An Ontology-Based Information Management System for Pharmaceutical Product Development

Leaelaf Hailemariam[1], *Chunhua Zhao*[2], *Girish Joglekar*[1], *David Whittinghill*[3], *Ankur Jain*[1], *Venkat Venkatasubramanian*[1], *Gintaras V. Reklaitis*[1], *Kenneth R. Morris*[4], *and Prabir K. Basu*[5].

(1) *School of Chemical Engineering, Purdue University, Forney Hall of Chemical Engineering, 480 Stadium Mall Drive, West Lafayette, IN 47907-2100,*
(2) *Bayer Technology and Engineering (Shanghai) Co. Ltd, Shanghai, 201507, China,*
(3) *Information Technology at Purdue, Ernest C Young Hall, 302 Wood Street, West Lafayette, 47907,*
(4) *Industrial and Physical Pharmacy, Purdue University, 575 Stadium Mall Drive, West Lafayette, IN 47907-2088,*
(5) *Purdue University, 315 Hovde Hall, 610 Purdue Mall, West Lafayette, IN 47907-2040*

Abstract

The development of pharmaceutical products and processes involves laboratory scale, pilot plant scale and commercial scale manufacturing. Through these steps, data on synthesis routes, material properties, processing steps, and scale up parameters are collected and used. Recent advances in process analytical technologies resulted in a large volume of heterogeneous data, which requires a systematic model of the associated information for optimal use. In addition, software tools are used to support decision-making and process modeling during process development. However, each tool creates and uses information in a specific form, making connection difficult amongst the tools and allowing little interaction with experimental data. In this work, an integrated information management system based on structured information is presented. This system is built to explicitly specify important concepts like physical properties and experiments using ontologies. Based on the developed ontologies, an information access and repository system is developed to allow convenient access to the repository for both the user and software tool. The infrastructure, designed to be accessible from the web, would allow simple search and browse as well as complicated queries, which can only be efficiently answered through the use of semantics in the information. This structure is extended to consider integration with software tools through the definition of an application-agnostic medium. Such an integrated information infrastructure provides a systematic approach to managing, sharing and reusing information, and is expected to result in considerable reduction in pharmaceutical process development time and better quality assurance.

1. Introduction

Once a chemical entity, viable for its intended pharmaceutical application, is discovered, further development takes place in several integrated scales. These are the laboratory scale, pilot plant scale and commercial scales. At the laboratory scale, experiments are performed to characterize various synthesis routes and obtain process parameter values. Information generated at the laboratory scale can be used to improve manufacturing. At the pilot plant scale, studies are done to provide a detailed understanding of processing steps in the selected route and data needed for scale-up to commercial manufacturing is collected. At the commercial scale, information related to manufacturing is applied in debottlenecking and productivity improvements. Problems identified at the manufacturing scale are communicated to the laboratory scale to identify root causes, and provide ideas on how to avoid similar problems in the future.

In transitioning between these scales, several key activities have to be performed, such as, recipe synthesis, design and operation of the pilot plant, and scale-up to commercial production, as shown in Figure 1. The boxes indicate the major source of information for the given activities, with some overlap.

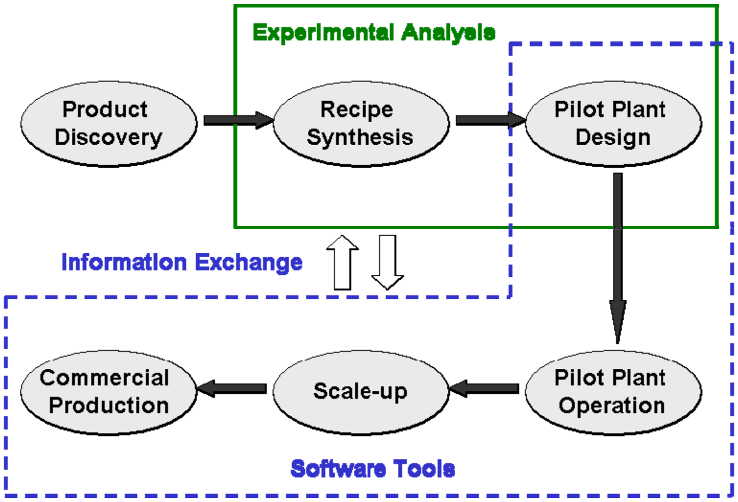


Figure 1: Tasks in pharmaceutical product development and information flow between them

