Process Technology to Facilitate the Conduct of Science

Leon J. Osterweil[1], Alexander Wise[1], Lori A. Clarke[1], Aaron M. Ellison[2],
Julian L. Hadley[2], Emery Boose[2], and David R. Foster[2]

[1] University of Massachusetts, Department of Computer Science,
Computer Science Building, Amherst, MA 01003
{ljo, wise, clarke}@cs.umass.edu
[2] Harvard University, Harvard Forest, P.O. Box 68,
Petersham, Massachusetts 01366, USA
{aellison, jhadley, boose, drfoster}@fas.harvard.edu

Abstract. This paper introduces the concept of an analytic web, a synthesis of three complementary views of a scientific process that is intended to facilitate the conduct of science. These three views support the clear, complete, and precise process documentation needed to enable the effective coordination of the activities of geographically dispersed scientists. An analytic web also supports automation of various scientific activities, education of young scientists, and reproducibility of scientific results. Of particular significance, an analytic web is intended to forestall the generation of scientific data that are erroneous or suspect, by using process definitions to prevent incorrect combinations of scientific results. The paper also describes experiences with a tool, SciWalker, designed to evaluate the efficacy of this approach.

1 Introduction

1.1 A Vision and a Caution

The Internet has created the need for a new focus on the processes by which science is done. Worldwide scientific collaborations such as Globus [1], and specific projects such as GriPhyN [2] are beginning to use the Internet to create opportunities for scientists to make data available to worldwide communities, thereby enabling expedited collaborations among geographically distributed researchers. While this creates opportunities through the broader availability of more comprehensive scientific analyses, it also creates the need for stronger and more effective control of dataset distribution and utilization. We believe that an essential component of this control is definition of the processes by which datasets and other key artifacts of science are developed, evolved, and promulgated.

If the vision of broad collaboration among geographically dispersed scientists is to be achieved, the scientists must be sure that they have the same view and understanding of the collaborative activity in which they are engaged. This suggests the need for some medium that is effective in supporting clear, complete, and precise communication about the scientific processes in which all are participating to help assure that the collaboration will produce correct results. Another benefit of such a

M. Li, B. Boehm, and L.J. Osterweil (Eds.): SPW 2005, LNCS 3840, pp. 403 – 415, 2005.

medium is its value as the basis for the development of definitions of scientific processes that might be promulgated and published, thereby facilitating community consensus and aiding in the education of younger scientists.

Whereas clear, complete, and precise process definitions can facilitate successful coordination and community education, executable process definitions can do much more, potentially helping define the way in which computers and communications technologies can be harnessed to take off of the shoulders of scientists many of the (especially more mundane and straightforward) steps of such processes. One immediate benefit of this is the possibility that such executable definitions of scientific processes might speed the rate of scientific discovery, by supporting automation of tedious activities such as dataset management and communication. This capability also offers the possibility that these processes might be used by independent scientists to validate published scientific results, thereby facilitating the reproducibility of results, an activity that is at the very core of the conduct of modern science.

Such a process definition capability could address another key concern, namely that scientific datasets might be used in misleading and incorrect ways if the precise context in which they were created is not communicated to, and respected by, other scientists. Scientific results are derived through increasingly complex sequences of scientific processes, such as sampling, cleaning, transformation, data mining, statistical inference, and evaluation. Often different processes are performed by different scientists at different times and in different places. And, while the data resulting from these processes is readily available, the processes themselves generally are not. We are concerned about the resulting difficulty of independent reproduction of scientific results, as the need for reproducibility is a bedrock requirement of modern science, and the possibility that different scientific teams may misapply results due to differences in their understandings of how scientific datasets have been produced. The lack of clear understandings of the processes by which these datasets have been produced thus stands to create less agreement, rather than more, and a reduced basis for being able to have the kinds of careful and precise debates needed to arrive at understandings of why differences exist, and how to resolve them. Ultimately, we are concerned that lack of understanding of the processes used to create datasets will inevitably cause some scientists to combine results in ways that will lead to incorrect or misleading conclusions.

