Social Informatics Data Grid

Bennett Bertenthal[1,5], Robert Grossman[3], David Hanley[3], Mark Hereld[1,6], Sarah Kenny[1], Gina-Anne Levow[2], Michael E. Papka[1,2,6], Stephen W. Porges[4], Kavithaa Rajavenkateshwaran[1], Rick Stevens[1,2,6], Thomas D. Uram[1,6], and Wenjun Wu[1]

[1]Computation Institute, Argonne National Laboratory and The University of Chicago
[2]Department of Computer Science, The University of Chicago
[3]Department of Mathematics and Computer Science, University of Illinois at Chicago
[4]Department of Psychiatry, University of Illinois at Chicago
[5]Department of Psychology, Indiana University
[6]Mathematics and Computer Science Division, Argonne National Laboratory

papka@ci.uchicago.edu

Abstract. The Social Informatics Data Grid is a new infrastructure designed to transform how social and behavioral scientists collect and annotate data, collaborate and share data, and analyze and mine large data repositories. An important goal of the project is to be compatible with existing databases and tools that support the sharing, storage and retrieval of archival data sets. It is built on web and grid services to enable transparent access to data and analysis resources from anywhere and to leverage new and emerging web-based technologies created by a large and growing community of developers around the world. At the heart of the SIDGrid design is a rich data model that captures notions of time, data streams, and semi-structured data attached to these streams to enable powerful manipulations of multimodal data spread across data resources. Through query and analysis services deployed against the data warehoused in the SIDGrid users can perform new classes of experiments. Shared data resources available from anywhere over the Web introduces new capabilities to the process of collection and analysis of data – collaborative annotation among them – without relinquishing control over sensitive data via an embedded security model. Through a series of workshops at which we engaged members of the broader community, and by cultivation of a few collaborative projects, we have steered the development process to provide the most important components and functions first.

Introduction

The study of human behavior encompasses a multiplicity of models and methods, but virtually all of them share the view that human behavior can be analyzed by decomposing the problem space into static variables or systems that are linearly related to each other. For example, the study of human memory emphasizes relationships between variables independent of time, even though memory is inherently a temporal process. Likewise, learning is a time-critical process, as new knowledge and skills are organized over time, but we tend to focus on the products or outcomes of this process. What is lacking in these and other domains is a way of modeling how behavior is dynamic, multi-causal and occurs over multiple time scales.

A much-needed solution to this problem is to address the study of human behavior as a dynamical system. By definition, such a system is dynamic, multi-level and multi-causal, and nonlinear. Although the study of dynamical systems has had a long and venerable history in the physical sciences, it has yet to have a major impact in the psychological sciences (Abraham, 1992). This seems somewhat paradoxical given that psychologists are interested in a wide range of phenomena that change over time, including learning, memory, thinking and development.

How can we explain this failure to explicitly incorporate dynamical systems in the study of these phenomena? The crux of the problem is that investigators studying the neural, cognitive, and social behaviors of humans lack the tools to assess multiple measures at multiple levels simultaneously and to store and analyze these measures in a common database. The discipline-based structure of traditional academic institutions, together with standard single-investigator approaches to research, is poorly suited to the study of multidisciplinary problems. Significant conceptual, technical, and analytic advances toward understanding and applying research on multimodal behaviors emerging at different time scales require multidisciplinary research and development on a larger scale than available to any individual, lab, or institution. This new field lies at the intersection of computer vision, psycholinguistics, cognitive neuroscience, neuroscience, psychology, linguistics, education, anthropology, speech and language processing, and high speed computing and networking. Successful collaboration among these diverse disciplines requires a 'material interface' (e.g., shared datasets and tools) and an intellectual interface (e.g., shared problems) to support multidisciplinary research.