Recipe synthesis and pilot plant design involve interaction with experimental data, for example, determination of material properties, while pilot plant design, operation and scale-up require use of software tools (e.g. ASPEN PLUS 2004 ®) which can also support decision making and process modeling in product development. In addition, there is information flow between these tasks; the software tools would make use of some experimental data (e.g. ASPEN PROPERTIES ®)and experimental analysis is assisted by software tools (e.g. computation of statistical trends). In addition, recent advances in process analytical technologies helped to provide a large volume of heterogeneous data, which requires systematic management for optimal use.

The integrated use and management of heterogeneous information present several challenges. Each software tool has a specific form of information creation and use, leading to islands of automation, which are difficult to connect and share functions among (Zhao et al., 2003). In addition, due to the lack of a versatile information model that can be used for data transfer, experimental data is likewise put in individual 'silos' as shown in Figure 2.

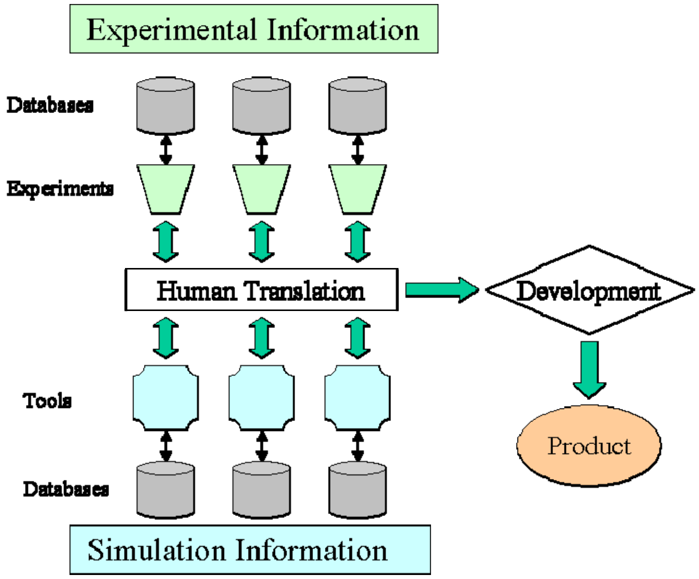


Figure 2: Use of information in a fragmented product development system

A large volume of heterogeneous data would require an efficient management system that allows for convenient access and retrieval as well as secure storage of information. The links between the pieces of available data (e.g. data from multiple experiments on the same sample) should be established in a systematic manner. Otherwise, pair-wise information translations would be required to cross these information divides, causing inefficiencies, uncertainties, costs, delays, and product quality concerns all along the product development pipeline. Thus, formal and explicit models of related information with well understood semantics (Schneider and Marquardt, 2002) are required. These information models should be easily accessible by humans and tools, and should provide a common understanding for information sharing

There had been several attempts to address these challenges, including the Multi Dimensional Object Oriented Model **MDOOM** (Burkett and Yang, 1995) and Process and Repository based Approach for Requirements Traceability/ Chemical Engineering **PROART/CE** (Jarke and Marquardt, 1996) for chemical process information integration associated with software tools. For laboratory information integration**,** one may cite **LIMS** (Paszko and Pugsley, 2000) and e-Lab Notebook (Zall, 2001). However, these solutions are frequently limited in their description of the relations between the different information entities and in their application to the pharmaceutical domain. The range of physical properties modeled was limited to fluid properties, while solid properties are of importance for pharmaceutical manufacture. In addition, integration of process information with experimental information was lacking. The objective of this research is to develop an integrated information management system that makes possible better data management, user view and edit capability of the data as well as efficient use of information in product development tools. This involves the development of an information structure, information storage systems, and interfaces for information exchange.

The remainder of the paper is organized as follows. In Section 2, the development of the information model is described. Section 3 includes the use of the developed information model for laboratory information management. Section 4 describes how the information model may be used to integrate experiments and tools. Conclusions and future directions are given in Section 5.

