Rensselaer

why not change the world?®

Scientific Workflows & Data Stewardship

Ahmed Eleish

Data Science – ITWS/CSCI/ERTH-4350/6350

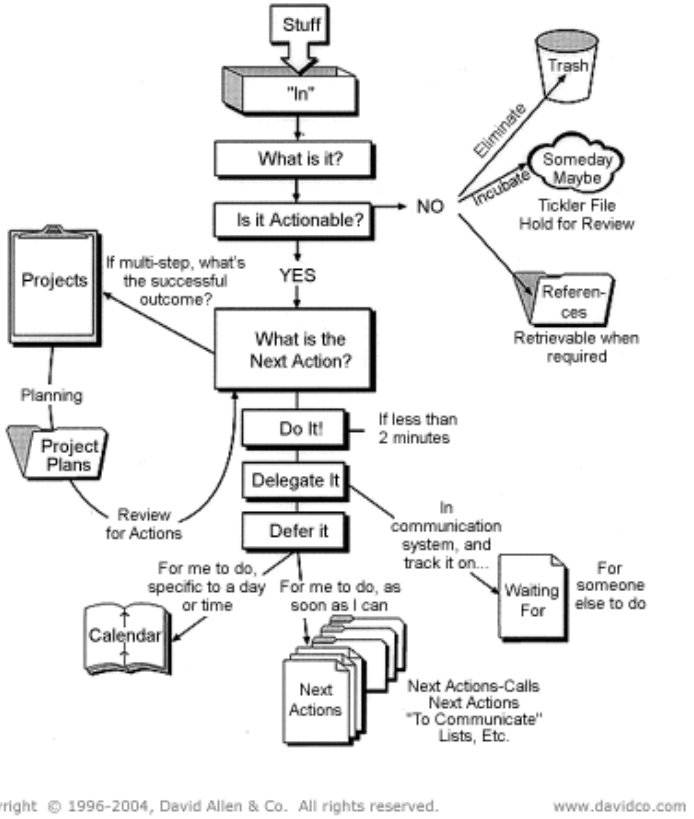
November 21th, 2024

Tetherless World Constellation
Rensselaer Polytechnic Institute

Contents

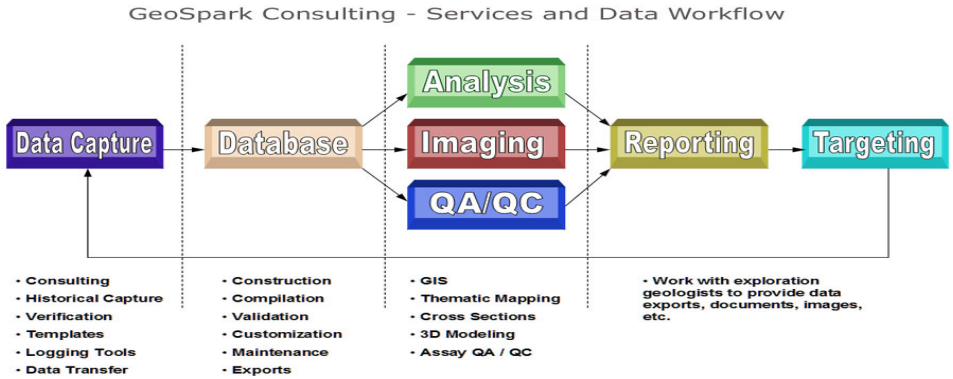
- Scientific Data Workflows
- Data Stewardship
- Summary

- Projects



Scientific Data Workflows

- What are they?
- Why use them?
- Some more detail in the context of Kepler
  - www.kepler-project.org
- Some pointers to other workflow systems



Scientific Data Workflows – Why Would You Use Them

Large Hadron Collider generates 15 petabytes (15 million gigabytes) of data annually...

The LHC: One of the world's most complex data systems

The $3.6 billion Large Hadron Collider (LHC) will sample and record the results of up to 600 million proton collisions per second, producing roughly 15 petabytes (15 million gigabytes) of data annually in search of new fundamental particles. To allow thousands of scientists from around the globe to collaborate on the analysis of these data over the next 15 years (the estimated lifetime of the LHC), tens of thousands of computers located around the world are being harnessed in a distributed computing network called the Grid. Within the Grid, described as the most powerful supercomputer system in the world, the avalanche of data will be analyzed, shared, re-purposed and combined in innovative new ways designed to reveal the secrets of the fundamental properties of matter.

LHC source: public.web.cern.ch/public/en/LHC/LHC-en.html
Source: public.web.cern.ch/Public/en/LHC/LHC-en.html

Scientific Data Workflows – Why Would You Use Them

With the acquisition of the human genome sequence and the advent of powerful new DNA sequencing technologies and analytical methods, it is increasingly possible to identify variations in human DNA that underlie particular diseases, conditions, or therapeutic responses. The National Center for Biotechnology Information (NCBI) has developed the database of Genotype and Phenotype (dbGaP) to preserve and distribute the results of studies employing these powerful new capabilities. The database represents the combined power of many different types of studies and analyses. As a result, clinicians and scientists from many fields can share their results and work together to investigate the interaction of genotype and phenotype, revealing new links between DNA sequence and a variety of diseases, from breast cancer to diabetes, blood pressure abnormalities, and age-related eye defects.

Source: ncbi.nlm.nih.gov/dbgap

Scientific Data Workflows – Why Would You Use Them

Tsunami related data archive has gone from 5 GB to 1700+ GB , with standards compliant metadata online to support the modeling, mapping and assessment activities required to minimize the effect of tsunamis.

NOAA's DART™ Tsunami Monitoring Buoys

As part of the U.S. National Tsunami Hazard Mitigation Program (NTHMP), the National Oceanic and Atmospheric Administration (NOAA) has developed and placed Deep-ocean Assessment and Reporting of Tsunamis (DART™) stations in regions with a history of generating destructive tsunamis to ensure early detection of tsunamis and to support real-time warnings. Currently DART™ stations are deployed and active in the Pacific, Atlantic and Indian Oceans, the Caribbean Sea, and the Gulf of Mexico.

The tsunami-related data archive has grown from five gigabytes to over 1,700 gigabytes, with standards-compliant metadata available online to support the modeling, mapping, and assessment activities required to minimize the effect of tsunamis.

