BRIL – Capturing Experiments in the Wild

Overview

This presentation describes a project to embed a repository system (based on Fedora) within the complex, experimental processes of a number of researchers in biophysics and structural biology. The project is capturing not just individual datasets but entire experimental workflows as complex objects, incorporating provenance information based on the Open Provenance Model, to support reproduction and validation of published results. The repository is integrated within these experimental processes, so that data capture is as far as possible automatic and invisible to the researcher. A particular challenge is that the researchers' work takes place in local environments within the department, entirely decoupled from the repository. In meeting this challenge, the project is bridging the gap between the "wild", ad hoc and independent environment of the researchers desktop, and the curated, sustainable, institutional environment of the repository, and in the process project crosses the boundary between several of the pairs of polar opposites identified in the call.

Background

The advantages that digital repositories can bring for researchers are widely recognised. On the one hand, they ensure that researchers' work (including both data and publications) can be found and made available, and is correctly attributed to its creators; availability of research data means that published results can be reproduced, and that the raw data can be re-used in new research. On the other hand, they provide effective mechanisms for archiving and digital preservation. These considerations apply to all digital research outputs: research data as well as the final publications. Despite these manifest advantages, however, it has proved to be difficult getting researchers to deposit. Various reasons have been adduced for this, such as difficulty of use, but a major factor is that the process of deposit is just one more demand on a researcher's limited time.

During and after an experiment, researchers deal with a variety of data, both raw and processed, together with the outputs of analysis, which form the basis of their scientific conclusions and published outputs. If these datasets are "managed", as they frequently are, on a combination of desktop machines, local servers and offline media, there is a significant risk that information will be lost, or become incomprehensible – information that is of great importance both as validation of the published results of the researchers that created it, and as the basis of further work by other researchers. The transitory nature of many academic jobs exacerbates this problem significantly.

An approach to this is to embed digital repositories in the researcher's everyday workflows, so that "deposit" becomes automatic rather than an explicit step that the researcher must take, and that metadata is as far as possible captured at the point of creation, rather than being demanded as an afterthought. This applies not just to publications but to the entire process of research, from carrying out experiments and capturing raw data, various stages of data processing, analysis and hypothesis, through to the final products of the researcher's work.

Use cases

Research practices vary immensely; it would not be a productive strategy to take a top-down approach that attempted to define and integrate such processes at an abstract level. Of course, some activities, such as writing journal papers, are relatively well defined; some researchers follow

fairly well demarcated workflows, performing experiments of a particular type, capturing data of a particular format, and applying particular processing and analysis. This is not in general the case, however.

The BRIL project is collaborating with a number of researchers in Biophysics and Structural Biology. This is a multidisciplinary area that interacts and collaborates with several research groups, both within the institution (e.g. Asthma, Cardiovascular, Cancer) as well as with industrial partners such as pharmaceutical companies. So that our efforts were not too diffuse, we addressed in the first instance four research groups, and are focussing for our implementation on two of these – macromolecular crystallography and biological nanoimaging.

Macromolecular crystallography addresses the determination of the structure of large molecules (such as proteins) using x-ray diffraction. In high-level terms, an X-ray beam is directed at a crystal of the substance under investigation from many angles, resulting in a set (typically 360, although possibly up to 720) of diffraction images. Each image contains several hundred spots, whose location and intensity are determined (using specialised software) and then combined to produce a model of the atomic co-ordinates of the protein. This process has multiple steps, dead ends and repetitions, which generate many interim files. While some files are deposited in a public database (the Protein Data Bank), the majority of them are not kept.

Biological nanoimaging involves the use of microscopes to capture high resolution images of biological samples, on the one hand to carry out research into cell and tissue structures, but on the other to developing new methods of digital imaging and processing. The images may be 3D representations, constructed from a large number of sections, as well as in vivo imaging where the sample is imaged at multiple time points. Datasets are processed several times using different image analysis techniques, and many raw images are processed when developing new analysis tools. Again, much of the information generated in this process is not retained.

The practices vary in detail, but from our use case analyses common patterns can be seen: capture of raw data from experimental equipment in a laboratory, various stages of processing and analysis (including many dead ends and loops/repetitions), and publication of outputs, which can be datasets (e.g. protein structure) as well as journal articles. Moreover, once the raw data has been captured, all subsequent processing takes place on the researcher's desktop PC (or Mac).

