Birth to Twenty
Taking data curation seriously

2009 HSRC Social Sciences Conference

16th September 2009

Lucia Lötter and Alison Bullen

Social science that makes a difference

HSRC
Human Sciences
Research Council

BT20 Data Curation Project

- Birth to Twenty obtained funding from the Wellcome Trust to curate its data collection, starting with the datasets from the first wave in 1989/90
- The HSRC was employed to assist in this process, setting up systems and developing best practice
- This paper will present an overview of the framework developed for this project, and then discuss some of the challenges in collecting metadata for a complex study that started 20 years ago, as well as the importance of doing this efficiently and affordably taking the pressures of the global economic crisis into account and the need for this to be an ongoing process

BT20 Study

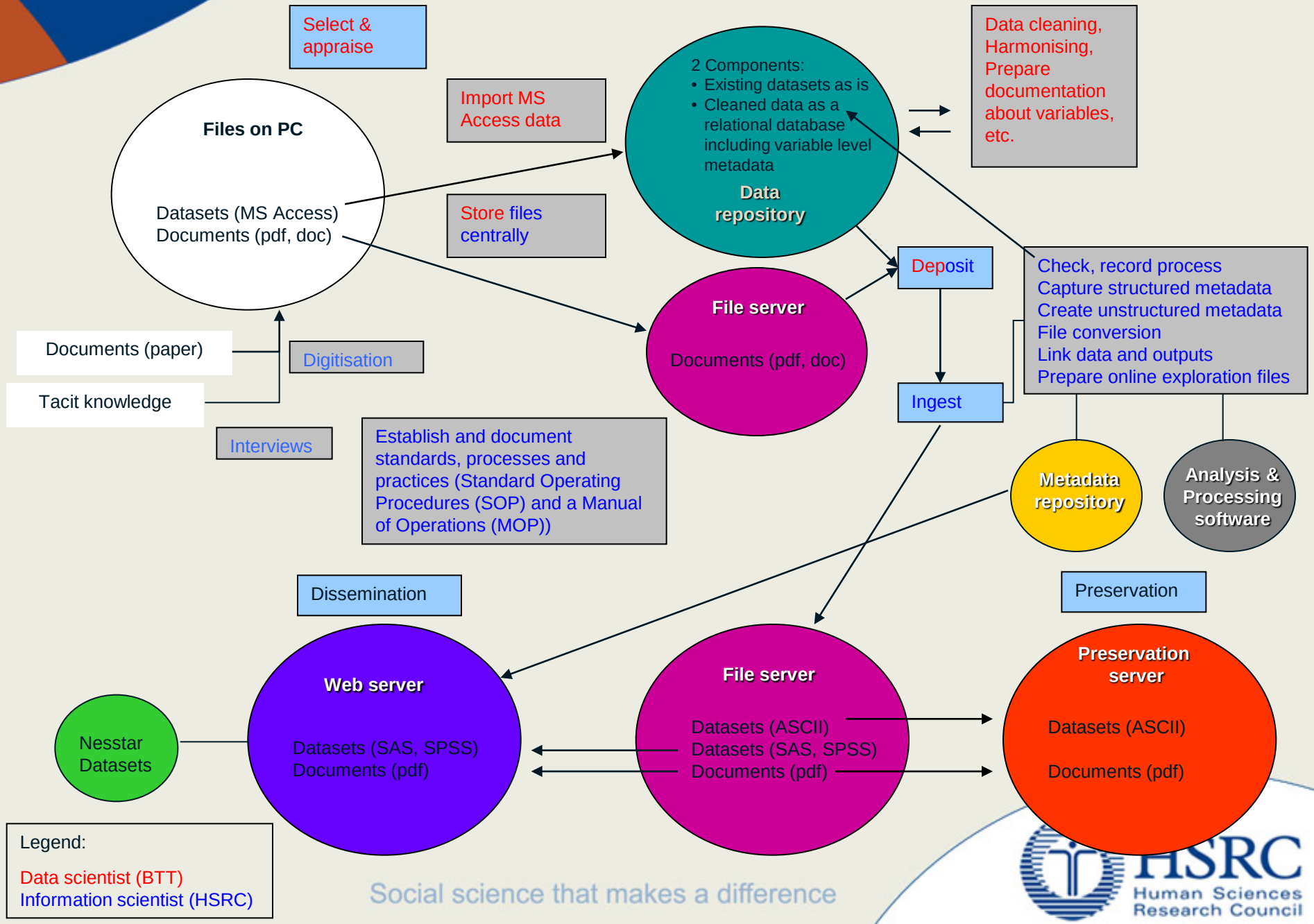
- Longitudinal birth cohort study starting with a pilot project in 1988 and antenatal survey in 1989
- Started tracking the development of 3,273 subjects/participants born between April and June 1990 in Johannesburg-Soweto
- Provides an important source of information on public health issues in an urban setting of a developing country.
- Links with other longitudinal studies in SA (Agincourt) as well as similar studies in other countries (Alspac, UK)
- Not simply a research study, also teaching (Phd students) and activities aimed at maintaining interest in the project – newsletters and events
- Analysis of data done in collaboration with researchers in other countries

