AIChE JOURNAL

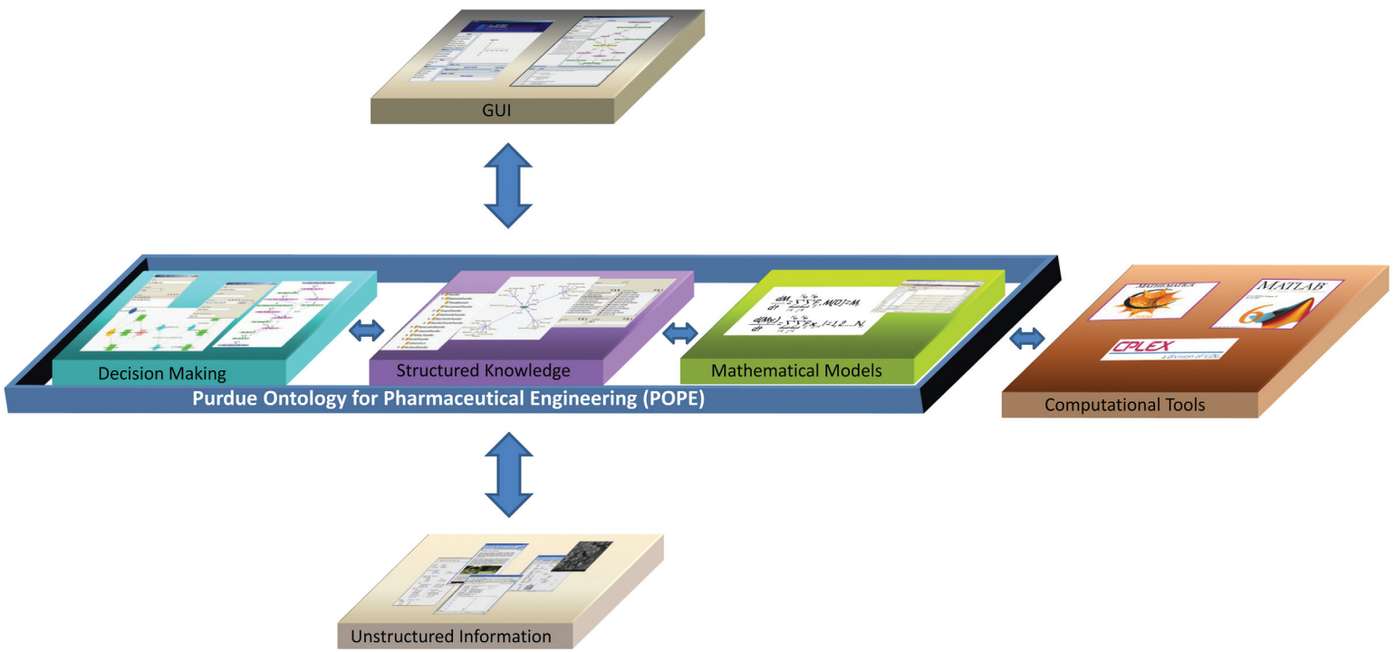
VOL.55, NO.1, PAGES 0000-0000

JANUARY 2009

AIChE JOURNAL

AN OFFICIAL PUBLICATION OF THE AMERICAN INSTITUTE OF CHEMICAL ENGINEERS

CHEMICAL ENGINEERING RESEARCH AND DEVELOPMENT **January 2009**



perspective

DROWNING IN DATA: Informatics and Modeling Challenges in a Data Rich Networked World

ARTICLES PUBLISHED ONLINE IN WILEY INTERSCIENCE,
OCTOBER 00, 2008 THROUGH NOVEMBER 00, 2008

WILEY-BLACKWELL

AIChE

Perspective

DROWNING IN DATA: Informatics and Modeling Challenges in a Data-Rich Networked World

Venkat Venkatasubramanian
Laboratory for Intelligent Process Systems, School of Chemical Engineering, Purdue University, West Lafayette, IN 47906

DOI 10.1002/aic.11756
Published online December 3, 2008 in Wiley InterScience (www.interscience.wiley.com).

Keywords: informatics, cyberinfrastructure, modeling challenges, process systems engineering

From a "Data Poor" to a "Data Rich" Paradigm

"May knowledge surround us in all directions!"—so prayed the ancient seers of the Rig Vedic era. It appears, however, we will all be surrounded by data, indeed drown in a deluge first, before our deliverance to the promised land of knowledge and understanding.

Chemical engineering is at important crossroads. Driven by a convergence of powerful forces such as the great progress in molecular sciences and computer/communications technologies, ever increasing automation of globally integrated operations of our enterprises, tightening regulatory constraints, and competitive business pressures demanding speed to market for products and services, our discipline is in an unprecedented transition. One important common outcome from this convergence is the generation, use, and management of massive amounts of diverse data, information, and knowledge.

