

The Chemical SmartLab: Intelligent Information Publication for Chemists

H. R. Mills[†], J. G. Frey^{*}, S. J. Coles^{*}, David De Roure[†]

[†]Electronics and Computer Science, University of Southampton, SO17 1BJ, United Kingdom

^{*}School of Chemistry, University of Southampton, SO17 1BJ, United Kingdom

{hugo,J.G.Frey,S.J.Coles}@soton.ac.uk, dder@ecs.soton.ac.uk

INTRODUCTION

The goal of the SmartLab project, a continuation of a part of the UK eScience-funded Combechem project, is to support the research chemist through the lifecycle of an experiment, and to facilitate the publication of the results of the experiment in a transparent, easily-accessible way. The lifecycle of a typical chemistry experiment (and, indeed, of most experimental science, whether chemistry or not), is what we describe as “Plan, Perform, Ponder, Publish”, and the SmartLab extends to every part of this process, from the chemist’s first experiment design, through to publication and dissemination of the experimental data as part of the scholarly publication process. The design and implementation of the SmartLab systems poses a number of questions in the areas of process/service automation, and of semantic description and inference. We first give a brief overview of the existing SmartLab systems, and then discuss some of the issues arising from various aspects of the SmartLab.

THE CHEMICAL SMART LABORATORY

We already have tools to support the chemist in (at least part of) the first three stages of an experiment. These tools, described in full elsewhere[1, 2], are the Planner, the “tablet” tools (Bench station and Weights & Measures), and the Viewer. Each tool communicates with a back-end RDF experiment store using an abstraction library (called libtea), and a SOAP interface which encapsulates the common requests for particular RDF substructures within an experiment. The SOAP interface also supports updating the experiment data store.

PLANNER

The Planner is a simple web-based form, allowing the user of the SmartLab system to list, view and create experiment plans. It replicates the functionality of the standard COSHH health and safety forms which must by law be filled in prior to every experiment. It also extends the requirements of the COSHH form slightly, so that the chemist can provide a step-by-step breakdown of the planned experiment. Additional services are provided by the planner to assist the chemist – a built-in molar quantity calculator, for example, and a ratio lock, so that the required quantities of each reagent can be updated automatically.

BENCH STATION

The Bench Station component of the SmartLab is designed to run on either a tablet PC carried with the experimenter in the laboratory, or on a fixed-location “embedded” screen in the laboratory. It shows the list of steps of the experiment originally entered in the Planner, and allows the experimenter to attach hand-written notes to each step. These notes can then be written up in typed form later (for easier indexing and search) if the experimenter feels that they are sufficiently useful within the context of the experiment record. The emphasis in the in-laboratory tools such as the Bench and the Weights and Measures (see below) is to minimise the user interface complexity, and to pare down the system whilst ensuring that the chemist expends the least possible effort in using the system. There is often little or no time available for the chemist to find or use a mouse or keyboard while performing an experiment,

and such equipment would suffer badly in the hostile environment of a chemistry laboratory. The simple pen-based tablet paradigm, however, is ideal for this situation.

WEIGHTS AND MEASURES

The Weights and Measures component is, as with the Bench station, designed to have a minimal interface usable on a keyboardless tablet PC or embedded display. The user need only select an experiment from the list of currently known and active experiments, then a chemical from the list of planned reagents. They are shown which reagents (and how much of each) they have already measured, and which they have yet to measure out. After measuring or weighing the sample which they will use, simply typing in the quantity used will record the information in the experiment record.

VIEWER

The Viewer application is currently in prototype form as a web-based application, similar in style to the Planner. It shows the list of reagents used in the experiment, with the planned and the actual quantities measured. It then shows, for each step in the experiment, the recorded results for that step and for the materials (chemicals, chemical mixtures, or data) in the step.

RDF EXPERIMENT STORAGE

We have developed a simple but powerful ontology schema in RDFS to store both the plan and the record of an experiment. It is necessary to be able to handle both since, in the laboratory context, the plan exists as a framework upon which to hang the notes and observations of the experimenter in the Bench Station.

The primary entities used to store an experiment in the SmartLab RDF store are the Process and the Material. A Process, as might be guessed, represents a process in the experiment. This might be anything from mixing two chemicals together, to observing a reaction, to performing some complex service-based computation on a data set from a previous analysis step. The Material concept covers both chemicals and data sets (and is broken down into subclasses as such), and describes the entities on which Processes operate and which Processes generate.

The core part of an experiment record in the SmartLab, therefore, consists of two chains of alternating Process/Material nodes. One chain represents the plan of the experiment; the other represents the record. In an ideal world, these chains will contain directly corresponding nodes in a 1:1 relationship. The realities of experimental science, however, mean that sometimes an additional step must be made (a solution needs additional work to crystallise it, say), or a step can be omitted. Thus, the plan and the record do not necessarily always match up. The purpose of the SmartLab experiment description schema is not as a process-control and automation language; neither is it designed to fulfil the tasks of a secure provenance-recording system[3]. Rather, it is designed to replicate the function of the scientist's laboratory notebook, recording information and information structures at a human scale, and allowing automated indexing, search and retrieval of that information in a way which has not previously been possible.

It is worth noting that, although the current user interfaces to the SmartLab system use a strictly linear flow of processes (since this is sufficient for the vast majority of the test cases which we have observed in real life), this is not a restriction imposed by the RDF information store. Our RDF storage schema can easily represent multiply branching experiment process flows, or even generalised process networks.

