<table>
<thead>
<tr>
<th>Pg</th>
<th>Nominating</th>
<th>User Company</th>
<th>Project Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>BioSoteria</td>
<td>University of MD School of Pharmacy</td>
<td>Pharmacovigilance eLearning curriculum</td>
</tr>
<tr>
<td>8</td>
<td>Biovista Inc.</td>
<td></td>
<td>Literature based discovery</td>
</tr>
<tr>
<td>12</td>
<td>Collaborative Drug Discovery Inc</td>
<td></td>
<td>ChemSpider</td>
</tr>
<tr>
<td>23</td>
<td>Coldsteam Laboratories</td>
<td></td>
<td>Strategic Analysis of Patent Landscape</td>
</tr>
<tr>
<td>33</td>
<td>City University of New York</td>
<td></td>
<td>Modeling, Mathematical Simulations and Data Mining: Making Sense out of Really Difficult Cancer Data</td>
</tr>
<tr>
<td>38</td>
<td>Dolcera</td>
<td>Dow Agrosciences</td>
<td>Innovation Dashboard</td>
</tr>
<tr>
<td>45</td>
<td>goBalto.com</td>
<td></td>
<td>The fastest, most effective way to connect with life science service providers</td>
</tr>
<tr>
<td>55</td>
<td>Hewlett-Packard Company</td>
<td>The University of Texas M. D. Anderson Cancer Center</td>
<td>The Translational Research Extensible platform (T-REX) Project (now referred to as ResearchStation)</td>
</tr>
<tr>
<td>67</td>
<td>KAI Research, Inc., an Altarum Company</td>
<td></td>
<td>Implementation of the Pediatric Common Data Elements and Data Repository</td>
</tr>
<tr>
<td>72</td>
<td>Merck and Co., Inc.</td>
<td></td>
<td>A Virtual Technology Symposium</td>
</tr>
<tr>
<td>77</td>
<td>Molecular Connections Private Ltd</td>
<td></td>
<td>A platform for discovery, analysis and modeling of published biomedical facts from PubMed.</td>
</tr>
<tr>
<td>86</td>
<td>Phase Forward GlaxoSmithKline K.K.</td>
<td></td>
<td>Changing the Game: Post-Marketing Surveillance in Japan</td>
</tr>
<tr>
<td>92</td>
<td>Recombinant Data Corp. Centocor R&amp;D, Inc. (a subsidiary of Johnson &amp; Johnson)</td>
<td></td>
<td>tranSMART</td>
</tr>
<tr>
<td>96</td>
<td>Sitrof Technologies Copernicus Group IRB</td>
<td></td>
<td>CGIRB’s DocuShare Paperless Environment</td>
</tr>
<tr>
<td>104</td>
<td>Tessella</td>
<td>AstraZeneca</td>
<td>CrysIS (CRYStallographic Information System)</td>
</tr>
<tr>
<td>110</td>
<td>Transinsight GmbH</td>
<td>Unilever</td>
<td>Unilever Knowledge Browser</td>
</tr>
</tbody>
</table>
Bio-IT World 2010 Best Practices Awards

Nominating Organization name: BioSoteria, Inc.
Nominating Organization address: 6425 Christie Avenue, Suite 220
Nominating Organization city: Emeryville
Nominating Organization state: CA
Nominating Organization zip: 94608
Nominating Contact Person: Sally Van Doren
Nominating Contact Person Title: President
Nominating Contact Person Phone: 510 295 6425
Nominating Contact Person Email: svandoren@biosoteria.com

User Organization name: University of MD School of Pharmacy
User Organization address: 220 Arch Street, 12th Floor
User Organization city: Baltimore
User Organization state: MD
User Organization zip: 21201
User Organization Contact Person: Sheila Weiss-Smith, PhD
User Organization Contact Person Title: Professor and Director, Center for Drug Safety
User Organization Contact Person Phone: 410-706-6989
User Organization Contact Person Email: sweiss@rx.umaryland.edu

Project Title: Pharmacovigilance eLearning curriculum

Team Leaders name:
Team Leaders title:
Team Leaders Company:
Team Leaders Contact Info:
Team Members name:
Team Members title:
Team Members Company:

Entry Category: Drug Discovery & Development

Abstract Summary:
Introduction: UMSOP offers a premier educational program in the United States for training pharmacists in practice areas, including drug safety and regulation, pharmacoepidemiology, and risk management. To enhance elective course offerings to pharmacy students, a pharmacovigilance eLearning course curriculum, eLadder Safety (BioSoteria, Inc.), was deployed on the UMSOP learning management system (LMS). Courses are self-paced and contain a test assessment for each topic; certificates of completion are issued with passing scores. The objectives of this program were to offer elective courses to pharmacy students on drug safety and regulation, which may in part address the anticipated requirements in the US for increased drug safety education in health care professional schools, based on The Patient Protection and Affordable Care Act, passed by the US Senate in 2009.

Course 1 pertains to drug and device development and regulation and marketing authorization. Course 2 addresses premarketing, clinical safety surveillance including US and EU regulations and guidance, ICH and CIOMS guidelines, serious
adverse event case management and premarketing risk assessment and analysis. Course 3 complements prior courses with the full lifecycle management of drug safety with postmarketing pharmacovigilance, including individual case safety report (ICSR) case management and reporting, regulations and guidances, MedDRA and WHODRUG coding, literature ADR reports, periodic reporting, special situations in pharmacovigilance, and risk management. The content in the 3 courses combined contain the vast majority of functional core competencies of pharmacovigilance practitioners described in the literature (Edwards B et al 2005)

Results: 3 elearning courses were deployed on the UMSOP LMS, Blackboard. The 3 elearning courses were designed to be Sharable Content Object Reference Model (SCORM) compliant and contained up to 40 hours of self-paced content and practice case exercises in pharmacovigilance practice topics. Courses were designed using Lectora eLearning Software (Trivantis) by the BioSoteria educational team, in collaboration with a multicompany faculty of subject matter experts in the area of drug development and regulation, and pharmacovigilance practice in the US and EU. 

ROI achieved: 
Conclusions: 


See eLearning course summaries submitted to aproffitt@healthtech.com
Click the logo to launch the eLadder™ Safety Course DEMO.

For more information, contact us today:

Email: training@biosoteria.com
Phone: 1.(866) 660-5553 ext. 013
www.biosoteria.com
Safety professionals in today’s rapidly changing global regulatory environment must have a strong foundation in safety surveillance and pharmacovigilance, which requires ongoing, comprehensive, and up-to-date training and education.

Courses in BioSoteria’s engaging, innovative, online eLadder Safety eLearning Program are:

- **Available immediately** through BioSoteria’s website—take any course in the series anytime and from anywhere you have an Internet connection, no travel required.

- **Available off-the-shelf** for rapid-deployment drug safety training—no travel or customized development fees required.

- **AICC/SCORM-compliant** and ready to run on your Learning Management System (LMS). No LMS? No problem. We offer a hosted LMS solution to deliver training to your employees or contractors.

- **Content-rich and truly multimedia**, with engaging graphics, animations, audio, video, and immersive application and reinforcement exercises.

- **Developed by an experienced multidisciplinary team**, including:
  - A faculty of subject-matter experts and drug safety practitioners
  - Instructional designers and medical writers who specialize in eLearning
  - Graphic designers, animators, and certified medical illustrators
  - eLearning programmers and Learning Management System (LMS) specialists

Course 1 provides a comprehensive overview of how medical products, including pharmaceuticals, biologics, and medical devices, are developed and regulated in the US, Europe, and other global markets.

**LEARNING OBJECTIVES:**

- Compare the differing perspectives of regulatory agencies and industry.
- Describe expectations for the safety of medical products.
- Identify the organizations and individuals who have a stake in the safety of medical products.
- Describe the roles and responsibilities of each stakeholder with regard to safety of medical products.
- Identify key events in the history of US regulation of medical products.
- Identify landmark legislation and describe its significance in regulatory history.
- Identify the phases of drug discovery and development.
- Describe the types of testing performed at each phase of development.
- Identify the regulatory milestones that occur during the development process.
- Identify the main regulatory agencies for the United States, European Union, and Japan.
- Describe the key regulatory documents and processes for drugs, biologics, and medical devices for the US and EU.
- Identify the leading organizations involved in efforts to harmonize global regulatory activities.

**Length:** 3 hours  
**Audience:** Professionals entering the pharmaceutical or biotechnology industry in the areas of nonclinical research, clinical research, drug safety and pharmacovigilance, project management, regulatory affairs, medical information or medical affairs, quality assurance/control

Visit [www.biosoteria.com](http://www.biosoteria.com), email training@biosoteria.com, or call 1.866.660.5553 ext. 13 for a demo today.
Course 2: CLINICAL SAFETY SURVEILLANCE

With interactive, real-world case studies, Course 2 covers the collection, assessment, and reporting of adverse events (AEs) during the clinical development phase of a drug or biologic. It includes US and EU regulations and related ICH guidance documents pertaining to clinical safety surveillance of an investigational medicinal product and provides a step-by-step overview of the responsibilities and activities of the study sponsor, from receipt of serious adverse event (SAE) reports from investigators, through all stages of case processing including causality assessment and reporting to regulatory authorities.

LEARNING OBJECTIVES:

- Describe the purpose of safety surveillance in clinical trials
- Explain the general regulatory oversight of clinical trials in the US and EU
- List the key stakeholders in a clinical trial and explain their role in assuring the safety of study subjects
- Differentiate between US and EU regulations and guidelines, and explain the influence of ICH and CIOMS
- Define key terms and concepts central to clinical safety surveillance (e.g., serious, expected, relatedness, dechallenge, rechallenge)
- Describe the purpose and content of Investigator’s brochure and the expected AE list
- List the responsibilities of the clinical trial investigator in AE reporting
- Understand the data fields collected for AEs and SAEs on reporting forms
- Identify the key steps in the workflow for processing an SAE report
- Describe in what situations a 7- and/or 15-day expedited safety report is required
- Determine how to triage SAE reports, and determine if an SAE report meets expedited regulatory reporting requirements
- Learn the important components of an SAE case narrative
- Demonstrate how and when to write a query to request followup information
- Describe when it is appropriate to unblind treatment assignment for SAE reports
- Understand the difference between the clinical and safety databases, and the purpose of database reconciliation
- Explain what is required for submitting reports to US and EU regulatory authorities and investigators
- Explain how overdose and pregnancy are typically reported in clinical trials

Courses in BioSoteria’s engaging, innovative, online eLadder Safety eLearning Program are:

- **Available immediately** through BioSoteria’s website-take any course in the series anytime and from anywhere you have an Internet connection, no travel required
- **Available off-the-shelf** for rapid-deployment drug safety training—no travel or customized development fees required
- **AICC/SCORM-compliant** and ready to run on your Learning Management System (LMS). No LMS? No problem. We offer a hosted LMS solution to deliver training to your employees or contractors.
- **Content-rich and truly multimedia**, with engaging graphics, animations, audio, video, and immersive application and reinforcement exercises
- **Developed by an experienced multi-disciplinary team**, including:
  - A faculty of subject-matter experts and drug safety practitioners
  - Instructional designers and medical writers who specialize in eLearning
  - Graphic designers, animators, and certified medical illustrators
  - eLearning programmers and Learning Management System (LMS) specialists

Visit www.biosoteria.com, email training@biosoteria.com, or call 1.866.660.5553 ext. 13 for a demo today.
Safety professionals in today’s rapidly changing global regulatory environment must have a strong foundation in safety surveillance and pharmacovigilance, which requires ongoing, comprehensive, and up-to-date training and education.

Courses in BioSoteria’s engaging, innovative, online eLadder Safety eLearning Program are:

- **Available immediately** through BioSoteria’s website—take any course in the series anytime and from anywhere you have an Internet connection, no travel required

- **Available off-the-shelf** for rapid-deployment drug safety training—no travel or customized development fees required

- **AICC/SCORM-compliant** and ready to run on your Learning Management System (LMS). No LMS? No problem. We offer a hosted LMS solution to deliver training to your employees or contractors.

- **Content-rich and truly multimedia**, with engaging graphics, animations, audio, video, and immersive application and reinforcement exercises

- **Developed by an experienced multidisciplinary team**, including:
  - A faculty of subject-matter experts and drug safety practitioners
  - Instructional designers and medical writers who specialize in eLearning
  - Graphic designers, animators, and certified medical illustrators
  - eLearning programmers and Learning Management System (LMS) specialists

**LEARNING OBJECTIVES:**

- Understand the purpose and importance of postmarketing PV
- Know the US and EU regulations and guidance for postmarketing PV
- Discuss roles, responsibilities and relationships of internal and external stakeholders with regard to postmarketing PV
- Identify the sources of postmarketing safety reports and distinguish between spontaneous ADR reports and solicited AE reports
- Define the key terms and definitions pertaining to postmarketing PV
- Distinguish between serious and non-serious ADRs
- Distinguish between expected/labeled and unexpected/unlabeled ADRs
- Identify the steps that occur at the Drug Safety Department of the MAH/license holder after ADR report receipt
- Describe how ADRs are triaged and reported to regulatory authorities
- Describe the components of a complete case narrative
- Describe how to evaluate and report ADR reports from the literature

**COURSE 3: POSTMARKETING PHARMACOVIGILANCE**

Course 3 delivers key information on the receipt, evaluation, and reporting of postmarketing spontaneous adverse drug reactions (ADRs), including the procedures that marketing authorization holders/license holders must follow to fulfill US and EU regulations. Similarities and differences between postmarketing pharmacovigilance (PV) regulations in the US and EU are covered.

**Length:** 4 hours

**Audience:** Individuals interested in entering the field of drug safety and pharmacovigilance; professionals already in the biopharmaceutical industry practicing in the areas of clinical research, drug safety, regulatory affairs, and GCP compliance.
Bio-IT World 2010 Best Practices Awards

1. Nominating Organization (Fill this out only if you are nominating a group other than your own.)

   A. Nominating Organization
   Organization name:
   Address:

   B. Nominating Contact Person
   Name:
   Title:
   Tel:
   Email:

2. User Organization (Organization at which the solution was deployed/applied)

   A. User Organization
   Organization name: Biovista Inc.
   Address: 927 Raymond Rd, Charlottesville, VA 22902, USA

   B. User Organization Contact Person
   Name: Dr. Aris Persidis
   Title: President
   Tel: 434 242 6514
   Email: arisp@biovista.com

3. Project
   Project Title: Literature based discovery
   Team Leader
   Name: Dr. Andreas Persidis
   Title: CEO
   Tel: 434 242 6514
   Email: andreasp@biovista.com
   Team members – name(s), title(s) and company (optional):

4. Category in which entry is being submitted (1 category per entry, highlight your choice)
   - [ ] Basic Research & Biological Research: Disease pathway research, applied and basic research
   - [ ] Drug Discovery & Development: Compound-focused research, drug safety
   - [ ] Clinical Trials & Research: Trial design, eCTD
   - [ ] Translational Medicine: Feedback loops, predictive technologies
5. Description of project (4 FIGURES MAXIMUM):

A. ABSTRACT/SUMMARY of the project and results (150 words max.)

The life sciences industry has been talking about an innovation deficit for quite some time now. At Biovista we believed that while our existing knowledge is indeed incomplete and is partially to blame for this deficit, if we could become extremely efficient at using all the publicly available information, we could increase our output of new therapies dramatically. To test this belief we launched a project to develop a generic literature-based discovery platform that would help users make ‘discoveries’ in a systematic way. Once developed, we proceeded to test the predictive accuracy of the platform, aiming to retrospectively predicted drug adverse events as these are reported over the years in the annual ASCO meetings. Our analysis showed a predictive accuracy of 79%. We subsequently tested 5 predictions of the system for repositioned drugs. In 3 of these we had positive efficacy results in the industry standard animal model tests.

B. INTRODUCTION/background/objectives

The life sciences industry has been talking about an innovation deficit for quite some time now. While increasing amounts of funds are being invested in R&D, the output of new therapies for a variety of indications has not followed pace. At Biovista we believed that this was and still is to be expected, given that all 3 knowledge components that are required for drug development, namely our knowledge of biology, of disease mechanisms and of drug mechanisms of action, are incomplete. At the same time we believed that our existing bodies of knowledge contain sufficient chunks of useful information that may be seemingly disparate, but that if combined in the right way may indicate useful, novel answers to yet unsolved problems.

At the same time we believed that existing technical approaches were focusing too much on optimizing algorithms that may be missing the grand picture by a long shot (in a sense getting extremely good at something that doesn’t really pay off). We therefore embarked on a project that would aim to develop a generic literature based discovery platform that would help users make ‘discoveries’ in a systematic way. In other words a computer system that could be used on a daily basis to shift through and use available scientific knowledge in the support of complex problem solving. From past experience we were
very much aware that we should not expect a fully automated system, but one where human experts are in the loop, can assess system output and make the final recommendations.

C. **RESULTS** (highlight major R&D/IT tools deployed; innovative uses of technology).

The platform we developed works with a variety of text-based resources, including Medline abstracts, the full text of USPTO Life science related patents, the AERS database and others. We developed tools that automatically extract information from these resources and create multi-dimensional profiles of every major disease, gene and drug development related concept. For example the gene “BRCA1” as its own profile, as does the disease “Wilms Tumor” as does the pathway “apoptosis”. The platform also uses semantic organization of these concepts but of a ‘light’ type. In other words we chose not to create complex ontologies since we felt they would be too restrictive both in terms of system performance and in terms of the assumptions they carry. Finally we spent significant resources in designing a custom underlying database management system that could handle very large amounts of data (currently our database contains over 4 billion correlations) and respond in real time. We had to reject the solution of object-oriented databases since they couldn’t handle the size of the data set. We have also done performance measurements that show a speed increase of 200X over columnar databases. We believe that this level of performance is critical because it is an important human interaction issue, affecting system usage and hence the ultimate success of the project.

In summary we believe that our platform is successful because we focused on certain basic requirements that are critical for the ultimate utility of such tools. These include:

1. Handling large data sets and responding in real time
2. Encompassing as few assumptions as possible. Since we wanted to develop a system that supported discovery work, we felt that the fewer specific viewpoints it encompassed, and the more direct access it have to the raw data, the better it would be.
3. Being able to handle incomplete and contradicting data.

Once the platform was developed and before embarking on its actual use, we wanted to assess its predictive accuracy and how this may vary with the amount of information that is available to the system. We chose to test the platform in the prediction of adverse events as these are reported in the annual ASCO events. Our analysis of over 850 data points showed an average predictive accuracy of 79%. Details are reported in a paper currently under review by a top tier journal.

The final stage of the project involved the development of system predictions in vivo. In late 2008 we ran 5 predictions for repositioning drugs and subsequently proceeded with animal model testing of these. We had positive efficacy results in for 3 of these predictions while the fourth was a failure and the fifth a border case. While the statistical significance of 5 predictions is still small, we believe that results so far are in agreement with our theoretical results. As a consequence, we are planning to continue our development work in 2010 and beyond.

D. **ROI** achieved or expected (200 words max.):
The expected ROI of this project is substantial. Developing a new compound to the clinical stage, using the traditional approach costs on average in the low double digit millions per compound and takes on average 3-5 years. Using this platform, Biovista was able to identify its repositioned candidates and have a first animal model efficacy test on average in 2 months and at a substantially lower cost. Moving into clinical trials does of course require the standard set of experiments to be performed. However, in the case of repositioned drugs, since the toxicity profile is already known and favorable, this is a low risk proposition.

E. CONCLUSIONS/implications for the field.

We believe that technologies such as Biovista’s literature-based discovery platform have the potential to transform the field. The lack of bias (other than what is in the raw data in the first place) the complete coverage of the field in combination with the interpretive capabilities of subject matter human experts create a resource that could help diminish a lot of the inefficiencies that are currently in the system. By going about knowledge discovery in a systematic, unbiased way, they promise to optimize our understanding and use of our hard won knowledge, and render drug development a more deterministic process than it currently is.

6. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)
Bio-IT World 2010 Best Practices Awards

1. **Nominating Organization** (Fill this out only if you are nominating a group other than your own.)

   **A. Nominating Organization**
   Organization name: Collaborative Drug Discovery
   Address: 1633 Bayshore Hwy, Suite 342, Burlingame CA.

   **B. Nominating Contact Person**
   Name: Sean Ekins
   Title: Collaborations Director
   Tel: 215-481-1059
   Email: sekins@collaborativedrug.com

2. **User Organization** (Organization at which the solution was deployed/applied)

   **A. User Organization**
   Organization name: Collaborative Drug Discovery Inc (as well as Multiple pharmaceutical and industrial communities as the software is available on the internet.)
   Address: 1633 Bayshore Hwy, Suite 342, Burlingame CA 94010

   **B. User Organization Contact Person**
   Name: Sean Ekins
   Title: Collaborations Director
   Tel: 215-687-1320
   Email: sekins@collaborativedrug.com

3. **Project**
   Project Title: ChemSpider
   Team Leader
   Name: Antony J. Williams
   Title: VP, Strategic development at the Royal Society of Chemistry
   Tel: 919-341-8375 (mobile 919-201-1516)
   Email: antony.williams@chemspider.com
   Valery Tkachenko, Chief Technology Officer, RSC
   Sergey Shevelev, Software Developer, RSC

4. **Category in which entry is being submitted (1 category per entry, highlight your choice)**

   - □ Basic Research & Biological Research: Disease pathway research, applied and basic research
   - □ Drug Discovery & Development: Compound-focused research, drug safety
5. Description of project (4 FIGURES MAXIMUM):

A. ABSTRACT/SUMMARY of the project and results (150 words max.)
ChemSpider (www.ChemSpider.com) is a chemistry database acquired by the Royal Society of Chemistry in April 2009, which was initiated and launched in March 2007. At the time of writing it contains over 23 million unique chemical entities aggregated from almost 300 diverse data sources including government databases, chemical vendors, commercial database vendors, publishers, all of the databases listed above and from individual chemists. It represents a free and valuable resource for the scientific community which contains calculated molecular properties. Importantly it represents an alternative to expensive chemistry databases (requiring a subscription) out of reach for many scientists. Secondly ChemSpider has single handedly contributed to improving the quality of structure curation in public resources, having a major impact on PubChem and Wikipedia. Third ChemSpider has become a repository for journals, hosting structures for several publishers. ChemSpider represents an innovative tool that is changing the way chemistry is delivered on the web.

B. INTRODUCTION/background/objectives
This nomination addresses how the Royal Society of Chemistry through the development of ChemSpider by Antony Williams and colleagues, have provided a freely available disruptive database technology to all for capturing molecule structures and data.

Biomedical research is fast moving towards a collaborative network of chemists and biologists and making knowledge available to the masses, enabling rapid sharing of information (1-4). Yet, pharmaceutical scientists (biologists and chemists in particular) commonly find themselves overwhelmed by the availability of information on the web, in primary commercial databases such as CAS SciFinder (http://www.cas.org/), journals and, commonly, a plethora of internally developed systems inside their companies. From another perspective that of the academic or those in the financially constrained developing countries, biology and chemistry information has long been limited by the tolls associated with accessing commercial databases. Even the calculation of relatively simple molecular properties (such as lipophilicity) has, up until very recently, required knowledge and ownership of informatics software. Structure searching of such chemically aware databases used to be restricted to computational specialists but now with user friendly web-based tools,
even the biologist or biomedical researcher (who is likely to be chemistry naïve) can find such tools of value for searching for interesting molecules from any of the commercial vendors. This is analogous to the power of the web in revolutionizing how people can book their own travel on the web without having to use third party travel agent

As an example, in the world of chemistry there are tens if not hundreds of chemical structure databases, many containing molecules of biological interest, yet until recently there was no single way to search across them. There are databases of curated literature data, chemical vendor catalogs, molecular properties, environmental data, toxicity data, analytical data etc. The only way to know whether a specific piece of information is available for a chemical structure is to have simultaneous access to all of these databases as well as journals and other commercial resources for mining and integrating them. Since many of these databases are commercial there is no way to easily determine the availability of information either within these or in the open access databases. The availability of molecule databases such as PubChem (http://pubchem.ncbi.nlm.nih.gov/) has changed scientists expectations of web-based databases in many ways but only goes part way to inform us about our chemical universe, and in particular those molecules we might be interested in for their pharmaceutical properties. For example, while the web has provided improved access to chemistry-related information there has not been an online central resource allowing integrated chemical structure combined with biology data-searching of chemistry or biology databases, chemistry articles, patents and web pages such as blogs and wikis. The development of ChemSpider has changed this.

