Universally accessible databases in the advancement of knowledge from psychological research[1]

David H. Johnson
Federation of Behavioral, Psychological and Cognitive Sciences, Washington DC, USA

Michel E. Sabourin
University of Montreal, Montreal, Canada

There is little data sharing in psychology, although many "private" data sets exist whose full capability to yield useful knowledge has not been realized. Their potential to become members of families of data that, together, could contribute knowledge that none could contribute alone is, likewise, unrealized. It is the tradition of our science to rely on the published word to announce the knowledge we have gleaned from our data. It is equally traditional to raise doubt about the reported findings of others. We have believed such critical review is the means to separate the true from the erstwhile knowledge of our field.

But these traditions grew out of a time in research, and in the development of our field, when the ability to communicate was limited by things like distance and publication lags. Scientific meetings presented a rare opportunity for those in a specialty area to talk to each other face to face and to learn about work that had not yet reached the journals. We do not live in that era of separation anymore.

Our era is one in which physical separation hardly matters. Should we choose to do so, we could place the data from our research in the hands of a colleague on the other side of the world with the touch of a button or the click of a mouse. Data sets that would have taken months to analyze in the earlier era can be analyzed as they are being collected. And, arguably, methods by which to validly compare discontinuous data are being developed apace. The era that generated our present traditions was reductionist. That is, we subdivided our work into ever narrower specialty areas. In our era, we have the means to take the knowledge being produced in each specialty area and begin to put the bits of knowledge together. We have the means to build the corpus of knowledge not through the frenzied winnowing that has characterized our evaluations in the past but through an orderly interlocking of the puzzle pieces contributed by the disparate sub-fields. We have the means for the first time in our history to begin putting together the full picture of human behavior.

This paper is an attempt to do three things: to consider why, despite current capabilities, data sharing is in its infancy; to examine the rewards and challenges of data sharing; and to look at some of the initial efforts in data sharing and reflect on an optimal developmental trajectory for these efforts.

BEHAVIORAL SCIENCE IS MANY THINGS

Just as the biological sciences can be classified along a continuum from molecular to integrational, the behavioral sciences can be placed along a scale that begins at the intersection of biology and behavior and ends at the study of group behavior. So, included among behavioral scientists are those who study single brain cells and those who study the behavior of nations toward one another. That is a truly awesome breadth of inquiry. And it suggests one of the reasons that data sharing is not a tradition in these sciences: The sciences have no common vocabulary even though the scientists themselves are engaged in a common effort to understand and explain human behavior. At one end of the spectrum, scientists are talking about excitatory and inhibitory processes at nerve synapses while scientists at the other end of the spectrum are talking about behavioral schemata, contextual cues, and free and forced-choice decisions. The two talk to each other as easily as speakers of Mandarin converse with speakers of Cantonese, which is to say, not very well.

But just as the health of a country depends in part on the ability of its citizens to communicate with each other, our ability to build a deep understanding of human behavior depends on communication and integration of knowledge. One of our major challenges in considering how best to share data is to find ways to communicate what we know with each other. One could settle for communicating well within a sub-discipline, and that would be a good and necessary thing to do. But in the end, behavioral science is about understanding why people behave as they do. No simply biological explanation is enough. Neither is an explanation that is purely cognitive, or based on individual personalities or on social influences. All of these

[1] Parts of this paper come from a presentation entitled "Behavioral Science Databases: Cultural Revolution, Scientific Necessity" that was given by David H. Johnson and Michel E. Sabourin during the 17th International CODATA Conference, Baveno, Italy, 17 October 2000.

http://www.tandf.co.uk/journals/pp/00207594.html DOI:10.1080/00207590043000081

determinants of behavior need to be understood in conjunction with each other. We haven't been able to achieve that level of communication yet. In fact, the lack of ability to communicate across fields is one of the impediments to an integrated understanding of human behavior.