Source: http://nctr.pmel.noaa.gov/Dart/dart_home.html

Scientific Data Workflows – Why Would You Use Them

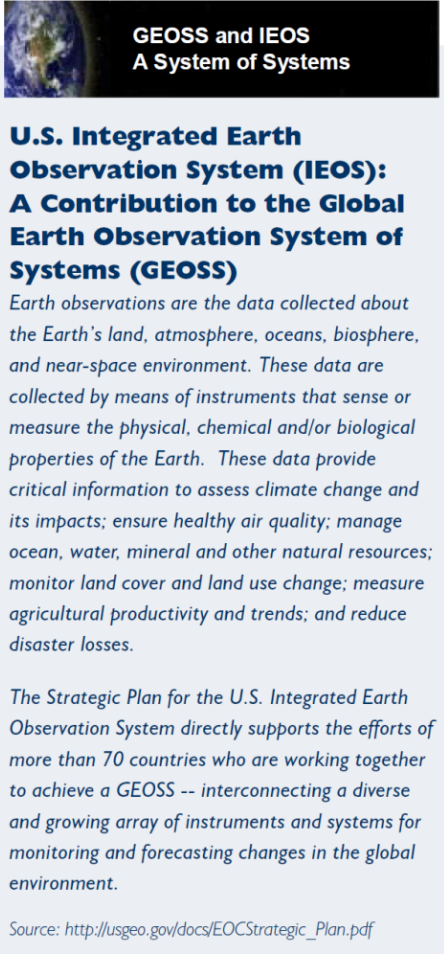
Barcode of Life

The Barcode of Life Initiative is an international effort to develop reliable and authoritative means for the global identification of biological species. Barcoding uses a short DNA sequence within an organism's genome as the equivalent of a barcode on a supermarket product to determine the species origin of a biological sample. Adoption of a standard format for barcode data allows a sample in a museum or collected in the field to be instantly linked to related information resources worldwide; to be tied in to relevant tissue, parasite, and other collections globally; and to reference DNA databases in the United States, Japan, and Europe. The result is the ability to conduct biodiversity, species migration and invasion, and population genetics studies that are more powerful because they can be reliably compared to and informed by other projects worldwide.

Source: http://barcoding.si.edu/

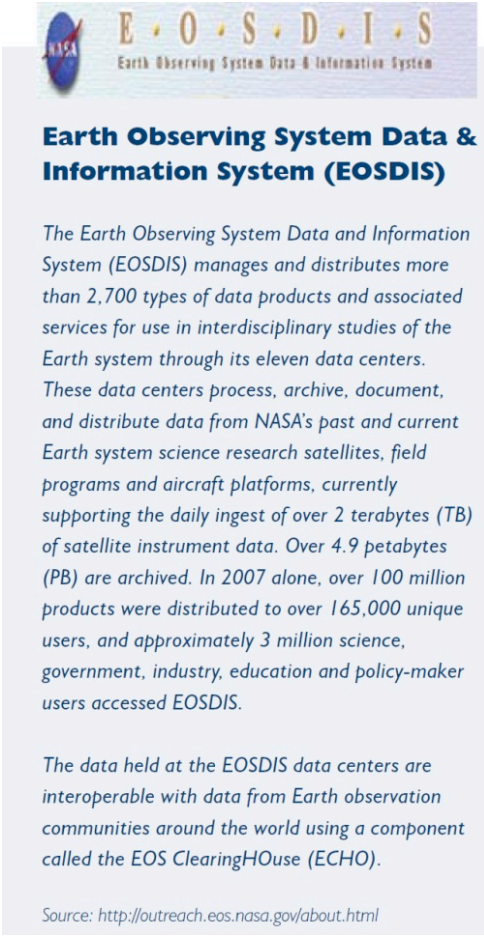
Scientific Data Workflows – Why Would You Use Them

The Strategic Plan for the U.S Integrated Earth Observation System directly support the efforts of more than 70 countries who are working together -- interconnecting a diverse and growing array of instruments and systems for forecasting changes in the 15 global environment



Scientific Data Workflows – Why Would You Use Them

Daily ingest of 2+ terabytes (TB) of Satellite instrument data... Over 4.9 petabytes (PB) are archived.



Scientific Data Workflows – Why Would You Use Them

IPUMS International

Digital Data Importance to Social and Behavioral Sciences

The study of powerful large-scale trends such as economic development, urbanization, expanding migration, population aging, and mass education by social, behavioral, and other scientists requires access to global-scale micro-data – data about individuals, households, and families collected by census offices around the world. The Integrated Public Use Microdata Series (IPUMS) provides researchers and educators with interoperable access to data from more than 111 censuses in 35 countries representing more than 260,000,000 person records. This powerful digital collection meets critical research needs while successfully preserving appropriate privacy and confidentiality rights, allowing researchers to construct frameworks for analyzing and visualizing the world's population in time and space to understand agents of change, to assess their implications for society and the environment, and to develop policies and plans to meet future challenges at local, regional, national, and global scales.

For additional information, see: https://international.ipums.org/international/

What were the key points in “Scientific Data Management”?

Scientific Data Management

Someone said: “We are drowning in data, but starving of information”. And this is particularly true for scientific data. This happens also for business data, but here they had more time to learn. They implemented data architectures, created data warehouse and used data mining to extract information from their data. So why don't study and implement something similar for scientific data? The solution can be to setup a Scientific Data Management architecture.

Scientists normally limit the meaning of Data Management to the mere physical data storage and access layer. But the scope of Scientific Data Management is much boarder: it is about **meaning** and **content**.

Below I listed common problem and opportunities in scientific data access. Then I collected what are considered the parts of a Data Management solution. A list of references and examples of data access and scientific data collections follow.

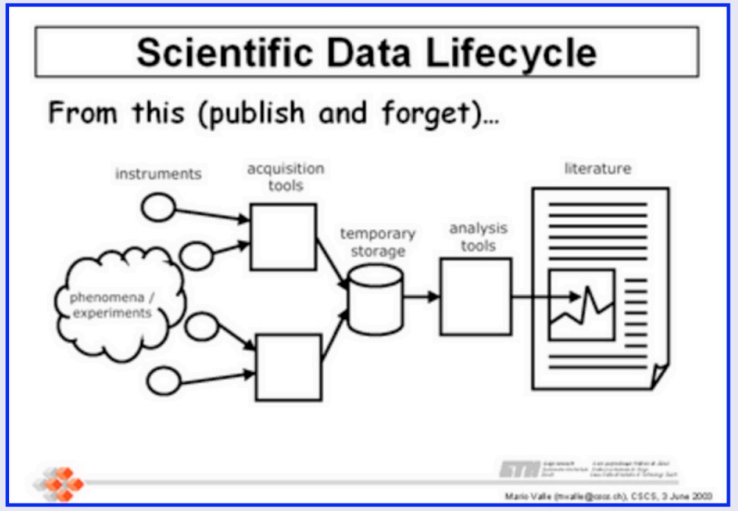
The paper ends with more implementation oriented issues: a survey of some scientific data formats, planning for a possible implementation and a survey of the supporting technologies available.

