Executive Summary: EarthCube Workflows Roadmap

September 15, 2012

Introduction

Advances in geoscience research and discovery are fundamentally tied to data and computation, although formal strategies for managing the diversity of models and data resources in the earth sciences have not yet been resolved or fully appreciated. Through this roadmap document we hope to motivate the importance of scientific workflows in support of geoscience research, and discuss a comprehensive path towards achieving the goals and practical benefits of leveraging scientific workflows in geoscience research and discovery.

Scientific activities can be seen as collections of interdependent steps represented as *workflows*. Gathering and analyzing data, coordinating computational experiments, and publishing results and data products are organized activities traditionally captured in research notebooks. Today we have the ability to digitally codify much of these activities, particularly for computational experiments, using workflow technologies. Workflows may be used to execute enormous computations, to combine distributed data and computing resources in novel ways, and to guide scientists through complex processes. When combined with metadata and provenance-capturing capabilities, workflows allow reproducibility of results, increased efficiency, and enhanced publications. The challenge before us is to make these tools ubiquitously available, enhanced, and adopted for the geosciences.

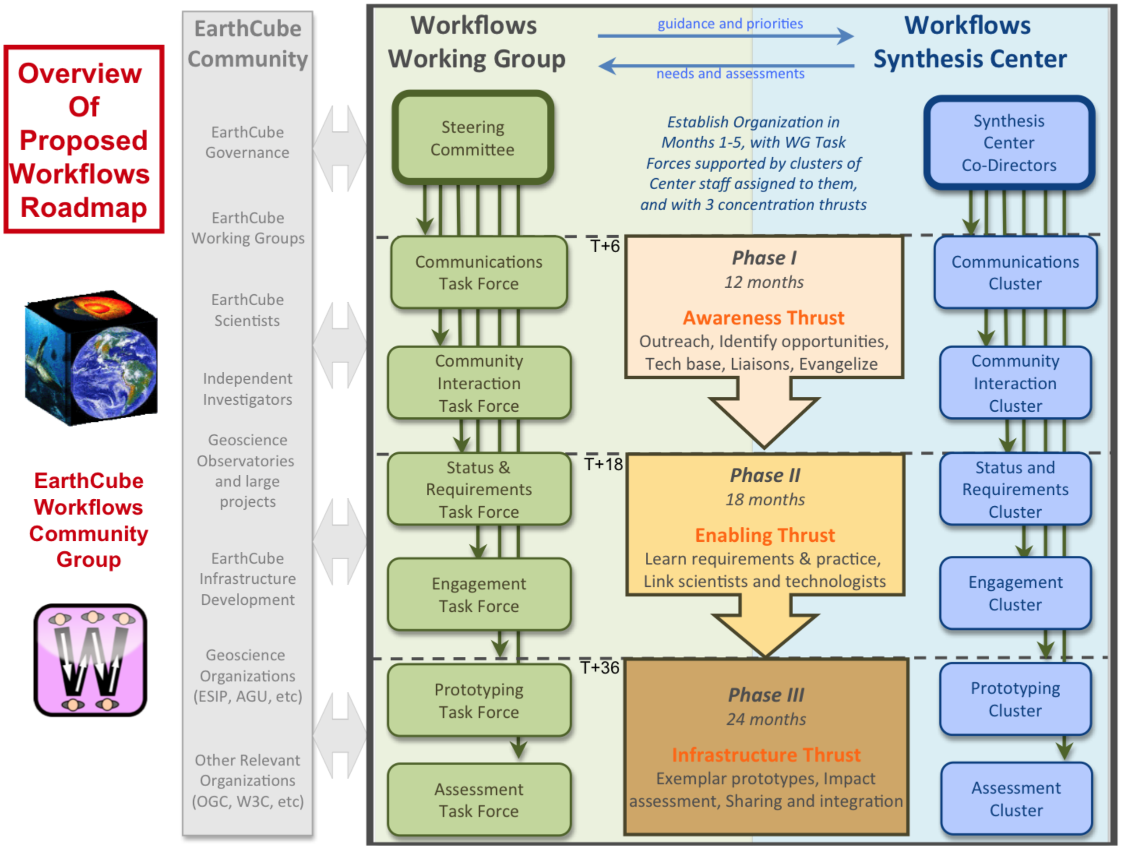
The EarthCube Workflows Community Group was created as part of the NSF EarthCube initiative. Its goal is to constitute a broad community within the geosciences that will identify both short-term problems and long-term challenges for scientific workflows. Addressing this goal is the central theme of this roadmap. Aspects of this goal include better education and outreach, better understanding of the different types of workflows, better collaboration between workflow software developers and geoscientists, the identification of gaps, and the vision for grand challenges that no workflow technology can currently address. The resulting workflow roadmap is considered a living document that will be extended and updated as future needs and our understanding of the problems evolve.

