

# PROCESS INFORMATICS—A NEW PARADIGM FOR BUILDING COMPLEX CHEMICAL REACTION MODELS

## Scope of Process Informatics

*Process Informatics* is a data-centric approach to developing predictive models for chemical reaction systems. It will be dealing with all aspects of integration of pertinent data of complex systems (industrial processes and natural phenomena) whose complexity originates from *chemical reaction networks*. The primary goal of the process informatics will be information gathering, validation, and transformation into useable form. The latter will include development of *predictive* (numerical/computer) models with quantified degrees of reliability. While such a scope may be applicable to any process, the immediate focus of Process Informatics will be on *chemically-based* processes, as they currently provide unique challenges and opportunities.

## Motivation

Chemical reaction models will never (for the foreseeable future, or ever) be complete. Yet, we do have large amounts of information and methods needed to produce such models. The problem is that these data are scattered over different sources and are not properly evaluated. Most importantly, these data cannot be applied directly to practical problems—they have to be “transformed” into useful models. Such models, however, cannot be created by simple “compilation” of the data. Chemical reaction model building is a time-consuming activity that requires expert knowledge. At present, most treat this activity as an “art” (or “religion”). The goal we pose, in short, is to convert such model building into science, automate the methodology, and make the results available in a prompt and convenient form for the user.

Immediate needs for predictive reaction models presently exist in combustion engineering, the petrochemical industry, and pharmaceuticals. (Around the corner, as quantity and quality of information in biological fields is increasing, is the need for predictive dynamic models of biological systems.) As computers become more powerful and more readily accessible to industry, the industrial interest in process simulation is continuously growing. While a decade ago, it would be a great deal to get a letter of endorsement from an industrial person in support of development of chemical reaction models for his/her industry needs, today they are pleading for our help. No such organized activity currently exists in the USA or elsewhere.

## Vision of Process Informatics

The Process Informatics infrastructure, referred to it as *Process Informatics Model* (PrIME), will have two principal components: a *database* and a collection of *tools*.

The Process Informatics Database will be actual or virtual and will represent the complete set of knowledge available in a given field. For instance, in the field of combustion, it will contain experimental data, on both combustion systems and on elementary reactions, collected in flow reactors, shock tubes, flames etc, molecular properties determined from quantum chemical calculations, reaction rates obtained from reaction-rate theories, and so on. The database will contain several hierarchical “layers” of information, for example, molecular species, elementary reactions, and reaction models.

The Process Informatics Tools will be of two general kinds, those enabling the collection, transfer, organization, display, and mining of the data as well as enabling the scientific collaboratory among all participants—i.e., computer science tools, and those enabling processing and analysis of the data along with assembly of the data into models—i.e., scientific and

numerical tools. The latter represents a shift in the paradigm of the scientific method—building reaction models in a consistent and systematic way by a process which incorporates all the data available and includes all members of the scientific community. In other words, *the scientific community builds the Process Informatics System and Process Informatics builds the community*.

To further understand the new paradigm, let us consider the two principal customers of the Process Informatics System: the *data provider* and the *model user*. During the development stage, there will also be a *model builder*, whose role will eventually be automated—and providing the means for this automation should be the PrIME objective.

A Data Provider (Experimenter, Theorist) makes a request to PrIME to deposit her/his new observations or new computational results. The PrIME protocol assures completeness of the data submission, i.e., that all required details about the experiment (apparatus, experimental conditions, uncertainties, local data processing, etc.) or calculations (method, inherent assumptions, uncertainties, local data processing, etc.). The deposited data are immediately analyzed for consistency with the PrIME database and the results are reported both to the data provider itself and to the *scientific council* (see below). Upon approval of the council, the database is modified (thus, the council members are users too; but only they will have the permission to modify the database). In other words, the database will be “liquid” and will be continuously modified as a result of the PrIME activity, and these modifications will be documented: who, why, when... The data provider will receive a feedback of whether her/his submission is accepted or rejected. If rejected a dialog will ensue, with comments on reasons, together with suggestions of what needs to be re-measured, or recalculated.

A Model User (design engineer, CFD researcher) requests a kinetic model (or a simulation with such a model). S/he specifies the conditions of interests, the desired level of accuracy, and the mathematic form of the model (detailed, reduced, parameterized, etc). PrIME checks if such model already exists, and if not generates it. It may report to the requester that some data are missing and ask if s/he is interested in generating the missing data and at what level of theory; PrIME may offer “advice” at each stage of the dialog. If an actual numerical simulation is requested, PrIME will execute such a simulation or a set of simulations, possibly post-process the results, and return the results in the requested form (tables, diagrams, etc). PrIME may store such models, or some of them, but in principle, the generation of models will be done on-the-fly and custom-built.