The Social Informatics Data Grid (SIDGrid)[1] is working to enable researchers to capture multimodal behavior in real-time at multiple levels simultaneously, and then to store and analyze different data types (e.g. voice, video, images, text, numerical) in a distributed multimedia data warehouse that employs web and grid services to support data storage, access, exploration, annotation, integration, analysis, and mining of individual and combined data sets. Previously collected corpora and data archives in raw or partially analyzed forms will be made compatible with the database. While the SIDGrid effort is funded as a testbed, we are focused on integrating data from three broad and complementary areas of research: (1) Multimodal communication in humans and machines, (2) Neurobiology of social behavior in human and animals, and (3) Cognitive and social neuroscience.

Since the scientific applications targeted by SIDGrid are manifold, have yet to bear fruit at this stage of the development program, and space for this paper is limited, we will focus our current discussion on the technical features and underpinnings of the project. Our discussion begins with an overview and is followed by more detailed explanations of the workings of SIDGrid at each of the three layers of its design architecture: the data and computing resource layer, the service layer, and the application layer. We will end with a discussion of the main connections that we have made to established research projects and the nature of our developing relationship with each.

SIDGrid Overview

The SIDGrid architecture provides transparent access to distributed, aligned, and annotated social informatics data. Multiple data streams capturing video, audio, and eye movement data can be acquired and automatically stored both locally and remotely in the SIDGrid. Once stored in raw form in the SIDGrid, these data streams can then be transformed into formats

[1] SIDGrid Project Website – sidgrid.ci.uchicago.edu

that are compatible with software tools for annotation, coding, integration, and analysis. Although the data may be collected in one location, web-based access to the data warehouse enables researchers all over the globe to participate in the annotation and analysis of the data streams.

Our design is driven in part by the expectation that easy access to rich, integrated, multimodal data will enable qualitatively new kinds of analyses and consequently to new discoveries. The services layer of the SIDGrid integrates data collected at multiple time scales including frame synchronized multi-camera video, multichannel audio, motion capture, eye tracking, physiological measures (e.g. heart rate, EMG, EEG), brain imaging data, bioassays, and single and multiple unit recordings from animal brains, as well as surveys, interviews and demographic data. Data streams may be sampled at different rates but organized in a common database with reference to a common time base. This organization enables comparisons within and between measures at different time scales. For example, speech, gesture, facial expression, and physiological measurements attending an event or interaction can be analyzed in the same context.

