Effectiveness of tagging laboratory data using Dublin Core in an electronic scientific notebook

Laura M. Bartolo[1], Cathy S. Lowe[2], Austin C. Melton[3,4], Monica Strah[5], Louis Feng[3], Christopher J. Woolverton[5]

[1]College of Arts & Sciences, [2] School of Library and Information Sciences, [3]Department of Computer Science, [4]Department of Mathematics, [5]Department of Biological Sciences, Kent State University, Kent, Ohio, USA

Summary

As a form of grey literature, scientific laboratory notebooks are intended to meet two broad functions: to record daily in-house activities as well as to manage research results. A major goal of this scientific electronic notebook project is to provide high quality resource discovery and retrieval capabilities for primary data objects produced in a multidisciplinary, biotechnology research laboratory study. This paper discusses a prototype modified relational database that incorporates Dublin Core metadata to organize and describe the laboratory data early in the scientific process. The study investigates the effectiveness of this approach to support daily in-house tasks as well as to capture, integrate, and exchange research results.

1 Introduction

Interdisciplinary scientific research efforts in academic, industrial and public settings need novel technologies to organize, access, and communicate research results, especially for a mobile society. This paper reports on one stage of a multi-stage project to construct an electronic scientific notebook for recording, storing, and manipulating multidisciplinary and multi-institutional scientific information from raw data to finished research papers. Long-term goals of this project include: 1) learning how to organize and store biotechnology information in formats which will encourage multidisciplinary use of the information; 2) applying the organizing knowledge gained and tools developed in storing biotechnology information to the storage of other scientific information; 3) developing an environment in which scientific information from different disciplines can be made more easily accessible by and meaningful to multidisciplinary research teams; and 4) constructing electronic scientific notebooks for the storage, retrieval, and dissemination of multidisciplinary scientific information.

Grey literature, such as data generated within scientific laboratories, has been recognized as an important area for innovation, new knowledge, and industrial enterprise (Jeffery 2000). In order to provide electronic access to laboratory data and to replace the standard paper notebook which scientists currently use in their laboratories, an electronic notebook needs to be able to perform all the functions of paper notebooks as well as facilitate greater operational flexibility. This paper presents a practical application of the software architecture utilizing laboratory data from an ongoing multidisciplinary, multi-organizational research program. The discovery process requires the recording of ideas, the identification of individual efforts, data acquisition, data analysis and presentation of data in scientific, lay and summary outputs (Buckland 1997). The investigation presented here used the traditional laboratory notebooks of the principal investigator to demonstrate the utility of the electronic architecture and Dublin Core (DC) in a scientific laboratory setting. Long term goals of this research include interfacing scientific notebooks with large data-

bases, facilitating the exchange of data in scientific notebooks among researchers, and making scientific notebooks central tools for presentations of results gained from laboratory research.

2 Interdisciplinary research and user needs

Scientific discovery can result in various final products such as published manuscripts, monographs, and patents. Our primary goal has been to create a software architecture that facilitates the discovery process across institutions and between investigators by creating an environment that mirrors the traditional laboratory notebook. Expanded features seek to enhance user access, data sharing and data mining while reducing or eliminating data integrity issues and redundant functions. A second goal of the project is focused on tagging data utilizing Dublin Core metadata so as to link individual data, analogous data types and non-chronological data entries across users and institutions. Tagging data by such elements as date, format, creator, relation, and rights is useful, if not necessary, to enable the retrieval of laboratory data. Data acquisition and analysis are facilitated and shared more readily electronically. Furthermore, data sharing enhances the team approach resulting in better quality control through enhanced data integrity and better data analysis.

