

# Scientific Annotation Middleware

“Lightweight Meta-data Framework”

Jim Myers, PNNL and Al Geist, ORNL  
[www.scidac.org/SAM](http://www.scidac.org/SAM)

## Summary

*The Scientific Annotation Middleware (SAM) system being created will provide significant advances in research documentation and data pedigree tracking required for effective management and coordination of the complex, collaborative, cross-disciplinary, compute-intensive research enabled through the Scientific Discovery through Advanced Computing (SciDAC) initiative. The SAM system presents researchers, applications, LIMS, electronic notebooks, and software agents with a layered set of components and services that provide successively more specialized capabilities for the creation and management of metadata, the definition of semantic relationships between data objects, and the development of electronic research records.*

Science is critically dependent on complete and accurate documentation of experiment processes and results. Scientific records enable confirmation of results, allow researchers to share findings and avoid duplication of work, and provide a means to establish credit and accountability for scientific discoveries. Traditional documentation methods have become much less effective because of the collaborative, cross-disciplinary, dynamically changing environment that is increasingly typical of modern scientific research such as SciDAC.

The Scientific Annotation Middleware (SAM) project is addressing the needs of modern scientific record keeping. The key concept behind SAM is “schema-less” data store that can accept arbitrary input and the use of dynamically registered translators to map data and metadata into the formats and schemas expected by applications and underlying data repositories. This allows researchers to capture records-related information using an arbitrary combination of electronic notebooks, applications, agents, and problem-solving environments. Scientists can later define how this information should be translated into forms

interpretable in other contexts, e.g. into the input format required by a collaborator’s software, the schema of a community database, or that of a records-management tool or automated workflow system. In the SAM model, it becomes possible to view all of the recorded information via a single interface and to simultaneously define views of that data that conform to the conceptual models of particular applications, groups, institutions, or communities.

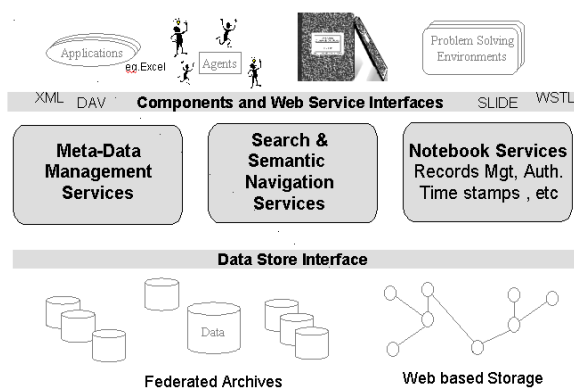


Figure 1. Interfaces and services being developed in SAM in order to improve the scientific record.

SAM, being developed by researchers at Pacific Northwest National Laboratory and

Oak Ridge National Laboratory, is a set of components and services supporting the creation and use of annotation metadata about data objects and the semantic relationships among them. SAM is built upon web, Grid, and semantic web technologies as well as the successful DOE2000 electronic notebook research.

SAM uses the Distributed Authoring and Versioning (webDAV) protocol. WebDAV is an Internet Engineering Task Force standard extension to HTTP that, like web services, uses XML to encode the content of service requests.

SAM is built on Jakarta Project's Slide content management system. SAM extends Slide to allow the generation of user-defined metadata and translation of data into other webDAV resources. To allow SAM to function as middleware, we have added interfaces to allow Slide to use external authentication and authorization mechanisms. Authentication can now be performed using any Java Authentication and Authorization Service provider, including an external username/password database, a Kerberos server, or public key certificate/Grid security infrastructure.

As an initial step in creating the notebook services layer, we have implemented the functionality required to generate a notebook. In the SAM-based notebook server, the chapters, pages, and notes within the notebook are stored as webDAV resources, and the chapter/page/note tree structure is stored as webDAV properties associated with those resources. Thus, the structure of the notebook and its contents are directly available to other webDAV-enabled applications.

**SAM Users:** The DOE Collaboratory for Multiscale Chemical Science (CMCS) is using SAM as a component of a portal-centric system designed to facilitate collaboration, data exchange, and provenance tracking across multiple

chemistry sub-disciplines. CMCS is also using SAM to link the webDAV-aware Extensible Computational Chemistry Environment and chemistry applications. The DOE Genomes to Life (GTL) program is leveraging SAM to create a "biology aware" electronic notebook that can store and manipulate biological data objects. The GTL research has many additional opportunities to leverage SAM services.

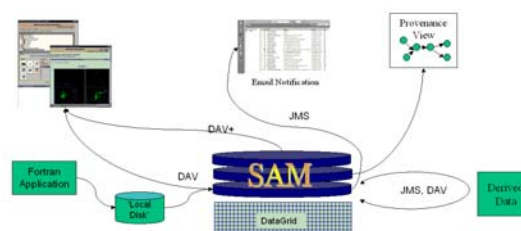


Figure 2. SAM Use Scenario. Multiple applications contribute to the research record.

**Future Work:** While SAM is already showing promise, we anticipate a variety of areas where additional work will be needed to provide a generally useful system. SAM's metadata generation and translation services will be simplified for scientists and registered as web services for easy access. SAM's semantic services layer will be implemented to minimize the extent to which applications must understand semantic languages. For example, it should be possible to make the results of a semantic inference available as an XML webDAV property. A final direction for SAM is to develop electronic notebooks that incorporate the richer, federated record SAM enables and allow scientific applications to access notebook functionality at the level of services.

**For further information on this subject contact:**

Dr. Mary Anne Scott, Program Manager  
DOE Office of Advanced Computing Research  
Phone: 301-903-6368  
[scott@er.doe.gov](mailto:scott@er.doe.gov)  
Or visit: [www.scidac.org/SAM](http://www.scidac.org/SAM)