There are many freely available chemical compound databases on the web and they assume different forms. These files generally contain the chemical identifiers in the form of chemical names (systematic and trade) and registry numbers. Since the files are assembled in a heterogeneous manner the resulting data are plagued with inconsistencies and data quality issues. Such an approach to gathering and merging data is a far cry from that taken by commercial database vendors who manually gather and curate data. While the commercial databases offer curated data there is certainly a price-barrier to accessing the information. A number of the free online resources are also manually curated and, as will be discussed later, can offer as high a quality as the commercial offerings. These resources are, however, constructed with a specific focus in mind and therefore commonly number in the low thousands of structures rather than the millions available in the larger online databases. Meanwhile, there are several large online database resources offering access to valuable data and knowledge.

The quality of chemical information in the public domain is also generally quite low. This does not mean that the data are not of value but that care needs to be taken in the nature of the provider as an authority. There is, of course, no central body responsible for the quality of data in the public domain. Databases of chemical structure information besides PubChem include ChemIDPlus and ChembioFinder which are commonly looked upon as authorities in terms of reliable information. However, these sources are also aggregators of information and are at risk of perpetuating errors form the original public data and depositions. Errors in structure-identifier pairs are common and inaccurate structure representations, specifically in regards to stereochemistry, proliferate across many databases. A definitive description of the challenges regarding quality in public domain databases, and the rigorous processes required to aggregate quality data was provided by Richard et al. during their assembly of the EPA DSSTox databases.
The creation, hosting and support of a curated compound database containing structures of chemical and biological interest with integrated content is an expensive enterprise. Historically these databases have been built as a result of hundreds if not thousands of man years of rigorous and exacting human effort and then, for some of the original founders in this domain, migrated onto computer systems. In the development of these systems some host organizations have created sizeable revenues. The hosting of large databases, the text-based searching of immense amounts of data and the ability to disseminate complex forms of graphical information via standard protocols provides an opportunity for future disruptive offerings in this domain whereby online offerings can also become authorities and, with the support and input of the community, can offer the benefits of crowdsourcing for enhancing the data.

ChemSpider (http://www.chemspider.com/): ChemSpider was initially developed as a hobby project by a small group of dedicated cheminformatics specialists. The intention was to aggregate and index available sources of chemical structures and their associated information into a single searchable repository and make it available to everybody, at no charge. ChemSpider was unveiled to the public in March 2007 with the intention of “building a structure centric community for chemists”. ChemSpider has grown into a resource containing over 23 million unique chemical structures. The data sources have been gathered from chemical vendors as well as commercial database vendors and publishers and members of the Open Notebook Science community. ChemSpider has also integrated the SureChem patent database collection of structures to facilitate links between the systems. The database can be queried using structure/substructure searching and alphanumeric text searching of both intrinsic as well as predicted molecular properties. The ChemSpider developers also added virtual screening results using the LASSO similarity search tool to screen the ChemSpider database against all 40 target families from the Database of Useful Decoys (DUD) dataset.

ChemSpider has enabled unique capabilities relative to the primary public chemistry databases. These include real time curation of the data, association of analytical data with chemical structures, real-time deposition of single or batch chemical structures (including with activity data) and transaction-based predictions of physicochemical data. The system developers have also made available a series of web services to allow integration to the system for the purpose of searching the system as well as generation of InChI identifiers and conversion routines. The system also integrates text-based searching of Open Access articles. The index is expected to increase dramatically as they extract chemical names from OA articles and convert the names to chemical structures using name to structure conversion algorithms. These chemical structures will be deposited back to the ChemSpider database thereby facilitating structure and substructure searching in concert with text-based searching.

ChemSpider has a focus on, and commitment to, community curation and ease of use. The social community aspects of the system demonstrate the potential of this approach. The team have committed to the release of a Wiki-like environment for further annotation of the chemical structures in the database, a project they term WiChempedia. They will utilize both available Wikipedia content and deposited content from users to enable the ongoing development of community curated chemistry. ChemSpider was acquired in May 2009 by the Royal Society of Chemistry and will continue to grow in its reach into the chemistry, biology and biomedical research communities.

ChemSpider provides a significant new knowledgebase and resource for chemists working in different domains (pharmaceutical, medical device, agricultural, nanotechnology, environmental, consumer goods
etc). As the diversity of chemistry information continues to expand and, coupled with an increasing awareness for quality, curation and improved tools for focused searches, chemists are now able to find valuable information within a few seconds using a few keystrokes with ChemSpider. This shift to publicly available resources offers great promise to the benefits of science and society while potentially being disruptive to commercial concerns that have held a stranglehold on chemistry knowledge and potentially inhibited innovation and progress. Such a public resource provides a rich source of molecule structures and physicochemical properties useful for the purpose of virtual screening experiments and potential drug discovery using Quantitative Structure Activity Relationship (QSAR) analyses and docking into protein receptor sites.

C. RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

ChemSpider has provided both a platform for searching as well as the deposition of new data and curation of existing data. By providing a platform for the deposition of new data ChemSpider has provided a way for chemists to expose molecules and associated data to the public thereby facilitating the rapid communication of chemistry not based on a classical publication model. Mainstream scientific publishers such as the Nature Publishing Group and, of course, the Royal Society of Chemistry also deposit their data into ChemSpider as an additional vehicle to provide access to their publications.

By providing a platform for public crowd-sourced curation ChemSpider has facilitated the cleansing of public domain data. Their system has allowed for the removal of poor data for data originating from ChEBI (http://www.chemspider.com/blog/collaboration-community-and-quality-in-chemistry-databases.html), from Wikipedia (http://www.chemspider.com/blog/a-need-to-improve-chemical-structure-handling-on-wikipedia.html) and they have reported errors to the FDA regarding their Daily Med website (http://www.chemspider.com/blog/dailymed-commentary-comment-1-structure-representations.html).

There have been several applications of ChemSpider for generating structure activity relationships. An example of this application (Ekins S, Iyer M, Krasowski MD and Kharasch ED, Molecular Characterization of CYP2B6 substrates, Curr Drug Metab, 9:363-373, 2008) used ChemSpider to provide structures and molecule properties for the human drug metabolizing enzyme CYP2B6. We were able to show how the property range for 64 substrates of this enzyme is very narrow (hydrophobic with one or more hydrogen bond acceptor) and seems limited to central nervous system active compounds. It should be noted that ChemSpider allows the user to download structures of interest and molecular properties so these could be used in other computational or analysis software. A second example used ChemSpider to derive molecular properties for machine learning models to predict biopharmaceutical characteristics of drugs (Khandelwal A, Bahadduri PM, Chang C, Polli JE, Swaan PW and Ekins S, Computational models to assign biopharmaceutics drug disposition classification from molecular structure, Pharm Res, 24: 2249-2262, 2007). A further application used ChemSpider to follow up a molecule selected by pharmacophore searching of vendor databases as a potential pregnane X receptor agonist (a potential target for modulating anticancer drug metabolism, transport etc). Substructure searching in ChemSpider indicated additional molecules of interest for testing which were validated in vitro and shown to have activity (Ekins S, Kholodovych V, Ai N, Sinz M, Gal J, Gera L, Welsh WJ, Bachmann K and Mani S, Computational discovery of novel low micromolar human pregnane X receptor agonists, Mol Pharmacol, 74: 662-672, 2008).
ChemSpider is acknowledged by scientists as a valuable resource for understanding chemistry (Figure 1). ChemSpider can be linked from other groups (academia or industry) software. For example Collaborative Drug Discovery (www.collaborativedrug.com) recently provided links to ChemSpider for molecules in this database. Again this enables the users to find out more information about their molecules (Figure 2). There are a multitude of other examples of databases and Wikis linking to ChemSpider. These include Wikipedia, PubChem and many others. Other databases such as WikiProteins and GeneWiki are presently developing their integration links to ChemSpider.

One current use for ChemSpider could be to find out as much about a compound as possible as researchers can eliminate undesirable leads early in the lead generation process by quickly accessing information on the pharmacological effects, side effects and drug-drug interactions for similar compounds or compound classes of interest, as well as their corresponding metabolites. For example for the compound gefitinib (Iressa®), what preclinical information exists? A search was initiated through ChemSpider and produced one hit. The results display for gefitinib includes the chemical structure, a series of intrinsic and predicted properties, links to a number of original data sources for associated information and a number of alphanumeric identifiers, some of which are validated. A number of the names, database IDs and synonyms connect to Wikipedia via the [Wiki] link, links to patents are immediately viewable and any articles containing Gefitinib or other synonyms in the title or abstract are linked through to Pubmed.

The above examples illustrate how the content in ChemSpider is useful to the scientific community involved in drug discovery and how free connectivity between tools via the web may enable a much broader impact.
Figure 1. Screenshots of a molecule record and predicted properties in ChemSpider.
Figure 2. A screenshot from the Collaborative Drug Discovery database showing the ChemSpider link below a molecule in the EPA ToxCast dataset.

D. ROI achieved or expected (200 words max.):

ChemSpider has become a hub for accessing structure-based information in over 200 databases by providing a centralized search engine. This has dramatically reduced the time it takes for chemists to resource structure-based information. Previously they would need to identify appropriate resources to query as well as launch the queries, each one through a different user interface with its own foibles and learning curve. In parallel with producing a database of chemical substances the ChemSpider team has taken on the challenge of curating and qualifying data, specifically relationships between chemical compounds and identifiers that are the basis of semantic markup in the life sciences literature.

ChemSpider web services are enabling many organizations to access quality data via a public resource. Companies such as Bruker, Waters, Thermo and Agilent, to name a few, have integrated to ChemSpider. The ROI for users of ChemSpider is very large as the software is free to them and their only investment is the time to use it

Prior to the acquisition of ChemSpider by RSC, approximately 6 man years of effort was invested. This should be balanced against the investment in PubChem of 5 years project lifetime, over 20 people on the development team and a multi-million dollar annual budget.

E. CONCLUSIONS/implications for the field.

The Internet has spawned access to unprecedented levels of information. For chemists the increasing number of resources they can use to access chemistry-related information provides them a valuable path to
discovery of information, one which was previously limited to commercial and therefore constrained resources. This newfound freedom has the ability to speed up research and sharing of results, develop extensive collaborations, conduct science in public, and in near-real time. Communication in chemistry is already witnessing a new revolution and ChemSpider in our opinion is a major contributor to innovation as illustrated by the following

1. Improving the quality of information available in online chemistry databases such as PubChem and Wikipedia by enabling a community-based curating process to help improve the association between a chemical compound and a set of identifiers (systematic names, trade names, synonyms, registry numbers).
2. Increased access to chemistry-related information such as the association of analytical data integration with patent searches and integration to QSAR-based modeling.
3. Increased access to online prediction of chemical properties and Web services (InChI and OpenBabel) for chemists.

6. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)

The following represent testimonials from ChemSpider users.

'The Open nature ChemSpider has allowed us to easily obtain chemical property data for inclusion into online spreadsheets. The ability to easily obtain such data in a programmatic fashion without having to go through authorization hoops, significantly enhances the possibility of novel, chemical mash ups'.

Rajarshi Guha – formerly Visiting Assistant Professor, School of Informatics, Indiana University, Bloomington, Indiana, now at the NIH.

'ChemSpider is a very useful tool for chemical structure verification. It is no secret that many public and even commercial databases are full of errors (wrong or duplicate structures, wrong structure-name association, etc.). ChemSpider is in the process of curating structures (as much as possible given very limited resources) as well as creating the libraries of synonyms associated with each structure. These services are very helpful in validating various specialized (e.g., toxicity or specific biological activity) databases mined from multiple available sources'.

Alexander Tropsha – Professor & Chair, Medicinal Chemistry and Natural Products, UNC School of Pharmacy, Beard Hall, Chapel Hill, North Carolina.

'ChemSpider has proven to be extremely successful, as evidenced by its recent acquisition by the Royal Society of Chemistry. It has become an essential tool for many research groups, including mine at Drexel University, and I now introduce it as a key resource in my chemical information retrieval class. Regardless of what project Antony pursues it results in success, due in no small part to his tenacity, creativity and resourcefulness. Those of us who have had the privilege of collaborating with Antony know how gratifying an experience it can be'.

Jean Claude Bradley - Associate Professor of Chemistry, Drexel University, Philadelphia.

'Love your site, solving lots of problems for us!'
James Little, Research Fellow, Eastman Chemical Company, TN, USA

‘ChemSpider has a number of advantages over a simple Google search. The variety of information about a compound provided at ChemSpider is hard to match on any other free web site, and the user can be sure that the data is provided by practicing chemists and in many cases has been reviewed for accuracy.’

Harry E. Pence, SUNY Distinguished Teaching Professor Emeritus

‘We hope that the ChemSpider / RSC partnership will both inspire more open-minded individuals to strike out on their own with similarly ambitious efforts and encourage various non-profit and government entities to actively recruit successful projects back into "the establishment" in ways which do not compromise project integrity and yet can enable even greater long-term positive societal impacts.’

Dr Warren DeLano (deceased)

The following publications (papers and presentations) highlight the value of ChemSpider.


Williams AJ, Internet-based tools for communication and collaboration in chemistry, Drug Discovery Today, Volume 13, Numbers 11/12, June 2008 502-506,

Williams AJ, A perspective of publicly accessible/open-access chemistry databases, Drug Discovery Today, Volume 13, Numbers 11/12, June 2008, 495-501


**ENTRY TIPS FROM THE JUDGES**

Bio-IT World has received and processed some 400 Best Practices entries since we first held the competition in 2003. Our experience has led to the following checklist of “do’s & don’ts” that will help you avoid simple mistakes or omissions that could detract from the quality of your entry. Follow these simple steps to improve your chances of recognition by our expert judges.

**Do**
- Describe the problem addressed and its importance
- List relevant technologies/products used on the project
- Identify key results and quantify the ROI associated with the project (e.g. man-hours or $ saved; new capabilities)
- Quantify the resources expended (hours, personnel, $)
- Explain broader impact on the life science community/biopharma industry

**Don’t**
- Enter promotional “brochure-ware” or rewrite press releases
- Make vague, unsupported qualitative claims
- Emphasize technology (product) features over project results -- integrate the two
- Forget to clearly identify positive results/benefits!
Bio-IT World 2010 Best Practices Award Submission

Knowledge Management Category

A. Nominating Organization (Fill this out only if you are nominating a group other than your own.)
Not Applicable

B. User Organization

A. User Organization
   Organization name: Coldstream Laboratories, Inc
   Address: 1500 Bull Lea Rd. Ste. 250, Lexington, KY 40511

B. User Organization Contact Person
   Name: Pramod Gupta
   Title: Vice President Quality & Chief Scientific Officer
   Tel: (585) 978-0008
   Email: pgupta@coldstreamlabs.com

3. Project
   Project Title: Strategic Analysis of Patent Landscape
   Team Leader: Joseph W. Wyse
   Name: Joseph W. Wyse
   Title: President & CEO
   Tel: (859) 948-4716
   Email: jwyse@coldstreamlabs.com
   Team members –
   Ken Zinda, President, Inspherion
   Pramod Gupta, Vice President Quality & Chief Scientific Officer, Coldstream Laboratories, Inc.

4. Category in which entry is being submitted (1 category per entry, highlight your choice)
   ☐ Basic Research & Biological Research: Disease pathway research, applied and basic research
   ☐ Drug Discovery & Development: Compound-focused research, drug safety
   ☐ Clinical Trials & Research: Trial design, eCTD
   ☐ Translational Medicine: Feedback loops, predictive technologies
   ☐ Personalized Medicine: Responders/non-responders, biomarkers
   ☐ IT & Informatics: LIMS, High Performance Computing, storage, data visualization, imaging technologies
   ☐ Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
   ☐ Health-IT: ePrescribing, RHIOs, EMR/PHR
   ☐ Manufacturing & Bioprocessing: Mass production, continuous manufacturing
5. Description of project

A. ABSTRACT/SUMMARY

Leaders in life science companies regularly face strategic decisions on developing or acquiring intellectual property to enter and protect their markets. The success or failure of these decisions requires insightful analyses of the patent landscape. A variety of patent tools and analytics used by companies to meet this demand often leave management with uncertainty that the landscape has been thoroughly evaluated. Inspherion has developed unique analytical algorithms and methods addressing this concern and their superiority is demonstrated in this submission. The case study presented asks the question: How does a mid-cap pharmaceutical company strategically develop “colloidal technologies” to deliver active ingredient across a variety of “mucosal membranes”? The results presented include a technology landscape and strategic information about market dynamics and competitor behavior and positioning that heretofore has not been available to corporate management. In addition, licensing/acquisition opportunities and potentially blocking patents are identified. The multi-dimensional information that can be leveraged from this analysis is compellingly superior and of high business value compared to other choices.

B. INTRODUCTION

Competitive Intelligence and strategic planning in the pharmaceutical, biotechnology, and medical device industries relies heavily on patent analysis. The business issues driving the need for patent analyses include, among others, merger and acquisitions, licensing, and new product or technology investments. But regardless of the specific business issue, a thorough patent analysis is vital for good decision making by management.

A number of patent search engines and analysis tools are available to companies. However a weakness in many of these tools is that they do not effectively analyze the entire patent landscape surrounding a technology of interest to provide the critical detail necessary for sound competitive and investment strategies. They typically provide a limited view of real threats and opportunities.

In this submission, Coldstream Laboratories presents a case study using unique strategic analysis tools and methods. These tools and methods were developed by Inspherion and PricewaterhouseCoopers over the past 11 years. Figure 1 summarizes a comparison between Inspherion’s approach versus conventional analysis.
Competitive intelligence/IP analysis typically follows four standard steps: Search, where you obtain the required data; Cluster, where the data is organized into meaningful groups; Analysis, where the clusters are analyzed for trends and insight; and Review, where the analysis results are interpreted and summarized. Most patent searches are completed by using one or more of the same type of tools: Free sites, such as the USPTO; Subscription sites, like Delphion; or Specialized Analytical tools, an example is Aureka.

The patent analysis process is generally ad hoc; repeatedly cycling between the search, cluster, analysis, review steps; and building up the analysis like assembling Lego blocks. Termination of the analysis activities generally occurs when resources or time have been exhausted. But unspoken questions remain: Is the analysis truly complete? How much time should have been spent?

Inspherion has systemized the step-wise analysis process similarly to how software companies have approached tax preparation and accounting. While Inspherion’s process utilizes the same four steps of Figure 1, navigating through the process is a comprehensive, rigorous, organized step-by-step approach with pre-defined duration and cost. Instead of cycling through to build-up the analysis, Inspherion starts by defining the analysis boundaries and fills in the pieces in a manner similar to building the boarder of a puzzle and filling in the center until the puzzle is completed. With a well defined process that has built-in quality control measures, it ensures the analysis stays focused on underlying business issues and provides the results sought from the patent landscape analysis.

How are the outcomes different? Unlike traditional document count analyses which only conveys how a company views itself, Inspherion’s technology assesses aggressive patterns of behavior that are independent of document counts or patent portfolio size. This is a true market comparison showcasing how each company’s activity compares to all others in the landscape. It enables identification of “under-the-radar”, emerging, aggressively transformational companies/inventors that could be a major threat because of their fierce focus or edge in their technology space.
Coldstream Laboratories, Inc. & Inspherion are presenting an example of the application of the Multi-Term Frequency Analysis (MTFA) tool to assess a pharmaceutical technology landscape. Similar analyses have been completed for a number of life science companies to achieve better strategic decisions. However, due to the confidentiality, no client project examples can be presented.

While all client names are confidential, Inspherion, and previously PricewaterhouseCoopers, have been providing analyses for 11 years to top Fortune companies, governments, and investment companies, both in the US and abroad. Inspherion also has a formal joint partnership with PricewaterhouseCoopers supporting their client engagements.

The partnership between Coldstream Laboratories, Inc. and Inspherion brings together an experienced team to support companies seeking to understand and develop strategies to navigate a competitive technology space. Specifically, our Team Leader, Joseph Wyse, developed the MTFA tool while serving previously as a Director at PricewaterhouseCooper’s Intellectual Asset Management practice. Inspherion President, Ken Zinda has spent nine years developing and optimizing strategic and tactical patent analysis tool and processes offered to clients. Pramod Gupta, Coldstream Laboratories Vice President of Quality & CSO, has served in senior positions at Abbott, Baxter, and Bausch & Lomb leading new drug and device development, intellectual property and licensing opportunities.

The example analysis presented here is built around a hypothetical pharmaceutical client’s business issue: How should they strategically develop “colloidal technologies” to deliver active ingredient across a variety of “mucosal membranes”?

The following analysis presents the service offering of Coldstream Laboratories, Inc. and Inspherion to this hypothetical client engagement.

C. RESULTS

The approach taken to develop a series of strategic and tactical actions for our “client” involved the following steps:

1. Frame the management decisions into a landscape structure;
2. Develop and refine the search criteria and perform worldwide search;
3. Clean up and analyze the data, taking into account corporate acquisitions and competitor consolidations;
4. Interpret and summarize the results in terms of Technology, Opportunity, Competition, the Technology Pipeline, Competitive Level, Innovation and Market Demand, individual company activity, and more.

**Technology Landscape**

The technology landscape was designed to evaluate both drug delivery routes and colloidal delivery systems as they are applied to five categories of active ingredients. Each of these three categories is assigned to a column or row in the landscape and multiple search criteria were developed for each row/column concept. The mucosal delivery routes, intranasal, oral, transdermal, and vaginal were assigned rows A-D. The colloidal drug delivery system of interest, emulsions, liposomes, and nanomaterials, were assigned rows E-G. The active ingredient categories of interest (DNA, hormones,
peptides, poorly soluble small molecules, and proteins) are assigned rows 1-5. Each of these 12 search terms, representing the 7 rows and 5 columns, were the only searches required to complete the landscape. Individual areas in the landscape are populated by patents identified in both areas comprising the individual area in the landscape. Figure 2 presents the technology landscape of interest for this hypothetical clients’ business opportunity.

**FIGURE 2:**

<table>
<thead>
<tr>
<th>DNA</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1989 to 1/30/2010</td>
<td>DNA</td>
<td>Hormones</td>
<td>Peptides</td>
<td>Poorly Water Soluble</td>
<td>Proteins</td>
</tr>
<tr>
<td>A</td>
<td>DR: Intranasal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>DR: Oral Mucosa</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>DR: Transdermal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>DR: Vaginal</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>DS: Emulsion</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>DS: Liposome</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>DS: Nanomaterials</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each area in the landscape shows the number of documents that have the corresponding row/column concepts coexisting in the same document. Using the patent activity as a proxy for market importance, quartiles help visualize what areas are more mature (red) or emergent (green). This enables construction of a technology pipeline and begins to frame an understanding of the market dynamics. From this chart we see that Hormones and Peptides are the most active ingredients and Liposome and Nanomaterials are the most active delivery systems (DS).

