Data Management Considerations for the Data Life Cycle

NRC STS Panel 2011

November 17, 2011, Washington DC

Peter Fox (RPI) foxp@rpi.edu, pfox@cs.rpi.edu
Tetherless World Constellation http://tw.rpi.edu

Working premise

Scientists – actually ANYONE - should be able to access and use a global, distributed knowledge base of scientific data that:

- appears to be integrated
- appears to be locally available

But… data and information is obtained by multiple means (instruments, models, analysis) using various (often opaque) protocols, in differing vocabularies, using (sometimes unstated) assumptions, with inconsistent (or non-existent) meta-data. It may be inconsistent, incomplete, evolving, and distributed AND created in a form that facilitates generation, not use (except by accident)

And … significant levels of semantic heterogeneity, large-scale data, complex data types, legacy systems, inflexible and unsustainable implementation technology…

Uh-oh

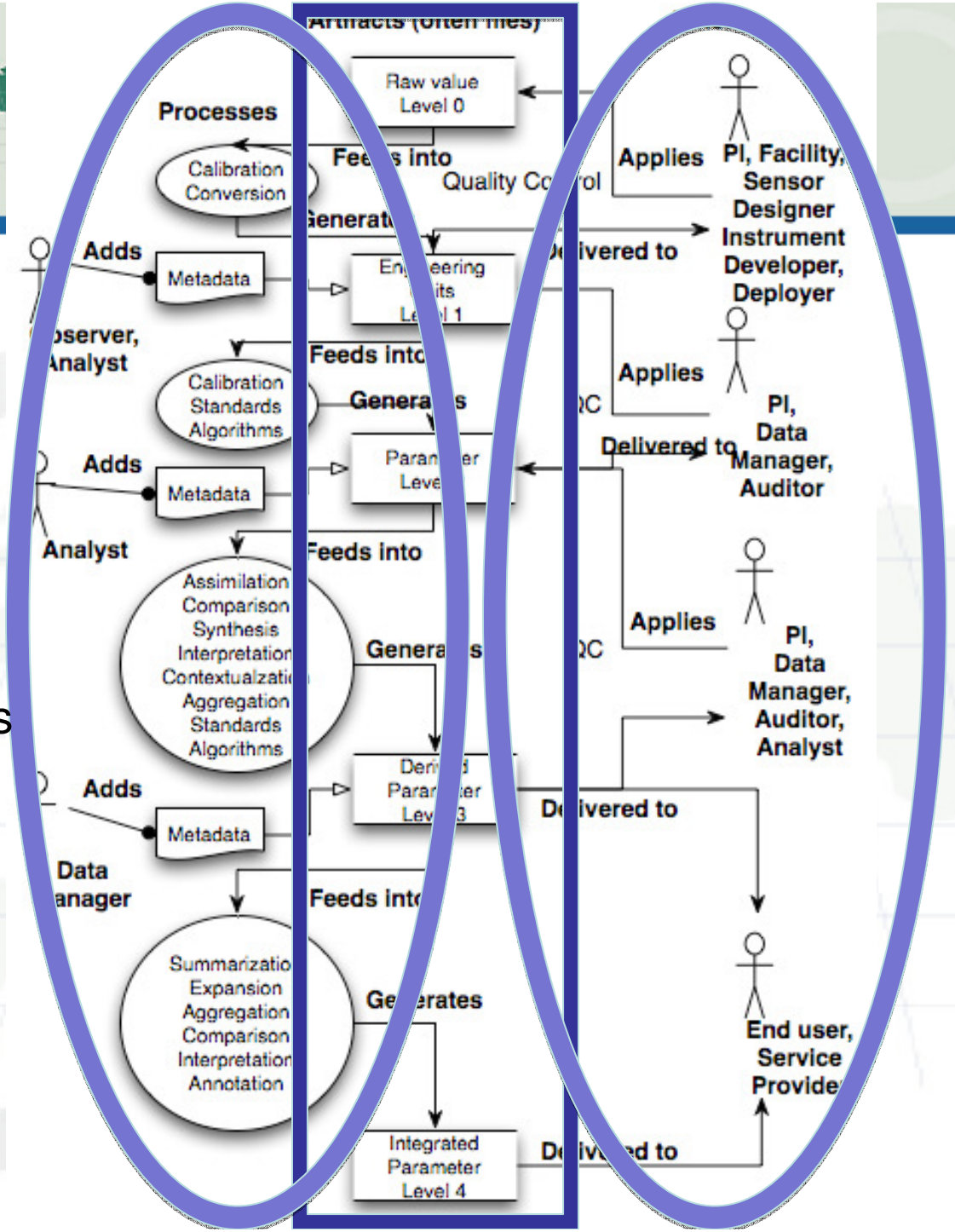
Life cycle – overused?

