Science of Design for Information Systems

Report of the NSF Workshop, Seattle, 2003[1]

ABSTRACT

The Workshop on Science of Design for Information Systems (SDIS2003) was held in Seattle, September 16 and 17, 2003. It was funded through a grant from the National Science Foundation[2], with the goal of assessing the state-of-the-art in information systems design and suggesting promising directions for future research and development in this critical area.

This short report is intended to provide an overview of the workshop report for the SIGMOD audience. In summary, we believe that there is a need to develop a new set of methodologies for information system design that cover advanced aspects of such systems. In particular, we are interested in techniques that offer guidance in the design of information systems that integrate data from multiple sources, handle dynamic aspects of the system (e.g., rapidly changing data, tracking provenance, version management), include aspects influenced by data location (e.g., cached objects and queries, peer-to-peer data sharing), model process-oriented issues (e.g., workflows, web-services), and account for the security and privacy of the data.

The interested reader can find a full version of this report at the workshop website, www.cs.wisc.edu/sdis03.

[1] The workshop was organized by R. Ramakrishnan (U. Wisconsin), Phil Bernstein (Microsoft), and Alon Halevy (U. Washington), and the participants were A. Borgida (Rutgers), M. Carey (BEA), S. Chaudhuri (Microsoft), P. Chrysanthis (U. Pittsburgh), C. Clifton (Purdue), U. Dayal (HP), D. Florescu (BEA), M. Franklin (UC-Berkeley), L. Gasser (UIUC), J. Gehrke (Cornell), S. Ghandeharizadeh (USC), H.V. Jagadish (U. Michigan)), H. Korth (Lehigh), A. Rosenthal (Mitre), S. Sripada (Boeing), J. Srivastava (U. Minnesota), J. Su (UCSB), D. Suciu (U. Washington), S. Thatte (Microsoft), B. Thuraisingham (NSF), V. Tsotras (UCR), J. Widom (Stanford), M. Winslett (UIUC), O. Wolfson (UIC), J. Yuan (Boeing), and M. Zemankova (NSF). This report was authored jointly by the above group; please address correspondence to Ramakrishnan (raghu@cs.wisc.edu).

[2] This workshop was supported by the National Science Foundation's Intelligent Data Management program under the grant IIS-0348699. All opinions, findings, conclusions and recommendations in any material resulting from this workshop are those of the workshop participants and do not necessarily reflect the views of the National Science Foundation.

1. Background and Overview

The Workshop on Science of Design for Information Systems (SDIS2003) was held in Seattle, September 16 and 17, 2003. It was funded through a grant from the National Science Foundation, with the goal of assessing the state-of-the-art in information systems design and suggesting promising directions for future research and development in this critical area.

This report contains two parts. The first, and main, part is a discussion of the research challenges in developing a rigorous science of information systems design and translating it to practice. Specifically, it seeks to identify existing areas of strength and weakness in the theory and practice of database and information systems design, place these within the broader context of software system design, and identify promising directions for future study. This part of the report is divided into an introduction (Section 2) and five subparts (Sections 3 through 7), corresponding to the five focus groups at the Workshop.

The second part of the report (Section 8) makes recommendations of a programmatic nature and covers the following:

1) The skill set and experience that must be brought together to advance the science of information systems design. Specifically, identify disciplines and groups within academia and industry that must work closely to achieve the desired advances.

2) Activities that could facilitate progress in this area. Specifically, what kinds of funding modalities are required, such as single-investigator or small-team grants, larger collaborative efforts, university-industry partnerships, long-term focus groups and studies of organizational behavior, and so on.

2. Introduction

The role of data, and the importance of technologies that support data access, maintenance, and evolution, is widely recognized as a central aspect of complex software systems. Understanding how to design the data-oriented aspects of an application is therefore an essential step in developing a science of software systems design.

Information Systems Design has traditionally encompassed methodologies for modeling the data underlying an enterprise and the relationships between data elements; the design of structures to store data; the design of a conceptual view of data as seen by other program components, which may differ from how the data is actually stored; and the design of efficient and robust techniques for accessing and modifying data, especially in the presence of data copies, distribution, and concurrent access. In recent years, there has been increasing interest in how to design systems that access and integrate multiple sources of data; systems that deal with complex data such as multimedia and text; applications involving spatiotemporal information and mobility; applications involving large-scale replication and distribution; applications involving complex workflows and notifications; and so on.

2.1 Focus Areas

To reflect the trends mentioned above, the Workshop was organized using the following five focus areas:

A. Information Systems Design for Integration. This includes semantic integration of legacy enterprise systems and processes, especially in the context of deploying complex, yet well-structured new applications, e.g., SAP or new infrastructure for scientific collaboration; and semantic integration in the context of novel settings such as Internet-scale integration and data-sharing data sharing in the bioinformatics domain. Many information systems today are built using off-the-shelf applications that are then customized and integrated with a user's existing systems, and we need better tools and design methodologies to support this process**.** This implies the need for better mechanisms to manage models of applications and data, and mappings between them.

