Digital Library Support for Public Health Simulation Infrastructures

Jonathan Leidig
Digital Library Research Laboratory, Dept. of Computer Science
Network Dynamics and Simulation Science Laboratory, VBI
Virginia Tech, USA
leidig@vt.edu

ABSTRACT

Epidemiological simulation environments produce extremely large quantities of scientific, numeric digital objects. Traditional digital libraries and services fail to effectively support simulation-based tasks. The requirements for this class of digital libraries can be described through formal definitions of content, users, and services. The set of definitions provide a basis for developing a simulation supporting digital library. Using formal definitions, services may be described, implemented, and reused. User studies are used to quantify the efficiency gains in DL generation, usability of services to complete tasks, and support for designing and conducting simulation experiments.

Keywords

Simulation Infrastructures, Domain-Specific Services, Research Communities

1. INTRODUCTION

Modeling and simulation research groups generate extremely large amounts of scientific content. Computational epidemiology is a simulation domain that studies the health of populations. With parallel simulation models and high performance computing (HPC) resources, epidemiological researchers produce large amounts of data, often without appropriate technologies to cohesively archive, manage, search, and disseminate scientific content. Data management is needed for public health research, medical and census records, synthetic populations, graphs and networks, and content from a simulation workflow. Common stages of a simulation workflow produce collections of model ontologies, input configurations, result datasets, result summaries, analyses, documentation, annotations, and publications. Traditional digital libraries and services are not well-suited to handle primarily numeric content. For example, indexing and retrieval algorithms for scientific content cannot rely on previous work in developing full-text indexing services that include term frequencies, thesauri, dictionaries, or partial-matching scores. Metadata generation and data mining applications also do not take advantage of the structured nature of simulation-based content (e.g., known, well-defined input and output formats).

Three computational epidemiology applications motivate our efforts in simulation content management; EpiSimdemics [1], EpiFast [2], and OpenMalaria [17]. This class of epidemiological models require large social networks (e.g., human populations at the continental level), HPC resources (e.g., grids and clouds), and large storage systems to store rapidly generated content. Petascale storage systems are in place for a variety of domains including the internet archive's PetaBox storage system, scientific research centers such as CERN, and national laboratories. As petascale simulation applications become widespread, large-scale content management and retrieval requirements are being tackled by research groups without prior experiences or expertise in designing and implementing these types of systems.

A *Sim*ulation *D*igital *L*ibrary (SimDL) is needed by simulation groups to manage the entire staged set of structured simulation-related content. Digital libraries in this context must archive previous results, enable collaboration at experiment stages, provide provenance for content, and provide a mechanism for discovery and data mining. Our goal in developing SimDL is to a generate a formal, reusable, and component-based digital library framework to be deployed by simulation groups. The focus of this work is on automatically managing large amounts of scientific content while enabling discovery and retrieval. This focus is motivated by the challenge of managing large-scale, rapidly-produced simulation content across multiple epidemiological domains. *What is needed is a DL that communicates with cyberinfrastructure components, manages simulation based content, and supports simulation research tasks.*

2. RESEARCH QUESTIONS

This work attempts to answer the following questions. What constitutes scientific and simulation content? What tasks for scientific content, simulation systems, and cyberinfrastructures can a DL support? What simulation-specific service implementations are required by researchers? How are these services formalized, discovered, and used by DL designers? How does a simulation-based DL perform in comparison with a traditional DL?

The process for answering these questions consists of for-

mally proving the following statements in the context of simulation-supporting digital libraries.

Statement R1: *Formal descriptions of simulation-based content, services, and users exist and can fully characterize this class of DLs.*

Statement R2: *Formal descriptions of DL components and functionality can be leveraged to produce and deploy DL instances with the stated capabilities.*

Statement R3: *Simulation-supporting DLs can provide interoperability through simulation-specific services for tasks such as UI generation, infrastructure functionality, component communication, and managing experiments.*

Statement R4: *Efficient service implementations can be specialized for simulation content (e.g., graph-based indexing in a cloud service).*

Statement R5: *Case studies and user studies effectively evaluate SimDL functionality and deployment processes.*

3. BACKGROUND AND RELATED WORK

Efforts to enhance scientific data management practices have been a major focus for eScience and cyberinfrastructure groups over the last several years. Several examples of scientific data management include earthquake simulation repositories [8], embedded sensor network DLs [3], community earth systems [6], D4Science II [11], mathematical-based retrieval [18], chemistry systems [12], national research data plans [9], and science portals [15].