INFORMATION CAPTURE AND CONTEXTUALISATION

The current SmartLab system effectively replaces most of the in-laboratory functions of the researcher's lab book, allowing the capture of hand-written notes and measurements. This is a manual process, capturing the experimenter's experiences and thoughts.

Towards the end of the laboratory-based part of an experiment, however, the chemist will usually perform a set of analysis steps. These steps are usually performed with the aid of computer-driven equipment, and generate a file or set of files containing the results of the analysis. A typical example would be the technique of spectroscopy, which analyses a small sample of the material under investigation, and produces a graph showing the material's response to different frequencies of light or magnetic field. The data forming the graph is typically saved to a file on the spectroscopic instrument's local hard disk.

One part of the current work on the SmartLab is investigating how best to handle the diverse types of automated and semi-automated analysis tools available to the chemist. One significant development problem is that almost all of these systems run some form of closed proprietary software, and most produce data in difficult-to-read system-specific data formats. Extracting information from these formats for use elsewhere can be awkward. Similarly, ensuring that the results from the equipment are captured and stored reliably (even if in a proprietary data format) when the machine is used can be difficult if the machine does not allow software hooks to be placed in its data-generation process. The derivation and/or adoption of standards in data formats is being undertaken so that proprietary formats can be converted, at source, into a community accepted common format for dissemination and reuse.

The primary problem, however, with the integration of these automated and semi-automated tools into the SmartLab data environment is the problem of context.

Each process, substance, and item of data recorded has a number of different pieces of information which together allow someone reading the experiment record to consider each part of the record of an experiment in its proper context. These contextual indicators include:

- experimenter,
- experiment,
- step within an experiment,
- subject of the datum,
- relationship of the subject and the datum,
- date and time of the record.

The above indicators may be viewed as the metadata for each individual part of the experiment. They can also be viewed as part of the data comprising the experiment record. If not provided explicitly by the user of the SmartLab at the time of the creation of the data, some of the above indicators can be inferred from others. Some others can be inferred from other ancillary information, or, with some level of probability, from pattern analysis. For example, several different tests are performed by the same person in a single hour – these are all likely to be part of the same experiment. As another example, a test on a material may be inferred to belong to one particular experimenter because they are the only person in the laboratory working on materials of that type or structure.

It is potentially onerous for the users of the SmartLab to have to enter all of the above metadata for each measurement or process that they make. One goal of the SmartLab is to ensure that the impact on the chemist of the necessary "bookkeeping" is minimised. Developing a distributed and environment-aware set of services both computational and physical which are also capable of recording and inferring sufficient information to place the record of every service invocation in its proper context with minimal human prompting is a (one might say the) major goal for the SmartLab.

PUBLICATION AT SOURCE

The accepted role of scientific and scholarly publication is to record research activity in a timely fashion and disseminate these results to the research community. A recent government select committee report published recommendations that the underlying data generated during

the course of a research experiment or exercise should be made publically available, and the Research Councils of the United Kingdom have now adopted this policy (<http://www.rcuk.ac.uk/access/index.asp>). Until recently, it has been the case that printed journals and conference proceedings were the most efficient method for the dissemination and archival of research results. However, the vast majority of articles published in the chemistry domain only contain a subset of data or the final result as opposed to all the data generated during an experiment.

The chemical SmartLab approach allows efficient capture, archival and retrieval of all the data and information produced by a chemical experiment, from a description of the processes undertaken in the synthesis lab to the raw and results data arising from analytical and characterisation instruments.

The SmartLab allows for two different mechanisms to publish data at source:

- 1) The publish / subscribe model [4], where an experimental instrument publishes data and operational information live and in real time as it is generated. This model is built around a central broker and a number of clients which connect to this broker. The broker is an agent that matches subscribers to information with publishers of information that is relevant to them. Clients can be publishers of, and/or subscribers to, data and can range from big enterprise-based servers to hand-held pervasive computing devices or unattended remote telemetry devices.
- 2) The open access archive model [5], where software tools are provided so that the chemist can recall all the data concerned with a compound or technique, transform them into a standard format, validate the data and then deposit in a publicly accessible archive. The archive describes data derived from specific techniques in domain-specific schemas, and publishes metadata about the dataset accordingly. The metadata is then published according to digital library (Open Archives Initiative, OAI) standards, which enables third parties to harvest and aggregate the data and build services based on a federation of these archives.

REFERENCES

1. Hughes, G., Mills, H., de Roure, D., Frey, J., Moreau, L., schraefel, m. c., Smith, G. and Zaluska, E. *The semantic smart laboratory: a system for supporting the chemical eScientist*. Organic and Biomolecular Chemistry (2004) 2:pp.1-10.
2. schraefel, m. c., Hughes, G., Mills, H., Smith, G., Payne, T. and Frey, J. (2004) *Breaking the Book: Translating the Chemistry Lab Book into a Pervasive Computing Lab Environment*. In Proceedings of CHI 2004, Vienna, Austria.
3. Groth, P., Miles, S., Tan, V., and Moreau, L. *Architecture for Provenance Systems*. Technical report, University of Southampton, October 2005.
4. <http://www.gridbus.org/escience/escience2005/workshop3.html#3.2>
5. Heery, R., Duke, M., Day, M., Lyon, L., Coles, S., Frey, J., Hursthouse, M., Carr, L., and Gutteridge, C., *Integrating research data into the publication workflow: eBank experience*. PV-2004, 5-7 October 2004 – Frascati, Italy.

Formatted: Bullets and Numbering