2. Development of the information model

Several approaches may be used to model product development information. However, the information models are usually in closed form and only provide a limited view of the information (Zhao et al., 2005). A promising approach is the representation of information models by a structured, formal and well understood semantics (Schneider and Marquardt, 2002), otherwise called *structured information*. Without semantics, it would be difficult to organize related information and systematically manage meta-data (information about information) and allow integration of product development tools (Bayer et al., 2000).

Ontologies, defined as “shared conceptualizations of a domain” (Gruber, 1993), are commonly used for modeling information in a formal and explicit manner. Compared to a database schema, which targets physical data independence, and an XML (eXtended Markup Language) schema, which targets document structure, an ontology targets agreed-upon and explicit semantics of information, and directly describes the graph structure consisting of the *concepts* and their relations.

In the pharmaceutical product development domain, several key *concepts* need to be defined. A central concept is the *material*, which represents substances and mixtures (which are characterized by pure substances and their compositions). A *material* has several physical and chemical *properties* (e.g. specific heat capacity), and can be involved in several *experiments* (e.g. Compression Tests). Conversely, each material property value is obtained from a *source of value,* which includes *experiments*, *literature* or *mathematical models*. The relationships between these concepts are shown in Figure 3.

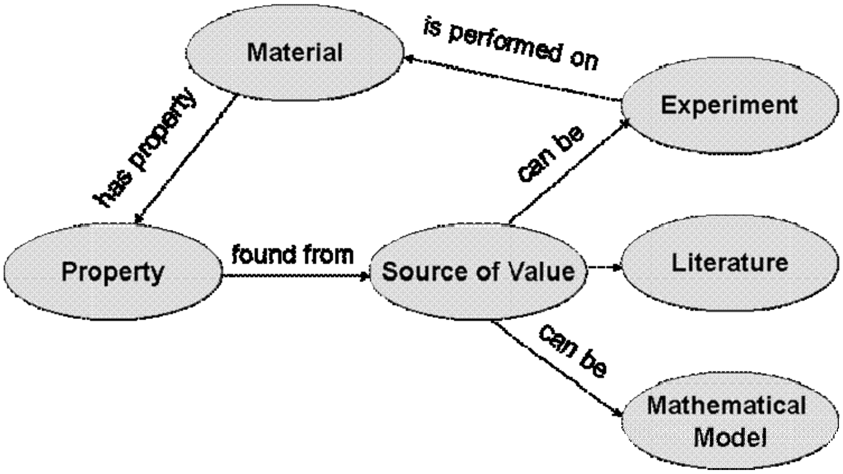
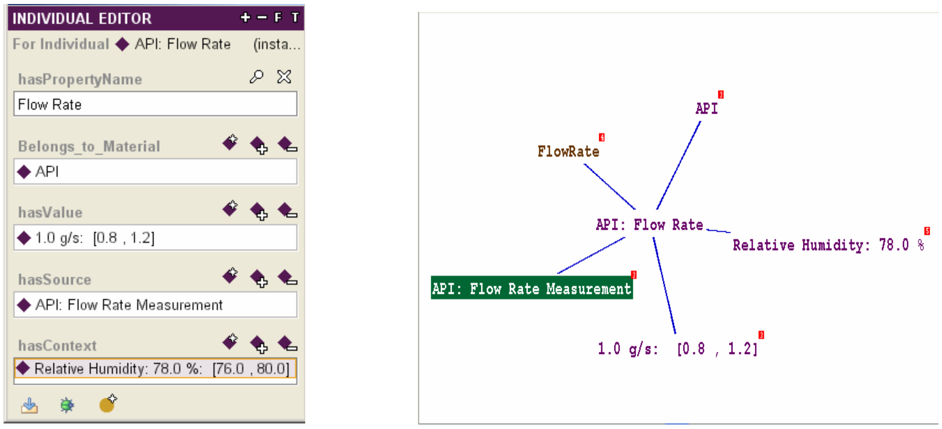


Figure 3: The relations between major concepts in pharmaceutical product development

The concepts defined above are described by a set of attributes. A *material* is described by its composition, phase, properties, roles (e.g. does it help flow of a powder of which it is a part?) and experiments it participated in. A *property* is represented by the *material* it belongs to, its *value*, the *source* of that value, the conditions under which that value was measured (called its *context*). The description of an *experiment* would include the materials involved, the property measured, the experimenter, location of the experiment, the date and time of experiment, the equipment used, the procedure followed and the experimental data obtained.