A cornerstone of the roadmap is to bring to life a **Workflows Synthesis Center** that will provide resources for an EarthCube **Workflows Working Group**, providing together an umbrella for all workflow-related Earthcube activities and for coordination with other activities that focus on other aspects of EarthCube. Specific task forces are identified in this roadmap.

The Workflow Working Group's Steering Committee will be the central organizer and broker with the EarthCube community. In its initial phase, the Steering Committee members were invited by the NSF to bootstrap the Workflows Group. In the next phase, this Steering Committee must expand to include new members, both funded and unfunded, who will be stakeholders in the working group. The Steering Committee will thus need to include representative end users, workflow researchers, representatives of other relevant projects, liaisons with other EarthCube working groups, and funding agency representatives.

A detailed roadmap including updates, and additional information can be found at https://sites.google.com/site/earthcubeworkflow.

A graphical overview of the Workflows Community Group Roadmap is shown below. The overall timeline is highlighted vertically in the middle, with the Workflows Working Group and the Workflows Synthesis Center interacting synergistically to support the roadmap activities.


Overview Of Proposed Workflows Roadmap
EarthCube Workflows Community Group
EarthCube Community
EarthCube Governance
EarthCube Working Groups
EarthCube Scientists
Independent Investigators
Geoscience Observatories and large projects
EarthCube Infrastructure Development
Geoscience Organizations (ESIP, AGU, etc)
Other Relevant Organizations (OGC, W3C, etc)
Workflows Working Group
guidance and priorities
needs and assessments
Workflows Synthesis Center
Steering Committee
Establish Organization in Months 1-5, with WG Task Forces supported by clusters of Center staff assigned to them, and with 3 concentration thrusts
Synthesis Center Co-Directors
T+6
Communications Task Force
Phase I
12 months
Awareness Thrust
Outreach, Identify opportunities, Tech base, Liaisons, Evangelize
Communications Cluster
Community Interaction Task Force
Community Interaction Cluster
T+18
Status & Requirements Task Force
Phase II
18 months
Enabling Thrust
Learn requirements & practice, Link scientists and technologists
Status and Requirements Cluster
Engagement Task Force
Engagement Cluster
T+36
Prototyping Task Force
Phase III
24 months
Infrastructure Thrust
Exemplar prototypes, Impact assessment, Sharing and integration
Prototyping Cluster
Assessment Task Force
Assessment Cluster

Communications

Effective communication plans and mechanisms will be essential for meeting the workflow roadmap goal for ubiquitous adoption of workflow technologies. Current communication shortcomings include lack of awareness of workflow technologies by geoscientists and lack of understanding of geoscience requirements by workflow researchers and developers. This leads to problems such as invention of redundant, individually unsustainable tools and lost opportunities to collaborate on long-term challenges such as scientific reproducibility and operational efficiency. To be successful, we must address the communication barriers between geoscientists and cyberinfrastructure researchers.

A Communications Task Force will create connections with the community, materials for dissemination of workflow concepts and opportunities, and engagement activities such as workshops and virtual meetings.