Such a data deluge is coming from smart sensors in process plants, *ab initio* quantum calculations, molecular dynamics simulations, and so on. We are moving from an era of limited data obtained through time-consuming experiments and simulations to one of a tsunami enabled by high-throughput experiments and TeraGrid computing environments— it is a dramatic transition from a "data poor" to a "data rich" paradigm. Further, the extensive monitoring of equipment, processes, and products at all scales, from individual units to globally integrated supply chains, enabled by revolutionary progress in sensing and wireless communication technologies such as RFIDs, is another source of such data overload.

For instance, in pharmaceutical drug development and manufacturing, the amount and complexity of information of different types, ranging from raw experimental data to laboratory reports to complex mathematical models, that need to be stored, accessed, validated, manipulated, managed, and used for decision making are staggering (McKenzie et al., 2006). A tremendous amount of information is generated in the form of raw data from analytical instruments, images, spectra, laboratory notes, various calculations from simulation tools, chemometric models, etc. This information is often in different formats, such as plain text files, Word documents, Excel worksheets, JPEG files, MPEG movies, mathematical models, and so on. A typical FDA filing for a new drug approval requires nearly half a million pages of documentation of such data and information.

However, it is not raw data that we are after. *What we desire are in-depth knowledge and mechanistic, first-principles based, understanding of the underlying phenomena that can be modeled to aid us in rational decision making*. However, knowledge extraction and model development from this data deluge are major challenges. Past approaches developed in a "data poor" era do not work well in this new world. The new environment requires imaginative thinking to address these challenges. This is where cyberinfrastructure (CI) and informatics will play a crucial role.

Now, cyberinfrastructure and informatics may mean different things to different people; therefore a definition, however, limited, can be helpful. The report from the National Science Foundation's Blue-Ribbon Advisory Panel on Cyberinfrastructure, the *Atkins Report* (Atkins, 2003), uses an analogy to industrial infrastructure such as transportation, communication, or power systems to define cyberinfrastructure as "*the infrastructure based on distributed computer, information, and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy*." Another definition, also put forward by NSF which may be considered as a *working* definition, states that CI is the "*integration of hardware, middleware, software, data bases, sensors, and human resources, all interconnected by a network*." Obviously, even this is quite broad, but that's the nature of the terrain. Nevertheless, it articulates the different components and stresses the importance of middleware, integration, and networking.

V. Venkatasubramanian's e-mail address is: venkat@ecn.purdue.edu.

In a similar vein, for informatics, a useful working definition is that it is the study of the structure, algorithms, behavior, and interactions of systems that store, process, access, manage, communicate, and use information. While the discipline of informatics has developed its own conceptual and theoretical foundations, it also utilizes foundations developed in other fields. Informatics is generally seen as a key component of cyberinfrastructure.

One might ask at this juncture, quite reasonably, "How are cyberinfrastructure and informatics approaches different from what we have been doing all along?"

Much of the past contributions only addressed different slices of the overall problem, but not the entire picture — data, information, and knowledge management issues were addressed separately, leading to stand alone systems with limited capabilities and significant integration barriers. In addition, the amount and complexity of data are orders of magnitude greater now, which is changing the scope of the challenge dramatically, resulting in data warehouses that often become data graveyards. The time has come, and the appropriate theoretical frameworks and practical technologies are emerging, to address this problem *in toto* by developing more comprehensive approaches.

From Data to Knowledge to Decisions: Beyond Differential Equations

Addressing these formidable modeling and informatics challenges would require a broader approach to modeling than what chemical engineers are used to.

In common parlance, one tends to use the terms data, information, and knowledge more or less synonymously. However, they do represent different concepts, and that distinction becomes important in developing formal frameworks to represent and reason with them. Simply stated, *data* address the question *what* (e.g., "What is the temperature T in reactor CSTR-108?"), while *information* answers the question *how* (e.g., "How are T and the concentration of product B (C_B) related?"), by revealing the relationships or correlations between the entities of interest (e.g., C_B increases with T). While *information* is more "informative" than raw data are, it does not, however, answer the crucial question of *why* C_B should increase with T. For this we require a more detailed mechanistic model of the phenomenon, which we refer to as *knowledge* in this context.

Thus, data, information, and knowledge do address different things, as answers to the questions *what, how* and *why*. Answering *why* (and *why not*) in detail is the most important and fundamental of all quests.

Such distinctions become important as we develop *formal* methodologies necessary for *automating* tasks such as acquisition, representation, storage, manipulation, modification of, and reasoning with, data, information, and knowledge. The formal representation techniques used for data, information, and knowledge are often quite different, the automation protocols for their exchange are also different, and so are the algorithms used for manipulating and reasoning with them for decision-making. In the past, these issues did not seem terribly important as there was a tremendous amount of *human intervention* in passing data and information from one program to another, and people took care of all these barriers. However, given the increased need to automate such computational tasks due to the data deluge, as well as due to the complexity of the problems we are addressing, these issues are now facing us front and center.

Generally speaking, when one mentions modeling to chemical engineers, people often think of a system of differential and algebraic equations (we will refer to this class of models as DAEs). However, there is a wider variety of knowledge representation concepts leading to other classes of models (Ungar and Venkatasubramanian, 1988) which will play an important role in this emerging future. While it is not the purpose of this article to have an extensive discussion on various modeling concepts, it is, nevertheless, useful for our theme to outline and summarize the issues involved.

