The Case for Data Archives

Professor John Bacon-Shone
Director, Social Sciences Research Centre, HKU
Acting Director, Centre for Criminology, HKU
Associate Dean (Research), Social Sciences, HKU

Key archives

- ICPSR (Michigan, US) - since 1962 - started out as sociology and politics, but now covers much wider range, but still mainly social sciences - funded by institutional subscriptions and funds linked with datasets
- UK Data Archive (Essex) - since 1967 - quantitative social and economic - funded by ESRC
- CESSDA - Portal to the many European national social science archives
- Asia: Australia, Taiwan, Korea, Japan have social science archives
- Other archives on a subject basis including astronomy, religion, history, geospatial,atmospheric, earth sciences etc.
- Generally cover data that is either very expensive or impossible to recollect.

Key Questions

- Are data archives a good public investment?
- If so, how best to fund/organise?
- Does Hong Kong (and every other place) need their own data archive?

Good public investment?

- What is the benefit?
- What is the cost?

Benefits

- Avoid data collection costs (often paid for by taxpayers)
- Guarantees datasets not lost
- Enable re-analysis of important datasets
- Facilitate training
- Many datasets cannot be replicated

Costs?

- Setup costs (hardware) (relatively small)
- Running costs (indexing, thesaurus, identifying, storing, distributing, space) (were large, but decreasing, particularly if reuse software and thesaurus developed elsewhere)

How best to fund/organise?

- Funding mechanism?
- How to organise?

Funding mechanism

- Michigan model - in order to obtain access, subscribe at university level, also seek funds to support specific datasets, link with summer school
- Essex model - research funding agency supports on behalf of all public universities, link with summer school
- Both have strengths and weaknesses

How to organise?

- Links with funding
- Key question is who decides what to archive (i.e. priorities)
- HK proposal was to have subject committees to review datasets for inclusion
- For Essex, funding agency requires datasets to be offered to archive
- Michigan chooses, unless linked funds, many government agencies require archiving

Case for local archives?

- Local knowledge (cultural, language) - important in deciding which datasets to archive
- Local access (for those who need physical or secure access) - important where secure access is critical, e.g. large fractions of census data where individuals might be identifiable, as done in Michigan & Essex

Attempts in HK

- Many attempts to persuade Research Grants Council, Vice-chancellors Committee (HUCOM) and Census and Statistics Dept over more than 10 years!
- RGC only agreed to add tickbox on grant applications asking if prepared to archive datasets (ignored in practice) and agreed in principle to fund archiving of individual datasets (never done yet)
- RGC has no record of how many applications have stated that they are prepared to archive!
- Refused to provide any setup costs for an acrhive at all, as argued
  - No demand
  - Lower priority than research grants
- Even though I presented a startup proposal that would cost about the average cost of 1 or 2 grants in addition to the per dataset cost they agreed in principle to cover long ago.

Attempts in HK

- Census and Statistics under the previous commissioner:
- Refused to archive any datasets
- Required anyone using C&SD data to sign a completely unrealistic contract:
  - no analysis based on a sample could be published
  - C&SD must replicate all analysis on the full dataset
  - C&SD could merge any cells that they judged too small
- Complained when people broke the contract!

Attempts in HK

- Classic example of dataset loss:
- The Sports Development Board was closed down
- All research data generated by them is no longer unavailable, including a Millennium Sports Study that cost several million dollars!
- Hong Kong does at least have a Social Science Methodology Summer School, which I have run for over 10 years, with RGC support!

Thanks!
Questions?