BT20 Data - Characteristics

- Almost 20 years of data collected – annual waves of collection with year 19 data collected end of 2008
- 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

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



Social science that makes a difference

HSRC Human Sciences Research Council

Data Description

- Data description includes
  - Context – in order to understand the data properly (both tacit and explicit from formal documents as well as informal communications)
  - Metadata: Tools for retrieval – key wording and abstracting
- Reuse is not just “the use of data collected for one purpose to study a new problem” and different types of reuse require different levels of metadata:
  - Researchers who are part of the research project
  - Collaborators
  - Scientific community at large

Standards and Best Practice Manuals

- What standards do you use
  - For data – Data Documentation Initiative (DDI)
  - For data description – Dublin Core
  - Classification – HASSET
- Processes - Need for Data curation manual of operations and standard procedures
  - Workflow and responsibilities
  - Creation of data files best practice – data management
  - Minimum supporting documentation required
  - How to organise documentation (File plans; file naming conventions)
  - Metadata best practice
  - Format and conversion of documents
  - Privacy and confidentiality

Challenges for Bt20 project

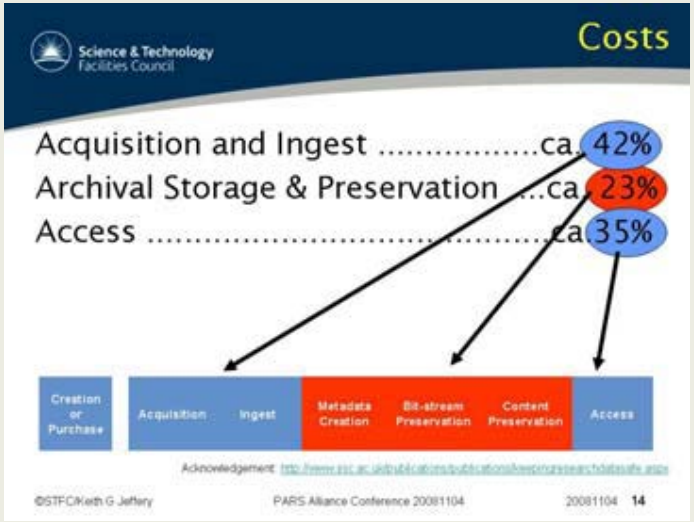
- Data description
  - Missing documents
  - Different versions of the same document
  - Tacit knowledge
  - Co-operative nature of the study management in the first ten years
- Data
  - Data audit
  - Qualitative data
  - Missing data
  - Version control
  - Datasets not harmonised

Data curation in the current economic recession

- Can we afford to curate the data? What are the benefits and the cost?

Benefits:

- Increase research efficiency
- Ensure data quality and integrity
- Preserve data for re-use
- Meet grant requirements
- Increase the visibility of research
- Access to a comprehensive data archive



Cost components

- Technology
  - Infrastructure: Storage, document repository, metadata repository, dissemination interfaces and software, long term preservation infrastructure
- People: Data management
  - Data creation, analysis (describe)
- People: Data curation
  - Standards
  - Reviewing and adding additional metadata
  - Format translation
  - Long term preservation

Limit cost

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

Maximise benefit

- Avoid unnecessary duplication of data collection
- Re-analyse for verification
- Do secondary analysis (different problem, same data)
- Do meta-analysis (same problem, several independent datasets)
- Refine (alternative analysis)
- Test the generality of research findings
- Create new enlarged datasets
- Apply new theories to existing data
- Monitor historical changes

Data curation in the current economic recession

- Can we afford not to curate data?

- Key benefits:
  - Improved data quality: longitudinal data cleaning
  - Better collaborative data sharing partnerships
  - Quicker analytical dataset construction
  - Long term use of data & documents

Long term benefits out weight short term costs – not having to redo research and being able to re-use data

HSRC
Human Sciences
Research Council