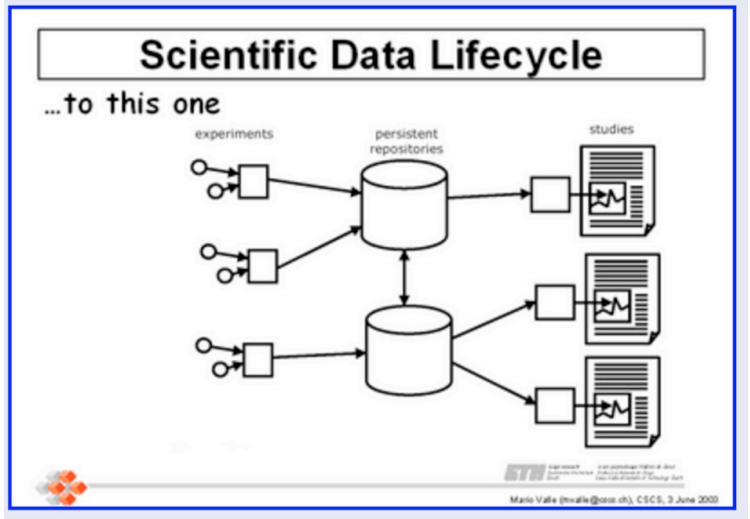
Most of this paper notes and information have been collected and studied for one specific project. But really the ideas collected are generally applicable to the kind of scientific projects that uses the CSCS computational and visualization services.

We are drowning in Data, but starving of Information...

Problems that can be found in current scientific projects include:

- Limited file and directory naming schemes. Some project data repositories are simply big flat directories (**No Logical Organization**!)
- **No access to important metadata in scientists' notebooks and heads.**
- **Un-owned data with dubious content after the end of project or PhD thesis.**





What is a workflow?

- General definition: **series of tasks performed to produce a final outcome**
  - E.g. *following a recipe to bake a pie*
- Scientific workflow – “data analysis pipeline”
  - Automate tedious jobs that scientists traditionally performed by hand for each dataset
  - Process large volumes of data faster than scientists could do by hand

Background: Business Workflows

- Example: planning a trip
- Need to perform a series of tasks: book a flight, reserve a hotel room, arrange for a rental car, etc.
- Each task may depend on outcome of previous task
  - Days you reserve the hotel depend on days of the flight
  - If hotel has shuttle service, may not need to rent a car

e.g. tripit.com

Workflow Elements

- **Task executions** – Specify a series of tasks to run – Outputs from one task may be inputs for another

- **Task scheduling** – Some tasks may be able to run in parallel with other tasks

What about scientific workflows

- Perform a set of transformations/ operations on a scientific dataset
- Examples (Data Mining examples)
  - Generating images from raw data
  - Identifying areas of interest in a large dataset
  - Querying a web service for more information on a set of objects
  - Many others...

More on Scientific Workflows

- Formal *models* of the *flow* of data among processing components
- May be simple and linear, or more complex
- Can process many data types:
  - Archived data
  - Streaming sensor data
  - Images (e.g., medical or satellite)
  - Simulation output
  - Observational data

Challenges

- Questions:

– What may be some challenges for scientists implementing scientific workflows?

– What are some challenges to executing these workflows?

Challenges

- Mastering a programming language
- Visualizing the workflow, end-to-end
- Sharing/exchanging that workflow, updating it
- Dealing with data formatting
- Locating datasets, services, or functions (input and outputs, metadata...)
- Example: What really happened to the software on the Mars Pathfinder spacecraft?

https://www.rapitasystems.com/blog/what-really-happened-to-the-software-on-the-mars-pathfinder-spacecraft

e.g. Kepler Scientific Workflow Management System

- Graphical interface for developing and executing scientific workflows
- Scientists create workflows by dragging and dropping
- Automates low-level data processing tasks
- Provides access to data repositories, compute resources, workflow libraries

Benefits of Scientific Workflows

- Documentation of aspects of analysis is much esier
- Visual communication of analytical steps helps the user
- Ease of testing/debugging
- Reproducibility
- Reuse of part or all of workflow in a different project

Benefits of Scientific Workflows

- Integration of multiple computing (deployment) environments
- Automated access to distributed resources via web services and Cloud technologies

Why not just use a script?

- Scripts do not (usually) specify low-level task scheduling and communication
- Often are very platform-dependent
- Can’t be easily reused
- May not have sufficient documentation to be adapted for another purpose

Why is a GUI useful?

- No need to learn a programming language
- Visual representation of what workflow does • Allows you to monitor workflow execution
- Enables user interaction
- Facilitates sharing of workflows

The Kepler Project - https://kepler-project.org/

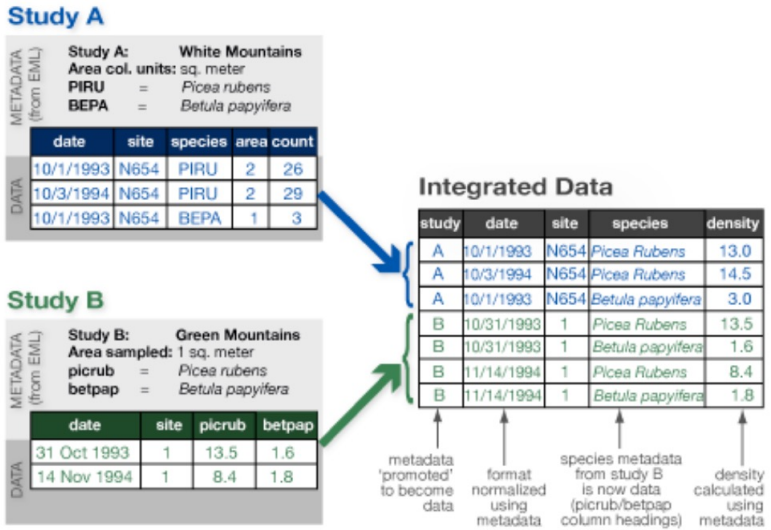
- Goals
  - Produce an open-source scientific workflow system
    - enable scientists to design scientific workflows and execute them
  - Support scientists in a variety of disciplines
    - e.g., biology, ecology, astronomy
  - Important features
    - access to scientific data
    - flexible means for executing complex analyses
    - enable use of Grid-based approaches to distributed computation
    - semantic models of scientific tasks
    - effective UI for workflow design