One may broadly classify models into (1) mechanism-driven models based on first-principles, and (2) data-driven models. Again, each of these classes may be further categorized into (1) quantitative and (2) qualitative models. Combinations of these classes lead to hybrid models.

DAE models are suitable for a certain class of problems that are amenable to such a mathematical description; chemical engineering has abundant examples of this class. However, there are other kinds of knowledge that do not lend themselves to such modeling. For example, reasoning about cause and effect in a process plant is central to fault diagnosis, risk analysis, alarm management, and intelligent supervisory control. Knowledge modeling for this problem class does not typically lend itself to the traditional DAE view of modeling. In some simple cases perhaps one can, but they are incapable of addressing real-life industrial process systems, which are often complex and nonlinear with incomplete and/or uncertain data. Furthermore, even for simple systems DAE based models are not suitable for generating explanations about causal behavior. This problem often requires a hybrid model, such as a combination of a graph theoretical model (e.g., signed digraphs), or a production system model (e.g., rule-based representations), and a data-driven model (e.g., principal component analysis (PCA) or neural networks) (Venkatasubramanian et al., 2003).

Thus, while we are quite familiar with ODE/PDE, statistical regression, and mathematical programming models, we are less so with other classes such as graph theoretical models (as noted, used extensively to perform causal reasoning in abnormal events identification and diagnosis, risk analysis etc.), Petri nets (used for modeling discrete event systems), rule-based production system models (used in knowledge-based systems for automating higher-order reasoning), semantic network models such as ontologies (used in materials discovery and design that utilize complex relational databases, domain-specific compilers, etc.), object-oriented models such as agents (used in simulating the behavior and decision-making choices of independent, interacting, entities endowed with complex attributes and decision-making powers), and so on. In addition, there are the data-driven quantitative models such as pattern recognition based models (e.g., neural nets, fuzzy logic), stochastic models (e.g., genetic algorithm, simulated annealing), etc.

Even though Rudd, Siirola, and Powers recognized, as far back as the 1960s, the need for such an alternative modeling philosophy in the context of process synthesis (Rudd et al., 1971), not much work on this subject appeared in the chemi-

cal engineering literature until the 1980s. This renewed interest was propelled by the progress in knowledge representation and search techniques in artificial intelligence (AI), as well as in computing hardware and software. Outstanding examples from that era are the DESIGN-KIT system for process engineering (Stephanopoulos et al., 1987), and the DECADE system for catalyst selection (Bañares-Alcántara et al., 1987). However, the progress was viewed with skepticism and its products as curiosities at the fringes of academic research. Nevertheless, recognizing the importance of this alternative philosophy, both Warren Seider (1993) and George Stephanopoulos (1994) highlighted this need in their CAST award acceptance speeches in the hope of stimulating further research.

Over the recent decade, much progress has been made as these methodologies proved their value by addressing problems of practical importance, which were previously hard, even impossible, to solve using the traditional modeling techniques. These AI-based modeling approaches have become more mainstream now, accepted as a part of the modeling arsenal of process systems engineers, with several important successes in the domains of abnormal events management, molecular products design, process synthesis, process safety analysis, scheduling, and so on.

Despite all of this progress, the number of academic researchers developing the alternative modeling methodologies in chemical engineering is very small and woefully inadequate considering the emerging challenges. This is mostly due to two factors: (1) the barrier of entry is quite high — requiring a considerable investment in time and effort to achieve a sufficient level of mastery in knowledge representation and search techniques, algorithm engineering, databases, compilers, and so on, which are not part of a typical chemical engineering education program, and (2) certain misconceptions about what the intellectual challenges are, as discussed later.

However, the crucial emerging trend, borne out of several necessitating factors, is the ever increasing *automation of higher-order reasoning and decision-making*. These activities that were once considered to be the exclusive domain of humans, are slowly, but surely, being taken over by computers. Such automated reasoning and decision-making will be driven by models, but these are not going to be limited to the DAE-based models that we train our students on. To be sure, the DAE-based models will play their useful role wherever they are appropriate, but the other kind will play an increasingly important role as we discussed above. This has important implications for our educational mission.

In some sense, we would be better off if all our modeling needs could be addressed by just the DAE models. It is a much more mature field of study, with decades of literature on how to define, formulate, and solve such models. The alternative modeling philosophy, on the other hand, does not enjoy these advantages as much. These alternative models are typically used for ill-posed or ill-structured problems (hence, do not enjoy the luxury of being amenable to DAE models), lacking a unique approach, and with a certain amount of heuristic element to them. They often lack the beauty and rigor of the DAE models. However, these deficiencies are sufficiently compensated for by their ability to address, although not always, messy, real-life, industrial problems where they often provide very good solutions, even optimal ones.

