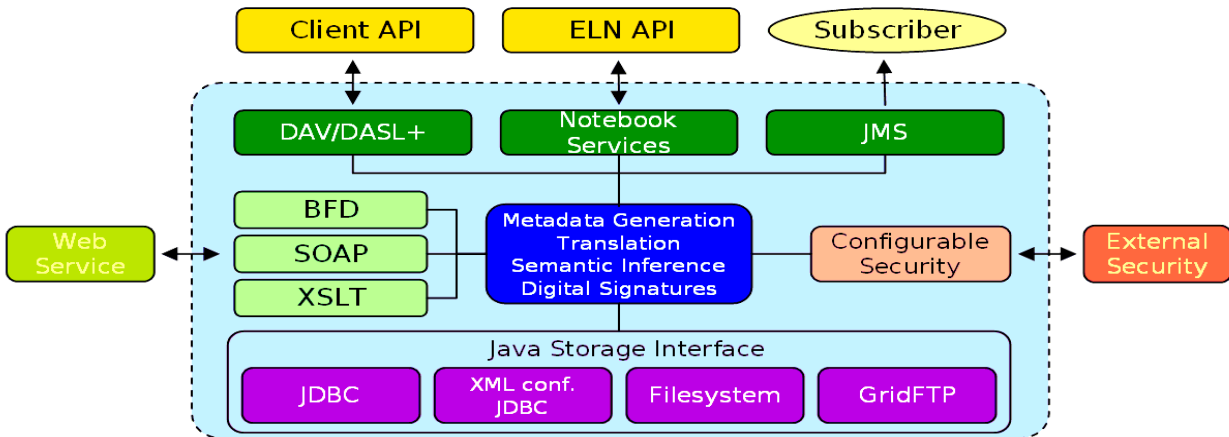


Scientific Annotation Middleware

Lead Investigator: Jim Myers, PNNL (Jim.Myers@pnl.gov)

ORNL: Al Geist, Jens Schwidder, PNNL: Alan Chappell, Carina Lansing, Mike Peterson, Tara Talbott



The Scientific Annotation Middleware (SAM) system (<http://www.scidac.org/SAM>) provides significant advances in research documentation and data provenance tracking to support the effective management and coordination of complex, collaborative, cross-disciplinary, compute-intensive research, such as that 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 among data objects (e.g. provenance), and the development of electronic research records.

Overview: The Scientific Annotation Middleware (SAM) project is addressing the needs of next-generation Grid-based scientific research to federate data and metadata, track pedigree, document research processes, and expose such information to a wide range of applications, agents, workflow services, notebooks, and scientific portals, within and across virtual organizations. The key concept behind SAM is a “schema-less” data store that can accept arbitrary input and the use of dynamically registered translators to map data and metadata into the formats, schemas, and ontologies expected by applications and underlying data repositories. This allows researchers to capture records-related information using an arbitrary combination of tools and to 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 virtual-data/workflow system.

In the SAM model, it becomes possible to view all of the recorded information via a single interface/protocol while simultaneously defining limited views of the data that conform to the conceptual models of particular applications, groups, institutions, or communities.

SAM is being developed by researchers at Pacific Northwest National Laboratory and Oak Ridge National Laboratory as a layered 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.

Technology: SAM is built on the Jakarta Slide content repository and implements 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 extends Slide to allow configurable automated metadata extraction and data translation from binary, ASCII, and XML inputs. XSLT and Binary Format Description (BFD) scripts can be registered with SAM and run dynamically using standard XSLT and BFD engines to extract metadata or create translations and data views. SAM also produces Java Message Service (JMS) events describing all data/metadata access and changes.

SAM is true middleware – it can be configured to use existing security (authentication/ authorization) and data storage services. Authentication can be performed using any Java Authentication and Authorization Service provider, including external username/password databases or public key certificate/Grid security infrastructure. Data and metadata can be stored in files, remote databases, or repositories accessed via GridFTP.

SAM also acts as an electronic notebook server compatible with the DOE2000 ELN 5.1 client. Chapters, pages, and notes within the notebook are stored/accessed directly via webDAV, and the chapter/page/note tree structure is stored as standard webDAV properties associated with those resources. Thus, the structure of the notebook and its contents are directly available to other webDAV-enabled applications.

Collaborations: 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. Several other projects are in earlier stages of investigating and using SAM, including the George E. Brown Network for Earthquake Engineering and Simulation Grid (NEESgrid). The DOE Genomes to Life program is leveraging SAM to create an electronic notebook for biological information. SAM is also being investigated for

potential integrated with PNNL's Lustre-based archive for use in biological data/metadata management.

The SAM team also participates in several standardization efforts through the Java Content Repository (JSR-170) Expert Group and Global Grid Forum groups including the Data Format Description Language (DFDL), Grid Information Retrieval (GridIR), Grid Computing Environments (GCE), and Semantic Grid groups.

Future Directions: SAM already provides powerful capabilities for data integration, metadata/pedigree management, and free-form annotation. As the project continues, we anticipate a variety of enhancements that will enable additional uses.

SAM's semantic services layer will allow configurable ontology-based transformations and inferences. It will be implemented to minimize the extent to which applications must understand semantic languages such as RDF and OWL in order to, for example, view the results of a semantic inference available, by exposing the results of semantic transformations as XML webDAV properties. We will also create web/Grid service access mechanisms.

Another direction for SAM is to develop next-generation electronic notebook capabilities for notebook lifecycle management, including digital signature and notarization capabilities, that interact with the richer, federated record SAM enables. These separable components and services will serve as the basis for a variety of third party annotation and peer-evaluation mechanisms in addition to advanced group notebooks and notebook-related capabilities embedded in other tools.

Status: SAM software, version 1.1 is available on the project website under an open source license.

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

Or visit: www.scidac.org/SAM