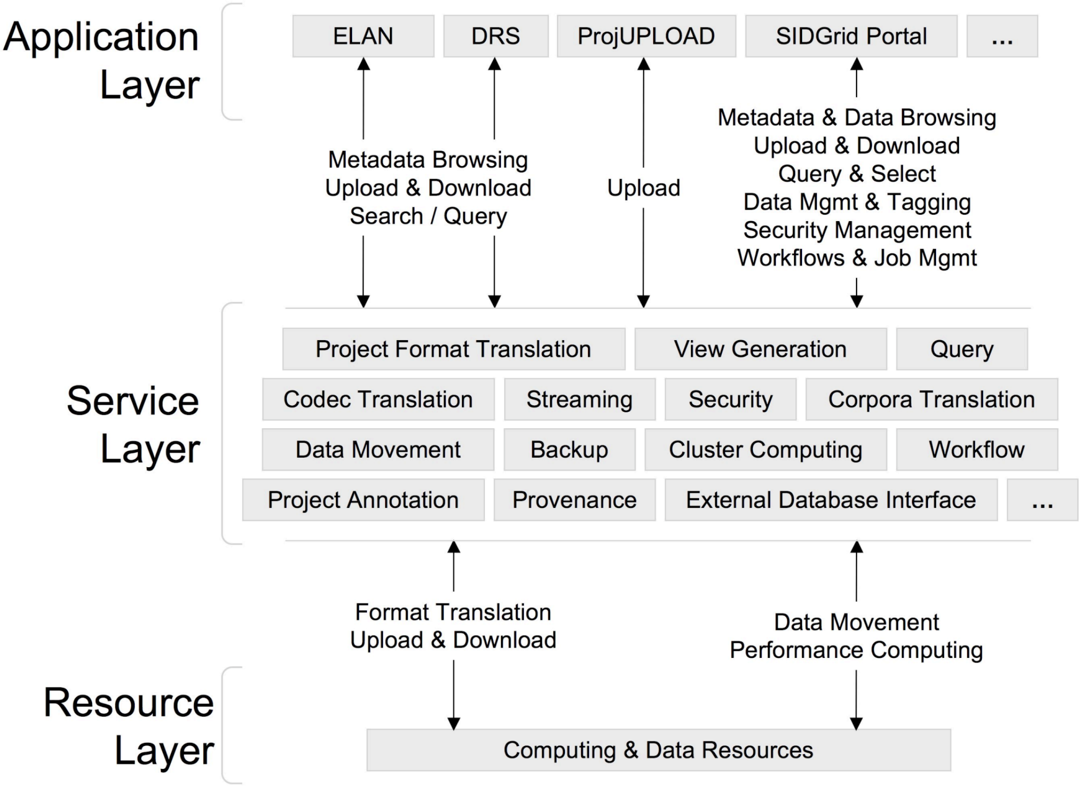


Figure 1. The SIDGrid architecture designed to incorporate new and existing bodies of data, enable analysis using key legacy tools, and provide a Grid services interface for flexible access by the distributed community of users. Component blocks in this schematic exchange data vertically through layer boundaries.

Another compelling aspect of the SIDGrid architecture is the possibility of advanced query against the repository, particularly query and exploration services that utilize semantic hierarchies. In recent years the power of data mining has been demonstrated using query and analysis tools that support discipline-specific concepts and abstractions. An example would be a semantic web query to a bioinformatics web site that explicitly uses tagged information describing genes, proteins, and biochemical pathways. We anticipate development of taxonomies and associated query and analysis services that include: i) physiological measures, such as heart rate, respiration, high-density EEG; ii) behavioral measures, such as eye gaze, posture, speech, and gesture; iii) single participants responding to visual or auditory events; and iv) multiple participants engaged in social interactions, such as casual conversation, problem solving, conflict resolution, meetings, tutoring. Query and analysis services would expose any of these layers and support integrated analysis of them.

The SIDGrid architecture shown in overview in Figure 1 is designed to provide a flexible and extensible testbed for research involving multimedia and multi-measure data. In the following sections we describe the architecture of the SIDGrid in terms of its component layers: resources, the data and computing assets available; services which expose the resources on the network; applications, the user software supported by SIDGrid services.

Resources

The resource layer includes facilities for data storage and large-scale computing. Because we employ grid technologies, these resources may be physically located anywhere.

Data in the form of audio, video, time series, annotations, and metadata form the core of the SIDGrid. Experiment data is stored in its original form, and in various other forms as required by higher-level services and applications. Access to externally available raw data is enabled through pointers to the data as part of the data model, reference to relevant access methods, and translation services provided by the database architecture.

The SIDGrid data model provides the lens through which raw data is seen by services and applications. Briefly, the model revolves around the core concepts of data streams and semi-structured data attached to the streams. The streams modeled include physiological data, audio, video, and environmental data. The semi-structured data attached to the streams are the building blocks for i) distributed annotations, ii) distributed user supplied tags to facilitate query and discovery, and iii) associations that can link any SIDGrid data resource or components or other resources. These may be added automatically by SIDGrid services or manually by researchers. With these constructs, data resources can be structured to express any and all of these: time aligned streams, system assigned metadata, distributed user supplied tags, distributed annotations, associations, events, derived features, and security metadata.

Complementing data resources described by a rich data model, SIDGrid provides access to compute resources for user-driven data analyses and for internal data translations. In addition to a data server and modest compute server to support the need for data archiving and processing, some data analysis applications will require access to more computing power. Leveraging efforts underway, we are working to deploy the SIDGrid computationally intensive services as part of a Science Gateway (Wilkins-Diehr, 2007). Users are able to access these resources via the Web using the SIDGrid Portal. The Gateway will enable large computing jobs to be constructed, submitted, and monitored with results deposited directly into the SIDGrid data warehouse. This greatly extends the class of automated analysis experiments that can be conducted and enables us to explore the feasibility of near-real-time analysis of experimental data.