It is important to recognize that these two approaches, in general, are complementary and not competitive, even though there are problems where they do compete. As noted previously, if indeed there is a DAE-based solution that would be satisfactory for a given problem, one should pursue it instead of the other methodologies. The strength of the alternate modeling philosophy lies is its ability to address a class of problems that are not amenable to the DAE framework. However, one often falls into the trap of using a particular modeling technique as the panacea for all problems — e.g., to a principal components analysis (PCA) specialist, every problem may seem to require that particular approach, reminding us of the adage "to a person with a hammer, every thing looks like a nail!" What we need is not such a "tool driven" outlook, but a "tool box" oriented modeling philosophy, where we are comfortable in using a wide variety of modeling tools, as determined by the features of the problem at hand. It is helpful at this point to remind ourselves of what Prof. George Box, the statistician from the University of Wisconsin, once famously said: "All models are wrong, some are useful."

Dispelling Some Misconceptions

All this would require innovative thinking, imaginative approaches, and getting over some misconceptions.

For instance, there is a tendency to *underestimate informatics as just programming or database management*. This would be akin to stating that chemical engineering is nothing but chemistry carried out in large vessels — i.e., "large-scale chemistry". About a hundred years ago, as our discipline was emerging, some people did think that, and, in fact, some of the early departments were called Deptartment of Industrial Chemistry, reflecting that attitude. Certainly, there is chemistry being done in those large process units, but, as we all know now, there is much more to it. Similarly, informatics certainly involves programming, but that's not all. Certainly, informatics involves data storage and management, but that's not all.

Another misconception is *this is not chemical engineering and that computer scientists/engineers will somehow address such problems for us*. Once again, invoking another chemical engineering analogy, this would be like thinking that since transport phenomena problems are solved by using differential equations, mathematicians would address these core problems of our discipline for us. Again, as we know, they cannot as they lack the application domain knowledge. They can provide us with generic concepts, tools, and techniques, but it is up to us to suitably adapt and extend these with our domain knowledge to solve our problems.

This leads to a related misconception that *this is all a matter of tracking new developments made by computer science researchers and applying them in chemical engineering*, implying that there is no creative contribution being made. But, then this criticism is equally applicable to the other modeling tools that chemical engineers use. It is important to realize that the use of informatics and intelligent systems modeling concepts and techniques in chemical engineering is no different from the modeling tools we have historically borrowed, adapted, and enhanced from mathematics (e.g., linear algebra, and differential equations, etc.), operations research (e.g., MILP, MINLP, IP, etc.), or statistics (e.g., PCA, PLS, etc.).

As it happened in these applications, the nature of our domain knowledge, constraints, and objectives provides us with opportunities to make novel intellectual advances on top of what we borrow from computer science. For instance, in the area of molecular products design (e.g., design of catalysts, fuel additives, rubber compounds, pharmaceuticals, etc.) or conceptual process design, the state-of-the-art techniques in computer science have not been adequate to address the representation and algorithmic challenges posed by the complexities in molecular structures, reaction networks and pathways, the need to automatically generate and solve a large set of model equations from reaction networks, and so on. This has led to advancements in the concepts, tools, and techniques themselves, such as the creation of domain-specific representations and languages (Bañares-Alcántara et al., 1987; Stephanopoulos et al., 1990; Prickett and Mavrovouniotis, 1997; Katare et al., 2004), compilers (Hsu et al., 2008), ontologies (Venkatasubramanian et al., 2006; Morbach et al., 2007), modeling environments (Stephanopoulos et al., 1990; Ghosh et al., 2006; Nagl and Marquardt, 2008), molecular structure search engines (Balachandra, 2008), and chemical entities extraction systems (Balachandra, 2008), etc. These references by no means constitute a comprehensive list, and there are other such contributions. These contributions go beyond the original computer science techniques which were developed for relatively simpler application domains that lack the richness of complex chemistries, nonlinear systems, process configurations, and decision choices seen in chemical engineering problems.

Cyberinfrastructure for Knowledge Modeling in Chemical Engineering: Current Trends and Future Outlook

The emerging "data rich" networked environment will eventually impact all aspects of our discipline and, in fact, the effects are being felt in some areas already. These effects are spawning new application domains and opportunities for our modeling skills. The following summary is not meant to be a comprehensive review, but only a representative survey of some recent developments that could serve as useful starting points for interested readers to explore further. We first summarize progress in general purpose environments, and then outline domain or application specific results.

Cyberinfrastructure developments have been along the lines of concepts, methodologies, and tools required for two broad categories of needs: (1) *Data Modeling and Management*, and (2) *Knowledge Modeling and Management*. Cyberinfrastructure typically involves (1) hardware, (2) software, (3) middleware, (4) networking, and (5) domain knowledge components in an integrated environment. Progress in hardware generally has been in developing high-performance and parallel computing environments such as the TeraGrid. Middleware are systems that connect multiple applications through the use of a common data model. Use of a common information model makes it possible for greater flexibility and efficiency. Software integration includes middleware between different software or programming languages. Networking usually exploits the Internet and wireless communication technologies.

