Transparent interaction; dynamic generation: context histories for shared science

mc schraefel, Sacha Brostoff, Ray Cooke
IAM Group
University of Southampton
Southampton, Hants, UK
http://www.ecs.soton.ac.uk
[mc, sb4, rc2] @soton.ac.uk

Robert Stevens, Andrew Gibson
BioHealth Informatics Group
School of Computer Science
University of Manchester
Manchester, UK
[robert.stevens,
Andrew.gibson]@manchester.ac.uk

ABSTRACT

Scientists who do *in silico* or computer-based experiments use general purpose computer tools, like Web browsers and word processors to carry out their tasks. As such, they have no formal file management support for collecting, coordinating, annotating and reflecting on their digital experimental traces. In this presentation we look at how we are exploring the use of implicit context histories to support scientists with both formal and everyday collaborations. We describe our goal to utilize the non-intrusive discovery and use of implicit contexts generated by task-based interactions in order to represent back, on demand, how one file or collection may be related to another. Such annotatable reports can then either be shared or used as inputs for further service requests for selected data.

Keywords

Transparent interaction, file management, metadata, semantic web.

INTRODUCTION

EScience is a new domain for HCI research. EScience seeks to use new networked computing opportunities such as the Grid to enable new science. [3] Part of these emerging requirements in this new field is to investigate ways to support a range of activities from the particular needs of scientific collaboration, to the requirements for demonstratable trustfulness of a system. One of the challenges in this space is to look at ways to support and enhance existing practice, as per Ubicomp's goals of transparent interaction [1, 2] in these rich lab-orineted environments, since much current practice is carried out with tools (from paper to mechanical devices to computer) which were not designed for data interchange or collaborative reflection. As a case in point, we have been looking at the practices of bioinformaticians, scientists who carry out their work almost exclusively *in silico* or on the computer, rather than *in vitro*, in the traditional wet lab. It would seem that in such an environment where work is already digital, integration and sharing of data would be less of a challenge than with their paper-bound colleagues. Alas, no. These disparate file traces have no medium through which they may be associated. Context histories, however, provide a possible vehicle for dynamic, sharable associations.

We have only recently completed our ethnographic studies and technologies review for the bioinformatics design space. In the following sections, therefore, we wish to report on an overview of these findings, the current design strategies based on them, their relation to context histories, and concerns surrounding the use/propagation of same.

Background: Experimental Recording in Bioinformatics

Bioinformaticians by way of background, are involved in molecular biological research. They run complex scientific experiments on myriads of biological data. Rather than running these experiments in the messiness of a traditional wet lab, their lab is generally a laptop computer connected to the Internet. This virtual lab is still, frequently, just as messy a space as their physical counterparts (see [14] for views of wet labs): digital files that are created in the heat of the experimental moment mayn't be saved with optimal names for later discovery. It is also up to the scientist to crack open a text editor in order to create annotations about a finding in progress. As has been shown elsewhere [15] copying data from the web into new files frequently leads to critical data, like descriptive names or originating URLs to be left off, making later recovery of information difficult to accomplish. Some bespoke services, such as myGrid, which run workflows of search patterns on gene databases have saved the scientist days of effort in having to run these web site crawls manually [16, 17] but the runs themselves still create legions of files associated with a given experiment which must be analyzed, assessed, and referenced as relevant or not.

As such, the recording of experiments is a largely ad hoc (or post hoc) and manual process, requiring the scientists to cadge together a variety of existing general applications (Web browsers, word processors, tools they may have written themselves for specific analytical tasks) to support their work. In other words, these new *in silico* based scientists do not have what their traditional wet lab colleagues routinely have to track the progress of an experiment: they do not always have or use lab books. We have observed that many bioinformaticians do not use lab books; implicit notes are taken in the creation of a file store (folder names, dates, file names and readme files). In addition, if data are lost or uncertain, an experiment can be re-run simply in a manner not possible at the wet lab bench.

