myTea: Connecting the Web to Digital Science on the Desktop

Andrew Gibson, Robert Stevens
BioHealth Informatics Group
School of Computer Science
University of Manchester, Oxford Road, Manchester
+44 161 275 6239
andrew.p.gibson@manchester.ac.uk

mc schraefel, Ray Cooke, Sacha Brostoff
IAM Group
Electronics and Computer Science
University of Southampton, Southampton, UK
http://mytea.org.uk
mc+www at ecs.soton.ac.uk

ABSTRACT

Bioinformaticians regularly access the hundreds of databases and tools that are available to them on the Web. None of these tools communicate with each other, causing the scientist to copy results manually from a Web site into a spreadsheet or word processor. myGrids' Taverna has made it possible to create templates (workflows) that automatically run searches using these databases and tools, cutting down what previously took days of work into hours, and enabling the automated capture of experimental details. What is still missing in the capture process, however, is the details of work done on that material once it moves from the Web to the desktop: if a scientist runs a process on some data, there is nothing to record why that action was taken; it is likewise not easy to publish a record of this process back to the community on the Web. In this paper, we present a novel interaction framework, built on Semantic Web technologies, and grounded in usability design practice, in particular the Making Tea method. Through this work, we introduce a new model of practice designed specifically to (1) support the scientists' interactions with data from the Web to the desktop, (2) provide automatic annotation of process to capture what has previously been lost and (3) associate provenance services automatically with that data in order to enable meaningful interrogation of the process and controlled sharing of the results.

Categories and Subject Descriptors

E.5 [FILES]: *Sorting/searching*

H.4.1 [INFORMATION SYSTEMS APPLICATIONS]: Office Automation - *Workflow management*

H5.2 [INFORMATION INTERFACES AND PRESENTATION]: User Interfaces - *Graphical user interfaces (GUI), Interaction styles, Theory and methods, User-centered design*

J.3 [LIFE AND MEDICAL SCIENCES]: *Biology and genetics*

General Terms

Management, Design, Experimentation, Human Factors, Theory.

Keywords

Bioinformatics, myGrid, jigsaw analogy, workbench, user interface, Web services

Copyright is held by IW3C2.
WWW 2006, May 22–26, 2006, Edinburgh, UK.

1. INTRODUCTION

Good scientific practice requires that a record of what has been done is preserved such that either experimental procedures can be cross examined by others and repeated if necessary, or aid the researcher who wishes to revisit something done some time ago. The traditional steps of an experiment are: Hypothesis, Methods, Results and Conclusions. These are deeply embedded in the scientific method, and in the hands-on laboratory environment, the lab-book is fundamental for recording details of each stage. Figure 1 highlights the role of the lab-book, and most importantly, the types of detail recorded which enable the researcher to look back, even after a considerable amount of time, and review what was done and what was learned.

Bioinformatics is the discipline in which computational and mathematical techniques are used to store, manage and analyze biological data in order to answer biological questions. Research in the field of bioinformatics typically puts the scientist in a position where experimentation is done exclusively on a computer (*in silico*), although a combination of laboratory and *in silico* work is not uncommon [9]. The prerequisites for bioinformatic analysis are some data with which to work and some tools with which to analyze the data. In many cases, bioinformatics analyses are made by passing data from resource to resource, filtering and transforming those data as the analysis progresses [17]. Workflows are seen as a mechanism for presenting a high-level view of the analysis procedure well suited for delivering bioinformatics analysis [16]. Access to such workflows is facilitated by the availability of Web Services for many bioinformatics resources. As well as this basic access to bioinformatics analyses, middleware, such as myGrid, also offer provenance services [23]. These record the origin and history of the runs of the workflows, the processes used, the derivation path of the data, the organization of who ran and created the workflow, and knowledge annotations for the analysis [22]. This provenance capture (what was done to what by what, whom and when) forms some of the basis of the functionality of a lab book, but only for the actual experiment itself. This kind of provenance does not capture either the planning or the more exploratory phases running up to the creating and running of the polished experiment itself. In other words, the process does not capture all of what traditional wet lab science practice captures in a lab book.

Lab book record keeping of an experimental process is a well understood model in the wet lab sciences in particular. When attempting to apply this model of research to bioinformatics,

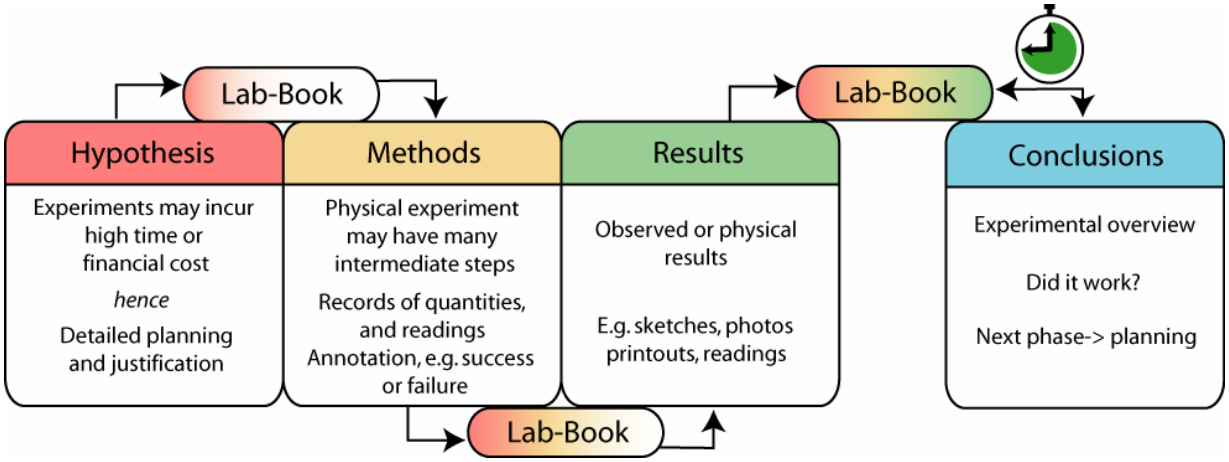

Figure 1. The crucial role of the lab-book in an idealized laboratory experiment. The experiment is captured in enough detail that conclusions can still be drawn after a considerable amount of time has passed.