In order to avoid this situation, the processes used to generate published data and results, including the tools and algorithms employed by those processes, must be clearly, completely, and precisely defined, and then made readily available. The magnitude of this task should not be underestimated. Modifications to any of the tools, algorithms, or subprocesses used in a scientific process may be inadvertent, as when a software package is updated or the underlying operating system is modified. Lacking awareness of these modifications, subsequent scientific processing (e.g., that done in order to reproduce results) may proceed under the incorrect assumption that the original scientific process is being executed. But if changes to the process have been made, then the original scientific process may indeed not have been repeated, leading either to different results or to the false conclusion that confirmation of prior results has occurred (see [3] for a recent example of the impact of changing algorithms on EPA's particulate matter standards for air quality).

We believe that sounder and more efficient science can facilitated by the Internet, but only if the expedited access to data that it allows is tempered by use of process definitions capabilities of the sort that we describe in this paper.

1.2 A Strategy

To ensure that scientific datasets are adequately well documented to support effective collaboration, education, automation, and reproducibility, and, moreover, to guard against misuse of datasets, potentially resulting in confusion and faulty science, we propose that every dataset generated by a research project should have attached to it structured process metadata that formally describes the processes by which the data were derived, including the sequence of tools, techniques, and intermediate datasets used. The representation of such process metadata information is intuitively what we refer to as an *analytic web*. In the next section we provide a more formal definition of this notion.

In the meantime, however, we can be more specific in stating the goals in developing the concept of an analytic web to be to:

- Facilitate scientific community understanding by providing a medium for the clear, precise, and complete communication about scientific processes;
- Promote effective collaboration in scientific discovery by larger, and geographically more dispersed, communities;
- Support expedited scientific activity by effective incorporation of computer and communications technologies into scientific processes;
- Forestall the possibility that scientific datasets will be misunderstood and misused, thereby leading to faulty scientific results.

The remainder of this paper describes our approach to creating technologies for defining analytic webs, and our early work to evaluate this approach.

2 Formal Description of an Analytic Web

An analytic web is a formal representation of a scientific process, in the form of structured metadata that completely and accurately describes the process, and is sufficient to support execution of the process. Our research suggests that an effective way to represent an analytic web is by means of a coordinated collection of three specific types of graphs – a dataflow graph, a dataset derivation graph, and a process definition graph – all of which were originally developed for use in defining and controlling software development projects (e.g., [4]). In this paper we demonstrate the use of these three graphs by applying them to the formalization of different aspects of a specific ecological data processing process of considerable scientific interest and importance. We argue that the analytic web represented by these three graphs makes an important contribution to assuring the understanding, executability, and reproducibility of this process, and to science in general.

2.1 Dataflow Graphs

A dataflow graph (DFG) defines which types of datasets are acted upon by which types of processes (tools, activities) in order to produce other types of datasets. A DFG documents the relationships among datasets and process types, which are inherently generic. Examples include “rainfall data”, “statistical package”, or “interpolation via regression”. A DFG is analogous to a recipe: “combine flour, eggs, seasonings, milk and water to make a batter”. Like a cookbook, the clarity and comprehensibility of a DFG facilitates the reproduction of scientific processes.

In the DFG shown in Fig. 1, the rectangular nodes represent the types of the datasets to be created and used, the rounded nodes represent the tools, techniques, and human activities that are to be performed, and the edges represent the flow of datasets into and out of these processes. Thus, this figure specifies that “DataModelA”, an artifact of type “Type1”, and “DataModelB”, an artifact of type “Type2”, are both required as inputs to an activity, called “Activity”, which then produces “DataModelC”, an artifact of type “Type3” as its output.

