QSAR & Combinatorial Science QCS

Integrated Informatics in Life and Materials Sciences: An Oxymoron?

François Gilardoni*, Vasa Curcin[a], Kanishka Karunanayake, Jonas Norgaard, and Yike Guo

Inforsense Ltd, 48 Princes Gardens, London SW7 2PE, UK, E-mail: fgilardoni@inforsense.com, Head of Cheminformatics; Phone: +44 (0) 207-594-6817; Fax: +44 (0) 207-594-6836
[a] Department of Computing, Imperial College, London, SW 72A2, UK

Keywords: Knowledge Management and Discovery, Integrated Informatics, Data Mining, Workflow, QSAR, QSPR, Enterprise Discovery Planning, Portable Chemistry, Webservices, XML, SOAP, Web Portal, Text Mining, Cheminformatics, Bioinformatics

Received: 08. 10. 2004; Accepted: 01. 12. 2004

Abstract

The pharmaceutical and chemical industries are facing significant internal and external pressure to boost the experimental efficiency and effectiveness by cutting the direct research costs and reducing the time to market for new sustainable products. Other key issues to secure a competitive return on investment are to enable the rescue of stalled product development projects, abort failing projects early, enhance collaborative multidisciplinary ventures, and ensure an effectual risk management and cost savings through safety testing and failure analysis. Typically, modern organizations rely on a variety of informatics solutions from well-established software vendors. These typically operate on different platforms and are controlled by different management systems using different data types and proprietary formats. Maintaining, integrating, updating and monitoring these powerful, but disparate, ensembles of tools, are an intricate and expensive operation. We present the concept of the Enterprise Discovery Planning (EDP) software platform design to facilitate the optimization of discovery activities at an enterprise level. We also briefly present how Inforsense addresses these issues.

1 Introduction

The pharmaceutical and chemical industries face an increasingly complex environment and are often searching for growth through mergers and acquisitions [1]. In light of decreasing numbers of blockbuster drugs and increasing development cost per drug, the industry as a whole is moving to cut costs [2]. Furthermore, the political and public [3, 4] pressure for controlling health care and drug costs, increased health care expectations [5], and more stringent regulatory environments [6] leading to longer approval periods causing a shorter effective patent life are compelling the industry to hunt for new areas for growth. This pressure in the pharmaceutical industry for stronger drug development pipelines and improved operational efficiency is a forceful driver for innovation and technology within the Information Technology (IT) sector [7].

The novel "-omics" era embraces new complex technologies and presents enormous informational, strategic and organizational challenges [8, 9]. For instance, genomics has led to new drugs and treatments, increased efficiency, greater use of intellectual property, new revenue models and enhanced distribution of products and processes. Technology will continue to pervade the life sciences industry and will change the way the industry operates [1, 10]. Where it was once focused primarily on science, the industry is now witnessing the impact of technology, primarily in the areas of product development and distribution. In every area of discovery and development, the pharmaceutical organization is forced to absorb techniques only months out of R&D and make decisions on how use to that data as part of their business practices [1, 7, 11]. Pharmaceutical and biotechnology executives have nowadays the intricate and perilous mission in assessing these new technologies that could make a crucial difference in the success and wealth of their organization. Even as IT and technology vendors continue to improve the utility of their products and services [12], they must sell them into a dynamic and highly fragmented marketplace.

Although mergers and acquisitions aim to reinforce competitiveness, they are ironically accountable for a great deal of fragmentation, ineffective communication and connectivity between the new departments because of the blend of legacy technologies, procedures and company culture. Therefore, an infrastructure that enables sharing data, information and knowledge within the organization is essential for biopharmaceutical companies hoping to capitalize quickly on new scientific discoveries. According

 DOI: 10.1002/qsar.200420056

to the International Data Corporation (IDC), by the year 2006, US$ 38 billion will be invested in IT in the life sciences sector [1]. Integration issues will drive a dramatic increase in spending by biopharmaceutical companies – IDC forecast $11 billion in 2004, see Figure 1. However, among the key barriers to more widespread adoption of the technologies that would help to shorten the drug discovery time line are the strict regulatory requirements under which the pharmaceutical industry operates [13]. Although most would agree that the US Food and Drug Administration (FDA) and similar regulatory agencies around the world serve a vital function, it is also clear that the regulatory environment tends to reinforce a conservative stance to the adoption of new processes and technologies. *De facto*, the acceptance and effectiveness of new technologies will drive IT spending in order to acquire, analyze, and integrate the disparate kinds of current tools dealing, for instance, with image processing, databases, data integration, storage and management, security, compliance, and standards. For instance, many informatics vendors currently provide training to the FDA free of charge in order to assure that discoveries made using their technologies can gain regulatory acceptance, and agency reviewers and pharmaceutical researchers will reach similar conclusions when reviewing their data.

Life sciences and biotechnology are widely recognized to be on the leading edge, together with the information technology sector, of the next wave of the knowledge-based economy, creating new opportunities for our societies and economies [14]. A revolution is taking place in the knowledge base of life sciences and biotechnology, opening up new applications in health care, agriculture and food production, environmental protection, as well as new scientific discoveries. This is happening globally and impacts R&D projects that today use nanotechnology, biology, chemistry, automation, and informatics [15]. There are a constellation of software packages for knowledge management and discovery available, but none encapsulate the process from inception to delivery of modern multidisciplinary R&D projects. We will unveil some issues facilitating access to knowledge through integrated informatics applied to life science and present how Inforsense addresses these issues with its workflow engine for scientific applications.

2 Integrated Informatics in Life Sciences and Materials Science

2.1 The Data and Process Flow in R&D

Modern R&D depends very much on data generated by high-throughput screening and experimentation for the identification of new chemical entities as drug candidates, new materials or catalysts [16]. In both materials and life sciences, computational methods are typically involved for