- *Data Life Cycle*: The data life cycle is a term coined to represent the entire process of data management.
- It starts with concept study and data collection, but importantly has **no end**, as data is continually repurposed, creating new data products that may be processed, distributed, discovered, analyzed and archived.
- Fully supporting the different steps in the life cycle puts demands on metadata, standards, tools and *people*.

Definition – simply put

- Data life-cycle elements (simple 3-level)
  1. Acquisition: Process of recording or generating a concrete artefact from the concept (see transduction)
  2. Curation: The activity of managing the use of data from its point of creation to ensure it is available for discovery and re-use in the future
  3. Preservation: Process of retaining usability of data in some source form for intended and unintended use
  - ❑Stewardship: Process of maintaining integrity across acquisition, curation and preservation + arranging for discovery, access and use of data, information and all related elements.
    - Involves fiscal and intellectual responsibility

Curation stages

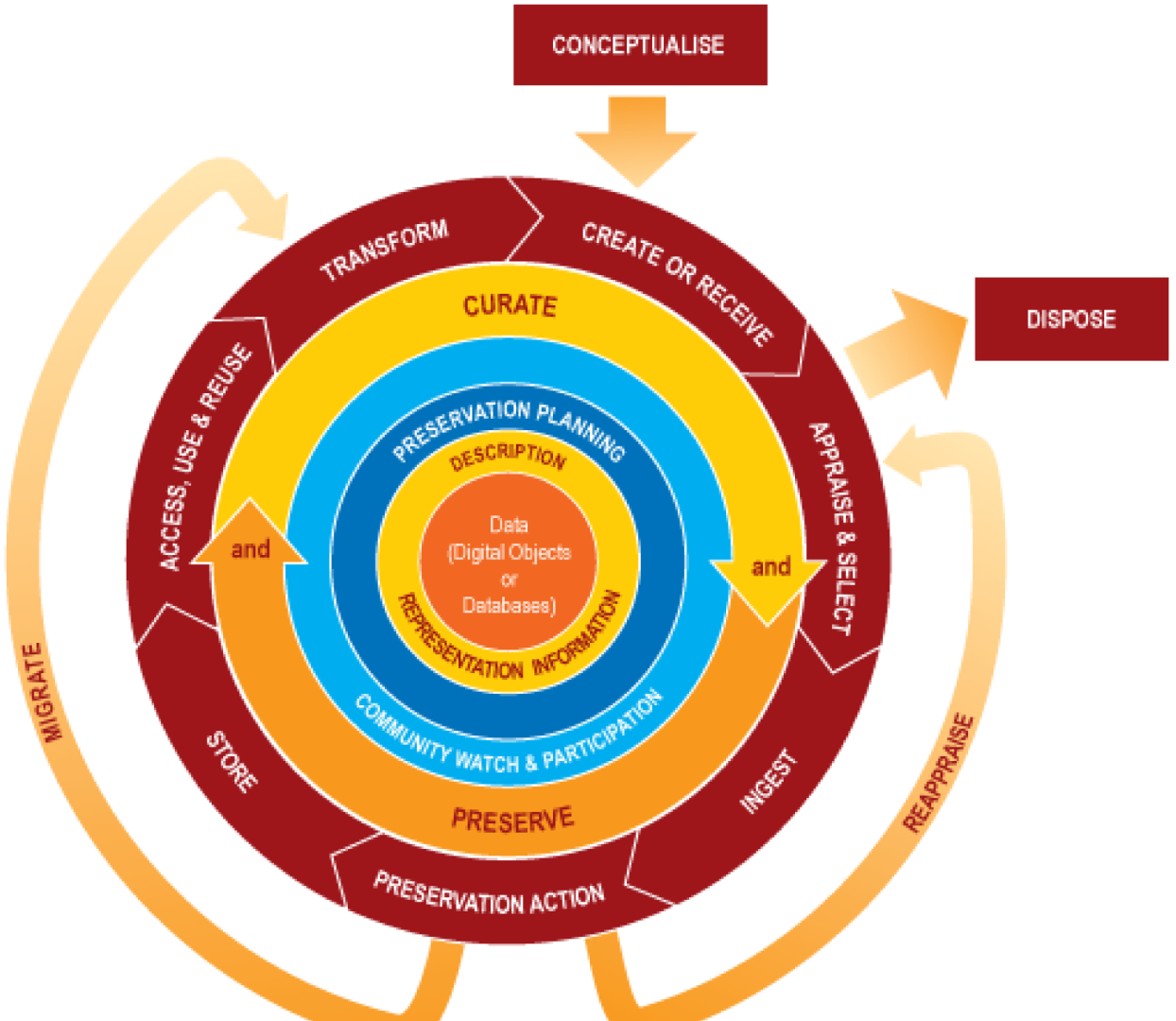


People!

People, organizations, norms – not so simply put

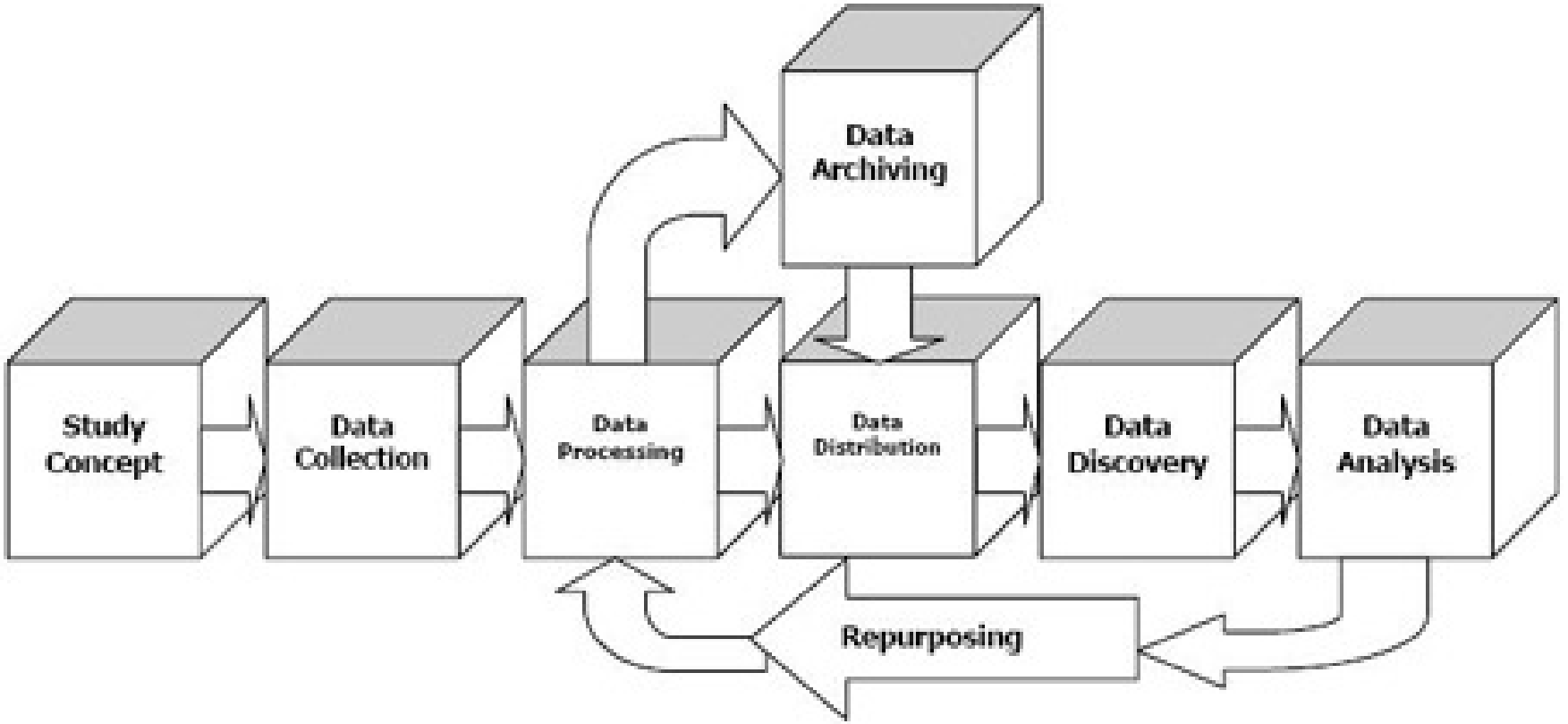
- Even cast into three broad stages
  - ✧ Acquisition/ Curation/ Preservation: The people/ roles, their organizations, means of generation, what matters to them can be very different across and (especially) between stages of the life cycle, even if they overlap
  - ✧ Stewardship: So how is an effective and coherent *process of maintaining integrity across acquisition, curation and preservation + arranging for discovery, access and use of data, information and all related elements,* achieved?
    - Especially as it *Involves fiscal and intellectual responsibility*
  - ❖ *Sustainability= resources + social + organizational constructs*

