DataONE Use Case Notes (Brainstorming)

2014-10-01

- Did some more brainstorming on the “lift script to workflow via structured comments” tool. More details on this to be added here …

2014-09-25

Present: Yaxing, Bertram, Josh, Mark S. (having trouble with connectivity-sorry)**,** Steve, Chris Jones, Christopher, Bob, Paolo

Regrets:

Agenda:

- End to end use case
- Review next week’s timetable/schedule (AHM Draft Agenda)

Notes:

- Review next week’s timetable/schedule:
  - Working Group discussions starting on Tuesday morning at 10:30 am ; Christopher and Debbie will be there around **noon**
  - Departing Wednesday after the meeting
- Should have a remote participation at DataONE through GotoMeeting or similar.
  - ⇒ local setup: Matt and Dave have looked into this. Mark Schildhauer et al will bring remote equipment, incl. projector; Dave to set up google hangout
- End to end use case:
  - ChrisJ:
    - #41 (Provenance Capture) need to be able to produce traces from scripts (Matlab, R, Python); client-side coding work; so that a scientist can “start recording”; then run; then the lib takes care of that; coarse-grained file artifacts, not fine-grained runtime level artifacts
    - #46 (Capture + DataONE “binding”) : similar, but without interaction with DataONE (#41 does the DataONE part, too)
    - #42, 43, 44: (Search, Discover, Explore Provenance) Search provenance info that was previously captured; to understand input-output relationships, lineage
- Bertram: Conceptual-level reverse-engineered workflow as a “jump off board” for various provenance artifacts
- Christopher:
  - “smart rerun” capability (cf. Make)
  - units used
  - phases and steps that lead up to standard output workflow (note: we need to list the inputs and outputs for each of the MsTMIP “phases”, each could be a stub to be added to the package, artifacts that come out of them (with identifiers) could be added to the package)
    - ingest phase
      - step 1: read raw model output (diversity of formats)
      - step 2
    - standardization phase
      - step 1: grid orientation/units/limits/semantics
      - step 2: data formatting (e.g. CF convention)
    - output phase
      - step 1: intermediate products
      - step 2: detail 2

- Paolo and others:
    - Within the script, push data into DataONE using the API -> this idea however is probably superseded by the idea of annotating the script itself with directives (structured comments). Such a generic mechanism with specialized domain vocabulary would have multiple usages, including indicating which data needs to have provenance associated with it. That is, you could say “when you upload this data to DataONE, its provenance should follow (eg in a data package).
- Chris:
    - What are the artifacts that are saved into the packages; need to have a mechanism to flag / export certain items (but not others)
- Steve: modeling language for MsTMIP -- see “phases or steps” item above
- Bertram: seconding Steve’s request; idea: instrument scripts via “structured comments” that then can be used to “harvest” some basic semantics and workflow structure
- Christopher: common vocabulary across MsTMIP
- Steve: collect the info from the scientists
- Christopher: start from ingest; standardize; output
-

2014-09-18
Present: Yaxing, Bertram, Paolo, Josh, Mark S.**,** Steve, Xixi Luo
Regrets: Bob, Christopher

Notes:

- Here is a link to the “MsTMIP Big Picture” figure with annotations.
- DataONE Provenance Use Case Document
- Josh:
    - need to support analysis
    - with multiple versions, to what extent are analyses comparable
    - if there’s always a new version coming out, how to deal with that?
    - distinguish what changes might need to be considered vs ignored
    - What are success stories and ideas about provenance?
        - showing Tianhong’s summer project presentation..
- Mark
    - develop overall workflow (as “interactive graphical flow chart”) for models and outputs; then enable drilling into specific runs-- to get outputs, data, code assoc with a specific run
    - role of workflow diagrams
        - use as an “interface” for the analysis study
        - simple “markdown-like” tagging (e.g. with PROV concepts) could be done in MATLAB, then parsed for a visual presentation (using some network-like diagramming tool such as CMAP, VUE -- both of which can depict RDF)
- Josh:
    - analysis done in Matlab and R; less so: Python
    - wf visualization is critical to understand the method, analysis
    - order of operations (two model outputs with different spatial resolutions, when compare them, which resolution to choose, the coarser one or the finer one? compute mean vs another function)
    - years later: replicate an “old” analysis

