What we provide

- **Networking**— we facilitate international networking through our active email discussion list, Facebook group, and links to blogs and networks. Archived RNLD discussions are publicly accessible and searchable through the Linguist List.

- **Resources**— we disseminate information about funding opportunities, capacity building programs, publications, films, news, and the methods and technologies needed to document, archive, revitalise and maintain indigenous languages.

- **Training**— we partner with indigenous communities and organisations to build capacity in language documentation and revitalisation. Our flexible training materials and practices ensure that the skills which people gain are used, retained and shared more widely within a community.

- **Train-the-trainer**— we promote autonomy and sustainability through workshop methods which mentor skilled language workers in becoming trainers in linguistic documentation and revitalisation methods within their own communities.

- **Advocacy**— we advocate on issues such as bilingual education, indigenous literacy, language policies and recognition for Indigenous peoples, raise awareness through public events on occasions such as International Mother Language Day, and organize informal gatherings for linguists and community language activists.

To read more about RNLD, please choose from the menu items above
Linguistics in the Pub
26th March 2013

Things you can do with outputs from language documentation projects

Time: 6:00 - 8:00 pm

Venue: Upstairs room, Prince Alfred Hotel, 191 Grattan St, Carlton (corner of Bouverie St)
Methods

• The methods described here need not add too much to your work
• They will add a great deal to what you can do with your work
• You need to choose how much you can do, but at least be aware of what is offered by new and emerging methods
Methods

• Ask for advice – use networks like RNLD
• If you know that some work that would take you two weeks can be done by someone with experience of the tools in a few minutes, then
  – learn those tools
  – or find that person!
We want

• To be able to prepare excellent data in the course of our fieldwork (without too much extra work!)

• Excellent data will endure, can be accessed and can be made into various forms for delivery to various users
We want

• To be able to prepare excellent data in the course of our fieldwork (without too much extra work!)

• Excellent data will **endure**, can be accessed and can be made into various forms for delivery to various users
How to make data endure

• Central tenet of language documentation that we create data for posterity
• Formats need to have longevity
We want

• To be able to prepare excellent data in the course of our fieldwork (without too much extra work!)

• Excellent data will endure, *can be accessed* and can be made into various forms for delivery to various users
How to access data?

• It has to be described adequately in a catalog that can be searched

• It has to have persistent identification
We want

• To be able to prepare excellent data in the course of our fieldwork (without too much extra work!)
• Excellent data will endure, can be accessed and can be made into various forms for delivery to various users
Reusing our data

- Automatically (not laboriously by hand)
- For our own research into the future
- For others to reuse in research
- So that others can produce material based on our data
What tools and processes do we need to do this?

- Tools that do not lock our data into proprietary formats
- Processes that allow our data to flow from recording through transcription to analysis and further reuse
Typical workflow resulting in well-formed data

1. Project planning
2. Recording - named
   - digital file captured
   - archival digital file
   - descriptive metadata added
3. Transcribed and linked (using e.g. Transcriber or Elan)
4. Archived
5. Concordance of texts, navigation tool
6. Media corpus instantiates links to media
7. Output to e.g. Toolbox or Fieldworks for interlinearising
8. Texts, dictionary etc

Archived
Tools

Metadata entry
– SayMore
– ExSite9

Transcription with time-alignment
– Transcriber: http://sourceforge.net/projects/trans

Annotation and interlinearisation
– Fieldworks http://fieldworks.sil.org

Lexical databases
– LexiquePro: http://lexiquepro.com
– Fieldworks http://fieldworks.sil.org
Issues with workflows

• Data is primary and needs to be able to flow through the various tools

• Need to adapt as new tools are developed

• Understand how each tool deals with the data
Typical outputs of documentation work

Primary data – audio, video, text

- transcripts
- published texts
- dictionary
- multimedia
Issues with workflows

• Lossless conversion of primary data is crucial

BUT derived views of the primary data *can* be compressed and reduced, e.g.,

• Subitled video clips may not have interlinear glosses
• Dictionaries in various forms derived from lexical databases
• Streaming media is in lower resolution formats
Filenaming

