COMMUNITY-BASED METADATA INTEGRATION FOR ENVIRONMENTAL RESEARCH

Joe Futrelle[1], Jim Myers[2], Barbara Minsker[3], and Peter Bajcsy[4]

ABSTRACT

The ability to aggregate information about environmental data and analysis processes across tools and services and across projects provides a powerful capability for discovering resources and coordinating projects and a means to convey the rich, community-scale context of data. In this paper, we summarize the science and engineering use cases motivating the metadata and provenance infrastructure of the Environmental Cyberinfrastructure Demonstrator (ECID) Cyberenvironment project at the National Center for Supercomputing Applications (NCSA) and discuss the requirements driving our system design. The user-level metadata and provenance capabilities being developed within ECID are described and we summarize the team's experiences in building them, and show how our experience can inform the continuing development and refinement of collaborative environmental science environments.

1. INTRODUCTION

The Environmental Cyberinfrastructure Demonstrator (ECID) project at the National Center for Supercomputing Applications (NCSA) is exploring cyberinfrastructure directions for environmental research, motivated in particular by the requirements being identified through the NSF's CLEANER/CUAHSI/WATERS planning activities. These requirements entail providing integrated access to community data collections, codes, sensor networks, literature, and community members, and tracking the relationships between them. National environmental observatories will soon provide large-scale data from diverse sensor networks and community models. While much attention is focused on piping data from sensors to archives and users, truly integrating these resources into the everyday research activities of scientists and engineers across the community, and enabling their results and innovations to be brought back into the observatory is also critical to long-term success of the observatories.

The notions of *provenance* and metadata are central to achieving integration at this scale. Metadata is often thought of as descriptive information that can support discovery of resources.

[1] Senior Research Coordinator, National Center for Supercomputing Applications, University of Illinois, Urbana, IL 61801, USA (futrelle@ncsa.uiuc.edu)
[2] Associate Director for Cyberenvironments, National Center for Supercomputing Applications, Urbana, IL 61801, USA (jimmyers@ncsa.uiuc.edu)
[3] Associate Professor, Department of Civil and Environmental Engineering and National Center for Supercomputing Applications, University of Illinois, Urbana, IL 61801, USA (minsker@ncsa.uiuc.edu)
[4] Research Scientist, National Center for Supercomputing Applications, University of Illinois, Urbana, IL 61801, USA (pbajcsy@ncsa.uiuc.edu)

Provenance – information about the activities and resources involved in the production of a resource – supports deeper collaboration and connects data and process-centric views of scientific efforts. For example, assessing the relevance and validity of a scientific data product (e.g., a visualization, model or workflow result, or a published paper) involves being able to describe the context in which a data product was produced. This can include: the conditions of the acquisition of the original, unprocessed data; the stages of processing that led to the production of the data product from the original data, and the associated rationale; and the relationship of the data product and process to the relevant community hypotheses, projects, and ongoing activities.

To support provenance at scale – across projects, across disciplines, and across independent tools (e.g., desktop data processing components, online databases, data acquisition systems), we believe that a focus on simple, high-level, real-world semantic contexts such as social networks, data processing procedures, and data collections is required, despite the fact that individual tools and subsystems may have much more structured and precise definitions of data sets, processes, and social organizations. Tracking provenance at such a scale then requires a mechanism capable of interacting with tools maintaining rich, evolving descriptive information about data products, people, and processes in their preferred format, level of granularity, and level of formality and abstracting sufficiently to provide a coherent picture of the end-to-end activities involved in the production of scientific knowledge. In ECID, we are investigating how generic, standard metadata technologies can be applied to provenance capture and integration throughout ECID's online environment with minimal a priori agreement on standard vocabularies or provenance-related programming interfaces or protocols. Although ECID is formally a project, as a Cyberenvironment, it is being constructed from middleware created by independent efforts and we anticipate the continuing inclusion of new third party tools over time, it thus is a reasonable initial proxy in which to develop provenance mechanisms scalable to a ubiquitous and persistent national/global cyberinfrastructure.

