Section 4. Requirements

This section includes a description of process(es) to be used to get the necessary technical, conceptual, and/or community (i.e., end-user) requirements at the outset and during the life of the activity, including approaches to achieving community/end-user consensus.

4.1 Challenges and Preliminary Findings

The Workflow Roadmap will need an ongoing process for obtaining and understanding requirements of the geoscientific community.   The diversity of users is an important challenge that must be addressed when obtaining these requirements.

User Diversity:

- Field researchers: these may be more interested in data collection and analysis. Their workflows may be only partially digital (ie need to make sure the field sample was properly annotated and stored).
- Computational researchers: these are researchers who are constantly building and modifying their models. Each one can be unique.
- Data-Intensive Computational Researchers: these are researchers who assimilate historical or real-time streaming data in models, or people who use other models as input (climate models as input to hydroecology models).
- Operational users: these are people who are responsible for creation of public data sets, forecasts, etc. These are probably good candidates for many current workflow tools.
- Classroom usage: Many workflow tools would expose students to analytic tools in science through workflows.
- Graduate and undergraduate research support: Geoscience student researchers spend an extraordinary amount of time attempting to find and access data necessary for their research. This can be everything from geospatial data, to time series, to publications (digital and paper), maps (digital and paper), relational databases, etc.

These may each require different types of workflow capabilities.

Preliminary Findings: As part of its March-June 2012 workshop series, the Workflows Community Group created a questionnaire (https://sites.google.com/site/earthcubeworkflow/questionnaire-for-the-community) as a way to capture community input. The survey format allowed essay responses to questions. From the community survey responses so far obtained, efficient sharing of multi-step data transformations, handling big data, projecting diverse geospatial/temporal data sets, integrating multiple data sets, managing complex executions, reproducibility of results, and interoperability with other tools and services (OPENDaP, NetCDF, OGC services, ArcGIS, etc.) are all capabilities mentioned by the responders.

Surveys will be an important mechanism for future broad requirements gathering. The initial survey was useful in making new contacts and in formulating initial conceptual requirements, but the format was not appropriate for gathering results that can be statistically analyzed. The initial questionnaire can be improved in successive versions (with help from experienced survey designers) as representative prototypes of geoscience workflows are designed, implemented, and evaluated.

The Workflows Roadmap will include a Status and Requirements Task Force that will be charged with formulating and executing plans for obtaining user, conceptual, and technical requirements.

The Workflow Roadmap will need to include methodologies for more systematic requirements gathering. The remaining sections outline these processes.

The Status and Requirements Task Force will follow well defined processes and assessment metrics to gather user requirements and to understand the impact of workflow technologies in geosciences research.

4.2 Processes for Obtaining User Requirements

We outline here a general approach to user requirements gathering. As discussed above, user diversity is a challenge for requirements gathering, so the process must take this into account. Since the user space is large, a guided study that samples important requirements space will be more useful than a brute force search.

Startup Processes

1. Broadly associate the range and types of geoscience data and model needed by scientists, grad students, K-12 teachers, environmental resource managers and the public. For example, geoscience data can mean something like "Essential Environmental Variables" (EEVS). The WMO uses the terminology Essential Climate Variables but that is too narrow. Most of the EEVS are already accessible but they reside on many different federal servers often with slow access and in formats not understood by geoscience users. (e.g. topography, historical climate or reanalysis, landuse and land cover, geology, soils, etc.).
2. Develop a matrix for aligning workflow technologies with the use-cases outlined earlier and the particular data and model needs in each use case. Partition the matrix into near-term (1-4 years) and long term 5-10 years) workflows.
3. Carry out a hypothetical prototype analysis on an existing workflows that can answer the

following:
    - How can workflows automate the management and sharing of data and models, experimental design, as well as efficient sharing of software?
    - Can workflows improve the geoscientists ability to track, visualize and analyze data, experiments and models results through a unified framework?
    - What is the role workflows in tracking provenance of data and models in the context of scientific reproducibility and for advancing scientific understanding?
    - How does a workflow facilitate data and model quality assurance?
4. Conduct a community survey to evaluate the matrix and the prototype.

Follow-up Processes

1. Design a technical strategy for aligning workflow technology for each geoscience use cases outlined earlier.
2. Design 2 to 3 pilot studies from the matrix of examples of workflow technologies over the next 1-4 years (near-term roadmap) and 5-10 years (long term roadmap)
3. Present results of preliminary design at the AGU/IEEE national meetings to solicit feedback from the larger geoscience computer science communities.