Services

The service layer provides the interface to SIDGrid raw data. This layer of the data warehouse enables user applications to browse the library, upload and download experimental data and metadata, create and manipulate metadata, and play back multimedia experimental sequences. The interfaces provided are compatible with applications typically found on desktop computing platforms, for easy integration into desktop applications and web browsers.

The core of these services is based on SOAP, an XML-based standard for web services. Descriptions of several of the SIDGrid services – query, media streaming, format translation, security, and grid compute – are given below.

Query. A sophisticated search system allows users to easily locate projects of interest among their data in SIDGrid. These facilities allow users to find projects based on particular identifiers such as filenames, project name, or project size. Equally accessible are searches based on the project content, such as linguistic or analytic annotations in projects. Searches of these types have become vital as the amount of data in SIDGrid has grown, and are particularly important when a user explores a dataset with which he is unfamiliar. Performing analyses on such selection sets enables a new level of exploration not available previously.

Streaming. By centralizing their data, researchers have more flexible access to it independent of which office or computer they are using. The disadvantage of this is that they would have to download an entire experiment to view its contents. SIDGrid uses a streaming server to let users preview their experiment data without having to download the entire experiment, preserving their lightweight access to data.

Translation. Three translation services are currently in operation:

- Accommodating the various compute platforms in use by SIDGrid users, and the associated variation in data formats required by these platforms, SIDGrid transforms the original media formats in the data warehouse to formats appropriate for each of the typical compute platforms, Windows, OSX, and Linux. This allows users to seamlessly collaborate on projects without concern for variations in their operating systems.

- Applications not specifically designed to work with SIDGrid will typically require some filtering to support upload of application-specific data formats to SIDGrid, and download of SIDGrid-specific data formats back to the application. SIDGrid performs these translations on the fly as needed.

- In integrating with external data repositories, adaptation of the data formats and access methods may be required before the data can be made available to SIDGrid users. This translation is done dynamically when the external data is accessed, to sustain SIDGrid's ability to offer access to the data, while the external data resides elsewhere and evolves independent of SIDGrid.

Security. Underlying these services is a group-based security model, which enables researchers to share their SIDGrid-resident projects, while preventing access by users outside the collaboration. Projects are owned by individual users, and can be shared with other users within the groups they control and specify. Access to the SIDGrid portal is dependent on authentication using standard password mechanisms. Network connections between client applications and the SIDGrid services are secured with SSL as a further privacy measure.

Grid Services. SIDGrid utilizes several Grid services in support of cluster computing and distribute storage. GridFTP services are used for high-performance Grid-based data transfer among data resources in support of compute jobs. These services are key in supporting the movement of the often large data sets found in SIDGrid workflows. Grid compute services are used in submission and monitoring of jobs to the high-performance compute clusters that backend the SIDGrid. These services are brought together as components in users' analysis workflows by the Virtual Data System (VDS). Grid Services are covered in more detail below in the section GRID computing with SIDGrid.

Applications

As noted in the discussion of the service layer, application programs interact with SIDGrid using a web services interface. The diagram in Figure 1 shows several applications and the functionality afforded by SIDGrid in each case.

SIDGrid Portal General access for browsing, preview, annotation, data upload, project management, time synchronization management, and grid workflow construction are provided in a user interface module. The portal requires no software to be installed on the user's workstation aside from a standard web browser such as Safari, Firefox, or Internet Explorer.