DataONE Use Case Notes (Brainstorming)

- Bertram:
  - Visual rendering of scripts
- Mark:
  - Capability to annotate notes of a workflow (revisiting the old flowchart approach to coding)
- Josh:
  - cutting edge in peer review: comment on papers
  - imagine people could annotate workflows, analysis scripts
- Steve:
  - timing may be right -- web annotation tools are coming online (example annotator js)
  - web-aware flowcharting Mark described resonated well;
    - lots of people describing something similar to a flowchart, not necessarily a more formalized workflow;
    - initially
      - would it be sufficient for DataONE and MsTMIP to simply document the inputs and outputs for each flow chart step?
      - develop mechanism to capture the information first then work on the implementation (of the graphic representation)?
  - GCIS documents data items and activities (GCIS activity example which links to an image (see the svg) and dataset (nb: 24 other images (b42fbac8, 6938ed9f, fb74813a, daed8535, 7cc0679a, 9e2a261e, 4f071049, 80f8effc, e3fb627a, 0c3eb1d1, 9d9aa7ac, f3f25c78, 2a8ed68d, c8484765, b8b652de, 0158fa86, 95fe2b26, c75d4166, 13129a6d, 51fd0ea1, b10ad0f2, 1976ca9b, 7614711e, 9e67e6e9) were also derived from this dataset))
  - how do we get the content updated and into DataONE
  - different levels of provenance: external, internal and detailed history (see the 2014-09-11 meeting comment “General, simplistic questions” below))
- Mark:
  - tantalizingly close
  - codeblocks
  - use markdown
  - use standard tags (PROV) could then be rendered as a workflow
  - if the analysis are like Make files
- Josh:
  - hit a button “capture here”
  - needs to be really easy for users

Misc:

- Location of the MsMTIP Version 1 data: http://nacp.ornl.gov/mstmipdata/
- Location of the MsTMIP pre-Version 1 data: http://nacp.ornl.gov/mstmip_model_output_inventory.html
- Tianhong’s summer project presentation /

2014-09-11

Present: Bob, Christopher, Josh, Yaxing, Steve, Bertram, Paolo

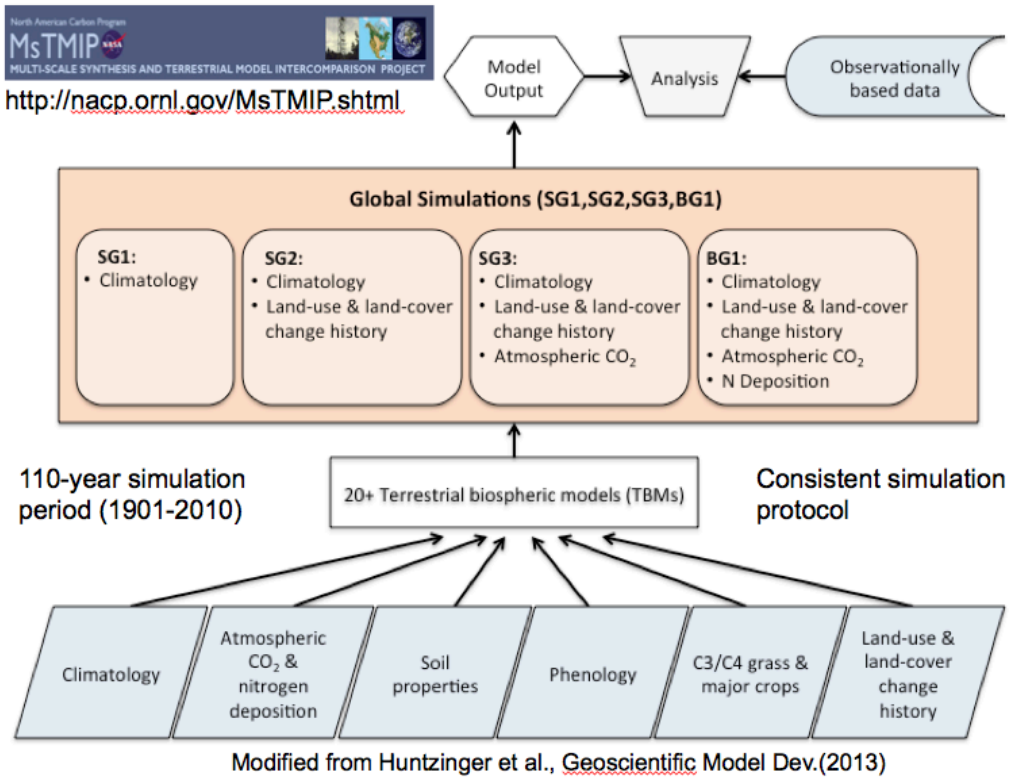
Notes:

- Bob: need some real-life examples from Christopher, Josh that fit in this framework; capture these examples, use cases; then see how DataONE can facilitate this. Should directly benefit their work and research.
- Christopher:
    - current analysis, looking at an ensemble of multiple outputs
    - 10 member ensemble; has changed now for 5th time: models revise simulation output; need to go back to his script/workflow, and update the analysis
    - now these 10 are “done” for now for MsTMIP
    - prov can be highly useful here
    - “can’t go back in time” (only most recent versions available)
    - data size and lack of version control are issues
    - archive only keeps most recent version(s)
    - didn’t anticipate to use the wf 5x; organic growth; use of “naive” provenance (naming conventions); looking for a more principled approach
    - notion of “frozen version number” in MsTMIP
    - some works were done prior to the release ⇒ working on different, pre-release versions
    - primary producers of v1 are the modeling teams
    - Christopher et al both creating and using v1
- Josh?
    - re. versioning issues: at some point (by fiat?) something is called v1
    - but over the years, there will still be desire to update parts of what used to be frozen v1
    - ⇒ does “full” versioning support address/solve this?
    - Who is doing the new versions? The modelers or the “integrators” (Christopher). E.g. there could have been a change in the “metadata” (definitions / units etc).
    - Not always clear who has done updating
- Christopher: MsTMIP needs a DOI
- Yaxing:
    - release of v1 as a bundle with comprehensive docu
    - several months later: v1.1 or v2, again bundled, docu, and changes
    - so we have different versioning issues here:
        - the official release bundles evolve
        - the inputs to the bundles evolve
- Josh
    - to support analysis across papers: difficult; cannot do Canada vs Alaska if they use different versions
- Bob
    - pre-release versions are not tracked
    - should be possible to keep previous bundles around
- Josh
    - goal is to create legacy for future projects
    - future vision goes beyond current duration
    - challenge how to sustain the provenance beyond current funding period
- Yaxing:
    - e.g. Visit model evolved, changes shown in the figure
    - updates often accompanied by little docu (what has changed)
    - we should not expose the detailed history to the end user
- Bob:

    - Harmonization steps are important and need to be captured and shared
    - part of the history should stay private
- Need for both internal and external provenance
    - e.g. Visit had various earlier versions, but those had bugs; needed to be discarded
- Paolo:
    - use of a version control system?
        - ? not beyond the “v1”, “v2”, etc. of released bundles
        - would git functionality be useful?
    - using DataONE packaging mechanisms?
- Yaxing: What would be overhead for using DataONE or git capabilities?
- Paolo: DataONE uses Resource Maps, describing what’s in the bundle; provenance can be included
- At the time of release, populate the manifest of the bundle with the versioning/provenance info should be very valuable
    - Yaxing: ORNL DAC is setting up a member node to publish MsTMIP data; not much prov included now. and metadata
    - but too costly to keep the file-level provenance for all files!?
- but we can manually produce provenance for selected files in MsTMIP output bundle version 1. Again, it’s hard to create a “reproducible” workflow/provenance. For a high-level workflow that describes steps involved in the data processing, it’s possible.
- Christopher:
    - challenge is generic use case of provenance, other than “db mgmt” issues
    - how can prov help me?
        - what I do w.r.t. paper, seems so specific, might not be generalizable
        - might be more at the “db mgmt” level
        - “long tail” of what all the different parties do with the data
- Josh:
    - rectifying problems in previous ?analysis?papers?
    - custom “version hell” at the time of preparing the manuscript; was difficult to sort out later and show the editor what was done and how to reproduce this
    - thinking about future papers: 5 years from now, MsTMIP output will be used, how do we deal with the “post-release” updates? E.g. someone wants to update a particular version, how to do that N years from now
- Paolo:
    - in software versioning: using e.g. cloning and branching; if things work, they can be merged back to the main branch
- Current V1 has ½ TB
- Future bundles will be multi TBs