• What needs to be named? What constitutes an *object* for our purposes?
  – Primary data
    • Tapes
    • Files - audio, video
    • Derived items
      – Transcripts
      – Texts
Filenaming

• Select a convention that works for you
• Be parsimonious (don’t get carried away!)
• Use simple characters
  – Various computer systems will have to be able to read your filenames
Simplicity is best:
200601A.wav
NT1-200601-A.wav

Not so good:
ERK200512ERAKORTOKELAUF.wav

Language code
Date
Place
Person
Gender
Filenaming

Every citation of a piece of data should be able to resolve to that data

Filenames must be consistent over time:
“Persistent Identification”

References:
RNLD FAQ on file-naming http://www.rnld.org/node/144
Principles

- Scale of data collection we are typically engaged in requires automating as much as possible of our workflow.

- To automate our work we need to understand the nature of the data and the tools we can use to work with it.
Principles

Automating means getting the computer to do most of the work
- Time-aligned transcripts
- Annotation
- Creating a corpus of texts
- Creating a lexical database
- Keeping track of all of that
Principles

e.g., in my work,
- 26 hours of recordings = roughly 125,000 words in transcripts
- Lexicon of 3,000 headwords
- Interlinear texts – 5,400 lines of text, with associated material results in a document that is 27,660 lines long, some 193,000 words
- Photographic images (>300)
- Media files (audio and video > 60)
- 423 objects in the archive
Outputs

- Thesis / book
- DVD of examples
- Online dictionary
- Printed dictionary
- Printed collection of stories
- iTunes installation of stories
- EOPAS stories
Outputs

- Thesis / book
- DVD of examples
- Online dictionary
- Printed dictionary
- Printed collection of stories
- ITunes installation of stories
- EOPAS stories
Principles

• Reuse
  – Of primary data
  – Of derived analysis

• By us

• By others, including the speakers/performers
Timeliness

• Leaving some tasks until later places an intolerable burden on the linguist who then does not manage to catch up with the backlog
  – types of tasks include: filenaming; entering metadata into a standard format
Project management

- Planning fieldwork and follow-up
- What kind of information will be recorded?
  - notes
  - audio/video
    - narratives; multi-participant conversations; songs …
  - still images
  - locations – GPS
  - genealogies
  - Other?
Project management

• What kind of equipment will be required?
  – e.g., music may require a better mic than spoken word
  – hard disks
  – backup equipment in case of failure?

Research equipment, get advice

RNLD list

http://www.rnld.org/join#email discussion list
Planning

- Consent forms for speakers to sign
  - Make it clear what can be done with your recordings and other outcomes of your fieldtrip
Planning

• Recording methods
  – learn to use your equipment, what it can and can’t do
    • in different conditions – background noise, wind etc
Planning

• Recording methods
  – Write notes, using a waterproof pen! Carry notebooks:
    • identify each recording in your notes
    • say ‘who, when’ intro to recordings
    • keyboard and scan written notes
    • use a laptop / portable device?
Seven Dimensions of Portability for Language Documentation and Description

LANGUAGE DOCUMENTATION - a record of the linguistic practices of a speech community, such as a collection of recorded and transcribed texts.

Planning

• Keeping equipment safe
• Humidity – use silica crystals (gel) for absorbing moisture
• Put hard disks, recorders, cameras into sealed containers with silica
• Use hard cases to carry equipment
Seven Dimensions of Portability for Language Documentation and Description

How can we ensure that this digital language documentation and description can be reused by others, both now and in the future?

Portability

• What is portability?
• Why is it important?

Fragility of digital records

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

• “BEST PRACTICES”

Best Practice

• "Le mieux est l'ennemi du bien."
  – Voltaire's Dictionnaire Philosophique (1764)

• "The best is the enemy of good."
• "The perfect is the enemy of the good."
  – Do the best you can!

The intertwining of content and format clearly makes this kind of language documentation difficult to maintain and reuse.

