

The Data Collective: A New Model for the Ownership and Use of Scientific Data

Timothy Koschmann
Southern Illinois University
School of Medicine

Traditionally, the ownership of scientific data has resided within an isolated laboratory. The laboratory itself could range from the working space of a single researcher to quite elaborate institutional structures involving many researchers, post-docs, and graduate students. Regardless, the traditional model has been one in which access to the data upon which scientific publication is based is locally controlled and tightly regulated. In most cases, the only public access to the data upon which research is based is provided in the form of published summaries and distillations. Pressures to change this model have come from a variety of quarters. Federal funding agencies have an interest in seeing data acquired at large public expense be put to maximal use. Also, various scandals involving the falsification of data in isolated laboratories has suggested the need for the development of more open data verification mechanisms. Finally, but perhaps most importantly, the notion of proprietary control of scientific data seems to clash with the view of science as an open and collaborative enterprise. As a result, a number of alternatives to the traditional model of lab-centered data ownership have emerged over time.

In certain disciplines such as linguistics there is a tradition of creating shared *data corpora*. Examples would include the London-Lund data corpus in communication studies, the CHILDES database in developmental psychology (MacWhinney, 1995), and the TIMSS video data in instructional science. All have international scope, both in terms of content and in use. The model here is one of collaborative effort to create a shared repository of data which then becomes a resource for multiple researchers and research groups. This is an efficient model for producing research and one that maximizes the use of the data. It is also a model that facilitates critical discussion within a field of inquiry, since the data constitutes an object for shared attention and discussion. But the model has certain built in limitations. Shared data corpora tend to be designed for use within a particular discipline and, once constructed, tend to be largely static in structure.

The National Science Foundation in the U.S. has been instrumental in promoting another model of data sharing known as the *collaboratory*. This is a more general model for sharing scarce resources among multiple research groups and facilitating coordination among these research groups. The resources may be data, apparatus, or facilities such as observatories and small-particle accelerators. One example would be NCSA, which was initially created to provide

shared access to a number of super computers. Research groups making use of a collaboratory need not be co-located and researchers from different disciplines may contribute to the work of the collaboratory. The existence of a collaboratory does not necessarily require shared ownership of scientific data, however. Furthermore, though the collaboratory model does require coordination across research groups, it does not necessarily entail collaboration among them.

We are working on developing yet another model for the ownership and use of scientific data. It is a model we term the *data collective*. Unlike data corpora designed to provide shared access to basic data for researchers within a discipline, we propose a model that involves sharing of both basic data and intermediate findings among researchers from different disciplinary traditions. A prototype of this sort of collaboration can be seen in the Professional Competency Project. This project was initiated to foster multidisciplinary research on clinical problem-solving and the assessment of medical competency. The medical licensing authorities in the U.S. and Canada have implemented (or are in the process of implementing) performance-based assessment for all licensure candidates. Such forms of assessment involve working up clinical cases presented by “standardized patients” (SPs), that is actresses or actors trained to portray a testing case. The National Board of Medical Examiners (NBME), for example, has announced its intention to add a test of clinical skills using SPs to the United States Medical Licensing Examination (USMLE) by the fall of 2004 (NBME, n.d.). Standardized patients have been used as a part of the certification process employed by the Educational Commission for Foreign Medical Graduates since 1998 (ECFMG, 2002) and such tests have been used in the licensure exams conducted by the Medical Council of Canada (MCC, 2002) for over a decade. Despite its importance in the training, licensure, and certification of medical practitioners, performance based assessment has been little studied outside of the institutions within which it is used because of the difficulty and expense involved in recruiting subjects, the lack of access to adequately-equipped testing facilities, and the unavailability of useable testing cases. The Professional Competency Project involves producing a corpus of performance-based assessment protocols for study by a diverse collection of social scientists.

We are in the process of developing a set of testing cases based on real clinical data. Research subjects will be undergraduate medical students and residents enrolled at a particular medical school. All subjects will be asked to take a history and perform a physical exam on SPs trained to present the testing cases in a special testing facility known as the Professional Development Laboratory (PDL). The PDL consists of a suite of rooms equipped as clinical examination rooms. Each examination room in the PDL contains all of the paraphernalia (e.g., exam table, otoscope, sphygmomanometer, eye chart) normally used in conducting a physical examination. After working up the patient, subjects retire to an adjacent computer lab and are given 30 min to compose a SOAP note and order lab tests. A SOAP note is a structured chart

entry consisting of the subjective data (i.e., the patient's reported symptoms), objective data (i.e., physical findings and lab results), assessment (i.e., a differential diagnosis), and a treatment plan. After completing their SOAP note and lab orders, the subjects were taken to an interview room in which a structured debriefing is conducted. The results of the laboratory tests ordered by the subject are presented to the subjects in the context of this interview. The videotapes from the encounters with the SP and the debriefing interviews are transcribed by a medical transcriptionist. These protocols will be studied by a group of researchers at different institutions to address different research questions.

One foundational issue pertains to the forms of knowledge subjects mobilize and employ in clinical problem solving. One approach to addressing this issue is to conduct a "cognitive discourse analysis" (Frederiksen, 2001). Cognitive discourse analysis (CDA) begins by constructing models of experts' procedural and declarative knowledge in particular domains of problem solving. The propositional content of subjects' discourse contributions can then be mapped with reference to the expert task model so constructed as a means of diagnosing possible deficiencies in the subjects' preparation or understanding.