Consensus Processes

1. Develop an evaluation strategy for testing the effectiveness of the 3 workflow prototypes
2. Demonstrate how the workflow improved the efficiency of the geoscience team for each workflow prototype

4.3 Processes for Obtaining Technical Requirements

Technical requirements will be obtained primarily from the cyberinfrastructure community and are derived from geoscientist community requirements (Section 4.2).

Startup Processes

1. Data: For example, develop a workflow structure for making essential terrestrial geo-data from many federal agencies accessible within the workflow (e.g. topography, historical climate or reanalysis, landuse and land cover, geology, soils, etc. see UN site http://www.fao.org/gtos/doc/pub52.pdf). This will also include agents/tools for automating data collection, transfers, data classification, data derived products and data management.
2. Models: Implement workflow technologies for geoscience models for the proposed workflow prototypes discussed above. This should include simple conceptual mathematical models developed in MATLAB or Mathematica and HPC-level models that

run on large clusters or the grid. Leverage exisiting community resources such as CSDMS (e.g. Community Surface Dynamics Modeling Systems), the NCAR/NOAA Community Model Weather Research & Forecasting Model and others.

3. Fault Tolerance, Quality Assurance & Provenance: The workflow environment will require the automated capability of identifying failures in system components, generally evaluating errors in data and models, versioning of models/data and automating fail-over strategies.
4. Research Planning and Scheduling: Implement technical requirements for user-directed scheduling of workflow technologies for data and models that align with the use cases outlined earlier and data and models needs in each case. This will depend on the geoscience problem or hypotheses but we might classify them as: retrospective simulation based on the geologic past or recent climatic changes, real-time simulation for data assimilation form streaming sensors, and finally prediction or projection-type simulations (e.g NCAR/NCEP WRF or IPCC projections).
5. Research Discovery: How do the proposed workflow tools support the intersection of data, models, analysis and visualization across geoscience disciplines?

Follow-up Processes

1. Develop 3 representative prototype examples from the matrix of examples that will most likely be enriched by workflow technologies over the next 1-4 years (2 prototypes) and 5-10 years (1 prototype).
2. Implement the representative workflow prototypes with strong community interaction and participation through funding from EarthCube.
3. Evaluate the protoype workflows
4. Carry out a community survey to evaluate the matrix and the pilot design prototype.

Consensus Processes

1. Present results of preliminary design and community survey at the AGU/IEEE (?) national meeting(s) to solicit feedback from the larger geoscience and information science communities.

4.4 Processes for Obtaining Conceptual Requirements

The user requirements gathering process will most likely result in gap analysis. Conceptual requirements gathering will face the more challenging problem of identifying conceptual omissions and shortcomings in the current workflow research landscape that have implications on geoscience workflows. These requirements would typically result in recommendations to the NSF on research opportunities suitable for solicitations.

Startup Processes

1. Refine the matrix for aligning workflow technologies with the use cases outlined earlier and with the particular data and model needs in each use case. Partition the matrix into near-term (1-3 years) and long term 4-10 years) to resolve the needs of the prototype workflows that can be achieved relatively quickly (1-3 years), mid-term (4-6) and longer term (7-10).
2. Evaluate the matrix for aligning workflow technologies with the use cases outlined earlier and the particular data and model needs in each use case. Evaluate and refine initial pilot studies from the matrix of examples that will likely be enriched by workflow technologies over the next 1-3 years and 4-10 years.
3. Refine initial pilot studies from the matrix of examples that will likely be enriched by workflow technologies over the next 1-3 years and 4-10 years.

Follow-up Processes

1. The workflow research team should carry out a community survey to evaluate the matrix and the pilot design prototype from their point of view.
2. Present results of preliminary design at the AGU/IEEE national meeting(s) to solicit feedback from the larger geoscience computer science communities.

Consensus Processes

1. Appoint and fund an independent community-based evaluation team for assessing and testing the effectiveness of the 3 workflow prototypes.
2. Quantitatively demonstrate how the automated workflow practices have improved the efficiency of the geoscience team (or not) for each workflow prototype
3. Create a national data-software-workflow synthesis center for the geosciences that acts as a clearing house for best practices and further provides post-doctoral and staff support for implementing best practices, hosts visiting researchers and other activities for dissemination of best practices. The synthesis center should support all NSF GEO science communities and not just communities that already have relatively mature data-software-workflow plans in place.

The Workflows Roadmap will include the establishment of a Workflows Synthesis Center for the geosciences that is a national center of excellence and acts as a clearing house for best practices. The Center will support the activities of the Workflows Working Group, by providing post-doctoral and staff support for implementing best practices, hosting visiting researchers, and pursuing community activities for dissemination of best practices. The synthesis center should support all NSF GEO science communities and not just communities that already have relatively mature data-software-workflow plans in place.