The software architecture seeks to integrate all the features necessary to support and augment the scientific discovery process. We demonstrate the utility of the software with a current biotechnology example uniting the disciplines of microbiology, chemical physics, and medicine to promote basic research as well as to develop new diagnostic tools. This interdisciplinary, multi-institutional research team, comprised of research faculty and personnel from the Department of Biological Sciences and the Liquid Crystal Institute at Kent State University, the Northeastern Ohio Universities College of Medicine, and Summa City Hospital, invented a novel microbial biosensor. Their needs for data acquisition, analysis and sharing are numerous. The team needs to collectively conceive new ideas, prevent redundant experimentation, swap results and write manuscripts while performing other daily activities in different physical locations. Thus, each team member needs to identify and view raw data, summary data, photographs, reactions, analog and digital equipment output, manuscripts, and visual presentations produced by other team members. The linear retrieval format of the traditional, chronological entry lab notebook hinders efficient data sharing between the team members. Optimal data sharing begged for an electronic database system that tagged data in multiple ways beyond the time/date format. Furthermore, discipline-based investigators tend to be restricted by domain-specific database systems that do not readily permit identification of relevant literature across multiple disciplines. The software we present uses Dublin Core as the common data cataloging format by which multidisciplinary, multi-institutional research teams can describe, identify and exchange relevant data in early stages of the scientific process.

3 Advantages of Dublin Core with Laboratory Data

A number of characteristics associated with DC make it a promising choice for describing scientific laboratory data (Duval et al. 2002, Weibel 1995, Weibel 2000). DC is relatively concise, simple, and easy to learn, increasing the likelihood that busy scientists, technicians, and students have the time to create and maintain metadata records for laboratory data as it is generated (Baker 2000, Greenberg et al. 2001). DC supports multiple formats including text, still images, video, audio, and datasets generated within scientific labs. Because DC is designed to facilitate Internet resource discovery, DC facilitates making scientific data rapidly and widely available at appropriate times. Plans for a DC metadata registry promise to ensure consistency of its application. Attention to international concerns increases the likelihood that DC will be used worldwide. While the current paper focuses solely on using only DC elements, in the future, possible

extensions of DC will be integral to this project to address the multidisciplinary, multi-institutional scientific research (Forsberg 2000). Elements from other metadata schemas will be added to enhance desired functions (e.g., administration) or provide more specific information for certain groups, e.g., interdisciplinary team members representing different domains. Finally, the American National Standards Institute in conjunction with the National Information Standards Organization approved the Dublin Core Metadata Element Set as a national standard (ANSI/ NISO Z39.85-2001) on September 10, 2001 (NISO 2002). This approval and efforts toward the acceptance of DC as an international standard are expected to strengthen confidence in the value of DC as a tool for improving resource discovery and information exchange leading to increased use of the metadata standard in the U.S. and abroad (Dekkers and Weibel 2002).

4 Database Design of the Scientific Notebook

A modified relational database has been constructed which provides all of the information typically incorporated in a print scientific laboratory notebook including raw data collected from numerous experiments; intermediate documents, such as procedures, memos, and technical documentation; and final products, such as completed research papers and multi-media presentations. The information that is most central to a scientific laboratory notebook is contained in the Main Notebook section of the database. The Main Notebook includes a number of tables which capture general descriptions of past, present and planned projects (Topic); experiment design concepts specific to a given project (Experiment Goals); procedures and materials used in actual experiments (Materials & Methods); and specific procedural components (Steps). Results would include drafts and finished papers (Topic Results), data tables and graphs (Experiment Results), or images and datasets (Materials & Methods Results). This section is organized hierarchically (see Figure 1). Higher level tables may be associated with an unlimited number of lower level table items (e.g., each Topic may be associated with an unlimited number of Topic Results and an unlimited number of Experiment Goals. An entry form has been created to facilitate recording of laboratory notebook data and to test the initial design of the Main Notebook. It is estimated that approximately 200 entries will be available to be described by DC from the paper laboratory notebook, including 1 Topic, 8 Experiment Goals, 61 Materials & Methods, and 127 Materials & Methods Results. Supplementary Tables support the scientific investigation by archiving additional information related to the Main Notebook that support the scientific investigation (See Figure 2). Dublin Core Records are being associated with items contained in each table of the Main Notebook except for Steps. See Appendix 2 for an example of a DC record associated with a Materials & Methods Results entry. Materials collects detailed information such as MSDS (material safety data sheets), Specification Sheets, and Materials Lot Analyses about organisms and liquid crystal substances involved with experiments. User includes basic contact information about individual researchers involved with the scientific investigation and specifies authentication and access rights. Memos includes entities such as correspondence, equipment issues, and notes for future experiments.