A host of related questions can be raised with respect to the ways in which subjects reason through diagnostic problems. How do they initially come to formulate the problem, for instance? How do they amend their understanding of the patient's problem as new information is developed? How is conflicting data dealt with? Such questions are related to the way in which the subject organizes the problem in memory. An analysis of memory-based reasoning can reveal past principles and experiences stored in memory, circumstances providing access to information, competing solutions, problem features used to test whether information is applicable, and support for the solution selected by examining subjects' problem-solving protocols (Seifert, Patalano, Hammond, & Converse, 1997).

Performance-based assessment requires that the subject and SP work together to produce a 'case.' Similarly, the purpose of the post-encounter interview is to reveal the subject's reasoning, but, again, the thing that the participants orient to (the 'case') is something produced interactionally. One might ask what methods do these members (subject, SP, interviewer) employ in carrying out this interactional task? The discipline that has taken up such matters as its central area of concern is Ethnomethodology (Garfinkel, 2002; Heritage, 1984). Conversation Analysis (CA) is an area of specialization within ethnomethodological research that focuses on the methods members use in interaction to make sense of each other and to, in turn, be seen and heard as sensible (Psathas, 1995; Heritage, 1984). Cicourel (1975) studied how interaction between a doctor and patient is transformed in producing a chart note. The corpus produced in this project constitutes a unique set of materials for studying the same process, one in which the 'case' is not only reproduced in the chart note, but also in an elaborate post-encounter interview.

The encounter between the SP and the subject results in two records, the subject's SOAP notes and the debriefing interview. Recent developments in computational linguistics have provided ways to statistically represent the semantics of these records allowing for a comparison of content. Latent semantic analysis (LSA) is one such method. For instance, these records can be compared with a gold-standard provided by an expert. Such a comparison has been successfully used in automated essay grading, where a student essay is compared with an expert essay in order to measure the student's performance (Landauer, Foltz, & Laham, 1998). LSA has generally been used for natural texts, however. Whether it can equally successfully applied to structured set of chart notes and dialog remains an important question to be investigated.

The four types of analyses (cognitive discourse analysis, analysis of memory-based reasoning, conversation analysis of doctor/patient interaction, latent semantic analysis) arise out of specific interests and address particular questions. They are not completely independent, however, and in many cases individual researchers can benefit from and build upon the analytic findings of others. Cognitive discourse analyses and analyses of memory-based reasoning, for instance, have considerable overlap in scope. The task model for a particular case developed to support cognitive discourse analysis would also be of great value for conducting analyses of memory-based reasoning. A detailed analysis of the reasoning done by a subject for a particular case would doubtless also be useful to an analyst doing a cognitive discourse analysis of the debriefing interview for that subject. Enhancements to the working transcripts produced by the CA researchers may prove to be useful to researchers doing analyses of the propositional content of dialog or memory-based reasoning. The more global measures of comprehension generated using LSA will be of value for all researchers in identifying particular assessment protocols in which problems in understanding were evidenced. The goal, therefore, is to go beyond the forms of sharing of "final research data" mandated by federal funding agencies (e.g., NIH, 2003) to include findings and other improvements produced by collaborating researchers. A number of questions arise, however, with respect to the construction of a data collective such as the one described here. When the research involves human subjects (as it does here), what are the ramifications of the data collective model with regard to confidentiality and informed consent? Also, what kinds of institutional arrangements need to be put into place to protect the interests of the researchers who contribute their work to the project? Finally, what are the requirements from a technological perspective for supporting this form of collaboration?

References

- Cicourel, A. (1975). Discourse and text: Cognitive and linguistic processes in studies of social structure. *Versus*, 12(11), 33–84.
- ECFMG (2002). *Clinical skills assessment: Candidate orientation manual*. Philadelphia, PA: Educational Commission for Foreign Medical Graduates. Retrieved August 28, 2002 as <http://www.ecfm.org/csa/csacom.pdf>
- Frederiksen, C.H. (2001). Propositional representation in cognitive psychology. In N.J. Smalser & P.B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences*. Amsterdam: Elsevier Science.
- FSMB (n.d.). Position of the Federation of State Medical Boards in support of adding a clinical skills examination using standardized patients to the United States Medical Licensing Examination (USMLE). Retrieved August 29, 2002 as http://www.fsmb.org/Policy%20Documents%20and%20White%20Papers/standardized_patient_support_white_paper.htm
- Garfinkel, H. (2002). *Ethnomethodology's program: Working out Durkheim's aphorism*. Lanham, MD: Rowman & Littlefield Publishers.
- Heritage, J. (1984). *Garfinkel and Ethnomethodology*. Cambridge: Polity Press.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk* (2nd Ed). Hillsdale, NJ: Lawrence Erlbaum Associates.
- NBME (n.d.). *USMLE clinical skills examination*. Retrieved August 28, 2002 from <http://www.usmle.org/news/newscse.htm>
- NIH (2003). NIH data sharing policy and implementation guidance. Bethesda, MD: Office of Extramural Research, National Institutes of Health. Retrieved May 30, 2003 as: http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm
- Psathas, G. (1995). *Conversation analysis: The study of talk-in-interaction*. Thousand Oaks, CA: Sage.
- Seifert, C. M., Patalano, A. L., Hammond, K. J., & Converse, T. M. (1997). Experience and expertise: The role of memory in planning for opportunities. In P. J. Feltovich, K. M. Ford & R. R. Hoffman (Eds.), *Expertise in Context: Human and Machine* (pp. 101 - 123). Menlo Park, CA: AAAI Press/ MIT Press.