BEHAVIORAL SCIENCE IS YOUNG

Chemistry, physics, and astronomy have been around a long time—millennia. Behavioral science as an empirical pursuit rather than a branch of philosophy has been around only about a century. Much of the research during that time has been theory-based. Every science goes through its periods of reductionism and its periods of integration. Behavioral science's period of reductionism has resulted in a variety of somewhat narrow specializations, with each specialization having its phenomena to study, its competing theories and its specialized tools. So within each specialty there has been a period of jousting among researchers. Each has a measurement tool of choice, an approach toward study of the phenomenon in question, and an explanation of results that attracts both supporters and detractors. In this contentious atmosphere, our data have been our gold standard. They underlie the currency with which we conduct our scientific business. So we have not readily given data away. We have preferred to carry out our transactions with publications, the paper proxy for data.

But even in the midst of contention, theories eventually begin to converge because triangulation begins to pinpoint the phenomenon under study. Different approaches arrive at similar results. As agreement is reached among the theoreticians, the knowledge base of the science is increased. There have been many repetitions of this process in each of the sub-disciplines of behavioral science such that each has a knowledge base. But perhaps until now the science had been too young, and the tools unavailable, to begin purposefully putting together the knowledge produced across all the behavioral science sub-fields. The rise of collaborative research teams and the growing encouragement by funding agencies to address broad problems through such multidisciplinary teams have made it clear that an evolution toward integration of the knowledge amassed in the sub-fields is needed to achieve full functionality in the new scientific workplace.

COMPETITION RATHER THAN COOPERATION

But behavioral scientists are fully human. They strive for recognition and reward just as do most humans, including most scientists. During our period of duelling theories, the greatest recognition and the greatest rewards have come from outdistancing the competition. When others find it difficult to refute one's theory, then one becomes a leader in a field. Others follow. Promotion, tenure, and a steady flow of research funding are assured. So we have built our system of rewards on individual prowess rather than a cooperative search for knowledge. Our scientific culture has grown to be one where we are eager to share the outcomes of research but are cool toward the notion of sharing the research itself, in other words, the primary data on which our analysis and hypothesis testing is based.

As we have moved along this path, the larger scientific world has been changing around us. To be sure, competition is intense in areas of immediate interest to the private sector such as gene manipulation or drug development. But those at the cutting edge of many sciences now, including behavioral science, are likely to be members of diverse research teams rather than lone scientists. It is becoming time to find ways to accelerate progress at the cutting edge of behavioral science, and one of the ways to do that is to share data as well as outcomes.

PRACTICAL DIFFICULTIES IN BEHAVIORAL SCIENCE DATA SHARING

In addition to historical and cultural reasons for lack of data sharing among behavioral scientists, there are some very practical challenges as well. Some of them are not easily overcome. Also, a good number of these issues have ethical connotations. We will explore the difficulties associated with privacy and confidentiality, with the heterogeneous nature of psychological data, with the differential perspectives of the data analyst and the data archivist, with the retrieval of data and with the joining of previously separate data sets.

- Privacy and confidentiality

In the behavioral sciences, respect for privacy and confidentiality is a central concept in the conduct of ethical research with human participants. Privacy refers to "a person's interest in controlling other people's access to information about himself or herself" (Sieber, 1992). On the other hand, confidentiality refers to "the right to maintain private information divulged in the course of a professional relationship with a researcher (Sales & Folkman, 2000).

Difficulties with privacy issues can lead to difficulties in properly conducting research. Thus the fact that a participant perceives that his or her privacy is threatened can make him or her worry about the consequences of participation and, in many instances, can lead to biased sampling, evasive and/or false responses, and many other impediments that can affect the validity of the results.

Protection of confidentiality usually implies informing participants about what may be done with their data. Data sharing is not expected to take place unless the participant has been informed and given consent to such sharing. In some instances, breaches of confidentiality can have serious consequences, especially if members of a vulnerable population such as HIV-positive individuals, are involved. Databases add an extra wrinkle to the issue of confidentiality. By linking previously separate data, it is sometimes possible to aggregate information about an individual that

would have been unattainable through examination of data from a single study. As databases are developed, confidentiality will become increasingly hard (though not impossible) to maintain. Simple stripping of the participant's name and its replacement with a code is no guarantee of complete confidentiality. Confidentiality can be threatened in many other ways, like legal action (subpoenas), access by third parties to data, technical lapses in security, etc.

