

National Science Foundation
Workshop for
**Documenting Endangered
Languages**



October 14 - 17, 2007
The New England Center,
University of New Hampshire,
Durham, NH

All materials contained herein are for viewing only.
Reproduction or dissemination without permission is strictly prohibited.
For permission or more information, write to DEL@NSF.GOV



Notes from ELAR

David Nathan
October 2007



What are the bottlenecks?

- ❖ Grantees do not always send data. About 60% of completed projects have deposited (some partial, eg audio waiting transcriptions); others in preparation
- ❖ Limited (human) resources for converting data to preservation formats
Not all linguists can be trained to or prevailed upon to make conversions effectively or without decreasing the amount or quality of their linguistic work. So the archive should ideally provide the service.
- ❖ Getting good IT staff for developing local systems
- ❖ Limited consensus within the language archives field about priorities and metadata standards
- ❖ Uncertainty in creator community about nature of documentation (to which archives need to respond)



Are archives underutilised?

- ❖ ELAR has no access system yet so unfortunately usage data cannot be provided until mid 2008. However, we are confident that its materials will attract significant usage for these reasons:
 - ❖ the archive is international in scope and so will provide an interesting range of resources
 - ❖ approximately 50% of materials will be “Open access”
 - ❖ we work closely with our sister programme, the Endangered Languages Academic Program, whose staff, students and visitors will find the resources useful
 - ❖ our granting programme required that materials are also co-archived in local archives accessible to speaker communities. This will provide further usage of the funded materials, even if not necessarily directly accessed from ELAR



To increase their effectiveness?

Recently the field's focus has been on access issues, eg "single portals", aids for resource discovery, standardisation of ontologies etc, which have achieved mixed success.

However, more attention needs to be paid to the nature of documentation materials, since there is currently low diversity in this area (generally the "trilogy" of linguistically annotated audio/video, grammar, dictionary). The nature of materials can be a driver for creating and serving new audiences.



ELAR's holdings

- ❖ ELAR currently hold 36 deposits with a total volume of approx 0.9 TB.
- ❖ The average deposit is about 25 GB, however, the sizes vary widely, with a few much larger deposits, and the median size is around 10GB.
- ❖ We expect this to nearly double over the next year
- ❖ See next slides for distribution of data types



ELAR holdings by data type

- ❖ This table analyses some data types of interest for a representative sample (70%) of holdings
- ❖ *Date type by volume and number of files, sorted by volume*

Data type	Volume (MB)	Files
audio	360,411	6,312
video	208,995	895
image	28,592	2,221
msword	223	404
pdf	196	134
eaf	33	176
text	32	781
lex	9	29
trs	5	246
xls	1	19
imdi	1	26



ELAR holdings by data type

- ❖ This table analyses some data types of interest for a representative sample (70%) of holdings
- ❖ *Date type by number of files and volume, sorted by number of files*

Data type	Files	Volume (MB)
audio	6,312	360,411
image	2,221	28,592
video	895	208,995
text	781	32
msword	404	223
trs	246	5
eaf	176	33
pdf	134	196
lex	29	9
imdi	26	1
xls	19	1



Other issues: access

- ❖ ELAR has researched and formulated an access rights classification system
- ❖ Balance between sensitivity and feasibility of implementation
- ❖ So far, high level of acceptance
- ❖ Currently applied at deposit level, will be extended to file or bundle level

... see excerpt next slide

...

Choose *one only* of the options P1, P2, P3, P4.

If you choose P2, choose *any combination* of P2A, P2B, and P2C.

P1. Anyone

Any person may view/listen to or receive a digital copy of any part of the deposit

P2. Certain people or groups

Choose any combination of P2A, P2B, and P2C:

P2A Research community members

What level of access (choose one only)?

P2A1. They can receive a digital copy of requested material

P2A2. They can view/listen but cannot receive a digital copy

P2B. Language community members

See below regarding identifying members

What level of access (choose one only)?

P2B1. They can receive a digital copy of requested material

P2B2. They can view/listen but cannot receive a digital copy

P2C. Particular named people or bodies

See below regarding identifying people/bodies

P3. Depositor is asked permission for each request

You will be contacted and asked for permission on each request.

How do you want to be contacted?

P3A. Requester is given address to contact you directly

P3B. ELAR will relay requests to you

P4. Only the depositor has access

Persons other than the depositor will not be able to request access.



Other issues: metadata

- ❖ ELAR has defined a metadata set
- ❖ It has about 50 field categories
- ❖ It is designed to be interoperable with IMDI, TEI, OLAC
- ❖ Metadata sets may be relative to the values of archiving institution and nature of holdings



Other issues: longevity, funding

- ❖ ELAR's funding is secured until at least 2012, probably longer
- ❖ Operating budget approx \$200,000 pa excluding high-cost items
- ❖ Arrangement with Oxford Text Archive for co-archiving
- ❖ Labour-intensive nature of our current model (training, advice, conversions, services etc) means that further funding will need to be found in the medium term