* will increasingly become the vector of language transmission (rather than intergenerational transmission)

- In the longer term, *documentations* and the *archives* that preserve, maintain and disseminate them, will be the only means of transmission of many languages
Definitions of language documentation

- subfield of linguistics ‘concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language’ (Himmelmann 2006:v)
- ‘creation, annotation, preservation, and dissemination of transparent records of a language’... by nature multidisciplinary, drawing on ‘concepts and techniques from linguistics, ethnography, psychology, computer science, recording arts, and more’ (Woodbury 2010)
Characteristics of language documentation

- **primary data** – collection and analysis of a wide variety of primary data on language usage
- **preservation and dissemination** – archiving to enable preservation and access
- **accountability** – access to primary data encourages evaluation of linguistic analyses
- **diversity** – projects have different goals, methods, individuals, contexts, languages, communities, cultures
- **ethical and collaborative** – work with community members as both producers and co-researchers
Protocol – inseparable from documentation

- “naturalistic” recordings – more people, more genres, private conversations, less researcher knowledge
- community participation – speakers can shape process and products
- dissemination and mobilisation – selection, juxtaposition, participation, control of IP and distribution
- focus on revitalisation – which language to teach? who to host and teach? who can learn? etc
- time – sensitivities change over time
Stakeholders in language documentation

- documenters
- documentation projects
- language consultants
- communities
- heritage
- users

What goes here?

Curiosity, journos, applied research

Language support, protocol, acknowledgement

Distribution, advice, acknowledgement

Preservation, outcomes, accountability, protocol

Users, communities, institutions

Data stories, cute sound bites

Identity, cultural & social resources
Stakeholders in language documentation

- Can documentation satisfy all of these stakeholders and needs?
- What exactly is language documentation?
- What are the forms of language documentation?
- What are the functions and contributions of a digital endangered languages archive?
tailored to key characteristics of endangered languages field and its stakeholders:

- diversity
- protocol

aims and agendas:

- “depositing data” as an ongoing activity
- level the playing field between researchers and community members
- archive is platform for conducting relationships between providers and users
eler.soas.ac.uk functions

- implemented through “Web 2.0” principles and functions
  - depositor composition/editing
  - depositor control and feedback
  - protocol = access control
  - discussion forums
  - Twitter integration
  - reports to depositors
the strange dynamic or pattern of electric information is to involve the audience increasingly ... as actor, as publisher, as writer

- Marshall McLuhan lecture
  “The medium is the massage”,
  May 7, 1966
ELAR - demonstration
Incremental additions and improvements

- Depositors share “user reputation” information – we need to build and maintain depositor trust even as we expand the scope of activities on the site.

- User comments, in appropriate ways, e.g., experts’ guided tours of archive.

- Enhancement of speaker/community representation and functions (cf. “informed consent”)

Issues: metadata
Metadata: data about data

- for identification, management, retrieval
- providing context and understanding of that data
- carries those understandings into the future, to others
- reflects knowledge and practices of providers
- defines and constrains audiences and usages
- the goals and practices of language documentation heighten the importance of metadata
OLAC example categories [link]

- Contributor
- Coverage
- Creator
- Date
- Description
- Format
- Identifier
- Language
- Publisher
- Relation
- Rights
- Source
- Subject
- Subject.language
- Title
- Type
What do depositors actually do?

- ELAR currently has a research assistant examining depositor’s types and use of metadata over 5 years.
- Results to be reported in future, but already clear that there is a wide variability of metadata and between researchers and projects (‘Better data about metadata: a survey of depositor metadata submitted to the Endangered Languages Archive’, 2011 LSA meeting).
A little heresy

- in addition to minimal sets, many depositors also have different metadata
- types of metadata are relative to each project, consultants, community ...
- many depositors are sending extensive metadata in a variety of formats including spreadsheets, tables, prose text
  - and in file/folder names, correspondence etc
- goal of ELAR: to maximise the amount and quality of metadata because quality and extent is more important than standards and comparability
Depositors add metadata such as:

- Protocol (access) including as audio
- Detailed locations
- Metadata in Spanish
- Indigenous title (eg of a song)
- Genres (emic, not just linguistic)
- Parents’ mother tongues
- L2, L3 and competencies
- Number of children
- Children’s language competence
- Birthplace
- Parents’ birthplaces
... more metadata

- Residence
- Occupation
- Clan/moiety
- Education level
- Date left home country
- Photographs (& captions) of consultants, field sessions etc
- Performer
- Equipment
- Microphone
- Workflow status
... more metadata

- Languages heard
- Naming and organisational codes and principles
- Recorder/linguist experience level
- Biography and project description (“metadata documentation”)
Issues: the form of documentation
What does documentation look like?

- what *form* do deposits take?
  - interface
  - granularities etc
  - navigation
  - functionalities

- current solutions
  - IMDI [link](#) (see Tofa, Language, Shor)
  - search, based on thin metadata (eg AILLA, emphasises genres [link](#))

- metadata is crucial
Issues: trends
Important influences

- **dlib** – from the libraries community, for preserving, managing and disseminating digital scholarly and other publications
- **small language-specific archives, eg ASEDA**
- **OAIS (open archive information systems)** – architecture for serving various audiences
- **EMELD, OLAC** – emphasis on portability, standards, preservation, discovery (key paper: Bird & Simons 2003, Seven dimensions of portability in Language 79).
- **DoBeS** – emphasis on corpus-like interoperability (key technologies: IMDI and ELAN)
Today – two main trends

- **services:**
  - software, data processing, aggregation of data, e.g. DoBeS focus on tool building for data producers, OLAC resource discovery

- **relationships:**
  - develop and support relationships between data producers and data users
  - mobilise knowledge and skills of these individuals and communities
Our broader agendas ...

- diagnosis of a systematic gap in the documentation discipline, between
  - ideology, goals, ethics
  - software tools
- a ‘revolution’ in language documentation, as a critical mass of documentations become accessible and drives a new commerce in and critical discourse of them
... TV uses as its content the film. Movies have become the content of TV. When the movie was environmental, it was ... imperceptible. Now that it has become the content of TV, we can see it clearly.

... if you notice that the motor car created road systems and vast servicing industries, you would have a better idea of the motor car than you could have from photographic images. The car is what it does to people and the environment

- Marshall McLuhan lecture “The medium is the massage”, May 7, 1966
Contact us

- about grants for language documentation
  - eldp@soas.ac.uk
- about the archive, and depositing
  - archive@hrelp.org
- about studying at SOAS or visiting our project
  - elap@soas.ac.uk

- David Nathan
  - djn@soas.ac.uk
End