The Kepler Project - https://kepler-project.org/

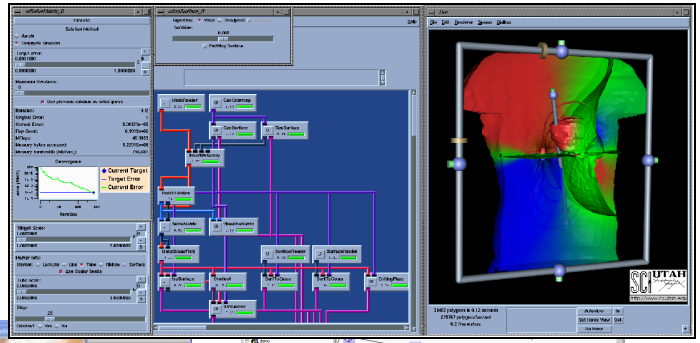
- Opportunities for parallel execution
  - Fine-grained parallelism
  - Coarse-grained parallelism
- Current ‘plumbing’ approaches to distributed execution
  - workflow acts as a controller
    - stages data resources
    - writes job description files
    - controls execution of jobs on nodes
  - requires expert understanding
- Scientists need to focus on just the computations
  - try to avoid plumbing as much as possible

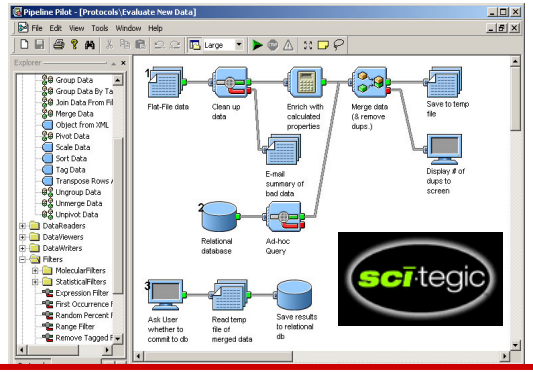
Managing Data Heterogeneity?

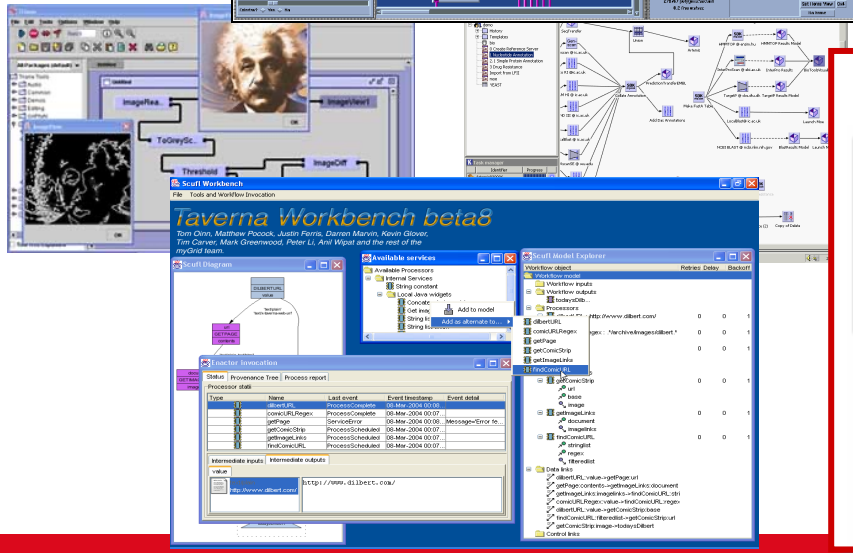
- Data comes from heterogeneous sources
  - Real-world observations
  - Spatial-temporal contexts
  - Collection/measurement protocols and procedures
  - Many representations for the same information (count, area, density)
  - Data, Syntax, Schema, Semantic heterogeneity
- Discovery and “synthesis” (integration) performed manually
  - Discovery often based on intuitive notion of “what is out there”
  - Synthesis of data is very time consuming, and limits use

