# Linguistic scholarship in the data-driven 21st century

Simon Musgrave

MONASH University

John Hajek

THE UNIVERSITY OF MELBOURNE

# Overview

- Changes in recent times:
  - Technology for collection and dissemination of data
  - Focus on collection of primary data
- Data as driver of all stages of scholarship
- Technological solutions are easy
- Sociological and cultural change are hard

# Data from digital fieldwork

- Digital techniques allow recording:
  - High-quality
  - Reasonable cost
- Is this transformative?
- Access to data is useful to individual researcher
- Transformation of discipline(s) comes with wide access - dissemination

# Transformative technology

- Aspects of digital technologies which are (can be) transformative:
  - Copying with no loss of fidelity
  - Ease and speed of copying
  - Non-destructive editing
  - Easy accessibility (e.g. via networks)
- In sum: dissemination

# Data-driven scholarship - access

- Large bodies of data can be made accessible
  - Repositories such as PARADISEC, The Language Archive
  - Federated discovery e.g.:
    - OLAC
    - ANDS
- Individual datasets tend to be not large
- Problems of aggregation still need attention

# Data driven scholarship – (re)using data

- Putting data together with useful tools
- Projects are emerging which do this
- In Australia:
  - HCS vLab
  - HuNI

# Value of dissemination

- More scholars having access to more data
- One possibility seems especially important:
- Linking of primary data to published analysis
- This would mean greater accountability in our scholarship
- But only effective if electronic publication becomes primary
- This raises challenges

# Data driven scholarship – data as a part of publication

- An example in a publication can link back to the original media
- This is a gold standard in accountability
- Example:
  Speakers of Sou Amana Teru at Liang palatalise [s] before [i]; thus [sia] at Tulehu becomes [sʸia] at Liang

# Technical challenges

- Such as allowing browsers to address specific sections of media files
- Solutions seem close:
  - Annodex was (is?) promising
  - HTML5 has the capability (although still not easy to use always….)
- But not something to worry about

# Cultural challenges

- Recognising making data accessible as academic output
  - FORCE11 - **DRAFT - Declaration of Data Citation Principles, Principle 1:** Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications (http://www.force11.org/datacitation)
- Developing new models of academic discourse
  - Or re-conceptualising existing models

# Data publication

- Australian Linguistic Society (ALS) has begun discussing issue with ARC
- ARC had no hesitation in acknowledging that curated data embodies research activity
- But discipline has to devise and administer processes for assessing collections
  - Sub-committee formed by ALS for this
  - Work continues

# New modes of dissemination

- Books are linear, hypertext need not be
- Electronic grammaticography:
  - 2012 edited volume (ed. Nordhoff, U Hawai'i Press)
  - NB – available electronically, but still conceived as written object
- Language description
  - Recognised as tri-partite since Franz Boas
  - Parts are richly interlinked
  - Natural for hypertext – Heath and Nunggubuyu
    - Ongoing project (Thieberger and Musgrave)

# Institutional challenges

- The difficult area!
- Various aspects and various groups to address:
  - Ourselves – the producers (but see previous slides)
  - Our peers – one type of consumer (but advantages should convince them)
  - Gatekeepers – publishers and academia as an administrative body

# Publishers

- We all use journals as electronic resources
- But publishers are slow to exploit the possibilities this offers:
  - Basic model – published text as pdf
  - Maybe additional online resources offered
- How long should we allow this to continue?
- How much effect can consumer pressure have?

# Academia

- But we need publishers
- They provide credibility for our work within academia as a whole
  - Job applications, promotion
  - Funding
- Self-publication of the material is not difficult
- But no recognition….
- Exposure of data as scholarly output needs persisting institutions

# Conclusion - Scholarly practice in 21C

- Digital fieldwork has implications for downstream activity
- These implications should be beneficial for our disciplines:
  - Better access to data
  - Better accountability
- It is not enough to only take up best data collection practices
- Dissemination practices, including publication, should also change

# Conclusion – Fostering change

- We should be prepared to articulate and defend new scholarly practices
- This may mean putting pressure on the institutions which act as gatekeepers
- But these are OUR disciplines – we should define what is best practice