System-Level Experimentation: Social Computing and Analytics for Theory Building and Evaluation

Tom McDermott, Molly Nadolski, Dennis Folds
Georgia Institute of Technology
Atlanta, Georgia
Email: tom.mcdermott@gtri.gatech.edu

Abstract— **This paper introduces the concept of shared data experimentation platforms as a means to transform access to and sharing of social science research data. Such platforms are becoming a central component of biomedical research, and are expanding into other fields. We discuss a framework for the development of data analytic experimentation platforms in the social sciences. Social situations are inherently complex adaptive systems that a difficult to generalize without explicitly documenting both the phenomena and related context. We introduce the concept of a "campaign of experiments" that focuses on purposeful exploration of social phenomena in order to evaluate generalizable, reproducible, and repeatable theory. We also propose sociotechnical systems analysis methods to define the appropriate conceptual models of social situations, which can then be used to structure the experimentation data in a form that promotes reuse and replication. We discuss challenges and opportunities associated with an experimentation platform concept, methodologies that can support development of such platforms.**

Keywords-sociotechnical systems; complex adaptive systems; data modeling; conceptual modeling; experimentation.

I. Introduction and Problem Statement

This paper introduces the concept of shared data experimentation platforms as a means to transform access to and sharing of social science research data. Such platforms are becoming a central component of biomedical research, and are expanding into other fields as diverse as international affairs, materials research, and system design. Digital network technologies supporting cloud computing, federated data architectures, knowledge graphs, data mining and machine learning, standardized web ontologies, digital annotation, experimental workflow sharing, computer visualization, crowdsourcing, and computer gaming are creating unprecedented capability for shared study of social behaviors. Although data sharing platforms like Harvard Dataverse are available to share the detailed results of scientific studies, in this paper we discuss the idea of federated data models for experimentation – platforms that allow geographically dispersed cohorts of researchers to work together on scientific experiments around a common problem or area of study. To our knowledge such platforms have not yet entered use in the social sciences community. This paper discusses challenges and opportunities associated with an experimentation platform concept, methodologies that can support development of such platforms, and an example case where a shared experimentation platform would be useful.

Unlike many other scientific areas of study, social situations represent complex adaptive systems that are characterized by independent agents who self-organize, adapt, and learn. In complex adaptive systems, broadly applicable models of behavior are difficult to generalize. The situation under study and the context of the situation must be studied together, and generalization across multiple contexts is not always wise or possible. Adaptation often makes generalized results short-lived. Intervention in social situations focuses heavily on causal relationships, but generalizing to purely linear causal relationships is often unsuccessful. Study of such systems must eventually account for *linear causal* relationships and also *circular causal* relationships, self-organization or *adaptive causal* relationships, and *reflexivity* which acknowledges the act of studying the system can effect causal relationships [1]. Generalization of results using linear regressions is most common and appropriate, but can only be accomplished by applying assumptions with respect to the other three causal models that are often not captured with the data. These assumptions are often about which of a number of potential causes aggregate to larger populations, making explanations of causality difficult.

Because of such "shifts in causality," reduction to linear models make the generalization of effects across multiple contexts difficult. They can also limit the reproducibility and replicability of social science study [2]. Issues related to reproducibility can be reduced by use of common datasets with access to original study data, models, and tools. Study replicability requires access to the original study methods, participants, instruments, and sampling approaches. Generalization requires access to sampling methods as well as both positive and negative results, and more difficult, the original assumptions and abstractions used by the researcher to conceptualize the study. However because many of these assumptions are related to selection of causal factors, effective conceptual models that capture context in the form of broader causal factors with hypotheses related to context-specific selections can help. The ability to do this has been until recently limited by the time and effort required to collect and analyze data, a condition which is changing rapidly.

Designing data analytic and computational models that accurately reflect performance measures at different layers of society, and the aggregation of measures from one layer to the next, is the primary conceptualization problem in social analysis and policy practice. Behavioral aspects of complex sociotechnical systems can be influenced at any layer of the system, but initiatives that try to analyze and improve factors at one level do not necessarily translate into positive influence at other layers. Moreover, the timeframes for measuring effects can vary greatly across different factors and societal layers [3][4]. Lack of common methods and tools to define model abstraction and aggregation of data create further barriers to generalization, which tie back to the original

ISBN: 978-1-61208-519-7

conceptualization of the study and related selection of constructs and dependent variables.