Numerous workflow management systems exist but are not tailored specifically for simulation workflows. Standard workflow technologies include business process model and notation, Kepler [13], Taverna [16], Triana [14], and Pegasus [5]. Computational epidemiology workflows consist of model design and software implementation by model developers as well as study design, input configuration design, simulation execution, result summarization, analysis execution, analyses gathering, publication, and policy decision making by public health researchers. Each portion of the workflow may be defined in an ontology that is used to define a minimal set of SimDL services.

Currently, there is a significant lack of deployable digital library options for simulation research institutions. Existing digital libraries lack a means of communicating specifically with cyberinfrastructure components, provisioning numeric-based services, automatically constructing content metadata records, supporting simulation-based tasks, and allowing federation across simulation systems, models, and model versions.

4. SIMDL

SimDL, once completed, will provide a digital library application with novel services to directly support the epidemiology community. The process for constructing SimDL includes formally defining content, services, users, and user tasks; identifying the minimal set of simulation-specific services; implementing simulation-specific services; defining a framework for composing generic and simulation specific services; developing multiple instantiations of the framework for a cyberinfrastructure and local institutional infrastructure; and evaluating the effectiveness of the set of functionality provided by the framework. A formalized description of SimDL and simulation content is presented in [10]. Formal definitions of simulation-specific services have been identified, implemented, and presented in forthcoming publications. These service definitions describe the minimal set of user tasks required by simulation-supporting DLs. The definitions are used to describe interactions between users and the infrastructure as well as the pattern of tasks. The service components of SimDL are shown in Figure 1.

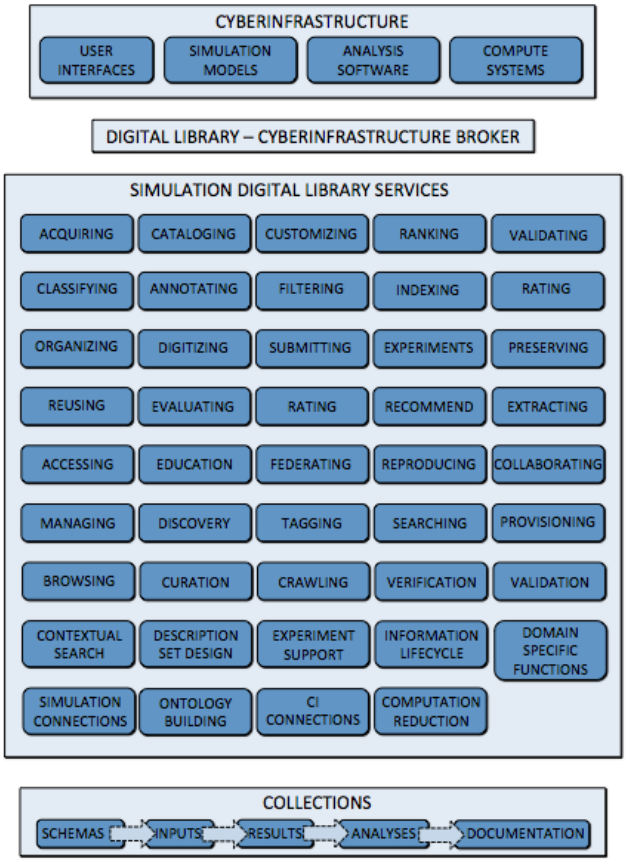

Figure 1: System components in SimDL.

Prototype installations of the SimDL framework are currently in development for (1) a cyberinfrastructure for network science applications including epidemiology, (2) two public health research institutions, and (3) immunopathology simulation models. Several case studies are used to guide the development and evaluation of SimDL. A specific cyberinfrastructure, CINET, and several local infrastructures at research institutions provide the initial set of environments for SimDL to support. These environments were selected as representative samples of possible use cases of SimDL. A brokering system is used for SimDL communication with other infrastructure components, see Figure 2. Evaluation of the sufficiency of formally-defined services, effectiveness at supporting user tasks, and comparison to traditional digital libraries remains an ongoing effort. All epidemiological

domain information is encapsulated in model version specific schemas and ontologies. SimDL may be extended into other domains through the simple provision of new domain ontologies. Future efforts will be aimed at evaluating the suitability of SimDL in other domains.

