CAMBRIDGE HEALTHTECH MEDIA GROUP
Bio-IT World
Published Resources for the Life Sciences

250 First Avenue, Suite 300, Needham, MA 02494 | phone: 781-972-5400 | fax: 781-972-5425

Bio-IT World 2013 Best Practices Awards

Celebrating Excellence in Innovation

1. Project Title:

Project Title: **Socrates Search**
Team Leader
Name: John Apathy
Title: VP, Data Analytics Strategy

2. Organization:

Organization name: GlaxoSmithKline
Address #1: 5 Moore Dr, Research Triangle Park, NC, USA
Address #2: Gunnels Wood Rd, Stevenage, Hertfordshire, SG1 2NY, UK
Name: Andrew Wooster
Title: Technical Director
Email: andrew.w.wooster@gsk.com

3. Entry Category:

Knowledge Management: Data mining, idea/expertise mining, text mining, collaboration, resource optimization

4. Project Summary/Abstract:

In 2012 GSK rolled out an application that profoundly improved our ability to find archived scientific knowledge.

Socrates Search is a Google-like application that has been enhanced for chemistry, biology and disease search. In addition to standard text indexing, the system uses sophisticated text analytics to identify chemical structure, gene, species and disease entities. This allows users to use a combination of free text keywords and chemical drawing to find relevant content, without worrying about which representation was used in the source document. The system currently indexes >20 terabytes of electronic lab notebooks (eLNBs), Documentum archives, Team Sites, Lotus Notes databases, file shares and databases.

Socrates is built on Autonomy's IDOL search engine and uses ChemAxon's JChem Oracle cartridge for chemistry indexing. The system also uses NextMove's LeadMine software for text entity extraction and their HazELNut package for eLNB crawling.

5. Introduction/Background/Objectives:

In 2011 GSK's R&D leadership sponsored a programme of work to maximize the value of the scientific data that we collect and to enable its reuse even after the data has served its originally intended purpose. They noted that it took great effort to answer the following types of questions:

- Who else has looked at these targets?
- We are about to in-license this compound. Have we ever looked at a similar structure?
- What tox issues should we anticipate for this compound?
- Find me all the PK data for this compound to answer a regulatory inquiry.
- Has this compound been synthesized before? At a CRO?

To better understand this we conducted a series of global voice of the customer workshops to assess how we could make better use of the data that we already collect. The feedback from these workshops was resoundingly clear – the greatest problems were in finding and accessing data. Feedback was often a variant on:

"Why can't we have something like Google."

6. Key Challenges:

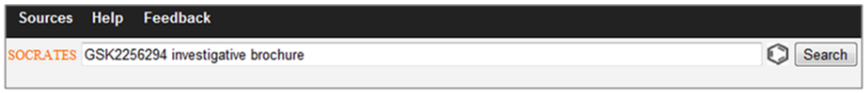
GSK had already made significant investments into its Autonomy enterprise search engine named GskSearch. Autonomy had been configured to search GSK Documentum archives and many file shares. Nevertheless, scientists were frustrated by two problems: 1) its lack of scientific data sources, such as electronic lab notebooks; 2) its lack of knowledge of chemical, biological and disease entities. The scientific community had given up on GskSearch as a source of scientific knowledge.

Upon careful analysis, GSK's IT team realized that the problem was not due to the search engine - it proved to be robust – rather it was that our scientific community had requirements that went beyond standard enterprise search. We set out to create a new web based front-end, named Socrates Search that leveraged the existing GskSearch engine but added the following features targeted at a scientific audience:

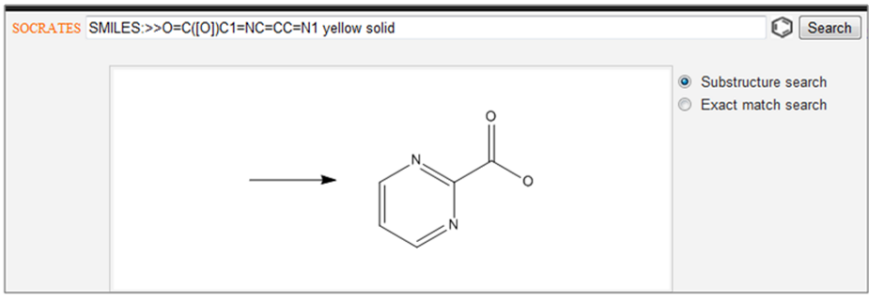
- **New R+D specific data sources**. The largest new source was >1M notebooks from our electronic lab notebook.
- Chemical entity recognition. The system should find chemical entities in a wide variety of formats: SMILES, IUPAC, ChemDraw drawings, Isis drawings, registration ids, trade names, generic names, and common names.
- **Reaction and substructure search** of chemistry in documents. Users should be able to draw a substructure to find documents that contain drawings or text identifiers that represent a matching structure.
- **Chemistry synonymization** based on chemical entity recognition.
    i. Compound aliases: It does not matter how a compound is identified in a document, or how a user specifies search compounds, the system must find matches on the basis of an identifier's chemical structure.
    ii. Parents/Salts: Searching for by a parent compound identifier should find all salt formulations of the compound.
    iii. Combination drugs: Users should find documents that reference combination drugs by searching for any component of the combination.
- **Gene synonymization** using NCBI gene aliases.
- **Disease indication synonymization** using several standard vocabularies: MeSH, ICD-9, ICD-10 and SNOMED.

7. Results:

The features described above were progressively rolled out to R&D during the 2nd half of 2012. By December 2012, Socrates had indexed >2M documents with >70M unique terms. Socrates currently averages ~500 users per month; this number is rapidly growing as we focus the roll-out on specific groups in the company.

