Semantic Laboratory Notebook: Managing Biomedical Research Notes and Experimental Data

Alexander Polonsky[1], Adrien Six[2,3], Mikhail Kotelnikov[1], Vadim Polonsky[1], Renaud Polly[1], Paul Brey[2]

[1] Cognium Systems SA, 15 rue Commines,
75003 Paris, France
{apolonsk, mikhail.kotelnikov, renaudpolly, vpolonsk}@cogniumsystems.com
[2] Institut Pasteur, 25-28 rue du Dr Roux,
75015 Paris, France
{askc, pbrey}@pasteur.fr
[3] Université Pierre et Marie Curie - Paris 6, 4 place Jussieu,
75005 Paris, France

Abstract. The main raw product of biomedical research is the information contained in laboratory notebooks and the associated computer files of individual researchers. Most of the problems in managing bioresearch information downstream stem from the way this information is initially recorded and stored. Electronic notebooks based on traditional knowledge management approaches have not been widely adopted by bio-researchers – the vast majority still use paper notebooks. We describe deployment of a software system based on the semantic tagging approach that successfully addresses the key adoption problems. This case study also indicates fruitful directions for the future R&D.

Keywords: Semantic Annotation, Semantic Tagging, Knowledge Articulation, Life Sciences.

1 Introduction

A recent article in Financial Times stated that "*R&D productivity* - not R&D investment - is the real challenge for global innovation" [1]. This is especially true for biomedical research, one of the largest global R&D sectors. Biomedical research is highly information intensive, and much of its information management aspects are inefficient due to a low degree of automation.

According to a study by Atrium Research, research chemists spend on average 2/3 of their time on information-intensive tasks such as meetings, literature analysis, writing papers and reports, and less than ¼ on conducting experiments [2]. Biomedical research generates at least as much information as chemical research, and hence we can expect that the work time distribution for biomedical researchers is skewed to at least the same degree. Indeed, getting the right information is critical for every step of biomedical research, from project planning to reporting the results.

Hence, automating its knowledge management aspects can result in a substantial productivity boost.

We have analyzed 40 articles from the relevant academic studies, analyst reports, and professional press to collect and prioritize the various knowledge management needs in biomedical research. We found that the adoption of structured Electronic Laboratory Notebook (ELN) systems is the key bottleneck to adequately addressing these needs. However, the ELNs based on the traditional knowledge management approaches force people to make a choice between 1) flexible but unstructured data entry or 2) rigid but useful organization. This is one of the main reasons why these systems have not been widely adopted by researchers [3]. As a result, a large amount of information discovered during research gets lost over time and is hard to retrieve, understand, analyze, and manipulate [4-7]. We have used a semantic tagging approach to develop a collaborative laboratory notebook software that allows both sufficient flexibility of data entry and thorough organization of the recorded information.

2 Knowledge Management Needs in Biomedical Research

We have randomly selected and analyzed 40 vendor-independent articles (including analyst reports, academic studies, and professional press) on the subject of knowledge management needs in biomedical research. The table below summarizes the results of the analysis. The needs in the right column were grouped into top-level categories located in the left column. The numbers in parentheses represent the number of articles that mentioned the corresponding needs.

Table 1. Literature analysis of KM needs in biomedical research.

Top-level needs with # of citing articles	Component needs with # of citing articles
Collective data management (29)	• search quality (recall and precision, multimedia) (11) • sharing experience, methods, data, analysis, resources (9) • simple and flexible data entry (5) • free access to scientific information (4) • partial (controlled) sharing (3) • information clarity (3) • real-time, persistent availability of information (2) • useful perspectives on information (2) • sharing knowledge organization methods (1)
Data storage (22)	• electronic, as opposed to paper (9) • storing all experimental data (including method details, full experimental history, negative results, "uninteresting" results, replications, unfinished projects) (8) • intellectual property protection (traceability, security) (8) • open and standard formats (5) • support for large data quantities and multimedia (3) • long-term archiving (2)