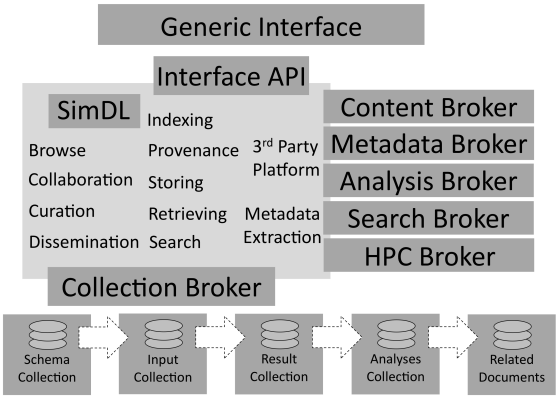

Figure 2: SimDL architecture and broker system.

Building an index structure for SimDL requires the federation of multiple domains, simulation models, types of content, and collections. An index graph structure has been selected to allow for generic indexing of documents and users, 'nodes,' and relationships of metadata terms, 'edges.' Cliques and motifs may be identified in the collection's index graph. Arbitrary overlay networks can also be created to build graphs to be used in specific services, e.g., a personalization service based on a graph of similar users and the documents they retrieve. This digital library is anticipated to support the minimal tasks required by the epidemiological research community.

5. METHODOLOGY AND EVALUATION

The five proposed research statements must be proven to evaluate the suitability of SimDL for scientific settings. Prototype designs, proofs, and implementations have been completed for R1-R3. Several simulation-specific services and the evaluation studies remain as current work.

5.1 R1. Formal Descriptions

Theoretical foundations for users, content, and services provide a means of explicitly defining digital libraries. With these definitions, services can be implemented and reused between DL efforts. The 5S framework [7] and the DELOS reference model [4] have been proposed for defining digital libraries. Here, the 5S framework notation is used to define SimDL while making use of previous definitions. Content including schemas, model-ontologies, input configurations, sub-configurations, results, analyses, and experiments have been formally defined for SimDL [10]. User roles were also defined for SimDL including tool builders, administrators, systems, study designers, analysts, annotators, and explorers. Formal definitions of simulation-specific services for browsing, collaboration, curation, dissemination, metadata extraction, provenance streams, and search have also been proposed in parallel work. Thus, full formal descriptions of users, content, and services have been produced.

5.2 R2. Implementations

Formal definitions of services, in 5S notation, include an informal description, required inputs, produced outputs, service pre conditions, and post conditions. The process for developing a service includes task identification, definition, and implementation. After service definition, developers have leeway to implement the service however necessary as long as the input format, output format, and conditions are met. Further work is required to determine if the current definition notation is sufficient for guiding the implementation of services.

5.3 R3. Interoperability

As shown by the multiple DL.org working groups, numerous types of interoperability exist. SimDL supports user, architecture, and functionality interoperability through simulation-specific services. SimDL provides an API to allow user interfaces to store and request content. A generic interface has been built using this API to display content related to individual epidemiological models. Thus, users have consistent interface presentations between models and systems. Architectural interoperability is provided through the brokering system communication with external components and infrastructures. With architectural interoperability, SimDL may be integrated into multiple simulation infrastructures that can provide adaptors to the brokering system. Functional interoperability is attempted through formal definitions of the inputs, outputs, pre conditions, and post conditions of services. With formal definitions of service inputs and outputs, it may be possible to form compositions of interoperable functions. However, formal definitions of services have not yet been proven to provide complete service interoperability. With sufficiently defined services, it may be possible to form service interoperability validation and registries.