General purpose tools and environments: Modeling and simulation

While efforts such as the CAPE-OPEN (Computer-Aided Process Engineering-OPEN) standard, which define interfaces between physical property packages, numerical solvers, and unit operation libraries, have been underway for a decade, issues of interoperability have been gaining a great deal of importance recently. In an extension of CAPE-OPEN, Braunschweig et al. (2004) proposed a framework (COGents) for combining and assembling CAPE-OPEN compliant components using software agents. Similarly, component-based hierarchical explorative open process simulator (CHEOPS) (Schopfer et al., 2004) is a conceptual framework for process model integration. CHEOPS includes a neutral model representation (conceptual object model), a tool wrapper, a modeling tool, a formulation bridge (to convert between two model formulations) and solution algorithms. More recent advances in this project are described in Nagl and Marquardt (2008).

General purpose tools and environments: LIMS and data warehouses

Vast amounts of data and information are frequently entered into Laboratory Information Management Systems (LIMS) and e-Lab Notebooks (electronic laboratory notebooks, ELN) (Paszko and Pugsley, 2000), which are repositories of raw data. Important recent progress toward storing and managing such large data sets include the virtual laboratory (VL) project at the University of Amsterdam (Frenkel et al., 2001), and XSIGMA (Kim et al., 2006). However, most current LIMS and ELN solutions utilize database schemas, which often limit the capacity for the description of complex relations between information entities. This is addressed by developing domains-specific ontologies as discussed below. Data warehouses (Gardner, 2005) use specialized database schemas to abstract and store a copy of data from several sources, and enable those data to be queried through a single query.

General purpose tools and environments: Ontologies, languages, and compilers

Addressing the challenges in modeling and informatics requires a departure from the *application-centric* approach of the past to an *information-centric* framework. In this new paradigm, the underlying data, information, and knowledge are modeled *explicitly*, independent of the tools that use these. Instead of encoding such information in objects in a particular programming language or tool specific constructs, the information is *explicitly described.* However, to describe the information explicitly, the syntax (i.e., structure), as well as semantics (i.e., meaning) of the information must be defined. The explicit description of domain concepts and relationships between these concepts is known as an *ontology* (Gruber, 1993).

Recent developments in the field of ontology have created new software capabilities that facilitate the implementation of the information-centric infrastructure. Compared to a database schema which targets physical data independence, and an XML schema which targets document structure, an ontology targets an agreed upon and explicit semantics of information. As a result, while the functionalities of this infrastructure can be implemented in a traditional client-server framework, the

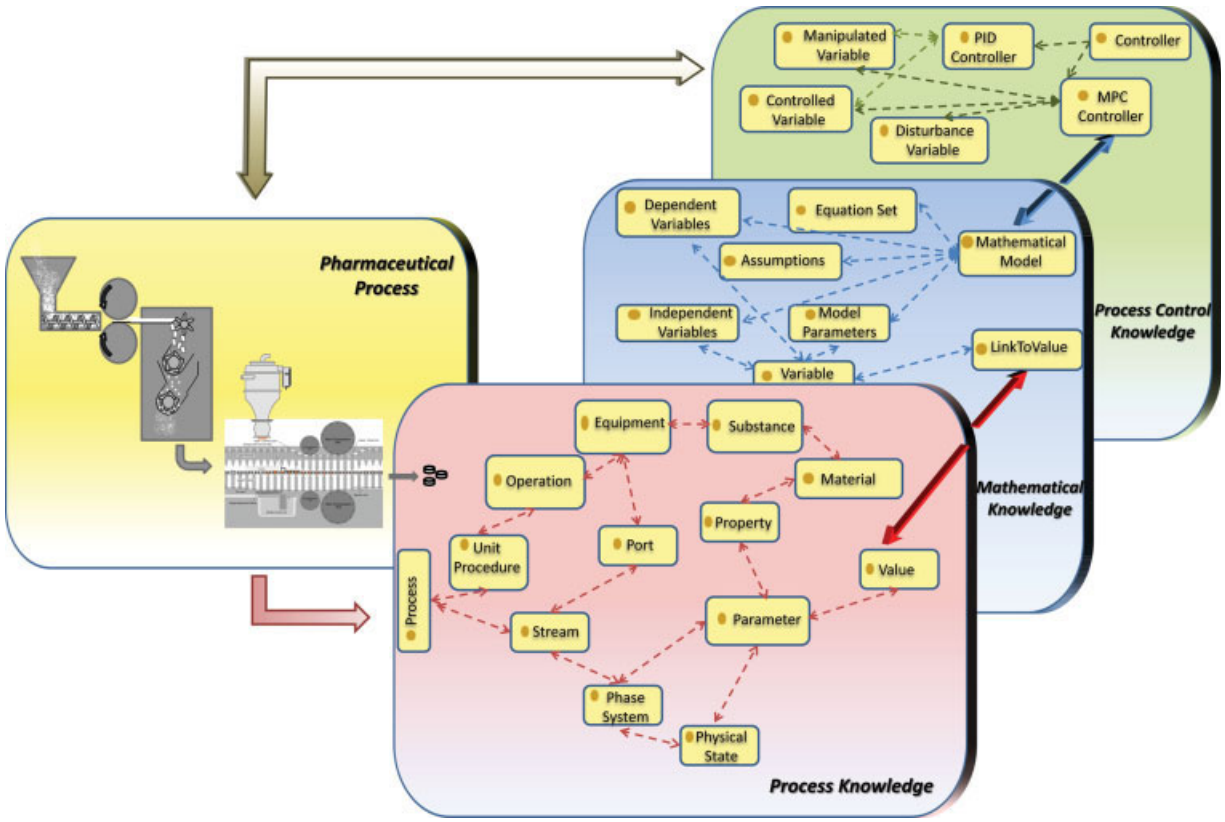


Figure 1. A more detailed view of some of the ontological relationships in POPE.