Digital Curation Center model



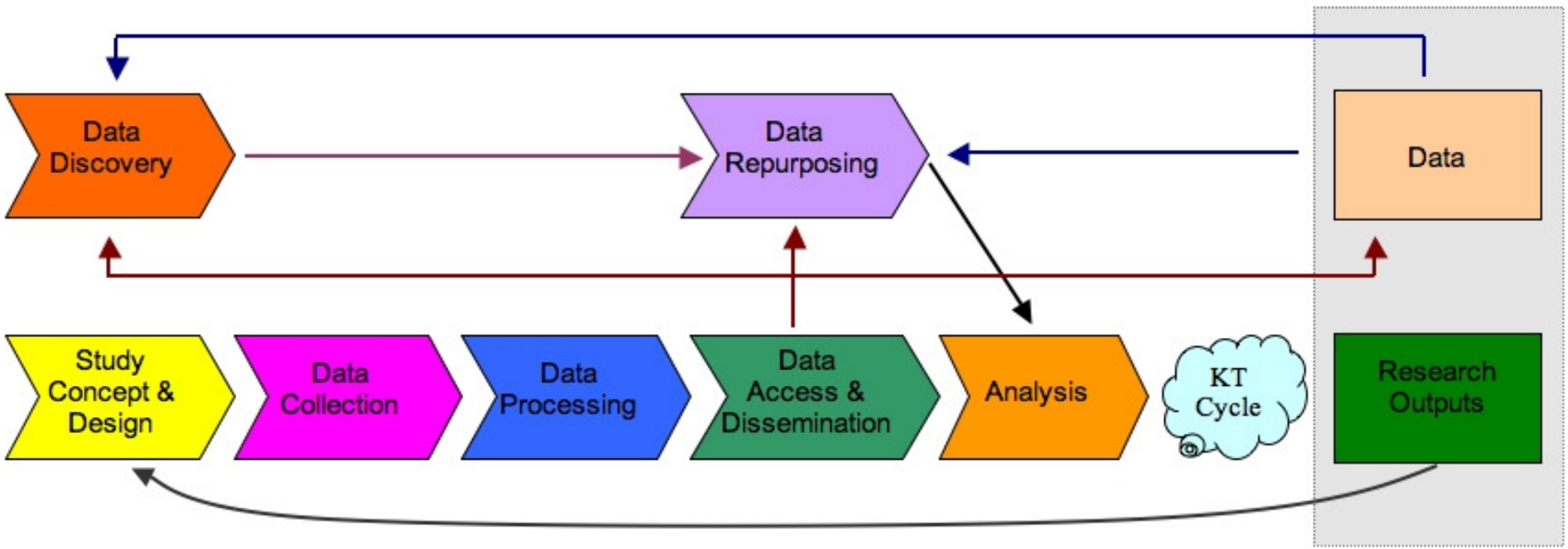
http://www.dcc.ac.uk/resources/curation-lifecycle-model

