Preserving scientific data in developing countries

Birth to Twenty data curation project

Scientific Information for Society: From Today to the Future

October 2010

Social science that makes a difference

HSRC
Human Sciences
Research Council

Data Curation in Developing Countries

Data curation has become increasingly important in the South African research environment

- Increased recognition of the importance of social science data for the South African government in the monitoring and evaluation of its policies
- The need for such data to be available for re-use and preservation
- Specific targets – MDG's ; 12 outcomes;

Structure of presentation

- Introduction to Birth to Twenty
- Description of Data curation project
- Challenges
- Possible solutions
- Role of government bodies

Birth to Twenty

- The Birth to Twenty (Bt20) cohort started in 1989 with pilot studies to test the feasibility of a long-term follow-up study of children's health and wellbeing (Yach et al 1991).
- Women were enrolled in their second and third trimester of pregnancy through public health facilities and interviewed during regarding their health and social history and current circumstances.
- Singleton children (n=3 273) born between April and June 1990 and resident for at least 6 months in the municipal area of Soweto-Johannesburg were enrolled into the birth cohort and have been followed up 16 times between birth and 20 years of age (Richter et al 2004; Richter et al 2007).
- During the last 7 years, young people have been seen twice a year, at the Bt20 offices and at home.
- Attrition over two decades has been comparatively low (30%), mostly occurring during children's infancy and early childhood, and approximately 2 300 children and their families remain in contact with the study (Norris et al 2007).
- The sample is roughly representative of the demographic parameters of South Africa with equal numbers of male and female participants.

Why is Bt20 Important?

- Unique source of information on social aspects of child development and health
- Public policy implications (eg smoking legislation influenced by smoking study)
- Methodological learning for future studies (eg tracking subjects in environment with very mobile population)

Data Curation Context

- Aim of the curation project
  - Preserve the context of the study
  - Create a repository for data & related documentation
  - Describe data & related documentation in terms of structured and un-structured metadata
  - Establish and document standards, processes and practices
  - Develop a capacity for data sharing
  - Preservation of data & documents
- Expected results
  - Better collaborative data sharing partnerships
  - Improved longitudinal data cleaning
  - Quicker analytical dataset construction
  - Long term use of data & documents

Data Curation Process

Steps

- *Audit of data and documentation*
- *Document collection*
- *Description*
- *Storage and preservation*
- *Dissemination*

Audit of data and documentation

- More than 20 years of data collected
- Various levels of cleaning done
- Mostly quantitative data with some qualitative
- Different themes covered over time
  - Infant, child and adolescent physical and mental health
  - Influence of home, school and family environment
  - Sexual and other risk behaviours during adolescence
  - Body composition, obesity, emerging non-communicable disease risk
  - Nutrition, bone health through childhood and adolescence
  - Methodological issues

What criteria should be used in the selection process?

What level of processing of data is required?

Document collection

- No systems in place for supporting documentation
- Three office moves
- No central administration for first two years
- Collaborative protocol and questionnaire development
- Perishing paper documents

Description

- DDI and Dublin Core
- Explicit and tacit knowledge
- Versioning
- Record of what took place and what did not
- Re-use by a variety of audiences

What types of metadata will be included and how much, and using what standards

Storage and preservation

- Filing systems
- Digital curation of documents
- File server for data and documentation and metadata
- Preservation – migration/ conversion

How will data (and related documentation) be preserved to ensure it is still available in the future?

Dissemination

- Different levels of access
- Challenges for researchers – need to ensure acknowledgement, potential for misrepresentation
- Ownership and copyright challenges
- Longitudinal and discrete data
- Principles of access

What criteria will there be for access?
confidentiality

Challenges

- Data curation process is multidisciplinary and requires participation from a wide range of individuals
- In countries where there are many competing needs, resulting in research being allocated less public funding than is ideal, is it really worthwhile to spend time and money on curating data, and are there any ways of cutting costs.

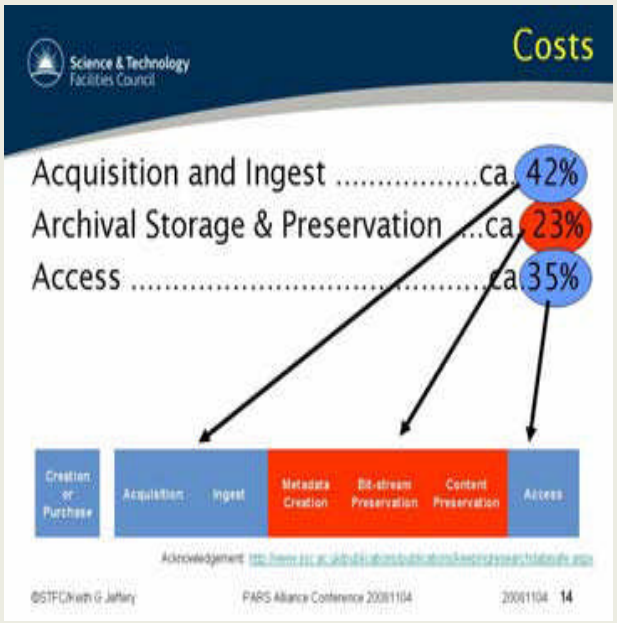
Sustainability

Technology

Infrastructure: Storage, document repository, metadata repository, dissemination interfaces and software, long term preservation infrastructure People:

Data creation, analysis (describe); Data management

Data curation – Standards; Reviewing and adding additional metadata; Format translation; Long term preservation



Limit cost

Individual datasets

- Appraise and select data
- Follow best practice in data management
- Curate data as soon as possible
- Efficiency curve effect
- Plan for re-use
- Share and collaborate – re-use

General

- Avoid unnecessary duplication of data collection
- Apply new theories to existing data
- Do secondary analysis (different problem, same data)
- Economy of scale effects
- Common standards

Maximise Benefits

- Network approach to data curation – sharing data infrastructure and standards, as adhering to these standards can greatly reduce the overall cost of data curation
- Standards for data management and metadata standards etc are the types of things which can be dealt with at a national or regional level, benefiting from economies of scale.
- Economies of scale in terms of the cost of infrastructure and the development of metadata and data standards would generate many benefits – including learning from the experience of others and possibilities of mentoring.
- If a Community of Practice is created with individuals from each data collecting site can be set up enormous amounts of learnings can be shared and replicated rather than re-inventing the wheel at each site

There are challenges with this approach – not least agreeing on standard terms despite regional, cultural and language differences

Role of government bodies

- According to Doorn (2007), most data archives sooner or later become affiliated with national research organizations or academies. If this is the case then it would be sensible for these research organizations and academies, as well as the regional bodies, to consider ways in which to co-ordinate data curation activities – either by geographical region, or by discipline.
- Despite Africa's diversity regionally, there are a number of regional bodies – both academic and political that could co-ordinate a network approach to data curation. Regional academic bodies such as CODESRIA and OSSRIA could co-ordinate efforts to create common standards and regional government co-ordinating bodies under the umbrella of both the AU and Nepad could assist with regional standards (for copyright for example)