2.0 FROM CYBERINFRASTRUCTURE TO CYBERENVIROMENT

The ECID project demonstrates the emerging concept of a Cyberenvironment. Cyberenvironments bring together distributed, heterogeneous tools for collaboration, information processing, content management, and collaboration to provide support for community activities with unprecedented tool integration. Unlike traditional cyberinfrastructure frameworks such as portals, workflow systems, and Grids, Cyberenvironments are intended to support complex work processes that span community interaction, data acquisition, content management, analysis, and publishing. Cyberenvironments emphasize integrating and supporting work processes rather than standardizing software components. ECID shows how agile software engineering can enable rapid integration of heterogeneous tools to support complex science work processes.

ECID demonstrates this type of Cyberenvironment by integrating several sets of components to support different aspects of the environmental observatory use case. The CyberCollaboratory provides a web-based portal environment comprising collaboration tools such as message boards and data repositories as well as science applications such as streaming sensor monitoring. CyberIntegrator provides a powerful, easy to use end-user application for developing, maintaining, and sharing complex analysis algorithms that can be run on the user's desktop or on remote compute nodes. CI-KNOW provides social networking analysis enabling users to locate heterogeneous resources related to a topic or resource of interest, including providing referral services integrated into multiple environments. The Dashboard provides a standalone application giving concise, live updates of the status of collaborations, resources, and sensors. All components can produce or use provenance information tracking user activity and resource use, and heterogeneous components are

"aware" of other activities in the Cyberenvironment via messaging and other forms of inter-application communication.

Behind the scenes, Tupelo 2 provides 'semantic content management' facilities to ECID components. Tupelo, a semantic content management system developed for the George E. Brown, Jr. Network for Earthquake Engineering Simulation, enables distributed management of datasets and RDF descriptions backed by a variety of storage implementations, including filesystems, relational databases, and RDF triple stores such as Kowari and Sesame. A key concept in semantic content management is that, at the level of Tupelo's operation, all information about any kind of entity is simply a combination of an opaque blob of bits and metadata associated with a globally unique identifier. Thus, at the repository level, people, scientific instruments, data, workflows, documents, etc. are all first-class, co-equal entities that can be managed and annotated by any application.

Tupelo 2 also solves a current problem that RDF API's and query languages are heterogeneous and non-standard by providing a unified, simple APIs for managing RDF data and moving it between multiple, heterogeneous implementations. It also provides components that enable asynchronous metadata harvesting and that allow integrated presentation, and further inference and analysis of the collected facts and assertions. The information generated by distributed, heterogeneous components can be displayed graphically or consumed by social networking analysis codes that can provide unique contextual and summary information to users about their relationship, and those of their collaborators, to the complex sets of community data, projects, annotations, provenance, and tools. The graphical browser capability being developed allows users to explore provenance and other metadata assembled by the ECID environment by navigating along specific semantic relationships. Social network analysis and rule-based inference systems add to the corpus of metadata available. These additional relationships can also be browsed or, as the ECID project is also exploring, used to generate recommendations about related resources (e.g., data, docs, people, workflows) that are made available from within a portal or workflow engine, or via a variety of notification techniques.

2.1 Collaboration as Content

The ECID CyberCollaboratory is a collaborative space where communities of researchers, practitioners, policy-makers, and other interested parties can come together to share knowledge and information, analyze data, solve problems, and collaborate. It is currently built on the LifeRay portal system, which allows each community to easily create their own customized workspaces where they can integrate and utilize tools, data, and collaboration teams. There will be a wide variety of data analysis, communication, and collaboration tools within the CLEANER CyberCollaboratory, supported by and integrated with generic Content Management System (CMS) capabilities allowing for data sharing between components.

Currently available tools include a demonstration workflow tool with Bayesian and decision support analysis capabilities, an oil spill simulation tool, RSS feeds, discussion forums integrated with e-mail, a blog for user announcements, document repositories, chat, searchable web databases. CLEANER is using the CyberCollaboratory to support their planning of the WATERS Network, which is providing a real-world testing environment as the product is being developed. The project is implementing new features and improving the usability of the CyberCollaboratory, focusing on desirable features identified by initial users through surveys, interviews, and formal usability studies. A Cyberdashboard provides notifications and alerts to users on their desktop and access to data, metadata, and events generated from collaborators' activity and live sensors.