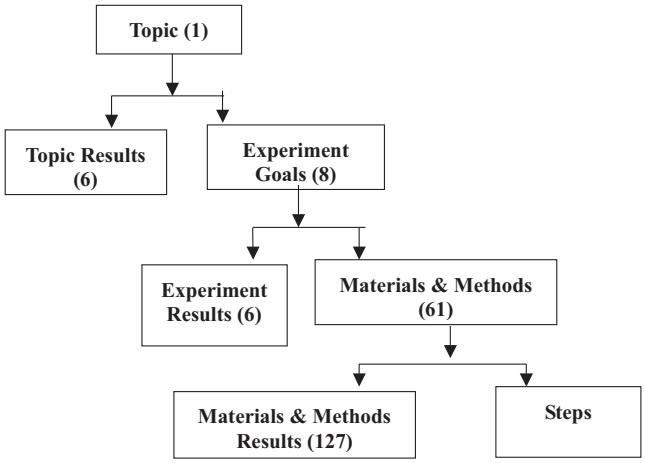


Figure 1: Representation of hierarchical database design of the Main Notebook and estimated number of entries available for Dublin Core description in parentheses

All tables within the Main Notebook will be associated with Dublin Core records except for Steps.

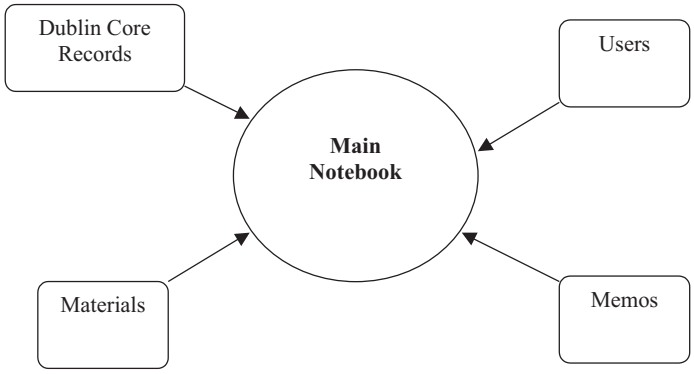


Figure 2: Representation of Main Notebook and Supplementary Tables

Supplementary tables are linked to the Main Notebook to provide additional information about the data.

5 Methodology

DC is being applied to a database representation of the laboratory data recorded in a paper notebook for a single project by one investigator over the course of one year. Microbiological experimental design, implementation, and result components are being described using DC. A content analysis will be used to investigate whether the DC schema is a feasible method of describing laboratory data. An expert user study will investigate the utility of DC for data description. Ex-

amples of the DC records with their associated laboratory data and results from the feasibility testing will be presented at the CRIS 2002 conference.

5.1 Feasibility Testing

During the DC application phase, any difficulties encountered in describing different types or aspects of the data are being noted (see Appendix 1). After application is completed and noted difficulties are addressed, if possible, a computerized content analysis (Murray 1998) of the DC records will be conducted in order to determine the frequency with which each element is used overall as well as for different types of information objects. Since the sample used in this study is much smaller and more heterogeneous than the sample used by Murray, much less content variability may be obtained for a given element. If results show that an element is used consistently but its content seldom or never varies, only unique instances of the element will be included in the frequency count in order to avoid artificially inflating results. The DC records will also be visually analyzed to identify any nonstandard DC usages. These measures provide an indication of the ease with which the DC element set can be applied to laboratory data as information objects. If most DC elements are used in accordance with established standards, then a good fit between the metadata schema and the data being described exists. If many unusable elements or inappropriate element assignments are identified, extensions or alternatives to the DC element set will be considered.