however, some large differences become apparent. First, most bioinformatic experimentation is a kind of light weight, rapid analysis using many resources. It is a kind of fast exploration where the bioinformatician is deciding yes this sequence of data may be of interest, this other may not. As we describe in greater detail below, this kind of bioinformatics practice can be likened to putting together a jigsaw puzzle where one has a vast number of pieces: one rapidly assesses first whether each piece is a candidate part of the puzzle of interest, and then assesses where that piece may fit: is it a corner, and edge; in the top right or lower left, and so on.

As a consequence, this kind of analysis is relatively low cost/low effort in the experimental planning stage. Many hypotheses likewise can be generated and tested rapidly. In this kind of analysis, the types of information produced *in silico* are not compatible with a lab-book style approach to experimental recording. Figure 2 illustrates how the distinction between the hypothesis, methods and results becomes blurred in bioinformatics, as one often feeds into another, and the movement between them is fast. As a consequence, when a significant result is obtained, a process of *post hoc* rationalization is required to tease apart which combination of data and tools produced that result. This is where the lack of annotation can and does have serious consequences. Because the cost of annotation is so high it is as if pieces of the puzzle are set aside or discarded and then are effectively lost in the mass of other data held on the filestore. Thus, the cost is simply lower to download the files again, to rerun an analysis, than to annotate work that has been done. That does not mean that there is no cost involved and that this is simple; it is only that going through re-finding and rerunning data is actually more efficient than the alternative. This is not a process the scientist prefers; it's simply better than the current alternative

In this paper we present an interaction framework and architecture, based on Semantic Web technologies, which provides mechanisms to support an automatic capture and annotation of components in a bioinformaticians exploratory analysis for easy retrieval and post hoc analysis. In the following sections, we describe in more detail the work practice of the bioinformatician in the context of existing tools and practices. We describe the design methodology informing our work. We then present the open Semantic Web-based architecture and tools we

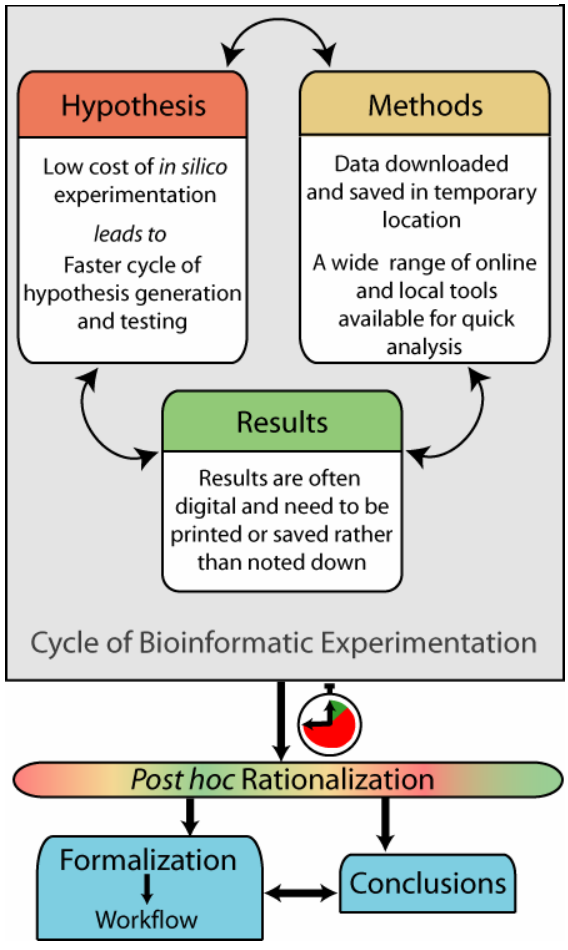

Figure 2. The low cost of bioinformatic experimentation can lead to a fast cycle of experimentation, and little of this is captured. This makes it difficult to look back, and requires an extra step of *post hoc* rationalization, which may include repeats of downloads and key analyses.

have developed to support this practice, and our plans for development and evaluation of the framework.

2. DESIGN METHOD

In order to understand how best to support the bioinformatics process we needed to understand the bioinformaticians' practice. To this end, we used the design methodology Making Tea which had been developed in a previous science-focused project where experts from one domain needed to understand and communicate with experts from another (very different) domain [11].

The Making Tea approach is designed specifically to model highly expert, loosely structured, potentially highly longitudinal tasks. It does this by leveraging the construction of an analogy which can be used to describe the process of the activity under consideration. The analogy is developed by a domain expert, and is then validated with other domain experts to ensure that the analogy will be useful and usable within the community. This latter point regarding acceptance of the analogy is important because part of the design process is to carry out iterative design reviews with domain experts beyond the confines of the project team. The use of the analogy itself to lead design reviews is an effective way to maintain consistent communication among groups of experts and the design team.

2.1.1 The Jigsaw Analogy of Bioinformatics

In the case of bioinformatics, we developed and validated the analogy of putting together a jigsaw puzzle as a way of describing the rapid assessment process of bioinformatics experimental work. In the jigsaw analogy of bioinformatics, the research goal of a bioinformatician can be considered to be the picture on the jigsaw puzzle box. Of course, this picture might be missing or only partially present, reflecting often ill-formed goals of exploratory analyses. Bioinformatic methods are used to either discover new pieces to the puzzle or improve their knowledge of the picture on the box. The jigsaw pieces themselves are therefore abstract entities representing knowledge, most typically data.

