Social Informatics Data Grid

Bennett Bertenthal[1, 5], Robert Grossman[3], David Hanley[3], Mark Hereld[1, 6], Sarah Kenny[1], Gina Levow[2], Michael E. Papka[1, 2, 6], Steve Porges[4], Kavithaa Rajavenkateshwaran[1], Rick Stevens[1, 2, 6], Thomas Uram[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 Psychology, University of Illinois at Chicago
[5]Department of Psychology, Indiana University
[6]Mathematics and Computer Science Division, Argonne National Laboratory

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 [1], it has yet to have a major impact in the psychological sciences. 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) [2] 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.

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.

- Raw data in the form of the many predicted and future types as well as references to data from external sources will be warehoused locally (Raw Data; External Public Databases).

- Access to externally available raw data will be 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 (Data Model; Translation Services; External Public Databases).
- The data model supports inclusion of standard data types and data descriptions. An implementation of time trees will enable management of detailed hierarchical time synchronization parameters based on data type, measurement circumstances, stream specifics, and quantitative comparison of stream-to-stream features (Data Model).
- Interface to SIDGrid raw data is provided by the SIDGrid Data Resources layer and supplementary translation services (SIDGrid Data Resources; Translation Services).
- The access layer of the data warehouse will enable user applications to browse the library, upload and download experimental data and metadata, create and manipulate metadata, and play back multimedia experimental sequences. It will implement interfaces for web and grid services as well as more primitive query and indexing services (SIDGrid Services).

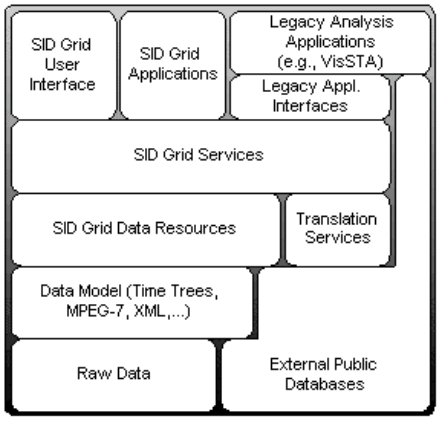

Figure 1 - The SIDGrid architecture designed to incorporate new and existing bodies of data, enable analysis using key legacy tools, and provide a Grid services interface for flexible access by the distributed community of users. Component blocks in this schematic exchange data vertically through layer

- General access for browsing, data upload, and time synchronization management will be provided in a user interface module to be developed for this proposal (SIDGrid User Interface).
- Native applications will interface using methods and functions in the access layer. Interesting examples of applications include: automatic processes to analyze experimental data and data mining interfaces (SIDGrid Applications; SIDGrid Services).
- Existing (i.e. non-native) applications (e.g. VisSTA) would interact with the SIDGrid through a custom interface designed to translate protocols and data formats as necessary. In this way, existing analytical tools will thereby be enabled to analyze SIDGrid data resources and to add metadata to the repository (Legacy Application Interfaces; Legacy Analysis Applications).

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 could then be transformed into formats 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 will enable 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 and layers of the SIDGrid integrate 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 will be sampled 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.

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 propose to develop 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.

Three projects currently working with SIDGrid are:

- We are importing the TalkBank database [3] 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 the University of Chicago's Human Neuroscience Laboratory [4] to grid enable analysis of their fMRI data. The group routinely launches Grid jobs from the SIDGrid portal onto the TeraGrid environment and has increased the amount of data they can process.
- Working with Professor Gina 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.

In this abstract we have presented motivation for the SIDGrid effort, an overview of the architecture, and a brief description of the projects that are currently using the infrastructure. The full paper will complete the description of the SIDGrid architecture, description of progress and state of the system. The paper will end with a full list of projects currently using the SIDGrid environment with an expanded description of a few of the projects.

Acknowledgements: 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:
[1] Abraham, F.D., R.H. Abraham, and C.D. Shaw, *A visual introduction to dynamical systems theory for psychology*. 1992, Santa Cruz: Aerial Press.
[2] SIDGrid Website – sidgrid.ci.uchicago.edu
[3] Talkbank Website – talkbank.org
[4] Human Neuroscience Laboratory Website - www.fmri.uchicago.edu