MIT DDA Alliance model

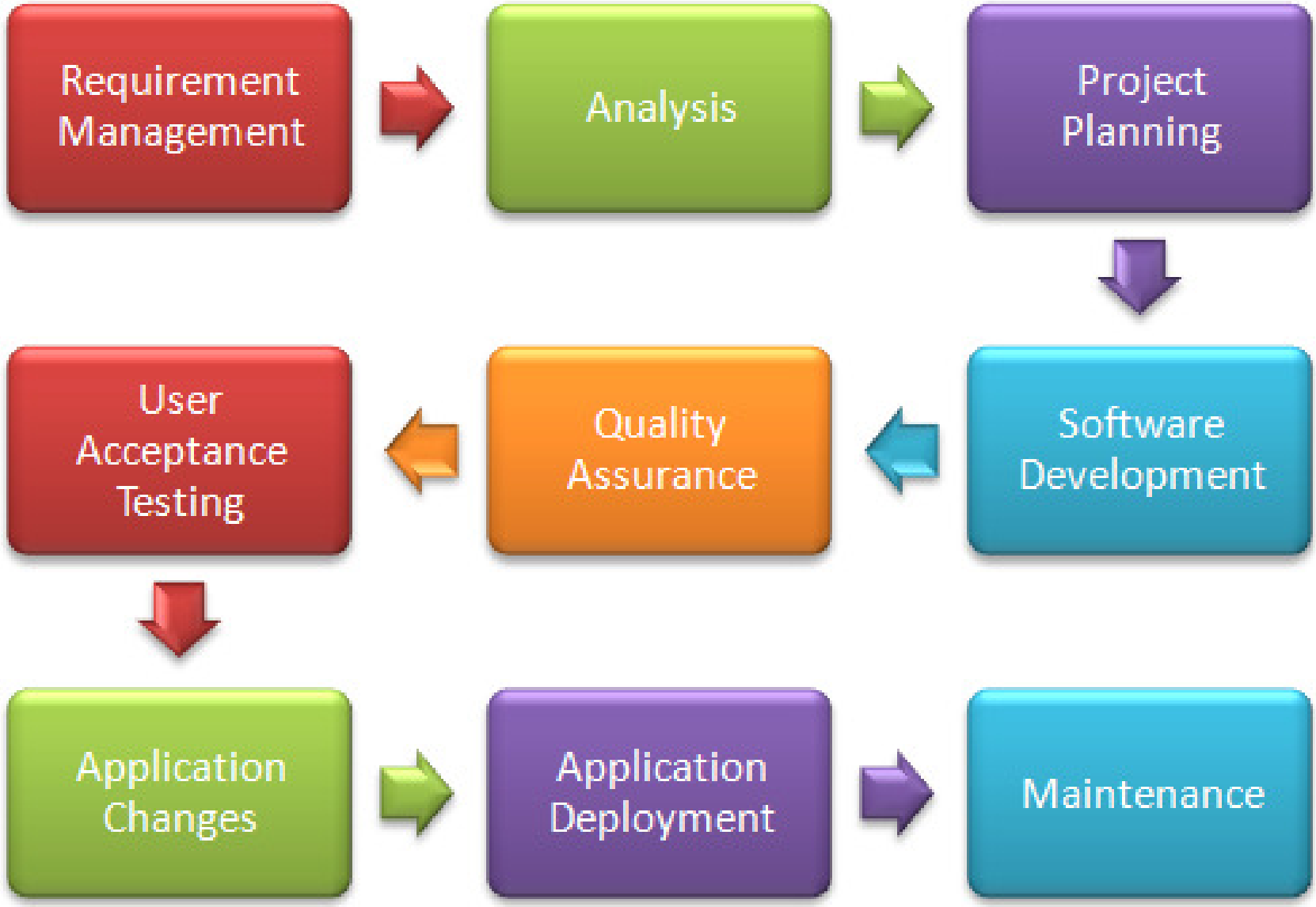


http://libraries.mit.edu/guides/subjects/data-management/cycle.html

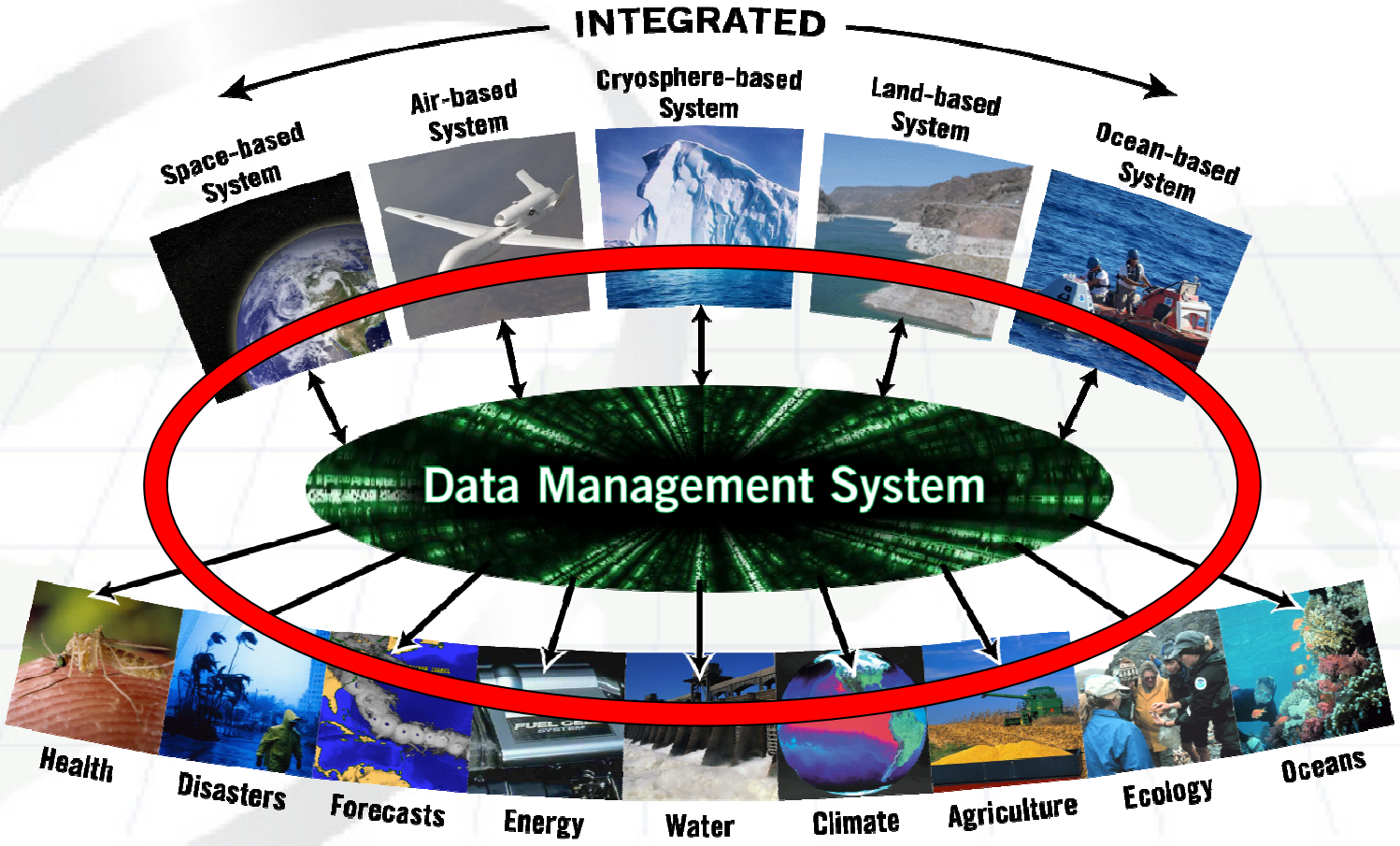
It does not go on forever…



Business or software model?

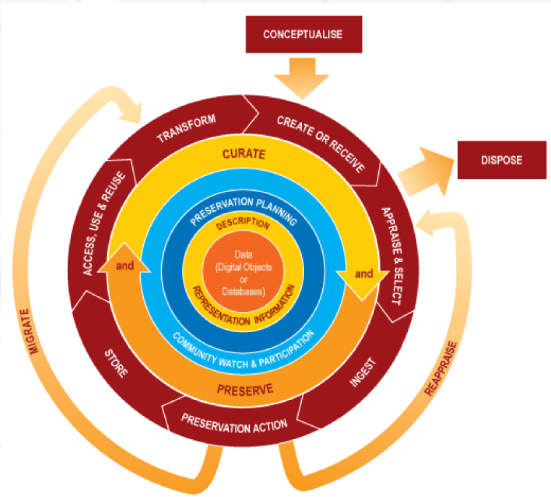


Can we really fulfil futures with diagrams?



