# Chemists are from Mars, Biologists from Venus

*Originally published 7th November 2006*

*Updated 2nd October 2013*

# Chemists are from Mars, Biologists from Venus

Andrew Lemon and Ted Hawkins, The Edge Software Consultancy Ltd

## Abstract

In this article we shall contrast the data management needs of biologists and pharmacologists with that of chemists working in drug discovery. We shall highlight why the structured approach used by BioRails is more suited to biological research and why the document-based design of electronic laboratory notebooks is more compatible with chemistry.

Structured and document-based approaches both provide value but only the structured approach offers the biologists real productivity benefits. We describe how BioRails combines the best of both worlds to support the process of drug discovery.

## Understanding chemists and biologists

Gaining an understanding of the differences between Biologists and Chemists is critical when trying to explain the reasons why their informational needs are different. Of course the term biologist and chemist covers a multitude of sub disciplines that don't necessarily think of themselves in these terms. For example pharmacologists, toxicologists, analytical chemists, medicinal chemists etc. Interestingly there are sub disciplines in chemistry such as formulation chemists that operate more like biologists and biologists from disciplines such as ADME that operate more like chemists but the general observations hold.

### Biologist

A typical biological study is often made up of several experiments and takes from days to months to perform. They are often done in collaboration with other groups to source animals or to assist in collection of histology data and analysis of blood or urine samples. Collaborative working is the norm in Biological research and coordination of all the work is why biological research is often resource and time intensive. The data generated is typically complex and interdependent.

Results will vary depending on a number of factors:

- Conditions such as temperature, time of day etc.
- Cell line or reagent batch
- Compound
- Technology used for the assay
- Animal, species or sex

- Test, or protocol used
- Biologist

The biologist will have regular occasion to analyse the results across one or more of these dimensions to check the quality of results and explain trends, correlations or artefacts that are caused by the environment of the experiment.

Typical searches may include:

- "Show me all results for compounds tested using this protocol under these conditions"
- "Show me the results for all compounds tested using this assay technology"
- “Show me every treatment applied to this animal”

Biological research tends to be organised into larger studies which comprise of lots of experiments. The results are typically written up as a study report which makes conclusions from the study. Results are frequently more about trends and weights of probability than hard and fast scientific determinations.

## Chemist

In contrast, whilst chemists work as part of a team generating compounds for a research program they tend to synthesise compounds in isolation. Similarly their interplay with analytical chemistry tends to be on a service-request basis. Traditionally chemists track all their work in a paper notebook which provides a record of the intellectual property they generate. Each chemist refers to his or her notebook for a record of the synthetic methods they've used and compounds they have produced. Chemists are often the drivers behind introduction of the electronic laboratory notebook (ELN) into research.

This has been for several reasons:

- Protection of intellectual property
- Improvement of productivity
- To provide a searchable record of their synthetic methods

A chemistry ELN facilitates searching by chemical structure across all the notebooks of the organisation allowing chemists to track and review all the synthetic methods used. Chemical indexing allows a link to be made between each product and the synthetic routes that can be used to obtain it. All information is related to either the reaction or the compound structure. This includes records of analytical spectra and purity measurements which are used to ensure the synthesised sample is correctly identified and of the required purity for testing. Such ELN can improve productivity by supporting the chemist to calculate stoichiometric quantities of reagents and to calculate theoretical yields from the experiment. In effect the ELN provides an electronic alternative to their paper notebooks. The ELN frequently provides connectivity to reagent and sample information for experiment design. Structure searching allows them to search across notebooks for the key information that is important to them.

Chemists are typically looking to answer questions such as:

- "What did I do on this date"
- "Show me all synthetic methods that generate this compound structure"

These types of questions will return all the pages of the ELN with relevant information. Chemistry experiments tend to be more deterministic than in biology, and are in general of shorter duration.

## Structured database or indexed documents

For the biologist all the dimensions of the data may be important and will need to be analysed not only within an experiment but across time, sites and potentially assays. Quality control of biological experiments is achieved with reference to historical results for controls and standards. This means that the biologist needs access to not only the data from the current experiment but also historical results from other experiments using the same protocol. This requires a well-designed structured database with ready access to data.

In contrast the chemist needs to search all experiments by time or chemical structure and retrieve the detail that allows them to understand the specifics of matching experiments. As long as a structure search index is provided across all the documents this requirements is straightforward. Their search requirements are relatively simple compared to biologists needs, as the data is relatively straightforward and more precise. Historical data maybe important such as access to the NMR spectra for previous samples of this compound structure, but in general these are easily accessible from individual notebook pages accessed through a structure search.