2.2 Social Networks as Content

Collaboration involves a variety of social interactions in addition to formalized work process steps. By capturing information about user interaction within the CyberCollaboratory and related ECID tools and managing it as content, the ECID Cyberenvironment can infer useful facts about the structure of collaborators' social networks, enabling advanced referral, recommendation, and analysis capabilities.

CI-KNOW (CyberInfrastructure Knowledge Networks On the Web) is a network referral system built into the ECID Cyberenvironment that identifies available resources of interest to users. Such resources could include other researchers, data sets, analytical tools, models, and visualization approaches. The system is built on NCSA's Science of Networks in Communities (SONIC) group's social networking technology, which provides referrals based on stated and inferred interests and behaviors. CI-KNOW can gather and collate data on user activities within the Cybercollaboratory (documents, data sets, visual-analytic tools) as well as in external digital repositories (such as web sites, bibliometric databases and online publications). CI-KNOW can use the data within social networking algorithms to generate and analyze multidimensional networks of connections between people, documents, data, etc., as shown in Figure 1. The results of these analyses are used to make proactive (unsolicited) as well as reactive (in response to a user query) network referrals. CI-KNOW has been developed as a portlet within the Cybercollaboratory, and also provides referrals within CyberIntegrator.

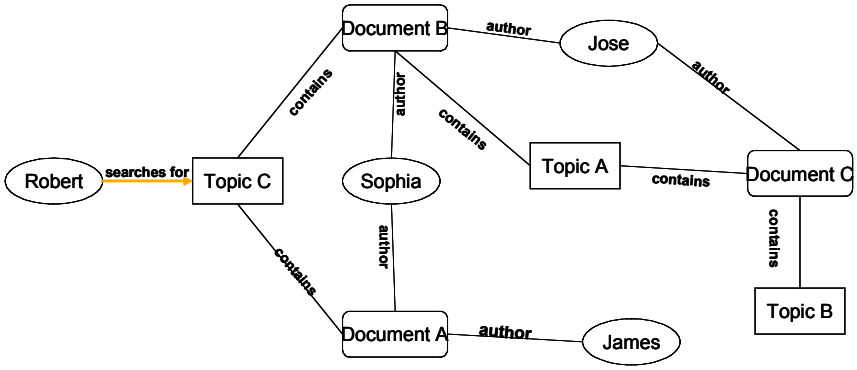

Figure 1. Example of CI-KNOW network, showing connections between users, topics, and documents in the CyberCollaboratory.

2.2 Process as Content

In order to gather metadata and provenance information about analytical and computational processes, we have prototyped a scientific process management technology called CyberIntegrator. CyberIntegrator is a highly interactive exploratory scientific process management environment to support earth observatories and to address the many needs of scientific processes. The objective of CyberIntegrator is to support exploratory analysis and to increase end-to-end scientific productivity. To address the scientific needs, CyberIntegrator has been implemented with the following key components: (1) support and integration of heterogeneous software tools, (2) provenance to recommendation pipeline, (3) event triggered execution, and (4) interactive and friendly human computer user interfaces for workflow creation and re-execution including search capabilities for data, tools and resources.

CyberIntegrator collects a hierarchy of provenance information about each execution. Each CyberIntegrator execution is described by its specific executor (e.g, MS Excel or Im2Learn), execution timestamps, a creator, and a step. Steps are characterized by a tool name, inputs, outputs and parameter descriptors. Parameter descriptors consist of a parameter name and multiple

properties described by property names and values, see Figure 2. The hierarchy of provenance information was designed based on general provenance gathering requirements for earth observatory applications. The provenance information is sent to a semantic content management system based on Tupelo 2.

Figure shows the CyberIntegrator editor and a simple graph of executions with stages (waiting, running and done) described using RDF statements. The workflow can be saved and re-run multiple times to generate provenance information with unique identifiers. By using globally unique identifiers for every execution, we are able to manage large collections of execution traces which can be queried to understand the underlying scientific work processes. The question mark button in the left upper corner of CyberIntegrator editor in Figure invokes the provenance to recommendation pipeline. In this case, the provenance information is utilized by retrieving a list of software tools based on their usage frequency of occurrence for a selected data type displayed in the left pane of CyberIntegrator editor. Other uses of metadata and provenance information for auto-completion of workflows and for establishing community standards are currently being explored.