GeoData 2011

- *Metadata at All Stages of the Data Life Cycle*
- *Best Practices in Data Life Cycle Management*
- *The Human Factor*
- *Training and Outreach*
- *Technology gaps*
- *A business model for the longer term*



http://tw.rpi.edu/web/Workshop/Community/GeoData2011 (NSF, USGS, NASA, NOAA, …)

The Human Factor in Data Life Cycle Management

- Tendency of people to continue working the way they have in the past: the inertia of habit.
  - Leads to slow growth in data submissions to archives.
- Both carrots (such as wider data use and data citations) and sticks (contingent funding and contingent publication of papers) were suggested as potential motivators to change this inertia.

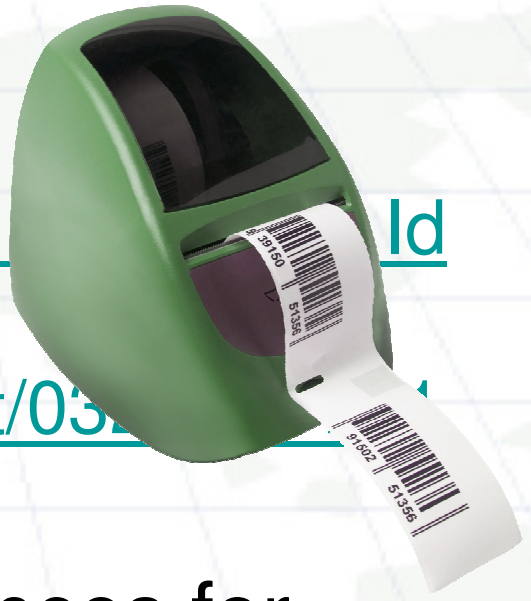
Personal view

- Over-emphasis on preservation
- Heavier, rather than supportable or scalable, approaches to curation
- Significant attention deficit disorder to acquisition

- Why – value? Return on investment but those that need to perform the key data management functions (across life cycle)
- A-I-G coalitions!

E.g. Identifiers

- Idea: not to pre-empt an implementation
  - DOI = http://www.doi.org/, e.g. 10.1007/s12145-008-0001-8
  - URI, http://en.wikipedia.org/wiki/Uniform_Identifier e.g. http://www.springerlink.com/content/03[illegible]338n85/fulltext.pdf



- But – has fundamental consequences for data in the life cycle (sharing, annotation)

To be sustainable…

- Bridge, or eliminate disconnects between people, organizations, norms
  - Yes, it's so easy to write
- Problem: the stakeholders are not all at the table…
- Lightweight solutions, e.g. linked data have a lot of potential but terrify data curators and preservationists…
- Heavyweight solutions, e.g. full life cycle OAIS approaches are robust Lightweight solutions, terrify (data) generators

Temptation

- To run screaming from the room?
  - Been there, done that

Or

- Look for a GOOD solution one that addresses the goal of the virtual organization across the life cycle (they exist ‘in the small’)
- Focus on value – the real and immediate value to the people their institutions/ communities and funders