Issues and concerns with use of data analytic methods in social experiments reflect the complex adaptive systems aspects of social phenomena. These include determining appropriate context, understanding both linear and non-linear causality, representing differing time scales, uncertainty about what constitutes entities that affect the system, and issues with agency or agent identification [5]. These can be overcome by viewing the social problem of interest as a system then conceptualizing both the problem system and response system as a set of conceptual and then dynamic models. Research related to enterprise systems of systems and sociotechnical systems analysis introduces a methodology to address these issues.

Shared experimentation implies agreement on paradigms that reflect the problem definition and contexts of interest, as well as the semantic descriptions of the sociotechnical system of interest, and the conceptual model of the current systems' behaviors and future states. The concept of an experimentation platform implies a set of methods and tools to define and address these agreements, which we discuss prior to descriptions of the tool framework.

In Section II, we introduce the concept of an experimentation platform, using references from a United States Air Force concept as an appropriate framework for this application. We describe emerging computer platforms that make this concept a viable approach, and a methodology for building community-wide models in these platforms. In Section III, we describe the characteristics of a tool platform for experimentation, and the technological approaches that might be used to build it. We do not at this point describe a complete toolset, but a call for research to create these tools.

II. EXPERIMENTATION PLATFORM CONCEPT

In this section, we discuss a set of methods and tools that can be applied to social situations in support of a system level experimentation platform.

A. *System Level Experimentation*

Alberts et al. [6][7] captured a useful vision for information age transformation of social theories and related analytics in pursuit of a set of methods we refer to as "System Level Experimentation." The authors define this as a "campaign of experimentation," or a "set of related activities that explore and mature knowledge about a concept of interest." Although developed as an approach for transforming military command and control, the general model of such a campaign provides a framework for joint experimentation in any social decision making domain. The framework is a scientific method for experimentation, which includes theory development, conceptualization or conceptual modeling, formulation of questions and hypotheses, collection of evidence, and analysis. The approach views system transformation as a campaign of multiple experiments that produces a body of knowledge that creates a foundation for future experiments. Such campaigns have leaders and goals, research cohorts who use and create knowledge aligned with the goals, and a shared knowledge capture framework that allows federated cohorts and experiments against a common knowledge model.

With respect to reproducibility, repeatability, and generalization of experiments, the idea of a campaign focuses the research process on aligned goals with deliberate urgency and resource allocation. Alberts and Hayes note, "*reuse here applies to ideas, information about investigations conducted, data collected, analyses performed, and tools developed and applied. In terms of experiments, it implies replication. Reuse, and hence progress, is maximized when attention is paid to the principles of science that prescribe how these activities should be conducted, how peer reviews should be executed, and when attention should be paid to the widespread dissemination of findings and conclusions*."

The authors stress the importance of a shared conceptual model as a key to generalization, reproducibility, and replicability. Although in many scientific studies there exists a shared paradigm of study and generally shared conceptualization, this is difficult to achieve in social situations where stakeholder perspectives, even those of research communities, are difficult to align. For example the community measurement paradigm for "standard of living" is moving from a Gross-Domestic Product (GDP)-based measure of production to more representative consumption-based representations. However, the GDP measure was conceptually simple, and consumption measures are conceptually complex. Although the community is accepting the paradigm shift, there do not exist common agreed upon conceptual models of standard of living that can drive shared and replicable experimentation. Thus an effective shared experimentation platform must address common conceptualization artifacts as well as data and potentially dynamic models.

B. *Emerging Data Analytics Platforms*

What we can do much more easily these days is collect the data. Public datasets that report social variables in both broad and localized contexts are becoming widespread. Shared community data warehouses and models for experimentation purposes are becoming more widely used in complex health and medical studies, leading one to believe that such approaches may also have use in social research and analysis. Notable examples of medical research platforms include the Global Alzheimer's Association Interactive Network (GAAIN) [8] and the Medical Informatics Platform (MIP) of the European Union's Human Brain Project [9]. Common features of these projects include a federated data model, shared schemas or data codings, machine learning tools for extraction and matching of data, and web-based interfaces to data, research cohorts, and visualizations. In all such projects, a shared database is created where an entity-relationship model defines the schema of the resultant "data warehouse," and agreed upon data codings provide a map between the larger sets of data and the phenomena of interest. We will further explore the possibility of designing similar projects for social data experimentation.