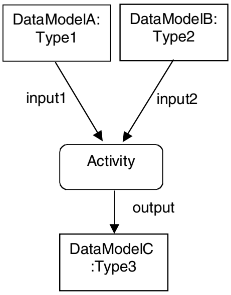

Fig. 1. Dataflow Graph

2.2 Dataset Derivation Graphs

In contrast, a dataset derivation graph (DDG) documents the instances of datasets produced by the actions of specific tools operating on other specific datasets. Dataset instances, which are uniquely specified, are the usual focus of attention in scientific processes. Examples include “rainfall data collected at the Harvard Forest on 1 June 2004 at hourly intervals”, “SAS version 6.1”, or “non-linear regression using nls2 [5]”. Continuing the analogy of thinking of a DFG as a specification of a recipe, then the DDG specifies the specific items resulting from following that recipe. This, the DDG might specify all of the final and intermediate products generated in baking a spiced chocolate prune cake for Jane’s 60th birthday following the Joy of Cooking 10th edition, 1978. Reproducibility demands the documentation provided by the DDG, namely the specific datasets and tools that were actually used.

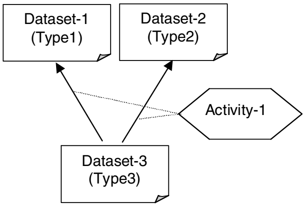

Fig. 2. Data Derivation Graph

The example data derivation graph (DDG) shown in Fig. 2 keeps track of the specific datasets that have been used and derived by the actions of the tools specified in a corresponding DFG. In the DDG, a clipped box represents each actual dataset (instance) created by executing a process. Each node is connected by an edge to the dataset(s) from which it was derived. The edge is annotated with a specification of the specific tool instance (e.g., the exact version of a statistical routine or software utility), or sub-process instance (e.g., a representation of another analytic web) used for the derivation.

Thus Fig. 2 specifies that “Dataset-3”, an artifact of type, “Type3”, was created by the actions of a tool recorded as “Activity-1”, using as inputs “Dataset-1”, an artifact of type “Type1”, and “Dataset-2”, an artifact of type “Type2”.

2.3 Process Definition Graphs

The augmentation of the information in a DFG with the information contained in a DDG does not provide sufficient documentation to always define the way in which datasets should be, and actually are, produced. The DFG defines the nominal way in which types of activities and tools are to be sequenced in order to produce specified types of datasets and other artifacts. The DDG does indeed record the exact dataset and tool instances that were actually used to produce various dataset results. But the DFG cannot be relied upon to incorporate sufficient checks and controls to assure that the actual instances chosen for participation in the DFG-defined process are consistent with each other, and suitable for use in the process of generating valid scientific results. The DFG only assures that such results are of the right type. Moreover, the DFG is effective for defining nominal processes, and is generally ineffective for defining how the process react when exceptional or unusual contingencies requiring non-nominal processing arise.

The detection and handling of such incompatibilities and non-nominal situations must be defined as part of any process if it is to be of genuine value to scientific investigation in the real world. Scientists generally are aware of such situations, and have appropriate remedies (although not always), but standard DFGs can make it hard or impossible to specify such contingencies and remedies clearly and completely. Therefore, our concept of an analytic web augments the information in a DFG with a more complete and articulate procedural description in the form of a process definition graph (PDG).

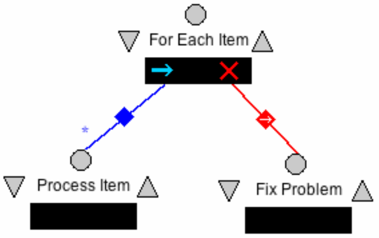

Fig. 3. Process Definition Graph in Little-JIL

A PDG defines the essential procedural details such as the order in which steps must be taken, but augments this with such additional features as preconditions for step execution, post-condition checks to determine whether or not processing has been successful, procedures to use when various exceptional conditions occur at various places in the process, conditions under which processing sequences are to be either iterated or terminated, and checking to assure that artifacts used are consistent with each other and with the activities employing them. In short, the PDG specifies the procedural flow of an analytic web, but also incorporates additional features to assure that dataset combinations are acceptable even when the process has to deal with exceptional conditions.