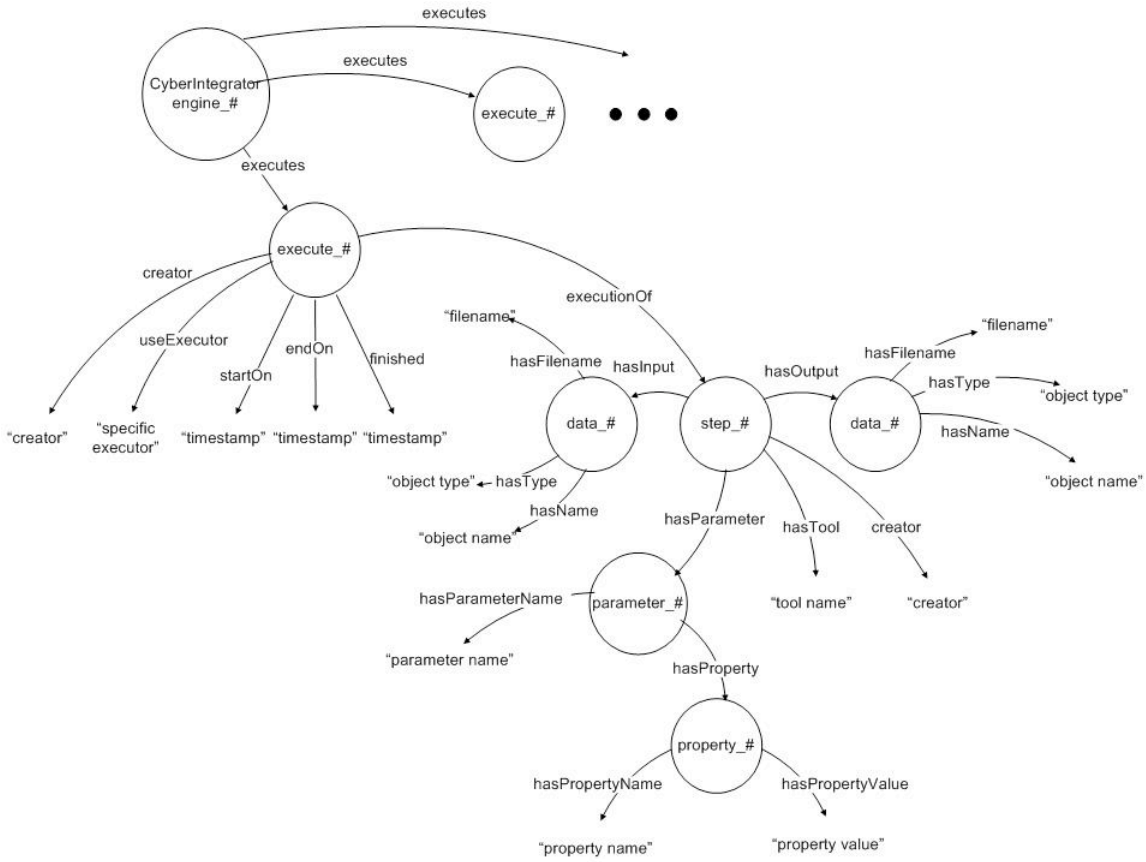

Figure 2: A schematic graph of RDF triple generation inside of CyberIntegrator.