The choice of bioinformatics tools available reflect the choice of strategies one may use to solve a jigsaw puzzle, such as finding "edge" or "corner" pieces, or the action of collecting together pieces that look like they might be "images from the same jigsaw".

2.1.2 Jigsaw Preprocessing

The Jigsaw analogy gives us a way of understanding the larger file and data management problems experienced by bioinformaticians: there is not only the problem of determining whether a given piece of a puzzle is part of a puzzle of interest to the scientist, scientists must often unpack dozens and dozens of pieces before they can even begin to assess whether or not they are part of the puzzle of interest.

Online databases, such as GenBank [19] or UniProt [1], provide Web-oriented search tools from which the bioinformatician can assemble collections of sequences on which to perform analyses. These large databases provide data formats that not only deliver the sequence of interest, but a rich set of annotations from relevant publications, to key contextual information about the sequence. Unfortunately, the annotation can be so extensive that these formats are ungainly for the purposes of experimentation. The much simpler FASTA format[1] is the common currency for sequence data. This format provides only the sequence and a customizable one line description of anything wanted by a bioinformatician. When the bioinformatician makes a collection of sequences the FASTA format is most likely to be used. It is difficult, particularly for large collections of sequences, to quickly assess how many sequences are in a FASTA file, when they were downloaded, where they were downloaded from and even what the sequences represent. The bioinformatician may have multiple copies of similar looking data, and may feel that they altered the data somehow during the course of their rapid experimentation in order to correct a problem. As a result, over time, the data can become distrusted, and therefore will need to be downloaded again from the Web. This is no guarantee that the data will be the same, as more or newer data may have become available since the initial experiment was performed. In addition, to see the information in the more extensive file formats, the user will have to browse back to a Web site, or will have to have downloaded the more extensive formats as well as having the sequences in FASTA format.

2.1.3 Solving Jigsaw Puzzles

Various bioinformatics tools are used to both gather pieces of a jigsaw and to create new pieces of jigsaw puzzle. The skill of a bioinformatician is to choose, design or create a tool for a particular jigsaw solving strategy. Work done using tools that have been downloaded and installed or written by the user present a similar problem to that seen with file organization and data management, fragments of files for which annotation is very limited and a lack of trust for files that were generated some time ago. Best practice would dictate that a particular strategy used and the tools used in that strategy should be recorded, as they are in a myGrid workflow [23]. In the exploratory phase of bioinformatics, however, this does not happen for the same reasons described earlier. Consequently, on re-visiting a solved puzzle at a later date, a bioinformatician has to re-think how that particular jigsaw was solved—perhaps re-writing history.

2.1.4 Solving Jigsaw Puzzles

One jigsaw can be assembled in different rooms, with the obvious attendant difficulties. To confound the already complex situation of the use of local programs on the desktop, data are often transferred back to Web based tools and back to the desktop again. The separation of desktop and Web makes recording strategy even more difficult.

2.1.5 Jigsaws on the Web

In contrast, work done on the Web presents different kinds of problems. Results are often graphical, displayed on dynamic Web pages, which makes them difficult to capture. The advantages of using the Web services, such as those available from Taverna [12], are not available to the bioinformatician from the desktop. In working out how to solve a jigsaw, a bioinformatician needs free and open access to such services, outside the confines of a workflow. Yet this stage of testing whether pieces of a jigsaw fit together and match a picture on the box need to be recorded. This use of Web based resources forces the behaviour of copy and paste, which ultimately leads to the fragmentation of data and the loss of any work context. This is much like having a box of jigsaw

[1] http://www.ebi.ac.uk/help/formats_frame.html

pieces, no picture and no where in which to lay the pieces out for inspection. Two pieces of jigsaw data are picked up, compared and put back in the box—leaving no record of what has happened. Consequently each time the problem is encountered it has to be re-created.

2.1.6 Requirements Uncovered via Analogy

This way of formulating the problem - the lack of the workspace or table for doing the jigsaw puzzle - helped us to develop the requirements for such a space - in this case, what might more effectively be recognized in the science community as a lab- or work-bench, rather than a table.

By understanding the associated problems with data and tools in terms of that analogy, we could begin to put together the picture of the kinds of tools we would need to develop (a) to help unpack and sort the pieces to be considered in the puzzle (b) to provide the table space for the analysis and (c) to enable mechanisms whereby each of the tools used in analyzing the pieces can automatically be tracked and recorded for later consideration.

The requirements for this virtual bench therefore, focus on enabling rapid recovery of lab-book like annotations such as why a piece of data was kept, set aside or discarded in a particular process.

Our analogy and subsequent design approach has effectively introduced a new space for work in the *in* silico bioinformaticians lab by introducing a virtual bench. This means that when scientists start up a virtual process, they will now initiate that process at the bench - rather than in the virtual ether of current laptop science. While this bench is putting one new step into the bioinformaticians process, the benefits we're already seeing with the tool set we're developing outweigh any perceived cost of starting a session with the bench. Indeed, the bench is largely transparent: the access of any tool which is associated with the bench automatically communicates its findings through the bench so they can be interrogated at the bioinformaticians convenience. Suffice it to say that by using user centred design methods in general, and Making Tea in particular, we are developing approaches to support bioinformaticians that they are rapidly able to assess, validate and reform.

3. RELATED WORK

This problem of solving the bioinformatics jigsaw is by no means new. Gathering all the jigsaw pieces together into one place so that a particular strategy can be deployed has been a long-standing goal in bioinformatics. Bioinformaticians have been users of the Web since its inception. As autonomous groups of biologists produce data, they wish to make these data available to a wider community. In this way many specialist databases have been produced and made available via the Web. In addition, many community wide databases provide large collections of protein and nucleic acid data. This autonomy naturally leads to heterogeneity and distribution. As a consequence, integration or interoperation between multiple resources has long been a goal within bioinformatics [5].