5.4 R4. Simulation-Specific Services

Simulation-specific implementations of a set of services are required by SimDL. As an example, a graph-based indexing service can be defined and implemented to support a variety of other internal DL services. With a graph of content, metadata, and users, cuts can be identified in the graph to produce service-specific graphs with a pertinent subset of nodes and edges. In a rating service, a weighted graph can be produced consisting of user and content nodes where each weighted edge represents a numerical rating given by a user for a particular document. Another graph with only user nodes can be used to identify cliques of similar users. Combining these two graphs can produce a recommendation service where similar users are suggested highly-recommended documents from similar peers. Efforts are currently underway to implement a graph-based indexing service using scalable graph mining algorithms to be executed as a cloud environment DL service. The full set of simulation-specific service implementations will be integrated within SimDL with the complementary set of generic services.

5.5 R5. Evaluation

The functionality of SimDL services and the deployment process require evaluation to prove R1-R4. User and case

studies are needed to answer the following questions.

- How effectively can the research groups used in developing the prototypes and other epidemiology simulation groups make use of SimDL?
- How does SimDL and its services improve efficiency in comparison to existing systems?
- What are the performance gains with comparing designing and running experiments with and without SimDL services?
- For the educational community, how well can educators carry out educational activities with managed content and simulation connections?
- What is the reduction in effort when integrating SimDL within an infrastructure versus implementing customized, institute specific processes, e.g., saved lines of code and development time?
- Is there a standard methodology to running experiments in simulation groups (e.g., study design, results, analyses, data mining, and publication)? How does SimDL support this from end-to-end based on the effort and hours required to create a study and execute an experiment?
- What additional functions are provided through SimDL in comparison with existing workflow, digital library, and management systems?
- How suitable is SimDL for non-epidemiological domains?

6. CONTRIBUTIONS

The ability to successfully integrate SimDL prototypes into multiple infrastructures has implications for simulation and DL research groups. Epidemiology research groups can evaluate the suitability for installing SimDL or identify site-specific requirements needing SimDL extensions. The DL research community is provided with an example of an end-to-end definition-to-production DL as well as novel designs and implementations of scientific services (e.g., network graph indexes).

This research makes the following contributions to DL and simulation communities:

- formal 5S definitions for a class of scientific digital libraries;
- definitions and implementations of novel digital library services for simulation management;
- efficient indexing and storage of structured simulation content using graph algorithms and cloud resources;
- foundation for a simulation-services registry;
- deployable SimDL framework to support simulation infrastructures and workflows; and
- three fully-fledged, practical SimDL instances.

7. FUTURE WORK

Formal definitions for content, services, and users have been completed along with many service implementations. Early prototypes of SimDL have been implemented for managing individual simulation models and the cyberinfrastructure. However, functionality for some services is currently in development, and integration requires tight coupling through assumed brokers. Quantification of the interoperability for users, architecture, and content has yet to be determined. Services are currently being organized into a registry for discovery and reuse. While the general simulation-supporting digital library framework has been well described, developing all three full deployable instantiations is ongoing. Case and user studies are planned to evaluate the completeness of SimDL services for supporting required user tasks.

Acknowledgements

I thank my external collaborators and members of DLRL and NDSSL for their suggestions and comments. This work has been partially supported by NSF Nets Grant CNS- 0626964, NSF HSD Grant SES-0729441, NSF PetaApps Grant OCI-0904844, NSF NETS Grant CNS-0831633, NSF REU Supplement Grant CNS-0845700, NSF Netse Grant CNS-1011769, NSF SDCI Grant OCI-1032677, DTRA R&D Grant HDTRA1-0901-0017, DTRA CNIMS Grant HDTRA1-07-C-0113, DOE Grant DE-SC0003957, US Naval Surface Warfare Center Grant N00178-09-D-3017 DEL ORDER 13, NIH MIDAS project 2U01GM070694-7 and NIAID & NIH project HHSN272201000056C.

8. REFERENCES

[1] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathe. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC '08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, pages 1–12, Piscataway, NJ, USA, 2008. IEEE Press.

[2] K. R. Bisset, J. Chen, X. Feng, V. A. Kumar, and M. V. Marathe. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *ICS '09: Proceedings of the 23rd international conference on Supercomputing*, pages 430–439, New York, NY, USA, 2009. ACM.