Figure 3 is a PDG specification, specified using the process definition language Little-JIL [6]. Space does not permit a full explanation of this language. Therefore, it must suffice to say that Fig. 3 specifies that the process “For Each Item” consists of repeated sequential executions of the “Process Item” activity, but that errors encountered in doing so are to be responded to by the execution of the “Fix Problem” activity, after which the next “Process Item” activity is to be initiated. More features of this language will be provided in the more comprehensive example presented in the next section.

3 An Example Analytic Web

We illustrate the need for, and the application of, an analytic web through an example drawn from the field of ecology. This example entails the processing and management of a type of data called eddy covariance data. The eddy covariance method is a commonly used technique for long-term measurement of the carbon exchange (e.g., the absorption of gases such as CO_2 into living organisms, such as plants) of whole ecosystems, and a useful tool in the study of global warming. Briefly, eddy covariance estimates CO_2 absorption by plant life such as forests from the covariance of CO_2 concentration and vertical wind velocity [7]. CO_2 measurements are taken continuously over an extended period by a structure, called a flux tower, located at a fixed location in the midst of a forest. Due to the variability in the accuracy of the data for a variety of reasons due to environmental conditions, researchers at Harvard Forest use a set of processes to identify unacceptable measurements and replace them with statistical estimates.

To identify and replace unacceptable measurements, first, they discard observed values of CO_2 flux if the wind direction is unsuitable for flux measurements. A particular wind direction may be unsuitable because local topography in a given direction from the flux tower creates unpredictable turbulence patterns, or because the forest of interest does not occur over a sufficient fetch in that direction. Second, the researchers examine the relationship between friction velocity, u* (a measure of turbulence in meters per second, which equals the square root of vertical momentum flux), and CO_2 flux for several weeks of nighttime measurements. Flux is plotted against u*, and a threshold value of u* ($u*_{threshold}$) is identified beyond which CO_2 flux does not increase significantly. Observed values of CO_2 flux are discarded if $u* < u*_{threshold}$. If data from all wind directions are suitable, the $u*_{threshold}$ criterion typically results in the discarding of <50% of the nighttime observations of CO_2 flux. On the other hand, if some wind directions are unsuitable, >75% of the nighttime observations may be rejected.

Finally, the researchers need to fill the gaps in the dataset that result from discarding observed values of CO2 flux by estimating the values that would have been observed if u* ≥ u*threshold. To fill these gaps, they fit regression models of the reliable observations (CO2 flux | u* ≥ u*threshold) to the measured environmental variables. For nighttime observations, the predictor variables are soil and air temperatures and, occasionally, soil moisture [8].

3.1 Dataflow Graph Model

The data flow graph (DFG) for the process described above is illustrated in Fig. 4. The boxes, “Tower Data”, “Environmental Data”, “Selection Criteria”, “Aggregated Data”, “Excluded Data”, “Rejected Data”, “Selected Data”, “Interpolated Data”, and “Row-Filled Data” all represent types of data used in creating a usable dataset. As models are also fixed data types, the model type, “Interpolation Model”, is also represented as a box. Processes are represented by ovals: “Create Aggregated Data”, “Segregate Data”, “Create Interpolation Model”, “Apply Interpolation Model”, “Merge Datasets”, and “Revise Selection Criteria” are all types of actions that are applied to particular types of datasets. Diamonds indicate points in which the same dataset is used as input to more than one action or subprocess.

Of particular interest is the action, “Revise Selection Criteria”, in which the criteria used to partition the data may be modified after examining the results of interpolating and merging the data. Intuitively, the DFG suggests that new criteria have been created, and that they are to be applied to previous datasets, generating new “Row-Filled Data”. The DFG also suggests that this iteration might be continued indefinitely, causing the successive generation of new criteria and new output data. While this intuition is probably correct, we note that it also illustrates a key inadequacy of the DFG, alluded to earlier. The DFG is incapable of specifying precisely which criteria are to be applied to which datasets. Indeed, it is conceivable that scientists may wish to apply new criteria to some previous datasets, or to all previous datasets, or to no previous datasets. The DFG provides no guidance about this. As we shall see, the DDG is capable of recording what datasets actually are created, and the precise datasets and activities that had been used in doing this. But the DDG and DFG together are still incapable specifying what should have been done, and what perhaps would be scientifically unsound. The need for such specification is provided in an analytic web by the PDG, as shall be seen.