**Opportunity, Concentration and Analysis Areas**

The patents in each area of the landscape are compared to other areas to determine valuable information about the overall landscape. Figure 3 showcases three categories of areas relative to others, areas with high concentration, high growth, and high opportunity.

Because the analysis metrics examine patterns of patent activity behavior, areas with sudden or large changes in activity can be readily identified. High opportunity areas are places of low level activity that has shown more sudden changes such as would occur in an emerging area that is beginning to see promise. Growth areas come after the emergent ones and are generally places where the market has identified and is pursuing a particular technology solution. Both these areas are targets for development and can help position a company early enough on the product adoption curve to maximize their return on investment (ROI).
Concentration areas are places where a small number of companies own most of the IP. This could either be because the market has matured and the IP is consolidated into a few major corporations, or this could be an area where only a few companies have been investigating and market pull does not yet exist. In either case, these areas pose higher barriers to entry.

Declining areas can also be identified and would be areas being abandoned by the market or where companies are shifting their resources out of the area. These places would naturally be avoided in most strategic decisions.

Innovation Metrics

The analysis also evaluates historical trends for overall innovation in the landscape (Table 1).

**TABLE 1**

Innovation metrics are easily calculated for all documents and companies in the landscape and compared to all others in the space, providing a relative ranking of their importance. Clients have back-tested and
verified these metrics by independently identifying important aspect of a technology being analyzed and then comparing those elements to those contained in the documents.

Experience has shown that when a technology is undergoing market pull there are very distinct and growing patterns. When a technology is not experiencing market pull, such as occurs for government programs involving advanced technologies, the metrics are low and generally a random pattern.

In the case of this analysis we see the metrics peaked in the late 90’s. This suggests that during that period many of the technologies were believed to be very promising. But since then the metrics have been steadily declining suggesting technical limitations have readjusted the market’s belief in the technology’s potential and ultimately the areas of profitability. More detailed examination of these technologies would therefore be in order by breaking apart individual columns into their fundamental elements and running the analysis from an expanded view.

In terms of investment, Company B reentered this technology landscape with new innovations in 2007 after testing this technology space in 1990 and 1998. This may suggest a new resurgence by Assignee B in this emerging technology landscape and possible appetite for licensing (Table 2)

**TABLE 2**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In contrast, Company A, invested heavily in creating new IP in 2000-2004 and has gradually tapered off filing new innovations since 2005. This may indicate a movement away from these targets. However, IP that they have developed should be carefully evaluated for potentially blocking client interests.
MTFA utilizes an algorithm to assign a competitive score across the landscape. By making almost 300 measurements per document, that algorithm assesses company aggressiveness and the interrelationship between concepts described in relevant sections of each patent. It then compares those results to all other companies in the landscape to prioritize their behavior, and ultimately their threat. It is important to note that the algorithm provides a more insightful way to prioritize companies for investigation than traditional current “highest document count” methods. Out of thousands of assignees, it surfaces small and relatively obscure ones who otherwise would go unnoticed because of the volume of data that exists. Essentially, the algorithm follows a method used in secure communications in which it separates low level, consistent patterns from high level, random patterns.

Figure 4 compares the more traditional patent count and MTFA’s competitive score for the top 10 assignees. These top ten assignees include three major pharmaceutical companies (Assignees A, B, and F), three academic research centers (Assignees C, D, and G), three small cap companies (Assignees E, H, and J), and one individual inventor (Assignee I).

**FIGURE 4**

A competitor gap analysis is a company centric view and shows where each company is focusing their efforts and therefore believes where it has a competitive advantage (Figure 5). This comparison helps to rapidly define a short list for licensing or acquisition opportunities by comparing where a company’s strengths are and how the gaps might be filled. This figure shows that there are several assignees that can fill gaps in Assignee A’s portfolio. Another observation is that Assignees A and B do not overlap in their most active areas of the landscape. However, with both large companies working on the delivery of different active ingredients by multiple delivery routes their IP could block others seeking to protect...
innovations across these areas of the landscape. Patents from these major pharmaceutical companies should be carefully evaluated because they are aggressively developing an IP position.

FIGURE 5

Underlying Patent Data Review

Based on the competitive score several tactical opportunities were identified after reviewing the underlying patent from a MTFA analysis of a patent landscape for the top 50 assignees. Examples are:

- Assignees H and another in the top 20, both small-cap companies developing strong patent portfolios with the potential for blocking nanomaterial delivery systems and 3 of 4 mucosal delivery routes. These company’s patents should be carefully reviewed. These assignees also present a potential licensing or acquisition opportunity to accelerate the development of our clients’ portfolio.

- Several top 50 assignees have developed product specific delivery systems for active ingredients in the technology landscape. Patents from these companies should be evaluated to determine if broader application of their technology is possible and could be licensed or developed internally.

Effort Required and Project Statistics

A landscape analysis such as the one illustrated in this submission involves at least two individuals to complete: an analyst knowledgeable of Inspherion’s analysis tools and process, and one or more subject matter experts to assist the analyst in framing the landscape around the business issue.

The above analysis hourly efforts are summarized below:

- Developing the matrix: 1 CLI, 1 Inspherion, 24 hours total
- Completing MTFA Analysis: 1 CLI, 1 Inspherion, 24 hours total
- Data Review and Report: 2 CLI, 1 Inspherion, 32 hours total

A project of this scope costs a client between 20,000 and 25,000 inclusive of labor, data costs, and software fees.
Brief statistics regarding this MTFA analysis

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patents in Landscape</td>
<td>11700</td>
</tr>
<tr>
<td>Unique Companies</td>
<td>4400</td>
</tr>
<tr>
<td>Search Terms Created</td>
<td>12</td>
</tr>
<tr>
<td>Technology areas in Landscape</td>
<td>35</td>
</tr>
</tbody>
</table>

The short, 2 week turn-around of this MTFA analysis, the strategic information provided, and the size of the patent landscapes cannot be matched in depth or scope by conventional patent search and analysis methods. It is for this reason that 90% of the Fortune companies using this approach engage in multiple projects in a single year despite having teams of experienced IP analysts and licensing professionals, and access to every commercially available patent analysis toolset.

D. ROI

As explained in this submission, the thoroughness with which a business issue related patent landscape is analyzed can have a significant impact on understanding the risks and opportunities facing a management team. Based on statistically sound, validated, and back-tested algorithms and processes, the MTFA tool provides not only valuable information about individual patents, it provides insightful details about market dynamics, competitive strength of assignees, and positioning of companies and inventors not otherwise obtainable. A typical MTFA analysis is completed in less than 4 weeks with 120-160 man-hours to evaluate 100 areas in a technology landscape (the above example was only 35 areas) comprised of an average of 25,000 individual, worldwide patent records, 10,000 assignees, and search criteria containing 2,000 - 15,000 characters. Clients report they have saved up to $200,000 on a single project and have used Inspherion’s approaches on high risk projects with starting investments as high as $1B.

E. CONCLUSIONS

Multi-Term Frequency Analysis (MTFA) is a strategically unique analysis tool and process which enables comprehensive and compelling assessment of patent activity in an area of business interest. This tool has proven highly effective over several years of assessment and has been utilized by corporations and government agencies. Its ability to conduct multi-dimensional analyses assures identification of all assignees engaged in a given area of interest, which can be missed using conventional search techniques. MTFA is a rapid and in-depth approach, which makes it even more attractive during times when information is desired in limited time.
Bio-IT World 2010 Best Practices Awards

1. **Nominating Organization** (Fill this out only if you are nominating a group other than your own.)

   A. **Nominating Organization**  
      Organization name:  
      Address:  

   B. **Nominating Contact Person**  
      Name:  
      Title:  
      Tel:  
      Email:  

2. **User Organization** (Organization at which the solution was deployed/applied)

   A. **User Organization**  
      Organization name: City University of New York.  
      Address: 160 Convent Avenue, New York, NY 10031  

   B. **User Organization Contact Person**  
      Name: Dr. Shubha Govind  
      Title: Professor, Genetics  
      Tel: 212.650.8476  
      Email: sgovind@scisun.sci.ccny.cuny.edu  

3. **Project**

   Project Title: Predictive Modeling, Mathematical Simulations and Data Mining: Making Sense out of Really Difficult Cancer Data  
   Team Leader  
   Name: Navin Sinha  
   Title: Independent Statistical Consultant  
   Tel: 952-942-5698  
   Email: sinhanavin@hotmail.com  
   Team members – name(s), title(s) and company (optional): Dr. Shubha Govind; PhD, Professor of Genetics at City university of NY. And Navin Sinha, MS, MBA.  

4. **Category in which entry is being submitted (1 category per entry, highlight your choice)**

   - [ ] **Basic Research & Biological Research**: Disease pathway research, applied and basic research  
   - [ ] **Drug Discovery & Development**: Compound-focused research, drug safety
5. Description of project (4 FIGURES MAXIMUM):

A. **ABSTRACT/SUMMARY** of the project and results (150 words max.) The data were produced by graduate students of Dr. Shubha Govind, PhD, Professor of Genetics, City University of New York (CUNY). The tumor shape and size characterization was presented by Chiu et al (www.sciencedirect.com<http://www.sciencedirect.com/>). How to quantify accurately the lower gene expression of Black cells (BC), especially when cancer researcher community believes that there is a competition to fill in the space created by cell death (Dr. Shubha Govind, personal communication, June 2009)? A dataset with two mutated genotypes, Bc (black cells) and lwr(-) containing frequency of 25K (1K=1000 micro meter) to 300K was undertaken. The genetic details are presented by FlyBase Database (Accessed August 2009). The area size from 25K, 75K, 100K… to 300K had only two frequency observations; one from each genotype. Application of Odds Ratio, Maximum likelihood, Simulated Regression Residual plot analysis, and Average effects of gene substitution resulted in hypotheses: 1. Operon or Tumor Gene expression occurs in a deterministic way from 25K to 300K in both mutants and would have longer survival probability, and 2. Log-normal distribution arose due to compensatory behavior by lowest size distribution (25K) at the expense of next few tumor classes. Dynamical Simulations (Symbolic Regression) and Reverse Engineering Algorithms simulations were being utilized to test those two hypotheses.

B. **INTRODUCTION/background/objectives** The end goal of Biotech-Pharmaceutical research remains Billion $ drug discovery patent, but a difficult dataset could make this objective elusive! Some research is accepted to have small sample size as information generation is quite laborious. Literature review resulted in no evidence on how to make sense of such a cancer data from fruit fly.

C. **RESULTS** (highlight major R&D/IT tools deployed; innovative uses of technology). A new data mining method was invented based upon graph of function to find reliable results from discrete, small sample data. Tools utilized were creative functions of EXCEL and VBA. These methods showed that tumor cell are somewhat linearly correlated unless the gene is over semi-over expressed.

D. **ROI** achieved or expected (200 words max.): Given a brain tumor type, the methods (Powerpoint) presented shows that a comprehensive treatment is possible for brain tumor, bringing variability in length of hospital stays for brain tumor patients (Healthcare outcome) under control ($).
E. CONCLUSIONS/implications for the field. Now, discrete, small brain tumor data from Systems Biology will have a reliable statistical/mathematical results.

1. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.) 1. Program Announcement
   > Title: Predictive Modeling, Mathematical Simulations and Data Mining: Making Sense out of Really Difficult Cancer Data
   > Speaker: Navin K. Sinha, CUNY, New York, NY
   > Healthcare Research & Development Division
   > Principal Statistical Advisor
   > Date/Time: Friday, October 23, 2009 / 5:30 - 7:30 p.m.
   > Location: University of Minnesota, 2-580 Moos Tower (Malcolm Moos Health Science Tower, 515 Delaware Street SE, Minneapolis, MN 55455 (south from Washington Ave.)
   > Meeting schedule:
     > 5:30-6:00 p.m. Social and refreshments
     > 6:00-7:00 p.m. Talk
     > 7:00-7:30 p.m. Open discussion

2. From: A9SHEMYAKIN@stthomas.edu
   To: sinhanavin@hotmail.com
   Date: Wed, 21 Oct 2009 17:57:35 -0500
   Subject: RE: October Meeting

   Navin,

   The room we will be in has standard U of M equipment – your camcorder operator might have to work it out whether the inputs needed are available. We will have no other technician at the time of the meeting and fairly basic projector/computer setup. I will have a laptop hooked up so you should be fine with a flash drive for your presentation or you can send your presentation files to me in advance.

   See you on Friday!

   Arkady
Hi Dr. Arkady,

Thanks for all the opportunity. Please find the video as promised. The presentation was 85 minutes, out which I have taken 33.5 minutes. But, it does start with a best looking person in the room - you.

Part 1 (introduction) - [http://www.youtube.com/watch?v=pOtYSwXVKt0](http://www.youtube.com/watch?v=pOtYSwXVKt0)

Part 2 (predictive modeling) - [http://www.youtube.com/watch?v=edIhS7kxucY](http://www.youtube.com/watch?v=edIhS7kxucY)

Part 3 (Data Mining) - [http://www.youtube.com/watch?v=zuTA1EelsTg](http://www.youtube.com/watch?v=zuTA1EelsTg)

Part 4 (Q & A) - [http://www.youtube.com/watch?v=wTDoZ17VpJQ](http://www.youtube.com/watch?v=wTDoZ17VpJQ)

Hope you liked it!

Regards,

Navin

---

From: A9SHEMYAKIN@stthomas.edu
To: sinhanavin@hotmail.com
Date: Mon, 26 Oct 2009 14:59:39 -0500
Subject: RE: October Meeting

It was our distinct pleasure to have you as a speaker at our meeting! I will be looking forward for the video.

Arkady

Arkady Shemyakin, Ph.D.

Associate Professor
Hi Navin, to keep the momentum of our discussion, I made up some idea slides. The first 2 are what I have used in the past--just for you to see. The third one has questions about what we are analyzing and what we are not. The remaining are what we talked about....

On Wed, Jun 3, 2009 at 12:25 PM, Navin Sinha <sinhanavin@hotmail.com> wrote:
Want to talk about two hypothesis: -
1. deterministic gene expression by operan system in the dominant genotype BC- All. Think of dynamical simulation by math equation $y=X^4+X^3+X^2+X$ and other one sent before.
2. Compensatory response gene expression by BC-All dominant genotype, on which I am working on. Lower initial frequencies and residual regression response on both genotype frequencies point to this hypothesis.

I want to go over all the outputs, and explain what i am doing here.

I am sure there are more but I do not understand the biology. If you help me in this area, I find a mathematical model to quantify that hypothesis for your June 18, Harvard Presentation.

More soon,
navin
Bio-IT World 2010 Best Practices Awards

1. Nominating Organization (Fill this out only if you are nominating a group other than your own.)

   A. Nominating Organization
   Organization name: Dolcera
   Address: Address: 201 A S. Delaware St #306, San Mateo, CA 94401

   B. Nominating Contact Person
   Name: Samir Raiyani
   Title: CEO
   Tel: 650-269-7952
   Email: samir.raiyani@dolcera.com

2. User Organization (Organization at which the solution was deployed/applied)

   A. User Organization
   Organization name: Dow Agrosciences
   Address: 9330 Zionsville Road
   Indianapolis, IN 46268

   B. User Organization Contact Person
   Name: Ray Roach
   Title: Technology Leader
   Tel: +1 (989) 636-0453
   Email: rproach@dow.com

3. Project
   Project Title: Innovation Dashboard
   Team Leader:
   Name: Ray Roach
   Title: Technology Leader
   Tel: +1 (989) 636-0453
   Email: rproach@dow.com
   Team members – name(s), title(s) and company (optional):
   o Samir Raiyani, CEO, Dolcera
   o B S Manikandan, Manager, Dolcera

4. Category in which entry is being submitted (1 category per entry)
   ☐ Basic Research & Biological Research: Disease pathway research, applied and basic research
   ☐ Drug Discovery & Development: Compound-focused research, drug safety
   ☐ Clinical Trials & Research: Trial design, eCTD
5. Description of project (4 FIGURES MAXIMUM):

A. **ABSTRACT/SUMMARY** of the project and results (150 words max.)

Dow Agrosciences has been using the Dolcera Collaboration Platform for a variety of Research and Development (R&D), and marketing innovation initiatives within the company. This collaboration platform helps Dow Agrosciences
  o Identify and brainstorm new product opportunities and whitespaces
  o Collaborate across teams and geographies for rapid development of new offerings
  o Track competitors and products from around the world
  o Monitor patent landscape to prevent any possible infringements

The Dolcera Collaboration Platform is a combination of a wiki and a custom-built document dashboard with social computing features (rating, tagging, sharing) that has been adopted by several of the largest divisions within Dow, and is used by their intellectual property (IP) managers, scientists and marketing professionals in US, Europe and Asia.

Using this platform, the Dow Agrosciences teams have:
  o Cut down the costs of tracking competitors’ patents, products and scientific literature by 50-60%
  o Organized and shared hundreds of documents in well-designed taxonomies with a wide range of stakeholders within the organization
  o Generated 2x more ideas during brainstorming sessions for new products
  o Discovered synergies across different divisions and geographically-separated teams because of better information sharing
  o Save Dow Agrosciences potentially costly litigation costs (several million dollars per lawsuit) from potential patent infringement lawsuits

B. **INTRODUCTION**/background/objectives

Technology-driven corporations around the world are under pressure to best manage and share their knowledge resources. Knowledge management tools are often adopted by the information technology (IT) savvy parts of the organization. Web 2.0 technologies – wikis, blogs, rich internet applications (RIA) – have
managed to appeal to non-IT savvy users too and have extended the reach of knowledge management and collaboration tools.

Dolcera has developed a knowledge sharing and collaboration platform using the popular Mediawiki technology (used by Wikipedia), combined with a set of widgets and a document-management application called the Dolcera Dashboard. Users can upload, categorize, rate, tag and share documents and links using these applications. The documents are presented alongside a variety of interactive features such as charts, interactive pathways and gene sequences. The charts and documents can be shared easily across teams. Dow Agrosciences saw potential in these technologies and rolled them out for a set of projects related to new product development and marketing innovation. The applications have gained rapid adoption since their initial roll-out in 2006.

The Dolcera Dashboard is an interactive Web 2.0 technology, used to review large quantities of patent, scientific and product literature in an aggregate fashion. It is used to collaborate with colleagues and partners around the world. The Dashboard is used by R&D teams around the world to monitor the competitive landscape and to find new opportunities for innovation. The dashboard content (patents, scientific literature etc.) is organized by Dolcera’s and Dow’s teams of experts, to reflect the business context of the user.

The dashboard can be used by executives in the organization who want high level information to benchmark research activity against competition. The same dashboard can be used research level folks who want the in-depth technical details of competitive approaches in their area of interest.
Published Resources for the Life Sciences

Screenshot 1: The Dolcera wiki for Dow Agrosciences

Screenshot 2: Interactive enhancements to Mediawiki for pathway mapping
Screenshot 3: The Dolcera Dashboard

- Patent categorization
- Patent timeline view by publication year
- Patent timeline view by application year
- Company name
- Patent type
- Full patent tab

Clicking on any bar will automatically lead to full patents relevant to that bar.

Screenshot 4: The dashboard – details available for research groups on the same dashboard at the click of a button

- Full patent .pdf link
- Patent title, abstract and claim
C. **RESULTS** (highlight major R&D/IT tools deployed; innovative uses of technology).

**Technology platform:**
- The Dolcera Collaboration Platform is built using the Adobe Flex platform and works inside the web browser. The server software is based on PHP and Postgres technologies.
- Wiki: Mediawiki platform with several custom extensions developed by Dolcera.

All the tools used by Dolcera including the dashboard and the wiki are using the on-demand model where the services are hosted by Dolcera and the clients use the web browser to access the information.

Dolcera’s innovations are:
- Adding domain-specific widgets and presentations (e.g. patent-pathway mapping) and security additions to the Mediawiki platform to create a wiki platform suitable for the enterprise
- Creating an easy-to-use document management dashboard that can handle enterprise-scale document needs with social computing features such as document tagging, rating, sharing etc.

D. **ROI achieved or expected** (200 words max.):

In the traditional world, competitive technology information like patents, scientific papers and other market data was stored in various silos in large organizations like Dow Agrosciences. No solution existed that could aggregate data from these sources into a single platform resulting in wasted labor to put together data for knowledge sharing across the organization. Document sharing technologies and portal platforms such as Sharepoint did not make the information readily accessible due to the inherent learning curve associated with them. The user interface of such portals was not very user friendly resulting in low usage of knowledge across the organization.

Easy access to relevant knowledge resources through the Dolcera Collaboration Platform has several benefits for Dow:
- Cut down the costs of tracking competitors’ patents, products and scientific literature by 50-60%
- Organized and shared hundreds of documents in well-designed taxonomies with a wide range of stakeholders within the organization
- Generated 2x more ideas during brainstorming sessions for new products, which, in turn, could generate billions of dollars in new revenues
- Discovered synergies across different divisions and geographically-separated teams because of better information sharing
- Save Dow Agrosciences potentially costly litigation costs (several million dollars per lawsuit) from potential patent infringement lawsuits

E. **CONCLUSIONS/implications for the field.**

The Dolcera Collaboration Platform has been proven to be an effective tool for competitive analysis and for generating new ideas in a variety of scientific domains. As the quantum of information increases, Dolcera’s combination of a trained team of scientific experts and a Web 2.0 technology platform is rapidly finding
acceptance in research-driven companies in biotechnology, telecommunication and alternative energy areas among others.

1. **REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)**


   If interested, Dolcera can provide reviewers access to samples of their work, and can answer any additional questions.
Bio-IT World 2010 Best Practices Awards

1. Nominating Organization (Fill this out only if you are nominating a group other than your own.)

   A. Nominating Organization
   Organization name: 
   Address: 

   B. Nominating Contact Person
   Name: 
   Title: 
   Tel: 
   Email: 

2. User Organization (Organization at which the solution was deployed/applied)

   A. User Organization
   Organization name: goBalto.com, Inc. (www.goBalto.com)
   Address: 260 King Street, Suite 759, San Francisco, CA 94107

   B. User Organization Contact Person
   Name: Jae Chung
   Title: Founder & CEO
   Tel: 415 875 9431
   Email: jchung@goBalto.com
   LinkedIn Group
   Blog: http://blog.gobalto.com
   Twitter: @gobalto

3. Project
   Project Title: Intentionally left blank – Not Applicable
   Team Leader
   Name: 
   Title: 
   Tel: 
   Email: 
   Team members – name(s), title(s) and company (optional):
4. Category in which entry is being submitted (1 category per entry, highlight your choice)

- Basic Research & Biological Research: Disease pathway research, applied and basic research
- Drug Discovery & Development: Compound-focused research, drug safety
- Clinical Trials & Research: Trial design, eCTD
- Translational Medicine: Feedback loops, predictive technologies
- Personalized Medicine: Responders/non-responders, biomarkers
- IT & Informatics: LIMS, High Performance Computing, storage, data visualization, imaging technologies

- Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
- Health-IT: ePrescribing, RHIOs, EMR/PHR
- Manufacturing & Bioprocessing: Mass production, continuous manufacturing

(Bio-IT World reserves the right to re-categorize submissions based on submission or in the event that a category is refined.)

5. Description of project (4 FIGURES MAXIMUM):

A. ABSTRACT/SUMMARY of the project and results (150 words max.)

goBalto.com. The fastest, most-effective way to connect with life-science service providers

Finding a trustworthy pharmaceutical service provider is way too hard. Asking colleagues and consultants might net you a recommendation for the largest CROs, but if you need a good regulatory consultant familiar with Chinese Investigator Document Review, you could be out of luck. And even when a pharma company does identify a service provider, the interaction isn’t straightforward: there is no simple way of searching for and comparing services, profiles and reputation.

