I am here today to talk to you about Open Data from the perspective of a data manager of a large, multi-institutional, multi-discipline research team.

Open data and you:
Data management in a large collaborative research project

Julia Colleen Miller
Senior Data Manager
ARC Centre of Excellence for the Dynamics of Language (CoEDL)
The Australian National University
julia.miller@anu.edu.au

20 October 2015

Researchers at the Centre represent four Australian Universities: The University of Queensland, Western Sydney University, the ANU, and the University of Melbourne. There are the numerous members affiliated with CoEDL and our research covers a wide range of language sciences.

We are investigating language as a diverse, dynamic system in constant change and how it interacts with our perceptual processes. Variability in language stems from social, geographical, biographical factors. Understanding why the world’s languages are designed so differently, and how our minds acquire and exploit them for different outcomes, will help generate important scientific insights and groundbreaking new technologies. To this end, we have four research programs:

- **SHAPE:** Why and how do languages differ? What can they tell us about ways of building communication systems that can address the needs of a human language?

- **LEARNING:** How do children learn language/languages so quickly? Are there some linguistic features that are harder than others?

- **PROCESSING:** How as adults can we process our native language effortlessly, but struggle when using a foreign language? What are the effects of being bi- or multi-lingual?

- **EVOLUTION:** Seeking to understand complexity and diversity of languages through studying processes of how they evolve. How might changing technologies influence directions of change?

One thread that runs through all of these programs is archiving.
In order to share these exciting linguistic insights, we need the following in place:

- A repository to store the data, in most of our research programs this will include video and audio recordings of language, but also elicitation materials and methodology for their use in order to replicate data collection, adding new languages for comparison.

- A plan to preserve and forward migrate the materials to keep in time with changing technologies. At one point reel-to-reel tapes were all the rage for anthropologists and linguists. These were kept in their original plastic or cardboard boxes, put into an archival box and placed in a temperature- and light-regulated room. However, now there are only a handful of institutions that can play, let alone convert these analog formats to digital formats to be accessed by later researchers, community members, family members, or other stake-holders.

- A way to create curated collections that offer meaningful gateways into the material.

Fortunately, CoEDL already has a plan for making our data publicly available. The PARADISEC archive is the chosen repository for CoEDL. PARADISEC was started in the early 2000s with ARC linkage funding (and other small grants) by an ethnomusicologist from University of Sydney, Linda Barwick, and a linguist from University of Melbourne, Nick Thieberger.
Researchers create collections, add items (such as an event or recording session) and then add the actual content files.

So, media files are at the lowest level in the three tiered structure (Collection>Item>Content File). The Item and Collection levels are organizational, metadata. The structured metadata found within the PARADISEC catalog is collected by online search engines, such as OLAC (Open Language Archives Community).
Training researchers to create rich metadata records, offering easy-to-use tools to create the records, keeping entries standardised as much as possible... All this allows for increased dissemination of their archived data.
In order to motivate researchers to actively archive their materials, we need to develop easy workflow models, such as this one, which maps out a flow of research material to the archive or to an external commercial service provider (such as DAMsmart! who transcode our video to an archival format).

We have our 4 CoEDL institutions on the right. There is an intermediate holding space for the transfer of data, archival audio gets shunted off to the archive immediately, while the video, a more complicated issue, get sent to our commercial service provider and then archival formats sent to PARADISEC and user-friendly, compressed versions of the raw video files are returned to the researcher.
Incentives and strategies for increasing access to data

- Transform large unwieldy data formats to more convenient sizes/formats
- Piece of mind that data is backed up in an archival format
- Provide shared workspace for collaborative efforts
- Develop technologies that allow crowd-sourcing of tasks
- Continue the push to have annotated corpora of small languages as referred publications and KPI/HERDC reporting (PARADISEC will offer DOIs)
- Many of the past language documentation projects explicitly tied funding to archiving the collected material
- Create curated collection of archived material for web presentation

So aside from facilitating the process and publishing/training researchers on the workflow, there are other incentives that can be offered to encourage researchers to increase access to the data:

- VIDEO TRANSCODING - RAW video captured by cameras in the field are large and in a format not easily used by researchers. If we offer this service within the workflow, we can archive the material early on in the program and researchers get professionally compressed files to use.

- CROWD-SOURCING - Similar to what the Atlas of Living Australia has done with their efforts to digitise specimen records from photographs to searchable text, all from the comfort of one’s own home. In our case, this could be extended to transcribing field notebooks, or narratives from audio/video files. We could come up with our own citizen scientist projects for use in high schools/undergrad courses/projects... Maybe even gamify the process.

- KPIs - Archived, annotated corpora including grammars, dictionaries, transcribed speech to be counted as KPIs (Key Performance Indicators) for HERDC reporting (Higher Education Research Data Collection). PARADISEC has successfully negotiated with the help of ANDS to get DOIs (Digital Object Identifier System) minted for all essence files in the archive.

- CURATED CORPORA - Designing web-portals that sit above the flat, rich archival material, drawing together media in a way that is meaningful
Discipline-specific challenge:

But at the end of the day, much of what we are collecting is the INTELLECTUAL PROPERTY of our colleagues/language consultants in the indigenous communities. So Prnda, here in the front, Qma and Doa ultimately guide us, the researchers, in how much of the recorded, original material we can make available. Informed consent and a clear discussion of access policies is paramount in keeping a healthy relationship with our consultants AND our data.
The majority of our data is intrinsically tied to people and culture. This offers us challenges, but not insurmountable ones—With openness, transparency, and continuing communication we should be able to continue with our trajectory of increased access to data and allow for good science to advance.

United Nations Declaration on the Rights of Indigenous Peoples

Article 31
Indigenous peoples have the right to maintain, control, protect and develop their cultural heritage, traditional knowledge and traditional cultural expressions, as well as the manifestations of their sciences, technologies and cultures, including human and genetic resources, seeds, medicines, knowledge of the properties of fauna and flora, oral traditions, literatures, designs, sports and traditional games and visual and performing arts.