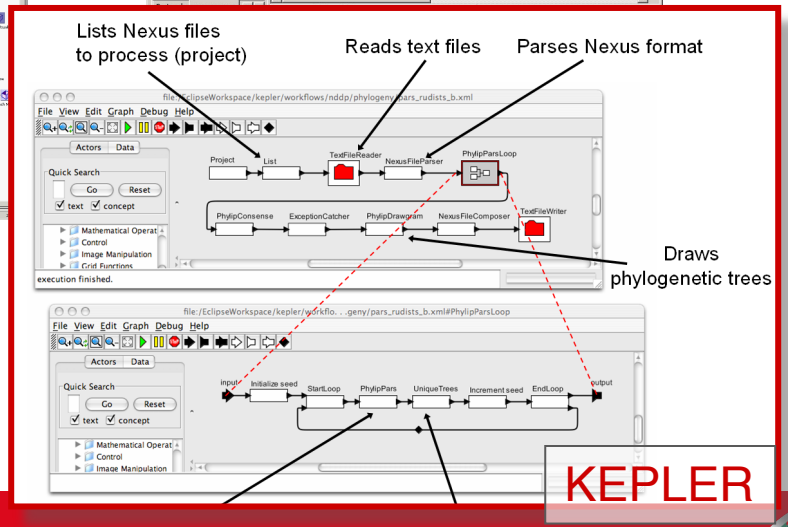


Scientific workflow systems support data analysis

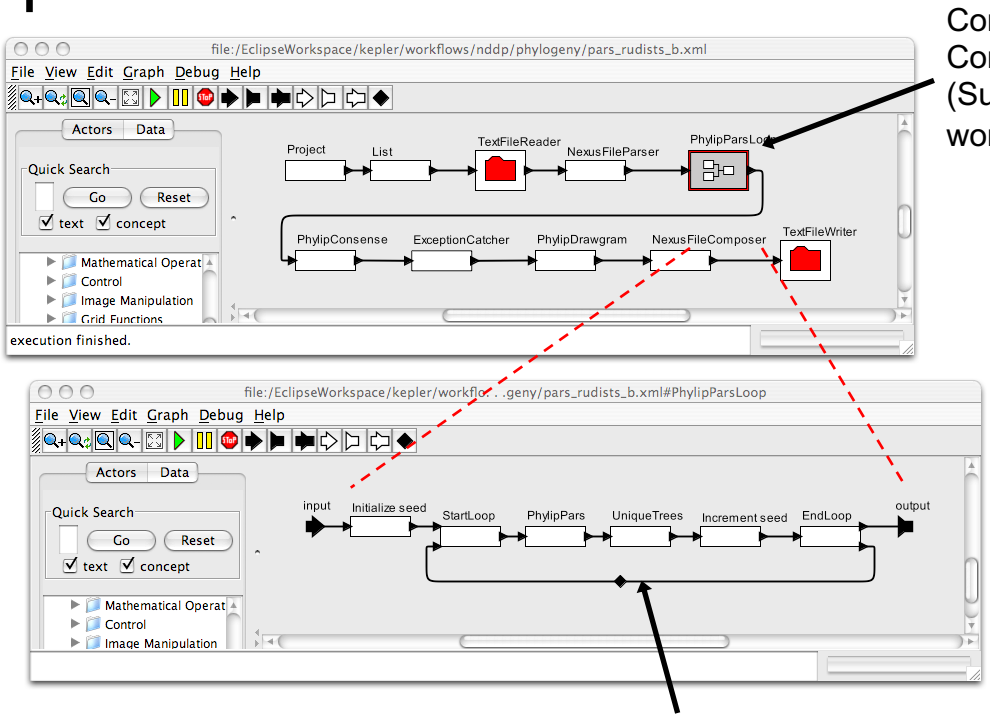








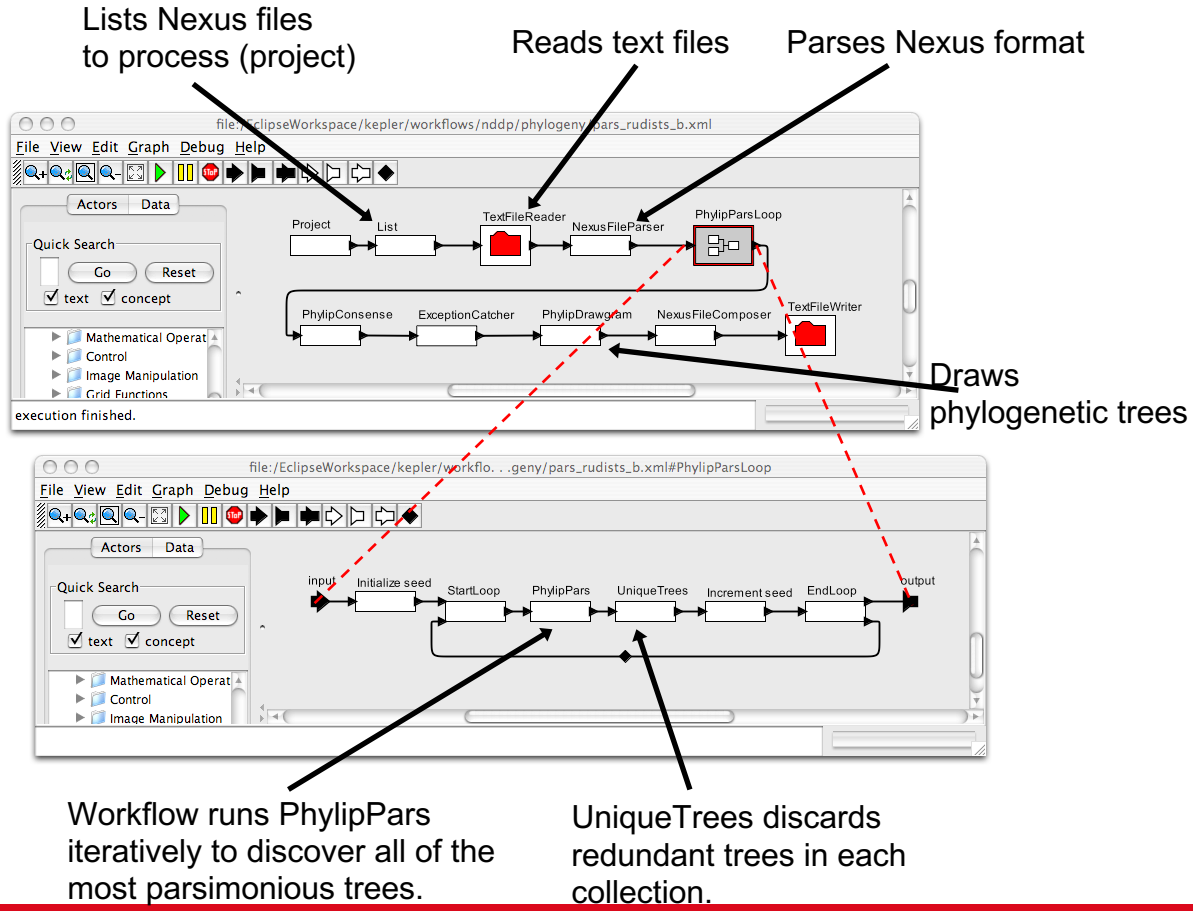
A simple Kepler workflow



(T. McPhillips)

Loops often used in SWFs; e.g., in genomics and bioinformatics (collections of data, nested data, statistical regressions, ...)