Thanks for listening

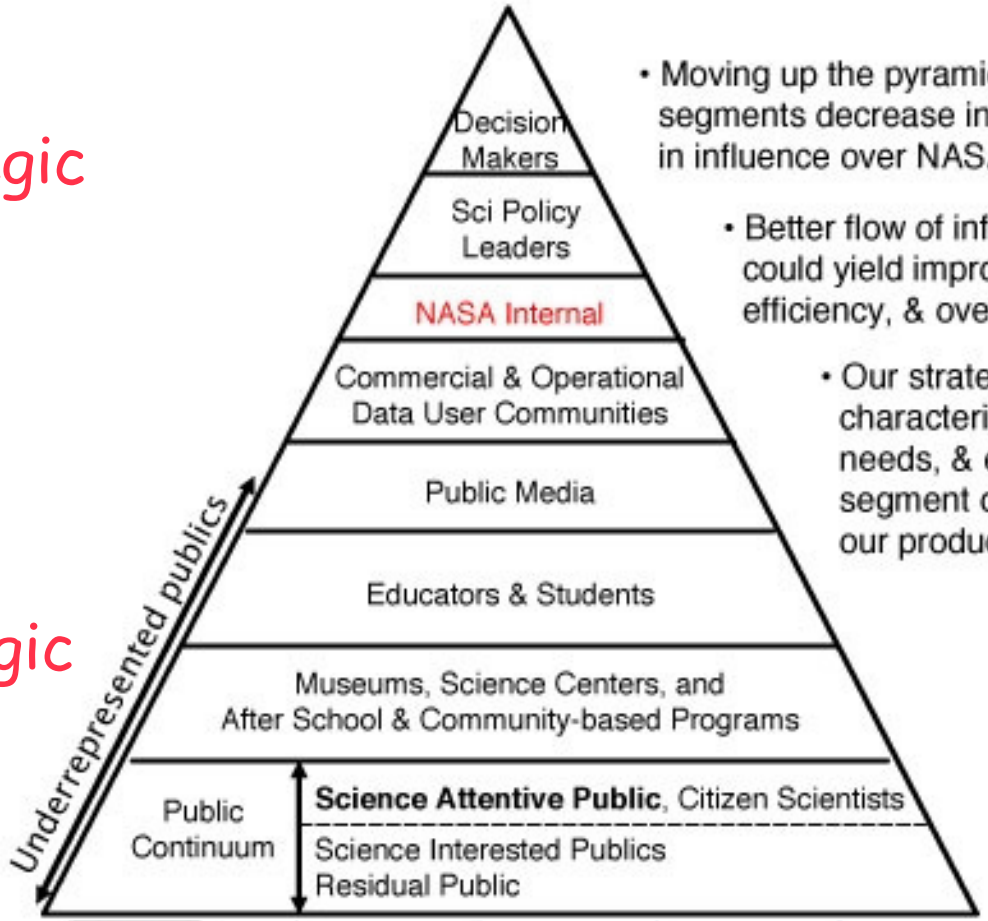
- pfox@cs.rpi.edu and @taswegian
- http://tw.rpi.edu

- Questions? Comments?

.. Data has Lots of Audiences

More Strategic

Less Strategic



- Moving up the pyramid, these audience segments decrease in size while increasing in influence over NASA's budget & agenda.
- Better flow of information within NASA could yield improved synergy, efficiency, & overall effectiveness.
- Our strategy should be to characterize the information wants, needs, & expectations of each segment of the public, & then tailor our products/programs accordingly.
- The 'science attentive' public considers itself knowledgeable & willing to participate in policy-relevant discussions. Thus, this is a particularly beneficial audience to target.

From "Why EPO?", a NASA internal report on science education, 2005

Science too!

Recommendations (1)

- Develop review criteria for data management plans that recognize the criticality of upfront planning, design of metadata and data design, as well as the importance of enabling future repurposing of the data.
- Develop proposal and review criteria for instrument development that ensures adequate metadata collection at the time of measurement.
- Develop a searchable repository for best practices in data lifecycle management for capturing best practices noted in Geodata 2011, the NSF Research Data Lifecycle Management Workshop in July 2011, and related work in the Earth Science Information Partners and similar organizations.

Recommendations (2)

- Identify Data Life Cycle Communities of Practice within NSF programs as well as other organizations such as ESIP and American Geophysical Union.
- Develop methods, such as workshops, AGU special sessions, etc. for bringing together overlapping Communities of Practice for information exchange.
- Develop domain-specific tools for adding and editing metadata, particularly within ISO metadata standards.

Recommendations (3)

- Develop incentives (both carrots and sticks) to induce data providers to develop metadata and data products that will be usable by both narrow initial users of data and the wider community of interdisciplinary users reusing data for other purposes.
- Develop curricula targeted at both the practicing researcher and science students in data science, including data collection, management and integration.
- Continue supporting work in recording provenance of data, while shifting some resources to the practical application of this work.

Recommendations (4)

- Support research and applications aimed at improving cross-disciplinary and inter-disciplinary discovery of data and related services.
- Develop Business Models for persistent long-term archives that take into account the funding cycle as well as the data life cycle.