Following the content analysis, the functionality of the DC element set will be examined by using four specific types of metadata classes that provide information about digital objects. The four metadata classes are: discovery, use, authentication and administration (Greenberg 2001). Metadata classes group element level metadata by the function(s) each element supports. For example, the DC elements *creator* and *subject* both contribute to the function of discovery. In applying Greenberg's definitions of the classes to the laboratory environment, it is assumed that discovery supports research needs of users while use, authentication, and administration support functional needs of the workplace. The DC schema's ability to sustain required information functions will be assessed by aggregating the content analysis element frequencies for each class. Information regarding average frequency of use for each metadata class will be compared across data types to identify any marked differences in DC's effectiveness regarding specific functions among data types. Results of the feasibility testing will be presented at the CRIS 2002 conference.

5.2 Utility Testing

After feasibility has been determined, the next stage of the investigation will be to conduct usability tests with experts to assess the effectiveness of the prototype database and DC in handling laboratory routines as well as scientific research. User subjects will comprise in-house researchers from a Kent State University biotechnology lab, including graduate assistants, technical assistants, and scientists. The subjects will query the database using two separate interfaces to complete predetermined tasks, representative of both laboratory workplace needs and scientific research needs. One interface will access only the data and the second interface will access the DC enhanced data. The tasks will correspond to discovery, one of Greenberg's four metadata classes. Database transaction logs, subject interviews, and analyses of result sets will be compared for the discovery metadata class to record user evaluations of DC's effectiveness in handling laboratory data.

Future studies will explore the utility of the prototype database and DC with scientific laboratory data through additional user studies. One type of user study will include researchers from other laboratories and in other disciplines. Such research will focus on the multidisciplinary, multi-institutional, collaborative components of the biotechnology project. The second type of

user study will extend the utility study to include Greenberg's remaining metadata classes: use, administration, and authentication. Such an investigation would provide quantitative analysis of DC's functionality in areas outside of discovery.

6 Discussion

This project is not limited to the scientific community but also seeks to address the growing need in all segments of our mobile society for tools where people can transform information into usable knowledge and share this knowledge effectively. The prototype scientific notebook would support distributed work environments, tying together people from multiple organizations to collaborate on complex projects from different locations and at different times.

Future developments for the scientific notebook project would enable a scientist using his or her notebook as an interface to upload data to national and international data repositories or to search literature databases relevant to his or her research questions. Whenever the researcher would wish to review results, compare them to the results of others, and read relevant, similar experiments and studies, he or she would use the notebook to access this information. In addition, the scientist would be able to use the notebook to interface with stored results in order to prepare papers and presentations for scientific journals and to give presentations at conferences. Further, in time, the scientist would be able to set up and run an experiment, using the electronic notebook to control a virtual laboratory. Such capability would allow for the refinement of already running wet lab experiments and the efficient planning of future wet lab experiments.

Though the scientific arena is a specialized one, the ideas, methods, and tools used to gather, organize, and present scientific information can also be used in other settings. In fact, what makes the ideas presented in this paper especially appealing is that the gathering, organizing, and presenting are done in a multidiscipline setting. Thus, these methods and ideas are potentially useful in almost any situation where data need to be gathered, assimilated, and presented or used.

7 References

Baker, T., 2000. A grammar of Dublin core. *D-Lib Magazine*. 6 (10). [On-line]. Available at: http://www.dlib.org/dlib/october00/baker/10baker.html

Buckland, M.K. 1997. What is a "document"? *Journal of the American Society for Information Science*, 48 (9): 804-809.

Dekkers, M. and Weibel, S. 2002. Dublin Core Metadata Initiative Progress Report and Workplan for 2002. *D-Lib Magazine*, Volume 8 (2) [On-line]. Available at: http://www.dlib.org/dlib/february02/weibel/02weibel.html

Duval, E. Hodgins, W. Sutton, S. Weibel, S.L. 2002. Metadata Principles and Practicalities. *D-Lib Magazine*, 8 (2) [On-line]. Available at: http://www.dlib.org/dlib/april02/weibel/04weibel.html

Forsberg, K., 2000. Extensible use of RDF in a business context. *Computer Networks*, 33:347-364.