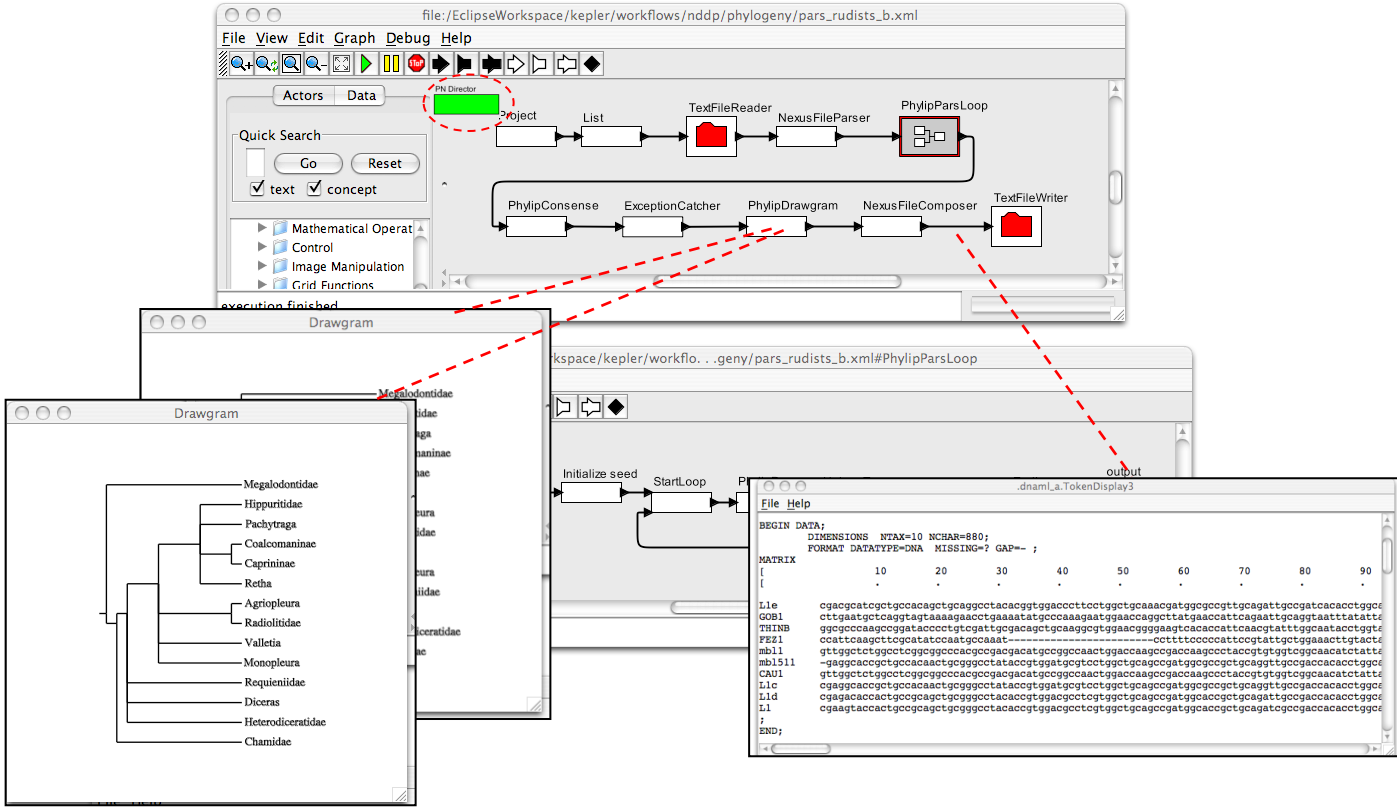
A simple Kepler workflow



(T. McPhillips)

A simple Kepler workflow

An example workflow run, executed as a Dataflow Process Network



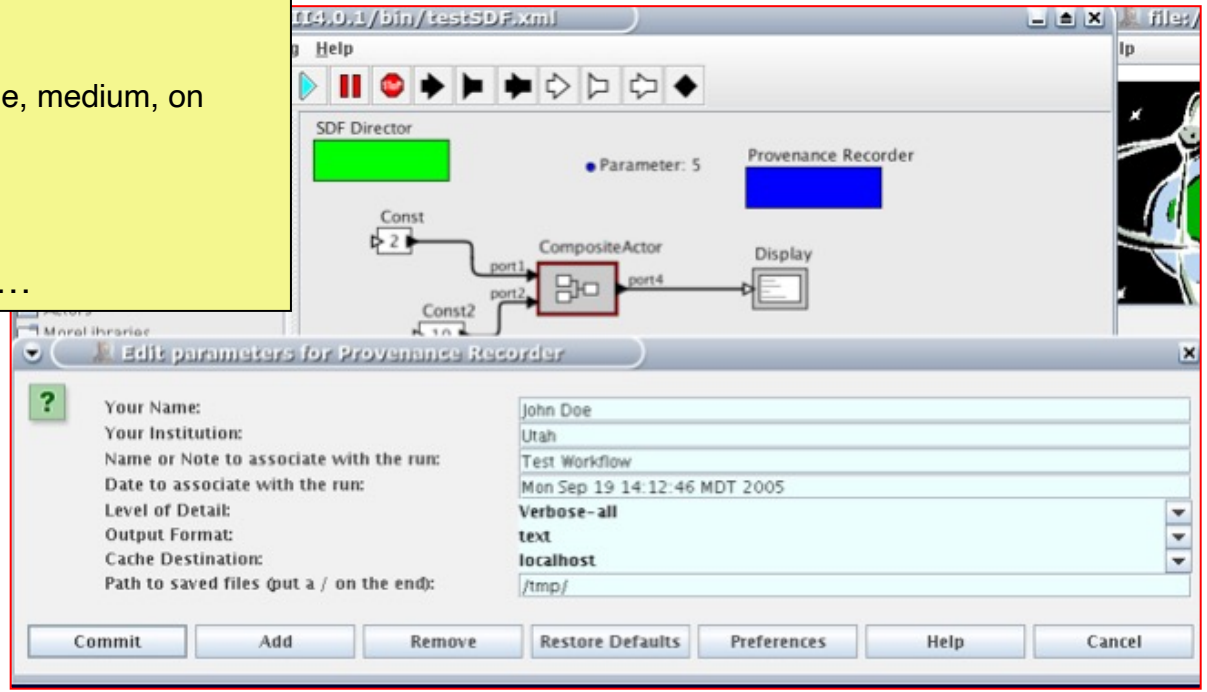
Provenance Framework

- Provenance
  - Tracks origin and derivation information about scientific workflows, their runs and derived information (datasets, metadata…)

- Types of Provenance Information:
  - Data provenance
    - Intermediate and end results including files and db references
  - Process (=workflow instance) provenance
    - Keep the workflow definition with data and parameters used in the run
  - Error and execution logs
  - Workflow-design provenance

Kepler Provenance Recording Utility

- Parametric and customizable
  - Different report formats
  - Variable levels of detail
    - Verbose-all, verbose-some, medium, on error
  - Multiple cache destinations
- Saves information on
  - User name, Date, Run, etc…



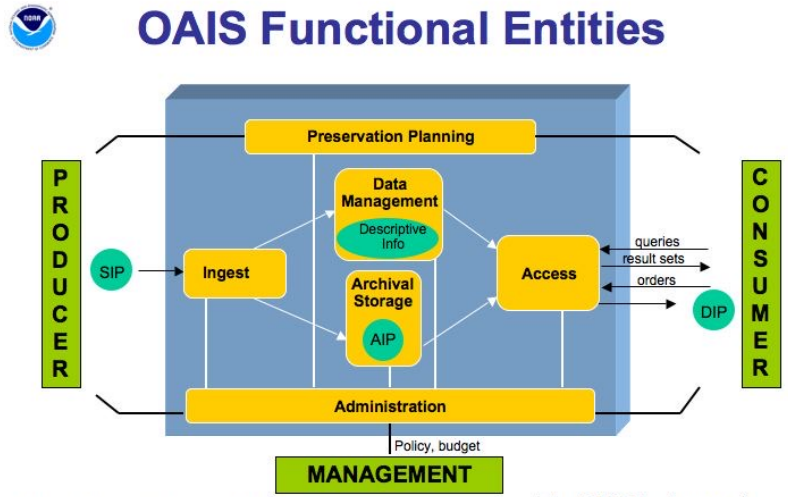
Some other workflow systems

- SCIRun
- Sciflo
- Triana
- Taverna
- Pegasus
- Some commercial tools:
  - Windows Workflow Foundation
  - Mac OS X Automator

http://www.isi.edu/~gil/AAAI08TutorialSlides/5-Survey.pdf

Now …. Data Stewardship

- Combining multiple data life cycle, management aspects together
- Keep the ideas in mind as you complete your assignments
- Why it is important
- Some examples



Why it is important

- Need ability to read the underlying sources, e.g. the data formats, metadata formats, knowledge formats, etc.

