WORKFLOWS AND DATA MANAGEMENT

(BITS, BYTES AND WHAT WE DO WITH THEM)

ADAM BRAZIER, SEPTEMBER 29TH 2014
Overview

Summary and scope

• Workflows
  • Automation, our friend and foe
  • How should we automate a workflow?
• Data management
  • From cradle to grave: the lifecycle of data
  • How should we make a plan?
• Scope
  • The (our) university research environment
  • Process, not specific software recommendations
“Workflow” may mean different things to different people.
“Workflow” may mean different things to different people. Avoiding dogma, we can consider “workflow” as:

A) What it says on the tin
“Workflow” may mean different things to different people. Avoiding dogma, we can consider “workflow” as:

- A) What it says on the tin
- B) A process which can be illustrated with a flow diagram
“Workflow” may mean different things to different people. Avoiding dogma, we can consider “workflow” as:

- A) What it says on the tin
- B) A process which can be illustrated with a flow diagram
- C) “A series of tasks which produce an outcome” (Microsoft)
“Workflow” may mean different things to different people. Avoiding dogma, we can consider “workflow” as:

- A) What it says on the tin
- B) A process which can be illustrated with a flow diagram
- C) “A series of tasks that produce an outcome” (Microsoft)
- D) “A workflow consists of an orchestrated and repeatable pattern of business activity enabled by the systematic organization of resources into processes that transform materials, provide services, or process information” (Wikipedia)
“Workflow” may mean different things to different people. Avoiding dogma, we can consider “workflow” as:

- A) What it says on the tin
- B) A process which can be illustrated with a flow diagram
- C) “A series of tasks that produce an outcome” (Microsoft)
- D) “A workflow consists of an orchestrated and repeatable pattern of business activity enabled by the systematic organization of resources into processes that transform materials, provide services, or process information” (Wikipedia)
What do our workflows look like?
What do our workflows look like?
Workflows

Or some part thereof...
Follow the data!

Workflows

Processing center → Webserver → Tracking Database

Data → Request → File data

Webserver → DDN Storage Unit

Data → Request to Data Products Database

Processing products

CyberSKA, AI rankings

Requested products → Candidate rankings
Model the processes!
Why automate?

- Cheaper, in the long run
- Speed
- Reliability, Robustness
- Repeatability!

Faster, better, stronger!
What *can* you lose if you automate?

- Hands-on involvement, the sense of what’s going on
- Grad student training ground
- Development time
- Development cost
- Don’t build HAL (or SKYNET!)
What do we need?

- Clear requirements.
- High-level, modular/loosely-coupled design
- Budget
Workflows

Who should do it?

• This is a decision which depends on scale
• Domain researchers:
  • Intimate understanding of the activities
  • Embedded into the workflow already
  • Typically involved in writing the proposal
• Software professionals
  • Generally more current with available technologies
  • More practiced
  • Outsider’s view
• Why not have both?
Managing the interface

- Software professionals and domain researchers both important

- Specification of the project’s scope and requirements necessary

- Communication between the individuals and teams is what makes or breaks design and development

- Quality of personnel obviously a big driver of output
Workflows

Case Study – PALFA

Diagram showing the connection between Arecibo, FedEx, Astronomy, Network, McGill, PALFA candidate ranking, Laptops, Desks, Internet, AEI Hanover, CAC.
Workflows

Case study – PALFA

• Workflow very heterogeneous

• Large set of actors: undergrads, grad students, facility staff, postdocs, faculty, sysadmins, software developers

• Very large data set (for the time!)

• End-to-end duration ~ 1 month, plus reprocessings
Workflows

PALFA – 2 key successful elements

• Management of the interface between researchers and IT professionals
  • Requirements, regular communications, short development cycles
  • Resulted in product which matched needs, with cost control

• Loosely-coupled workflow elements with defined interfaces
  • Independent development by people with the expertise
  • Resulted in robust and adaptable design
Workflow

PALFA – 2 key areas for development

• Monitoring of workflow
  • Strengthens automation, improves debugging
  • Make report-production much easier

• Documentation
  • Easier to bring new researchers on board and survive people leaving
  • Makes modification and enhancement of the workflow much easier
Some rules of thumb

- Put aside time for planning.
  - Separate requirements from design. Do requirements first! Evaluate what is needed

- Assign responsibilities to individuals and teams

- Ensure communications

- Documentation and monitoring/QA should be defined deliverables
What is data management?

- One view (congruent with NSF guidance)
  - Description
  - Control
  - Policies
  - Storage/preservation

- Another way of looking at it:
  - Data management is the workflow, cradle to grave.
  - Your workflow will/should/can achieve NSF/other data management requirements
One view (congruent with NSF guidance)
  • Description
  • Control
  • Policies
  • Storage/preservation

CODE IS DATA, TOO!
Data Management Plans (DMPs) now required by many RFPs (including all NSF RFPs)

Taking planning seriously makes sense:

- It allows costing it into a budget

IT OFTEN IS THE WORKFLOW, END-TO-END

A proposal DMP is a higher-level description, but further planning should take place before implementation begins
Research Data Management Service Group (RDMSG, http://data.research.cornell.edu/) provides DMP consulting and other services to Cornell researchers

For those planning to use CAC services, we will provide help writing Data Management Plans and cyberinfrastructure sections of Proposals

Many people are addressing similar questions, both inside and outside Cornell.
• Enumerate your data products!
  • Include code, documentations, visualizations, online content
  • Metadata is also data!

• Decide on formats, including considerations of:
  • Format longevity
  • Access to the content elements
  • Ease of use, including by others
Control includes things we *do* to our data.

- I/O
- Transport
- Pipelining/processing
- Versioning
- Tracking
- Quality Assurance
- Sharing and security

Many functional requirements arise here.
Policies constrain and guide control, generating non-functional requirements/design constraints

Key policy issues include:
- Who can have our data?
- When can they have our data?
- Under what conditions can they have our data?
- Licensing and attribution requirements
- For how long must we keep our data?
Data Storage/preservation

- **Storage:**
  - Persisting the data during the project’s duration

- **Preservation:**
  - Persisting the data after the project is completed

- There can be some hard decisions!
  - Cost broadly scales with volume
    - On-campus: CAC’s Archival Storage facility, eCommons, CIT’s EZ-Backup and department facilities – each serves different needs
  - For code, documents, audio-visual material, lower-volume data and data products, free solutions exist
Data

What to keep, long-term?

- Material which supports publications should have the highest importance

- Take advantage of free resources:
  - eCommons (a Cornell service)
  - Github, sourceforge, etc
  - Youtube
  - Journal supplementary data resources
  - Department resources
  - Keep your eyes open!
Conclusion

And, in summary

• Workflows and Data Management are inextricably linked

• Planning is key!

• It takes a team to build a solution; provision the expertise before you start