Lab books themselves, however, are not an optimal solution. Going back to paper in a digital field re-introduces the disadvantages of paper, the lack of sharability of results being key.

REQUIREMENTS

In observations of and meetings with bioinformaticians, it is clear that they would like a utility that would allow them to

- Generate dynamic reports referencing and linking to related files on a particular experiment, both the data and supporting material
- Allow multiple views on how one file relates to another
- Supports annotation of files by meaningful markers, both the for the biology, bioinformatics as well as the process of discovery and investigation itself
- Supports sharing a subset of these notes and files for collaboration, itself producing further annotations

The scientists have asked us for these types of controls not only in order to help them find previously potentially mislaid files (experiments can run for months or years), but also to help them share the state of their work, or subsets of it, readily with other collaborators.

They have also asked us to provide not only machine support for dynamic report generation, but human support, such as the ability to define a naming convention for a series of files and to have that convention (date, gene family for instance) applied automatically. Richer kinds of labels have also been requested, so that they can see at a glance what files are active which are potential candidates and which have been used and discarded.

Our frame of reference for these requirements has been to find a way to put some of the benefits of the lab book into the bioinformatitan's process. In particular, we wish to support the lab book's functionality to provide in one place a view on the processes and annotations on those processes associated with an experiment or collection of experiments, and the ability to browse through previous work. It is clear, however, that asking the scientist to carry out the file management tasks they would need to do to create these views manually is unacceptable. We also do not wish to ask them to change their favorite tools in order to use a "digital lab book" that would attempt to be part browser, part email client, part scrap book and part word processor. We would rather leverage the input/output created in using these tools and make such reports which reference this I/O available on demand.

(Implicit) Context Histories

The use of context histories is a means towards creating just such transparent, reusable tracking of associated information. In this case, we understand context histories to be the history of interactions traceable within the interactions with the computer which can be seen to be associated with a given experiment. We have been thinking of these as implicit or possibly latent context histories since they will be teased out from the history of input/output interactions logged in the system as files are created, manipulated and deleted. To support transparency, these histories will be made available on demand, linked to the appropriate files, and providing opportunities for annotation on the context as well as annotation of a specific artifact. These context histories can then be viewed from multiple perspectives, shared and altered by scientists with their communities to reflect on the progress of a study for feedback, or to share the evidence of a specific conclusion.

Our challenges are

- to derive the correct/required contexts from the available interaction history of a scientist's laptop,
- to provide appropriate forms of representation for viewing these histories along multiple perspectives.
- To annotate and/or tag the files in ways which support organization in these contexts
- To ensure that manual effort can remain at the level of a secondary task rather than be forced regularly into primary attention.

The last point in particular is inspired by concepts like marking menus [9] which support secondary interaction of tasks like copying or pasting by allowing a simple gesture to invoke the action anywhere on the screen rather than requiring a person to acquire a specific target, navigate the associated menu, and activate the command. We wish to support any required manual annotation of files in a similarly transparent, context-based approach. Our goal, however, is that we will be able to deduce sufficient value from a scientists' interactions that we will be able to build up a context history and use this for constructing appropriate associations. It will then be easier to subtract mis-additions or flag/annotate collected files than either to construct all contexts and additions manually.

RELATED WORK

Our approach is informed by three related efforts: innovative research in desktop replacement or desktop assistant models, virtual notebook applications and Semantic Web frameworks. We describe each in turn.

Desktop Replacements

The closest related work to the type of transparent interaction we are describing are desktop replacement systems which either replace or enchance the traditional desktop. Reikimoto's Timescape is perhaps one of the most oft-sited examples of such a system. In Timescape, the paradigm of file-based hierarchies is changed to temporal views of spatially associated filed for exploring information contexts [12]. A person can therefore travel backwards or forwards in time to watch how an interaction with a file may have progressed.