Presumably it is vital that there be a precisely defined relationship between each dataset and the model from which it was created. Relationships of this sort are quite familiar to software configuration management practitioners, who rely upon configuration management (CM) tools and technologies to assure needed consistency. Up until the popularization of remote access to data, researchers were better able to exercise informal configuration management processes in their own domains, generally being capable of assuring consistent application of models and tools to appropriate datasets, and thereby assuring that they could themselves reproduce the results of their scientific investigations. However, a number of forces are encouraging the sharing of data, models, and tools among scientists in disparate locations and research groups, including pervasive access to the Internet, and mandates from funding agencies such as US National Science Foundation. This sharply increases the likelihood that a scientific investigator might access datasets remotely, and then use incorrect or inappropriate tools or models to process these datasets. Indeed, as noted above, the ability of other scientists to reproduce published results is central and essential to the establishment of the validity of such results. Thus, Internet access to datasets and models should ideally expedite and facilitate such reproduction, thereby improving the quality and the rate of scientific progress. But, configuration management mishaps clearly increase the risk that just the opposite might happen, with inappropriate combinations of datasets and tools causing an inability to reproduce scientific results, adding to uncertainty.

In our example, Harvard Forest researchers are continually getting new datasets from their flux tower, and creating new models, often based upon analysis of the outputs from previous models. In their work they have created sizeable bodies of "Row Filled Data" datasets, predictive models, and datasets produced by those models. Informal internal configuration management procedures tend to assure the scientific integrity of their results. But, Harvard Forest datasets or models are accessible by the operators of other flux towers, increasing the opportunities for validation of scientific results through their reproduction. Moreover, Harvard Forest researchers access datasets generated by other flux towers in an attempt to validate or improve their own models. In both cases, it is vital that the remote accessor of such data have the benefit of documentation or descriptions (such as definitions of the processes by which the datasets were created) in order to assure configuration mismatches do not cause the risk of creating invalid scientific analyses and datasets.

To illustrate the problem, Fig. 5 depicts the state of the execution of the process whose DFG is shown in Fig. 4, at the beginning of the second iteration. As in Fig. 2, boxes with clipped corners denote specific dataset instances. Each clipped box represents the dataset derived by the application of the activity from which it emanates to the dataset(s) input to that activity.

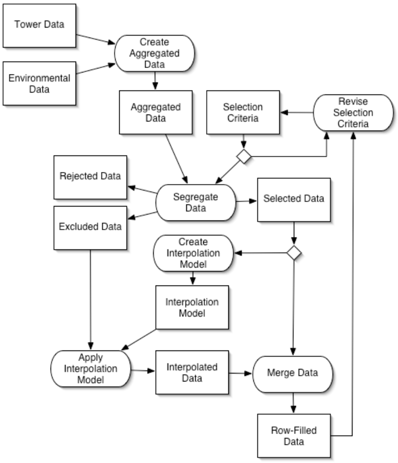

Fig. 4. Example Dataflow Graph

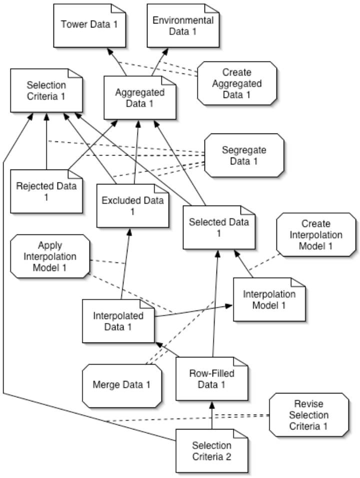

Fig. 5. Example Data Derivation Graph

Fig. 5 seems to provide a clear view of how certain datasets have been created, but the continued execution of this process will lead to the creation of increasing numbers of instances of datasets and models of the various types depicted. Thus, with iteration, there will be a growing number of instances of “Tower Data”, “Environmental Data” and “Interpolation Model”.