If one compares the challenges of managing biological and chemical information.

- *Biological information is a complex arrangement of simple data types*
- *Chemical information is a simple arrangement of complex data types*

These differences are at the heart of the alternative solutions used to server each type of user.

## The rise of the electronic laboratory notebook

Currently there is a drive in many pharmaceutical companies to protect their intellectual property electronically by transitioning from paper laboratory notebooks to electronic versions. As we have discussed this drive is frequently initiated by the chemists, as the business case for the productivity benefits of an electronic laboratory notebook for the chemists has long been established. The synthesis of the potential drug molecule is the cornerstone of the intellectual property for most pharmaceutical and biotechnology companies. This movement to ELN technology has been viewed by many discovery IT departments as an opportunity to bring the research of all the scientists into a managed environment. This especially includes those working in the less routine preclinical research disciplines such as *in vivo* and therapeutic research areas. The hope is that by ensuring

that biological information is captured in document form and under change control, it will be made available to the organisation for knowledge sharing and exploitation.

## Does IP protection mean data management?

> *The question is, can an ELN alone provide the necessary data management to support biological research?*

ELN technology essentially manages unstructured data. They support structure within but not across documents. This provides enough flexibility to apply an ELN in a heterogeneous research environment and ensures that documents are managed and brought under change control. An ELN does not however, support mining and exploitation of the data, which is left to knowledge management experts. Although there is some promise being shown by technologies such as the semantic web and text mining tools, they essentially still leave the same problem.

This problem can be likened to a library catalogue search system; one can search by keyword and find all the books which match your query of interest. Perhaps some advanced systems can even identify the pages that contain the data. However, because the data is unstructured, this still leaves the researcher to read and understand all the references before they can answer their specific question. It is impossible to answer structured questions across the entire knowledge-base. This is where ELN fails to address the need.

> *In protecting knowledge, an ELN essentially hides the same information from the biologist.*

There is therefore a clear requirement for a fully structured approach to data management. One which will help to manage biological data across the whole of discovery research and facilitate knowledge exploitation. The task centric approach of BioRails provides productivity benefits to the biologist by helping them to organise and schedule their work in a collaborative environment, whilst the structured approach facilitates knowledge extraction and exploitation later.

## Two paradigms

### Structured data

There are two elements which are essential to the discipline of data management.

1. **Quality**: It must assist the scientist to capture their data in a managed and structured environment, with controlled vocabularies and data capture mechanisms that ensure the data is consistent, well organised and of high quality

2. **Accessibility**: All data management tools must allow the scientists to search and report on their data making use of the same structure to facilitate analysis and utilisation of the data. This means that the results of their experiments and other processes are stored in a searchable and reportable database.

BioRails assists biologists to capture, QC and analyse data, providing support for ensuring that data is of consistent high quality at the point of capture and removing the need to cleanse and harmonise data later. This approach facilitates more re-use of the data after capture by ensuring that it is captured with a strong dimensionality ready for extraction to corporate data warehouses as part of the overall data life cycle.

## Organised documents

In contrast the approach used by electronic laboratory notebooks is to organise data into documents that associate actions of the scientists with their data. The data remains essentially unstructured across these documents. Any consistency across documents is only imposed by use of a standard document structure (or template) and the addition of meta-data tags to the data. This is a consequence of the fact that the primary goal of ELNs is intellectual property protection and not knowledge exploitation.

## Consequences

In other words an ELN document stores data as text. The author can add meta-data to the pages including project name, author, experiment id, and subjects of the experiment but the data is essentially unstructured. Searching can be extended beyond this meta-data using text search engines that provide keyword searching across the documents.

An example query may be

> *'Where has compound XX1 been tested?'*.

The answer would be a list of pages containing the compound identifier XX1. The ELN has no knowledge if the experiment is using compound XX1 as a control, standard or test compound and will indeed include the experiment even if there is a simple comment saying this compound is nothing like compound XX1. The scientist must then navigate these pages, read and understand all the details of each experiment. This leaves a lot of work for the scientist to make sense of the data before they can answer specific questions. This approach does not scale to larger volumes of data and consideration of many studies and compounds.

In contrast BioRails uses a structured approach providing the ability to build and run very precise queries against the data.