Data integration (21), across:	• databases and publication archives (8) • disciplines (biomedical research, chemistry, high-throughput screening, drug development, clinical/patient evaluations) (6) • individuals and groups within an organization (different departments, globally distributed sites) (4) • applications and websites (4) • organizations (subcontractors, partners) (3) • subfields (e.g., brain mapping, genomics, transcriptomics, proteomics, metabolomics) (3) • personal information (research notes, data files, emails) (1) • access rights levels (private, group, corporate, public) (1) • domain concepts (pharmaceutical compounds) (1) • business processes (1)
Personal data management (18) (a subset of Collective data management)	• search quality (recall and precision, multimedia) (11) • simple and flexible data entry (5) • information clarity (3) • useful perspectives on information (2)
Project management (18)	• quality of process and innovation (e.g., quality assurance, experimental design) (5) • work evaluation (work/contribution-based as opposed to publication-based, accountability) (4) • task management (efficient coordination, planning, and reliable implementation of a preset sequence of hierarchical tasks, e.g., protocol implementation) (2) • keeping up to date on a project (1) • resource sharing (cost, time, expertise) (1)
Automatic Analysis (12)	• inference rules, validation (compliance, safety checks, and other validity checks) (6) • consistency analysis (results, methods) (3) • decision support (2) • quantitative analysis (2) • interdisciplinary concept mapping (1) • discovery (e.g., new inter-object relationships) (1) • statistical bias analysis (1) • hints, autocompletion (1)
Communication (7)	• clarity (e.g., format consistency) (3) • automatic report and publication-draft generation (2) • multi-channel publishing (2) • open (review-independent) communication channels (1)

The above summary represents the needs as *perceived* by the domain analysts and the bioresearch community. The citation frequency indicates the degree to which a particular need is perceived, thereby giving a rough sense of the need's priority. However, most of the needs in the table are in fact inter-dependent. For example, better integration would lead to improved search which would in turn lead to improved project management and can indirectly improve data integration.

In order to derive the core user requirements in the domain, we propose to classify the expressed needs into 4 requirements categories: 1) constraints: properties that

must be present in a software solution; 2) simplicity or ease of use; 3) direct benefits from individual and collective use of the system; and 4) desired side-effect benefits. We can redistribute the needs from Table 1 according to the 4 categories as follows:

1) intellectual property protection;
2) simple and flexible data entry;
3) all the remaining needs from Table 1;
4) free access to scientific information.

The needs in the 3rd category would be best addressed via manipulations of structured data [8]. The classic approach is for the information to be entered into structured forms or templates thereby becoming much clearer to humans and more accessible to computer-aided operations, such as structure-based search, integration, and analysis. Although this approach has worked very well for certain kind of data (structured data), it has not worked well for all data (unstructured data). Indeed, a lot of information entered in a document format is difficult to input into a form. The same is true for information represented as a network, image or sound. Hence, the traditional approach of structured forms creates a conflict between the requirements categories 2 and 3 above.

This is the main reason why the vast majority of information in a bioresearch organization remains unstructured [9]. As we discuss in the next section, the requirements categories 1 and 4 also depend on the degree of information structure, and therefore, are also in a conflict with the 2nd category. Yet, due to the complex and unpredictable nature of research information, the category 2 is key for a successful adoption of an IT solution by researchers [3]. As we show below, a semantic approach can substantially diminish the conflict between these critical requirements.

2.1 The Key Role of Electronic Laboratory Notebooks

The vast majority of biomedical researchers store the raw information obtained in the course of their experimental work in paper laboratory notebooks and private computer files. Only a small fraction of this information remains during the transformation into scientific reports or publications, leading to a large information loss [4,5]. Hence, an ELN plays two key roles: 1) as the first point of information entry, and 2) as a comprehensive repository of research information, where both research notes and associated electronic files can be stored.