Greenberg, J. 2001. A quantitative categorical analysis of metadata elements in image-applicable metadata schemas. *Journal of the American Society for Information Science*. 52 (11): 917-924.

Greenberg, J., Pattuelli, M., Parsia, B. Robertson, W. 2001. *Author-generated Dublin Core Metadata for Web Resources: A Baseline Study in an Organization*, Journal of Digital Information, 2 (2) [On-line]. Available at: http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Greenberg

Jeffery, K. G. 2000 An architecture for grey literature in a R&D context. *The International Journal on Grey Literature*. 1 (2): 64-72.

Murray, K. 1998. CIMI DC Simple Testbed *Record Content Analysis*. [On-line]. Available: http://www.cimi.org/old_site/documents/CIMI_DC_Simple_RCA.html; accessed January 25, 2002.

National Information Standards Organization (NISO). 2002. NISO standard: Dublin Core Metadata Element Set (ANSI/NISO Z39.85-2001). [On-line]. Available: http://www.niso.org/standards/standard_detail.cfm?std_id=725

Weibel, S. 1995. Metadata: The foundations of resource description. *D-Lib Magazine*. 6 (07). [On-line]. Available at: http://www.dlib.org/dlib/July95/07weibel.html
Weibel, S. 2000. The Dublin core metadata initiative: Mission, current activities, and future directions. *D-Lib Magazine*. 6 (12). Available at: http://www.dlib.org/dlib/december00/weibel/12weibel.html

8 Contact Information

Laura Bartolo
Kent State University
College of Arts & Sciences
Kent, Ohio 44242-0001
USA

e-mail: lbartolo@kent.edu

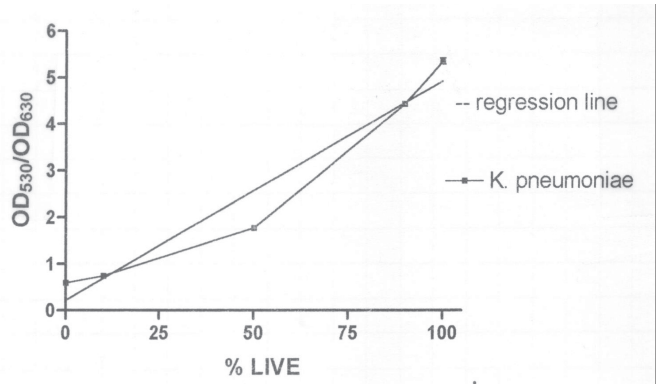
Appendix

APPENDIX: 1 Problems with DC Application

Description of Difficulty	*Item Identification*
	*Area: **Info Object: Record #:
	Area: Info Object: Record #:
	Area: Info Object: Record #:
	Area: Info Object: Record #:
	Area: Info Object: Record #:
	Area: Info Object: Record #:

*Area: Goal, Procedure, Goal Result, Procedure Result
**Info Object: Text, Table, Graph, Image, Dataset, Video, Audio

APPENDIX: 2 Example of Laboratory Data and Associated Dublin Core Record



Title = "Bacterial Toxicity Assay of CPCl treated Klebsiella pneumoniae"

Creator = "Woolverton, Christopher J."

Subject= "Bacterial Toxins—analysis" (MeSH)

Subject= "Klebsiella Infections—immunology" (MeSH)

Subject= "Klebsiella pneumoniae" (MeSH)

Description = "Graph of Bacterial Toxicity Assay of CPCl treated Klebsiella pneumoniae.
% live standard curve used to evaluate CPCl effects."

Date = "2000-09-06"

Type = "image"

Format = "image/jpg 183 KB"

Identifier = "CJW2_043001.JPG"

Identifier = "Materials and Methods Result #39"

Language = "en-us"

Relation = "IsPartOf Materials and Methods #18"