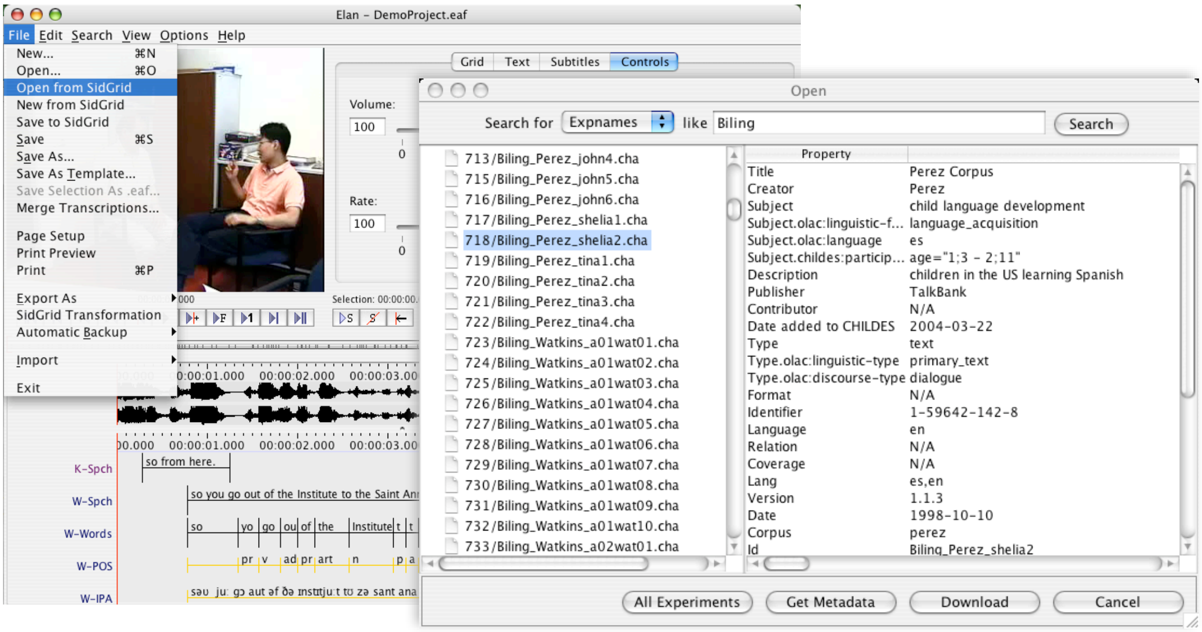
SIDGrid Applications Native applications interface to SIDGrid using methods and functions in the access layer. Interesting examples of applications include: automatic processes to analyze experimental data and data mining interfaces. An example of a SIDGrid application is the platform independent standalone upload tool.

Legacy Applications Existing applications (ELAN[2], DRS, MacVisSTA, Praat, and many others) could interact with the SIDGrid through a custom interface designed to translate protocols and data formats as necessary. In this way, existing analytical tools are thereby enabled to analyze SIDGrid data resources and to add metadata to the repository. Figure 2 shows an example of SIDGrid function added to an existing application. The ELAN window shows a project with synchronized video, audio, and annotation tracks. The user has selected Open from SIDGrid on the File menu, which opens the dialog in the foreground on the right. The search bar at the top enables selection of data conforming to query criteria with results in the left column. Metadata for the highlighted item is shown in the right column.

Grid Computing with SIDGrid

SIDGrid is a science gateway to the TeraGrid, which allows researchers to access the powerful computational resources of a large Grid cluster from within the user-friendly portal environment. Thus, users can not only search, retrieve, and store data but they can also offload compute-intensive jobs from their own labs and local machines onto the TeraGrid. SIDGrid provides batch submission and parallelization of jobs which can greatly reduce the time needed for data processing. Once this remote computation is complete, the resulting dataset is transferred back to SIDGrid where it can be tagged, queried and shared easily with other research groups. Therefore, a researcher can run either a new script or analysis on data he has just uploaded or on the output of a previous run. SIDGrid implicitly addresses the issue of standardization within the social sciences by providing a framework whereby researchers, labs, and entire communities can easily replicate one another's findings by