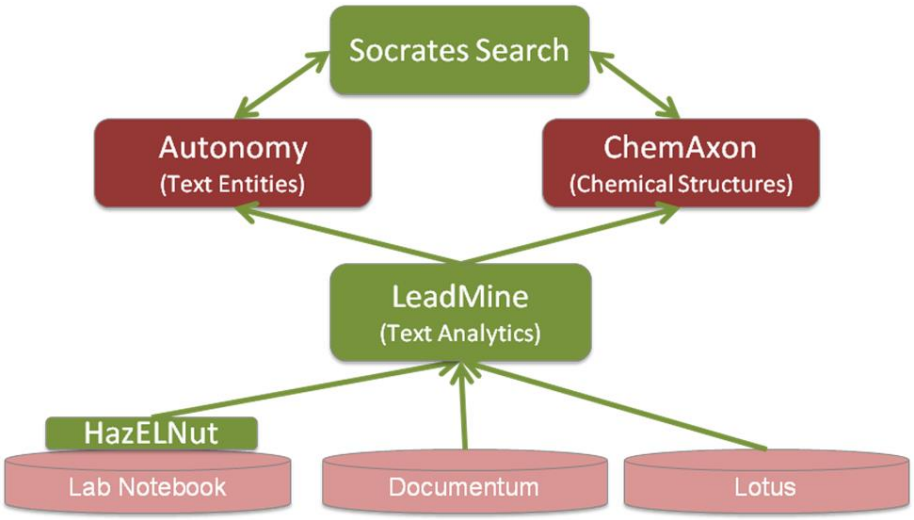


Keyword search with compound synonymization

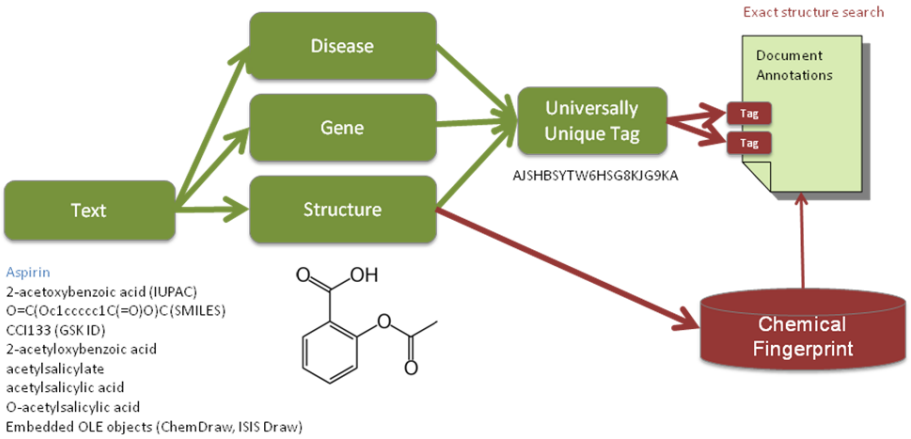


Reaction + keyword search

Socrates integrates a number of commercial technologies. Two of the technologies were already in place at GSK: the Autonomy text search engine and ChemAxon's JChem Oracle cartridge. Additionally, GSK licensed two products from NextMove software: LeadMine for text analytics and HazELNut for eLNB data extraction. Web application and integration components were written using Microsoft's C# ASP.NET libraries.



LeadMine uses a combination of algorithms, dictionaries and regular expressions to identify entities of interest to extract from the text. Socrates then resolves these entities into a canonical form. Chemical identifiers are resolved to canonical SMILES, genes are resolved to a NCBI gene identifier, and diseases are resolved to a MeSH identifier. An encryption algorithm is used to generate a universally unique tag, which is applied to the document for indexing. To enable chemical substructure searches, Chemical entities are also stored in a ChemAxon database with a reference to the source document.



When a user enters search keywords or drawings, the web user interface intercepts the query and runs a similar LeadMine analysis on the inputs. The input criteria are then enhanced to include the universally unique tags as part of the search criteria.

8. ROI (achieved or expected):

Development costs for Socrates were £750K. This included labor, hardware and software licensing. We were able to keep these costs low by reusing the GSK's existing Autonomy Search infrastructure and existing ChemAxon database cartridge licenses.

Efficiency benefits from being able to search electronic lab notebooks were calculated to be £2M per annum. This included the time savings from being able to find successful synthesizes and the time saved in responding to audits. We have not yet tallied up the benefits of being able to search all other archival systems, but we expect to realize several million GBP of benefit from these systems in 2013.

9. Conclusions:

Internet search engines, like Google, are critical to how we all find information on the Internet; it is now impossible to imagine the Internet without them. Enterprise search engines, however, are not held in such high regard. The key reason for this disparity is that Internet community is anchored by content providers with a strong commercial interest to provide the metadata to make their information findable, while the enterprise is made up of people who trying to get their day job done and who are not focused on re-use of their data. In order to overcome this disparity, GSK invested in making its enterprise search smarter so that it could infer the necessary metadata to make content more findable. Socrates Search is now integral to how GSK scientists find and re-use knowledge.

Our primary focus in 2012 was chemistry. However, we were also able to leverage our text mining technology to enhance disease and gene search.

In 2013, GSK will focus on further enabling clinical and biology search. We will invest in integrating a number of late stage sources, such as our clinical trial and adverse event databases. We also expect to add features to support ontology mining and biological sequence indexing.

10. References:

1. "I had to find out the solubility in Fassif for 38 compounds. I had very little progress for almost 3 weeks, until I started to use the Socrates last week….took me about 3 days to find all the information with about 3-4 hours per day. *Investigator*
2. "Socrates Search is an amazing tool and a great advance in our ability to leverage our internal data. Our investigators have been heavily reliant on external and anecdotal data for designing new experiments. This tool allows us to more fully apply our considerable experience, link internal expertise, and design more robust experiments." *Director Animal Research Strategy*
3. "30 seconds to find with Socrates would have taken 5 – 20 minutes without Socrates." *Legal Counsel*
4. "Socrates has just saved us a lot of time today regarding regulatory question around an impurity in Ventolin." *Director Computational Toxicology*
5. "DMPK get requests from Scientists for data, which is in their eLNBs. Since eLNB is indexed in Socrates Search, DMPK can now refer requesters to Socrates Search since eLNB searching is much better than the native eLNB." *Manager Oncology Epigenetics*
6. "I am very impressed with the speed of the searches. …would be totally impractical to do directly with the current eLNB interface." *Chemist, Green Chemistry*
7. "I was able to quickly retrieve program documents for programs that I was working on in the early 1990s!" *Director Chemistry, Protein Dynamics, Oncology*
8. "When I search a registration number for a structure, the chemistry and biology experiments are linked….Socrates gives us a great way to go to the biologists notebooks directly." *Investigator, Metabolic Pathways*