- E.g., dictionaries hand formatted, ignoring logical internal structure
- Headwords
- Part of speech
- Definition
- Semantic relations
- Semantic field
- etc

Seven dimensions

- CONTENT
- FORMAT
- DISCOVERY
- ACCESS
- CITATION
- PRESERVATION
- RIGHTS

Content

• Coverage (not just text)
  – E.g., multimedia
    • pay attention to quality of recordings, both in technical sense and in terms of content

• Accountability
  – Ability to check claims against (unedited) primary data

• Terminology
  – Interpretive labels (need for standard usage, e.g., Leipzig rules)

Format

• Open format
  – NOT proprietary formats
• Encoding of characters
  – Custom characters devised by clever linguists
  – Unicode helps
• Markup of structures
  – Marking explicit structural elements, e.g., Toolbox
  – Descriptive vs presentational / procedural markup
• Rendering
  – Ability to render or convert marked-up data into human readable forms

Format

• What media to use
  – We have learned over the past 20 years
    • Punchcards
    • Magnetic tape
    • Floppies (12 inch, 5 inch, 3 inch etc)
    • CD, DVD
    • HD

No media is adequate!

Migrate

Format

- Proprietary formats
  - Can vary depending on the needs of the company
  - Can become illegible over time (e.g. earlier versions of MS Word, FMPro require specialised techniques for reading, HyperCard stacks are virtually unusable) – "Digital Dark Ages"

Format

• Open formats
  – Saved out of MS Word as rtf
  – MS Word docx format
  – Created in Open Office
  – Markup of structures (using e.g., XML)
  – Need for good models and standards

Format - Well-formed data

• Will endure into the future
• Will be reusable

Format - Well-formed data

- Will endure into the future
- Will be reusable

- Characteristics
  - Structured: Content and form separated
  - Media in standard and archival formats

Format - Well-formed data

• It is in a form that will be legible over time (not locked into transient proprietary formats) and documents the use of any special fonts
• encoding of characters – use Unicode
• markup of structures

• Central to data management
• Distinguish archival from delivery forms of data, e.g.,
  – TIFF images with JPG delivery versions

*Language* 79:557-582.
• Central to data management
• Distinguish archival from delivery forms of data, e.g.,
  – WAV audio files with MP3 delivery versions

• Central to data management
• Distinguish archival from delivery forms of data, e.g.,
  – Complex interlinear annotations with simple 2 or 3 line outputs

• Central to data management
• Distinguish archival from delivery forms of data, e.g.,
  – Lexical database with MS Word / rtf / Open Office delivery versions

• Popular software like Microsoft Word and database applications are not suitable as archival forms.

• Neither are popular presentation forms (like dynamic web pages) suitable as archival forms.

• Unfortunately, linguists tend to focus on working form and presentation form when they think about using digital technologies; instead, they must look beyond these forms to the archival form if they want to create work that will endure.

This is important!

• Distinguishing
  – Form and content
  – Information and format
  – Structure and presentation

Create well-formed data and allow delivery versions to be derived from it

Why structure data?

• Outputs (including formatting) can flow from structure (text output from MDF in Toolbox)

Why structure data?

- Outputs (including formatting) can flow from structure (text output from MDF in Toolbox)
- Formats will be consistent

*Language* 79:557-582.
Why structure data?

• Outputs (including formatting) can flow from structure (text output from MDF in Toolbox)
• Formats will be consistent
• You can make global changes to the material
Why structure data?

• Outputs (including formatting) can flow from structure (text output from MDF in Toolbox)
• Formats will be consistent
• You can make global changes to the material
• Different outputs from the same underlying data

Why structure data?

• Outputs (including formatting) can flow from structure (text output from MDF in Toolbox)
• Formats will be consistent
• You can make global changes to the material
• Different outputs from the same underlying data
• Documents are portable

Discovery

• **Existence**
  – How to locate existing resources?

• **Relevance**
  – Is the resource what I expect it to be?
  – Can I read it (is the information given about its format sufficient?)