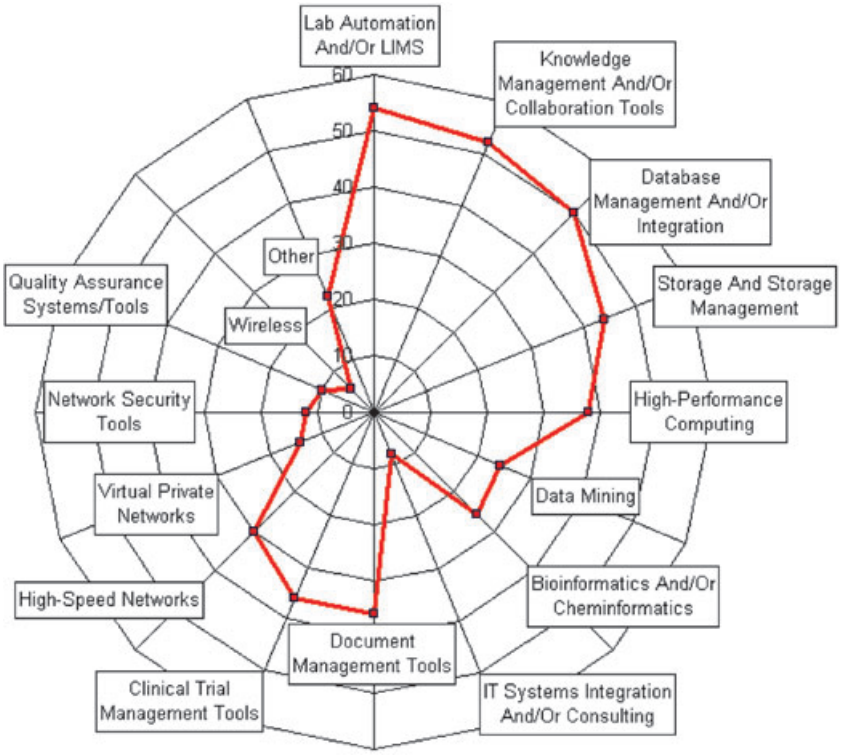

Figure 1. The ability to share data across a broad base of researchers is essential for pharmaceutical and chemical companies hoping to capitalize on the new scientific discoveries. Integration issues will drive dramatic increases in information technology spending by these industries stimulated by the great deal of fragmentation, and the little communication and connectivity between departments.

rapid screening and experimental design prior to the move to small scale laboratory methods first, then scaled-up for large-scale production [17]. This rising reliance on informatics generates an ever-increasing and overwhelming amount of data, ranging from chemical structures, biological sequences and assay data to related text information complemented with experimental data and setups. The recurrent and old paradigm is to transform data into knowledge, but discovery processes are rarely secured in a tangible and consistent structure, and are often lost. Henceforward, we will refer to "workflows" to describe data and process centric operations.

The R&D roadmap in drug discovery and materials science is analogous (Figure 2), albeit the pharmaceutical industry is constrained by a stringent regulatory framework. It usually starts by assessing and mining in-house and publicly available data sources stored in different locations and disparate formats. Experimental planning, screening and optimization campaigns, data validation and analysis, visualization and reporting are the subsequent steps in the process. We will not discuss these issues here that are very well described elsewhere [16, 18]. The workflow can be reiterated several times and objectives and constraints reassessed accordingly. *A priori*, the discovery cycle to derive QSAR/QSPR models substantially differs whether this is a life or materials sciences framework. In the "-omic" world, data are very heterogeneous and discovery processes are still evolving rapidly as compared process chemistry and materials science area – this is a strong differentiator. To a large extent, both the semantic description of the problematic and the experimental techniques of each field are often dissimilar, but the underlying concept is alike. Often, the data-mining strategy and the corresponding statistical, descriptive and predictive tools are the same [19]. In life sciences, the strategy relies on validated QSAR models developed for the available series of biologically active compounds. In time, these collections of local models describing classes of compounds or reactions are linked with "-omics" models. The cycle starts with the numerical representation of chemical compounds with a set of descriptors or fingerprints. Eventually, QSAR models are built and all the relevant descriptors are then identified. Pre- or post-processing steps involve clustering techniques or multivariate analysis. Then, genetic algorithms can be em-

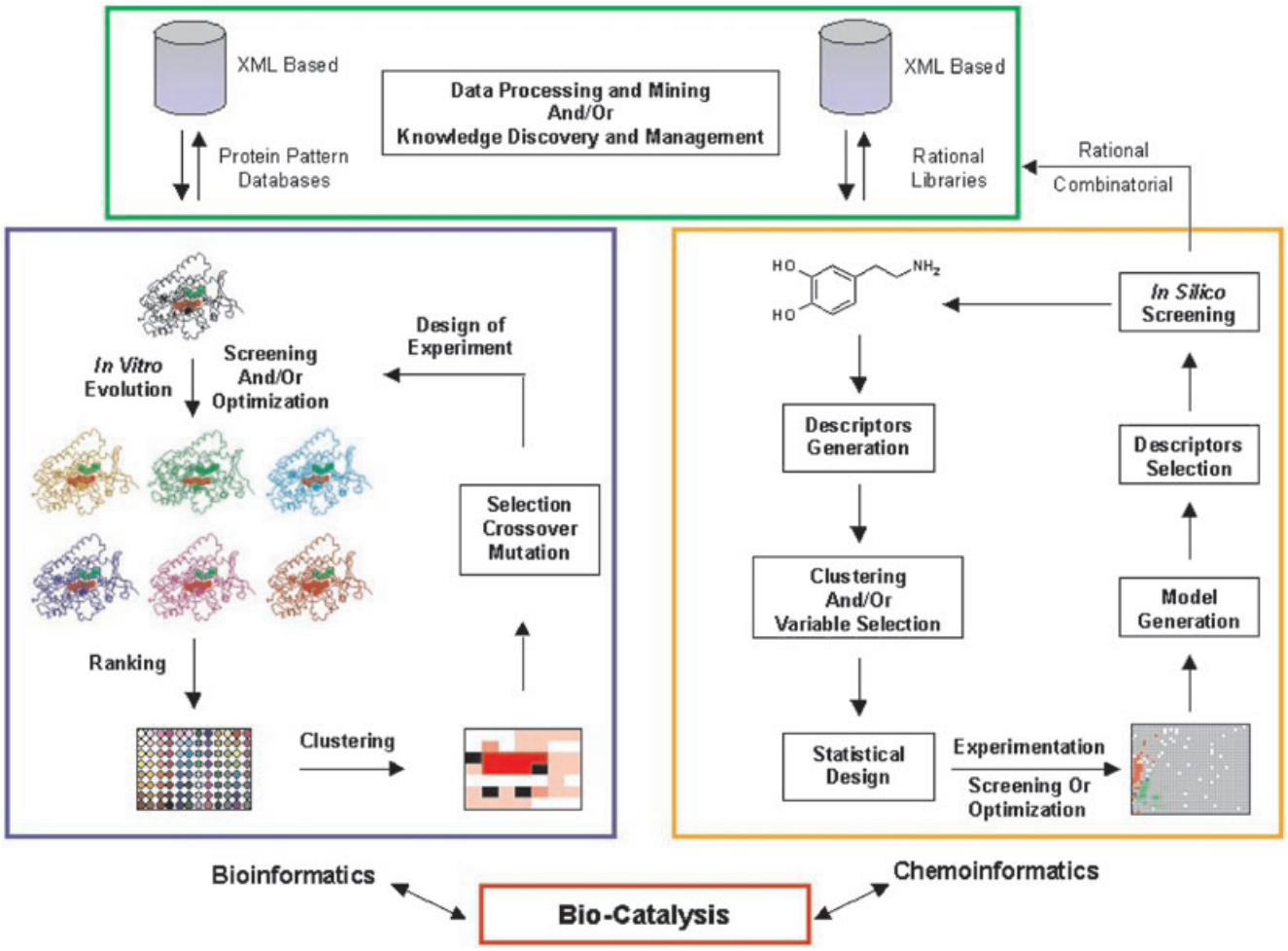

Figure 2. an example of a process outline that blends chemoinformatics and bioinformatics applied to biocatalysis.

ployed to design new structures with specific properties. In materials science, solids, like catalysts, are also described by a set of descriptors that are specific to these application areas. Variable selection and material optimization rely on identical techniques used in the life sciences environment.