INDIVIDUAL EDITOR

For Individual ◆ API: Flow Rate (insta...

hasPropertyName: Flow Rate

Belongs_to_Material: ◆ API

hasValue: ◆ 1.0 g/s: [0.8 , 1.2]

hasSource: ◆ API: Flow Rate Measurement

hasContext: ◆ Relative Humidity: 78.0 %: [76.0 , 80.0]

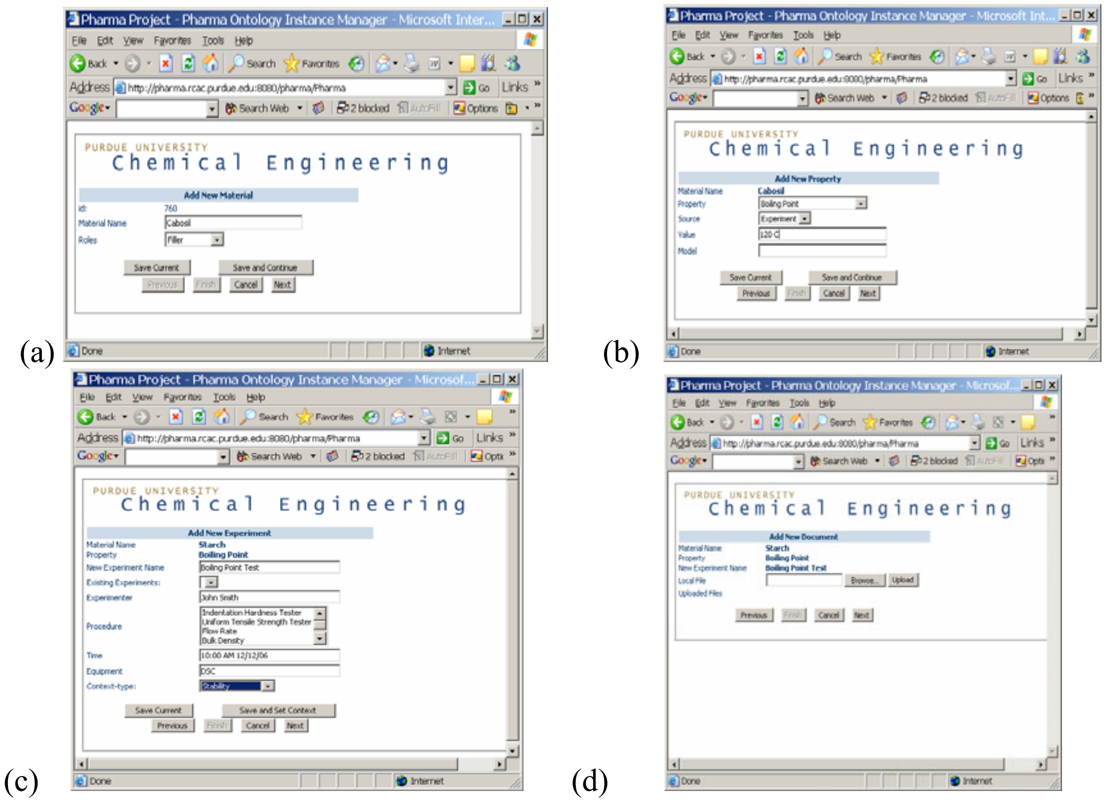
(a) (b)

Figure 4: (a) Protégé input Interface for creation of material properties (b) Graphical view of the relations between concepts

Based on explicit expressions of the relations between the concepts described above, *material, property* and *experiment* ontologies were developed. Ontologies were also developed for 'secondary' level concepts like value, source, context and other descriptors as necessary. Web Ontology Language (OWL, 2004), was used in this work to encode ontologies. The ontologies were developed in close collaboration with colleagues in the Department of Industrial and Physical Pharmacy at Purdue University with several revisions during development. The visualization tools provided by the ontology editor, Protégé (see URL) and the plug-ins in the editor including the view of class hierarchy, the graph view (shown in Figure 4) were found to be very convenient in the collaboration.