Discovery

• Standard metadata vocabularies facilitate automated retrieval
• Computers need standard forms for searching
• Who establishes standards and how do we find out what they are?
• Open Language Archives Community (OLAC)

Discovery

• Exposure of metadata to standard search tools

• Critical that research data be discoverable
  – Analog or digital data can be discovered if the metadata is properly constituted and published
    • e.g., OLAC metadata set
    • http://catalog.paradisec.org.au

Access

• Scope of access
  – Is the resource actually available?

• Process for access
  – Has to be some way of applying for and gaining access

• Ease of access
  – Appropriate to potential users (e.g., CD/DVD delivery)

Citation

- Citation form of data, e.g., in a bibliography
- Persistence
  - Uniform Resource Locator (URL), Persistent Identifier (PI), Digital Object Identifier (DOI), Handles

Citation

• Information in digital form (digital content)
• Access via the Internet available over long periods of time
• Likely that the content's location will change during that time

Citation

• Persistence
  – Uniform Resource Locator (URL), Persistent Identifier (PI), Digital Object Identifier (DOI), Handles
• Handles (redirect to the actual location)
  – http://hdl.handle.net/10125/4422

Citation

• Immutability - versions of resources, e.g. dictionary files
• Granularity
  – Ability to cite down to selections of the resource

Granularity of Citation

- Time aligned transcripts provide utterance level reference points in the data
- Potential for future streaming access to media at lowest level of granularity
- E.g. EOPAS

Preservation

• Longevity

  – “digitized materials are evanescent because they are based on binary formats.” (“tending to vanish like vapor”)

  (Bird and Simons 2003:567)

Preservation

- Media

  - To date, none of the digital recording systems developed specifically for audio has achieved a proven stability in the market place, let alone in an archive.
  
  - (International Association of Sound and Audiovisual Archives 2001).

Preservation

• Digital images, sound, and text have a binary representation

• Must be decoded and displayed before the document can be viewed.

• None of these translation schemes has stood the test of time to compare with plain old paper.

• All of them were invented in the last quarter of the past century.

Preservation

Safety

LOCKSS (Lots of copies keep stuff safe)

Do they?

Migration of data to new formats over time

Archive your records

Rights

• Terms of use
  – Need to identify who has rights in the material and make it explicit

• Benefit
  – Researchers may discover too late that legitimate but unanticipated uses by unforeseen users are unintentionally jeopardized when permissions are tightly circumscribed

Rights

• Sensitivity
  – Community sensitivity about content of recordings
  – Relationships with communities rely on proper shared understanding of how to use recordings

• Balance
  – Beware of locking up material with no time-sensitive endpoint

Rights

• Speakers have signed consent forms outlining their wishes for how recordings can be used

• Deposit forms with an archive outline deposit condition

Cataloging our data

- Keep a list of files created as you create them
- Text file for fieldnotes
- Structured data for the catalog
  - Spreadsheet
  - DBMS - database
Cataloging our data using tab-delimited text
Cataloging our data using a spreadsheet

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1010405.jpg</td>
<td>arthills and trees in landscape</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010404.jpg</td>
<td>arthills and trees in landscape</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010403.jpg</td>
<td>RK on ridge</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010402.jpg</td>
<td>Landscape</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010401.jpg</td>
<td>Kang looking for bush food</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010400.jpg</td>
<td>Muuki drinking the walyouriny infusion</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010399.jpg</td>
<td>Straining the walyouriny infusion</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010398.jpg</td>
<td>Straining the walyouriny infusion</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010397.jpg</td>
<td>Straining the walyouriny infusion</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010396.jpg</td>
<td>Walyouriny after being brought to the boil, sleeping.</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010395.jpg</td>
<td>Muuki preparing walyouriny</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010394.jpg</td>
<td>Muuki preparing walyouriny</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010393.jpg</td>
<td>Discarded hard wood of walyouriny</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010392.jpg</td>
<td>Muuki preparing walyouriny</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010391.jpg</td>
<td>Muuki preparing walyouriny</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010390.jpg</td>
<td>Sturt desert peas and car wreck</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010389.jpg</td>
<td>Camel leg pefood</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010388.jpg</td>
<td>RK's house and vehicle (on left) with moon.</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010387.jpg</td>
<td>Sunset</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010386.jpg</td>
<td>Sunset</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010385.jpg</td>
<td>Blurry moonrise</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010383.jpg</td>
<td>Camels</td>
<td>29/05/10</td>
</tr>
<tr>
<td>P1010382.jpg</td>
<td>Minvi ouriny at the saddle on this hill</td>
<td>29/05/10</td>
</tr>
</tbody>
</table>
Cataloging our data