Prior knowledge is critical to reduce the time to market and ensure a high return on investment. Beside the technological and scientific aspects to consider, tools for enterprise resources planning, business intelligence and strategic planning can be utilized to secure and forecast business assets [20]. For instance, identifying core competencies within the organization, or assessing risks in delivering a project. Modern R&D is a blend of science and technology driven by effectiveness and business rules.

This introduces the technical project definition with its experimental part, which could combine various branches of computational biology and molecular and materials science. The ability to facilitate collaborative operations, i.e., workflows, between specialized personnel and their informatics platforms is critical to the success of the project. From this perspective, web services hold the potential to address these concerns by allowing the use of on-demand processing power and best-of-breed components, in a collaborative and transparent fashion [21]. Webservices [22] are a modular collection of web-protocol-based applications that can be mixed and matched to provide business functionality through an Internet connection. These services represent black-box functionality that can be used and reused without regard to how the service is implemented. Webservices use standard Internet protocols such as HTTP, (eXtensible Markup Language) (XML) and SOAP to provide connectivity and interoperability between business units and eventually virtual organizations.

Automatic data-processing demands continue to grow rapidly due to increased use of *in-silico* tools, high-throughput platforms and laboratory automation [23]. For decades the traditional research process within the biopharmaceutical industry was a sequential operation where, after many months of target validation, the process would lead to assay development followed by high-throughput library screens for hits, and then on to lead optimization. This *modus operandi* is becoming obsolete with the introduction of parallel and designed experimentation.

The use of robots allows the life sciences industry to screen an ever-increasing number of compounds. At first, compounds for testing may be selected from the large, readily available collections of products accumulated over years of synthetic effort in industrial or academic research laboratories. A second step consists of synthesizing new and large combinatorial libraries, hitting the wall of a combinatorial explosion, having potentially billions of chemical compounds. Although modern laboratory automation and its informatics infrastructure have progressed dramatically during the last decades, they still cannot be handled by modern experimental platforms. The challenge is to select of a representative subset of samples to be experimented upon from billions of possible candidates. Data is certainly not Information. Besides this quandary, experimentation should be designed to deliver relevant and applicable information aimed to guide discovery in an effective and fast manner. This is next to impossible to achieve without the use of an efficient computational infrastructure, and without some prior knowledge of the researched domain.

An alternative to the combinatorial data explosion builds on an investigation of the greatest diversity of the experimental space in the least number of experiments to create a performance-based model [18]. This model delivers the highest density of information per experiment at higher speed and favors the transformation of information into knowledge. This methodology combines the advantages of clustering techniques, molecular modelling, statistical design, multivariate statistics and data visualization and mining. Eventually, a model capable of guiding discovery is built; it can correlate a collection of properties, or descriptors, of a protein, a biocatalyst, or new material, and the corresponding process conditions, such as temperature or solvent, to its end performance. The strength of a good model is its predictive power, its usability and its versatility, but it can also deliver erroneous and misleading information if not used properly, especially when the working set is very different to the training set. However, expert users – chemometricians, statisticians or modellers – can engineer very complex processes and data-centric workflows that encapsulate the heuristic, ensuring that novices properly use these models though a web portal, by accessing a collection of webservices. For instance, in production phase, laboratory technicians or bench chemists would have the most suitable QSAR/QSPR model automatically selected from a model warehouse depending on the problem they need to address. This practice ensures capture, dissemination and use of Knowledge throughout the organization that *de facto* maximizes its return on investment and reduces the total cost of ownership of its Informatics. It also enables seamless blending of technology, like bioinformatics, cheminformatics, materials science or robotics, with business rules, such as business intelligence, enterprise resource planning or business process management.

2.2 Data Harmonization

Best-practice data-mining techniques are ineffective without high-quality data, fast and reliable access to the information and a consistent capture of data and processes. The experimental issue is addressed with an apposite methodology by the experimentalist. The second topic is more challenging because it involves coping with the disparate data structures and data-exchange protocols. This heterogeneous information can be overwhelming to maintain and requires tailored tools to be utilized. This drastically impacts the total cost of ownership of the Informatics infrastructure, precludes a proper dissemination of knowledge and delays scientific breakthroughs [1 – 3].

Commercial and public organizations dealing with life and materials sciences start dedicating collegially important resources to harmonize and integrate this incongruent information. Part of the solution is to use webservices that access and open up legacy or proprietary applications. For instance, Laboratory Information Management System (LIMS) providers support these protocols but also promote and support the development of open-standards using the XML framework. XML provides syntax and generic mechanisms to structure data in documents. It aims to provide a reliable transmission of versatile information regardless of the platform used, which guarantees portability of data between different operating systems and machine architectures. XML schemas, i.e., dictionaries not only restricted to scientific applications, are increasingly developed worldwide and international regulatory institutions, such as International Union for Pure and Applied Chemistry (IUPAC) [24], act to prevent anarchy. This technology complements and rationalizes the utilization of data-mining techniques by combining data standardization, accessibility, portability and modularity with new computational techniques. It also eases system-integration services that permit the rapid deployment of flexible solutions to its adopters. An immediate and obvious corollary is the consistent capture of data and processes. Because the semantic and the underlying ontology are defined in the XML schema, the transformation of data into knowledge is facilitated.

Ontologies [25] are a formal, explicit specification of a shared conceptualization that emerged in artificial intelligence as an alternative to represent knowledge and provide meta-information that describes data semantics. Ontologies enable shared knowledge and reuse where information resources can be communicated between human or software agents. Semantic relationships in ontologies are machine-readable, in such a way they enable making statements and asking queries about a subject domain due to the use of a conceptualization, which describes entities and their relationships. This conceptualization enables those software agents of a vocabulary to represent and to communicate knowledge. Properties, activities and characteristics of a material, for example, a catalyst, are highly correlated with its constituents and its preparation. The reliable and consistent capture of the catalyst preparation recipe would facilitate the identification of subtle factors responsible for the material performance. Eventually, the combination of process conditions, elemental descriptors and recipes enable purely *in-silico* materials design, including design of experiment, screening and optimization.