3. Development of an integrated information infrastructure

An effective information management system should manage the large amount of experimental information and knowledge generated during experimental analysis and facilitate efficient decision-making. Users of the system should be able to conveniently upload, browse and search the available data. This requires the information management system to have the following basic components: an information model, an information repository and interfaces that can communicate with the user and allow the user to perform such tasks as search and browse. To that end, the developed ontology has been used as the foundation for the information repository (Zhao et al., 2005). Folder structures to store experimental files are created based on the concept hierarchy defined in the ontology. The system allows transition between related instances and search by keyword as well by hierarchy. A web-based interface to the information repository (as shown in Figure 5-6) is currently being developed using the Java programming language so that users can access, view, and modify the information.


(a)
Pharma Project - Pharma Ontology Instance Manager - Microsoft Inter...
File Edit View Favorites Tools Help
Back Search Favorites
Address http://pharma.rcac.purdue.edu:8080/pharma/Pharma
Go Links
Google Search Web 2 blocked AutoFill Options
PURDUE UNIVERSITY
Chemical Engineering
Save Current
Save and Continue
Previous
Finish
Cancel
Next
Done
Internet

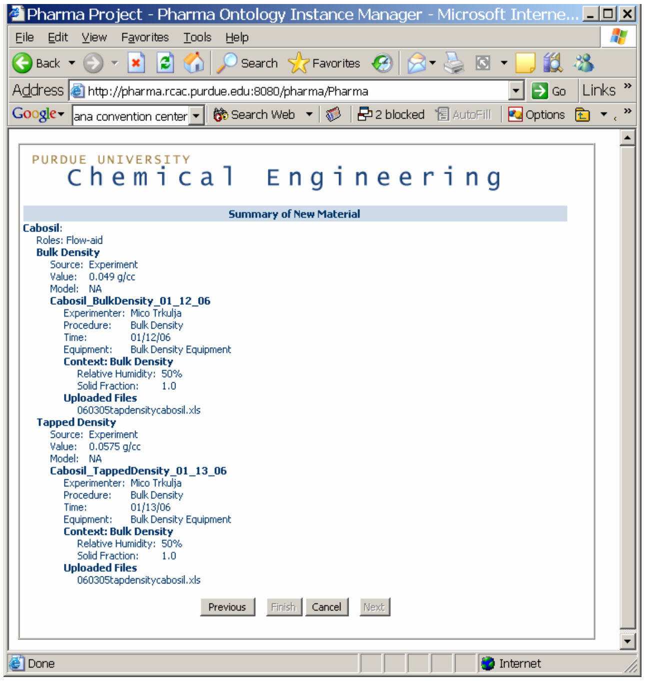
Add New Material
id: 760
Material Name Cabosil
Roles Filler

Add New Property
Material Name Cabosil
Property Boiling Point
Source Experiment
Value 120 C
Model

Add New Experiment
Material Name Starch
Property Boiling Point
New Experiment Name Boiling Point Test
Existing Experiments:
Experimenter John Smith
Procedure Indentation Hardness Tester
Uniform Tensile Strength Tester
Flow Rate
Bulk Density
Time 10:00 AM 12/12/06
Equipment DSC
Context-type: Stability

Add New Document
Material Name Starch
Property Boiling Point
New Experiment Name Boiling Point Test
Local File
Uploaded Files

Figure 5: Upload screens in the User Interface for a material (a), a property (b), an experiment (c) and an experimental file (d)


Pharma Project - Pharma Ontology Instance Manager - Microsoft Interne...
File Edit View Favorites Tools Help
Back Search Favorites
Address http://pharma.rcac.purdue.edu:8080/pharma/Pharma
Go Links
Google ana convention center Search Web 2 blocked AutoFill Options
PURDUE UNIVERSITY
Chemical Engineering
Previous
Finish
Cancel
Next
Done
Internet