• Spreadsheets stop being a good framework when you have repeated fields, e.g.,

> 1 language
> 1 speaker

Can’t do complex queries over spreadsheets

Need to use a database (DBMS) for your own work and then typically in an archive’s catalog
What metadata to use?

• Various metadata schemes to choose from, including:
  – IMDI
  – OLAC

• You can have your own terms, but ideally they will feed into an archive’s terms at some point
Brief aside – why metadata?

• Try a google search on a language name
• lots of extraneous stuff that obscures what we actually want to find
Standard terms

• Language names (Ethnologue)
  – ISO-639-3
  – List of names and codes for 7,105 languages
    – http://www.ethnologue.com/language/hau
  – Allows searching on just that language
Which metadata to use?

• Funded by DoBeS or using the MPI as the archive? – IMDI
• Funded by ELAR? Use their catalog
• Your own funding? Use whatever works that conforms to the metadata we have discussed
ExSite9

Regular expressions
Find/replace

• MS Word examples
• Using the 'Special' and 'Wildcards' option in find and replace in MS Word
Using REGEX in MS Word
Using REGEX in MS Word
Using REGEX in MS Word
Using REGEX

Find:
```
\lx [^s]+.
```

Replace:
```
&\d
```

Matching:
- [ ] Case sensitive
- [ ] Entire word
- [x] Greps

Search in:
- [ ] Selected text only
- [x] Wrap around
Regular expressions

[xyz] any one of the characters x, y, z
[^xyz] any character except x, y, z
[a-z] any character in the range a to z
[aeiou] any vowel
[^aeiou] any character that is not a vowel
[a-zA-Z0-9] any alphanumerical character
[^aeiou0-9] any character that is neither a vowel nor a digit
Regular expressions

. any character except a line break (i.e. a carriage return)
^ beginning of a line
$ end of line
\r line break (carriage return)
\n Unix line break (line feed)
\t tab
\f page break (form feed)
\ backslash
## Regular expressions

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Text is...</th>
<th>Matches...</th>
</tr>
</thead>
<tbody>
<tr>
<td>.*</td>
<td>Fourscore and seven years</td>
<td>Fourscore and seven years</td>
</tr>
<tr>
<td>[0-9]+</td>
<td>I've been here since 1983 or so. 1983</td>
<td>1983</td>
</tr>
<tr>
<td>\d+</td>
<td>I've got 12 years on him.</td>
<td>12</td>
</tr>
<tr>
<td>A+</td>
<td>BAAAAAAAAB</td>
<td>AAAAAAAA</td>
</tr>
<tr>
<td>A?</td>
<td>Andy joined AAA</td>
<td>the &quot;A&quot; from Andy</td>
</tr>
<tr>
<td>A+</td>
<td>Ted joined AAA yesterday</td>
<td>&quot;AAA&quot; and the &quot;a&quot; from yesterday</td>
</tr>
</tbody>
</table>

**Note**

* zero or more
? zero or one
+ one or more
Regular expressions

Applications that use fully featured Regex:

Libre Office

PERL – Python
**Microsoft**: EditPad Pro, PowerGrep, Notepad++
**Mac**: BBedit; Text Wrangler

MS Word has a weak sort of Regex in its 'find/replace' mechanism, and with wildcards.

**Useful sites:**
http://en.wikipedia.org/wiki/Regular_expression
http://www.regular-expressions.info/
http://etext.virginia.edu/services/helpsheets/unix/regex.html

**Online tutorial:**
http://www.zvon.org/comp/r/tut-Regexp.html