2.3 The Missing Link

Science, technology, politics and economics are globally interconnected and one does not evolve without affecting the others. The incentives for an integrated Informatics (infrastructure) strategy is to create very clear and tangible business benefits [26]. It enables the definition of the strategic direction within corporate and business development structures, the identification of future therapeutic areas and technologies to generate growth and both isolating and valuing discrete project or company-based opportunities that align with this vision drive the investment in R&D. A perfect and natural symbiosis between the new global economy and the emerging global R&D is vital to ensure sustainable and profitable scientific breakthroughs. Given the scarcity of profitable R&D projects in biopharmaceuticals, companies need to narrow their focus and maximize their efficiency and to create value within very real, but short-term constraints. This is not only restricted to the pharmaceutical and biotech worlds. Other industries, like oil and gas, or fine chemicals, are also tightly linked with business imperatives depending, for instance, on geopolitical conditions that can affect their operational structures. Thus far, tools for dealing with comprehensive strategies for portfolio analysis, decision-making and optimization for the pharmaceutical and biotechnology industries, and that take into account the technological constrictions and risks are scarce or even nonexistent. The need to combine data- and process-centric workflows encapsulating business rules and scientific applications is growing strongly.

2.4 Enterprise Discovery Planning

An Enterprise Discovery Planning (EDP) software platform facilitates the organization and the optimization of discovery activities at an enterprise level and is designed to address the issues discussed above. An EDP-oriented discovery platform is an open infrastructure that emphasizes the management of collaborative discovery projects and contrasts with conventional more static systems. EDP platforms are containers for integrated services and analysis of dispersed information. This concept is analogous to the successfully realized principle of Enterprise Resource Planning in business information process management.

The main characteristics of an EDP software platform are:

- an open architecture that supports, as plug-ins, new resources, such as output from high-throughput devices, databases or analytical software components,
- a dynamic information-integration framework enabling a flexible, seamless access to heterogeneous scattered data,
- an exhaustive collection of generic analytical and modelling tools that encompass all required processing functions, supplied as building blocks for composite auditable discovery applications,
- an intuitive application composition framework to facilitate the creation of cross-domain data- and process-centric workflows, eventually exposed and adapted for novice users,

- an apposite application performance, reliability, security, availability, scalability and manageability,
- compliancy with the apposite regulatory requirements and other typical legal requirements, such as confidentiality, internal control processes and document retention,
- a searchable workflow warehouse permitting the monitoring, management, optimization and reporting of scientific discovery processes and global corporate performance.

Such EDP-oriented platforms operate as a technology which is a collaborative and social bridge and are the backbone for overseeing R&D activities ranging from the life sciences and the oil and gas industries, to contract research organizations.

2.5 Adoption of EDP

The incentives for adopting an EDP-based solution are found in the current very competitive commercial and technological landscape. For instance, sustainable high return on investment, minimizing the total cost of ownership, scattered complex organizational company structures, post-merger assets management and increased R&D outsourcing, volatile core competencies and poor knowledge- and intellectual-property management tools motivate the implementation of EDP-based solutions.

Although EDP-based solutions can help organizations mitigate the impact of a slowing economy and lower corporate profit margins, and provide tactical cost effectiveness and sophisticated results without increasing complexity, they are often perceived as too complex and perilous to implement. Tangible expected benefits are often counterbalanced by prior disastrous over-promised and under-delivered IT projects and by rigorous regulatory environment tending to reinforce conservatism with regard to adopting new processes and technologies. Unquestionably, data and services integration remain a key challenge for the pharmaceutical and chemical industry because companies are still learning how to integrate new informatics constructs into the world of chemistry, biology and robotics.