Expedia and Yelp have helped consumers easily find hotels and local businesses --and in the process, made these transactions much more transparent--but the life sciences industry has lagged behind: service providers still mainly generate leads by “word of mouth” and attending expensive conferences! Even where services have moved online, the directories are not search friendly and highly disorganized. goBalto.com is filling this void by creating an online marketplace for pharmaceutical contract services of all kinds.

goBalto.com is a trusted online portal for pharmaceutical services, from Contract Research Organizations ("CROs") to Investigator Sites and Contract Manufacturers (CMOs) and Preclinical Researchers. For service providers, goBalto.com offers a virtual home, verifiable reputation, transaction support, and marketing tools. For sponsors, goBalto.com provides a central marketplace to easily find, compare, and contact service providers.
B. INTRODUCTION/background/objectives

If you were to ask a clinical operator of any drug-development sponsor for a list of why clinical trials are delayed, there’s a good chance “Patient Recruitment” would be near the top of the list. There are plenty of reasons why drug-development is damn expensive (and lengthy), but there’s usually one overarching theme: they’re inefficient.

From inflated patient recruitment rates to ill-prepared investigators, clinical trials could probably be held in a fraction of the time if everyone just got their act together. Cue goBalto.com, a new site that hopes to help streamline investigator recruitment, helping you find and connect with investigators in a time efficient manner.

goBalto.com was built by the founder behind Celltrion - a leading biopharmaceutical contract manufacturer - listed on the KOSDAQ with a market value of over $1.6 billion. It was during his experience leading business development, Jae became intimately aware of the inefficiencies associated with managing drug-development collaborations with multiple partners around the world, and how web-based collaboration tools can help bring greater efficiency. To help build the product, he says that company consulted with numerous pharmaceutical companies, CROs, start-ups and investigators to develop the user interface. Of note is Dr. Cary Queen (Co-founder of Protein Design Labs) who serves as a goBalto.com advisor.
Click on this link to view a video introduction to goBalto.com.

C. RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

goBalto.com’s core functionality is to help you quickly get to a “short-list” of qualified investigator sites and distribute multiple Request for Information (“RFI’s”) for your clinical trial. First, you filter from over 20,000 investigators based on phase, region, therapeutic indication and site type.
Then you send a template RFI and add a few comments outlining the trial you’re currently working on. The logic behind this is simple: everyone knows ahead of time what basic questions you’re going to need answered, so there’s less of a chance of off-topic replies.

goBalto.com also has a set of features you can use after you send the RFI. During the planning phase, you can view a dashboard tool that helps you track inquiries and monitor response rates. This is particularly useful to see if you will meet your patient recruitment targets. You can resend inquiries to those who haven’t responded and mark unqualified investigators as Inactive. You can print, download and share your inquiry status, so your co-workers know exactly where things stand, which can later be compiled into a management progress report. The dashboard also includes a snapshot view at the top of the screen that helps you keep track of how long you’ve spent on patient recruitment so that you don’t run over time (and budget).

goBalto.com is intuitive and should be easy to pick up for just about anyone. Given that the service revolves around collaboration, future versions will allow users to create a customized Site Feasibility Questionnaire (“SFQ”). Users not wishing to go through the lengthy exercise of building a SFQ from scratch, will also be able to start from a template, which will include most of the commonly asked questions.
D. **ROI achieved or expected (200 words max.):**

Looking for a service provider and subsequently managing the interaction (ex. RFIs, RFQs, SFQs) can take anywhere from 12~18 months. Using goBalto.com, a user can save 2~3 months, resulting in cost savings of hundreds of thousands of dollars in a drug-development program. Founder Jae Chung likens goBalto.com’s model to Web 2.0’s products, explaining that it’s a straightforward tool that anyone in a company can start using without any involvement from management or IT.

E. **CONCLUSIONS/implications for the field.**

The way I describe goBalto.com’s implications for the drug development field is to liken what the world would be like without travel websites like Expedia.com, Orbitz.com and Travelocity.com. Imagine if you had to find a hotel in a country, where you’ve never been before - say for example, a hotel in Korea. If the websites did not exist, how would you go about finding such a hotel?

- Ask a friend?
- Hire a travel agent?
- Search for the term “Korean hotel” on Google?

All of the approaches above are limited in their ability to provide you with a complete list of hotels with the specific criteria you are looking for. Plus, this also assumes that your friend has been to Korea! If you were to search on the term “Korean hotel” on Google, you would see a list of directories, which summarize the hotels alphabetically from A-Z. Not very user-friendly in helping you filter your results to find precisely what you’re looking for. Finally, you would not know which hotels were good.

All of these issues relate to efficiency and transparency. You would spend days - if not weeks - trying to find a hotel, which meets your needs.

This is precisely the situation we are facing right now when you try to find a life-science service provider. We surveyed over 200 life-science professionals and asked them how they go about finding a service provider or partner. Their responses parallel how users go about searching for a hotel in a pre-Expedia age.

- Word of mouth
- Hired a consultant
- Searched on Google

With the pharmaceutical industry under enormous pressure to increase productivity and reduce costs, they are increasingly turning to the FIPNet\(^1\) model. Due to the “Flattening of the World”\(^2\) as postulated by Thomas Friedman –

---

1. FIPNet--for fully integrated pharmaceutical network. Nowhere is the strategy more apparent than in Asia, where Big Pharma hopes to tap into the increasingly high quality research available in India and China--countries that also should prove to be a major source of future customers. Merck and Eli Lilly are two companies leading the FIPNet charge.

the sheer number of life science providers are growing and are increasingly sprouting in remote regions of the world. Try searching on the term “clinical research organization” on Google, and you will be presented with a list of alphabetized directories, articles and corporate websites. Not very useful in helping you find a qualified service provider.

Imagine how much time you would save, if you could use a search-friendly database, which allows you to find qualified service providers and then read ratings or reviews on that service provider. Similar to how Expedia.com has made it easier for travelers to find hotels around the world, goBalto.com will greatly accelerate development cycle times by enabling sponsors to quickly find a list of life-science providers and help manage the interaction between them.

1. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)

“In today’s highly competitive marketplace, it’s critical to be able to find quality partners that can reduce your costs of drug development and time to market. The goBalto.com platform provides transparency, the ability to quickly identify and evaluate new partners via a rating system and rich vendor profiles.”

- Dr. Cary Queen, Co-Founder of Protein Design Labs

“goBalto.com provides a great platform for showcasing our drug development and manufacturing capabilities to biotech and pharmaceutical companies around the globe. The goBalto.com partnering web site provides us with a highly cost-effective channel to reach our target customers compared to traditional options such as industry conferences and advertising.”

- Kiran-Mazumdar Shaw, Chairman and Managing Director of Biocon Limited

“Have already signed up for goBalto.com. Looks like a great start – many hours work for sure. While Celltrion is a great creation, goBalto.com will make the biggest impact in biopharma, so for the industry’s sake good luck and best wishes for its success.”

- Ross Horsburgh, Kendle International

“Great website! I was skeptical at first but after trying out your search engine for a specialized preclinical CRO & I was sold! Will definitely recommend goBalto.com to my research colleagues.”

- Phong Tran, In-vivo pharmacologist at Regulus Therapeutics

“I believe goBalto.com has a good lead here, which currently does not have much equivalent and could be a definite first mover advantage in the current climate with pressures being exerted on the industry. A proposal ripe for it’s time!”

- Brian White-Guay, Professor, Faculty of Pharmacy at Université de Montréal

---

2 The World Is Flat: A Brief History of the Twenty-First Century is an international bestselling book by Thomas L. Friedman that analyzes globalization, primarily in the early 21st century. The title is a metaphor for viewing the world as a level playing field in terms of commerce, where all competitors have an equal opportunity.
“You triggered a rather interesting web project which will hopefully become a valuable resource and tool to R&D professionals.”
- Aliya Arinova, INC Research

“Love your attempt to capture number of pharma service providers, especially 1100 CRO figure.”
- John Lewis, Vice President of Public Affairs at ACRO (Association of Clinical Research Organizations)

“It took me a while to check out goBalto.com. But now that I have, I must say it is an impressive start. I like that it’s very professional, but friendly and non-intimidating. It is easy to maneuver through and I like the easy to read largish font (I’m getting old). I wondered where the name goBalto.com came from and once I learned of it’s origin, I loved it!
- Martha Onasch, Clinical Trial Management Consultant, Elan

I think goBalto.com has identified an undeveloped niche in the process of drug, biologic and medical device development, and perhaps any development program that relies on hard to find and manage external resources.
- Robert Michalik, Esq, Graduate School Lecturer at Northeastern University

“With the ever increasing pressures to reduce the drug development and manufacturing costs, global outsourcing has now become an integral part of development strategy for research based pharmaceutical companies. To that end, goBalto.com offers a unique and comprehensive set of tools in order to make the process of outsourcing much more efficient and effective. The instant access to CMOs, CROs, Consultants and other service providers will dramatically change the global outsourcing landscape”.
- Dr. Umesh Dalvi, Former Senior Director, Global External Manufacturing, BMS

“goBalto.com’s timing couldn’t be perfect. I am currently an independent Quality Contractor and one of the firms I am working with is looking for a GMP CMO in Asia and/or India. API Manufacturing for a small molecule, however complex process.”
- Elaine Lee, Quality Assurance Representative at Pharmacyclics

“goBalto.com is going to make a tremendous impact in the R&D market.”
- Nicholas Lilmbocker, Regional Director at Chiltern International Limited

“goBalto.com is an interesting concept. During the 17 years I worked with researchers seeking human tissue for their studies, I often observed the delays they encountered in connecting with the proper resources. I will look forward to using goBalto.com.”
- Kelly Feil, Consultant and Writer/Nonprofit Management and Biorepository Professional at Independent
“This is a great idea. Why didn’t we think of this??!! In the short six month, its founders have been able to rack up some impressive partners, including Baxter BioPharma Solutions and India’s Intas Biopharmaceuticals. Again, awesome idea.”

- IguanaBio.com, April 2009

“goBalto.com will be very helpful for my consulting services especially since I don’t have a website yet.”

- Vincent Wu, Principal Consultant & CEO, VLW Associates

“I like the service that goBalto.com is offering. Definitely something that will be of use to the CRO and biotech/pharma industries. I have had good responses from LinkedIn, however goBalto.com is industry specific so it will probably result in more business being generated.”

- John Bishop, CEO, Numeric Focus

“goBalto.com is a great idea and I am sure that it will be very successful and provide a lot of valuable service to many people. I am thrilled about being listed and having premium service.

- Marion A. Howard, Principal at Cambridge BioStrategies

“I am quite impressed with your new venture so far-something that has been long over due in the industry.”

- Patricia Seymour, BioProcess Technology Consultants

“I have seen goBalto.com, its really a nice piece of work for all the people involved in clinical research.”

- Catalyst Clinical

“You triggered a rather interesting web project which will hopefully become a valuable resource and tool to R&D professionals.”

- Aliya Arinova, INC Research

“goBalto.com is a great platform and we have our first inquiry. I wish you all the best.”

- Nagraj Lanka, Vivo Biotech

“Interesting concept – goBalto.com. I’ve got a lot of learning to do and I’m loving it.”

- Barry Prom, Collaborative Drug Discovery, Inc.
“Good stuff you guys are doing. Keep it up and good luck!”
- Steve Ayala, President & COO, Clinverse

“I came across your blog post earlier today and intrigued by it, I explored more on your website. I think this is a great idea. I must say that rating service providers has occurred to me and my colleagues before but I really like what you have done with this. Congratulations on a great start. I believe this will be a great success.”
- Jay Reddy, AAP Pharma Technologies

“I went to your site and viewed the video and found your group to be very interesting, as well as with a young staff and great ideas.”
- Marlene Llopiz, Icon Clinical Research

“I am a great fan of goBalto.com and find it quite useful and impressed by the user friendly interface.”
- Kirsten Suurkask, PharmaBioSource

“I have to congratulate you on a great site.”
- Andy Parrett, Chairman, Pharmaceutical Contract Management Group

“Our industry is changing so rapidly. How do we adapt? You are driving many of the changes, which is really great!”
- Rita Wilby, Planetconnect

“I have seen your website (it is very good), and I saw your You-Tube explanation of how you came up with your business model and plan. You have done a lot of good work.”
- Karl Gallegos, MD, MPH, CPI, Pharmaceutical Medicine Consultant
Bio-IT World 2010 Best Practices Awards

1. **Nominating Organization** (Fill this out only if you are nominating a group other than your own.)

   **A. Nominating Organization**
   Organization name: Hewlett-Packard Company
   Address: 3000 Hanover Street, Palo Alto, CA 94304-1185

   **B. Nominating Contact Person**
   Name: David Medina
   Title: Worldwide Life Science and Pharma Segment Executive, HP Enterprise Business
   Tel: +1 281-493-1582
   Email: david.medina@hp.com

2. **User Organization** (Organization at which the solution was deployed/applied)

   **A. User Organization**
   Organization name: The University of Texas M. D. Anderson Cancer Center Houston, Texas
   Address: 1515 Holcombe Blvd, Houston, Texas 77030

   **B. User Organization Contact Person**
   Name: Lynn H. Vogel, Ph.D.
   Title: Vice President and Chief Information Officer
   Tel: +1 713-745-7960
   Email: LHVogel@mdanderson.org

3. **Project**

   Project Title: The Translational Research Extensible platform (T-REX) Project (now referred to as ResearchStation)
   Team Leader:
   Name: Krishna Sankhavaram
   Title: Director for Research Information Systems and Technology Development
   Tel: +1 713-745-7758
   Email: ksankhav@mdanderson.org
   Team members – name(s), title(s) and company (optional):

   **Key MDACC Team**
   Krishna Sankhavaram
   Momodu Sanyang
   Dr. Latha Ramdas
4. Category in which entry is being submitted (1 category per entry, highlight your choice)

☐ Basic Research & Biological Research: Disease pathway research, applied and basic research
☐ Drug Discovery & Development: Compound-focused research, drug safety
☐ Clinical Trials & Research: Trial design, eCTD
☐ Translational Medicine: Feedback loops, predictive technologies
☐ Personalized Medicine: Responders/non-responders, biomarkers
☐ IT & Informatics: LIMS, High Performance Computing, storage, data visualization, imaging technologies
☑ Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
☐ Health-IT: ePrescribing, RHIOs, EMR/PHR
☐ Manufacturing & Bioprocessing: Mass production, continuous manufacturing

(Bio-IT World reserves the right to re-categorize submissions based on submission or in the event that a category is refined.)

5. Description of project (4 FIGURES MAXIMUM):

A. ABSTRACT/SUMMARY of the project and results (150 words max.)

M.D. Anderson Cancer Center is dedicated to eliminating cancer with integrated programs in cancer treatment, clinical trials, education programs and cancer prevention. Personalized, translational medicine remains a core focus, but M.D. Anderson struggled with common challenges – developing “personalized molecular medicine” that treats cancer and integrating genomic data with clinical
data to enhance the quality of clinical decisions. In a strategic partnership with HP, M. D. Anderson’s IT division developed a unique SOA-based IT platform to easily access, integrate and analyze genomic and clinical data. This application, called ResearchStation, enables collaboration between life science researchers utilizing disparate data sources and the analytic tools of their choice, and demonstrates the effectiveness of a SOA-based architecture in enabling interoperability in support of translational research.

B. INTRODUCTION/background/objectives

In 2005, the National Cancer Institute (NCI) initiated a collaboration with the National Human Genome Research Institute (NHGRI) to pursue a 3-year pilot project known as The Cancer Genome Atlas (TCGA) initiative, to determine the feasibility of comprehensively cataloging the genomic alterations associated with a set of human cancers.

A central Human Cancer Biospecimen Core Resource (BCR) established by TCGA became the biospecimen resource center and the primary interface with the clinical sites at which donor samples and clinical data are collected and M.D. Anderson was selected to be a primary contributor.

At the time, the current environment at M.D. Anderson did not support comprehensive analysis of genomic and proteomic data. In 2008, the T-REX (Translational Research Extensible Platform) Project was born as a way to:

- Accelerate the identification of cancer cures
- Tie in systems, tools, algorithms to perform analyses and connect patient clinical and genomic data
- Consolidate disparate toolsets and promote uniformity
- Provide ability to integrate research tools and formulate intelligent queries
- Establish better collaboration among researchers, clinicians
- Through IT, change the clinician’s behaviors and in the process change the health of MDACC patients

C. RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

In an innovative project, M.D. Anderson Cancer Center (MDACC) collaborated with Hewlett-Packard to develop a platform for collaboration, data management and analysis interfaces for their researchers. The result of this is the development of the T-REX (Translation Research Extensible) platform and ResearchStation applications for collaborative translational research. T-REX is a service-oriented architecture platform that integrates genomic, transcriptomic, and clinical data and access to analytic and research applications that address the needs of researchers with the capability and plans for adding proteomic data in the near future. Sitting atop the T-REX platform is Research Station, an application that provides life science research workflow and collaboration tools. Research Station provides life science researchers with an integrated view of data sources
such as TCGA, MDACC or other public data as well as ability to connect with open source and proprietary analytic tools and applications commonly used by life science researchers at MD Anderson.

The T-REX platform encompasses the following elements:
- Extensible platform based on a robust service-oriented architecture (SOA)
- SOA allows for easy addition of additional data sources. The formats of the data sources are configurable in the application and hence allow extensibility to new data types.
- SOA allows for easy addition of external analytic tools and applications allowing MDACC to leverage current investments in analytic tools as well as providing researchers access to analytic tools that meets their needs.
- T-REX’s SOA architecture allows MDACC to connect the research applications and data sources with other MDACC systems and data, such as clinical, laboratory, etc.
- T-REX has a plug-in facility to add analysis modules written by the bio-analysts to analyze the genomic data from these data sources. This allows the platform to grow along with the new analysis techniques from the bio-informatics world.

ResearchStation builds on the following key factors:
- Provides an integrated view across data types through user query mechanisms
- Provides unique user interface that allows researchers to visualize available data sources
- Provides researches with a basic collaboration workflow allowing them to easily share data, analyses, findings, hypotheses and comments with other researchers.
- Allows for capture of information from external analytic tools and applications
- Supports privacy, security and confidentiality requirements of life science research
- Supports at least three disease data sites, and allows for more to be added (version 1 addressed only one disease site – Glioma)
- A personalized system with an intuitive interface
- Combines TCGA and MDACC data – maps between research & clinical data; and
- Provides visualization capabilities for results of genomic analyses

The scope included the following work:
- Develop a method for seamless integration of genomics, and transcriptomics data from different platforms into a common data structure with the capability of adding proteomics data in the future
- Create the foundation of integrated data management and analysis.

Technologies Used:
- All the backend services are developed using JavaEE5 stateless session beans and exposed as SOAP services. The long running transactions such as analysis modules used JavaEE5 message driven beans
The Research Station front-end and Visualization are developed using Eclipse RCP. The web based front-end development is in progress.

The administrative interface was built using JSF.

Informatica is used to transform and load data from the data sources.

The analysis modules written in R are plugged in via XML based configuration registry.

Security between the Research station clients running on the end user machines and the services on the server side is implemented using WS Security and SAML.

The data services are exposed as RESTful interfaces using JAX-RS. This allows the bio-analysts to use the data directly from their analysis modules without going through the complex SOAP interfaces.

Hardware/software activities for ResearchStation include:

- Computing platform requirements and specifications
- Network requirements and specifications
- Operating system requirements and specifications
- Storage/database requirements and specifications

Software development activities include:

- Iterative design and development of components
- Implementing database schema and other objects
- Assisting the users during acceptance testing
- Creating test plans and test cases
- Creating and executing manual test scripts
- Preparing installation and configuration documentation.
- Transferring knowledge to selected MDACC personnel on the usage and administration of the system
- Testing the releases in the MDACC
- Testing/UAT environment and providing installers, packaged components and relevant installation documentation to enable MDACC to deploy the system in their staging/production environment

Transition (implementation) activities include:

- Transition of Source Repository, build procedures, and Development Environment to the MDACC developer community.
- Transition of any needed operations (security, installation, maintenance, backup, recovery, configuration, and systems operations) and monitoring to the MDACC operations and support community and to RIS Administration.

The HP total man-hours devoted to the project approximate 48,000. Over the course of the project MD Anderson Cancer Center expended approximately 4500 man-hours, including approximately 2,000 – 2,500 for Dr. Latha Ramdas alone.
D. **ROI** achieved or expected (200 words max.):

ResearchStation enables researchers and other consumers to frame intelligent questions and perform further analyses, test hypotheses, etc. The application provides a platform for various cancer researchers to access, analyze and display results from various data sources in an intuitive and integrated fashion including:

- Graphical visualization to find specific genetic markers, helping to accelerate the research process
- Enhancing research capability to advance finding cures
- Enhanced collaboration in the cancer research & treatment community
- SOA platform with re-usable services that can be built upon

E. **CONCLUSIONS**/implications for the field.

Successful completion of the pilot project focused on three tumor types (glioblastoma multiforme, squamous cell lung cancer, and serous ovarian cancer). Collectively, genomic and clinical data generated by all the components of the pilot project provide the initial contributions to a comprehensive Web-based resource describing the genomic “fingerprints" of specific cancer types. This resource is known as The Cancer Genome Atlas (TCGA).

The pilot demonstrated that a platform and application based on a service-oriented architecture approach provided significant benefits to the research process. SOA allows life science organizations to easily add data sources, reducing the need to move data, ensuring that data has a single source of truth and allowing the development of a “flexible" virtual data model. The pilot also demonstrated that analytic tools and external applications (open source and proprietary) can easily be integrated into a research platform, ensuring that life science organizations are able to leverage existing investments in analytical tools so that researchers have access to the analytic tools and applications that meet the needs and requirements of their researchers.

6. **REFERENCES**/testimonials/supporting internal documents *(If necessary; 5 pages max.)*

“The Human Genome Project and the Cancer Genome Atlas have the potential to transform the way in which we treat patients by delivering on the promise of personalized medicine. However the information needed to have an impact is hidden from those who can make a difference by the inclusion of too much data.