The Sequence Retrieval Service (SRS) [8] provides a Web based mechanism for querying indexed flat-file data resources. The Web pages also offer access to many standard analysis tools. SRS has been one of the most successful integration mechanisms used in bioinformatics. Whilst SRS provides a common access to many resources, it still relies on a human operator to direct the data between resources.

More recent integration attempts have moved in the direction of automation. iSys [15] provided a common bus into which bioinformatics services could be plugged in order to build applications. The BioKleisli system [6] offered middleware and a query language that could enable sophisticated queries and the mechanism by which these could be passed between resources. These and other systems have, to a greater or lesser extent, integrated bioinformatics resources and enabled complex, multi-source queries to be made. Yet none of these services have provided any support for the wider scientific process and nor did they claim to do so. Whilst integration may be achieved, there is no scope for the development of the analysis to be performed and no recording mechanism for either this exploratory stage or the running of the final experiment.

The myGrid middleware [12, 16] has attempted to address one of these omissions in its provision of workflows to interoperate between multiple bioinformatics resources. Amongst the many services provided by myGrid, there are several pertinent to the wider scientific process. myGrid offers workflow creation and enactment through Taverna [12], and FreeFluo [2], its enactment engine. Through these services, highly sophisticated workflows can be built [16]. One problem encountered when building a workflow is the discovery of services which are to be incorporated into the workflow [18]. myGrid offers a semantic service discovery called Feta [10]. A controlled vocabulary is used to annotate the inputs, outputs and task performed by services. These can be used to query a service registry to find Web Services and other workflows. myGrid also offers a provenance service that records information about the Web Services called by the workflow; the derivation path for the data generated in the workflow; the organizational metadata for the workflow[23] and any semantic annotations made by the user using terms from the myGrid ontology[21], which is also used in Feta [10]. This provides a basic lab book for the analysis itself. A bioinformatician can use these data, stored in RDF[2], to explore their experimental holding, debug an experiment, undertake impact analyses, etc.

The myGrid services, currently seen via the Taverna workbench, offer the ability to create and run bioinformatics analyses and record rich metadata about these experiments. There still remains, however, little support from Taverna or any of the other Web Service based bioinformatics systems or workflow tools such as PathPort [7] or Discovery Net [14]. The myTea user interface reported in this paper does not seek to replace tools such as Taverna, but to provide a wider support for their use. The myTea workbench supports a bioinformatician in the stages leading up to the use of Taverna.

This exploratory phase is currently supported only by the users file store usage. Web pages of results may be saved to disk, perhaps to a folder for an investigation. Naming of these files is usually non-systematic, making any review difficult.

This problem has also been recognized in the Utopia project [13]. In Utopia the aim is to build a toolkit for building bioinformatics applications. Underlying the toolkit is Ambrosia—the Utopia file store. Ambrosia blurs the distinction between the desktop and the outside world so that local data and tools, together with those found on the Web all appear to be on the desktop. By hiding the plumbing necessary for handling bioinformatics data and

[2] Resource Description Framework http://www.w3.org

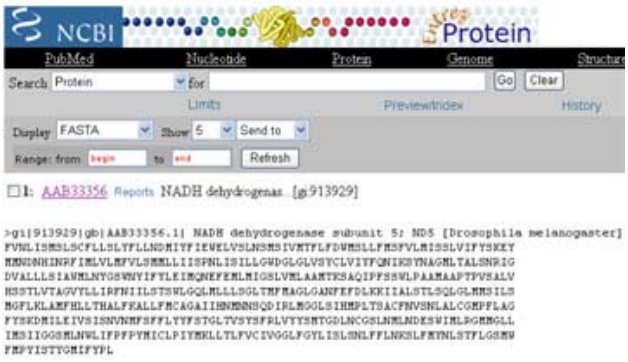

Figure 3. Gene sequence accessed from a Web database.

combining this with metadata for handling, for instance, changes in versions of data, much of the overhead of building *ad hoc* applications is removed. Our workbench can be seen to be a general application for performing *ad hoc* bioinformatics tasks built over such a system.

4. FRAMEWORK ARCHITECTURE

Based on this analysis of bioinformatics practice, the architecture we have developed supports three core components: (1) the **Report**, which functions as an automated but annotatable lab book; (2) the **Bench** which provides a mechanism for tracking processes which scientists may wish to have reported to their Report/virtual lab book, and (3) the **Datastore** which acts as a repository for data produced by the Bench and is used by the Report. We refer to these components collectively as the **myTea system.**

A way of imagining the system working is the following scenario: A scientist downloads a sequence from a Web database (Figure 3). Rather than copying and pasting this sequence into a text

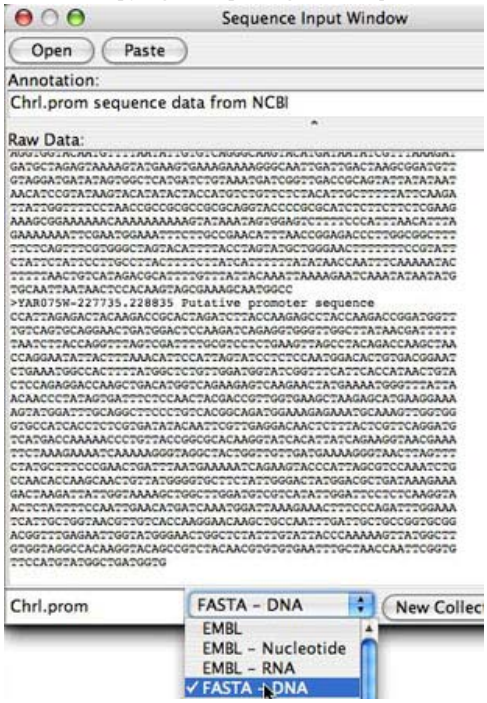

Figure 4. Sequence Editor with raw sequence pasted in from sequence copied from Web database.