To reach this point, the community must develop not just common data, but also methods for agreement on research paradigms, related stakeholder perspectives of problem and solution spaces, associated viewpoints, and shared conceptualizations. Thus long-term success in social analytics must address the capture of both the data and conceptual relationship models that make the data meaningful. These conceptual relationships are often determined using soft systems approaches, which is appropriate, but existing methods and tools do not adequately connect the conceptual artifacts with the data-driven analytics. In the social analytics field, there is a need for research that connects the resulting collected data to its conceptual model artifacts. Without these problems with abstraction, generalization, reproducibility, and replicability cannot be resolved. Research from the systems engineering community centered on management of enterprise systems-of-systems provides a set of useful methods and tools.

C. *Enterprise Systems of Systems Methodology*

Sociotechnical systems analysis is a specific methodology that supports assessment of multiple factors across all layers of a complex enterprise or societal construct using sets of tools derived from system science and system modeling. The methods recognize that factors arise from the interaction of many and diverse enterprises that can be defined by their entities, relationships, established processes, pursued strategies, and emergent phenomena. The sociotechnical systems analysis attempts to capture the combined conceptual, data, and analytical modeling artifacts necessary to completely describe the problem [10][11].

With respect to social situations, the method produces a set of artifacts that describe the system context and boundaries, system entities and relationships, primary construct variables, potential causal variables, and phenomena of interest. The process is conducted such that insight can be fed into dynamic computer models. Hypotheses that intervene in lower level causal factors can then be viewed as they aggregate up into larger population behaviors. The sociotechnical systems analysis produces artifacts that communicate the abstractions and aggregation of behaviors across different scales, helping to explicitly document both the assumed and modeled variables.

At the core of a sociotechnical systems model are entities and their relationships, which can be organized into associated databases and warehouses. The entity-relationship model can be created, modified, and refined over periods of short and long term study. Standardized codings of the data entities then make relevant data elements accessible to researchers and analysts. One use of this is for data collection and analysis, but the sociotechnical systems analysis methods are focused on development of experimentation platforms. Experimentation requires that not only the data but also the underlying conceptual models context of study be updated over time.

The conceptual model representations produced by the sociotechnical systems analysis serve as a bridge between the soft systems aspects of the problem (systems thinking) and the quantitative analysis approach (design). This is an area that needs significant additional research as related to methods and tool design. However recent advances in machine learning and semantic graphs can bring the semantic model and mathematical model artifacts into the same toolsets. The bridge between the two is a conceptual model that uses semantic models to specify the analytical models. We identify these as metamodels as they should describe broader conceptual models and data, while individual experiments explore a subset of executable models and constructs related to central questions of interest. Fig. 1 describes that bridge.

We define the soft systems aspects in Figure 1 as "*System Metamodeling*" using three fundamental abstraction approaches: system metamodels, system constructs, and system architecture models. These are determined in a participative, inquiry-based process. We describe hard system aspects as "*Executable Metamodeling*" determined by a specification and design workflow using conceptual models, executable metamodels, and data visualization. It is useful to think about this as a tool framework. The tools support structuring the systems metamodel, creating the conceptual models, creating the executable metamodels, analyzing and visualizing the decision space, and managing the contained knowledge over time [12].

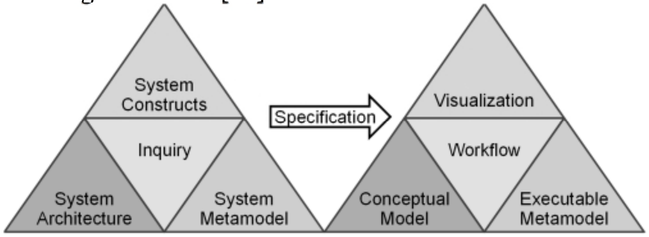

Figure 1. The bridge between soft systems analysis and social analytic model specification.