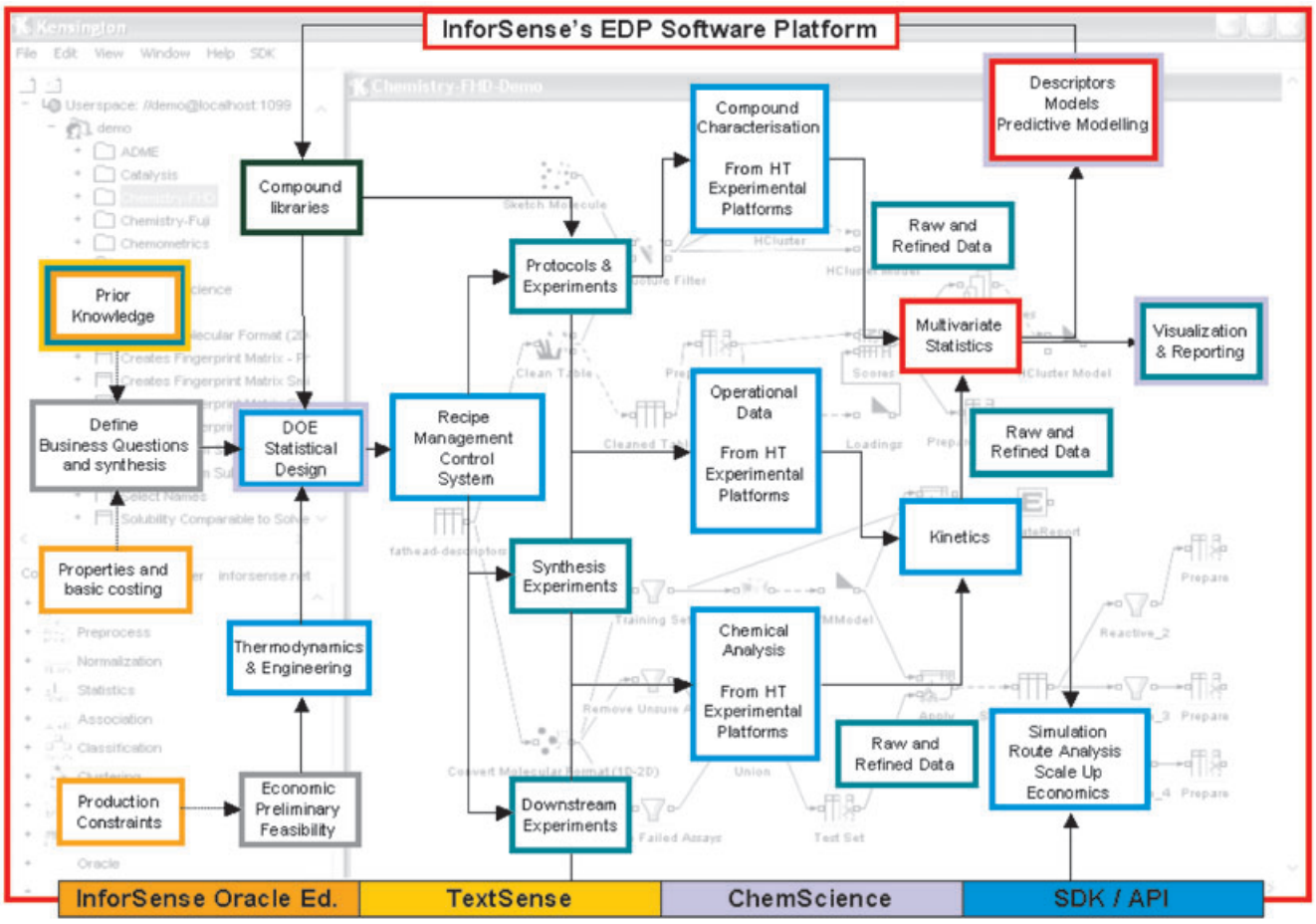

Figure 3. InforSense's Knowledge Discovery Environment (KDE) proposes a full set of modular and interlinked software components that enacts the Enterprise Discovery Planning (EDP) concept. It combines text mining, cheminformatics, bioinformatics and database access. The SDK empowers the user desiring to deliver a "best-of-breed" EDP software system.

2.6 InforSense's Approach to EDP

InforSense's Knowledge Discovery Environment (KDE) proposes a full set of modular and interlinked software components that enact the EDP concept (see Figure 3). InforSense technology has been deployed in pharmaceutical and biotech industries and academic institutions as the new infrastructure for discovery informatics.

Powerful Architecture: KDE's architecture (see Figure 4) provides computational power, versatility and scalability. An individual discovery station provides a basic single-user system that can be augmented gradually to address new requirements. Advanced modules designed for collaborative campaigns and large-scale exploitation of workflows, within a high-performance computing framework, allo one to tailor the system to fulfil precise analytical and performance requests.

Integrated Access to Data Assets: KDE ensure seamless and secure access to heterogeneous and scattered data sources from data warehouses. Flexible and intuitive interfaces are provided to relational databases, specialized informatics data warehouses and to semi-structured data such as text, images and web sources.

Extensible Discovery Algorithms: KDE proposes a comprehensive portfolio of high performance discovery tools to cope with modern R&D requisites. Its open architecture, with its Software Development Kit (SDK) that consents the effortless integration of third-party applications and its in-buildt webservices access framework, ensure that problems are addressed properly and efficiently.

Text Mining and Ontology Solutions: This extensive suite of tools is designed to analyze large bodies of text in order to complement other discovery material. Ontology support augments the text-mining offering and is integrat-

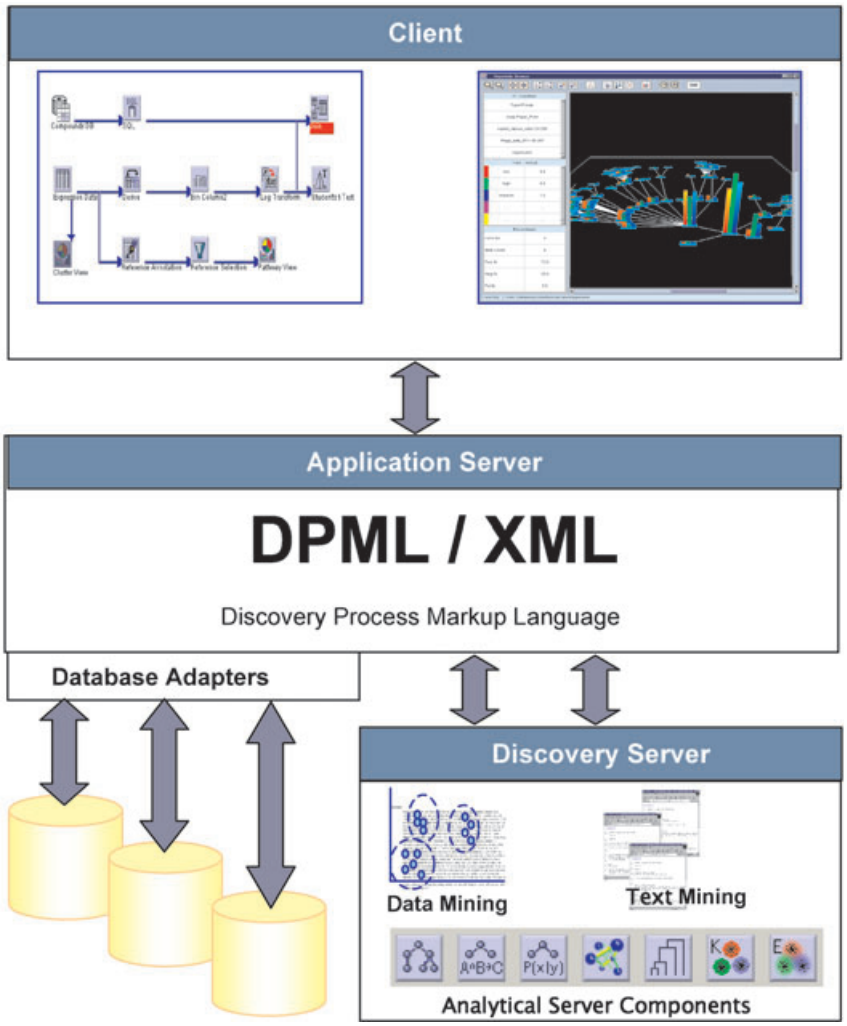

Figure 4. The InforSense EDP software system is a three-tier client-server architecture enabling large-scale, distributed creation and management and execution of discovery processes. KDE includes integrated access to data assets through an extensible set of Database Adapters. The Client Suite provides an integrated environment for the creation of discovery processes (top left) as well as interactive mining and visualization (top right). Computationally intensive processes are automatically executed on the Discovery Server.

ed with cheminformatics and bioinformatics Inforsense modules.

Collaborative Discovery Workflows: This framework facilitates the dissemination of knowledge and intellectual property that is encapsulated in data- and process-centric workflows (see Figures 5a – 5c). Eventually, workflow templates can be shared and exchanged between different business units to secure the best use of core competencies with the organizations. This *modus operandi* is extendable to virtual organizations or subcontractors, for instance.