Prepared by Pradeep Suresh, Laboratory for Intelligent Process Systems, School of Chemical Engineering, Purdue University, W. Lafayette, IN 47906.

main benefits of this ontology-driven architecture are its openness and semantic richness.

One recent development is the Purdue Ontology for Pharmaceutical Engineering (POPE) to support automated decision-making by intelligent systems for pharmaceutical product and process development and manufacturing (Venkatasubramanian et al., 2006). POPE is composed of several ontological subsystems that formally model data, information, and knowledge regarding experiments, materials, chemical species and reactions, expert knowledge, unit operations, and mathematical models using the web ontology language (OWL) (OWL, 2004). A particularly innovative contribution in POPE is the creation of a mathematical models ontology, which represents mathematical models, as well as their underlying assumptions. This framework separates the declarative and procedural components of mathematical models creation, manipulation, and solution. The cover picture of this journal shows the overall architecture of POPE, which integrates formal models of data and information with mathematical models and guidelines to support automated decision-making. Figure 1 shows a more detailed view of some of the ontological relationships in POPE.

The utility of such an environment might be illustrated with an example. Consider a typical problem scenario in pharmaceutical manufacturing, one where the recent batch of tablets have failed the dissolution test — a critical test where one evaluates how well a sample tablet dissolves in 250 mL of buffer. Current approaches to diagnosing and resolving this important product quality problem requires a detailed and thorough investigation, comparing current batch records with those of previous batches, proposing failure hypotheses, analyzing and checking hundreds to thousands of pages of documentation on raw material properties, process operating conditions, design specifications, modeling assumptions (made a few years ago when the models were originally developed for this manufacturing process by an engineer who is no longer with the company), and even the original laboratory data on dissolution (from experiments conducted several years ago with data recorded as entries in hard to read laboratory notebooks). Using a POPE-like ontological informatics system, such data, information, and models would be readily accessible through the ontological relationships for an operator or a design engineer to make the right decisions at the right time. Such a system would greatly empower human decision-making in real-time by screening vast quantities of data and information to find better solutions, thereby reducing the time to complete a complex reasoning task.

Let us now summarize recent progress in some application domains of interest to chemical engineers.

Computational chemistry

There have been several attempts in the chemistry domain. For instance, GridChem (2005) is a Java desktop application that includes a client, a grid middleware server and a set of distributed, high-end computational resources. Computational

chemistry tools are connected, through the GridChem client portal that undertakes job management, to a middleware server that handles data management. The middleware server is connected to computational grids like GLOBUS, Condor and MyProxy. Both the middleware server and GridChem client portal are connected to long-term storage.

The Collaboratory for Chemical Kinetics (ChemSeer, 2006) provides a web environment which includes a database of chemistry documents, tools for search like Tableseer and Chemistry Entity Search, as well as access to a Globus grid environment for grid computation. NorthWest Chemistry (NWChem) (Kendall et al., 2000) is another important software system for computational chemistry on a grid.

Molecular products design and discovery

Molecular products design deals with the important and difficult problem of discovering and designing new materials and formulations with desired properties. This encompasses a wide variety of products such as fuel and oil additives, polymeric composites, rubber compounds, paints and varnishes, catalysts, etc. Recent research efforts are beginning to address the informatics and multiscale modeling challenges (de Pablo, 2005) in this domain. One such attempt is the multiscale model-based informatics framework called *Discovery Informatics* (Caruthers et al., 2003). The discovery informatics framework has led to the successful development of automated, rational, materials design systems in several industrial applications, such as gasoline additives (Sundaram et al., 2001), formulated rubbers (Ghosh et al., 2000), and catalyst design (Katare et al., 2004). As noted above, the richness and complexity of the underlying chemistries in this domain has led to important advancements in knowledge representation, languages, compilers, molecular structure search engines, chemical entities extraction systems, and so on. The recent contribution by Dion Vlachos and his group (Prasad and Vlachos, 2008) is another example of the value of the informatics based approach in catalytic chemistry.

There is considerable literature in domains such as drug discovery, bioinformatics (Floudas, 2007; Moore and Maranas, 2004), etc., which we have not even touched due to space constraints. The modeling and informatics challenges in other application domains such as smart manufacturing plants (Davis, 2008; Edgar and Davis, 2008), enterprise-wide optimization (Grossmann, 2005), sustainable energy and environmental systems, pharmaceutical engineering, nanoscale engineering, engineering virtual organizations (pharmaHUB, 2008) etc., require cyberinfrastructure based solutions.