The system metamodel is described as the set of constructs and rules used to define semantic relationships across information sets, associated data sets, and methodologies or processes [13]. The metamodel definition on the semantic side is an architectural description of the system using modeling views and stakeholder viewpoints. The executable metamodel is the dataset design and any associated computational models.

D. *Metamodels and Federated Data Models*

The emerging medical community models link together research cohorts by providing a common data model for integrating federated datasets. As experimentation platforms they provide a cohort discovery tool to link research communities, a federated data model integration architecture, and a common data visualization toolset that allows data exploration across multiple cohort data. The federated approach to data model integration allows individual cohorts to maintain their own working datasets while sharing and using data from other cohorts via a common data model representation. State of the art tools for data discovery, transformation, and integration automate most of the source data integration into the common data model. The common data model is implemented as a schema in a relational database using agreed upon codings for data tables and variables.

In a federated data model design, metadata or data descriptions are essential to data harmonization – integrating data from different sets and integrating experimental data back into the common data warehouse. Emerging data mining and machine learning tools can automate data harmonization assuming the metadata has a rich enough natural language description of the data elements to link multiple sets. Mapping variables between federated datasets and the common data model is accomplished by extracting and matching the data entities via descriptive data mapped from element descriptions in data dictionaries, a component of metadata. Adequate metadata provides a path to harmonizing the often cryptic tags placed on data elements in databases. Transformation tools are provided to map data between the common model representation and federated datasets [14].

The conceptualization of most existing common data model examples were developed initially from manual coding and integration of existing datasets [15][16]. In the social analytics area, a common conceptual definition of the data tables and entities would be a huge undertaking due to the tremendous differences in terminology, conceptual data relationships, and assumptions made around data generalizations across societal scales. Emerging approaches for graph representation of data entities and relationships should be explored in the social sciences arena as a tool for amassing large volumes of linked data and knowledge supporting both generalized and contextual research results.

III. Social Experimentation Tool Framework

We present a generalized concept for social experimentation and analytics using both bottoms-up software environment and top-down conceptual architecture descriptions. The purpose of this discussion is not to present the design of an existing tool (none exist), but to describe the characteristics and architectural constructs of future frameworks for social experimentation and analysis. Fig. 2 presents our high level system and process architecture.

Alberts et al. note that "*For purposes of building knowledge, the most important elements are (1) consistent language (clear and operational definitions and measures), (2) explicit use of metatags (meta-data) on data, and (3) clear and complete descriptions of assumptions. These are part and parcel of an explicit conceptual model.*"

A *consistent language* and *use of metatags* relate to the semantic model of the system of interest. This is often described as an ontology, but the term "System Metamodel" is more appropriate. The *description of assumptions* refers to appropriate documentation of construct variables and associated contextual assumptions of lower level abstractions.

The use of inconsistent language to name the data elements in the resulting database is the major limitation of a common data model, it can take years to agree on data element definitions and a static data schema can make the data model difficult to modify. Data element names are often useless to infer meaning. These issues can be abated by consistent mapping generated from data element descriptions in data dictionaries, a primary component of metadata. Data providers that create rich metadata and share this across the data federation will aid in effective model and data sharing. Metadata has additional benefit as it can hide the actual data if it is restricted, without impacting the federation [15]. Data value ranges and units must also be consistent or readable from the metadata.

Three general developments emerging from modern web standards aid in linking different data collections from different domains. The first is the Web Ontology Language (OWL) and widely used Resource Description Framework (RDF) stores such as Google's FreeBase. The standard subject-predicate-object or object-attribute-value framework and semantic linking ease in the standardization of semantic terms and relationships. Various domains are rapidly creating large RDF stores or web ontologies describing their domain. To date relatively little development and standardization of common web ontologies have been undertaken across the social sciences domain. However as researchers opt to use existing ontologies and create domain specific ones, conditions will improve. A consistent language representation is the foundation of a good system metamodel.

A second development is extensive use of web linked data standards. Most database schemas remain defined in eXtensible Markup Language (XML) form but the web community is transitioning to JavaScript Object Notation (JSON) format for standard document annotation and linking of data to research. JSON is a computer language independent format for sharing objects and attribute-value relationships across different datasets, documents, etc. in addition, the use of annotated Hyper-Text Markup Language (HTML) documents to describe research experiments and link input data and results will aid in broader community sharing.