Workflow Warehousing: Inforsense provides a palette of tools to search exhaustively within proprietary collections of workflows gathered in a workflow warehouse (see Figure 6). Queries can operate on annotated and audited building blocks composing workflows. That framework enables to identify core competencies, track the use of informatics resources, and provides limited functionalities for portfolio management.

Discovery Portal: Workflows can be complex entities build by several specialized groups. The discovery portal module authorizes IT managers and expert users to expose specific properties of a workflow to novice users. The workflow is published as a service that is available to the authorized personnel via a normal web access. For instance, a set of workflows mapping the entire complexity of an R&D process and its corresponding heuristic can be

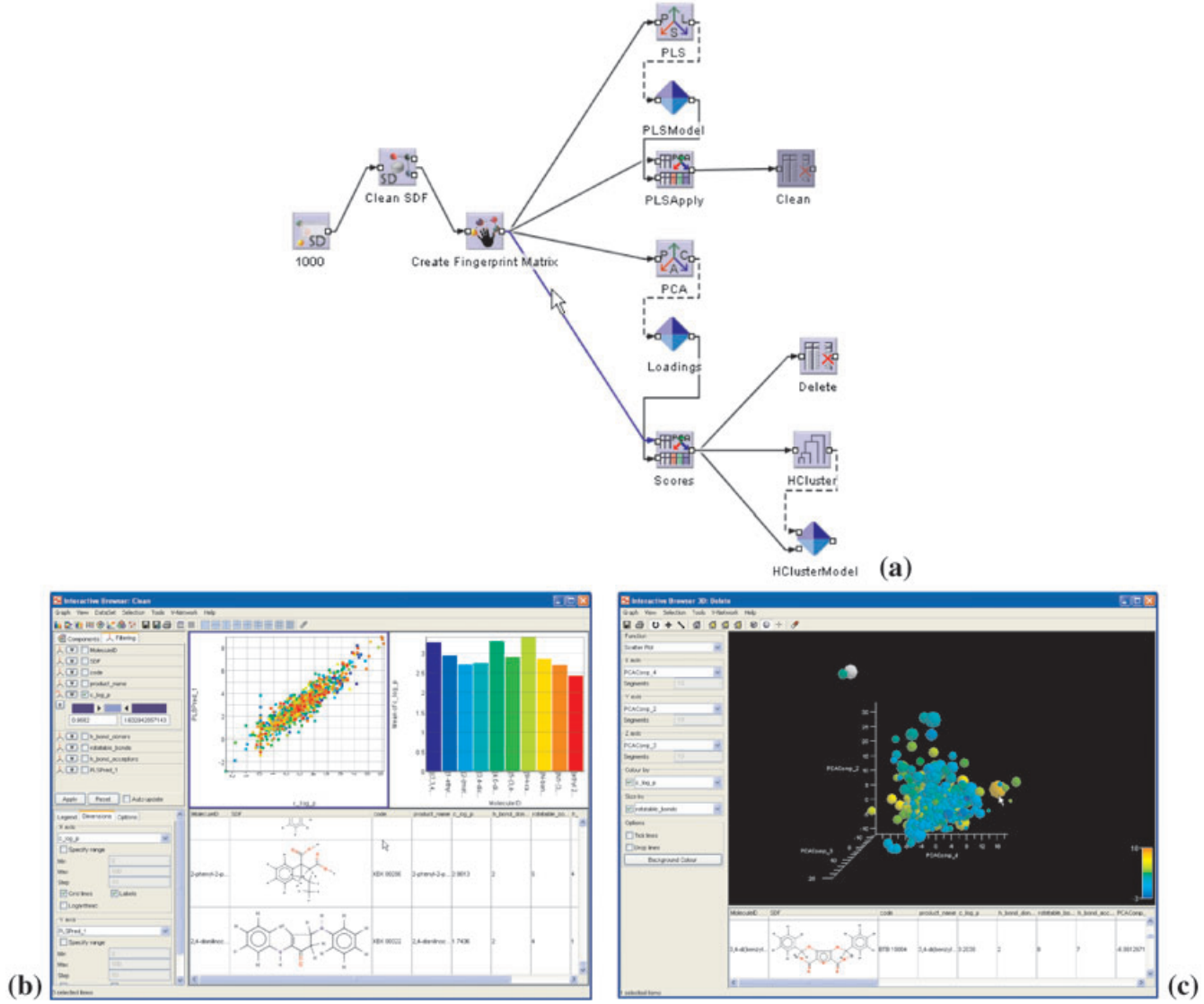

Figure 5. a) This workflow encapsulates the multivariate analysis performed on the fingerprints calculated for library of chemical compounds. The system has an embedded visualization framework. Workflow templates b) – c) can eventually be shared and exchanged between different business units to secure the best use of core competencies with the organizations. Ultimately, the workflow is published as a service that is available to the authorized, and maybe, novice personnel via a normal web access.