Summary

Driven by powerful convergent forces, the coming data deluge poses unprecedented challenges and opportunities in modeling and informatics fronts. While computer scientists and information technologists can help, the demands imposed by the chemical engineering domain-specific knowledge and constraints, which are unlikely to be understood and appreciated by outsiders, make it clear that only chemical engineers can address these challenges— in particular, these need to be addressed by the process systems engineering community as it is the one best positioned to do so. However, addressing these would require discarding some traditional misconceptions about informatics and non-DAE based modeling methodologies, and fostering innovative approaches toward a wider class of knowledge modeling. Great opportunities for making *field defining* intellectual contributions await us in inventing chemical engineering domain-specific cyberinfrastructure components such as ontologies, compilers, molecular structure and semantic search engines, chemical entities extraction systems, languages for modeling chemical reaction pathways, modeling and knowledge management environments, visualization, virtual organizations, cybersecurity systems, and so on. Naturally, all these also provide us with opportunities for new business ventures in high-end modeling and informatics products and services. This is a long, adventurous, and intellectually exciting journey that we have only barely begun, but progress in this will revolutionize all aspects of chemical engineering for years to come.

Acknowledgments

In preparing this perspective, the author has benefited from suggestions from Professors Pablo Debenedetti, Thomas Edgar, Ignacio Grossmann, Sangtae Kim, Rex Reklaitis, Warren Seider, and George Stephanopoulos, all of whom he gratefully acknowledges. The author is also grateful to his former and current students who have worked on various aspects of the cyberinfrastructure work cited in this perspective. Finally, the financial support from the National Science Foundation's Engineering Research Center on Structured Organic Particulate Systems and the Indiana 21st Century Science and Technology Fund are also gratefully acknowledged.

Literature Cited

Atkins Report, "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the Blue-Ribbon Advisory Panel on Cyberinfrastructure," *the National Science Foundation* (2003).

Balachandra, B. K., "Information Retrieval and Knowledge Management in Catalyst Chemistry Discovery Environments," PhD Thesis, Purdue University (2008).

Bañares-Alcántara, R., A. W. Westerberg, E. I. Ko, and M. D. Rychener, "DECADE: A Hybrid Expert System for Catalyst Selection. I: Expert Systems Considerations," *Comp. Chem. Eng.*, **11**(3), 265-277 (1987).

Braunschweig, B., E. Fraga, Z. Guessoum, W. Marquardt, O. Nadjemi, D. Paen, D. Piñol, P Roux, S. Sama, M. Serra, I. Stalker, and Yang A, "CAPE Web Services: The COGents Way," *Proceedings of the European Symposium on Computer Aided Process Engineering - 14*, A. P. Barbosa-Póvoa and H. Matos, Eds., Elsevier, 1021–1026 (2004).

Caruthers, J.M., J. A. Lauterbach, K. T. Thomson, V. Venkatasubramanian, C. M. Snively, A. Bhan, S. Katare, and G. Oskarsdottir, "Catalyst Design: Knowledge Extraction from High Throughput Experimentation." In: "Understanding Catalysis from a Fundamental Perspective: Past, Present, and Future," A. Bell, Che, M. and W. N. Delgass, Eds. (40th Anniversary Issue), *J. Catal.*, 216/1-2, 98-109 (2003).

ChemSeer (2006). Available at http://dirac.chem.psu.edu/index.htm

Davis, J. F., Smart Process Manufacturing Workshop Report, NSF Roadmap Development Workshop, April, Arlington, VA (2008).

de Pablo, J. J., "Molecular and Multiscale Modeling in Chemical Engineering - Current View and Future Perspectives," *AIChE J*., **51**(9) (2005).

Edgar, T. F., and J. F. Davis, "Smart Process Manufacturing - A Vision of the Future," *Ind. Eng. Chem. Res. Dev*., Centennial Issue (2008).

Floudas, C. A., "Computational Methods in Protein Structure Prediction", *Biotechnol. Bioeng*, **97**(2), 207–213 (2007).

Frenkel, A, H. Afsarmanesh, G. Eijkel, and L. O. Hertzberger, "Information Management for Material Science Applications in a Virtual Laboratory," *Lecture Notes in Computer Science*, **2113**, 165–174 (2001).

Gardner, S. P., "Ontologies and Semantic Data Integration," *Drug Discovery Today*, **10**(14), 1001–1008 (2005).

Ghosh, P., S. R. Katare, P. R. Patkar, J. M. Caruthers, V. Venkatasubramanian, and K. A. Walker, "Sulfur Vulcanization of Natural Rubber for Benzothiazole Accelerated Formulations: From Reaction Mechanisms to a Rational Kinetic Model," *Rubber Chem. Technol*., **76**(3), 592–693 (2003).

GridChem,*Computational Chemistry Grid* (2005). Available at https://www.gridchem.org/.

Grossmann, I. E., "Enterprise-wide Optimization: A New Frontier in Process Systems Engineering," *AIChE J*., **51**(7), 1846–1857 (2005).

Gruber, T. R., "A translation approach to portable ontology specification," *Knowledge Acquisition*, **5**(2), 199–220 (1993).