A third area of exploration is the evolution of linked graphs of semantic and mathematical information, an area that

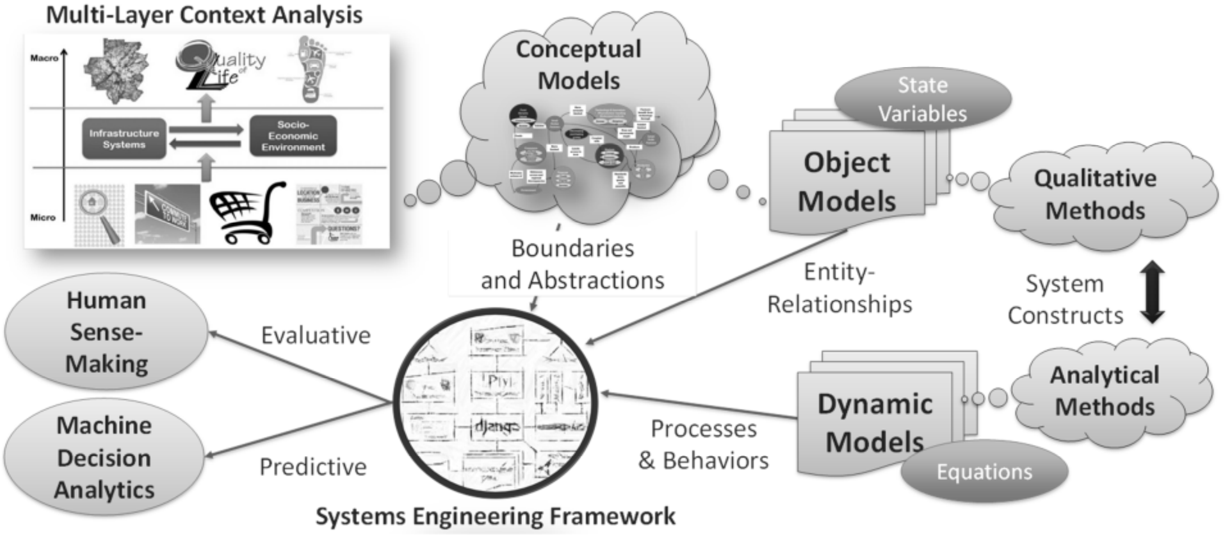

Figure 2. Conceptual Architecture.

is rapidly developing due to Google's introduction of Knowledge Graph and similar entity-driven stores of large information sets. Graph structures support semantic integration and structuring of linked data by compiling text into linked nodes and then relating these to concepts that provide shared meaning to the text. In the graph structure the metadata of our data federation could be linked into a semantic network that can be grown over time with new data. This is an area of needed research; the ability to create large curated sets of community shared and agreed upon causal data and linked experimental results could transform social science research.

A significant hurdle in social science use of these tools is reconciling the linking of different actors' viewpoints to the standard object-attribute-value ontologies. Different actors assign different meaning to social entities and relationships, making contextual features of language by the actor an important variable. The specific meaning associated with the language used by different actors requires a different structuring of shared ontologies than used in most of these applications today. This is an area for further research.

Finally, the use of these new technologies does not inherently capture the conceptualizations that defined that data to be important in the first case, and it does not capture assumptions made about missing data elements in the graph. Discerning real causality from experimental measurement of a social construct often requires a qualitative analysis of the underlying causal variables that cannot be measured directly. This is an underlying conceptual model that is often not fully documented in the research results, particularly those potentially causal variables that were purposefully not assessed in the research. This is where context becomes critical – discussions of why these variables are assumed to be causal in this context versus different variables in another context – becomes a key component of the knowledge base. Existing computer-based data models and analytical models are not linked to their conceptual parent models, primarily because the available modeling tools have not been built. A related area of research is specific to this problem, which is how to formally link more freeform conceptual diagramming or facilitation artifacts with more constrained formal modeling and simulations tools.