[3] C. L. Borgman, J. C. Wallis, M. S. Mayernik, and A. Pepe. Drowning in data: digital library architecture to support scientific use of embedded sensor networks. In *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '07, pages 269–277, New York, NY, USA, 2007. ACM.

[4] D. Castelli, Y. Ioannidis, S. Ross, H. Schek, and H. Schuldt. Reference Model for DLMS. *DELOS Network of Excellence on Digital Libraries*, 2006.

[5] E. Deelman, J. Blythe, Y. Gil, C. Kesselman, S. Koranda, A. Lazzarini, G. Mehta, M. A. Papa, and K. Vahi. Pegasus and the Pulsar Search: From Metadata to Execution on the Grid. In *Applications Grid Workshop at the Fifth International Conference on Parallel Processing and Applied Mathematics (PPAM)*, pages 821–830, Czestochowa, Poland, 2003.

[6] R. Dunlap, L. Mark, S. Rugaber, V. Balaji, J. Chastang, L. Cinquini, C. DeLuca, D. Middleton,

and S. Murphy. Earth system curator: metadata infrastructure for climate modeling. *Earth Science Informatics*, 1:131–149, 2008. 10.1007/s12145-008-0016-1.

[7] M. A. Gonçalves, E. A. Fox, L. T. Watson, and N. A. Kipp. Streams, structures, spaces, scenarios, societies (5S): A formal model for digital libraries. *ACM Trans. Inf. Syst.*, 22(2):270–312, 2004.

[8] T. H. Jordan. SCEC 2009 Annual Report. *Southern California Earthquake Center*, 2009.

[9] S. Kethers, X. Shen, A. E. Treloar, and R. G. Wilkinson. Discovering Australia's research data. In *Proceedings of the 10th annual Joint Conference on Digital Libraries*, JCDL '10, pages 345–348, New York, NY, USA, 2010. ACM.

[10] J. Leidig, E. Fox, M. Marathe, and H. Mortveit. Epidemiology experiment and simulation management through schema-based digital libraries. In *Proceedings of the 2nd DL.org Workshop at ECDL, Making Digital Libraries Interoperable: Challenges and Approaches*, pages 57–66, 2010.

[11] P. P. Leonardo Candela, Donatella Castelli. D4Science: an e-infrastructure for supporting virtual research. In *Proceedings of IRCDL 2009 - 5th Italian Research Conference on Digital Libraries*, pages 166–169, 2009.

[12] N. Li, L. Zhu, P. Mitra, K. Mueller, E. Poweleit, and C. L. Giles. oreChem ChemXSeer: a semantic digital library for chemistry. In *Proceedings of the 10th annual Joint Conference on Digital Libraries*, JCDL '10, pages 245–254, New York, NY, USA, 2010. ACM.

[13] B. LudŁscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao. Scientific Workflow Management and the Kepler System. In *Concurr. Comput. Pract. Exper*, page 2006, 2005.

[14] S. Majithia, M. S. Shields, I. J. Taylor, and I. Wang. Triana: A Graphical Web Service Composition and Execution Toolkit. In *Proceedings of the IEEE International Conference on Web Services (ICWS'04)*, pages 514–524. IEEE Computer Society, 2004.

[15] R. W. Moore, A. Rajasekar, M. Wan, Y. Katsis, D. Zhou, A. Deutsch, and Y. Papakonstantinou. Constraint-based Knowledge Systems for Grids, Digital Libraries, and Persistent Archives: Yearly Report. In *SDSC TR-2005-5*, 2005.

[16] T. Oinn, M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe. Taverna: lessons in creating a workflow environment for the life sciences. In *Concurrency and Computation: Practice and Experience*, volume 18, pages 1067–1100, 2006.

[17] T. Smith, G. F. Killeen, N. Maire, A. Ross, L. Molineaux, F. Tediosi, G. Hutton, J. Utzinger, K. Dietz, and M. Tanner. Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of Plasmodium falciparum malaria: Overview. In *Am. J. Trop. Med. Hyg.*, pages 1–10, 2006.

[18] J. Zhao, M.-Y. Kan, and Y. L. Theng. Math information retrieval: user requirements and prototype implementation. In *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '08, pages 187–196, New York, NY, USA, 2008. ACM.