Presto [5] is a java based networked desktop replacement, enabled by a sophisticated infrastructure to trap changes to documents/data, and which allows much greater flexibility in document organization than traditional hierarchical file systems. It interoperate with Solaris, Windows NT, and common applications like MS Word, and uses automated (through feature extraction) and manually generated attributes to group documents. It concentrates on dynamic reorganization of objects on the desktop, rather than generating a history that can be shared (although collections are shared). It has multiple inheritance - documents can appear in more than one category or collection. It has a centralized metadata store, that runs across the network extracting features from document contents and existing metadata (creation time, owner, filename, etc.) from where they are stored locally or on shared resources and visualizes them on each user's Presto desktop via an application called Vista. These documents can be launched and worked upon with the user's usual tools, but need to be manually associated with particular projects or categories.

Desktop Supplements for Context

As an alternate to desktop replacements, there are a set of applications which may be considered to be desktop supplements which endeavor to derive contextual associations or support their discovery.

UMEA [8] is an application that tracks activity and the objects of those activities, and creates a History log organized according to projects. Metadata describes the context in which the work is being carried out, which can then be used for retrieving contexts. UMEA, however, requires users to set up projects and then manually to switch between them in the UMEA interface. If a document is opened during a particular project context, then the document is associated with that project by UMEA. This can lead to mode errors, where the user forgets to switch project contexts before performing an action. This leads to the action or object being mistakenly classified, for example as belonging to the "workshop" project rather than the "funding proposal" project., UMEA therefore allows manual reclassification. Like Timescape, the interface allows different views, including a calendar view. It also allows the launching of PIM applications such as. Sticky notes, to do lists, and emails to project related contacts.

Milestones [13], is a visualisation for Stuff I've Seen [6] – a Microsoft research desktop search tool. It uses events and images from the user's wider context (such as headlines from world or local news, digital photographs the user took and stored on the computer at that time, etc) to act as landmarks and cue and orient the user in a timeline view (of search results). Episodic memory is therefore used to cue the user's recall of context, and was found to speed retrieval of desired items from search results compared to a view with no landmarks. The tool does not currently support user-authored annotation of the things shown for cuing context.

OnCue [4] rather than watching file I/O, monitors the clipboard in order to provide associated available services from postal code look ups to historgram generation from table data. OnCue is inspirational in the kinds of context-aware services it provides, and with which we would wish to supplement any contextual association of information.

Virtual Notebooks

Virtual Notebooks, like their physical cousins, support note taking and artefact pasting. They also provide additional digital features which enhance their data collection value. Some exemplars are Tinderbox (eastgatesystems.com) NoteTaker (aquaminds.com) and NoteBook (circusponies.com) which support direct entry of information, such as pasting in screen shots or web information, making notes or outlines, and publishing contents of pages or whole notebooks to the web. Most can output to XML and provide indexing for rapid searches. Tinderbox adds the interesting feature of providng agentware to data mine collections of information in order to find new possible associations in the data not previously noticed. We are strongly interested in the features which these notebooks provide for freely associating and cataloging multiple types of media. These applications are designed, however, for user-determined addition of content to the books. Our approach will be an effort to generate much of the content by the discovery of implicit contexts, supplemented by opportunities to add, subtract or annotate content manually.

Semantic Web Frameworks

The Semantic Web utilizes metadata that is represented in triples of subject-predicate-object. This simple structure can then be associated with ontologies representing classes of entities. The power of this ontology-informed approach to metadata means that we can use inference to derive new knowledge not explicitly stated in the data. Two different files which say nothing in their content that would relate them may still be inferred to be related based on some other association apparent in the metadata, as mediated through an ontology. It is this power of association that we think can be most valuable in helping to connect a scientists' local information with global contexts. As it stands, the eScience project in its utilization of the Grid (or what the NSF in the States refers to as "cyberinfrastructure) has been developing technologies which support the Semantic Web as a communication layer for eScience Grid applications. In an earlier project with synthetic chemists, we were able to capture their experiments in plain English and translate these into Semantic Web parlance for concurrent publication of results to the Grid [7]. The use of an ontology for mapping the data meant that other services could use that ontology for interpreting these results against their own, and thus know how to process this data for their own requirements. It is because of this local/global flexibility for data reuse that we are interested in supporting a semantic web layer as a way to mediate context histories.