But, a process definition graph (using for example, the Little-JIL shown in Fig. 6) enables the specification of the configuration management information needed to ensure that specific process executions are consistent with rules or properties derived from correct process executions. Again, without delving too deeply into the syntax and semantics of Little-JIL (see [6] for a complete description), there are two details of particular import in this diagram. First, a single instance of “Aggregated Data” is used to create an instance of “Row-Filled Data”, but the ‘+’ on “Create Row-Filled Data” permits multiple instances of “Row-Filled Data” to be created from that instance. This specification removes ambiguity left by the DFG (note that other specifications, resolving the ambiguity in other ways, can also be specified using a PDG. This specification is offered only as an example). Second, the “reference” to “Apply Selection Criteria” that appears as part of “Evaluate and Revise” ensures that when the selection criteria are revised, they are applied to the same instance of “Aggregated Data” as in the previous iteration, again clearing up ambiguity left by the DFG. Without the ability to add these clarifying specifications, there would seem to be little or no protection from the improper selection of datasets as inputs to process activities, with the consequent production of results that may be incorrect or of questionable validity.

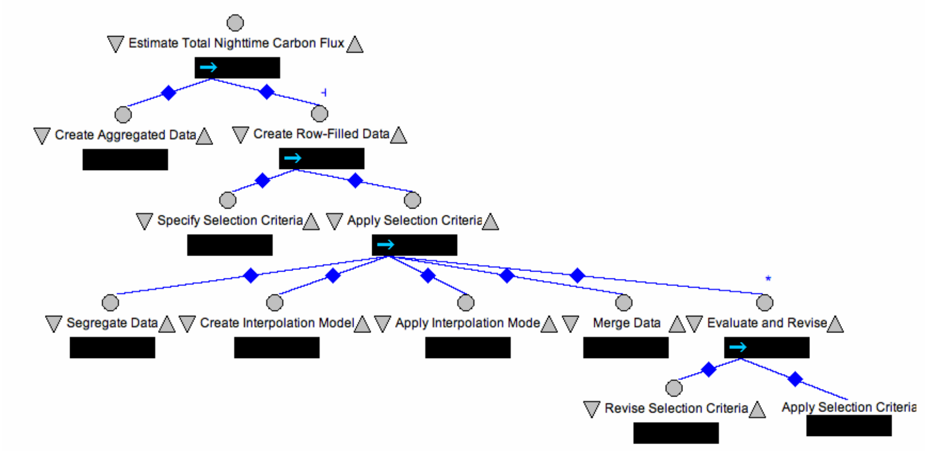

Fig. 6. Example Process Definition Graph

It seems important to note at this point that others (e.g., Estublier and his colleagues [9, 10]) have long ago noted that process definitions should be essential components of software configuration management systems. The work we describe here confirms that observation, and demonstrates that it extends beyond software configuration management, and also applies to scientific dataset configuration management.

4 Experimental Evaluation Through the SciWalker Tool

To gain some experience in assessing the value of the analytic web concept, we have developed a prototype tool, called SciWalker, as a vehicle for exploring the value of analytic webs. SciWalker supports the creation of two of the three analytic web graph representations (namely DFGs and DDGs). This capability is intended to demonstrate the value of the analytic web approach in supporting clear communication among scientific collaborators, as well as supporting education.

SciWalker also supports the execution of DFGs that it has been used to define, supports the ability to access datasets remotely across the Internet, and makes locally produced datasets available to others via the Internet. These capabilities are intended to demonstrate how analytic webs can speed the development of scientific results, and serve as facilitators for supporting reproduction of scientific results.