Summary of New Material

Cabosil:
Roles: Flow-aid
Bulk Density
Source: Experiment
Value: 0.049 g/cc
Model: NA
Cabosil_BulkDensity_01_12_06
Experimenter: Mico Trkulja
Procedure: Bulk Density
Time: 01/12/06
Equipment: Bulk Density Equipment
Context: Bulk Density
Relative Humidity: 50%
Solid Fraction: 1.0
Uploaded Files
060305tapdensitycabosil.xls
Tapped Density
Source: Experiment
Value: 0.0575 g/cc
Model: NA
Cabosil_TappedDensity_01_13_06
Experimenter: Mico Trkulja
Procedure: Bulk Density
Time: 01/13/06
Equipment: Bulk Density Equipment
Context: Bulk Density
Relative Humidity: 50%
Solid Fraction: 1.0
Uploaded Files
060305tapdensitycabosil.xls

Figure 6: Summary of the known information on a given material

The semantic richness of the information can provide many benefits. The information infrastructure described above makes possible the effective management of experimental files, which may exist in different forms (spreadsheets, movies etc.) with non-descriptive names and folder locations. Additionally, information can be queried based on user criteria; for example, a user may like to find out all the experiments that have been done on micromeritics (properties related to the nature of surfaces making up a solid: Brittain (1995)), which are potentially affected by the relative humidity. In a typical relational database this would be difficult to implement. Micromeritics is defined as the super-class of a set of properties. Hence, a keyword-based search would not be able to navigate to the pertinent subclasses. In addition, the effect of relative humidity is observed only in some of the micromeritic properties. Such a query would be difficult to process without semantically rich information. Once the properties of interest are identified, one would then have to identify the relevant experiments. The material, property and experiment ontologies define the relations needed and thus can provide semantic search capability. Such search features are currently under development.

4. Integration of experiments and tools

Experiments and software tools are both integral parts of pharmaceutical product development. With an explicit information model, the tools and experiment environments can be seamlessly integrated, thereby considerably accelerating product development. A schematic representation of the envisaged integrated information environment is shown in Figure 7.

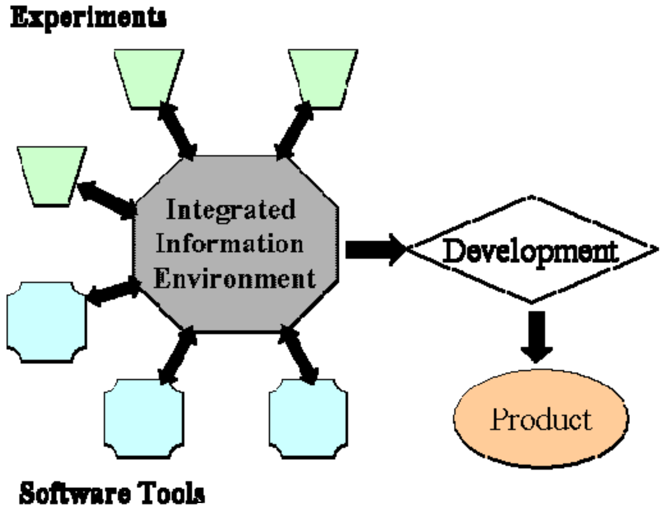


Figure 7: The integrated information management environment

There are several ways in which tools and experiments may interact, including an experimental device accessing an external software tool for statistical computation e.g. computation of standard deviations of readings. The use of a software tool to determine an important physical property using experimental data is described. A good example from the pharmaceutical domain is the computation of the slope of the Heckel Plot (a description of the compressibility of a powder: Heckel (1961)) from experimental data points. The data points would be entered into the system through a suitable interface (manually or automatically), which, using the defined ontology, assigns the proper structure to the data.

This structured data is then sent to an external tool, which can read the structured data through an interface, perform the computations and release the results (in the process generating information) through its interface to the data repository for later access. A schematic representation of the steps is shown in Figure 8.