Of course, every requester will be able to peruse the entire database: thermochemistry, kinetics, transport, reactions, mechanisms, experimental profiles, etc., and their respective sources; view the history of data deposition and modification: how the data were changed, why, when, by whom; examine the performance of the current data; e.g., the level of prediction of the current model against all or part of the experimental data; retrieve data; perform some analysis of the data interactively, e.g., compute sensitivity; and participate by submitting her/his comment to PrIME.

Another user of PrIME might be a project manager (a member of the Scientific Council, or a researcher) who wants to know whether a proposed experiment/calculation will improve the current database, and by how much. S/he can play a “what-if” scenario with the envisioned experiment or calculation (Note: just adding another data point does not necessarily reduce the model uncertainty—revealing the inherent correlations in the data is one of the objectives of the Process Informatics). A similar question can be posed differently: what needs to be done to improve the predictability of a given model? Repeat old experiments? Under what conditions? New experiments? Which ones? New Calculations? What level of local error must be maintained to accomplish the stated goal?

The goal of Process Informatics is to enable and sustain all of the above functionality (and more). In other words, the goal is not merely a collection of tools, but a shift in the paradigm of the scientific process: building targeted knowledge by the entire community and providing the wealth of information in its entirety to every user. To attain the vision described above will require the following: a) creation of enabling software infrastructure; b) development and implementation of scientific methods taking the full advantage of the approach; and c) establishment of a new paradigm of scientific collaboration for data collection, evaluation, and utilization.

### **Road to Process Informatics**

Development of the above vision requires a concrete system. It is proposed to consider reaction chemistry of combustion as the initial system. The reasons for this choice are as follows. Nearly all of the energy currently used in the industrialized world comes from burning fossil fuels and chemistry is the essence of combustion systems, from internal combustion engines to gas turbines. Knowledge of the chemical mechanism is at the center of device design to limit combustion-generated environmental pollution. Societal demands for cleaner and more efficient combustion are rapidly bringing the chemical aspects of combustion processes to the forefront.

Combustion systems are represented by complex dynamic models that serve as input for the design of equipment so as to maximize the efficiency of fuel utilization while minimizing unburned hydrocarbons, greenhouse gases and pollutants such as nitrogen oxides, sulfur oxides, soot or air toxics. The complexity of these models originates from high dimensionality, nonlinearity of the underlying differential equations, multi-dimensional correlations, and multifaceted dependence on temperature and pressure. This complexity is beyond simple analysis. At the same time, because of the important societal dependence on the knowledge of combustion processes, decades of research have been devoted worldwide to the subject matter. As a result, large amount of knowledge has accumulated and the level of fundamental theory and numerical solution has been advanced to stage of practical application.

However, the history of combustion chemistry model development, unfortunately, has been sporadic and piecemeal. Models are often constructed from a single investigation and have little chance of allowing reliable extrapolations to other conditions. Attempts to correct such individual models lead to a larger number of models, each failing in one respect or another.

Thus, although specific properties of a combustion process, such as ignition, flame speed, temperature, and both desirable and undesirable combustion products, are subject to calculation through the model, accuracy is severely limited by the reliability of the input rate parameters. A consistent, reliable set of values for these input parameters is best obtained by a systematic constrained optimization, based on an evaluated database of elementary reaction rate coefficients. This procedure adjusts key input rate parameters within their error ranges to obtain the best self-consistent model predictions of the experimental data.

A first effort at developing the new paradigm for dynamic models of complex chemical systems has been demonstrated with a quantitative chemical model for natural gas combustion, "GRI-Mech". That model was optimized using the method of "Solution Mapping". The approach starts with the proposition that although measured combustion properties, such as ignition, temperature, flame speed and concentrations of species are predictable by models containing the chemistry of many basic reactions, the combined errors in the input quantities and in the experiments require a systematic optimization procedure in order to produce a consistent reliable mechanism. Thus the result is more than a collection of individual reaction pathways, but rather an integrated model that accounts for the uncertainties.

This natural gas combustion mechanism, while an important first step, had several shortcomings: it was limited to a small number of research groups, and did not have the advantage of the tools needed to handle large amounts of information. Hence, as the next step of building the Process Informatics paradigm, it is proposed to undertake, as a target, development of the chemical reaction mechanism for the combustion of natural gas, such that it will include now *all* the information available worldwide. Doing this will accomplish several objectives: first, it will build the necessary infrastructure and tools that will form the foundation of the PrIME collaboratory; second, it will create the most reliable natural-gas combustion model that will find an immediate practical use; and finally, and perhaps foremost among all, it will bring the community together, to a new paradigm of scientific method. Once this prototype has been tested and accepted by the community, the approach will be extended for general use by chemists, chemical engineers, atmospheric chemists, pharmacokineticists, and biologists.

