Quality Control of Data in Data-Sharing Practices and Regulations

Paul Wouters and Anne Beaulieu

Networked Research and Digital Information (Nerdi)

Background

- Fraud cases in science:
  - Peer review no guarantee
  - Purification measures
  - Harbingers of future of data quality control
  - Responsibility of quality control at stake
- Networked Research and Digital Information:
  - Interaction ICTs and knowledge creation
  - Interdisciplinary social science
  - Research methods
  - Research themes

Increasing role data

- Exponential rise of amounts of data
- Research increasingly data oriented and dependent
- Developments vary by discipline
- Quality control of data more crucial
- Ethics of research also focused on data (human subjects)
- Different configurations

Conceptual issues

- Definition of data
  - Units of information in research that can be isolated from their context of production in order to be used in another context.
  - In digitized science: digital records of scientific measurements or observations.
  - Distinction raw data and processed data blurred
  - definition of data always context specific and related to research question

- Flows of data instead of data bits (Hilgartner)
- Data structure is field specific
- Quality as constructed and context specific
- Quality control both produces and represents quality

Data Sharing

- Data Sharing as a Good Thing (policy)
  - Good Stewardship of public knowledge
  - Strong value chains of innovation
  - The creation of value from international co-operation
  - Quality control implicit
- Data Sharing as Extra Work (practice)
  - privacy of subjects;
  - too much work
  - being scooped
  - long-running squabbles
  - paper work
  - losing volunteers
  - career
  - collaboration with industry

Data Sharing Configurations

- Different actors:
  - Peer to peer
  - Data archives and repositories
  - Centralized data production
- Different mechanisms:
  - Face to face
  - Mediated by ICTs
- Different data types

Peer to Peer Data Sharing

- Discrete research groups
- Data location not self-evident
- Researcher is keeper/steward of the data
- Data tied to specific research project
- Trust among researchers key element

Data Archives

- Centralized repository
- Data annotated and formatted (meta-data)
- Focused on one field or sub-field
- Uncertain budgets due to system of research funding

Centralized data production

- “Big science” institutions or networks
- Close coordination of data production
- Data sharing *not* a separate issue: data availability limited to groups involved in production
- Highly processed (interpreted) data available for public (education)

	PI	journals	repos.	institute	fund ag	user
p to p	X	X BMJ				X
p to p ict	X	X NCHS				X
archives	X	X	X Steinmetz		X	X
archives ict	X	X fMRIDC	X EBI		X	X
big Science	X VLBI			X CERN	X	

Implications for Quality Control

- Increased pressure on investigators
- Increased pressure on peer review system
- Different data sharing configurations require differential approach to quality control
- Pressure on research funding mechanisms
- Big science networks/institutions: business as usual?
- ICT tools and skills and investments part of research infrastructure

Further Research Questions

- What are the limitations of this matrix in the analysis of other case studies?
- Can we see emerging interfaces between the actors?
- How are the actors developing qc mechanisms?
- Which dimensions of social relations such as trust are crucial in the different contexts?
- What role can be played by ICTs in qc of shared data?
- Which trade offs are being made?