OPEN ACCESS TO SCIENTIFIC DATA:

WHY RESEARCHERS ARE RELUCTANT TO SHARE

Dr Dale Peters
Academic Computing
ICT
Outline

• Increasing acceptance of open access policies as a way to enhance global research
• Concerns of academics that encumber the full potential of OA
• Priority action areas for future development of global research data infrastructures
OA mandates

http://roarmap.eprints.org/
The genome generation

By comparing the elephant genome to the genomes of other mammals, scientists will be able to identify some of the oldest pieces of DNA that are common among placental mammals.

What you are seeing is the sequence of genetic letters that makes up the elephant genome. Shown here are 96 simultaneous reads from a single DNA sequencing machine. Each read provides 700-800 letters per hour. 117 sequencers running 24 hours a day generate billions of bases per week. The elephant genome is estimated to be 2.3 billion letters long and scientists will sequence and re-sequence it six or seven times to ensure accuracy.
Concerns of scholars

• Self-archiving an administrative burden
• Few rewards
• Separate from IRMA
• No services on top of repositories
• Usage metrics as quality research indicators
• No legal requirement – yet
• Responsible data management of primary and secondary research data
Repository Services

- OAI-PMH / OAI-ORE
- Reliable Metadata OAI_DC
- Best Practices for OAI_DC
- Compound Object Wrapping
- Vocabularies and Semantics
- Quality Labels
- Long Term Preservation
- Persistent Identifiers
- Usage Statistics Exchange
- Intellectual Property Rights (IPR)
Usage metrics

• Repository usage
  – Content (extent, author, subject, publication type)
  – Hits vs. downloads
  – COUNTER/ PIRUS / SUSHI

• Research impact
  – Journal Impact Factor
  – Citation analysis

• Research assessment
  – Enhance research performance?
• assessment of the impact of scholarly communication

• semantic store that cross-validates a range of bibliographic, citation and usage data
Metrics of scholarly status

Complexity: what are we citing?

- Journal
- Article
- Workflow
- Data
- Visualisation
- Model
- Annotation
- Concept

Attribution granularity

Legal requirement

Open Access Pilot in FP7

"If I have seen further it is by standing on the shoulders of giants." (Isaac Newton, 1676)
• Support the diffusion and adoption of the European Commission Open Access mandates
• Focus on publications in seven pilot disciplines - including datasets
• Orphan works repository supported by CERN
• Global dissemination
Overview

The NIH Public Access Policy ensures that the public has access to the published results of NIH funded research. It requires scientists to submit final peer-reviewed journal manuscripts that arise from NIH funds to the digital archive PubMed Central upon acceptance for publication. To help advance science and improve human health, the Policy requires that these papers are accessible to the public on PubMed Central no later than 12 months after publication.

How to Comply

All of your papers that fall under the NIH Public Access Policy, whether in press or in print, must include evidence of compliance in all of your NIH applications and reports.

1. **Determine Applicability**
   Does the NIH Public Access Policy apply to your paper?

2. **Address Copyright**
   Ensure your publishing agreement allows the paper to be posted to PubMed Central in accordance with the NIH Public Access Policy.

3. **Submit paper to PMC**
   Submit papers to PubMed Central (PMC) and approve public release. Enter your journal name into the box on the right side of the screen to determine how you paper will be posted to PMC.

4. **Include PMCID in Citations**
   Include the PMCID at the end of the full citation in your application or report.
Data management

• Policies and practice
• Creating & organising data
• Storage and access
• Back-up
• Preservation
• Sharing and re-use
Enhanced Publications

An enhanced publication is a compound data object which combines an “ePrint article” with:
– one or more metadata records
– one or more data resources
– or any combination of these
Compound Data Object
The diagram illustrates the migration process between different research domains. From the Private Research Domain, data is migrated to the Shared Research Domain and then to the Public Domain.

**Private Research Domain**
- Laboratory Information Management System/Research Management System
- May link to data objects stored in Research Data Store

**Shared Research Domain**
- Collaboration Support System (Plone, TWiki, Sharepoint)
- Involves object selection, migration, assignment of access controls, augmentation of metadata
- May link to data objects stored in Collaboration Data Store

**Public Domain**
- Institutional document repository
- May link to data objects stored in Public Data Store/Repository

**Migration Process**
- Authorised by research team leader
- Performed by research team IT support

**Private to Shared Migration**
- Performed by research team IT support

**Shared to Public Migration**
- Performed by research team IT support/IR staff
- Involves object selection, migration, assignment of persistent identifier, creation of fixity metadata, augmentation of other metadata

**Curation Boundary**
- This domain involves the core research team as they undertake the research, usually within a single institution. Access is often tightly controlled as hypotheses and analyses are developed.

- This domain involves researchers outside the core team as they collaborate with colleagues, often across institutions. Access is more open, but not everything is shared.

- This domain involves the public sphere (publication in the sense of making public). Access will usually be open to all.

“While many researchers are positive about sharing data in principle, they are almost universally reluctant in practice. ..... using these data to publish results before anyone else is the primary way of gaining prestige in nearly all disciplines.”

“Data sharing was more readily discussed by early career researchers.”
Desmond Tutu's genome sequenced as part of genetic diversity study
Archbishop Desmond Tutu has had his genome sequenced in research to reveal the true breadth of human genetic diversity

They have shared their data....
The majority of people felt that some form of policy or guidance was needed....
e-Infrastructure policy @ UKZN

- Research networks
- e-Infrastructure governance
- Authentication, authorisation, access
- Energy management
- Exa-scale computing
- e-Infrastructure services
- Data management

E-IRUG White Paper, 2011
High Level Overview of Centrally Managed Storage Utilization for UKZN

Diagram showing the storage utilization across different campuses with various data repositories and replication methods.
Priority action areas

• Open Access mandate
• Establish an e-Infrastructure governance model
• Enable technical solutions for open data platforms
• Recognise and promote data ownership, data management and attribution granularity
• Develop research skills in data analysis, data access and data sharing
References

• ROARMAP (Registry of Open Access Repository Material Archiving Policies) http://www.eprints.org/openaccess/policy/signup/
• Registry of Open Access Repositories (ROAR). http://roar.eprints.org/cgi/roar_search/advanced?location_country=&software=&type=&order=-recordcount%2F-date
Attribution-Noncommercial 2.5 South Africa

You are free:

to Share — to copy, distribute and transmit the work

to Remix — to adapt the work

Under the following conditions:

Attribution — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).

Noncommercial — You may not use this work for commercial purposes.