### **Organizational Structure**

The success of the undertaking will also depend on organization and management of the collaboratory. Based on the experience gained from the GRI-Mech project, the organization of the PrIME collaboratory is envisioned as following.

To initiate the collaboratory, several working teams will be created, organized by subject and each containing representatives of different disciplines as necessary. For instance, a data assembly team will have physical scientists with the expertise in reaction kinetics along with computer scientists with the expertise in databases and communication.

The underlying principle of the collaboratory will be a periodic release (at least once per year or per 6 months) of “the best current model” via a dedicated web site.

Each of these teams will be small enough to enable effective working interactions on a daily basis via internet and video-conferencing.

The leadership of each team will be on a rotational basis, with a leader selected (or elected) based on the expertise mostly required for a given round of release.

Current leaders of all teams will form a management team, with a rotationally assigned leader.

The entire team will meet periodically, probably coinciding with or as a part of a professional and contractual meeting. Decisions on the direction of the PrIME collaboratory will be discussed and decided at these meetings.

### **Scientific Council**

The concept of the *Scientific Council*, introduced here, is perhaps the cornerstone of the entire proposal. Its mission is “quality control” of the knowledge buildup in the scientific community (i.e., buildup of the PrIME database), with, say, combustion chemistry as the initial focus. The Council membership will begin with a few experts, those involved in the establishment of PrIME, with the intention to encompass and engage the entire community as PrIME develops.

In many ways the Scientific Council is similar to a Data Evaluation Panel, an established practice today for database quality control. The difference—and hence the novelty underlying the shift in the scientific paradigm—is that the Council activity will be based on the analysis of the *entire knowledge* available in the field. It will be the goal of Process Informatics to develop tools and infrastructure to enable such operation of the Scientific Council. With such tools in place, the Council will be able to act responsibly, with the aid of objective numerical and statistical measures provided by PrIME, and to respond rapidly with a thorough documentation of Council’s analysis and decision.

The members of Scientific Council must have, among them, expertise in all areas pertinent to the scientific underpinning of PrIME, with a good understanding of the processes involved in generating the data inputs and the applications. For this purpose, the membership may be extended to include people outside of PrIME, those with the expert knowledge required for a given stage of PrIME development. Generalizing this further, the membership and the structure of the Scientific Council may and will change, as necessitated by the evolving PrIME needs. On the other hand, the continuous evolution of PrIME (development of the tools and expansion of the “audience”) will transform the role of Scientific Council from being an evaluation panel composed of a small group of experts to a facilitator of consensus building by the entire scientific community.

## **Team**

As this document is prepared, we have commitment in support of launching Process Informatics from research teams of the following organizations (with the contact person listed in parentheses):

*University of California at Berkeley, USA* (Michael Frenklach, [myf@me.berkeley.edu](mailto:myf@me.berkeley.edu))

*Stanford University, USA* (David Golden, [david.golden@stanford.edu](mailto:david.golden@stanford.edu))

*SRI International, USA* (Greg Smith, [smith@mplvax.sri.com](mailto:smith@mplvax.sri.com))

*Leeds University, UK* (Michael Pilling, [mikep@chem.leeds.ac.uk](mailto:mikep@chem.leeds.ac.uk))

*National Institute of Standards and Technology, USA* (Wing Tsang, [wing.tsang@nist.gov](mailto:wing.tsang@nist.gov))

*Argonne National Laboratory, USA* (Al Wagner, [wagner@anchim.chm.anl.gov](mailto:wagner@anchim.chm.anl.gov))

*University of Karlsruhe, Germany* (Horst Hippler, [horst.hippler@chemie.uni-karlsruhe.de](mailto:horst.hippler@chemie.uni-karlsruhe.de))

*University of Bielefeld, Germany* (Katharina Kohse-Höinghaus, [kkh@pci.uni-bielefeld.de](mailto:kkh@pci.uni-bielefeld.de))

*CNRS, France* (Philippe Dagaut, [dagaut@cnrs-orleans.fr](mailto:dagaut@cnrs-orleans.fr))

*Lawrence Livermore National Laboratory, USA* (Charlie Westbrook) [westbrook1@llnl.gov](mailto:westbrook1@llnl.gov))

*Technical University of Denmark* (Peter Glarborg, [pgl@olivia.kt.dtu.dk](mailto:pgl@olivia.kt.dtu.dk))

*Princeton University, USA* (Fred Dryer, [fdryer@phoenix.Princeton.EDU](mailto:fdryer@phoenix.Princeton.EDU))

*National Institute of Statistical Sciences, USA* (Jerry Sacks, [sacks@niss.org](mailto:sacks@niss.org))