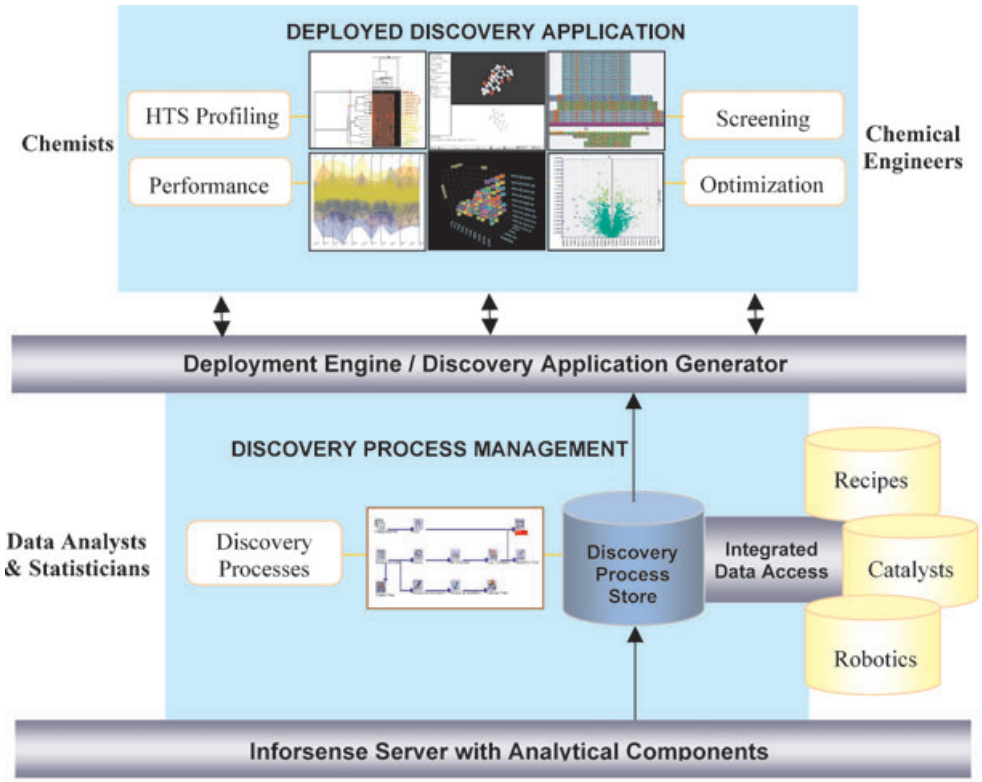

Figure 6. The InforSense Knowledge Discovery Environment as an Enterprise Discovery Planning platform. Workflows are enabling technologies that facilitate collaborative activities of multi-disciplinary business units to share either Discovery Processes or Standard Operating Procedures (SOP) – this is accomplished within the "Discovery Process Management" environment. Eventually, core competency groups can produce, share and publish functional knowledge within the company to therefore empower novice users. This task is achieved with the "Deployment Engine". Reducing the human intervention boosts performance and effectiveness. Repetitive tasks are minimized to secure time for analytics.

presented as a collection of web pages having a specific layout and exposed settings.

3 Conclusion

We have showed that the Enterprise Discovery Planning software platform is the first significant step to facilitate the optimization of discovery activities at an enterprise level. An EDP software platform empowers modern organizations that deal with new, diverse and rapidly evolving cross-domain challenges. Important investments are made both at the public and private levels to accelerate the development of such systems that are not restricted anymore to "early-adopters". However, these efforts are not the panacea to extract information and knowledge from unstructured and unreliable data sets. Data harmonization is also essential to achieve a fully integrated and powerful discovery framework.

Acknowledgements

The authors express their appreciation to the many enthusiastic and competent co-workers, who, during the last years, have taken up the design of Inforsense's software packages. We would like to thank particularly Anthony Rowe and Judith Bandy for their contribution. We are also grateful to Mr. Tim Smith, student at the London Business School (LBS), for educating us in business intelligence.

References

[1] a) D. S. Goldfarb, *IDC White Paper* **2003**, Doc #29242. b) K. Yokley, D. S. Goldfarb, M. Swenson, *IDC White Paper* **2003**, Doc #29353. c) D. S. Goldfarb, *IDC White Paper* **2003**, Doc #DR2003 B6DG. d) D. Chung, *IDC White Paper* **2003**, Doc #AP282111K. e) D. Chung, *IDC White Paper* **2003**, Doc #30166. f) D. S. Goldfarb, *IDC White Paper* **2004**, Doc #pr2004 06 24 145543. g) D. S. Goldfarb, *IDC White Paper* **2004**. h) D. S. Goldfarb, *IDC White Paper* **2004**, Doc #IDC P9025. i) D. S. Goldfarb, *IDC White Paper* **2004**, Doc #IDC P8485. h) Michael Swenson, *IDC White Paper* **2003.**