The 1st role is important since it is simpler and more efficient to organize information at the time of its entry than afterwards. A semantic structure created at the stage of note-taking can be propagated to the subsequent stages of processing, such as reports or publications.

3 iPad: Semantic Laboratory Notebook for Biomedical Research

The most natural and straightforward way to represent research notes is using the document representation (that is how they are currently recorded). Imbedded in the

note documents can be other data formats such images, video, tables, forms, and network data. Hence, what is needed ideally is an integrated environment for structuring and working with structured information that would address the specific needs of all these numerous data types and would allow to view the same information in different representations (e.g., in a tabular format or in a document format). The ELN system we have developed so far, named iPad, uses the semantic tagging approach to allow people to easily structure and work with structured documents.

3.1 Semantic Tagging Approach

Electronic forms have carried over into the electronic environment many constraints associated with their predecessors, the paper-based forms. For example, it is not possible to enter information in between form fields, copy, delete, or move several form fields at a time, add a form field inside another form field, etc. As a result, electronic forms are often not well-suited for structuring complex hierarchical information such as research notes. However, the constraints of the forms are purely historical, they are not required to give information a semantic structure.

The semantic tagging approach is based on the inverse paradigm: as opposed to forcing information into a given form structure, information can be recorded in a free document format and then labeled with the corresponding semantic tags. This approach allows to combine flexibility with structural organization during document authoring.

3.2 iPad Overview

The system is based on a three-tier software architecture comprised of the client applications (standalone iPad Editor and iPad Web Portal), iPad middle-layer Server, and a database. The information is entered via iPad Editor (Fig. 1) and can be viewed either in the Editor or the Portal. It can be stored either in the database or on the users' computers (although, in the latter case, the functionality is quite limited to encourage central storage). The middle layer mediates the transfer of information between the client applications and the database. Different middle-layer adapters allow connecting the system to any database, although currently only the relational database adapter has been developed for connecting to any major relational database (the default is MySQL). The Editor and the Server are implemented in Java (JDK 1.4.2).

Multimedia information (formatted text, tables, images, etc) is entered into the Document Editor (Fig. 1) using traditional document editing functionalities. External files can be attached to the document by drag-and-drop and are displayed as hyperlinks.

The Tag List proposes relevant semantic tags (e.g., project, result, method) depending on the cursor's position within the document. At this time, the proposed tags model the organizational concepts of biomedical research projects and not the discovered biological knowledge (a much more complex problem). For example, the tags are used to clearly mark which method was used to obtain a given result as a part of what experiment or project. iPad also offers a free tag mode in which users can

define their own tags. However, these modes are kept separately in order to avoid confusion between the preset tags proposed by the system and those created by users. The free tag mode can be used for ad-hoc organizational needs that were not taken into account by the preset tags.

Only semantically valid tags are proposed and they can be inserted at any place in a document where they are valid. Each tag has a set of attributes that can be filled out in a pop-up form. The possible tags and their attributes are specified in external XML Schema documents and are visualized within the document as specified in external XML-based templates.

The subsequent document structure appears in the Document Browser window. The structure is a hierarchy of inserted tags and can be used to browse the document by clicking on the tag of interest and viewing the corresponding document section in the Document Editor. In addition to the hierarchical relationships, tags (from the same or different documents) can be interlinked with hyperlinks.

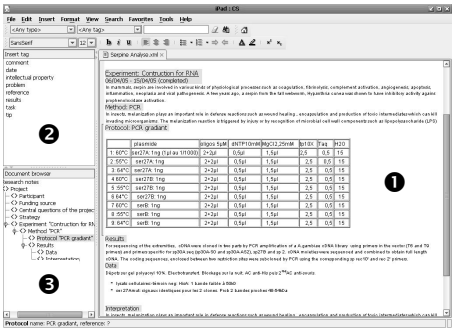

Fig. 1. iPad Editor: (1) Document Editor, (2) Tag List, (3) Document Browser.