- Need ability to know the inter-relations, assumptions and missing information

What to collect?

- Documentation
  - Metadata
  - Provenance
- Ancillary Information
- Knowledge

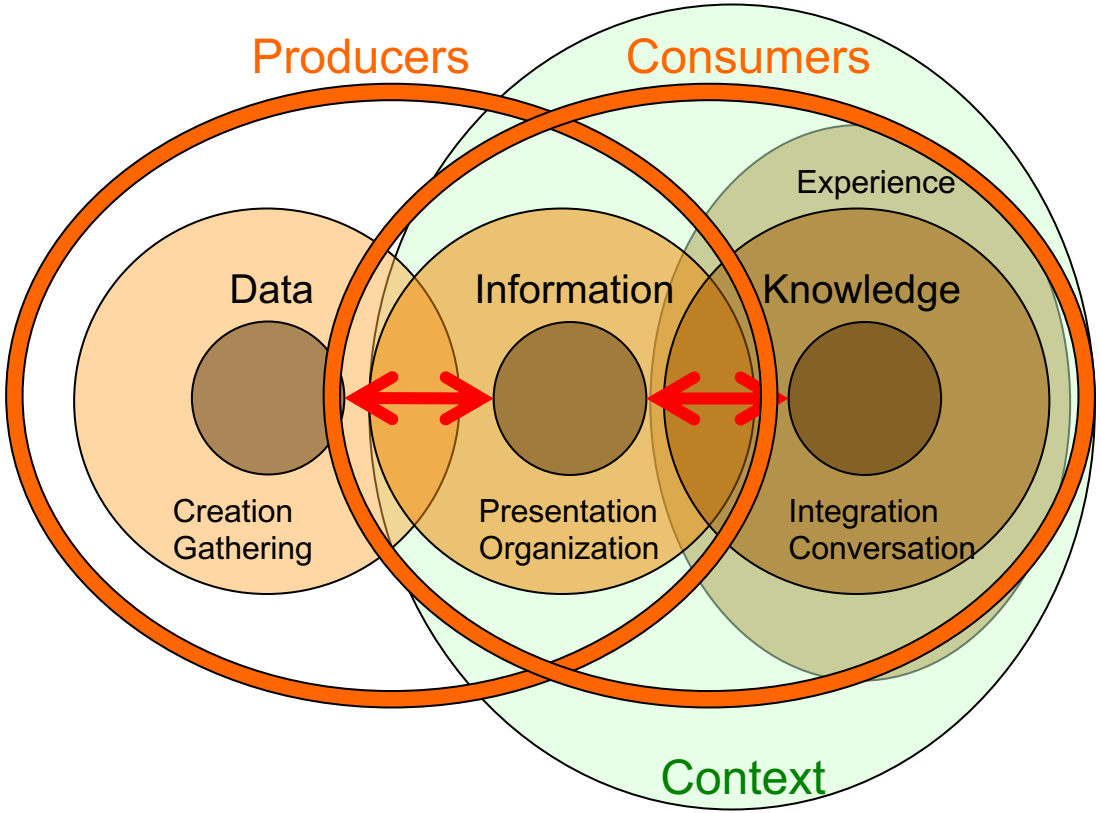
Who does this?

- Roles:
  - *Data* creator
  - *Data* analyst
  - *Data* manager
  - *Data* curator

How it is done

- Opening and examining Archive Information Packages! Yes, people look at them.
- Reviewing data management plans and documentation! Yes, people look at them.
- Talking (!) to the people:
    - *Data* creator
    - *Data* analyst
    - *Data* manager
    - *Data* curator
    - Sometimes, reading the data and the code

Data-Information-Knowledge Ecosystem



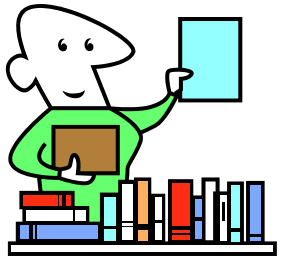
Remember - Acquisition

- Learn / read what you can about the developer of the means of acquisition
  - Documents may not be easy to find
  - Remember **bias**!!!
- Document things as you go
- Have a checklist (see the Data Management list) and review it often

Management

- Creation of logical collections
  - The primary goal of a Data Management system is to abstract the physical data into logical collections. The resulting view of the data is a uniform homogeneous library collection.

- Physical data handling
  - This layer maps between the physical to the logical data views. Here you find items like data replication, backup, caching, etc.



Management

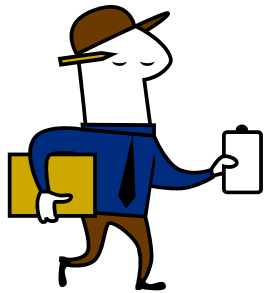
- Interoperability support
  - Normally the data does not reside in the same place, or various data collection (like catalogues) should be put together in the same logical collection.
- Security support
  - Data access authorization and change verification. This is the basis of trusting your data.
- Data ownership
  - Define who is responsible for data quality and meaning

