Challenges for Multi-Disciplinary Research Data Infrastructures

The Private Domain

Dieter Van Uytvanck,
MPI for Psycholinguistics / Radieschen
forschungsdaten.org
Domains

1. Data management plan
2. Incentives

Group Domain B
3. Quality control
4. Selection
5. (extended) Data management plan
6. (New) incentives

Persistent Domain C
7. Service Level Agreements
8. Trust & Certification
9. APIs, metadata standards
10. Persistent Identifiers

Access & Reuse D
14. Community portal
15. Scientists
16. Openness
17. Searchability (e.g. Google)

11. Preservation Policy

12. Community-specific organisational structures

13. Access
Life cycle

Life cycle: creating data

Creating data

- design research
- plan data management (formats, storage etc)
- plan consent for sharing
- locate existing data
- collect data (experiment, observe, measure, simulate)
- capture and create metadata
Creating data – Issues

- Community-dependent
- Not always that structured (excel files, loose notes, recordings, missing metadata)
- Lacking awareness of existing data
Creating data – some RI answers

- Archive training
- Start organising early
- Sharing best practices
- Offer easy data discovery options
  - metadata search
  - content search
Life cycle: processing data

Processing data
- enter data, digitise, transcribe, translate
- check, validate, clean data
- anonymise data where necessary
- describe data
- manage and store data
Processing data – Issues

• invalid data (and/or formats)
• missing and incomplete metadata
• confidentiality (privacy)
Processing data – some RI answers

- Automatic data checks with useful feedback
- Semi-automatic metadata generation
  - balance between what the researcher knows and what is useful now
- Anonymisation support (non-trivial!)
- Efficient workflow to store data into a repository
  - in line with research workflow
Life cycle: analysing data

Analysing data

- interpret data
- derive data
- produce research outputs
- author publications
- prepare data for preservation
Analysing data – Issues

- derivation: reproducability (versioning, software used, ...)
- interpretation (existing data): often requires idiosyncratic insight
Analysing data – some RI answers

- support to document workflows + versioning
- allow to document ways of interpreting the data
Life cycle: preserving data

- Preserving data
  - migrate data to best format
  - migrate data to suitable medium
  - back-up and store data
  - create metadata and documentation
  - archive data
Preserving data – Issues

- archiving data often takes quite some additional (specialized) work
- creation of metadata / documentation: might be too late already
Preserving data – some RI answers

- easy-to-use interfaces
- Live archive principle: support for incremental archiving
- offer support in earlier stages to document the data
Life cycle: giving access to data

- creating data
- processing data
- analysing data
- preserving data
- re-using data
- giving access to data

- distribute data
- share data
- control access
- establish copyright
- promote data
Giving access to data – Issues

- sharing: mentality ("but my data is not ready yet")
- access control: often many technical hurdles
- data promotion: another drop in the tsunami?
Giving access – some RI answers

- community interaction, group training (peer support for sharing – archiving day)
- federated login
- smart ways of citing: embedding, part identifiers in publications
Life cycle: re-using data

Re-using data

- follow-up research
- new research
- undertake research reviews
- scrutinise findings
- teach and learn
Re-using data – Issues

- the thrill of re-inventing the wheel
- technical hurdles to get access to the “real” data
- lack of awareness
Re-using data – some RI answers

- making it really easy to re-use data
  - good index/exploration
  - address technical barriers from default work environment researcher
  - training and best practices