T-REX and its transition to ResearchStation is designed to provide methods for the physician and the researcher to visualize and query the data present in The Human Genome Project and the Cancer Genome Atlas and the patient record in a manner that will allow them to develop and test hypotheses. By making the data accessible and intuitive to the user, T-REX and ResearchStation have the ability to help us realize the potential envisioned by those who dreamed big and planned and implemented The Human Genome Project and the Cancer Genome Atlas."
Dr. Gordon Mills, Professor and Chairman, Systems Biology, Ann Rife Cox Chair in Gynecology. University of Texas, MD Anderson Cancer Center, Houston, TX

Key notes:
**T-REX Platform Functional View of Research Supported Processes**

This diagram is a functional representation of the major actors, systems and processes that the T-REX platform and ResearchStation interact with. This diagram demonstrates that the platform’s service-oriented architecture allows for the integration of external applications and websites. For example, integration of DNA copy data results can be visualized using the UCSC Genome Browser. This also enables the use of the capabilities of the UCSC Genome Browser through the T-REX platform, such as the ability to click on a SNP of interest and gain access to SNP details on the NCBI website through the browser.
Research Workflow Process Example

Process
- Review Available Data
- Select Cohort
- Run Analysis
- Review Results
- View Results
- Analyze Results
- Develop Report
- Send to Collaborator

Services
- Genomic Data
- Clinical Data
- R
- Matlab
- JAVA
- T-REX Visualization
- UCSC Browser
- IPA

External Applications

Presentation Layer

Services Layer for Integration

Business Layer
- Application Logic
  - I2B2 Hive Components
  - Java/EE5
  - Analysis Logic
- R, Matlab, C, Java
- Analysis Registry
- LIMS
  - Workflow Mgmt, Inventory Mgmt, Sample Mgmt

Data Access Layer
- MDACC
  - Genomic Data Archive
- Other MDACC
  - Data Sources
- External
  - Data Sources

Clients (MAC, Linux, Windows)
- Web Browser
- RIS Workbench

Glioma/TCGA App
- TMA Lab App
- Genomics Lab App

Common Lab Services

External Application
- UCSC Browser
- IPA
- Nexus

*based on early work on Translational Research
Extensible Platform -

ResearchStaton Architecture
ResearchStation Sample Screen Shots

Browse Map of Available Data Sources in ResearchStation

Browse Table of Available Data Sources in ResearchStation
Whole Genome Analysis Visualization in ResearchStation

Manage Projects in ResearchStation
View Collaborators’ Analysis Results in ResearchStation

Integration with Third Party Application (Ingenuity Pathway Analysis)
Bio-IT World 2010 Best Practices Awards

A. Nominating Organization (Fill this out only if you are nominating a group other than your own.)

A. Nominating Organization
Organization name: KAI Research, Inc.
Address: 11300 Rockville Pike, Suite 500, Rockville, MD 20852

B. Nominating Contact Person
Name: Selma Kunitz, Ph. D.
Title: President
Tel: (301)770-2730
Email: skunitz@kai-research.com

B. User Organization (Organization at which the solution was deployed/applied)

A. User Organization
Organization name: KAI Research, Inc., an Altarum Company
Address: 11300 Rockville Pike, Suite 500, Rockville, MD 20852

B. User Organization Contact Person
Name: Selma Kunitz, Ph. D.
Title: President
Tel: (301)770-2730
Email: skunitz@kai-research.com

3. Project
Project Title: Implementation of the Pediatric Common Data Elements and Data Repository
Team Leader: Patti Shugarts & Yun Lu, Ph.D.
Name: Selma Kunitz, Ph.D.
Title: President
Tel: 301-770-2730
Email: skunitz@kai-research.com
Team members – name(s), title(s) and company (optional):
Rene Kozloff, Ph.D., Exec. Vice President
Patti Shugarts, Director, Government Operations
Yun Lu, Manager, Clinical Management Systems
Ben Piper, Research Associate
4. Category in which entry is being submitted (1 category per entry, highlight your choice)

- **Basic Research & Biological Research**: Disease pathway research, applied and basic research
- **Drug Discovery & Development**: Compound-focused research, drug safety
- **Clinical Trials & Research**: Trial design, eCTD
- **Translational Medicine**: Feedback loops, predictive technologies
- **Personalized Medicine**: Responders/non-responders, biomarkers
- **IT & Informatics**: LIMS, High Performance Computing, storage, data visualization, imaging technologies
- **Knowledge Management**: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
- **Health-IT**: ePrescribing, RHIOs, EMR/PHR
- **Manufacturing & Bioprocessing**: Mass production, continuous manufacturing

(Bio-IT World reserves the right to re-categorize submissions based on submission or in the event that a category is refined.)

5. Description of project (4 FIGURES MAXIMUM):

   A. **ABSTRACT/SUMMARY** of the project and results (150 words max.)

   To provide investigators with a “universal language” to utilize within Pediatric Pharmacologic Research Unit (PPRU) studies, streamline the implementation of pediatric clinical studies and facilitate data aggregation across studies with small populations, the National Institute of Child Health and Development (NICHD), National Institute of Health (NIH) together with KAI Research, Inc. (KAI), an Altarum Company initiated the PPRU Toolbox and the Pediatric Data Repository (PeDaR) project. KAI, as the Coordinating Center for NICHD’s PPRU Network, has developed a systematic approach to the development and implementation of pediatric clinical trial data standards for data collection, analysis, sharing and aggregation. A Web portal that houses a Toolbox and data repository was developed and implemented.

   B. **INTRODUCTION/background/objectives**

   The Pediatric Pharmacology Research Units Network (PPRU), funded by the National Institute of Child Health and Human Development (NICHD), National Institutes of Health (NIH) recognized that it could improve efficiency of conducting clinical studies if it developed a set of common data elements (CDEs), related tools, and a Data Warehouse to assist its investigators.

   In the past several years, several factors have coalesced to stimulate interest in CDEs and data warehouses: recognition that there is much redundancy across clinical trials with respect to development of study forms and clinical data management systems; constriction of funds for clinical research; and introduction of mandatory data sharing plans in 2003 for studies receiving over $500,000 in one year.
Typically, there is great overlap in data requirements across clinical trials and the potential benefits of identifying CDEs are numerous. The use of CDEs can:

- **Facilitate study start up** – Investigators can access the CDEs and forms to greatly reduce the time spent in developing them for each study.

- **Simplify presentations** – CDEs facilitate preparation of data for Data and Safety Monitoring Boards (DSMBs) as there is a standard format for these reports.

- **Simplify data sharing** – Investigators are required to share study data and the common data elements and definitions will simplify the submission of these data sets.

- **Facilitate data aggregation** – Common forms and data definitions allow analyses across studies providing access to a larger study population. There is a wealth of pediatric clinical data in the clinical studies NICHD has funded that could provide a resource for the design of future clinical studies.

KAI, as the PPRU Coordinating Center, has worked with Network investigators to develop CDEs and related tools that form the PPRU Toolbox and the Pediatric Clinical Data Repository (PeDaR). Together they provide a useful research resource for the Institute and its investigators.

The **Toolbox** is a central resource of templates and guidelines for protocol development, standardized data elements and data collection forms, manuals of procedures (MOP), as well as tools for study and data management.

PeDaR provides a secure storage and retrieval environment for completed study data using a common set of terms and definitions to promote data sharing and metadata analysis. It contains data and protocol descriptors to facilitate queries and analyses across studies. Data elements are created following standards developed by the Clinical Data Interchange Standards Consortium (CDISC) to allow for improved data portability. We have developed processes for easily transforming study data into the warehouse and have successfully loaded data sets from completed PPRU studies.

In its development activities with the PPRU and other Institutes, KAI has addressed issues related to the stringent protection of individual privacy as mandated by Health Insurance Portability and Accountability Act (HIPAA) and with PeDaR has created a useful data warehouse for archiving and disseminating data sets.

C. **RESULTS** (highlight major R&D/IT tools deployed; innovative uses of technology).

The toolbox is built upon existing and widely adopted CDISC SDTM and CDASH standards that can easily interoperate with other standardization efforts. It contains 242 Common data elements.
(CDEs), library of 25 forms, form instructions, study and data management templates and

guidelines.

Three modules have been developed for the data repository and ten PPRU legacy studies mapped
to PeDaR.

- **Study Archives**: Allows researchers to download data files from completed studies to conduct
  analysis. Study and data documentation describing the data source is linked to each study.
  Information will be available through the PeDaR Website.

- **PeDaR Datasets Extraction**: Enables data extraction and filtering of datasets from PeDaR for
  use by researchers.

- **Queries and Reports**: PeDaR will not only provide a single secure location for the storage of all
  completed study data, but also will promote data sharing and meta data analysis across studies.
  KAI has worked with the PPRU community to gather and develop requirements on the user
  interface for Query and Analysis tools. With 10 PPRU legacy studies mapped to PeDaR, we
  developed a multiple dynamic queries and graphical reports related to subject profiles,
  pharmacokinetics and safety evaluation based on SDTM v3.1.1 standard.

The PPRU Web portal for the toolbox and data repository were developed and implemented with
MS SQL database, Microsoft’s .net technology, MS SQL reporting service and CDISC define.xml
model. The toolbox database schema and tables were built upon existing and widely adopted CDISC
SDTM and CDASH standards that can easily interoperate with other CDE efforts. The Pediatric Data
Repository (PeDaR), utilized the SDTM v3.1.1 standard as part of a current MS SQL database
standard for the repository. Additional pediatric data elements were created following CDISC SDTM
standards. Protocol elements were defined based on ClinicalTrials.gov and used as parameters for
users to select and aggregate interested studies. Currently, we are expanding the data repository to
enable translational and basic research interface and pharmacokinetic/ pharmacodynamic analysis
as well as development of analytical tools across studies.

In addition to the “general” CDEs, we are developing data elements for specific disease areas
including pediatric hypertension and neonatology. These disease-specific CDEs will be added to the
toolbox library as they are finalized. To define CDEs for these specific pediatric diseases, the PPRU
has convened working groups of experts to participate in the identification process.

D. **ROI** achieved or expected (200 words max.):
The implementation of the PPRU Toolbox and PeDaR has demonstrated that it promotes data
quality, decreases the time and resources needed to develop a study database, and helps customize
the Data Management Plan. The incorporation of standards upstream in the study life cycle
expedites the transformation of study data into the PeDaR, improves the accessibility of study data,
and helps standardize the Extraction, Transformation, and Loading (ETL) processes. The PeDaR
Published Resources for the Life Sciences

provides warehouse capability for completed studies, facilitates data exploration across studies and promotes interoperability with other standards.

- Common toolbox libraries with 242 CDEs, 25 CRF templates and study and data management tips
- Study implemented within 4 weeks using the tools
- Ten completed pediatric NICHD legacy studies have been mapped into PeDaR
- Five graphical/tabular dynamic queries and reports were developed for across study analysis, more will come soon as we are working with PK and other therapeutic working groups
- PPRU Repository governance documents and policies were in place to ensure human subject protection according to HIPPA requirements as well as appropriate procedures and processes for the submission, storage and usage of the data stored in the PPRU repository
- Interoperation with CDISC, caBIG and other standards

E. CONCLUSIONS/implications for the field.
The PPRU Toolbox and repository are generic products and can be integrated into clinical trials by other pediatric networks. This is the first system of its kind that specifically addresses pediatric studies and interoperates with widely adopted standards such as CDISC and caBIG. The developed PPRU portal can be used for both labeling studies and academic network studies.

1. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)

Please find the following supporting document attached to the email:

Nominating Organization:

- Organization Name: Merck and Co., Inc.
- Address: 770 Sumneytown Pike, WP97A-133, West Point PA, 19486

Nominating Contact Person

- Nominating Name: Jaime Melendez
- Title: Innovation Experiment Lead
- Phone: 215-652-0351
- Email: jaime_melendez@merck.com

User Organization

- Organization Name: Merck and Co., Inc.
- Address: 770 Sumneytown Pike, WP97A-133, West Point PA, 19486

Contact Person

- Contact Person: Jaime Melendez
- Title: Innovation Experiment Lead
- Phone: 215-652-0351
- Email: jaime_melendez@merck.com

Project

Project Title: A Virtual Technology Symposium

Team Leader:

Name: Jaime Melendez

Title: Innovation Experiment Lead

Company: Merck and Co., Inc.

Contact Info: jaime_melendez@merck.com

Team Members (Optional)

- Christopher J. Welch, Merck, Distinguished Senior Investigator, Process Research
- Sanjoy Ray, Merck, Director MRL-IT Innovation
- Thomas Fare, Merck, Director Technology Licensing Integrator
A: ABSTRACT/SUMMARY of the project and results (800 characters max.)

Life in a scientific environment is fast-paced and requires a significant amount of information sharing which include face-to-face interactions, e-mails, and net-. While these forms of interactions are successful for most meetings, they also have their limitations when collaborating globally. The purpose of this proof-of-concept experiment was to evaluate the effects and benefits a virtual environment solution would have on internal meetings, specifically observing the effects on collaboration, travel costs and scheduling constraints. The results of the experiment demonstrated that a virtual environment solution can effectively support a meeting through the use of the additional features it can bring in addition to demonstrating that participants were supportive of such a technology.

B: INTRODUCTION/background/objectives

The Merck Technology Symposium is a premier conference for technology sharing and collaboration held every year. This event, which is sponsored by Merck’s New Technology Research Licensing Committee [NT-RLC], enables employees with an opportunity to learn about new research technologies while providing them a chance to compare products and services of various technology solutions.

By providing a medium where employees can all attend in one common location with vendors and other employees, an increase in collaboration and discussion can arise. While discussions may persist beyond the initial event, most of the knowledge/information generated from the discussions is not easily retained or shared for future reference. Additionally, many events that occur at this event are constrained by time and require time commitment in order for interested participants to attend. Some of these interested participants aren’t able to make the necessary plans and schedule readjustments required to travel out of the office and therefore are unable to attend the event.

Consistent with its mission of continuously reviewing new ways to provide effective knowledge sharing / collaboration, the NT-RLC identified a capability that would allow employees to attend these events virtually and interact with other participants remotely – a virtual environment. While the idea seemed promising, it was unknown whether the solution would be successful in supporting virtual scientific collaboration. To understand the effects a virtual environment would have on meetings, a project team was assembled to evaluate specific elements:

- Do the attendees see any improvement regarding collaboration and information sharing
- Does the virtual event make it easier for attendees to participate
- What impact does it have on travel costs

C: RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

The results of the experiment were very positive in that the solution demonstrated valuable information regarding the capabilities of a virtual environment and its effect on supporting a meeting. The first goal – Increasing Collaboration and Information Sharing – was achieved through the use of the Voice over Internet Protocol [VoIP] technology. Through the use of headsets with microphones, users were able to have both formal and informal conversations through the Virtual Environment software. This approach would allow participants to interact verbally in the same free-form manner they would normally communicate in a real-world scenario with the added convenience of being in a remote location.

The format of this virtual scientific poster session required that technical poster displays be formatted and on display in the virtual environment prior to the event starting. In a traditional poster session, multiple posters are on display that allows free-form viewing and discussions to occur. While there were a few challenges in designing a space to support a dynamic event, many of them were well addressed by crafty design of the virtual meeting spaces. One particular challenge was providing users the ability to share documents and presentations that were not originally placed within the virtual world. To address this situation, an application sharing feature was incorporated into the virtual environment solution that allowed users to flexibly add supporting information to the live discussions.

It was discovered during this event that participants felt it was easier to communicate and collaborate with others while in the virtual world because they felt more comfortable having conversations with others relative to a real conference. Survey responses indicated that scientists were more willing to approach senior executives in the virtual environment than they would in a traditional face-to-face meeting. The ability to present yourself as a slimmer, younger, and more attractive avatar when compared to your real-life counterpart could be a reason why participants felt more at ease. Another possible reason for the increased communication and collaboration amongst virtual attendees could be the lack of visible display of rank for the various avatars thereby making all attendees be viewed as equals. If these effects continue beyond the experiment it could prove valuable as information and ideas from the bench level can be shared between junior team members and senior executives.

Following the success of the evaluation, many business areas understand and saw the value of using such a solution to support their business activities. Example implementations of the solution currently include the use of the system as a place for teams to meet and discuss project activities while not having to worry about travel or room availability. Additionally, this solution has also been implemented as an internal training tool for new employees. Simulating a daily, work environment for the employees allows them to get an understanding of the best practices and safety protocols associated with their job activities. Using an actual work environment for
training purposes would be costly, and depending on the role, even impractical. Virtual environments allow for the new employees to see this work environment in its true form but safely and conveniently from their computer.

D: ROI achieved or expected (1000 characters max.):

While it is important to mention that the virtual environment tends to be more expensive to develop, having the ability of the virtual world persist long after the meeting itself has ended proves to be significantly valuable due to the new thoughts or ideas are captured at a later time. Since this event was only an experiment to examine the capabilities of a virtual environment solution, the participant pool was limited to 50 so that it was small enough for a fast and cheap experiment but large enough to provide qualitative and quantitative data. Based on travel estimates from previous events, it was calculated that this small event had contributed to a time saving of over 60 hours and a cost saving of over $8,000. Extrapolating those time and cost savings to the real-world scenario of 500 participants, the virtual environment has the potential to save over 680 hours with a potential cost saving of over $90,000 – this is just for travel alone! Adding the cost of hoteling, food, airfare and associated printing costs that would be eliminated by utilizing a virtual environment would further increase the savings.

E: CONCLUSIONS/implications for the field (800 characters max.)

This experiment proved valuable in understanding how a virtual environment solution could support a business meeting when compared to a traditional face-to-face event. Example findings from the event were easier information sharing, an observed increase in interactions amongst scientists and senior executives, reduced travel costs, reduced time commitment required to attend a virtual event in comparison to a physical meeting resulting in an increase in participation – the end result is a higher likelihood of time-constrained senior scientists interacting and sharing their knowledge with junior scientists.

Collaboration will always be needed to discover and share scientific information. Providing an always on, always available virtual environment will allow scientists to leverage its capabilities to work more efficiently.

F: REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)

Figure Description: A virtual meeting environment. a. A schematic of the meeting space layout showing the gathering vestibule (1), poster halls (2 and 3) and roundtable discussion rooms (4-11). b) The vestibule area and the entrances to the poster halls. c) A poster hall containing 10 double-sided poster display areas. d) Private conversation areas within a poster hall — light blue rings delineate conversation zones.
Bio-IT World 2010 Best Practices Awards

1. **Nominating Organization** (Fill this out only if you are nominating a group other than your own.)

   A. Nominating Organization
   Organization name:
   Address:

   B. Nominating Contact Person
   Name:
   Title:
   Tel:
   Email:

2. **User Organization** (Organization at which the solution was deployed/applied)

   A. User Organization
   Organization name: Molecular Connections Private Ltd.
   Address: Kandala Mansions,
   #2/2, Kariappa Road, Basavanagudi,
   Bangalore - 560 004. India.
   phone: +9180-4120 5016
   fax No: +9180-4120 8956

   B. User Organization Contact Person
   Name: Ms. Sapna Y. R.
   Title: Sr. Manager Business Development
   Tel: +9180-4120 5016
   Email: sapna@molecularconnections.com

3. **Project**

   Project Title: XTractor System: A platform for discovery, analysis and modeling of published biomedical facts from PubMed.
   Team Leader
   Name: Guruprasad Rao
   Title: Principal Scientist
   Tel: +91 80 4120 5016
   Email: guru@molecularconnections.com
   Team members – name(s), title(s) and company (optional):

4. **Category in which entry is being submitted** (1 category per entry, highlight your choice)
5. Description of project (4 FIGURES MAXIMUM):

A. **ABSTRACT/SUMMARY** of the project and results (150 words max.)

Molecular Connections’ XTractor (http://www.xtractor.in) and XTractor Premium (http://www.xtractor.in/premium) - a Semantic Knowledge discovery & Expert curated Knowledgebase platform of published biomedical facts from PubMed. XTractor enables researchers to cut short their time on data analysis and helps them to stay up-to-date with the most recent published scientific data from PubMed with its advanced semantic search, concept linking analytics for effective knowledge discovery and modeling. It provides manually annotated and classified relationships on biomarkers, knockout studies, mutations, disease mechanisms, pathways, clinical trials and much more.

We use a hybrid text mining approach to generate the content in XTractor, which involves NLP based text mining followed by manual annotation. This enables us to reduce the turnaround time as well as maintain the highest standards of accurate annotation. The underlying knowledgebase consists of more than 500,000 relationships on interactions, regulations, and modulations between the biomedical entities and is updated everyday with relationships from the latest publications in PubMed. Presently more than 2500 researchers are using XTractor across the globe.

B. **INTRODUCTION/background/objectives**

Published Literature findings act as a key driver for decision making in drug discovery and biotechnology industry. Majority of the researchers prefer PubMed for published literature since it is a free resource for dissemination of scientific information and data mining biomedical information. Utilizing this information from PubMed has been a long-standing problem for many researchers across the globe. On an average more than 50,000 scientific abstracts get published in PubMed every month and for a researcher it is a highly time consuming task to pick the most relevant abstracts everyday and annotate them accurately for research purposes. Searching PubMed usually results in lots of results- so sorting the relevant ones and making scientific conclusions becomes a time-consuming task.
Though many text mining NLP (Natural Language Processing) engines are used to address this problem, manually annotated data is the best in terms of accuracy and quality of the data that is captured. However, one of the major disadvantages with manual data mining still happens to be the processing time and its cost intensive nature.

So as an example, for a drug discovery researcher interested in target identification or biomarker research or toxicity based studies, it presently costs time and money to go through all the regular relevant updates of PubMed and also to keep oneself updated with the latest findings in their areas of research.

To a researcher the accuracy and being current in scientific information is of utmost importance. To address this need we designed XTractor (http://www.xtractor.in).

XTractor is proven to be highly accurate and more efficient than many Natural language Processing (NLP) engines since a hybrid technology of “semi-automated data mining”, is used. This process technology involves a layer of Natural Language based text mining followed by a layer of manual validation and thus enables us to generate the most accurate facts for genes/proteins, diseases, drugs and biological processes.

Since the annotation is accurate it is possible to perform complex semantic searches and retrieve the most complex relations in PubMed, which is currently not possible with the conventional NLP systems or PubMed.

We have been able to achieve up to 99% accuracy in term pickups and relationship extraction with the XTractor system.

A few advantages of the XTractor system are as follows:

- Sentences/facts are manually validated and classified into 24 categories such as Biomarker-Disease, Drug- Gene, Gene- Process and many other relevant categories. So searching PUBMED and extracting relationships becomes simpler and more effective.
- With XTractor, the entities/terms in the sentences are manually categorized to public biological ontologies (MeSH, GO, Pub Chem and SwissProt) and it also provides users with the ability to create their own databases of sentences and relations for their sets of Keywords.
- Perform Semantic searches and high end analytics over the knowledgebase
- XTractor acts as an alert service and keeps researchers up-to-date with the latest publications, as and when it gets published at PUBMED for their choice of Keywords.
- XTractor also provides the user with ability to change Keywords preferences from time to time.

C. RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

A unique 3-step process to mine Literature abstracts from PubMed.

Step 1- Downloads: The abstracts from PubMed are downloaded every single day using an automated protocol.

Step 2- Text mining: Later these abstracts are processed through a text-mining engine to carry out annotation of abstracts for Proteins, Drugs, Diseases and Biological Processes.

Step 3- Manual Annotation and Categorization:

a. As a next step, XTractor curation team consisting of Masters and PhDs from Life sciences screen through the annotated abstracts, add/edit the annotations that are applied to a text mined output.
b. The curators also ensure that each of the entities are mapped to their respective ontologies/standards such as Swiss Prot for Proteins, PubChem for Drugs, MeSH for diseases and Gene Ontology for Biological Processes. This step enables us to ensure the accuracy of the annotation.

c. Following, which the XTractor curation team picks the most relevant facts, which represent relationships between biomedical entities and categorize these findings to multiple categories such as clinical trial related information, biomarker studies, knockout studies, mutation studies and 24 such categories.

**Figure 1: Sentence Display panel and Disease Report**

A curation-training manual that has been developed covers all the aspects pertaining to processing of information from biomedical information. The curators follow a definite set of guidelines on which they are trained before they initiate the curation efforts.

The curation cycle operates on a weekly basis with regular updations from PubMed for every week that get added to the knowledgebase. The team works towards gathering this set of information on an everyday basis. The workflow is stream lined in such a way so as to ensure that the process takes minimal turnaround time.

The information thus generated is stored in the XTractor knowledgebase, which presently contains more than 500,000 manually extracted biomedical relationships. On an average more than 600-1000 facts get added to the XTractor knowledgebase on a daily basis.

The information that is stored in the XTractor knowledgebase can be easily accessed for FREE using XTractor alert and analysis system (http://www.xtractor.in). The users can register for FREE and set email alerts for their favorite biomedical entities. As soon as the information is updated in the XTractor knowledgebase the users receive an email alert. The users can then analyze the stored information for their queries and also use analytical tools for figuring out the most highly co-occurring entities and saving the required set of facts. Tag clouds feature enables the user to get instant frequency of occurrence of the terms and it also acts as a collaborative platform where in one could tag and share the information with peers, research groups or common interest groups.

XTractor is also available with advanced search and analysis futures for users, who wish to carry out in depth content analysis called XTractor Premium (http://www.xtractor.in/premium).