The "clear and operational definitions and measures" noted by Alberts et al. [7] in the military context is a difficult hurdle in less well governed social situations. Operational definitions and measures in social situations tend to be an area of great debate between different communities of interest. A GAAIN-like common data model is doomed to fail unless we can also define methods and tools to reach agreement on the conceptual models that drive entities, relationships, data definitions, and assumptions. Much of this disagreement involves data conceptualization, definition, and abstraction/aggregation at different scales (for example macroscale measures like "GDP per capita" versus microscale measures like "owning a dishwasher" – both used to describe standard of living). Emerging computer approaches to semantic integration offer hope for much richer microscale measurement sets, as long as the community can clearly see the need for research in this area.

IV. Conclusions

We discussed a concept for a social experimentation and data analytics platform based on emerging data and model federations that are emerging in medical and other research areas. This type of platform has not been explored for use in social science research, although the type of tools and technologies that can be applied are finding broad use in other disciplines.

The differences between social science research and other domains of research make a platform of this type much more difficult to envision and build. Problems of data abstraction and aggregation, differing actor viewpoints, and differing conceptualizations of system models make traditional data federations too expensive and time consuming to maintain. However emerging technologies associated with linked data, knowledge graphs, machine learning, and conceptual design tools provide a research base to explore implementation of social data experimentation platforms. This summary paper describes the concept as a means to encourage such exploration.

References

[1] S.A. Umpleby, "Second-order science: logic, strategies, methods," Constructivist Foundations 2014, vol. 10, no. 1, pp. 16-23, 15 November 2014.

[2] K. Bollen, J Cacioppo, R.M. Kaplan, J.A. Krosnick, and J.L. Olds, Social, Behavioral, and Economic Science Perspectives on Robust and Reliable Science, Report of the Subcommittee on Replicability in Science, Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Science, May 2015.

[3] J. Rotmans, R. Kemp, and M. van Asselt, "More evolution than revolution: transition management in public policy", Foresight, vol. 3, no. 1, pp. 15-31, February 2001. ISSN 1463-6689.

[4] F.W. Geels, "Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study." Research Policy, vol. 31, pp. 1257–1274, 2002.

[5] R. Wagner-Pacifici, J.W. Mohr, and R.L. Breiger, "Ontologies, methodologies, and new uses of Big Data in the social and cultural sciences," Big Data & Society, vol. 2 iss. 2, pp. 1-11, December 2015. DOI: 10.1177/2053951715613810.

[6] D.S. Alberts, R.E. Hayes, D.K. Leedom, J.E. Kirzl, and D.T. Maxwell, Code of Best Practice for Experimentation, Washington DC: CCRP Publication Series, 2002.

[7] D.S. Alberts and R.E. Hayes, Code of Best Practice for Campaigns of Experimentation: Pathways to Innovation and Transformation, Washington DC: CCRP Publication Series, 2002.

[8] www.gaain.org, retrieved: July 2016.

[9] www.humanbrainproject.eu/mip, retrieved: July 2015.

[10] W. B. Rouse and D. Bodner, Multi-level modeling of complex socio-technical systems – phase 1, A013 - final technical report, SERC-2013-TR-020-2, Systems Engineering Research Center, 2013.

[11] W. B. Rouse and M. Pennock, Multi-level modeling of socio-technical systems a013 - final technical report, SERC-2013-TR-020-3, Systems Engineering Research Center, 2013.

[12] T. McDermott and D. Freeman, Systems thinking in the systems engineering process: new methods and tools, in Systems Thinking: Foundation, Uses and Challenges, Eds. Frank, Shaked, Kordova, Nova Publications, 2016.

[13] J. Ernst, "What is metamodeling, and what is it good for," http://infogrid.org/trac/wiki/Reference/WhatIsMetaModeling, retrieved: November 2015.

[14] N. Ashish and A.W. Toga, “Medical data transformation using rewriting,” Frontiers in Neuroinformatics, vol. 9, no. 2, pp. 1-8, 20 February 2015. doi: 10.3389/fninf.2015.00001

[15] N. Ashish, P. Dewan, JL Ambite, and A.W. Toga, GEM: The GAAIN Entity Mapper, in Data Integration in the Life Sciences, 11th International Conference, DILS 2015, Eds. Ashish, N. and Ambite, J., Springer 2015.