Certain frameworks already exist which we are exploring for adaptation in the current project. Haystack from MIT [11], at two years old, is the most mature. It provides a framework for developing Semantic Web applications. Its core demonstrator has been a personal information management system. Like virtual notebooks, it relies on the manual capture of information, but its use of a semantic back end, through ontologies, allows inferencing over data. In this way, making a plane booking will result in a calendar being updated with new location information for the dates away. To date their have been known issues with speed in applying to real world data, and predictable resistance to using one monolithic tool rather than being able to use one's own communication and scheduling tools. Recently, Haystack has been refining its framework and working on speed so we looking forward to exploring this further.

UTOPIA is another eScience Semantic Web project which can monitor activities in a defined virtual disk/work environment [10]. While such monitoring is potentially ideal for deriving context histories and translating them to a Semantic Web layer, it requires scientists to use a network disk mounted on their desktop. Files are saved to this virtual disk. As we will look at later, there is considerable apprehension in the community to having data stored on a remote device rather than first and foremost on one's own hard drive. Utopia is also as yet an early technology, not yet released as a framework. The Utopia group, however, is keen to have feedback from the interaction community in order to understand better what services/interactions in needs to support.

One of the chief concerns relayed to us from the bioinformatics community is the need to have flexible visualizations. mSpace (www.mspace.fm) is an interaction model currently implemented on Semantic Web protocols. The model supports user-determined arrangement of an information space in order to support exploration of relationships of the data from multiple perspectives. We are looking at adapting the mSpace software framework to provide local visualizations of the relations in the information.

APPROACH

As can be seen, there are already a variety of tools we can draw on for supporting the types of transparent interactions we wish to explore in utilizing context histories. We are not committed to any particular tool or framework, nor do we need to be, since our main goal is to explore the interactions we may be able to support in using context histories.

In keeping with the EScience lean towards Semantic Web technologies we do wish to create a semantic layer that can translate activities into the appropriate formats for semantic web service interaction.

In terms of interaction, while we want to be able to generate notebook type reports, it is clear from our early ethnography that requiring scientists to manage digital notebooks while carrying out digital experiments has a higher cost in terms of required steps than using a paper lab book. This kind of forced divided attention between file management and experimental activities is counter productive. Therefore we will investigate leveraging the type of transparent capture of file I/O activities demonstrated in Timescape and Presto. We will not be replacing the desktop, though, but will want the kind of project-sensitive associations found in UMEA, but without the required manual context switching. While we can leverage certain cues for context discovery -- a search in a gene database is likely part of an experiment; a search for a bike is less likely to be part – other kinds of cues, such as time, are more problematic. What is unimportant today may prove important tomorrow. Therefore tracking versioning on digital artifacts in a way similar to Timescape may be significant for recovery of context.

Risks of Contexts: Concerns for Design

In bioinformatics, privacy/security of data is a critical concern: any contextual history will almost always be reflecting traces of privileged data. This engagement with privileged data also relates to notions of perceived confidence in any deliverable system. Currently privileged bioinformatics data and related material is kept locally by individual scientists on computers they control. Solutions like Utopia which can only trace file I/O by using networked services rather than locally controlled machines are viewed with suspicion. Likewise, companies sometimes provide privileged data to bioinformatics scientists. These stakeholders will also need to be convinced that their data will be secure. Exposing the context of what scientists are doing with this data could be considered a risk. We will therefore examine current sharing practices, contexts and investigate our users' desired levels of data confidentiality, integrity and availability in order to design lightweight authorization models and authentication protocols. This may include expanding our design space from users to stakeholders, so that it includes data donors as well as data recipients. Where privacy is a concern, deploying encryption is initially compelling. However, it is a powerful technology that is often poorly implemented, causing the best a false sense of security, and at worst severe risks to data availability. Finally, we may need to investigate the effect of transparency of security solutions on bioinformatics users' trust in and desire to use our system - to tread the fine line between security's visibility and intrusiveness.