Once the information is entered in a structured way, iPad gives the user a large set of functionalities to benefit from the resultant structure, including structure-based browsing and search (with user-friendly interface), information perspectives (also called "semantic lenses" [10]), automatic report and publication draft generation (using mapping between XML Schemas), and ability to visualize the information in a variety of ways on iPad Web Portal (using XSLT).

It turns out that the ability to structure documents also addresses the issue of intellectual property protection (requirements category 1). Parts of documents that contain sensitive information can be specially tagged and processed appropriately (e.g., printed and signed). This substantially decreases the amount of work since only a small part of the electronic information has to be processed in this way.

In addition to the valuable functionality, formal and guided document structure improves information clarity. This has a side-effect benefit: research notes could eventually be shared freely on the Web since they would be sufficiently structured to be understood by other scientists and to retrieved using structure-based search [8] (requirements category 4).

4 Case Study at Institut Pasteur

iPad has been developed in collaboration with Institut Pasteur (Paris, France), a world-renowned biomedical research Institute. It has been used for over a year by a research group of 4 people as well as 3 individual researchers at the Institute. The number of users is gradually expanding.

Although we have not yet conducted a formal evaluation, user feedback has been positive and confirms our assumptions. Users have noted the improvements in 1) information clarity (both within their own and others' notes), 2) research quality due to useful perspectives on their work, 3) report and publication writing, 4) information retrieval, 5) information sharing, and 6) integration.

We have also confirmed our view that the semantic tagging paradigm is not straightforward for users from the beginning and requires a tutorial. The User Interface is the key factor determining the learning curve. Despite being new, the tagging paradigm as implemented in iPad becomes sufficiently intuitive after a couple hours of training and practice.

The system has been used by several individual researchers independently from a research group, showing that it provides benefits that are independent from collective utilization. This is important for its adoption since it avoids the prisoner's dilemma issues commonly present in collaborative systems.

5 Future Work

Directions for future development are numerous:

- Migration from the read-only Web Portal to a Web-based editing environment (a Semantic Wiki) to improve information availability
- Integration of an environment for structuring and working with network information (i.e., biological processes)
- Migration from XML to RDF in order to better support semantic relationships
- Adoption of the peer-to-peer paradigm to increase collaborative flexibility
- Integration with linguistic algorithms for semi-automatic tagging
- Adoption of falksonomy techniques for constructing dynamic collaborative ontologies of biomedical concepts and using this ontology to achieve a greater degree of information structure.

In addition to the technical development, we plan to complete a formal evaluation of iPad in both academic and industrial research settings.

Although iPad's functionality has been focused on the needs of biomedical scientists, due to the generality of iPad's approach and architecture, it can add value in domains other than biomedical research. For example, iPad has already been used to structure and manage generic (non-research) project notes. More work needs be done to evaluate the scope of iPad's applicability.

Acknowledgements. This work was in part conducted within NEPOMUK project supported by IST FP6 grant from the European Community.

References

1. Schrage, M.: For innovation success, do not follow where the money goes. Financial Times (2005)
2. Michael, E.H.: It's Not About the Paper. Scientific Computing & Instrumentation (2004)
3. Michael, E.H.: Electronic Study Management. Scientific Computing (2006)
4. Butler, D.: A new leaf. Nature Vol 436 (2005)
5. Knight, J.: Null and void. Nature Vol 422 (2003)
6. Phillips, M.L.: Do you need an electronic lab notebook. The Scientist (2006)
7. Sarini, M., Blanzieri, E., Giorgini, P., Moser, C.: From actions to suggestions: supporting the work of biologists through laboratory notebooks. Proceedings of the 6th International Conference on the Design of Cooperative Systems (2004)
8. Berners-Lee, T., Hendler, J.: Scientific publishing on the 'semantic web', Nature Web Debates (2001)
9. Building Blocks of an Enterprise Content Management Business Case for Life Sciences. First Consulting Group (2004)
10. Neumann, E.: A Life Science Semantic Web: Are We There Yet? Science Vol 2005, Issue 283 (2005)