An example question may be

> *"Which compounds are selectively active against the CCKB receptor and have been tested with FLIPR technology?"*

This question crosses the boundaries of individual experiments. It may also cross the boundary of different assay definitions and sites on which the experiments were conducted. The result of the query delivers specifically results for compounds that are tested using a FLIPR assay and are active in CCKB and not active in the other tested receptors. The returned results are extracted with their full context; they speak for themselves and do not require the scientist to drill down into the data to discovery his results. This type of query is what makes a research informatics system useful, and differentiates them from, for example, web search engines or document management systems. The data is organised and in context.

## Combining the paradigms

When comparing the two approaches for biological research, the ELN seems, at first sight, inherently more flexible and less intrusive to implement. Other than pre-defined meta-data, the biologist can add any information to their notebook they see fit, annotate it with free text and it will all be stored and, to a limited extent, searchable. It does not constrain the type of data that can be captured or impose restrictions on how the data is represented. The data can be captured and is secured with an effective audit trail that ensures it was generated at the date of saving and has not been tampered with. Whilst this is a very valuable function, the purpose of the experiment is to understand more about the system under investigation and to manage the knowledge, not just to capture data.

> *But what is the direct benefit to the biologists from using an ELN other than document management?*

BioRails combines the two approaches bringing the best of structured data management and unstructured ELN. By linking the structured data to the unstructured information, biologists can explore data and answer specific questions whilst finding the unstructured observations and reports that are linked. This provides the best of both worlds with the freedom to use BioRails like an ELN knowing that the results are firmly linked to the unstructured information for searching later.

ELN have been used in chemistry for over 15 years and are now considered essential software for synthetic chemists in research discovery. This level of acceptance by the chemists is a direct result of the productivity benefits delivered from their use, such as stoichiometry calculators, drawing tools, connectivity to compound databases, inventory and reaction searching. Their use of the data is different from the biologist and the document centric ELN approach works well. This is partly because their key dimension the chemical structure is indexed and searchable.

## So what productivity tools are required by biologists?

As we discussed biological research is much more collaborative combining work from multiple groups and individuals from across the organisation coordinated and culminating in a complex study report which is shared with project teams.

Biologists therefore require support for:

- receiving and making work requests
- planning and scheduling experiments
- tracking an monitoring workloads
- capturing and processing data from instruments
- statistical analysis
- generating study reports

Many of these are traditionally achieved manually by biologists using Excel, Word and shared folders. The reason traditional ELN do not greatly intrude on the working life of a

biologist is because they neatly sidestep data management requirements and provide little in the way of real productivity benefit to the scientists.

In contrast BioRails has been designed to address these issues directly. Not only does it support direct capture of structured data supporting quality control and processing of results but it also offers a collaborative environment for planning and scheduling. By breaking complex workflows down into simple tasks it can address a large range of experimental techniques and facilitate coordination of work across teams and individuals.

## BioRails fits the biologist

BioRails can be used either to conduct the experiment in a structured environment or to bring the results from an unstructured experiment into a structured study environment. It supports the scientists to exploit data through reporting and analysis. BioRails supports the scientist running the studies reducing the time it takes to analyse and make conclusions from data. Decision documents can then be secured and brought under change control within an ELN or document management system. All data captured is under audit control and can be electronically signed proving the data was generated on the date recorded.

## Conclusions

Unlike most ELN systems BioRails provides a direct productivity benefit to the biologists and those analysing the data. It assists the biologist to plan, schedule and execute their work, whilst handling requests from project teams and results delivery to the chemists. The intellectual property generated during biological studies is secured and can be combined with other knowledge in document management systems or other ELN .

Whilst a traditional ELN can provide a tangible benefit to chemists by facilitating structure searching across data, and providing productivity tools for stoichiometric calculations, they offer nothing like the same productivity benefit to biologists. In fact the document based approach used by traditional ELN hinder research and lock away data leaving researchers without the capability to mine and explore data across experiments.

BioRails supports the biologist in organising and tracking preclinical studies, coordinating tasks across groups and individuals to improve productivity and provide more visibility on progress of study requests from project leaders. Its structured approach ensures data is accessible and facilitates knowledge exploitation..

> *...so BioRails is for Biologists and traditional ELN are for Chemists,*
> *...because Chemists are from Mars and Biologists from Venus*

## Contact

The Edge Software Consultancy Ltd,

77 Walnut Tree Close,

Guildford, Surrey,

UK GU1 4UH

http://www.edge-ka.com/

Tel: +44 2380411098 Email: info@edge-ka.com