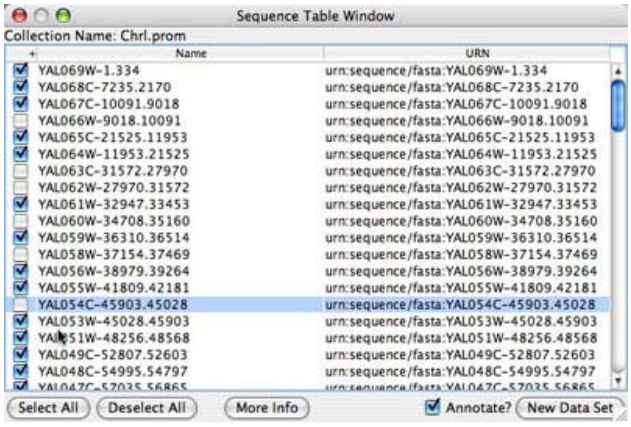

Figure 5. Sequence Editor: once sequences have been automatically parsed into components which can be turned on or off. The resulting collection of "on"s can then be processed by a tool such as a sequence alignment editor. A record is generated noting which components were turned off and why.

editor as they would now, and manually pour through the data to pull out useable sequences they invoke a myTea tool, the sequence editor (Figure 4) which automatically parses the sequence into meaningful components (Figure 5) – a task which the scientist previously did manually.

Opening the sequence alignment editor, a local tool we have developed as part of the myTea system, automatically invokes both the data store and the Bench. The Bench tracks the job being performed – an analysis of a sequence – and the datastore keeps track of what data is used, where it is located, and what manipulations have been performed. At any point in the process, the scientist can pull up the Report to see a record of what processes have been invoked on what (Figure 6).

For example, they might see that before running an alignment on the sequence that was captured in the sequence editor, one part of the sequence collection was turned off, and according to the annotation, it was turned off because it was deemed to be poor data. A link to the source data is also available so that the scientist can recover the original sequence. The Bench transactions, therefore, are automatically recorded in the Report. The Report can then be annotated at the bioinformaticians' convenience.

The workbench maintains the contextual history of all the data gathered. This is a necessary step in re-creating the lab book, but it is not sufficient for a full record of what has been done. To do this adequately, a scientist needs to take notes. Therefore, at any

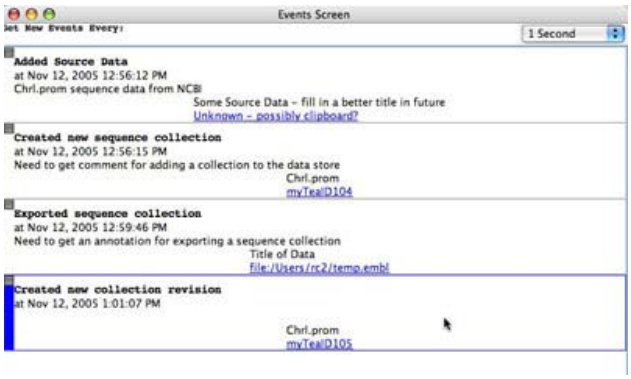

Figure 6. Summary Event View of Report

dialogue with the user, an opportunity to semantically annotate is offered. Terms from the myTea ontology and myGrid services ontology, which includes general domain concepts from biology [21] be offered. We offer terms that describe the state of work – "finished", "unsatisfactory", "useful", "needs attention", etc. these will be used to keep track of the state of work. Such semantic annotations can add value to the automatically recorded provenance data by allowing more informed querying over those data [23]. In this respect, our design exploits well-understood principles of user interaction design: reduce forced divided attention [20]. That is, our approach lets the scientists focus on the fast processing of sequences, while the system automatically tracks what happens to their data. Then when the scientist wishes to focus on annotation, they can do so.

The concept of the Report generator combined with pervasive annotation opportunities goes beyond our primary aim of reintroducing the benefits of the lab-book into bioinformatics. From the bioinformaticians with which we have spoken, it seems that the ability to generate reports from the recorded context will be useful in two major ways. The Report itself has provisions for sharing findings with others: the scientist can decide whether they wish to share just the description of the processes carried out, or