B. Design for Dynamic Environments. The focus is on design issues arising from rapidly changing data, either because of continuous data monitoring or acquisition (e.g., sensor streams, data warehouses), or the need to incorporate information from ongoing analysis, such as data mining results. How can we update the knowledge base while still permitting reasoning with the old "state-of-the-world"? Specific issues here include tracking provenance of data objects, processes, schedules, etc.; conflict-resolution; version management; data aging and archiving; and impact of real-time and continuous processing requirements on new data.

C. Design Issues Based on Data Location. This topic covers design of caching and replication in middle-tier software systems; integration of highly distributed data collections such as web data and peer-to-peer systems; and dealing with mobile environments and objects, including monitoring, tracking, and reactive systems. The focus is on design considerations based on, or strongly influenced by, the *location* of data objects (or copies) and queries, and related issues such as distributed ownership.

D. Process-Centric Design. Examples of topics here are workflow and web services. The key issue is how to extend the scope of design beyond the traditional focus on data, to encompass patterns of *activity*, such as business rules and application-level interactions that govern how the data is generated and utilized.

E. Designing for Security and Privacy. Security and privacy are increasingly important for data and information. Security and privacy design issues include the specification, analysis, and implementation of security- and privacy-related policies. Perhaps the most critical challenges are how to scale to the enterprise level, how to support inter-organizational collaborations, and how to coordinate multiple types of policies during the design process. Other challenges include design-time and run-time support for fine-granularity access control, individual privacy preferences, and enforcement approaches that are verifiably robust against relevant threats.

2.2 A Framework for Discussing Information Systems Design

Information Systems are broader in scope than Database Systems. They also include vertical applications that make heavy use of a database (such as enterprise resource planning (ERP) and customer relationship management (CRM)), document management systems, web-based repositories and search engines, and workflow systems that capture and support data-oriented business processes. While the term "database design" has traditionally denoted design and tuning of database schemas, including normalization and performance-oriented tuning, there are many other aspects of an information system to be considered in a design.

Design of a complex system is an ongoing exercise in finding the right configuration in a high-dimensional space of design choices. Each focus group was asked to identify the main “axes” that characterize design choices from the perspective of issues central to that group. The report reflects this, with a subsection identifying the space of design choices from each group’s perspective; this section includes a discussion of concepts, design criteria, and design principles especially relevant to that group’s concerns.

We use the term *design aid* broadly to denote design models and theories, design methodologies, and tools that build upon them. Design aids facilitate exploration of the space of design choices in a principled manner. Such aids can help us to understand the space of choices and how different choices interact, enable “what if” analyses to understand the consequences of given choices, translate design choices into implementations, validate a design with respect to requirements, and validate an implementation with respect to a design. They are especially valuable in dynamic environments where the system must constantly react to changes in the environment, or be re-configured frequently to respond to changes in the objectives of the enterprise.

Research on the Science of Information Systems Design should support activities that lead to the development of a comprehensive body of design aids that quantifiably advance our ability to design and manage complex Information Systems. With a careful focus on a range of motivating applications to ground the research, we believe it is possible to investigate techniques that will significantly advance the state-of-the-art in Information Systems design. We observe that developing more flexible technologies for underlying system components can enable new and better design configurations, and that it complements research on how to better utilize existing system components to meet design objectives.

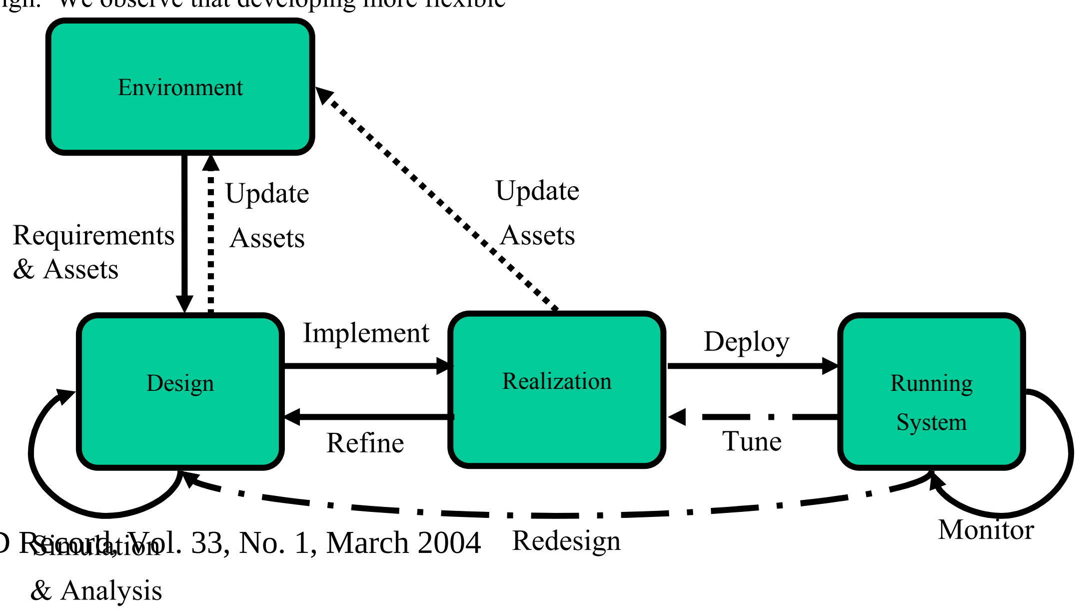
The process of design has some fundamental characteristics that cut across the different focus areas, in fact, across the design of all aspects of a complex system, going beyond the Information Systems aspects alone, as illustrated in the following simplified process diagram shown below.