ACKNOWLEDGMENTS

This work is supported by the myTea eScience Best Practice Project, mytea.ecs.soton.ac.uk, funded by the EPSRC.

REFERENCES

1. Aboud, Gregory. Common Features of Ubicomp Applications. ICSE99.
2. Jakob Bardram Olav W. Bertelsen Supporting the Development of Transparent Interaction Lecture Notes In Computer Science; Selected papers from the 5th International Conference on Human-Computer Interaction. 1015, (1995):79-90.
3. De Roure, D., Jennings, N., Shadbolt, N. Research Agenda for the Semantic Grid: A Future *e*Science Infrastructure, in *Grid Computing: Making the global infrastructure a reality*, Berman, F., Fox, G., and Hey, T. (eds), Wiley Europe, 2003, 437-470.
4. A. Dix, R. Beale and A. Wood (2000). Architectures to make Simple Visualisations using Simple Systems. Proceedings of Advanced Visual Interfaces - AVI2000, ACM Press, pp. 51-60.
5. Dourish, P., Edwards, W. K., LaMarca, A., & Salisbury, M. (1999). Presto: An experimental architecture for fluid interactive document spaces. ACM Transactions on Computer-Human Interaction.
6. Dumais, S., Cutrell, E., Cadiz, J. J., Jancke, G., Sarin, R., & Robbins, D. C. (2003). Human interaction: Stuff i've seen: A system for personal information retrieval and re-use. Paper presented at the SIGIR '03: 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, Canada.
7. Hughes, G., Mills, H., de Roure, D., Frey, J., Moreau, L., schraefel, m. c., Smith, G. and Zaluska, E. (2004) The semantic smart laboratory: a system for supporting the chemical eScientist. Organic and Biomolecular Chemistry 2:pp. 1-10.
8. Kaptelinin, V. (2003). Umea: Translating interaction histories into project contexts. Paper presented at the CHI '03, Ft. Lauderdale, Florida, USA.
9. Kurtenbach, G., Buxton, W. Issues in combining marking and direct manipulation techniques, In Proc. of UIST, 1991, pp.137-144
10. S. Pettifer, J. R. Sinnott, and T. K. Attwood. UTOPIA: user friendly tools for operating informatics applications. Comparative and Functional Genomics, 5:56-60, January 2004.
11. Dennis Quan, David Huynh, and David R. Karger. Haystack: A Platform for Authoring End User Semantic Web Applications in ISWC 2003.
12. Rekimoto, J. (1999). Time-machine computing: A time-centric approach for the information environment. Paper presented at the 12th annual ACM symposium on User interface software and technology, Asheville, North Carolina.
13. Ringel, M., Cutrell, E., Dumais, S., & Horvitz, E. (2003). Milestones in time: The value of landmarks in retrieving information from personal stores. Paper presented at the Interact 2003, Zurich.
14. 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
15. schraefel, m. c., Wigdor, D., Zhu, Y. and Modjeska, D. (2002) Hunter gatherer: within-web-page collection making. In Proceedings of CHI '02 extended abstracts on Human factors in computer systems, pages pp. 826-827.
16. Robert Stevens, Hannah J. Tipney, Chris Wroe, Tom Oinn, Martin Senger, Phil Lord, Carole Goble, Andy Brass, and May Tassabehji. Exploring Williams-Beuren Syndrome Using myGrid. Bioinformatics, 20:i303-i310, 2004.
17. Jun Zhao, Chris Wroe, Carole Goble, Robert Stevens, Dennis Quan, and Mark Greenwood. Using semantic web technologies for representing e-science provenance. In Proc. of the Third International Semantic Web Conference, Lecture Notes in Computer Science, pages 92 - 106, Hiroshima, Japan, 2004. Springer.