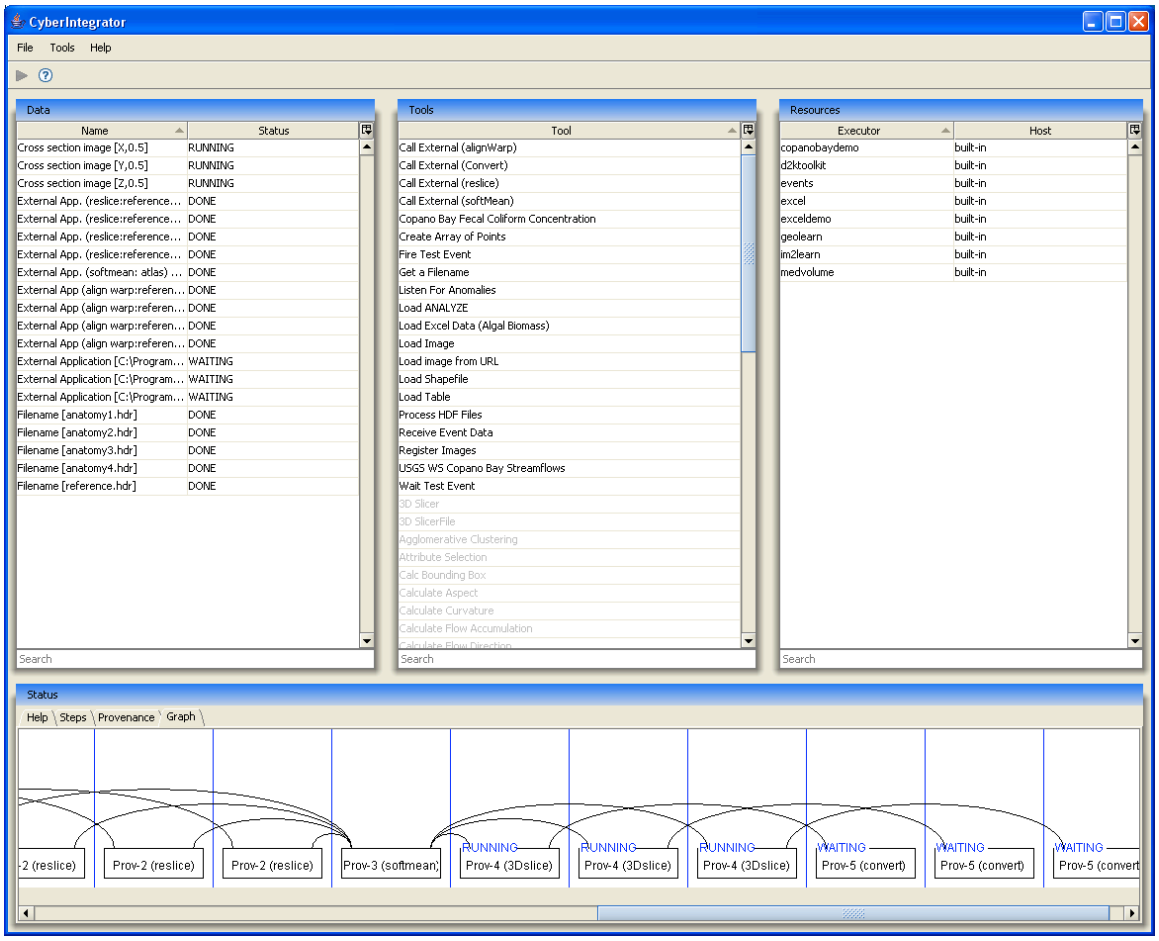

Figure 3: Workflow execution in CyberIntegrator that captures visually the provenance information about completed, running and waiting stages of different execution steps.

3. EXAMPLE USE CASE

ECID has developed a set of capabilities within the CyberEnvironment that support environmental observation via live sensor networks with collaboration and referral capabilities based on our provenance infrastructure. The CyberCollaboratory organizes community efforts into shared workspaces. When users log in, they can select the community workspace they want to enter and can access relevant resources, tools, and people.

In our example use case, developed with the WATERS Corpus Christi Bay testbed observatory community, researchers can monitor remote sensor platforms deployed in the bay that provide live data streams for a variety of environmental conditions (e.g., windspeed). Users can see a geographic representation of the sensor network using Google Maps, whose interface has been extended to enable users to select sensors by location and indicate their interest in the sensor data stream via a subscription interface. In our example use case, sensor data is processed continually via a remotely deployed CyberIntegrator application to detect anomalies, and anomaly events are delivered as notifications to the dashboard and CyberCollaboratory.