[2] ELAN Website – www.let.kun.nl/sign-lang/echo/ELAN/ELAN_intro.html



Search for Expnames like Biling

713/Biling_Perez_john4.cha
715/Biling_Perez_john5.cha
716/Biling_Perez_john6.cha
717/Biling_Perez_shelia1.cha
718/Biling_Perez_shelia2.cha
719/Biling_Perez_tina1.cha
720/Biling_Perez_tina2.cha
721/Biling_Perez_tina3.cha
722/Biling_Perez_tina4.cha
723/Biling_Watkins_a01wat01.cha
724/Biling_Watkins_a01wat02.cha
725/Biling_Watkins_a01wat03.cha
726/Biling_Watkins_a01wat04.cha
727/Biling_Watkins_a01wat05.cha
728/Biling_Watkins_a01wat06.cha
729/Biling_Watkins_a01wat07.cha
730/Biling_Watkins_a01wat08.cha
731/Biling_Watkins_a01wat09.cha
732/Biling_Watkins_a01wat10.cha
733/Biling_Watkins_a02wat01.cha

Property	
Title	Perez Corpus
Creator	Perez
Subject	child language development
Subject.olac:linguistic-f...	language_acquisition
Subject.olac:language	es
Subject.childes:particip...	age="1;3 - 2;11"
Description	children in the US learning Spanish
Publisher	TalkBank
Contributor	N/A
Date added to CHILDES	2004-03-22
Type	text
Type.olac:linguistic-type	primary_text
Type.olac:discourse-type	dialogue
Format	N/A
Identifier	1-59642-142-8
Language	en
Relation	N/A
Coverage	N/A
Lang	es,en
Version	1.1.3
Date	1998-10-10
Corpus	perez
Id	Biling_Perez_shelia2

so from here.

so you go out of the Institute to the Saint Ann

Figure 2. The *Open from SIDGrid* dialog added to ELAN enables data and metadata browsing, query and search, and selection for download to the local workstation.

accessing shared datasets along with the analysis tools and scripts that were used to produce them.

SIDGrid interfaces with the TeraGrid in the following way:

- User selects the input files he wishes to run a transformation on and the script to process them. Currently, researchers are running scripts in R, MATLAB, AFNI, FreeSurfer, Praat, Python and Perl.

- These input files are then registered with the Globus Replica Location Service (RLS) running on SIDGrid which keeps track of where replicas exist on physical storage systems. The RLS can then be queried when file transfers are staged. Transfer is done using GridFTP and is generally the first job in a user's workflow.

- The Virtual Data System (VDS) is used to manage the components of the workflow. A workflow is made up of a series of transformations. For example, a user may have a script written in R that he wishes to run on 2000 files. The R script itself is known as a transformation and each call to it is considered a 'derivation'. More generally, each derivation can be thought of as a job and all of the jobs together constitute a workflow. If these jobs are not dependent upon one another's output, they will automatically be run in parallel. VDS has a set of Java classes for representing these jobs and their dependencies in XML format. Also, all of the transformations are added to the Virtual Data Catalog (a Postgres database), so that the provenance, or history, of a workflow is persistent. Future plans involve making this provenance data available to the user so that he can query data derivation.

- At this point the workflow (in XML) is still considered abstract until each job is assigned to be run on a particular Grid site. This is done using a planner called Pegasus. Pegasus checks for file dependencies and queries the RLS for needed files and verifies the existence of necessary resources on the remote sites. This results in the concrete workflow submit file being generated. This file is known as a DAX (for directed acyclic graph, or DAG, in XML).

- Condor is the batch system used by SIDGrid that provides job queuing and resource management. The workflow is submitted to Condor using the DAX and Condor manages the correct submission of jobs within the workflow to the remote Grid sites.

- The final job in each SIDGrid workflow is the upload of results from the remote Grid site back to SIDGrid as a new project.