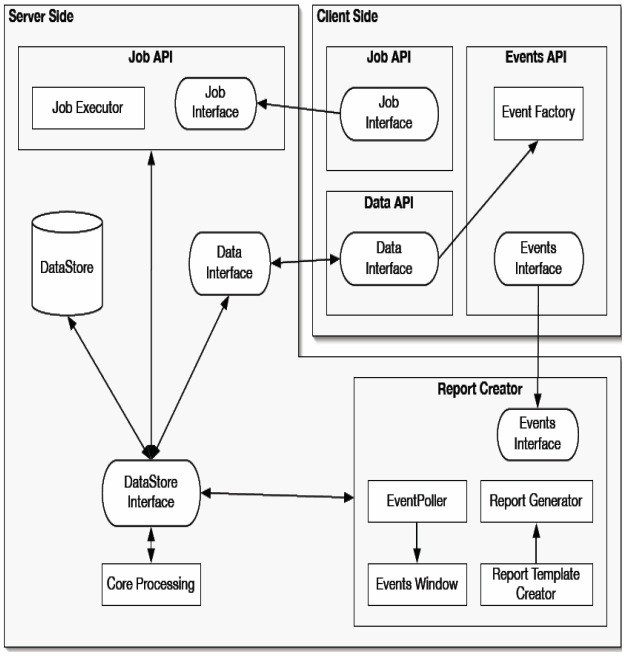

Figure 7. Client-Server Architecture for myTea System

provide access to the source data. The user defining the report for someone else to view may wish to exhibit some important result, yet the viewer of the report may want to see more. This is made possible with a Report , as the viewer, if they have the appropriate permissions, may drill down into the underlying data, annotations and metadata. Consequently the viewer may be able to ask more pertinent questions regarding the methods chosen or the data used that they were previously forced to merely assume were reliable. We are keen to understand if this lightweight approach both to annotation and sharing creates new opportunities for scientific collaboration. Also, we understand that reports will be run by the user for their own benefit. This creates a lab-book that is searchable and can present the user with the information they require, including their own thoughts in the forms of annotation and context in the form of captured provenance data.

The goal of our approach in the myTea system is to provide myTea as an open platform for e-Science developers. To that end, the system is deployed as plug-in style client-server architecture, as shown in Figure 7. This approach allows application developers in the e-Science space either to utilize the services provided by the myTea environment by writing wrappers for applications that already exist, or by directly integrating myTea services into their own application architectures. Communication between the myTea service and the client tools is implemented using Java RMI. This means that tools in myTea environment can use either Java RMI or Web Services to communicate with the myTea system.

In terms of implementation, we are using Semantic Web technologies and languages for data communication, discovery and storage. This includes triple stores for storage, RDF for describing the data, ontologies to support inference over the data and OWL to describe the ontologies. For storage, we are using a local triple store the contents are represented by a combination of the myGrid ontology [21], the myTea ontology and what we are calling the myTea-BioJava ontology that is based on the BioJava[3] class hierarchy. The myTea ontology represents concepts unique to myTea, such as jobs, sequence collections etc. The second is a bioinformatics ontology that uses the properties of the well-known and well-used BioJava class data (such as sequence data) exactly as BioJava stores them in memory, just in triples and hence semantically accessible. The use of the myGrid ontology as well as our BioJava ontology means that any application that is written for the widely used myGrid workflows and myGrid data stores, or uses the BioJava libraries can then easily also access data in the myTea data store.

The rationale for using the Semantic Web approach rather than a database only is encapsulated in the potential for the Semantic Web to make it easier for applications developers to connect researchers with other data sources and researchers with other researchers. For instance, we are connecting concepts from a variety of services that we wish to be able to integrate in the Bench. By using ontologies to define these concepts and the triple store to hold these concepts, it becomes easy for developers to build on top of these collections, and infer new knowledge from what is stored.

When data is asserted into the triple store and is annotated in one of our ontologies, then the triple store can infer links between them automatically, rather than having to create the link manually. This is a powerful effect. When reviewing his or her experiment holdings, for example, the aggregation of triples through mechanisms such as the Life Science Identifier (LSID) [4] it might be noticed that much activity is centered about a particular sequence. This can reveal the importance of that sequence to a bioinformatician and the context and semantic annotation recorded by the myTea system can enable him or her to realize why.

4.1 Client Side

The client side interface to the myTea architecture consists of three distinct components, the Events API, the Job API and the Data API as shown in Figure 3, above. In terms of the scenario of a scientist processing a sequence, data flows through the client side architecture in the following way: the data is stored using the myTea data store API. The events API is used to generate an

[3] http://www.biojava.org

event that says that some sequence data was retrieved from the Web (this is associated with the data in the data store automatically). The system then lets the user run a process on the sequences, such as an alignment Web-service that tries to automatically align the sequences. This is executed using the Bench's Job API. The application generates an event again to say that the user has performed an alignment and then stores the results using the data store API.

4.1.1 Events API

The events API allows the client application to post event notifications to the myTea environment. These events are recorded in the data store to be used by the myTea system to generate the user reports. An example of an event may be "a collection of sequences was created" accompanied by some annotations made by the scientist about the reasoning behind this and a link to the data in the form of a URI, a file path, an LSID, or a MyTeaID.

4.1.2 Job API

The job API allows external applications to execute jobs through the myTea environment using data stored in the myTea repository (or any data specified externally). Also, applications using the myTea environment can execute jobs within applications that implement this API.

4.1.3 Data API

The data API allows applications to store and retrieve data from the local myTea data store. The data store uses the myTea ontology, myGrid ontology and LSIDs to provide as unified an approach to classifying objects within the bioinformatics domain as possible.

4.2 Server Side

4.2.1 Report Creator

The Report creation system works on the server side, as the Report is created from events registered with the myTea system using the Events Interface. Events consist mainly of a meaningful title, an annotation added by the scientist and data associated with the event. An example might be "a number of sequences were downloaded from a database". The scientist then puts together a Report template which is a structured display of selected events. The Report Generator then takes the template and fills out information with data stored in the myTea data store or the contents of files or Web pages. The Report can be used as a reference for scientists about what work they've done recently or in the past, or as a means of creating reports for their supervisor for example.

4.2.2 Job Executor

The job executor provides the means by which jobs (which can be local applications, Web services or myGrid workflows) can be given data, executed, and the results retrieved. The advantage of executing these through the myTea environment is that the chain of provenance between the source data and any final results can be maintained throughout an entire project while not having to pre-specify what jobs will be done. It allows the research scientist flexibility in the work practices.

4.2.3 Data Store

The data store is a triple store based on the Sesame API [3]. Data is stored and inferences made across it using the myGrid and the myTea OWL ontologies.