Objectives

We have two broad objectives:

- To integrate the repository with these experimental/research processes (including laboratory equipment), so that, as far as possible, capture of data and metadata occurs automatically, invisibly to the researcher, and with no (or very little) change to the researchers' practice.
- To capture not just individual datasets but entire experimental workflows, modelled as compound objects including datasets, metadata and provenance information, so that it is possible to trace back from published results and conclusions to the processing and data on which they are based and which justify them. It would also facilitate re-use of data, and allow parts of the workflow to be repeated, for example to verify the earlier results, or to apply a new analysis techniques or software.

Issues and Constraints

The environment in which this is taking place is not a tidy one. Researchers are not using tools that are provided within the repository environment, and thus to some extent controlled by the repository managers. In fact there are two quite distinct and independent environments, which are by no means tightly integrated. On the one hand, there is the researcher's desktop environment – which is typically located in the department and is under the control of the researcher (subject to whatever requirements the department places on that environment). On the other hand, the repository environment is managed centrally (although it could in principle be managed by a school or department – similar issues will apply though).

Moreover, the tools are typically developed by people working in the discipline (either by the scientific communities themselves, or by e.g. suppliers of lab equipment). They are designed to run on a local machine on data accessible via the file system – they don't know anything about web protocols (although there are some web-based services). Consequently, the repository staff has no control (or influence) over them – it is necessary to take them as they come.

Researchers' workflows are complex, but are also quite unpredictable, and they are taking place outside the environment that we control. These processes aren't automated workflows (as implemented in various workflow engines), but are highly interactive processes that pass through several stages, and the researcher uses various tools – some interactive, some not – in the course of this workflow, and there can be a lot of dead ends and looping back when something is tried and doesn't work.

Implementation

A given researcher typically works through an experiment at the same desktop machine, which has simplified our initial prototyping by allowing us to focus on capturing the researcher's process (i.e. by looking at information flow in one direction only). The approach we are taking is to use a local, lightweight client to "scavenge" data (any any information about the data, such as pathname, timestamps) from the researcher's work area on their desktop, and transfer it to the repository environment, where the information is interpreted and use as the basis for creating digital objects, and relationships between objects, which are then ingested into the repository. Much of the work here is concerned with analysing the information that is available and exploiting it to capture the workflow. Although the researcher's workflow is outside our control, it generates as a by-product a lot of information that can be exploited, for example in file headers and log files.

The mechanism for modelling the experimental processes is based on the Open Provenance Model (OPM), although in a slightly simplified form. We define domain-specific predicates for defining provenance relationships between specific objects in the repository. For example, in the case of crystallography, we represent casual dependencies between objects by binary predicates such as *wasDerivedFromReflectionFile* and *wasDerivedFromDiffractionImageSet*. Mechanisms will be provided to use these to capture complete provenance graphs of the experimental processes, based on OPM, which will enable the researchers to review and manage their experimental processes, and in some cases to recreate them as well.

The environment in which the project is operating may be viewed as bit messy, but this makes it an interesting and challenging environment in which to operate. Much work on what might be called "virtual research environments" looks at integrated environments where things are under the

control of the service providers. The sort of model encountered here – where much of the activity is the control of and only loosely coupled to the repository environment, is something that can be encountered in many research situations, and is on with which repository developers will have to engage. In attempting to do this, the project crosses (and bridges) the boundary between several of the pairs of polar opposites identified in the call:

- wild and curated content: the wild content of the researcher's desktop; the curated content of the repository.
- linked and isolated data: isolated data in the working environment of the researcher; joined-up data in the shareable environment of the repository.
- disciplinary and institutional systems: the discipline-focused working environment of the researcher; the institutional systems providing data management and preservation services.
- scholars and service providers: the researchers managing their own material locally; the repository and preservation services provided by the institution.
- ad-hoc and long-term access: limited (in time and space) access to data on the researcher's desktop (during an experiment) or archived to DVD or some other perishable medium (after the experiment is complete); long-term access provided by integration with the institutional preservation.
- the cloud and the desktop: the researcher's desktop; the institutional "cloud" (although not a cloud in the technical sense, it displays cloud-like properties to the researcher).