**Figure 2: Using Summary Search to perform a quick view analysis**

**Figure 3: Concept Linking and Semantic Search Features of XTractor**

This version comes with the following features such as:

1. Summary Search- to obtain downloadable reports and graphs
2. Semantic Search- includes ontology based search for easy refining
3. Bibliographic Search- to track competition and the hot areas of research
4. Concept Linking- to associate entities & discover newer concepts / relationships, which were earlier, not related
5. Watch List- to track favorite proteins or drugs as & when latest data gets added to XTractor
6. Network Visualization- integrated with Cytoscape
7. Save session and retrieve history of search
8. Downloadable reports in XML & pdf formats

**IT Architecture:**

XTractor is a 100% java web application that is developed using the Model-View-Controller architecture. All external technologies used in the implementation of XTractor are open source and java-based allowing the flexibility to run the software on all major hardware platforms with no proprietary restrictions. The application is deployed on servers running Linux Operating system.

For data storage XTractor uses MySQL because of its consistent fast performance, high reliability & ease of use. The core data model consists of four elements: biomedical entities (approx 23,000), relationship between the biomedical entities (also referred to as facts i.e. approx. 500,000), meta-data (supporting the entities and relationships eg: synonyms, ontologies, evidence sentence from source article, etc) and links to popular external public databases. The powerful query and analysis features work with this large volume of data elements using the open source Lucene index engine.

As described above the IT infrastructure caters to two user groups i.e. the data submission by domain experts over the LAN (post the information aggregation and NLP processing) and researchers performing complex analysis of facts over the Internet. The jakarta tomcat servers support the three function of annotation submission, posting email alerts and querying the knowledgebase. The middle tier of the application architecture uses the Jakarta Struts library for clearly separating application logic from HTML/XML presentation. The Web 2.0 functionalities of online social collaboration and email are achieved using mature java libraries (namely Jmail, Jgraph & Xerces)

The data model anticipates significant increase in data volume and regular updation/maintenance of the above described four data elements. Therefore the XTractor data model is fundamentally represented in generic Extensible Markup Language (XML) format with extensive use of Object relation mapping implementation involved Java Architecture for XML Binding (JAXB) and Castor tool kits. This approach provides an added advantage of easy data exchange and interoperability with third party softwares (eg: cytoscape for network analysis).

To optimize data transmission across the internet for knowledgebase queries that have large results sets, the XML data is streamed in Javascript Object Notation (JSON) format and also offers research teams the option to integrate XTractor datasets into their native analytics workflows.

Further to improve the system’s response time to queries (faster searches), XTractor uses in memory caching provided via the open source Memcache tool kit and real time logging is achieved using Apache Log4J.

Addressing the need for wider distribution on the rich, granular information, XTractor allows users to generate comprehensive PDF reports on biomedical entities. This is achieved using the Apache Formatting object processor (FOR) that is driven by XSL transformations of the XTractor data subsets.

**D. ROI achieved or expected (200 words max.):**

Using the semi-automated approach we have been able to compress the time required for manual curation by more than 50%. Regular manual reading and extraction of the relevant facts from an abstract would involve a minimum of 6-8 mins for an average reader. Since we used text mining engines at the first level followed by enhancements in our work flow we have been able to reduce
the time required to generate the facts on an everyday basis. Also, the employable manpower for carrying out this job is reduced by more than 40% as against carrying this process entirely manually. So, it reduces the cost involved in obtaining high quality manually annotated facts in the shortest possible time.

Accuracy of the information has a high significance when it comes to researchers & we ensure that the content that gets into the knowledgebase is accurate and manually verified following text mining. This results in highly accurate information in comparison to some well known text mining engines. XTractor also scores in terms of reducing the false positive results by more than 12-15%.

E. CONCLUSIONS/implications for the field.

The email alert service and high-end analytics coupled with accuracy of information and the faster turnaround time ensure that the researcher is able to get the most accurate scientific information with in the fastest turnaround time.

The [http://www.xtractor.in](http://www.xtractor.in) is being currently used by more than 2500 researchers from 350 organizations across the globe. Recently we have even been approved of an outlinking from NIH Entrez Gene, so that every relevant Gene record in entrez gene links out to manually curated XTractor reports.

6. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.):

Testimonials:

My supervisor introduced me to XTractor a few months back. Since then XTractor became the No.1 tool in my mind when comes to literature mining. We are working on protein lipid interaction, and we can use XTractor to validate our predictions. I have also introduced XTractor to my colleagues and many PhD students, and XTractor instantly caught their attention. I would give full marks to such great tool for its comprehensive knowledge base, friendly user interface and immense power in literature mining.

Li Bowen, NUS, Singapore

We are very impressed with your product. It would definitely help us with our literature searching needs and save time.

Jay Doughty, PMP, Project Manager, Division of Biomedical Statistics & Informatics (BSI)

Your search engine (XTractor) is superb. I was able to get updates for my query in less than 24 hours. Thanks for your wonderful (free) work.

Dr. G. Jagadeesh- US FDA, Silver Spring, Maryland, USA

XTractor - Discover Newer Scientific Relations Across Abstracts Internet expert, author, keynote speaker and consultant

Marcus P. Zillman, M.S., A.M.H.A.

Case studies: http://www.xtractor.in/case_study.do

Blogs: http://xtractorpremium.wordpress.com/
Bio-IT World 2010 Best Practices Awards

1. Nominating Organization (Fill this out only if you are nominating a group other than your own.)

   A. Nominating Organization
   Organization name: Phase Forward
   Address: 77 Fourth Avenue, Waltham, MA 02451

   B. Nominating Contact Person
   Name: Elaine Maynard
   Title: Director, Global Marketing
   Tel: 781-902-4371
   Email: elaine.maynard@phaseforward.com

2. User Organization (Organization at which the solution was deployed/applied)

   A. User Organization
   Organization name: GlaxoSmithKline K.K.
   Address: GSK Building
           6-15 Sendagaya 4-Chome
           Shibuya-ku, Tokyo 151-8511

   B. User Organization Contact Person
   Name: Ted Chin, PhD, MBA
   Title: Vice President & Technical Advisor, Development & Medical Affairs; GSK Japan
   Tel: 610.787.3085
   Email: Ted.Y.Chin@gsk.com

3. Project
   Project Title: Changing the Game: Post-Marketing Surveillance in Japan
   Team Leader
   Name: Ted Chin, PhD, MBA
   Title: Vice President & Technical Advisor, Development & Medical Affairs; GSK Japan
   Tel: 610.787.3085
   Email: Ted.Y.Chin@gsk.com
   Team members – name(s), title(s) and company (optional):
   Masaru Iwasaki, MD — Managing Director, Development and Medical Affairs, GSK Japan
   Akiyoshi Uchiyama, MD, PhD — Deputy Director, Development and Medical Affairs, GSK Japan
   Yasunori Trashima, MD — Director, Clinical Safety and Pharmacovigilance, GSK Japan
   Shogo Inoshiri — Director, Post-Marketing Surveillance, GSK Japan
Hiroshi Kambayashi — Post-Marketing Surveillance, GSK Japan

4. Category in which entry is being submitted (1 category per entry, highlight your choice)
   - Basic Research & Biological Research: Disease pathway research, applied and basic research
   - Drug Discovery & Development: Compound-focused research, drug safety
   - Clinical Trials & Research: Trial design, eCTD
   - Translational Medicine: Feedback loops, predictive technologies
   - Personalized Medicine: Responders/non-responders, biomarkers
   - IT & Informatics: LIMS, High Performance Computing, storage, data visualization, imaging technologies
   - Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
   - Health-IT: ePrescribing, RHIOs, EMR/PHR
   - Manufacturing & Bioprocessing: Mass production, continuous manufacturing

(Bio-IT World reserves the right to re-categorize submissions based on submission or in the event that a category is refined.)

5. Description of project (4 FIGURES MAXIMUM):

   A. ABSTRACT/SUMMARY of the project and results (150 words max.)

      In Japan, pharmaceutical manufacturers are required to conduct Post-Marketing Surveillance (PMS) on their products and provide the results to the government regulatory agency for review. These studies can reveal unexpected adverse events and drug interactions that did not appear during the drug approval process, which can improve a drug’s overall safety profile.

      PMS studies typically entail a larger, more diverse patient base than those in premarket studies—it’s not unusual for several thousand patients to be included. Most companies today conduct PMS on paper—a time-consuming process, yielding inaccurate and incomplete information. To address this, Phase Forward has collaborated with GlaxoSmithKline Japan to develop an Integrated Post-Marketing Surveillance (iPMS) solution to capture accurate data without the lag associated with paper.

      The iPMS solution also has possible global implications. While PMS is not yet compulsory in the U.S. and Europe, it is frequently mandated.

   B. INTRODUCTION/background/objectives

      In Japan, pharmaceutical manufacturers are required under the Pharmaceutical Affairs Law to conduct Post-Marketing Surveillance (PMS) on every new drug sold in the country. These studies must be launched quickly as they are critical to patient safety and the go-to-market timing of a new drug. Post-marketing studies also can be requested on drugs already on the market.
To adhere to this requirement, most pharmaceutical companies today are conducting Post-Marketing Surveillance on paper. Unfortunately, once live, the use of paper for PMS data collection is typically a time-consuming, arduous process. When the data finally becomes available – usually months or even years after the patient was seen – it may contain inaccurate and incomplete information. This can slow down safety signal detection and make it difficult to draw meaningful, conclusive results. From the sponsor perspective, the data is too late in the product cycle to be fully effective.

To overcome this challenge, some pharmaceutical companies have tried to leverage the sophisticated Electronic Data Capture (EDC) systems they use for clinical research for Post-Marketing Surveillance. Although EDC significantly improves the quality and availability of data, it is designed to handle the rigor and complexities of clinical research workflow generally associated with Phase II and III studies, and unfortunately does not adequately match some of the unique requirements for Post-Marketing Surveillance. Some of these include:

- Rapid deployment in weeks, not months
- A simple, streamlined and intuitive application designed to promote easy adoption by general practitioners and medical representatives
- Automated production of the required regulatory and medical reports via a standards-based clinical data repository
- Real-time patient enrollment visibility to adhere and maintain cohort groups (such as pregnant/non-pregnant, immuno-compromised/healthy, very young/elderly) and adjust recruitment strategies as necessary
- Earlier notification of potential safety events
- More cost-effective compared to existing options

To address the unique needs for Post-Marketing Surveillance, Phase Forward has worked closely with GlaxoSmithKline Japan to develop the Phase Forward Integrated Post-Marketing Surveillance (iPMS) offering. The first iPMS implementation was deployed for a post-marketing study of GSK’s Relenza flu treatment in just five weeks.

C. RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

The iPMS project leverages unique functionality within Phase Forward’s Integrated Clinical Research Suite (specifically the OutcomeLogix and clinical data repository tools), which enabled the company to develop the iPMS solution quickly and cost effectively. Tailored specifically for Post-Marketing Surveillance, the solution can be quickly deployed and enabled GSK to efficiently collect surveillance data via an easy-to-use Web interface and generate the required reports for the Ministry of Health, Labour and Welfare.
The solution uses a standard library of highly intuitive forms that required minimal training of the general practitioners who were entering the data. As soon as information was entered into the system, GSK staff had immediate access to the collected data.

In addition to streamlining data collection, the Phase Forward iPMS solution provides study management and tracking capability as well as a standardized data repository to address the long-term data integration, storage and management requirements for surveillance studies. This repository allows different stakeholders to easily access data and view standard reports on all surveillance data from this central environment, as well as allow sponsors to indefinitely store and report on data in different formats. This functionality allows the repository to automatically generate standard regulatory safety reports and facilitate information sharing.

The speed of data availability and ability to analyze and correlate data can help the team to proactively identify and address any safety risks as they emerge. While this not only improves the safety profile of the drug, it could potentially shorten the duration of the PMS study.

D. **ROI achieved or expected (200 words max.):**

The project was truly a game changer, both in terms of the expected delivery of the safety report for the Ministry and the efficiencies derived from the new process.

**Overall Efficiency** *(See comparison chart in References section – iPMS study vs. paper study)*

1. The first iPMS implementation was deployed in just five weeks, compared to months for paper-based solutions.
2. Enrolled 100% of target enrollment within 10 days after go-live.
3. Within approximately eight weeks, over 80% of the CRFs were collected; 50% of these were locked down; and the interim safety report was delivered to the Ministry.
4. To date, the project cost has been a fraction of that associated with a similar paper-based study.
5. Early study data was provided to the numerous stakeholders within GSK regarding the safety profile of the product rapidly upon deployment.

**Immediate User Adoption**

6. Within six hours of roll out, general practitioners were entering data into the system.
7. Help desk calls were minimal. In the first two months of the study, only four calls were received at the help desk related to application support.
8. Greatly reduced the number of physicians required to conduct such a study.
9. Minimal queries from the system users.

E. **CONCLUSIONS/implications for the field.**
We believe this technology is truly transformational to the existing business processes surrounding the way in which PMS studies are conducted. Using the iPMS solution, organizations will have more immediate data access and greatly reduced operational costs. The speed of data visibility and a repository-driven approach supports comprehensive safety analysis and a more thoroughly understood safety profile.

In addition, patient recruitment is not only quicker, but an early review of data may show that fewer subjects are needed. This real-time visibility into subject enrollment allows organizations to quickly adjust recruitment strategies to adhere to and maintain cohort groups.

The availability of clean data much earlier in the process is seen by many sponsors as an opportunity to help accelerate the acceptance of their products in the market and to help shorten the durations of post-marketing studies while giving patients greater confidence in the safety of the products they are using. As these PMS studies become increasingly required across the globe, the need to more efficiently conduct these massive studies will become more acute.

6. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)

"iPMS is truly a game changer. The integrated solution allows a PMS organization to efficiently design, collect, manage, analyze and report data in near real time, so that general practitioners and medical representatives can get their PMS reports in a timely manner. This allows for more accurate and timely access to safety data, enabling the team to more promptly address any safety problems. Moreover, iPMS accomplishes this feat at a significant cost advantage."

Ted Y Chin, PhD, MBA
Vice President & Technical Advisor, Development & Medical Affairs; GSK Japan

The chart below details the speed of enrollment and CRF collection using the iPMS solution in the Relenza Flu PMS study (Rz-SSMP) versus a similar study (AADUI) that used paper-driven processes.

PMS (Rz-SSMP vs. AADUI)
Comparison of iPMS (RZ-SSMP) and Paper-driven (AADUI) Processes
Bio-IT World 2010 Best Practices Awards

1. **Nominating Organization** (Fill this out only if you are nominating a group other than your own.)

   **A. Nominating Organization**
   Organization name: Recombinant Data Corp.
   Address: 255 Washington Street, Suite 235, Newton, MA 02458

   **B. Nominating Contact Person**
   Name: William Stetson
   Title: Marketing Manager
   Tel: (617) 243-3700 x229
   Email: wstetson@recomdata.com

2. **User Organization** (Organization at which the solution was deployed/applied)

   **A. User Organization**
   Organization name: Centocor R&D, Inc. (a subsidiary of Johnson & Johnson)
   Address: 145 King of Prussia Rd., Radnor, PA 19087

   **B. User Organization Contact Person**
   Name: Dr. Eric Perakslis
   Title: VP, R&D Informatics
   Tel: (610) 240-8476
   Email: eperaksl@cntus.jnj.com

3. **Project**
   Project Title: tranSMART
   Team Leader Name: **Eric D. Perakslis, PhD**
   Title: Vice President, R&D Informatics
   Tel: (610) 240-8476
   Email: eperaksl@cntus.jnj.com
   Team members – name(s), title(s) and company (optional):

4. **Category in which entry is being submitted (1 category per entry, highlight your choice)**
   - [ ] Basic Research & Biological Research: Disease pathway research, applied and basic research
   - [ ] Drug Discovery & Development: Compound-focused research, drug safety
   - [ ] Clinical Trials & Research: Trial design, eCTD
   - [ ] Translational Medicine: Feedback loops, predictive technologies
Personalized Medicine: Responders/non-responders, biomarkers
IT & Informatics: LIMS, High Performance Computing, storage, data visualization, imaging technologies
Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
Health-IT: ePrescribing, RHIOs, EMR/PHR
Manufacturing & Bioprocessing: Mass production, continuous manufacturing

(Bio-IT World reserves the right to re-categorize submissions based on submission or in the event that a category is refined.)

5. Description of project (4 FIGURES MAXIMUM):

A. ABSTRACT/SUMMARY of the project and results (150 words max.)
tranSMART is a research data management project that was implemented at Centocor R&D, Inc., a subsidiary of Johnson & Johnson with assistance from Recombinant Data Corp. The project created a federated connectivity architecture, providing Johnson & Johnson researchers with the ability to query the company’s vast amount of drug research and development data located in disparate source systems using a single tool. The tranSMART architecture minimizes the complexity and resources required to aggregate disparate research data, improving the productivity of the company’s researchers.

B. INTRODUCTION/background/objectives
Modern drug development is expensive, time consuming— and data intensive. To be competitive, pharmaceutical companies need to find new drug targets, develop prospective compounds that act on those targets, and judiciously select compounds to move forward through the drug delivery pipeline. Drug development requires hundreds of experiments, resulting in thousands of pages of documentation and terabytes of data.

Often there are missed opportunities for collaboration that could provide insight into indications, improved effectiveness through pharmacogenomic tests, or identification of factors leading to toxicity. Researchers may be investigating the same target or pathway in different therapeutic areas, or investigating different compounds affecting the same pathway. Scientists doing basic research may want to analyze results from clinical trials to better understand human diseases. Users with different roles and skill sets may want to share data.

Many organizations find it difficult to share data internally. Bringing a new product to market requires focused work of hundreds of people from many different disciplines: chemists, biologists, pharmacologists, physicians, statisticians, and bioinformaticians. Different specialists use different analysis tools, making it difficult for specialists from other areas to find and share data. Information is often trapped in “silos,” broken down by functional area or therapeutic target.

tranSMART was designed to help share drug research and development data across Johnson &
Johnson. It enables scientists, physicians, informatics professionals, and executives to access information from one source, reducing redundant work, facilitating collaboration, and encouraging hypothesis formation.

C. RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

The tranSMART project includes the following tools/functionality:

a. **Search**

   The search engine provides a simple way to query many different types of data, displaying the results in an appropriate format. We designed the search engine to effectively index many different types of data: internal white papers, raw data sets, licensed data sets from third parties, and public databases. A scientist can search by pathway, gene, disease, trial, compound or just a keyword. Results from a query are classified by type (literature references, internal analyses from pre-clinical studies, clinical trial analyses, and curated gene expression analyses, etc.). Matching data results include summary statistics, heatmaps, and links for further analysis by external applications like Ariadne Genomics Pathway Studio®, while text results include text clippings and links to the original documents. Some of the text sources were hand-curated and the resulting biomedical facts are stored in a structured format. The hitlists from these curated text sources can then be exported in semantic triples and visualized in network diagrams.

b. **Dataset Explorer**

   The Dataset Explorer allows users to query clinical trial and experimental medicine data without assistance from a biostatistician or bioinformatics professional. By consolidating many different types of data (patient records, gene expression data, proteomics, metabolomics, and SNPs, etc.) from many different sources (Affymetrix CEL files, SAS files, Excel spreadsheets) into a common system, the solution makes it possible for scientists and clinicians to focus on scientific questions and not data processing. A streamlined interface is provided to generate queries by phenotypes, genotypes or combination thereof. End users can view summary statistics about their queries, analyze gene expression, proteomic, or RBM data through a link to the Broad Institute’s Gene Pattern application, or view SNP data using Linkage Disequilibrium plots.

c. **Data Trust**

   Both the search engine and the Dataset Explorer leverage the same back-end data warehouse: the Recombinant Data Trust. It includes the processes and software for efficiently, securely, and accurately curating and loading clinical records, biomarker data, research results, and other information into a data warehouse, and for making this information accessible to end users.

d. **i2b2**

   Informatics for Integrating Biology and the Bedside (i2b2) is an open source, NIH-funded, scalable informatics framework that bridges clinical research data and the vast data banks arising from basic research in order to better understand the molecular underpinnings of complex diseases.
D. **ROI achieved or expected (200 words max.):**
Prior to tranSMART, the typical delay between a query across disparate source systems and a result was measured in days or weeks. Using tranSMART, that same query will deliver a result in approximately two minutes. This improvement in productivity is expected to foster collaboration and expedite future discoveries.

E. **CONCLUSIONS/implications for the field.**
tranSMART eliminated the challenges around consolidating disparate data repositories for translational research. Its relatively fast deployment (approximately 12 months to a production release from a team of nine engineers), and inclusion of open source components suggests that the technology could be leveraged by other institutions.

6. **REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)**
Bio-IT World 2010 Best Practices Awards

Celebrating Excellence in Innovation

ENTRY FORM
Direct questions about entry criteria/process to:
Allison Proffitt, Managing Editor, 617.233.8280 or aproffitt@healthtech.com

Please email completed entry to:
Allison Proffitt, Managing Editor, aproffitt@healthtech.com
Subject: 2010 Best Practices Entry
Early bird deadline: December 18, 2009; Deadline: January 18, 2010

1. Nominating Organization (Fill this out only if you are nominating a group other than your own.)

   A. Nominating Organization
   Organization name: Sitrof Technologies
   Address: 700 Alexander Park, Princeton, NJ 08540

   B. Nominating Contact Person
   Name: Bryan Reynolds
   Title: CEO
   Tel: +1 (866) 545-8954
   Email: breynolds@sitrof

2. User Organization (Organization at which the solution was deployed/applied)

   A. User Organization
   Organization name: Copernicus Group IRB
   Address: One Triangle Drive, Suite 100, Research Triangle Park, NC 27709

   B. User Organization Contact Person
   Name: Jennifer Sodrel
   Title: Director, Information Management
   Tel: 919-465-4310 Ext. 119
   Email: jsodrel@cgitb.com

3. Project
Project Title: CGIRB’s DocuShare Paperless Environment

Team Leader
Name: Jennifer Sodrel
Title: Director, Information Management
Tel: 919-465-4310 Ext. 119
Email: jsodrel@cgirb.com

4. Category in which entry is being submitted (1 category per entry, highlight your choice)
- Basic Research & Biological Research: Disease pathway research, applied and basic research
- Drug Discovery & Development: Compound-focused research, drug safety
- Clinical Trials & Research: Trial design, eCTD
- Translational Medicine: Feedback loops, predictive technologies
- Personalized Medicine: Responders/non-responders, biomarkers
- IT & Informatics: LIMS, High Performance Computing, storage, data visualization, imaging technologies
- Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
- Health-IT: ePrescribing, RHIOs, EMR/PHR
- Manufacturing & Bioprocessing: Mass production, continuous manufacturing

(Bio-IT World reserves the right to re-categorize submissions based on submission or in the event that a category is refined.)

5. Description of project (4 FIGURES MAXIMUM):

A. ABSTRACT/SUMMARY of the project and results (150 words max.)
Through employee-driven change management and valuable software and process assistance from Sitrof, CGIRB was able to transform itself from a paper-reliant company with millions of pages of legacy documentation into a completely digital paperless organization in less than two years. As a result, the company dramatically reduced expenses and created a truly "green" work environment by nearly eliminating paper, printing, storage, shipping and other hard document handling costs.

Through a three-phase implementation approach, coupled with employee-led review and oversight committees, CGIRB fostered a positive, productive atmosphere for long-term change management. This empowered CGIRB to increase business without increasing staff and improved the company’s ability to retain staff and reduce turnover.

CGIRB gained competitive advantage and improved service to pharmaceutical clients by becoming more efficient in review and approval process—thus enabling clients to accelerate their study start up and increase the speed in getting drugs approved and into the market.