Community

We are motivated by a number of factors to cultivate working relationships with projects that might have scientific or technical goals aligned with ours. Of prime importance is to ensure that the cyberinfrastructure we develop is relevant to the scientific problems. Consequently, we seek to add analytical capability to research programs using existing corpora. We are also aware of the range of issues that attend standardization of metadata, formats, and methodologies. To guide our development and selection of priorities we have been diligent about developing and maintaining contacts with a wide variety of research programs in the fields touched by SIDGrid. For the duration of our prototyping phase we have maintained a series of workshops with the goals of informing potential collaborators and users of our progress, hearing about practices and priorities in other research programs, and soliciting guidance. From this effort, we have engaged several groups directly in ongoing collaborations:

- We are importing the data in the CHILDES[3] and TalkBank[4] datasets into the SIDGrid environment to enable increased searching capabilities as well as integration with the SIDGrid Grid services. TalkBank is an international multimedia database of spoken language interactions in areas ranging from child development to classroom discourse to language disorders. Nearly 3,000 journal articles and books based on these databases have been published. Integration with SIDGrid will improve access to this data and will enable qualitatively and quantitatively expanded analyses.

- We are working with the University of Chicago's Human Neuroscience Laboratory5 to Grid-enable analysis of their fMRI data. The group routinely launches Grid jobs from the SIDGrid portal onto the TeraGrid environment and has significantly increased the amount of data they can process.

- Working with Professor Gina-Anne Levow of the University of Chicago's Department of Computer Science, the SIDGrid environment is being used for the storage and management of data collections. These data are analyzed on Grid computational resources to understand the role of prosody, from the lexical level to the pragmatic in the structuring of discourse and dialogue.

- Researchers at the National Centre for e-Social Sciences at the University of Nottingham, UK, are studying the embodied nature of language while building the Nottingham Multimodal Corpus (NMMC). We are collaborating with them on a project to develop

[3] CHILDES Website – childes.psy.cmu.edu

[4] Talkbank Website – talkbank.org

[5] Human Neuroscience Laboratory Website – www.fmri.uchicago.edu

and understand integrated methods that combine their analysis and computational components with the SIDGrid cyberinfrastructure (Adolphs et al. 2007).

Conclusions

In this paper we have presented a new infrastructure that has been designed to transform how social and behavioral scientists collect, annotate, collaborate, share, and analyze data. We have leveraged tools and data already in use in the community and shown how when combined with the SIDGrid environment they enable new capabilities such as advanced search and data management. We described SIDGrid as a service oriented architecture, one that exposes a growing collection of interlinked services —including streaming, securing, transcoding, and transforming data – via standards-based Web and Grid interfaces. We showed how the environment enables access to advanced computing resources while hiding a tremendous amount of complexity from the user. We have included segments of the community in our development process through a series of workshops that not only report progress of the SIDGrid effort but also gather issues and requirements to be fed into the development process. SIDGrid provides a rich data environment that is capable of capturing the notion of time, data streams, and semi-structured data attached to these streams in order to enable powerful manipulations of multimodal data.

Acknowledgments

This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-ENG-38, and in part by NSF under Grant No. BCS-05-37849.

References

Abraham, F.D., Abraham, R.H., and Shaw, C.D. (1992): ‘A Visual Introduction To Dynamical Systems Theory For Psychology’, *Basic Approaches to General Systems, Dynamic Systems, and Cybernetics*, Aerial Press.

Adolphs, S., Bertenthal, B., Boker, S., Carter, R., Greenhalgh, C., Hereld, M., Kenny, S., Levow, G., Papka, M. E., and Pridmore, T. (2007): “Integrating Cyberinfrastructure into Existing e-Social Science Research,” *Proceedings of the e-Social Science 2007 Conference*.

Wilkins-Diehr, N. (2007): ‘Science Gateways – Common Community Interfaces to Grid Resources’, *Concurrency and Computation: Practice and Experience*, 19(6), John Wiley & Sons, pp. 743 – 749.