5. WORKBENCH IN USE: DATA MANAGEMENT

A key part of the Workbench design has been the implementation of the Dataset Manager, which treats biological sequences much like a reference manager treats references. The aim is to provide the bioinformatician with the tools needed to browse, search, annotate, rearrange, use and reuse downloaded sequences, thereby greatly reducing the need for locally stored files that contain sequence data. The motivation for this is to reduce both the need for a new file to be created every time a new set of sequences is needed and also the tendency for reacquiring data from the Web. In addition, semantic annotation is preserved, and even provenance data of the sequences adds value as versions of sequences can be checked for consistency when reviewing results. In this way, through retention of provenance or contextual information, trustworthiness is built into the system.

Currently this application allows the user to gather together collections of sequences by importing any of the common formats supported by the BioJava API[4]. Import can take place via a copy and paste action from a Web site or as an import of a file. The next version will see the incorporation of Web Services that will mean that the user does not have to visit the Web through a Web page in order to import sequences, in line with our goal of a unified environment. Sequences are central to myriad bioinformatic tasks, but they are not the whole of bioinformatics. We have started with sequences as proof of concept and will expand to other categories of data.

One of the rate-limiting factors for bioinformatics research occurs when the bioinformatician has limited knowledge about online tools that already exist. There are two main ways in which a lack of knowledge about available online tools can hinder bioinformatics research. In the first instance, the bioinformatician may choose to go and look for a service online. There are several indexes of online bioinformatics tools (e.g. www.expasy.org) which may be visited to help discover the some of the more commonly used tools. The bioinformatician often does not know if the service they require is at all available and therefore these searches may be short lived. If the service is discovered, it may not be adequately described, such that the bioinformatician has to invest more time understanding how the tool works, including what sort of inputs and outputs are acceptable. Another barrier to using tools that already exist is often that the data needs to be transformed in some way in order for it to be used. The ability to find the right tool is of high importance, as bioinformaticians are often capable programmers and can write bespoke software to satisfy their analytical needs. Often they will do by looking over online tools that are capable of the same task, resulting in a large replication of effort within the field.

Our workbench design tackles this behavior, both by providing the framework in which data can be prepared in the most widely used formats, and by incorporating the FETA service discovery tool [10]. By reducing the amount of time needed to find, learn about and use a tool that is already available as a Web Service, we

[4] http://www.biojava.org

aim to provide a way in which bioinformaticians can quickly adopt new approaches to their everyday work. Also, by providing a unified place in which to discover these services, much repetition of effort is removed.

Lab books are the traditional way in which scientists record what they have done. The *in silico* nature of bioinformatics means that the lab book does not transfer well to this domain. This is true of both paper and electronic lab book; neither offers sufficient connection between what is recorded on the desktop and the notes a scientist would naturally take during a wet experiment. By making things on the Web available on the desktop and keeping links between the things done in this exploratory phase of bioinformatics, we make it easier for a scientist to re-establish the context of what he or she has done; this is often the job of a lab book. By keeping this context the data stored becomes more reliable and the annotations show the user this context.

As the data become instantiated within the framework, the costs associated with their normal everyday use are reduced. The workbench imposes only those constraints on what the user does that are fundamental to bioinformatics, such as only applying protein services to protein data. This leaves the bioinformatician free to perform any action, whether or not it is currently understood or accepted in bioinformatics practice. For instance, the authors can think of no reason why one would wish to create a sequence collection containing both nucleic acid and protein sequences, so the open world design of the workbench does not prevent this from happening. It is a clear principle within this design not to block creativity.

Indeed, our approach throughout has been to ground our design in supporting bioinformaticians practice. This has meant frequent design iterations, feedback and evaluation with bioinformaticians from a variety of approaches. Because the evolution of our approach has been first facilitated by the shared jigsaw analogy, and second developed between scientists and designers through that analogy and third frequently assessed with light weight reality check style evaluations with the practitioners, we have had successful take up of these tools during each phase of the development.

6. FUTURE WORK

The basic design of the workbench is in place and it supports the essential practices of a bioinformatician. We will extend the workbench to a wider form of bioinformatics. That is, currently the myTea system is focused on one of the most common applications of bioinformatics, that of sequence analysis. We will extend the capabilities of the system to deal with other forms of biological information, such as protein structures, and provide access to the appropriate tools from the Job manager

We will also provide more “out of the box” connectivity of the Bench with more of the myGrid services that support *in silico* analysis in bioinformatics. Eventually we expect that a tool such as Taverna will be available directly on the workbench and all the Web services Taverna exposes will be available to the bench as well, rather than restricted to within Taverna’s workflows as they are currently.

One goal of this work is to enable bioinformaticians to formulate experiments and then move them to a formal workflow environment seamlessly, all the time retaining records of context in a lab-book. In the future, as bioinformatics and other *in silico* disciplines, such as Chemistry Informatics, move from specialist disciplines back into the wet lab, designs such as our workbench should become an extension of the current desktop. The openness of this design, based on how bioinformaticians perform their work – the exploratory phase – makes the design extensible to any discipline that works primarily on the Web.

To this end, we are working towards an SDK for software developers who wish to construct tools from scratch that will work with the myTea framework. We are also building a suite of APIs and associated wrappers so that developers can wrap their existing tools to take advantage of the Bench and associated framework services. We are also about to begin work with ChemInformatics to investigate requirements for porting myTea to this space.

7. CONCLUSION

In this paper we have presented an overview of the myTea system. We have shown that this system makes three contributions to Web-based e-Science: (1) facilitating the “last mile” of connecting Web-based services with desktop/local processing; (2) providing mechanisms to process and store desktop analysis within Web-accessible/sharable Semantic Web technologies such as triplestores and ontologies; (3) providing new mechanisms through these services to facilitate easy manipulation and owner-determined sharing of reports and/or reports and associated data.