B. INTRODUCTION/background/objectives
**Copernicus Group IRB** (CGIRB) serves an important role in the drug development and clinical trial process. Its primary responsibility is to ensure that the rights and welfare of human research subjects are protected. To assure regulatory compliance, the CGIRB Board reviews research protocols and study-related information, as well as investigator qualifications and resources. CGIRB abides by the federal regulations that pertain to research conduct, exceeding these requirements in some areas.

CGIRB was established in July 1996 as an independent institutional review board (IRB) and is organized and operates in compliance with regulations governing institutional review boards set forth in 21 CFR and ICH guidelines, as well as 45 CFR when applicable. CGIRB is a member of the Consortium of Independent Review Boards, a group of independent IRBs that regularly meet to review issues of importance in the protection of human research subjects. In 2004, CGIRB was awarded full accreditation of the human research protection program by AAHRPP®, The Association for the Accreditation of Human Research Protection Programs, Inc., and was awarded full re-accreditation in 2007. CGIRB is registered with the Office for Human Research Protections as IRB00001313, enabling the review of HHS-supported or -conducted human subjects research under the Federal-Wide Assurance of a submitting body.

**What is an IRB?**
Under FDA regulations, an IRB is an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects. In accordance with FDA regulations, an IRB has the authority to approve, require modifications to secure approval, or disapprove research. This group review serves an important role in the protection of the rights and welfare of human research subjects.

The IRB review assures, both in advance and by periodic review, that appropriate steps are taken to protect the rights and welfare of humans participating as subjects in the research. To accomplish this, IRBs use a group process to review research protocols and related materials (e.g., subject information and consent forms and investigator brochures) to ensure protection of the rights and welfare of human subjects of research.

**Project Background:**
Running the day-to-day business at an IRB is incredibly paper intensive and requires a great deal of internal and external collaboration. But by 2006, 96% of incoming documents to CGIRB were received electronically via digital fax, portals and email—and then printed out because CGIRB did not have a FDA 21 CFR Part 11 certified process for electronic document handling.

In the fall of 2006, CGIRB sought a document management solution to handle incoming electronic documents so they did not have to be printed and circulated in hard copy. It approached Sitrof Technologies, a company specializing in unstructured document management with extensive experience Part 11 compliance. CGIRB’s goal was to install a 21 CRF Part 11 compliant document management system with workflow, version control and electronic signatures to handle incoming documents. The goals were to improve efficiency, reduce the need for paper (cost, storage and environmental reasons) and most of all, improve their competitive advantage.
Initially, CGIRB only sought a way to handle incoming electronic documents, and nothing more. However, early in the investigation process, CGIRB realized it was entirely possible to become a completely paperless company and so established a very aggressive goal of doing so. This paperless evolution would involve a massive legacy document scanning operation, including 1.5 million legacy documents totaling 5 million pages, as well as all internal documents and all inbound paper and electronic documents.

The project would also require extensive system integration with CGIRB’s internal Protocol Tracking Systems and its customer-facing portal, Connexus. It would also necessitate reengineering and converting of the company’s existing paper processes and workflows—which in turn necessitated a critically important cultural/change management solution for CGIRB employees.

C. RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

CGIRB understood that the change management aspect of the project was the single most important factor in long-term success if the company were to become truly paper-free. The company worked with Sitrof to create an efficient and successful implementation process using a three-phase approach. Due to the requirements of 21CFR Part 11, CGIRB must validate all regulated systems. Each of these phases included documentation and validation efforts to ensure CGIRB systems remained in an appropriately validated state. All changes to existing systems were implemented through a formal Change Control Process and all aspects of SDLC were adhered to.

- **Phase One:** Using DocuShare out of the box.

- **Phase Two:** Scanning and uploading 5 million legacy pages to an e-file room for viewing purposes only.

- **Phase Three:** Adding Sitrof’s Compliance Module to DocuShare to automate the workflow and decision process for electronic records while maintaining Part 11 compliance.

**Phase One (mid 2008):**
From the very start, CGIRB improved employee adoption and buy-in immeasurably by creating subcommittees, focus groups, team meetings and best-practices teams from lines of business and IT. In mid-2008 Sitrof installed Xerox DocuShare software for use with CGIRB’s non-regulated documents, such as vacation requests, business development information, training information, PowerPoint presentations, spreadsheets, status reports, RFIs and contracts. Employees quickly became familiar with the DocuShare interface, easing them out of their comfort zone in a non-threatening, time-relaxed manner.

Subcommittees, focus groups, team meetings and best practices teams involved as many staff members as possible. These subgroups examined existing SOPs and processes throughout the transition process, bringing people along gradually. This group buy-in approach proved pivotal to the success of the entire project as people brought solutions rather than problems to the table.
Phase One also allowed CGIRB and Sitrof to fine-tune processes and customize the software before launch, again minimizing risk before full implementation.

Phase Two (mid-late 2008):
To become truly paperless, a major legacy scanning operation was necessary. CGIRB commissioned a local established outsource firm, SCDATA, Inc., to work on-site, scanning, indexing and archiving 5 million pages of legacy documents (more than 200,000 pages scanned per month). Proper check-in/check-out of these live paper documents was essential throughout the conversion process, because at any point in time, CGIRB had to be able to locate any specific document—even if it had already been pulled for scanning.

Initially, the scanned documents were available through a DocuShare e-file room for “read only” purpose. In other words, they functioned just like the paper version. This allowed users to become familiar with the system and working with electronic records. However, because the system was not yet Part 11 compliant for decision-making, the original paper was still routed to board members. Thus by the end of phase two, and before full launch, users already had nearly a year’s worth of hands-on experience with the system.

In addition to making a certified and trustworthy duplicate through scanning, Optical Character Recognition (OCR) was applied to each page, converting it to digital text and allowing full text search of all 5 million pages of documentation. This strategic innovation added minimal incremental cost while significantly increasing the long-term value, usability and ROI. Adding OCR and full text search was very forward thinking on part of the team and resulted in further streamlining of the clinical research process. Each document was also compressed to within 5% of its original size. The OCR and compression process continues today for all newly entered hardcopy and electronic documents.

Phase Three (mid 2009):
Both CGIRB and Sitrof knew from the start that Xerox DocuShare alone would not meet all functional requirements for a new paperless system, including Part 11 compliance. Yet, by design, it wasn’t until a full year later—after thorough testing with non-regulated documents, working through process changes and getting buy-in—that Sitrof installed its Compliance Module to ensure 21 CFR Part 11 compliance. Until that point the mission critical documents were still being handled in paper.

DocuShare, when combined with the Sitrof Compliance Module, is a robust document management collaboration tool supporting CGIRB’s mission critical function. Critical features include electronic signature capabilities and the “Change Status” function, as much of what a review board does can be thought of as “case management,” and each “case” goes through a series of status changes throughout its lifecycle.

Summary: CGIRB’s paperless evolution is green, efficient, cost-effective, compliant and competitive.
With over 96% of documents received electronically at CGIRB, eliminating the need to print documents increased efficiency, reduced cost and maintained a green work environment.

Today 95% of CGIRB is paperless. (The exception being human resources and SOPs which must meet other federal requirements in addition to 21 CFR Part 11, and will be paperless in the near future.) By comparison, in the past, working with digital documents meant printing and copying
packets to be couriered to board members for review. Signed outcome letters would then be copied, shipped back and filed. Today, board members simply apply their e-signature and send them to the pharmaceutical client, or make them available on CGIRB’s electronic portal. This entirely eliminates the cost of paper, copier equipment and courier service.

But beyond these important considerations, electronic packets can now be received instantaneously by board members, and the entire collaboration, signature and document lifecycle is handled electronically. All decisions and collaboration are now based on certified, validated, paperless records. By scanning and performing OCR all the 1.5 million documents in the legacy repository can now be searched in a Google-style manner by the DocuShare software. This enables board members to quickly find the right passage in typically 200-page long protocols and investigative brochures. In addition to the time savings, the risk of missing something is virtually eliminated.

In the past, thousands of approval documents and correspondence were shipped out to sites and clients each day. Now, with updated SOPs and the technology in place to exchange electronic documents, paper is conserved, shipping costs are reduced and approval documents can be accessed instantly—dramatically improving CGIRB’s ability to achieve more efficient and speedy clinical trials to get drugs approved into market faster.

Access to study documents also underwent significant change. Because study documents are dynamic documents that are never fully retired, they can be in use for more than ten years. Converting all paper to digital dramatically expedited the retrieval process of these valuable documents and nearly eliminated hard storage costs. CGIRB no longer risks losing paper documents, and a single document can easily and securely be shared among numerous board members.

Mission critical “case” documents no longer have to be printed to maintain compliance. There is complete buy-in from the knowledge workers (Project Managers, IRB Administrators, QA and IRB Board Members) who today are more effective and efficient than ever. The review process has been streamlined through innovative use and integration of numerous digital technologies including scanning, CGIRB legacy systems, off-the-shelf software, Sitrof Compliance module and custom software.

The resulting across-the-board efficiency has allowed CGIRB to increase business without increasing staff and has also allowed clerical staff (whose primary role was to file paper) with new opportunities to expand their skillset into other areas, increasing the company’s ability to maintain our staff and reduce turnover. In short, CGIRB is now able to do more with less.

The competitive advantage is another significant benefit. Having fully validated and certified paperless review process gives CGIRB a tremendous competitive boost. As one of the top five IRBs in the country, and being over ten years old, a full conversion to paperless was a huge undertaking, and CGIRB’s competitors have not been able to follow suit. New competitors can move in the direction of paperless more easily, as they do not have the wealth of legacy documentation to convert, but these upstarts do not come to market with the level of expertise of the more established CGIRB.
D. **ROI achieved or expected (200 words max.):**

Making the IRB operations paperless has enabled CGIRB to improve efficiencies internally which has resulted in noticeable increase in turnaround time for study and site start up. It has also enabled the review board to perform more thorough and efficient review by enabling them access to all electronic documentation and data for current, pending and past studies.

An ROI of 51% represents the implementation of CGIRB’s DocuShare paperless environment. This does not include the additional Return on Investment for the CGIRB Connexus™ Client Portal to be released in March 2010. This calculation was based on efficiencies in personnel, reduction in resources and expenses and other associated costs. (formulas and individual numbers can be provided upon request)

<table>
<thead>
<tr>
<th>Project Costs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Personnel</td>
<td></td>
</tr>
<tr>
<td>Development</td>
<td></td>
</tr>
<tr>
<td>Temp services</td>
<td></td>
</tr>
<tr>
<td>Information mgmt</td>
<td></td>
</tr>
<tr>
<td>Expenses</td>
<td></td>
</tr>
<tr>
<td>IM consulting</td>
<td></td>
</tr>
</tbody>
</table>

Total Invested   

<table>
<thead>
<tr>
<th>Project Savings (Lifetime estimates)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Office expense savings</td>
<td></td>
</tr>
<tr>
<td>Personnel savings</td>
<td></td>
</tr>
</tbody>
</table>

**ROI**  

51%

E. **CONCLUSIONS/implications for the field.**

- Accelerated the Clinical Research startup and maintenance cycle by improving study and site approval turnaround times, thus increasing the speed in which a drug can make it to market.

- Enabled more efficient and precise Board Reviews and allowed unprecedented access to all legacy and new documentation during reviews via electronic searches and data access.

- Increased IRB visibility and further enhanced the protection of human research subjects by increasing efficiency and bringing more focus to IRB tasks rather than admin functions.
• Increased site monitoring capabilities by allowing remote access to all study and site documents and submissions.

• Integrated with Sponsor and CRO systems in order to make the IRB Submission and Approval process seamless.

• Reduced paper, printing, shipping and document handling costs.

6. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)
Bio-IT World 2010 Best Practices Awards

1. **Nominating Organization** (Fill this out only if you are nominating a group other than your own.)

   **A. Nominating Organization**
   Organization name: Tessella plc
   Address: 26 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, OX14 3YS

   **B. Nominating Contact Person**
   Name: Dr Adrian Roach
   Title: Project Manager
   Tel: 01925 286800
   Email: adrian.roach@tessella.com

2. **User Organization** (Organization at which the solution was deployed/applied)

   **A. User Organization**
   Organization name: AstraZeneca
   Address: Alderley Park, Macclesfield, Cheshire, SK10 4TG

   **B. User Organization Contact Person**
   Name: Dr Richard Pauptit
   Title: Senior Principal Scientist
   Tel: 01625 516135
   Email: richard.pauptit@astrazeneca.com

3. **Project**
   Project Title: CrysIS (CRYStallographic Information System)
   Team Leader
   Name: Richard Pauptit
   Title: Senior Principal Scientist
   Tel: 01625 516135
   Email: richard.pauptit@astrazeneca.com
   Team members – name(s), title(s) and company (optional):
   David Hargreaves
   Ludovic Otterbein
   Derek Ogg
   Judit Debreczeni
   Kenth Hallberg
   Helena Kack
4. Category in which entry is being submitted (1 category per entry, highlight your choice)
   - Basic Research & Biological Research: Disease pathway research, applied and basic research
   - Drug Discovery & Development: Compound-focused research, drug safety
   - Clinical Trials & Research: Trial design, eCTD
   - Translational Medicine: Feedback loops, predictive technologies
   - Personalized Medicine: Responders/non-responders, biomarkers
   - IT & Informatics: LIMS, High Performance Computing, storage, data visualization, imaging technologies
   - Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
   - Health-IT: ePrescribing, RHIOs, EMR/PHR
   - Manufacturing & Bioprocessing: Mass production, continuous manufacturing

(Bio-IT World reserves the right to re-categorize submissions based on submission or in the event that a category is refined.)

5. Description of project (4 FIGURES MAXIMUM):

A. ABSTRACT/SUMMARY of the project and results (150 words max.)

CrysIS was developed as a workflow tool and archiving system for the X-ray crystallography group at AstraZeneca, who determine atomic resolution structural information on how lead-like compounds made by chemists bind to their protein targets, allowing insight to drug improvement. The need for a single, integrated IT system was highlighted by a Lean-Sigma process improvement analysis on the crystallography group. CrysIS is now used globally by chemists at all research sites in AstraZeneca in drug design, it features easy retrieval of protein-ligand structures which are automatically compared to all other structures in the data base, providing the chemist with unprecedented links to similar work in other projects. CrysIS features drop and drag interfaces that mimic laboratory procedures for ease of use. The information in CrysIS is now being transformed into knowledge by introducing an annotation wiki which details the application of the structure to drug design, enhancing the value to chemists enormously.

B. INTRODUCTION/background/objectives

3D structural information on proteins and bound ligands at atomic resolution is an important tool in the design of new drugs, having a large impact on drug discovery projects. Therefore, having an efficient and effective structural chemistry group within a pharmaceutical organization is of
paramount importance. The information is obtained using single crystal X-ray crystallography. Most large pharmaceutical companies have a dedicated structural chemistry department that generates such information, involving production and purification of proteins, crystallizing the proteins in complex with drug leads, and carrying out the full X-ray crystallographic analysis revealing the 3D coordinates of all atoms in the protein-drug lead complex. This allows chemists to visualize how compounds are binding their targets and how binding properties can be improved.

AstraZeneca employed the use of a Lean Sigma exercise to identify areas for process improvement within their structural chemistry group to improve the service offered to their customers: computational and medicinal chemists within drug-hunting project teams.

As one output from this exercise, it was recognized that a new, single, integrated I.T. system was required in order to meet the growing needs of an already large structural chemistry group for the management of their demand, work-flow and archive of structural data.

The group developed the overarching vision into requirements for an I.T. system, which was then designed, developed and deployed across their global organization. The system allows requests for experimental information from the customer, highlights whether any similar work has been previously done, and provides a data base for storing a complexity of experimental information which represent the structural chemist’s work flow. The system is accessed at two levels, depositor and user, where the user as customer feeds in requests and can see and compare results, but is not burdened by the technicalities of the internal workflow.

The CrysIS system allows for integration with several other important data management systems within AstraZeneca, integration with third-party international synchrotrons and is tailored to the work flow needs of protein crystallography.

Previously, an increasing number of ad hoc systems had been used that contained data inconsistent with other systems; a lack of overall visibility resulted in poor planning of work; the nature of systems also meant that data was not accurately captured. An important issue was the fact that crystallization experiments were being repeated using a wide array of initial conditions, as the existing I.T. systems could not highlight earlier determined optimal conditions that would provide a hit.

The objectives of the project were to gather requirements, design, build and deploy a global I.T. system that could be used to manage the workflow of protein structure determination and to allow customers of the group to plan their work better.

The AstraZeneca team worked to define requirements, align working practices and highlight areas where flexibility was required in order to keep disruption to effective, existing local working practices to a minimum.
C. **RESULTS** (highlight major R&D/IT tools deployed; innovative uses of technology).

The CrysIS system is a global system used by ~300 users (chemists) across AstraZeneca major research sites (in USA, UK, Sweden, France and India) and covering all therapeutic areas.

The system allows crystallisers and crystallographers to track experiments from request for a structure and the receipt of protein through to final deposition of a solved structure. Through integration with compound databases, the system allows the user to scan a barcode of a compound plate for example; information on the plate (compounds, volumes, concentrations etc) are then extracted for the user without them having to key in such information.

Through integration with protein synthesis databases, the CrysIS system extracts data regarding the protein that is relevant to the crystalliser. Previously, this information would have been contained in inaccessible paper-based laboratory notebooks; the information would have to be requested from the individual concerned and be emailed to the crystalliser concerned.

The system utilizes modern web-based technologies to allow users to have a drag-and-drop style interface within a web browser that mimics what scientists are doing the laboratory. The AstraZeneca teams were crucial for this aspect of the user interface design to ensure that it was aligned with the work that they perform in the laboratory and that it improved their effectiveness. The system therefore allows users to drag compounds from plates, mix with proteins and subsequently drag to sample holders for X-ray diffraction data collection, signaling the creation of protein-compound crystals.

Once crystals have been defined, CrysIS includes several workflow features to allow users to export data according to various data formats for the purposes of automated x-ray diffraction data collection. CrysIS exports data to ACTOR robot formats, external synchrotron formats such as ISPyB used at the ESRF Synchrotron facility.

Following collection of diffraction data, the data are processed extensively using a large variety of software external to CrysIS. CrysIS is then able to import the resultant data reduction and refinement data from a variety of programs typically used for this purpose: Mosfilm, D**Trek, etc.

Finally, once a protein-ligand structure has been solved, users are able to deposit their structure coordinates into the system, in a format compatible with the world wide Protein Data Bank (PDB), providing a repository for AstraZeneca protein-ligand structures. The replacement of a legacy system for the storage of structures has meant that users have improved quality of data through validation of deposited structures and the ability to trace experiment s from the final structure back through to data collection and even production of the protein. Also, the users have increased functionality in retrieving and comparing (overlaying) structural information.

Throughout this whole process of request management, crystal growth, crystal screening, data collection, refinement, reduction and structure solution, status information on the experiment is
published to the AstraZeneca global assay platform: IBIS. Within such a customer-focused environment, this allows the customers of the structural groups to determine readily the progress of their requests in an environment they already use, to ensure they have up-to-date information and ultimately to gain ready access to the most recent protein-ligand structures.

Within AstraZeneca, the organization has recently upgraded its vast number of desktop PCs to Windows Vista, a major exercise that has had minimal impact on the CrysIS system due to the forward planning of the CrysIS team and infrastructure based teams within AstraZeneca.

Features within CrysIS enable scientists to not just to search for data but for the system to intuitively display other data that the user may be interested in. Search and subsequent display of a particular structure within the system results in the system suggesting other structures where the compound is chemically similar to the one that the user is interested in or other protein-ligand structures that have a similar binding site to the one of interest. This provides the user with unprecedented cross-project learning.

Through the tracking of structures from a request being entered through the experiment to a final structure being deposited, CrysIS allows various metrics or key performance indicators to be generated. This in turns allows for proactive risk management: difficult structures with low expected impact can be avoided for example.

D. ROI achieved or expected (200 words max.):
The CrysIS system has provided many benefits to the structural chemistry group within AstraZeneca and their clients:

- Improved quality of service for customers of the group through improved quality of data and quicker turnaround times for structures, as well as fast and easy access to new and legacy structural information.

- Improved decision making through having data available and in a timely manner. Knowledge of which structures on which to focus effort.

- Process improvement through knowledge of what has been done previously such that structure determinations are not repeated. Ability to see the optimal conditions for an experiment ensure that these conditions can be repeated for similar experiments.

- Benefits from cross project learning may reduce drug development timelines and improve compound quality

- Risk Reduction through analysing which experiments result in few structures or have long turnaround times.
• Providing a strategic I.T. fit through integration with existing systems and hosting on dedicated infrastructure.

• A protein-ligand structure can take several man weeks to determine. For each structure that is not required to be solved (through CrysI$S$ informing the user that such a structure already exists or a similar one exists), this effort can be saved.

E. CONCLUSIONS/implications for the field.

Within the field of structural chemistry a number of options exist for the pharmaceutical company in terms of hardware, software, external facilities and the use of contract research organizations. Through the use of initiatives such as CrysI$S$ at AstraZeneca, it has been possible to integrate and amalgamate these sources of information, facilitating greater efficiency in working practices and improved quality of data. This has all been enabled in an environment where flexibility in working practices has been of paramount importance: different geographical sites that were previously different organizations having created local working practices to meet their specific local needs, flexible use of external organizations to generate structures as needed, and use of internal and external facilities to make optimal use of available infrastructure for each project.

AstraZeneca are now able to focus efforts on two further initiatives: the transformation of information contained within CrysI$S$ into acquired knowledge regarding use of the information in structure based drug design, through the development of an annotation wiki, and the subsequent application of this knowledge into awareness of project developments, teaching and importantly ensuring that AstraZeneca remain innovators in this exciting area of pharmaceutical research.

6. REFERENCES/testimonials/supporting internal documents (If necessary; 5 pages max.)
1. **Nominating Organization** (Fill this out only if you are nominating a group other than your own.)

   **A. Nominating Organization**
   Organization name: Transinsight GmbH
   Address: Tatzberg 47-51
           D-01307 Dresden
           Germany

   **B. Nominating Contact Person**
   Name: Dr. Michael R. Alvers
   Title: CEO
   Tel: +49 351 796 57 80
   Email: malvers@transinsight.com

2. **User Organization** (Organization at which the solution was deployed/applied)

   **A. User Organization**
   Organization name: Unilever
   Address: Unilever Discover
           Colworth Park
           Sharnbrook Bedford MK44 1LQ
B. User Organization Contact Person
Name: Wendy Filsell
Title: responsible for Information Management
Tel: +44 1234 222561
Email: wendy.filsell@unilever.com

3. Project
Project Title: Unilever Knowledge Browser
Team Leader
Name: Wendy Filsell
Title: responsible for Information Management
Tel: +44 1234 222561
Email: wendy.filsell@unilever.com

Unilever - collaboration with Transinsight in the field of semantic searching
Unilever, one of the world’s largest consumer products companies, is working with Transinsight towards a tailored solution from GoPubMed that meets Unilever Research’s diverse scientific data mining requirements. Today’s searching is mostly split between Internet and intranet searches. The integration of both worlds under one semantic umbrella is the goal of a continuing collaboration between Unilever and Transinsight, provider of solutions for knowledge-based semantic search technology. Unilever has chosen Transinsight for corporate semantic searching.