Each group was therefore asked to consider the following overarching themes:

I. Change patterns in user/system requirements. Beyond the initial system design, we must address *system evolution* in response to changes in the individual or organization that the information system serves, including process changes and changes in the nature and volume of data and its access patterns. Specific issues here include archiving, schema evolution, tuning challenges, object-oriented design and software reuse, and simulation of legacy software and business logic.

II. Information Systems design within the broader context of software design. Many of the issues touched upon here are not unique to information systems, but apply to software in general. We need to understand the issues in this broader context, and gain insight into novel additional aspects introduced by our focus on the information dimension.

Each group wrote a section of the report that summarized their discussions, including illustrative problem scenarios. These sections are omitted from this summary version of the report, and the reader is directed to the Workshop website, www.cs.wisc.edu/sdis03, for the full version of



the report.

3. Recommendations

This section of the report makes recommendations of a programmatic nature, based on discussions at the Workshop, and covers the following issues:

- The skill set and experience that must be brought together to advance the science of information systems design. Specifically, identify disciplines and groups within academia and industry that must work closely to achieve the desired advances.
- Research activities that could facilitate progress in this area. Specifically, what kinds of funding modalities are required, such as single-investigator or small-team grants, larger collaborative efforts, university-industry partnerships, long-term focus groups and studies of organizational behavior, and so on.

3.1 Related Groups

Design of Information Systems is inter-disciplinary by nature, and advances will require the combined efforts of many disciplines within Computer Science. The field of software engineering is addressing the design of software systems in general, and while Information Systems design offers some unique challenges and opportunities because of its data-centric emphasis, there are close ties between these fields. It is important to recognize the commonality in goals, and to leverage relevant techniques, such as component-based development and UML modeling, whenever appropriate. Similarly, it is worth exploring the broader applicability of declarative database paradigms (e.g., data models and mappings), which focus the design of Information Systems, to software systems design.

Beyond Databases and Information Systems, and Software Engineering, other areas of Computer Science that are relevant to Information Systems design include:

- Artificial Intelligence (machine learning, knowledge representation)
- Programming Languages
- Operating Systems
- Networking
- Security
- Workflow

There are many opportunities to apply Artificial Intelligence techniques to improve or adapt the design of an information system. The Programming Languages and Workflow communities have much to contribute in the areas of process design and realization. Networks and Operating Systems are an essential part of the infrastructure underpinning any software system, and as such, they have a great influence on the design of such systems. For example, any consideration of security aspects of an Information System must address network and operating system vulnerabilities.

The landscape for software systems is changing rapidly, influenced by the proliferation of the Internet and a trend towards service-oriented architectures that re-use and inter-connect existing systems. This change is reflected—and accelerated—by emerging standards for data exchange, resource discovery, and remote, autonomous process interactions. In order to meet the short-term demands of information systems design in this new landscape, and to lay the foundations for a sound and comprehensive science of Information Systems design for the future, we need contributions from a wide range of groups:

- Research groups in academia and industrial laboratories, who can address foundational issues with a long-term horizon.
- Application developers in academia (e.g., scientists developing applications using information systems) and industry, who identify the key bottlenecks and who use and validate proposed design methodologies and design aids.

3.2 Research Activities to Advance SDIS

We strongly recommend funding the development of a comprehensive and well-founded Science of Information Systems Design through a variety of initiatives. The funding objectives should be to identify the key bottlenecks in the current state of the art and to support research into the development of design methodologies and design aids that can eliminate or mitigate these bottlenecks. There should be specific initiatives that foster close collaboration across different Computer Science disciplines, and that bring together teams of developers and researchers from academia and industry.

We suggest the following specific kinds of funding initiatives:

- Cross-disciplinary workshops with a design focus.
- Single-investigator research grants.
- Larger grants that fund research involving multiple PIs, especially teams that are inter-disciplinary.
- Larger grants that fund research with an "end-to-end" focus that covers development of concepts and testing in real applications.
- Grants that leverage industrial funding and interest in design-related research.

4. Summary

Design of complex information systems raises many unresolved issues, and progress in understanding how to build robust, reliable, secure software systems that leverage and protect large information repositories is a central challenge for the field of Computer Science. We think a significant investment by NSF and other federal agencies is called for, and will be amply justified by the practical impact of the results. The interested reader can find a full version of this report at the workshop website, www.cs.wisc.edu/sdis03.