In the dashboard, users can quickly check the status of subscribed data streams by viewing a summary of events related to the stream, as well as seeing the status of collaborators and other community activity. In the CyberCollaboratory, users can view visualizations of incoming data and anomalies. Provenance information in the CI-KNOW knowledge network links data streams, sensors, and collaborative discussions with published CyberIntegrator workflows, so that a

researcher can follow links in the network to discover which application is responsible for the upstream processing that produced a derived data product (in this case, anomaly annotations on data streams). If the researcher wants to modify the anomaly detection algorithm, CI-KNOW enables them to directly open up CyberIntegrator with the workflow, modify it by altering parameters or inserting additional processing steps, and redeploy the workflow in the anomaly event subscription.

This represents a significant advantage over traditional science portals that provide end users only with the capability of parameterizing domain-specific algorithms rather than iteratively developing real-time applications by tracking data back to the algorithms that participated in its production, and then changing the application on the fly.

4. UBIQUITOUS PROVENANCE

The provenance of a digital artifact typically consists of descriptive information that spans heterogeneous work processes over the lifecycle of the artifact, which has traditionally made it difficult to capture and share provenance information. Digital Libraries typically strongly bind metadata to artifacts (e.g., using markup languages), creating problems when knowledge is embodied in multiple, heterogeneous artifacts (e.g., a scientific paper and the data referred to in the paper). Attempts to combine all information management functions relevant to some domain into a single, monolithic system inevitably fail to capture important processes that take place before an artifact is placed in the system or when it is used upon being retrieved from the system. This is by no means an inevitable result of the heterogeneity of work processes but rather a symptom of the kinds of descriptive modalities typically employed in collection management systems, which require multiple pieces of information to be co-located structurally and/or physically before the relationships between them can be managed. This approach may scale to large collections, but it doesn't scale to multiple domains, since new structural and physical relationships typically need to be designed and implemented before the salient relationships can be managed.

Provenance is therefore ubiquitous in that there is no single locus of control at which all the information relevant to the provenance of an artifact can be produced, managed, or located. RDF and other "open-world" descriptive modalities (e.g., Topic Maps) are an ideal fit for this kind of problem, since they're based on strong global identification and extensible description logics. These key features enable metadata descriptions to be assembled incrementally from independently generated parts using simple, generic mechanisms that require neither complete information nor *a priori* knowledge of the structural relationships between first-class entities that are being described (e.g., artifacts, properties, relationships such as dependency and precedence, etc.) This takes provenance management systems out of the business of enforcing domain-specific structural validation rules and enables them to become efficient and easily maintainable "dumb pipes" for information flowing from distributed, heterogeneous, lightweight components, independently of whether those components are deployed as desktop applications, portal components, web services, or supercomputer jobs.

4.1 Capturing and using Provenance

Provenance can be modeled as what Groth, Luck and Moreau (2004) call "process documentation," and to the extent that processes are automated, so can the documentation of those processes be automated. Unlike traditional manually produced documentation, which typically takes the form of a narrative, automated process documentation can be modeled as sets of discrete events implicating explicitly identified entities such as users, digital artifacts, processing steps, algorithms, parameters, and processing environments (e.g., hardware and software configuration.) Capturing these events amounts to capturing provenance, as long as the event descriptions are robust enough to be able to

be composed into a coherent account of what happened to the artifact during its participation in a complex work process. As was mentioned above, RDF and other open-world descriptive modalities meet this requirement.

Consider the derivation of a data product. In most scientific work processes, "raw" data from observations goes through several initial processing steps (e.g., cleaning, calibration, registration, etc.) before being analyzed. As in workflow systems, each of these steps can be modeled as a parameterized process that consumes one or more input datasets and produces one or more output datasets. From this abstract model we can infer a derivation or dependency relationship between the datasets. With RDF descriptions, we can describe this derivation without modifying the datasets or the processing algorithm, provided all of the relevant entities can be referenced with strong global identifiers. In ECID we accomplish this by instrumenting CyberIntegrator and other data-processing components with code that logs parameters, processing steps, and dataset identifiers as time-scoped RDF triples that can be managed independently (i.e., in Kowari) from the datasets themselves (i.e., in a data repository). We need only strongly bind a dataset to its global identifier in order to be able to integrate RDF descriptions with the data repository.