B. User Organization
Organization name: Federal Institute for Risk Assessment (BfR)
Address: Diedersdorfer Weg 1
D - 12277 Berlin
Germany

B. User Organization Contact Person
Name: Dr. Barbara Grune
Title: Head
Tel: +49 30 8412 2271
Email: Barbara.Grune@bfr.bund.de

3. Project
Project Title: Go3R - semantic search to avoid animal experiments
Team Leader
Name: Dr. Barbara Grune
German Federal Institute for Risk Assessment, BASF SE and Transinsight - collaborate in exploring alternatives to animal testing

The German Federal Institute for Risk Assessment (Bundesinstitut für Risikobewertung BfR), BASF SE and Transinsight have agreed to establish close cooperation in the area of semantic analyses to search for alternatives to animal testing. Transinsight will expand the first specialized semantic search engine for this research area called www.Go3R.org to become a global internet platform. In the future, experts from all over the world will be able to use this platform to bring together their knowledge on alternative and supplementary methods to animal testing.

4. Category in which entry is being submitted (1 category per entry, highlight your choice)

- Basic Research & Biological Research: Disease pathway research, applied and basic research
- Drug Discovery & Development: Compound-focused research, drug safety
- Clinical Trials & Research: Trial design, eCTD
- Translational Medicine: Feedback loops, predictive technologies
- Personalized Medicine: Responders/non-responders, biomarkers
- IT & Informatics: LIMS, High Performance Computing, storage, data visualization, imaging technologies
- Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization
- Health-IT: ePrescribing, RHIOs, EMR/PHR
- Manufacturing & Bioprocessing: Mass production, continuous manufacturing

(Bio-IT World reserves the right to re-categorize submissions based on submission or in the event that a category is refined.)

5. Description of project (4 FIGURES MAXIMUM):

GoPubMed - Searching is now sorted!

A. ABSTRACT/SUMMARY of the project and results (800 characters max.)

The biomedical literature grows at a tremendous rate and only the standard search system for biomedical topics PubMed comprises already almost 20 million abstracts. Finding relevant literature is still an important and very difficult problem due to the high complexity of the matter. We introduce
GoPubMed, a search engine which allows users to semantically explore the famous MEDLINE data base by facilitating the Gene Ontology (GO), a hierarchically structured vocabulary for molecular biology and Medical Subject Headings (MeSH), a hierarchically structured vocabulary for medicine. The two semantic networks contain about 250,000 concepts. The whole background knowledge including geo locations, journals and authors holds about 15 million concepts. This background-knowledge can be considered the “brain” of the system.

GoPubMed enables users to search from the get-go in a straightforward WHAT, WHO, WHERE and WHEN fashion but then enables them to drill down to ever increasing levels of specificity. GoPubMed facilities contacts between authors of scientific papers they are reading via social networking.

B. INTRODUCTION/background/objectives

GoPubMed provides semantic navigation through a novel approach where content and meta data are treated seamlessly: the WWWW concept.

The WHAT Category:
It gives an overview over the scientific paper abstracts by semantically categorizing them according to its content by using GO and MeSH and thus allowing users to quickly navigate through the abstracts by category. Second, it automatically shows terms from the semantic background knowledge related to the query, which often do not appear directly in the text. Third, it enables users to verify its classification, because GO and MeSH terms are highlighted in the abstracts so that they can be seen at a glance. Fourth, exploring abstracts with GoPubMed is useful, as it shows definitions of GO and MeSH terms without the need for further look up.

The WHO Category:
This category helps users to find leading scientists and centers in the biomedical area. After performing a search authors of all relevant abstracts are listed under this category. If two articles share the same author, GoPubMed evaluates their similar properties. The system thereby takes into account that the author of each paper often publishes about similar research topics, with the same co-authors and in the same journals. The research topics are thereby connected to the concepts of the semantic network in the background. The more concepts two articles have in common and the shorter the semantic distance in the network is, the more likely it is that the articles were written by the same person. This approach leads to impressive accuracy. If at any point the system is not correct, it can be corrected by the scientific crowd.

The WHERE Category:
Geographic localization of persons, centers, universities, etc, are found under the "Where"-category. Moreover journals relevant for the query are also listed in this category. This allows to easily identify relevant institutions for certain topics in an unseen way.
The WHEN Category:
The date of the publications for your query are shown in the "When"-category. Using the section "Publication date" by clicking on today, last week, month, year, etc., the time window for the search can be changed.

The semantic approach of GoPubMed is fundamentally different to that of a key-word search engine like PubMed.org or google.com. GoPubMed uses background knowledge in the form of semantic networks – also called ontologies – to make search intelligent – brain-like. For example, a search in PubMed for “Heart Diseases” yields today app. 54,200 results. GoPubMed delivers app. 858,000 results. Why does GoPubMed get so many more results? It’s because it includes all 544 ontological sub-concepts from MeSH and their synonyms found under the concept “Heart Diseases.” This is the big difference between semantic search and key-word search: it guarantees completeness of results and allows drilling down to relevant articles in no time. GoPubMed goes beyond traditional approaches by providing a unique navigational instrument. The tree on the left bar, allows users to narrow down huge result-sets with only a few mouse clicks. It also works as a facet filter; it shows users something like a statistical distribution of the results. The Top Categories list shows the most important concepts in the semantic network.

Based on this semantic analysis, GoPubMed also provides very useful statistical information on top authors, journals and cities, and networks of co-authorship. None of these features are available with in other systems. Using GoPubMed makes searching more efficient and saves time, which is important to all scientists. GoPubMed’s semantic search technologies stand for the next generation of search and goes beyond key-words. Again: by using knowledge in the form of semantic networks GoPubMed has two significant benefits: the completeness of results is guaranteed and ultra-fast filtering of relevant results becomes possible.

GoPubMed is a system where users can also search across all information in a multidimensional concept space: actors (WHO, WHAT, WHERE), activities (WHAT), topics (WHAT) and time (WHEN).
Transinsight, the company behind GoPubMed develops novel semantic algorithms, state-of-the-art natural language processing, and ontologies to provide better and faster search capabilities for complex queries. The ultimate goal of Transinsight's line of research and product development is the establishment of a semantic web not only for the life sciences. GoPubMed, is an ontology-based literature search engine, is a first example of a knowledge-based semantic search engine to improve literature searching. It paves the way for a migration from web to semantic web.

The firm works in close collaboration with the Dresden University of Technology. In acknowledgement of the technologies developed by the company, Transinsight has repeatedly been honored with international awards.

GoPubMed is online at www.gopubmed.com free of charge.

C. RESULTS (highlight major R&D/IT tools deployed; innovative uses of technology).

Using GoPubmed’s technology different semantic search engines were developed by Transinsight:

- GoGene is a fantastic new semantic online-tool when it comes to the elucidation of gene-interaction networks.
- Go3R helps scientists find all relevant information when planning to conduct animal trials.
Published Resources for the Life Sciences

- GoWeb: Transinsight creates GoWeb to explore how a general semantic search engine could look like and how typical Google users react to the semantic approach.

**GoGene: gene annotation in the fast lane.**
Plake, C, Royer, L J, Winnenburg, R, Hakenberg, J, Schroeder, M
PMID 19465383

High-throughput screens such as microarrays and RNAi screens produce huge amounts of data. They typically result in hundreds of genes, which are often further explored and clustered via enriched Gene Ontology terms. The strength of such analyses is that they build on high-quality manual annotations provided with the Gene Ontology. However, the weakness is that annotations are restricted to process, function and location and that they do not cover all known genes in model organisms. GoGene addresses this weakness by complementing high-quality manual annotation with high-throughput text mining extracting co-occurrences of genes and ontology terms from literature. GoGene contains over 4,000,000 associations between genes and gene-related terms for 10 model organisms extracted from more than 19,000,000 PubMed entries. It does not cover only process, function and location of genes, but also biomedical categories such as diseases, compounds, techniques and mutations. By bringing it all together, GoGene provides the most recent and most complete facts about genes and can rank them according to novelty and importance. GoGene accepts keywords, gene lists, gene sequences and protein sequences as input and supports search for genes in PubMed, EntrezGene and via BLAST. Since all associations of genes to terms are supported by evidence in the literature, the results are transparent and can be verified by the user. GoGene is available at http://gopubmed.com/gogene.

**Go3R - semantic Internet search engine for alternative methods to animal testing.**
PMID: 19326030

Consideration and incorporation of all available scientific information is an important part of the planning of any scientific project. As regards research with sentient animals, EU Directive 86/609/EEC for the protection of laboratory animals requires scientists to consider whether any planned animal experiment can be substituted by other scientifically satisfactory methods not entailing the use of animals or entailing less animals or less animal suffering, before performing the experiment. Thus, collection of relevant information is indispensable in order to meet this legal obligation. However, no standard procedures or services exist to provide convenient access to the information required to reliably determine whether it is possible to replace, reduce or refine a planned animal experiment in accordance with the 3Rs principle. The search engine Go3R, which is available free of charge under http://Go3R.org, runs up to become such a standard service. Go3R is the world-wide first search engine on alternative methods building on new semantic technologies that use an expert-knowledge based ontology to identify relevant documents. Due to Go3R's concept and design, the search engine can be used without lengthy instructions. It enables all those involved in the planning, authorisation and
performance of animal experiments to determine the availability of non-animal methodologies in a fast, comprehensive and transparent manner. Thereby, Go3R strives to significantly contribute to the avoidance and replacement of animal experiments.

**Collaboration**

BASF, Transinsight and German Federal Institute for Risk Assessment, Berlin are working closely together to use semantic analyses to search for alternatives to animal testing. The group will expand the first specialized semantic search engine Go3R for this research area to become a global platform. The search engine finds related terms and web pages that the consortium hope will help researchers bring together knowledge on alternative and supplementary methods to animal testing.

The origins of Go3R and the meaning of the name and slogan reflect the ideal situation and touches on a more specific case. In regards to research with sentient animals, the EU Directive 86/609/EEC for the protection of laboratory animals requires scientists to consider whether any planned animal experiment can be substituted by other scientifically satisfactory method not entailing the use of animals, or entailing fewer animals or less animal suffering, before performing the experiment. Thus, a collection of relevant information is indispensable in order to meet this legal obligation. The 3R concept developed by Russell and Birtch in 1959 addresses this directive and is a “gold standard” in the community:

- Refinement of scientific techniques
- Reduction in the numbers of animals used
- Replacement of animal procedures with non-animal procedures

Go3R helps scientists find all relevant information when planning to conduct animal trials. That’s why we called it Go – traditionally for GeneOntology – and 3R for the 3R concept.

**GoWeb: a semantic search engine for the life science web.**

Dietze, H, Schroeder, M
PMID: 19796404

Current search engines are keyword-based. Semantic technologies promise a next generation of semantic search engines, which will be able to answer questions. Current approaches either apply natural language processing to unstructured text or they assume the existence of structured statements over which they can reason. With GoWeb we introduce a third approach which combines classical keyword-based Web search with text-mining and ontologies to navigate large results sets and facilitate question answering. We evaluate GoWeb on three benchmarks of questions on genes and functions, on symptoms and diseases, and on proteins and diseases. The first benchmark is based on the BioCreAtivE 1 Task 2 and links 457 gene names with 1352 functions. GoWeb finds 58% of the functional Gene Ontology annotations. The second benchmark is based on 26 case reports and links symptoms with diseases. GoWeb achieves 77% success rate improving an existing approach by nearly 20%. The third
benchmark is based on 28 questions in the TREC genomics challenge and links proteins to diseases. GoWeb achieves a success rate of 79%.

GoWeb's combination of classical Web search with text-mining and ontologies is a first step towards answering questions in the biomedical domain.

GoWeb is online at: http://www.gopubmed.org/goweb.

An example using GoPubMed

A patient in the UK with a broken arm went to the doctor and a blood sample is taken. Later he/she receives a letter stating the alkaline phosphatase levels were elevated. The patient wants to find out whether this has important implications. He/she visits Transinsight’s semantic search engine GoPubMed at www.gopubmed.com.

The patient starts typing "alkali" and the input box offers among others Alkaline Phosphatase.

GoPubMed retrieves over 62,000 relevant articles. It offers to filter the results through the table of contents – the GO and MeSH ontologies - on the left or to get an overview via the statistics page.

The patient selects the statistics page and learns among others that over 2,000 papers were published in the UK and that "bone" is a key journal for research involving alkaline phosphatase.

The patient wants to learn about associated diseases and opens the disease entry in the table of contents on the left. It shows e.g. that there are over 1,900 papers on a disease named Cholestasis. The tool tip shows plain english description of the disease and defines it as impairment of bile flow.

The users filters the results with Cholestatis and finds an abstract stating: "...features of biochemical cholestasis...high serum concentrations of alkaline phosphatase...". So the patient has learned from a peer-reviewed publication that a possible cause for elevated alkaline phosphatase is Cholestasis.
However, in the disease section of the table of contents on the left there are terms such as bone resorption which possibly relate to the patient’s broken arm. Clicking bone resorption the patient retrieves an article which mentions "...alkaline phosphatase (ALP), a bone formation marker"
The patient is relieved since a likely reason for the elevated alkaline phosphatase levels is the healing of the broken arm. Nonetheless, the patients, who lives in the UK, wants to find leading experts on bone fractures in the country. He/she types bone fr... and selects from the menu the term bone fractures and continues typing unite... and selects the term United Kingdom. GoPubMed retrieves over 6.000 papers and the user clicks the on “top author” to learn that J Kanis in Sheffield is the leading UK author on bone fractures.
A click on the statistics tag shows other top authors and places as well as a network of co-authors. Besides Kanis, Giannoudis and Reeve are top UK scientists in the field and the graph shows that Kanis and Reeve have worked together.

D. **ROI achieved or expected** (1000 characters max.):

The central expertise of Transinsight is the ontology based semantic large scale analysis of written human language. Transinsight facilitates the annotation of millions of texts with ontological concepts, enabling users to rapidly browse and filter the quantity of texts by defined semantic background knowledge, thus speeding up the finding of relevant information significantly. The achieved time savings are 90% and more. By reducing the 12.4 hours per week by only 50% for only 5% of all users gives a potential of savings of 120 Million € per year. If 10% of all users only in Germany would lead to about 12 Million per year. Applying these semantic technologies to all industries the amount of savings can be estimated to be in the range of 4-6 billion Euros.

Our solution GoPubMed PRO – the commercial version of GoPubMed - helps customers like Unilever, BASF, AbCam, European Centre for Disease Prevention and Control, MedTrust, RWE.

E. **CONCLUSIONS**/implications for the field.

The name Transinsight stands for gaining insight through transparency. Our philosophy is therefore to allow selected customers as much insight into their data and information as possible. We faithfully follow this philosophy with our products. In the search technology used in GoPubMed, for example, we don’t rank results using an opaque algorithm. Instead, we put users in the driver’s seat by allowing them to rank or filter results according to their needs. The use of a knowledge base makes this possible. Sorting search results into categories is the key to fostering transparency. In accordance with our vision "Towards Finding Answers", transparency plays the central role.

Our users should be able to understand why certain search results are presented and how the software has applied reasoning techniques in order to arrive at the presented conclusions. This is a new paradigm in knowledge management: semantic and meaning based computing.
6. REFERENCES

GoPubMed delivers an information structuring methodology enabling the semi-automatic development of a knowledge model

The described WWW concept allows for elegant semantic drill down of millions of documents and millions of terms. But this is not enough when it comes to changing knowledge! Key is therefore the daily extension of the used background knowledge. Therefore we developed a tool which allows the semi automated update and the generation of large scale ontologies in very short time. The idea behind is simple but powerful: newly arrived documents are screened for characteristic concepts setting this document apart from general text. These documents are screened for previously unknown facts. These facts are than connected to the already existing ontology. The user can – but does not have to – choose from a list of provided concepts. In a second step selected concepts are sent to a variety of selected web sites like i.e. wikipedia, medpedia, for the generation of definitions by means of hypernymic propositions (is-a relations). Again the user may choose by the system proposed definition. In a third step the system fully automatically suggests the most likely positions in the existing ontology. As an example, with this tool a specialised ontology for the area of alternative methods for animal trials was created. The network contains around 20.000 concepts and was created in two month by one domain expert and it is currently used by www.Go3R.org, another specialised search engine for alternatives to animal testing. Using traditional technologies this work would have taken at least two years. As an example the Gene Ontology can be named. Its creation took up to date around 15 million U.S. dollar. GO contains also around 20.000 concepts.

Scale to technology domains containing millions of terms

The whole system is used by one of the largest energy provider in Germany to create a specialised information platform dealing with about 100 million documents and a background knowledge containing at least 1 million concepts. The system already runs and safes the enterprise huge amounts of time and therefore money.

A minimum of human interaction and correction is required

It is rather a philosophical question if the generation of knowledge can at one day done by machines itself. We strongly believe in a semi automated process where the user has the chance to model her or his view onto the world. The above described system for ontology generation follows this principle: put the user in the driver’s seat but free him from all tasks which can be done by a machine. Another aspect is also the often quoted wisdom of the crowd. Our system allows to collaboratively extending the background knowledge. A feature highly appreciated by users and customers.
Knowledge bases are similar to the complexity of classical thesauri knowledge structures (i.e. controlled vocabularies with three different kind of relationships: hierarchical, equivalency, and associative)

Today’s best and fastest reasoners can still not cope with the complexity of today’s state of the art ontologies. Therefore reasoning is still a future topic. We think – and fully reflected this in our technology – that for today’s knowledge systems the above mentioned relation types (hierarchical, equivalency, and associative) are sufficient enough to still gain a rich and unparalled search experience. Also on base of neuro-scientific insights it can be assumed that these types of relation can act as base for reasoning on a higher level of knowledge aggregation. The pure associative but context sensitive structure of concepts can be seen as the back bone for tasks on a higher semantic level.

GoPubMed provides bridges between apparently unrelated domains

The kings discipline! Finding previously unknown bridges between putative unrelated concepts. Our “discovery feature” allows exactly this. The system identifies document sets for two (or more) given concepts containing either A OR B (no! A&B) and identifies those concepts mentioned together with A and B. Basically this is best described with the phrase “finding evidences for unknown paths in the semantic knowledge network”.

System description

Based on GoPubMed architecture, the GoPubMed PRO – the commercial application - consists of three main parts: a component for handling documents from various sources, the searchable index and the web-interface for the user.

The documents for the systems can be originated from different sources. Depending on the requirements the system can handle a wide range of document sources. Text-extraction and meta-data handling algorithms are employed to prepare the documents. The full-text index is created from the prepared documents. The search capabilities include the extracted textual content, available meta-data and text-mined ontology annotations.

The web-interface client in the web-browser for GoPubMed PRO communicates with the web-server. The server coordinates the requests and handles the searches from the user.

In the context of documents there are two main aspects to consider: Where are the documents and what content do these documents contain. GoPubMed PRO implements a crawling system to retrieve documents from various sources. This includes the capability to handle documents from web-servers (internet and intranet), fileserver or a direct access to the local file system. A direct access to a document repository based on Microsoft Office SharePoint is possible, if required.

After the access to the documents a text-extraction step takes place. For text- or XML- based formats like plaintext, HTML, OpenDocument or Office Open XML documents this is straight forward. For binary formats like the older MS-Office formats or PDFs extractors are used.
In particular the many different possibilities for the PDF-Format in its different versions pose a problem for text extraction. In the context of long-term archiving this has lead to a new ISO-Standard PDF/A. Its goal is to provide long-term reproducibility with self-contained documents. This is achieved by restricting the allowed constructs in PDF. A PDF/A file created from a digital text document, the text will be recognized as searchable.

Meta-data like authors, title, sub-title, last-modified, publish-date are an important tool for the filtering of documents. The automatic extraction of meta-data, e.g. authors and title is possible. Unfortunately not all document formats support embedded meta-data. Even if it is possible the information is often not up-to-date or incomplete. The system can include the option to add or edit meta-data by hand. If available, e.g. through the file-system, GoPubMed PRO acquires information about dates (e.g. last-modified) and other meta-data.

It is of great importance to create a system which:

1. Is easy to interact and communicate with and – equally important –
2. Allows modelling all relevant context sensitive semantics.

Based on big customer installations e.g. for Unilever or the German Federal Institute for risk assessment, the most important factor is communication design in all means. The Web 2.0 is new to users and goes beyond “The Google experience”. Users need to be carefully guided towards a new paradigm which also goes beyond the real world where documents exist once (on paper) and can be “stored” only once at one location.\(^1\)

It is fair to mention that GoPubMed just received the prestigious international reddot design award: best of the best for communication design. This shows that a leading edge international jury has selected the communication design of GoPubMed as one of 56 out of 6113 (!) submitted works.

---

Develop knowledge repository

Saving the knowledge in an organization is one aspect. Making it accessible is the more difficult step. The often used way to “force” staff into structure mostly (if not always) fails in even small organisations. Building up a “semantically networked knowledge repository” is the only way to overcome the limitation of e.g. file-system-like structures. The problem is always that a structure created at point A in time is most likely not or only partly valid at point B in time. Another limitation which is more in the brain of the users is the fact that “documents” are practically always associated to ONE category (e.g. a certain folder). But documents contain always many

\(^1\) The possibility of making copies of documents makes things more complex in terms of consistency and are avoided in huge “still paper driven” organisations.
concepts and needs to be semantically associated to many folders, in order to be found under different points of view. GoPubMed PRO provides the semantic sorting by design and is according to our knowledge the only system world wide which fills it.

**Individualized KM support services**

Individualization is seen as an important part of acceptance by users. It is also important to adopt the KM system to different scenarios (in different times). Thereby the easiness of use coupled with multi user capabilities is of great relevance for smooth joint operation. In GoPubMed PRO this is realized through a jointly useable ontology editor and – more important – a semi automated ontology generation platform. These two features allow customers to adapt the system to their specific needs. In conjunction with a rather small feature at first allows users to hide (for the moment) unnecessary ontology branches, the platform enables to tailor “the brain” with few mouse clicks without restructuring the whole ontology.

Hide concepts helps personalize the Ontology to certain needs

The exclusive automated term-generation tool allows: querying GoPubMed, the Web, yours files, etc and extracting and ranking terms, as well as their abbreviations; saving and loading terms from clipboard; filtering terms with regular expressions; finding likely definitions for terms by querying the Web; and adding terms to the ontology.
Discovering knowledge by semantically bridging concepts ...
GoPubMed PRO includes the above described Discovery feature that allows users to find concepts bridging two different unconnected concepts. This unparalleled feature which becomes possible through semantics allows users true knowledge discovery. Using an easy discovery interface users visualize that i.e. tau-protein kinase activity, a molecular function is a potential bridge between ALS (Amyotrophic Lateral Sclerosis) and AD (Alzheimer Disease).

Relevant Literature from Technical University of Dresden and Transinsight until 2008

Patents

Relevant Publications


[8] Unraveling Protein Networks with Power Graph Analysis.
Loic Royer, Matthias Reimann, William Andreopoulos, and Michael Schroeder.
Plos Computational Biology, 4(7):e1000108, 2008

Dimitra Alexopoulou, Thomas Wächter, Laura Pickersgill, Cecilia Eyre, and Michael Schroeder. BMC

Clustering.
Bill Andreopoulos, Dimitra Alexopoulou, and Michael Schroeder. International Journal of Datamining and

Heiko Dietze and Michael Schroeder. In Albert Burger, Adrian Paschke, Paolo Romano, and Andrea Splendiani,
Edinburgh, UK, 2008

Heiko Dietze, Dimitra Alexopoulou, Michael R. Alvers, Liliana Barrio-Alvers, Bill Andreopoulos, Andreas Doms,
Jörg Hakenberg, Jan Mönnich, Conrad Plake, Andreas Reischuck, Loic Royer, Thomas Wächter, Matthias
Zschunke, and Michael Schroeder.

Thomas Wächter, Dimitra Alexopoulou, Heiko Dietze, Jörg Hakenberg, and Michael Schroeder. In Albert Burger,
Richard A Baldock, and Duncan R Davidson, editors, Anatomy Ontologies for Bioinformatics: Principles and