[2] Deloitte & Touche LLP (Eds.), *Life Sciences Quarterly*, 1st Quarter **2004.**
[3] a) D. Chung, *IDC White Paper* **2003**, Doc #AP282119K. b) M. Claps, *IDC White Paper* **2004**, Doc #MS05L. c) M. Claps, *IDC White Paper* **2004**, Doc # PP06K.
[4] R. J. Finamore, *CSSC White Paper*, **2003.**
[5] a) M. McGeary, P. M. Smith, *State Support for Health Research: An Assessment*, Report to the Mary Woodard Lasker Charitable Trust and Funding First, **2001.** b) M. McGeary, P. M. Smith. *Leadership by Example: Federal Agency Programs to Enhance Health Care Quality.* National Academy Press, Washington, DC, **2002**. c) Z. Zimmerman, *IDC White Paper* **2004**, Doc # 30968.
[6] a) Food and Drugs Administration (FDA) (Eds.), *Electronic Records; Electronic Signatures; Final Rule*, Code of Federal Regulations, Title 21, Food and Drugs. Part 11. Federal Register, **1997**, 62, 54, 13429–13466. b) Food and Drugs Administration (FDA) (Eds.), *21 CFR Part 11; Electronic Records; Electronic Signatures, Validation*. Draft Guidance for Industry, **2001**. c) Food and Drugs Administration (FDA) (Eds.), *Computerized Systems Used in Clinical Trials*, Guidance for Industry, **1999**. d) J. Hanover, M. Swenson, J. B. Golden III, *IDC White Paper*. 2004, Doc #30946. e) L. Draper, *IDC White Paper* **2004**, Doc #31320. f) J. Hanover, *IDC White Paper* **2004**, Doc #31365.
[7] a) M. Swenson, *IDC White Paper* **2003**, Doc # 29554. b) M. Swenson, C. Rancourt, *IDC White Paper* **2003,** Doc # 30112. c) G. D. Wilson, *IDC White Paper*. **2003**, Doc #30037. d) Z. Zimmerman, R. Fisler, B. Consulting, *IDC White Paper* **2004**, Doc #31405. e) D. Byron, *IDC White Paper* **2003**, Doc #29837.
[8] a) *Nature* **2001**, *409*, 745–964. b) *Science* **2001**, *291*, 1145–1434. c) D. Korn et al., *Science* **2002**, *296*, 1401–1402.
[9] Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of the Regions, *Life Sciences And Biotechnology – A Strategy For Europe*, Commission Of The European Communities, COM(2002) 27 final, **2002.**
[10] a) C. Stratowa, *Curr. Drug Disc.* **2002**, *2*, 29–33, b) J. A. Dimasi, *Pharmacoeconomics* **2002**, *20*, 1–10. c) E. Davidov, *Drug Disc. Today* **2003**, *8*, 75–183. d) G. Pisano, S. Wheelright, *Harvard Business Review* **1995**, *September–October*, 93–105.
[11] a) J. Morris, *KPMG White Paper* **2002**. b) Cap Gemini Ernst & Young, *Life Science & Chemicals*, **2001**, p. 3. c) Cap Gemini Ernst & Young, *Life Science & Chemicals*, **2002**, p. 4. d) Industrial Research Institute (Eds.), *R&D Trends Forecast for 2003*, Washington, DC, **2002**. e) S. Feldman, R. Winett, *IDC White Paper* **2002**, Doc #198SOFTWA3440.
[12] a) M. Hall, *IDC White Paper* **2003**, Doc # 29322. b) D. Chung, *IDC White Paper* **2003**, Doc # AP282111K, c) M. Swenson, Z. Zimmerman, *IDC White Paper* **2004**, Doc # 31061.
[13] K. Yokley, M. Swenson, *IDC White Paper* **2003**, Doc #03C3832.
[14] UBS Global Life Sciences Conference; D. Senf, *IDC White Paper* **2003**, Doc #29244.
[15] a) EU-FP6 Projects, information van be found on European Community website www.cordis.lu. b) M. S. Lesney, *Today's Chemistat work*, January **2004**, 18–20. c) J. C. Maxwell, A. Pitagno, C. Trupp Gil (Eds.), *Chemical Sciences in the FY 2005 Budget*, American Chemical Society, **2004**. d) K. Koizumi (Ed.), *AAAS*, Department of Homeland Security, **2004**. e) F. M. Ross Armbrecht, Jr, (Ed.), *R&D and Innovation in Industry*, Industrial Research Institute **2004**. f) K. Koizumi (Ed.) R&D Trends and Special Analyses, *AAAS*, Department of Homeland Security, **2004.** g) Information can be found on the website at http://www.aaas.org/spp/rd/.
[16] N. Adams, U. S. Schubert, *J. Comb. Chem.* **2004**, 6, 12–23.
[17] a) A. Holzwarth, P. Denton, H. Zanthoff1, C. Mirodatos, *Catal. Today* **2001**, *67*, 309–318. b) D. Akporiaye, I. Dahl, A. Karlsson, M. Plassen, R. Wendelbo, D. S. Bem, R. W. Broach, G. J. Lewis, M. Miller, J. Moscoso, *Microporous Mesoporous Mater.* **2001**, *48*, 367. c) M. Baerns, C. Mirodatos, *NATO Science Series, II: Mathematics, Physics and Chemistry* **2002**, *69*, 469. d) D. Farrusseng, L. Baumes, I. Vauthey, C. Hayaud, P. Denton, C. Mirodatos, in: E. G. Derouane, V. Parmon, F. Lemos, F. R. Ribeiro (Eds.), *Principles and Methods for Accelerated Catalyst Design and Testing*, Kluwer Academic Publishers, Dordrecht, The Netherlands, **2002**, pp. 101. e) B. Jandeleit, D. J. Schaefer, T. S. Powers, H. W. Turner, W. H. Weinberg, *Angew. Chem., Int. Ed.* **1999**, *38*, 2494. f) W. F. Maier, *Angew. Chem., Int. Ed*. **1999**, *38*, 1216. g) B. K. Lavine, J. Workman, Jr., *Anal. Chem.* **2002**, *74*, 2763–2770. h) S. Wold, M. Sjostrom, P. M. Andersson, A. Linusson, M. Edman, T. Lundstedt, B. Norden, M. Sandberg, Multivariate Design and Modeling in QSAR, Combinatorial Chemistry, and Bioinformatics, in: K. G. Jorgensen, (Ed.), *Proceedings of the 12th European Symposium on Structure–Activity Relationships–Molecular Modeling and Prediction of Bioactivity*, Kluwer Academic/Plenum Press, New York, **2000**, pp. 27–45. i) L. Xue, J. Bajorath, *J. Comb. Chem. High Throughput Screening* **2000**, *3*, 363–372. j) E. Giraud C. Luttmann, F. Lavelle, J.-F. Riou, P. Mailliet, A. Laoui, *J. Med. Chem.* **2000**, *43*, 1807–1816.
[18] a) B. R. Beno, J. S. Mason, *Drug Discov. Today* **2001**, *6*, 251–258. b) R. Benigni,L. Passerini, D. J. Livingstone, M. A. Johnson, A. Giuliani, *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 558–562. c) G. J. M. Gruter, A. Graham, B. McKay, F. Gilardoni, *Macromol. Rapid Commun*. **2003**, *24*, 73–80.
[19] a) M. Shen, C. Beguin, A. Golbraikh, J. P. Stables, H. Kohn, A. Tropsha, *J. Med. Chem.* **2004**, *47,* 2356–2364. b) N. Brown, B. McKay, F. Gilardoni, J. Gasteiger, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1079–1087. c) J. D. Holliday, S. L. Rodgers, P. Willett, M.-Y. Chen, M. Mahfouf, K. Lawson, G. Mullier, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 894–902. c) J. Hert, P. Willett, D. J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, and A. Schuffenhauer, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185. d) R. P. Bywater, T. A. Poulsen, P. Røgen, P. G. Hjorth, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 856–861. e) J. K. Wegner, H. Frohlich, A. Zell, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 921–930; *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 931–939. f) E. Byvatov, G. Schneider, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 993–999. g) N. Baurin, R. Baker, C. Richardson, I. Chen, N. Foloppe, A. Potter, A. Jordan, S. Roughley, M. Parratt, P. Greaney, D. Morley, R. E. Hubbard, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 643–651. h) P. Constans, J. D. Hirst, *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 452–459. i) D. K. Agrafiotis, W. Cedeno, V. S. Lobanov, *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903–911.
[20] J. B. Golden III, *IDC White Paper* **2004**, Doc # TB20040304.
[21] a) S. Rogers, K. Dick, *IDC White Paper* **2003**, Doc # 30222. b) B. O'Donnell, *IDC White Paper* **2003**, Doc #30016. c) F. Gens, M. Melenovsky, S. Rogers, V. Turner, *IDC White Paper* **2004**, Doc # 31371. d) F. Gens, M. Melenovsky, S. Rogers, V. Turner, *IDC White Paper* **2004**, Doc # 30946. e) J. Hanover, M. Swenson, *IDC White Paper* **2003**, Doc # 31275.

f) Information can be found on the website at http://www.i3-c.org/.
[22] G. Alonso, F. Casati, *Web SerVices – Concepts, Architectures and Applications*, Springer-Verlag, Berlin, Heidelberg, New York, **2003.**
[23] a) K. Dick, *IDC White Paper* **2004**, Doc # pr2004 08 24 165803. b) C. Ahlberg, *Drug Discov. Today* **1999**, *4*, 370 – 376. c) N. Fay, D. Ullmann, *Drug Discov. Today* (information biotechnology suppl.) **2002**, *7*, S181 – S186. d) B. L. Claus, D. J. Underwood, *Drug Discov. Today* **2002**, *7*, 957 – 966. e) T. Peakman, S. Franks, C. White, M. Beggs, *Drug Discov. Today* **2003**, *8*, 203 – 211.
[24] http://www.iupac.org/projects/2002/2002-022-1-024.html/.
[25] D. Fensel, *Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce*, Springer-Verlag New York, Inc, **2003**.
[26] J. B. Golden III, *IDC White Paper* **2004**, Doc # TB20040304.