We performed some experiments using SciWalker to develop analytic webs that define the carbon flux process discussed in Section 3 of this paper. This experiment was designed to determine how readily such analytic webs could be defined and modified, how effective they were in communicating scientific processes to other scientists, and how easily they could be used to support remote access to datasets. A subsequent version of SciWalker will incorporate the third type of graph (the PDG), and will then be the basis for further experiments aimed at determining how effective an analytic web is in preventing inappropriate or incorrect combinations of datasets and models.

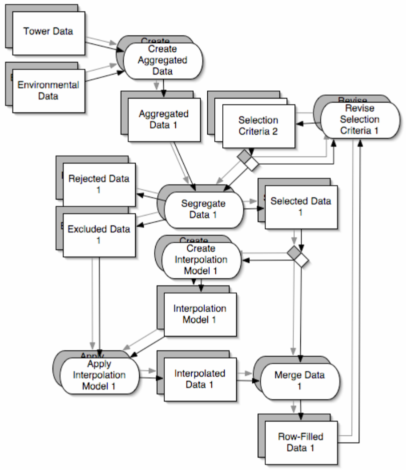

Fig. 7. Stacked DFG/DDG View

Rather than depicting DDGs independently, SciWalker depicts dataset instances as stacks piled atop the boxes (types) in the DFG representation (Fig. 7). Our early experience in using this depiction of instances has confirmed our expectation that this approach is indeed helpful to working scientists in clearly showing the specific dataset instances that have been created in successive (iterative) applications of an analytic web. The instances in SciWalker are all accompanied by specific metadata annotations, viewed through clickable menu items that provide exact and specific information about how they were generated. This approach to providing such key metadata has also proven to be useful and well received.

By using SciWalker to estimate nighttime carbon exchange from eddy covariance data, the Harvard Forest researchers were able to quickly determine the effect of varying $u*_{\text{threshold}}$ on estimated nighttime carbon flux from a forest, without employing specific statistical tools that might be inaccessible to others who are interested in recreating or modifying their analysis. Simultaneously, the tool created a complete audit-trail of the process that is easily accessible via the Internet. With this audit-trail, data that were included or excluded can be easily retrieved and examined. Other researchers have examined effects of $u*_{\text{threshold}}$ on estimates of carbon flux [11-13], but not through procedures that are easily accessible or repeatable. The researchers have indicated that SciWalker is a step forward both in ease and speed of data processing and analysis for them as individual researchers, and also a great leap forward in communicating their data analysis procedures to others.

While influences of $u*_{\text{threshold}}$ on ecosystem carbon flux estimated from eddy covariance data have been examined in several papers, effects of other meteorological variables have been examined less frequently. Wind direction is of particular interest, because forest composition is rarely uniform around a flux tower. In general, one cannot relate carbon flux to a specific type of forest without limiting the range of wind directions that provide acceptable data for processing. As carbon exchange estimates and statistical models of carbon exchange for one forest cannot be applied to other forests unless the forest composition is similar in the two areas, it is often important for researchers to partition eddy covariance data by wind direction in order to confine measurements to a specific forest type. SciWalker seems to be a perfect tool for supporting this, by allowing for the specification of the range of wind direction for included *versus* excluded data. In the case of the Harvard Forest estimates of carbon exchange by a hemlock forest, they included data only if the winds were from the southwest (180-270° compass bearing) because of the relatively small size of the hemlock forest they were studying, and the position of the flux tower in the northeast corner of hemlock-dominated forest. Using SciWalker, they are now examining the effects of using other ranges of wind direction, thereby including other forest types within the eddy covariance footprint.

5 Discussion, Conclusions, and Future Directions

There is an important need to develop tools and techniques that will facilitate the production of high quality scientific results. Internet access can clearly help, by making important datasets more accessible to more scientists, thereby facilitating broader collaborations. But such capabilities must be balanced by additional capabilities for helping scientists to understand the ways in which the datasets they access have been developed. In this paper we demonstrate that the concept of an analytic web can be used as the basis for providing process metadata capable of providing scientists with the information that they need in order to assure that their use of remotely accessed datasets is safe and correct. An analytic web consists of three different graphs that together provide this capability, but also offer the promise of facilitation of education, effective application of computer support for scientific investigation, and catalysis of community debate about most effective scientific methods.