Challenges

The overall goal of the Workflow Roadmap will be to make workflows ubiquitous within the geosciences and to further develop or enhance the workflow tools to meet the needs of geosciences. Several challenges must be overcome to reach this goal.

Technical Challenges: Workflow researchers are constantly gathering requirements from the scientific community, which are sophisticated and beyond reach of current technologies. Basic research needs to be done in the context of EarthCube requirements, as the capabilities required to support the EarthCube vision only exist in part. The group will have to develop mechanisms to facilitate early transition of new capabilities to users. To ensure success, these activities need to occur as a partnership between scientists and developers as new workflow capabilities are addressed.

Broader Adoption: While there are a number of workflow systems that are used and/ or well-known in the geosciences community, there is also much reinvention and lack of use. The tension between encouraging adoption of mature workflow systems versus development of lightweight customized systems or simple scripting solutions will need to be addressed. A large percentage of geoscientists are not using any workflow tool. This has a number of consequences: lost efficiency, lost metadata, lack of reproducibility, limited or no access to national geoscience datasets, problem of national geo-spatial/temporal data on secure federal servers with many different formats, etc. The challenge is to increase the access and efficiency of access of workflow technologies to geoscientists.

Reproducibility: Reproducibility, a cornerstone of the scientific method was identified as an important problem in interactions to date with the community. Reproducibility will require using semantic representations that document enough details about scientific processes in a reusable form, so they can be easily re-run by others and adapted to new problems. True bit-by-bit reproducibility may be an impossible problem as heterogeneous execution platforms may generate slightly different results. However, the more coarser reproducibility--the scientific reproducibility needs to be attained.

Rapidly Evolving Technologies: Resources available to scientists are changing rapidly, challenging cyberinfrastructure (and particularly workflows) to stay in step. While evolving infrastructure enables more powerful computations and the storage of more data, it also introduces impedances to integration, such as the difficulties of moving data, provisioning adequate storage, computational resources, dealing with various security mechanisms, etc.

Requirements

We will need an ongoing process for obtaining, understanding and evaluating the requirements of the geoscientific community. The diversity of users is an important challenge that must be addressed when obtaining these requirements. Our first step was to outline typical use-case workflows that form the organizing principal of the Workflow Roadmap. As part of its March-June 2012 workshop series, our next step was to create a the Workflows Community Group questionnaire 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. The responders covered a full range of geoscience research.

The next step will be to begin to develop/design a basic strategy for aligning broad user requirements determined from the surveys and workshops with existing and novel workflow technologies. We will need to develop a matrix associating use-cases with workflow technologies that recognize the particular data and model needs in each case, that allow for automated management and sharing of information, integrate resource planning and scheduling, data quality assurance and generally provides a test-drive of a new vehicle for research discovery.

A Prototyping Task Force will test the technical requirements posed by the community with prototypes of typical use-cases. The process will require follow-up, evaluation, and community consensus on all phases of the Workflow Roadmap.

Status

In general, it is difficult to assess the current state of the art in the various fields and commercial sectors. This type of assessment needs to happen as part of an ongoing earthcube activity (both because of the scope of the activity and the dynamic nature of the state-of-the art technologies). This activity can be done by the Status and Requirements Task Force through interactions with the science-focused workshops planned for the Fall of 2012. The roadmap surveys workflow solutions from the geoscience community, the cyberinfrastructure community, and commercial vendors.

Solutions and Process

The Workflows Working Group will include an Engagement Task Force that will: 1) provide guidance to geoscientists in identifying approaches to address their workflow needs, 2) assist scientists in evaluating potential workflow technology solutions, 3) request the support of the Status and Requirements Task Force and the Prototyping Task Force when necessary 4) disseminate their expertise in workflow solution approaches.

The underlying basis for identifying approaches to address workflow needs is the creation of a situation specific workflows capability maturity model. The process to identify technology solutions is based on a technology evaluation framework.