Hsu, S-H., B. B. Krishnamurthy, P. Rao, C. Zhao, S. Jagannathan, and V. Venkatasubramanian, "A Domain-specific Compiler Theory Based Framework for Automated Reaction Network Generation," *Comp Chem Eng*., **32**, 2455–2470 (2008).

Katare, S., J. M. Caruthers, W. N. Delgass, and V. Venkatasubramanian, "An Intelligent System for Reaction Kinetic Modeling and Catalyst Design," *Ind. Eng. Chem. Res. Dev*., **43**(14), 3484–3512 (2004).

Kendall, R. A., E. Aprà, D. E. Bernholdt, Eric, J. E. J. Bylaska, M. Dupuis, G. I. Fann, Harrison, J. Ju, J. A. Nichols, J. Nieplocha, T. P. Straatsma, T. L. Windus, and Wonget, "High Performance Computational Chemistry: An Overview of NWChem a Distributed Parallel Application," *Comp. Phys. Commun*., **128**, 260–283 (2000).

Kim, D, K. Jeong, K, Hwang, S. and Cho, KW, " X-SIGMA: XML based Simple data Integration system for Gathering, Managing, and Accessing Scientific Experimental Data in Grid Environments," *Proceedings of the Second IEEE International Conference on e-Science and Grid Computing (e-Science'06)* (2006).

McKenzie, P., S. Kiang, J. Tom, A. E. Rubin, and M. Futran, "Can Pharmaceutical Process Development Become High Tech?" *AIChE J*., **52**(12), 3990–3944 (2006).

Morbach., J, A. Yang, and W. Marquardt, "OntoCAPE-A large-Scale Ontology for Chemical Process Engineering," *Eng. Appl. Artif. Intell*., **20**(2), 147–161 (2007).

Nagl, M., and W. Marquardt, Collaborative and Distributed Chemical Engineering,Springer-Verlag Berlin, Heidelberg (2008).

Moore, G. L., and C.D. Maranas, "Computational Challenges in Combinatorial Library Design for Protein Engineering," *AIChE J*., **50**(2), 262–272 (2004).

OWL: *Web ontology language overview* (2004), Available at http://www.w3.org/TR/owl-features/.

Paszko, C., and C. Pugsley, "Considerations in Selecting a Laboratory Information Management System (LIMS)," *Am. Lab*., **9**, 38–42 (2000).

pharmaHUB (2008). Available at www.pharmahub.org.

Prasad, V., and D. Vlachos, "Multiscale Model and Informatics-Based Optimal Design of Experiments: Application to the Catalytic Decomposition of Ammonia on Ruthenium," *Ind. Eng. Chem. Res*., **47**(17), 6555–6567 (2008).

Prickett, S. E and M. L. Mavrovouniotis, "Construction of Complex Reaction Systems-I. Reaction Description Language," *Comp. Chem. Eng*., **21**(11), 1219–1235 (1997).

Schopfer, G., A. Yang, L. vonWedel., and W. Marquardt, "CHEOPS: A Tool-Integration Platform for Chemical Process Modeling and Simulation," *Int. J. Software Tools Technol. Trans*., **6**, 186–202 (2004).

Rudd, D. F., G. J. Powers, and J. J. Siirola, Process Synthesis, Prentice-Hall, Englewood Cliffs, NJ (1973).

Seider, W., "The Quantitative-Qualitative Dichotomy in Process Engineering," *CAST Communications*, **16**(1), (1993).

Stephanopoulos, G., J. Johnston, T. Kriticos, R. Lakshmanan, M. Mavrovouniotis, C. Siletti, "DESIGN-KIT: An Object-Oriented Environment for Process Engineering," *Comp Chem Eng*., **11**(6) (1987).

Stephanopoulos, G., G. Henning, and H. Leone, "MODELLA. A Modeling Language for Process Engineering-I. The Formal Framework," *Comp. Chem. Eng*., **14**(8), 813–846, 1990.

Stephanopoulos, G., "Knowledge, Computers, and Chemical Engineering: A Critical Synthesis," *CAST Communications*, **17**(1) (1994).

Sundaram, A., P. Ghosh, J. M. Caruthers, and V. Venkatasubramanian, "Design of Fuel Additives Using Neural Networks and Evolutionary Algorithms," *AIChE J*., **47**(6), 1387–1406, 2001.

Ungar, L., and V. Venkatasubramanian, *Advanced Knowledge Representation*, CACHE Monographs Series, CACHE, Austin, Texas (1988).

Venkatasubramanian, V., R. Rengaswamy, K. Yin, and S. N. Kavuri, "A Review of Process Fault Detection and Diagnosis - Part I: Quantitative Model-Based Methods," *Comp. Chem. Eng*., **27**(3), 293–311 (2003).

Venkatasubramanian, V., C. Zhao, G. Joglekar, A. Jain, L. Hailemariam, P. Sureshbabu, P. Akkisetti, K. Morris, and G. V. Reklaitis, "Ontological Informatics Infrastructure for Chemical Product Design and Process Development," *Comp. Chem. Eng*. **30**(10-12), 1482–1496 (2006).