Early evaluation of these services has shown that scientists are keen to embrace the features these tools are enabling. The Sequence editor alone has been met with eager use. By having new tools which support the way the scientist works, which support automatic annotation and enable sharing back to the Web, we now have a platform that will let us investigate some of the core motivating premises of the e-Science agenda: better science will result from better capture, annotation and sharing of data. With this framework in place, and with tools built to take advantage of its automated reporting features via the Bench, we will be able to carry out longitudinal studies to let us assess the degree to which not only do we facilitate the individual scientist’s practice, but the degree to which sharing and possibly new science emerges from such activities.

8. ACKNOWLEDGEMENTS

myTea is funded by EPSRC grant EP/C002180/1

9. REFERENCES

1. A., B., R., A., C.H., W., W.C., B., B., B., S., F., E., G., H., H., R., L., M., M., M.J., M., A., N.D., C., O.D., N., R. and S., Y.L. The Universal Protein Resource (UniProt). *Nucleic Acids Res, 33* ((Database Issue)). D154-159.
2. Addis, M., Greenwood, M., Oinn, T., Li, P., Wipat, A., Ferris, J. and Marvin, D. Experiences with workflow specification and enactment for bioinformatics.
3. Broekstra, J., Kampman, A. and Harmelen, F.v., Sesame: A generic architecture for storing and querying rdf and rdf schema. in *Proc. of the First International Semantic Web Conference*, (Sardinia, Italy, 2002).
4. Clark, T. and Liefeld, S.M. Globally Distributed Object Identification for Biological Knowledgebases. *Briefings in Bioinformatics, 5* (1). 59--70.

5. Davidson, S.B., Overton, C. and Buneman, P. Challenges in Integrating Biological Data Sources. *Journal of Computational Biology, 2* (4). 557--572.
6. Davidson, S.B., Overton, C., Tannen, V. and Wong, L.S. BioKleisli: A Digital Library for Biomedical Researchers. *Journal of Digital Libraries, 1* (1).
7. Eckart, J.D. and Sobral, B.W.S. A Life Scientist's Gateway
to Distributed Data Management and Computing: The PathPort/ToolBus Framework. *OMICS: A Journal of Integrative Biology 7*. 79-88.
8. Etzold, T., Ulyanov, A. and Argos, P. SRS: Information Retrieval System for Molecular Biology Data Banks. *Methods in Enzymology, 266*. 114--128.
9. Kaminski, N. Bioinformatics. A user's perspective. . *Am J Respir Cell Mol Biol 23*. 705-711.
10. Lord, P., Alper, P., Wroe, C. and Goble, C. Feta: A light-weight architecture for user oriented semantic service discovery *European Semantic Web Conference*, Lecture Notes in Computer Science, Springer-Verlag, Heracleon, Crete, 2005.
11. m. c, s., Gareth, H., Hugo, M., Graham, S. and Jeremy, F. Making tea: iterative design through analogy *Proceedings of the 2004 conference on Designing interactive systems: processes, practices, methods, and techniques*, ACM Press, Cambridge, MA, USA, 2004.
12. Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Glover, K., Pocock, M.R., Wipat, A. and Li, P. Taverna: A tool for the composition and enactment of bioinformatics workflows. *bioinformatics, 20*.
13. Pettifer, S.R., Sinnott, J.R. and Attwood, T.K. UTOPIA - user-friendly tools for operating informatics applications*Comparative and Functional Genomics5*(1). 56-60.
14. Rowe, A., Kalaitzopoulos, D., Osmond, M., Ghanem, M. and Guo, Y. The discovery net system for high throughput bioinformatics. *Bioinformatics, 19* (90001). 225i-231.
15. Siepel, A., Farmer, A., Tolopko, A., Zhuang, M., Mendes, P., Beavis, W. and Sobral, B. ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics resources. *Bioinformatics, 17* (1). 83-94.
16. Stevens, R., Tipney, H.J., Wroe, C., Oinn, T., Senger, M., Lord, P., Goble, C., Brass, A. and Tassabehji, M. Exploring Williams-Beuren Syndrome Using myGrid. *bioinformatics, 20*. i303-i310.
17. Stevens, R.D., Goble, C.A., Baker, P. and Brass, A. A Classification of Tasks in Bioinformatics. *Bioinformatics, 17* (2). 180--188.
18. Tran, D., Dubay, C., Gorman, P. and Hersh, W. Applying task analysis to describe and facilitate bioinformatics tasks *MEDINFO*, Ios Press, 2004.
19. Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., Helmberg, W., Kenton, D.L., Khovayko, O., Lipman, D.J., Madden, T.L., Maglott, D.R., Ostell, J., Pontius, J.U., Pruitt, K.D., Schuler, G.D., Schriml, L.M., Sequeira, E., Sherry, S.T., Sirotkin, K., Starchenko, G., Suzek, T.O., Tatusov, R., Tatusova, T.A., Wagner, L. and Yaschenko, E. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res, 33* (Database issue). D39-45.
20. Wickens, C.D., Hollands, J. D. *Engineering Psychology and Human Performance*. Prentice Hall, 2000.
21. Wroe, C., Stevens, R., Goble, C., Roberts, A. and Greenwood, M. A Suite of DAML+OIL Ontologies to Describe Bioinformatics Web Services and Data. *the International Journal of Cooperative Information Systems, 12* (2). 597--624.
22. Zhao, J., Goble, C.A., Greenwood, M., Wroe, C. and Stevens, R. Annotating, linking and browsing provenance logs for e-Science.
23. Zhao, J., Wroe, C., Goble, C., Stevens, R., Quan, D. and Greenwood, M. Using Semantic Web Technologies for Representing e-Science Provenance *Proc 3rd International Semantic Web Conference ISWC2004*, Hiroshima, Japan, 9-11 Nov 2004 , Springer LNCS Hiroshima, Japan, 2004.