Our proposal to create analytic webs as syntheses of three specific types of graphs seems quite promising, based upon our initial work with the SciWalker prototype and its application to eddy flux data used to estimate whether forests are sources or sinks of CO_2. In its current implementation, SciWalker incorporates only two graphs, data flow graphs and data derivation graphs. Our preliminary use of this prototype has indicated that these two graphs can be used effectively to support process definition, computerization of some process steps, and reproducibility of results. In addition, our application of SciWalker already has led to new scientific insights and interesting new results.

Future versions of SciWalker will incorporate the PDGs necessary to support assessment of the correctness of the use of datasets, and process reliability. It is our goal that scientific analyses eventually be accompanied by process certification metadata derived from the (presumably successful) application of formal process analyzers to our process metadata. These certifications would then be usable by other scientists to guide them away from dangerous misuse of datasets or combinations of processes. The net result will be science that is not only more rapid and efficient, but also more reliable and reproducible.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Award No. CCR-0205575. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

We are also grateful to many colleagues who supported this work and contributed key ideas that have led to our analytic web concept and our SciWalker prototype tool. In particular, we wish to thank Ed Riseman, Al Hanson, David Jensen, Paul Kuzeja, Howard Schultz, Bert Rawert, George Avrunin, and Mohammed Raunak for their advice, support, encouragement, and many stimulating conversations.

References

1. *Globus consortium description*. http://www.globus.org/
2. *GryPhyN project description*. http://www.globus.org/about/news/GriPhyN.html
3. Dominici, F., A. McDermott, and T.J. Hastie, *Improved Semi-parametric Time Series Models of Air Pollution and Mortality*. Journal of the American Statistical Association, 2004(99): p. 938-948.
4. Ghezzi, C., M. Jazayeri, and D. Mandrioli, *Fundamentals of Software Engineering*. 2nd Edition ed. 2003, Upper Saddle River, NJ: Pearson Education, Inc.
5. Huet, S., et al., *Statistical Tools for Nonlinear Regression: A Practical Guide with S-Plus and R Examples*. 2nd Edition ed. 2004, New York, NY: Springer-Verlag, Inc.

6. Wise, A., *Little-JIL 1.0 Language Report*, in *Computer Science Technical Report*. 1998, University of Massachusetts: Amherst, MA.
7. Baldocchi, D.D., B.B. Hicks, and T.P. Myers, *Measuring Biosphere-Atmosphere Exchanges of Biologically Related Gases with Micrometeorological Methods.* Ecology, 1998(69): p. 1331-1340.
8. Savage, K.E. and E.A. Davidson, *Inter-annual Variation of Soil Respiration in Two New England Forests.* Global Biogeochemical Cycles, 2001(15): p. 227-350.
9. Belkhatir, N. and J. Estublier. *Software Management Constraints and Action Triggering in Adele Program Database*. in *1st European Software Engineering Conference*. 1987. Strasbourg, France.
10. Belkhatir, N., J. Estublier, and W.L. Melo. *Software Process Modeling in Adele: The ISPW-7 Example*. in *Proceedings of the 7th International Software Process Workshop*. 1991. San Francisco, CA: IEEE Computer Society Press.
11. Hollinger, D.Y., et al., *Spatial and Temporal Variability in Forest-Atmosphere CO2 Exchange.* Global Change Biology, 2004.
12. Barford, C.C., et al., *Factors Controlling Long- and Short-term Sequestration of Atomspherics CO2 is a Mid-latitude Forest.* Science, 2001(294): p. 1688-1691.
13. Saleska, S.R., et al., *Carbon in Amazon Forests: Unexpected Seasonal Fluxes and Disturbance-induced Losses.* Science, 2003(302): p. 1554-1557.