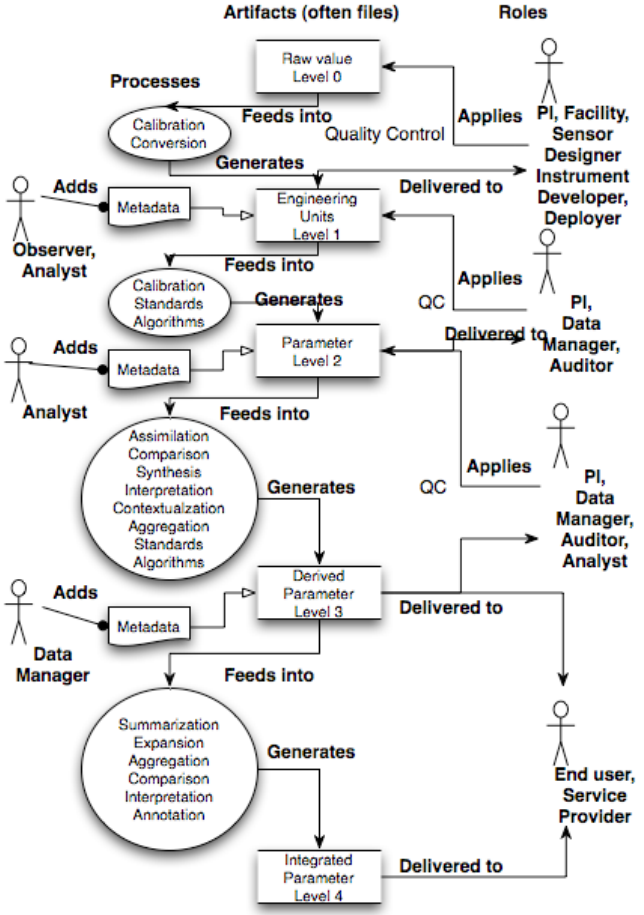
Management

- Metadata collection, management and access.
  - Metadata are data about data.
- Persistence
  - Definition of data lifetime. Deployment of mechanisms to counteract technology obsolescence.
- Knowledge and information discovery
  - Ability to identify useful relations and information inside the data collection.

Management

- Data dissemination and publication
  - Mechanism to make aware the interested parties of changes and additions to the collections.




Artifacts (often files)
Roles
Raw value Level 0
Processes
Calibration Conversion
Feeds into
Quality Control
Applies
PI, Facility, Sensor Designer Instrument Developer, Deployer
Generates
Delivered to
Adds
Metadata
Engineering Units Level 1
Observer, Analyst
Feeds into
Calibration Standards Algorithms
Generates
Applies
QC
PI, Data Manager, Auditor
Delivered to
Adds
Metadata
Parameter Level 2
Analyst
Feeds into
Assimilation Comparison Synthesis Interpretation Contextualization Aggregation Standards Algorithms
Generates
Applies
QC
PI, Data Manager, Auditor, Analyst
Adds
Metadata
Derived Parameter Level 3
Delivered to
Data Manager
Feeds into
Summarization Expansion Aggregation Comparison Interpretation Annotation
Generates
End user, Service Provider
Integrated Parameter Level 4
Delivered to

Curation

- Consider the “changes” in the organization and presentation of the data
- Document what has been (and has not been) done
- Consider and address the provenance of the data to date, you are now THE next person in the workflow…
- Be as technology-neutral as possible
- Look to add information and meta-information

Preservation

- Usually refers to the latter part of data life cycle
- Archiving is only one component
- Intent is that ‘you can open it any time in the future’ and that ‘it will be there’
- This involves steps that may not be conventionally thought of
- Think 10, 20, 50, 200 years (or 1 hour!) …. looking historically gives some guide to future considerations

Some examples and experience

- NASA, NOAA
- http://wiki.esipfed.org/index.php/Preservation_and_Stewardship
- Library community
- Note:
  - Mostly in relation to publications, books, etc. but some for data
  - Note that knowledge is in publications but the structural form is meant for humans not computers, despite advances in text analysis, NLP
  - Very little for the type of knowledge -> data we are considering: i.e. in machine accessible form

However…

- Even groups like NASA do not have a governance model for this work
- Governance: is the activity of governing. It relates to decisions that define *expectations, grant power, or verify performance. It consists either of a separate process or of a specific part of management or leadership processes. Sometimes people set up a government to administer these processes and systems. (wikipedia)*

Who cares…

- Stakeholders:
  - NASA for integrity of their data holdings (is it their responsibility?)
  - Public for value for and return on investment
  - Scientists for future use (intended and un-intended)
  - Historians

Back to why you need to…

- E-science uses data and it needs to be around when what you create goes into service and you go on to something else

- That's why someone on the team must address life-cycle (data, information and knowledge) and work with other team members to implement organizational, social and technical solutions to the requirements

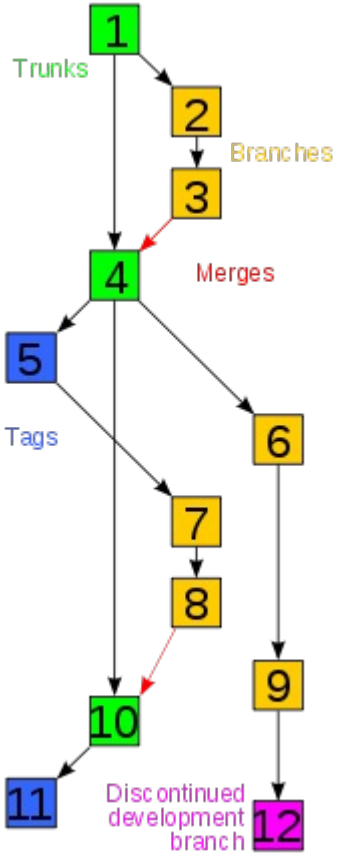
- And, you ask how do we know what is what?

(Digital) Object Identifiers

- Object is used here so as not to pre-empt an implementation, e.g. resource, sample, data, catalog
  - DOI = http://www.doi.org/, e.g. 10.1007/s12145-008-0001-8 – visit crossref.org and see where this leads you.
  - URI, http://en.wikipedia.org/wiki/Uniform_Resource_Identifier e.g. http://www.springerlink.com/content/0322621781338n85/fulltext.pdf
  - XRI (from OAIS), http://www.oasis-open.org/committees/xri

And not least … Versioning

- Is a key enabler of good preservation
- Is a tricky trap for those who do not conform to written guidelines for versioning
- http://en.wikipedia.org/wiki/Revision_control



Summary

- The progression toward more formal encoding of science workflow, and in our context data-science workflow (dataflow) is substantially improving data management

- Awareness of preservation and stewardship for valuable data and information resources is receiving renewed attention in the digital age

- Workflows are a potential solution to the data stewardship challenge

What is next

- Next week, Nov 28th – Thanksgiving Thursday (no lecture!!!).
- Thursday Dec 5th – Presentations of Team project (Assignment 4) during class
- Tuesday Dec 10th – Written reports of Team project (Assignment 4) due AND individual Assignment 5 submissions due

Thanks!!!

We made it!!! (almost lol)