Preservation Tools, Techniques, and Policies IG: Initial Meeting

Co-Chairs: Mike Hildreth (Notre Dame), Ruth Duerr (Ronin Inst.) + ?
PTTP IG: Introduction

• Co-Chairs:

  – Ruth Duerr: Research Scholar, Ronin Institute, PI or Co-I on several NSF, NASA, and NOAA cyberinfrastructure and informatics grants, current President of AGU Earth and Space Science Informatics group

  – Mike Hildreth: Experimental Particle Physicist (CERN), PI of US-based DASPOS project, co-leader of various NSF workshops on open access to research data

  – + ? Looking for a non-US-based co-chair
Some Thoughts...

- This IG is intended to host a conversation that has been mostly lacking in the RDA discussions so far.
Some Thoughts...

- This session is intended to start a conversation that extends the RDA discussions so far.

Researcher/Data-Generator

Archivist/Data Scientist
Some Thoughts...

- This session is intended to start a conversation that extends the RDA discussions so far.
From our (Draft) Charter:

• What data/software/artifacts/documentation (“knowledge products”) should be preserved for sharing, re-use, and reproducibility for a given research domain? For other domains?

• What tools are available for researchers to preserve these elements in a manner that does not obstruct or hinder their research?
  – What are the strengths and weaknesses of these tools?
  – Are there common features that could allow tools from one domain to be re-used elsewhere?
  – Are there tools that archives/repositories could provide that could make preservation much easier for researchers?
  – What are the longer-term development goals of each of these tools?

• What preservation policies exist, imposed by government agencies, publishers, or other actors? How are they changing? How are they implemented? What are their strengths and weaknesses?

• How can preservation policies be implemented in a way that aids research both now and in the future?
  – How does this depend on the tools provided?
Tools & Techniques

• Preservation Tools & Techniques are essential for data sharing, curation, and reproducibility
  – you have to preserve the data before you share it!
  – decidedly non-trivial, even daunting for researchers

• Overlap between preservation and the needs of computational portability
  – e.g. if you can wrap up your workflow for remote execution, you can preserve it
  – potential synergies with data sharing here
Provocative(?) Statements

• No generic preservation tool(s) exists
  – workflows, data structures, use cases etc. are very discipline-specific

• But at some base level, all preservation is the same
  – must capture:
    • data generation and/or processing and filtering algorithms (sometimes called data provenance)
    • algorithm input parameters (metadata, workflow information)
    • repeat until result is obtained
  – capture should be “complete” for reproducibility
  – how do we find middle ground of commonality?
Provocative(?) Statements

• No tool will be adopted by researchers unless there is an “Economic Incentive” (enlightened self-interest)
  – tool makes doing science easier, more efficient
  – use of tools is mandated in order to obtain $$$
  – use of tools is mandated in order to publish
  – use of tools improves the training of students
  – re-use possibilities worth the effort
  – outreach/training worth the effort
Provocative(?) Statements

• Policies determining preservation and open access are often created without considering the impact on the researcher.

• Often different funding agencies within the same country have different policies.

What can we do to improve this situation?
Agenda for the Session

• Introduction/Overview (done)
• Talks:
  – Natalie Meyers
  – Jan Brase
• Discussion
  – Need additional co-chairs
  – Need to update charter as per TAB
Discussion

TAB Comments on Group Charter:

• We need to make the communities involved more obvious (which was hard as a startup we don’t really have a lot yet). The number of domains covered by the membership was not evident, beyond the earth and geo sciences, and attracting domain researchers may take some recruiting effort. The statement does not indicate which organizations or professional communities are represented among the leaders or membership, but there are many that could make vital contributions, such as the very active international iPRES community, the Open Preservation Foundation, and the National Digital Stewardship Alliance and Data-Pass in the US.

• Need a clear timeline for the cataloging activity as well as to ensure that it captures previous work in this area. The catalogue activity appears to be a priority but the timeline is not clear and previous related work is not discussed. (for example: COPTR http://coptr.digipres.org/Main_Page and the POWRR tool grid http://digitalpowrr.niu.edu/tool-grid/)