The roadmap outlines processes for identifying workflow approaches and identifying technical solutions. Processes will also be used for identifying appropriate standards, developing use cases and reference implementations through open community processes.

A Community Interaction Task Force will document use cases for existing and potential uses of workflows in the geosciences. It will also identify and prioritize needs for basic research in workflows motivated by grand challenges in the geosciences, and facilitate transfer of new advances in workflows research into geoscience infrastructure and adoption by scientists. It will also document existing standards and recommendations for adoption and interoperability.

An Assessment Task Force will track and assess the impact of workflow technologies across geosciences through the collection of quantitative and qualitative data at the early stages of EarthCube and as the roadmap activities progress.

Timeline

The overarching goal of the workflows working group is to make workflows ubiquitous within the geosciences community. This roadmap is motivated towards addressing the challenges we see in the community as extensively discussed in the previous sections. The biggest issues against achieving the overarching goal of ubiquity is the lack of awareness on how to map science challenges into workflow technologies that would improve the process, and the diverse and dispersed workflow community. Hence our activity prioritization, our milestones and the associated timelines are heavily influenced towards addressing the major issues early on, in the roadmap execution.

The timeline is divided into three overlapping thrusts: 1) Awareness Thrust, focused on community outreach and requirements gathering, 2) Enablement Thrust, focused on prototyping proofs of concept and working with the community to disseminate workflow technologies, 3) Infrastructure and Services Thrust, focused on developing community infrastructure that would include workflow publication and citation, workflow sharing, workflow execution resources, and other substantial community resources concerning workflows.

Management

The goal of management is to execute the roadmap and its principal goal of making workflows ubiquitous within the geosciences.

The working group management must be as effective as possible. Traditional management processes are inadequate for the roadmap evolution and execution, since we must coordinate multiple independent organizations and individuals. Thus the problems that we need to solve can be categorized into two perspectives - organizational and individual.

The challenges from an organizational perspective include establishing an effective organizational structure that would enable efficient strategizing; establishing an effective operational structure that would ensure smooth and timely operational activities; establishing effective processes for creating groups, organizations, etc; and establishing effective processes to facilitate consensus and enable efficient decision-making.

The problems that need to be addressed from an individual's perspective include efficient and productive use of participants' time, creating incentives beyond funding to encourage participation, and supporting and rewarding initiative by individuals;

The organizational goals will primarily be achieved through the organizational structure of the workflows working group and the associated open community process model. The goals from an individual's perspective will to a large extent, be facilitated by the substructures within the overall organizational structure, and the associated open community process model.

The strategy will be to establish a central Steering Committee for the Working Group with the flexibility that allows its members to take initiative to address problems. We also plan on establishing an institute that would function as a Synthesis Center that will support the Working Group, and specific Task Forces that would enable the implementation of the strategies proposed in this roadmap. Each of these structural components and their operational processes and primary responsibilities are discussed in detail within the roadmap.

Risks

A number of risks have been identified, including not establishing meaningful requirements, substantive differences in user requirements, not addressing workflow requirements, inadequate communication with the scientific user community, lack of adoption, and choosing the wrong software engineering methodology. The primary mitigation mechanism is the implementation of a community-based governance model that will be specifically charged with representing the community.

Acknowledgements

The 2012 Workflows Community Group Roadmap activity was co-chaired by Ewa Deelman (Information Sciences Institute, University of Southern California), Ibrahim Demir (Department of Civil and Environmental Engineering, University of Iowa), Christopher Duffy (Department of Civil and Environmental Engineering, Pennsylvania State University), Yolanda Gil (Information Sciences Institute, University of Southern California), Suresh Marru (Pervasive Technology Institute, Indiana University), Marlon Pierce (Pervasive Technology Institute, Indiana University), and Gerry Wiener (National Center for Atmospheric Research).

This work was supported through National Science Foundation under grant # EAR-1238216 as part of the NSF EarthCube initiative.