Science & Technology Facilities Council

Key attributes for a Research Data Infrastructure

Juan Bicarregui
Head of Data Division, STFC
RDA Organisational Advisory Board
EOSCpilot project Coordinator

- **Provenance and Persistence (PIds)**
  - For Data, Pubs, People, Organisations, Projects, etc.
    - for linking data with publications, people, organisations, projects. Etc.
  - Vision is to be able to build a graph between any objects
    - find all related objects
  - Baseline is the PID providers infrastructure,
    - For Pubs, Data, People, …
  - But is that enough?
    - Person with two papers and two datasets
      - which relates to which?
  - Do we need repositories of links
    - Or are repositories of PIds is enough?
    - Dynamic, on the fly build of graph to do a task
    - How do we find the reverse links

- **Effort and Efficiency**
  - Lowering the **human effort** required to open up data
  - Openness throughout the whole research process
    - rather than as something done later as part of publication.
  - Not just about motivation (later)
    - researchers can find the effort to write papers
  - Not just about automation
    - require the context
  - Metadata: built into supporting infrastructure
    - No effort beyond what researchers are doing anyway

- **Accessible and Assessable**
  - Making data **Assessable** as well as **Accessible**
  - Provide provenance – through PIds
    - PIds need semantics
    - Reproducible chains of processing
  - Validation
    - accreditation of repositories
    - accreditation of FAIRness of data?
  - Trusted
    - Tripadvisor or Wikipedia?
  - Quality v fit for use
    - Quality is not the same as Value
      - high quality data: 99.999% of entries are zero!

- **Culture and Credit**
  - *everyone* wants maximise benefit from research funds
    - Making a better world – together
    - Productive researchers
  - Funders want to maximise benefit
    - irrespective of who delivers it
  - Researchers want to maximise benefit
    - arising from *their* own projects

- **Effect and Evaluation**
  - What is **Quality**?
    - **Semantics** or **Provenance** or scientific **Value**
    - Have some very poor metrics for each of these.
  - **Effect** is **Impact?**
  - What is a sound **methodology** for making decisions?
    - any system can be gamed
  - Counting papers is hopeless
    - too many papers, varying standards
    - Impact Factor is approximate at best
  - Numerical comparison across fields is dangerous
    - Constantly changing – but hard to change.
  - Generational latency measured in decades.

Recap

- **P**rovenance and Pids
- **E**ffort/Efficiency
- **A**ccessible and Assessible
- **C**ulture and Credit
- **E**ffect/Evaluation