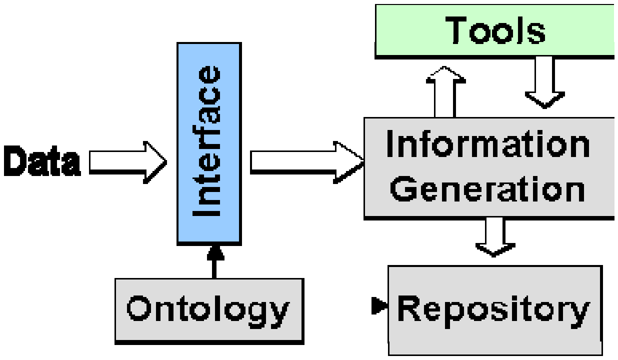


Figure 8: Flow of data between a tool and an experiment in an integrated information management system

The advantage of a platform-agnostic information model becomes apparent when one remembers that for n entities each with its own information model, it would required n^2 translators whereas with the entities all translating to the same information model, at most 2n translators would be required; an order of magnitude savings in computational effort. An added advantage is that software tools would not be required to 'open' up their internal data structure as the inputs would be configured to the input information model they accept and the output information model translated from their original output model.

5. Conclusions and future work

Efficient transfer, management and use of information during product development are crucial for the continued competitiveness of a pharmaceutical company. Effective management of heterogeneous experimental data as well as integration between the experimental work and software tools used to assist product development were key challenges to be resolved. Several attempts had been made to provide an integrated environment for information management across manufacturing processes but they are mostly limited in scope and not directly applicable to the domain at hand. To meet these challenges, an integrated information management system, with constituent units included an information model, information repository and user interface. The development of the information model and the utility of the repository and user interfaces were described.

This infrastructure, named POPEI3 (Purdue Ontology for Pharmaceutical Engineering: Integrated Information Infrastructure) is still in its early developmental stage and is expected to benefit from further research in ontology development, construction of interfaces with external databases and experimental analysis of materials. At its conclusion, POPEI3 is expected to benefit not only the pharmaceutical industry but through its development of a paradigm for handling information in product development, related areas like chemical product and process development.

Acknowledgement: The authors would like to acknowledge Jonathan Gortat, Dr. Steef Boerrigter, Dr. Nicholas Poendaev, Mico Trkulja, Dr. Carl Wassgren and Vince Hoon for their assistance.

References

Bayer B., Schneider R., Marquardt W. (2000), Integration of data models for process design-first steps and experiences, Computers and Chemical Engineering 24 599-605

Brittain H.G. (1995), Physical characterization of Pharmaceutical Solids, Marcel Dekker Inc. New York, NY

Burkett W.C., Yang Y. (1995), The STEP integration information architecture, Engineering with Computers 11 136-144

Grauer Z. (2003), Laboratory Information Management Systems and Traceability of Quality Systems, American Laboratory, 9, 15.

Gruber T.R. (1993), A Translation Approach to Portable Ontology Specification, Knowledge Acquisition, 5, 2, 199.

Heckel R.W. (1961), An Analysis of Powder Compaction Phenomena
Transactions of the Metallurgical Society Of AIME 221 5 1001-1008

Holly F. (2002), XML and Pharmaceuticals: Regulatory and non-regulatory applications, Proceedings of the XML Conference and Exposition Dec 8-13 Baltimore MD 1-8

Jarke M. , Marquardt W. (1996), Design and Evaluation of Computer–Aided Process Modeling Tools. In: Davis, J.F., Stephanopoulos, G., Venkatasubramanian, V.: Intelligent Systems in Process Engineering, AIChE Symposium Series 312 92 97–109

OWL (2004). Web Ontology Language Overview, http://www.w3.org/TR/owl-features/

Paszko C., Pugsley C. (2000). Considerations in Selecting a Laboratory Information Management System (LIMS), American Laboratory 9 38-42

Protégé (Version 3.1). http://protégé.stanford.edu.

Schneider R. , Marquardt W.(2002). Information Technology Support in the chemical process design life cycle, Chemical Engineering Science 57 1763-1792

Zall M. (2001), The Nascent Paperless Laboratory, Chemical Innovation, 31, 2-9

Zhao C., Bhushan M, Venkatasubramanian V. (2003), Roles of Ontology in Automated Process Safety Analysis, Proceedings of ESCAPE 13.

Zhao C., Joglekar G., Jain A., Venkatasubramanian V. and Reklaitis G. V. (2005), Pharmaceutical Informatics: A Novel Paradigm for Pharmaceutical Product Development and Manufacture, Proceedings of ESCAPE15