Data Commons and the Humanities

Michael Haugh (University of Queensland)
Simon Musgrave (Monash University)
Knowledge as commons

• There is a changing view of knowledge and how it is created developing around the idea of openness: Open Access, Open Science, Open Data

• Knowledge is conceptualized as a commons, that is, as a shared resource built, maintained and used by a community

• A data commons is a part of research infrastructure in such a model

• The ideal of Open Data does not necessarily align with research traditions in all areas of humanities or with the wishes of communities

• Does a humanities data commons need to be different to a data commons in other disciplines?
Humanities data - differences

• Data in humanities research increasingly are:
  • Digital objects
  • Manipulated with computational tools

• Data is less likely to be created in the research process and often sourced from GLAM sector
  • Strengthens view of data as part of commons
  • No one owns the text of Shakespeare but many base their research on this data

• “Big data” in humanities: differences in volume, variety, value
Volume

• Data objects are typically smaller than in many other disciplines
• But high resolution imaging or multimodal material require significant storage space
  • Large proportion of storage
  • Small proportion of objects
• There are a limited number of very large humanities data collections (e.g. Australian Web Archive: 9 billion records, 600TB)
Variety

• Humanities research can use many different kinds of material:
  • Text
  • Images
  • Audio
  • Video
  • Quantitative data
  • Others....

• This variety is an essential feature of humanities research
Value

• Reuse is an important aim of sharing data
• In some disciplines, only specialists in one particular discipline can reuse data
• Humanities data is:
  • Usable by multiple disciplines
  • Accessible to non-specialists
• Humanities data has significant value to communities and so can require access restrictions (FAIR versus CARE principles)
• The value of humanities data does not diminish over time – indeed its value can progressively increase
Language data as example

• Language data exemplifies the points made previously:
  • Volume – media important for much research today, but still not enormous amounts of data
    • Media only valuable if annotated, further restricts amount of data that needs to be stored
    • Text is cheap to store
  • Variety – text, audio, video, images all used in language research
  • Value – language data can be used by researchers in other disciplines and vice-versa
    • Text collections may be of interest to media studies or literary studies
    • Oral history collections are of interest to linguists
    • Non-scholars may be interested in any of these materials
      • Special case of speaker communities and endangered languages
Some implications

• Storage level:
  • Dealing with many formats is more important than dealing with large volumes

• Access/interface level:
  • Enabling interaction with different data types is desirable
  • Access control is very important

• In building the interface to a language data commons, appreciating the complexity (i.e. granularity and interconnectedness) of language data is critical
A Language Data Commons as part of a Humanities Commons

• The interests of language researchers are diverse but:
  • Many already deal with data in digital forms
  • Costs of producing (some kinds of) data mean that aggregation is the best strategy to build scale
  • Many kinds of language data are reusable

• A language data commons should represent the massive diversity of languages in Australia and its region

• Ethical and equitable:
  • Access as wide as possible while respecting the rights and sensitivities of those who have contributed to the data
  • Barriers in accessing the data should be minimised as far as is consistent with the ethical and cultural commitment
Language Data Commons Policy Framework

• Dialogue is needed between researchers and communities to reconcile the principles of FAIR and CARE with respect to language data (and cultural data more generally)

• A sustainable national language data commons requires the formalisation of institutional relationships between the research sector (esp. universities) and the GLAM sector (esp. libraries and archives)

• The key to building a national language data commons is developing a policy framework that navigates rights restrictions (cultural, moral, copyright)
Conclusion: Big Data in the Humanities

• A data commons is implicated in each level of Kaplan’s (2015) model:
  • Big Cultural Datasets
    • Preservation and Curation are part of the processing cycle at this level
    • A data commons is part of that processing
  • Digital Culture
    • Includes a *control* domain - covers the relationship of communities and global actors with massive digital objects and the software medium
    • A data commons is one model for how control relationships can work to the benefit of broad communities
    • Finding an access model which balances rights of contributors with other communities is crucial
  • Digital experiences:
    • Interfaces contribute to digital experiences
    • Humanities data in a commons poses special challenges for interface construction
Humanities and Computing Potential

• There is vast potential for the humanities through extending computational methods to humanities research
  • Multimodal / multipurpose big data (e.g. Trove)
  • Mapping (combining data and geographical space)
  • Real world applications
    (applying corpus-based research in classrooms)