2014-09-04 Brainstorming Call (Steve, Paolo, Yaxing, Bertram)
Possible DataONE use case:

- Starting point: MSTMiP figure:



Modified from Huntzinger et al., Geoscientific Model Dev.(2013)

- ⇒ here is a link to this figure with annotations.
- Summer internship:
  - Take a simple Matlab script from Christopher
  - Migrate to Python
  - Run noWorkflow “provenance” (profiling) tool
  - Execute some provenance queries over the noWorkflow collected provenance
  - learn about iPython provenance
    - see how changes w.r.t. provenance
- Two kinds of uses of provenance:
  - “externally facing” documents a published product
    - publish provenance (e.g. along with the paper) to make transparent what you did
    - respond to queries from colleagues and interested parties (I refer to my “internally facing” provenance information)?
  - “internally facing” documents the workflow steps used to produce the published product
    - what the heck did I do here?
      - use “well-designed” folder structure, audit logs, …
      - facilitate collaboration among colleagues in the same team
  - relation to GoldenTrail
    - main idea is that lots of prov gets created in the “unstable state”, but once the data object is ready for publishing, the prov trail you want to associate with it (the “externally facing”) is a subgraph on the internally facing prov.

PM: can we state this as a problem of [automatically] generating an (externally visible) provenance view from a low-level, larger (internal and private) prov document?

Attempt at a story: Christopher et al publish MsTMIP paper along provenance. What can we do with this?

DataONE Use Case Notes (Brainstorming)

- Different levels of “validating” (being satisfied with) the paper / results:
    - Ralf reads the paper, looks at the method section and the published provenance and determines: this is legit! (the paper has already been peer reviewed and published at this point)
    - Anne reads the Christopher et al paper and the external provenance. Anne is interested in learning more so she can apply the results to her ongoing research. She contacts Christopher and has specific questions about his “internal” provenance.
    - Yaxing, as a reader of Christopher’s paper, wants to reproduce the figures in the manuscript without anyone’s help (externally facing?)
    - Yaxing, as a user of DataONE, found a data product from DataONE, read its metadata, wants to know what data were used to generate this data product. He also sees a high-level workflow described in the metadata. He then contacted creator of this data product to find more details about the algorithm (internally facing?)
-

Bertram: Andrews and Andrews from recomputation.org

DataONE Provenance Use Case Document

- Use Case 41 (capture & share provenance)
    - (provenance capture) In DataONE-enabled client software, investigators can easily provide tracking information as they create new products from existing data files.
    - (provenance upload) Investigators can upload derived datasets to a Member Node and provide traceable links to the primary resources used to create them.
- Use Case 42 (examine provenance)
    - Scientists examining a synthetic dataset in DataONE are able to determine which dataset from a DataONE Member Node was used in the synthesis and can examine that dataset.
    - A scientist that has searched for data relevant to their studies has found a synthetic dataset in DataONE. They are able to view the relationships between the original and derived datasets, and can download the originals for examination.
- Use Case 43 (attribution: who uses my stuff?)
    - To provide a traceable link to derived works for each dataset.
    - A scientist that has uploaded their dataset to DataONE and has allowed derived works in their intellectual rights statement can view and understand which derived works have used their dataset.
- Use Case 44 (reproducibility)
    - To assist reproducible science by providing a link between script or models, the input data used, and the generated output.
    - A scientist reviewing a data table or figure in DataONE can discover the script or model that was used to generate it. The scientist can subsequently download and rerun the script to reproduce the same results as the original run.