Using provenance information amounts to making queries against process documentation. For instance, if a researcher suspects that an anomaly in a derived dataset resulted from a mis-configured calibration step, having the derived dataset's identifier in hand (say, by selecting the dataset in a data repository browsing or search interface) is sufficient to enable the researcher to locate not just the source dataset, but also the algorithm-specific parameterization on the calibration step. RDF's extensibility means that process documentation can be extended to support any number of ontologies for describing domain-specific processing steps without affecting the code that is responsible for managing and querying process documentation.

Ubiquitous provenance with explicit semantics decouples descriptive information from rigid, tool-specific control flow, enabling new kinds of integration. For example, when ECID users select datasets to process in the Cyberintegrator tool, descriptive information about who they are, which dataset they are processing, and what tools they are using to process it is captured and harvested into an RDF triple store. The RDF store also includes information about ECID users' interactions in shared message boards accessed through the ECID Cybercollaboratory portal, harvested using the same approach. The harvesting mechanisms are generic; domain-specific customization resides instead in the ontologies made explicit in the provenance metadata itself.

5. CONCLUSIONS

Provenance information is critical to validating the evidence upon which scientific knowledge is based. Accordingly, scientific work environments can be made more usable, robust, and relevant by incorporating the production, management, analysis, and use of provenance information into diverse collections of heterogeneous tools used throughout a complex domain-specific work process. Data repositories, sensor networks, analysis algorithms, and collaboration tools that scientists use as part of a community work process can all participate in process documentation provided that extensible description modalities such as RDF are used that can incorporate domain-specific explicit semantics and ontologies without the intervention of software developers or system administrators. ECID demonstrates the viability of this approach by showing how an RDF-based provenance strategy can link previously dis-integrated types of tools such as social networking analysis and workflow.

Incorporating RDF harvesting into ECID portal components did not require extensive recoding or redesign of the components, since RDF can be harvested from ordinary logging information using simple, generic tools or by instrumenting isolated parts of an application with an API that closely resembles logging. Nor did it require prior agreement on an ontology or XML schema; we have found that agreeing simply on how to represent a portal user ID as a node in an RDF graph is

sufficient to enable several novel integration features. For instance, if two sets of RDF statements representing a user's interaction with data collections and a user's interaction with other users on a community message board are merged, a shared user ID can be used to link discussion threads a user participates in with data that the same user has used as part of a scientific workflow. In ECID we have used these kinds of links to rapidly design social network analysis components that can predict user preferences based on the aggregate behavior of other users (e.g., "users who selected dataset X used tool Y to process it"). RDF enables components to work together with only partial agreement on ontologies, and this enables us to evolve metadata descriptions without recoding application components or requiring hard-to-agree-on changes to a community-level union ontology or schema. This approach enables the creation of Cyberenvironments that scale to new scientific communities, domains, and work practices, resulting in richer artifacts and rapid scientific progress.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contributions of the ECID project team and many researchers at NCSA and at partnering institutions involved in Cyberenvironment efforts and in developing their component technologies, to the development of the provenance work reported. The National Center for Supercomputing Applications is funded by the US National Science Foundation under Grant No. SCI-0438712. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

Bajcsy P., R. Kooper, L. Marini, B. Minsker and J. Myers, "A Meta-Workflow Cyber-infrastructure System Designed for Environmental Observatories," Technical Report, NCSA Cyberenvironments Division, ISDA01-2005, December 30, 2005.

Bajcsy P, R. Kooper, L. Marini, B. Minsker and J. Myers, "CyberIntegrator: A Meta-Workflow System Designed for Solving Complex Scientific Problems using Heterogeneous Tools," the Geoinformatics conference, May 10-12, 2006, the USGS National Center in Reston,Virginia.

Futrelle, J. (2006) "Harvesting RDF Triples", International Provenance and Annotation Workshop, May 3-5, Chicago, IL USA.

Groth, P., Luck, M., and Moreau, L. "Formalizing a protocol for recording provenance in grids," Proceedings of the UK OST e-Science second All Hands Meeting 2004.

Marini L., R. Kooper, B. Minsker, J. Myers and P. Bajcsy, CyberIntegrator: A Meta-Workflow System Designed for Solving Complex Scientific Problems using Heterogeneous Tools, the NSF EO Modeling Workshop , poster, May 16-18, 2006, Tucson, AZ.