Behavioral scientists share with geneticists the need to maintain the privacy and confidentiality of those who participate in research. Some of the information participants share is extremely personal and thus very sensitive. In fact, in most behavioral research, a participant must sign a consent form telling the individual what to expect in the course of the research. The content of such form is based on the informed consent doctrine, and typically includes a clear statement of the purposes of the study, a description of the procedures, of the potential risks involved and the benefits to be expected, as well as the obligations and commitments of both the participants and the researcher. This form also commonly states that the confidentiality of the participant will be preserved and that the data collected may be shared with other researchers. It is, in fact, a kind of pact: If you participate in this research, we the researchers, will not let any information traceable to you be made public.

A question currently under debate among behavioral scientists is whether a consent form stating that personal data will not be shared precludes sharing of the data even if identifying characteristics are removed. As indicated earlier, one of the reasons is that the removal of identifying information from data gathered on an individual may not be enough since, in some cases, identities can be reconstructed from disparate data sources. In most cases, however, such detective work is unlikely because it is not worth the effort. But in cases where the health status of members of a group could affect their ability to keep or obtain health insurance, or where admitting use of a banned drug amounts to admission of a crime, or where knowledge of one's mental health status or sexual preference could be a factor in determining one's employability—in cases such as these there is no guarantee that simply removing identifying information before sharing data is adequate to safeguard the privacy of respondents.

Fortunately, there are solutions to the challenge of maintaining confidentiality including substituting numerical identifiers for names, aggregating data so that the performance of individuals is not obtainable, encryption, or layering data so that researchers who need identifying information can obtain it only after signing a legal document that requires honoring the confidentiality of individuals. Researchers who do not need identifying information can have free access to aggregated data.

This is not to say that the solutions will always be easy. Later in our discussion, we will look at some databases that are now under construction. One of these pioneering databases is comprised of videotapes of children's behavior. It is a database that is used by developmental scientists to study the development of social behaviors. Essential to this database is the ability to see the children's faces because some of the behaviors to be coded involve facial expressions. This database was created for secondary analyses, but it is not a database where it is possible to obscure even so personal a piece of information as one's physical appearance. This points up a requirement for rendering some types of behavioral data useful. That is, some data have little value unless personal information is retained with the data. Longitudinal studies that follow individuals throughout their lifetimes are prime examples of this requirement. To make sense of the data, it is necessary to follow the same person through repeated measures of his or her life events. Some data sets in behavioral genetics require for analysis not just the identity of an individual but that of the individual's family members over many generations. Included in some of these data sets are prominent public figures and even internationally recognized criminals. So, for behavioral scientists preserving confidentiality as databases are created is a non-trivial task that in some cases is also technically difficult to accomplish satisfactorily.

- Heterogeneous distributed data

Another practical problem that the archiving of psychological data can face is the heterogeneous nature of the data collected. It is well known that psychological data are not as neat and well constrained as DNA sequences, and can take a variety of forms (Boker, 1997): categorical or continuous numerical tables or matrices, mixed alphanumeric and numeric tables, tables with relations among them, natural language text, video and audio data. And these different data types can have different storage formats and even different query formats. So, very soon, we can find ourselves with a scientific Tower of Babel phenomenon.

- Perspectives of the analyst and the archivist

As described in a paper delivered by Steven Boker in 1997 at a Symposium on Data Archiving presented during the Annual Convention of the American Psychological Association, problems associated with psychological data archiving can be defined from the data analyst's perspective and from the data archivist's standpoint if the Web is to be used as a data archiving tool.

From the *data analyst's perspective*, the following questions need to be answered:

(1) How does one find the variables of interest? Since the data are likely to be distributed in a wide variety of sites, secondary data analysis can only proceed if a good deal of information is known concerning the methodology used for data collection and the coding of the variables.
(2) How does one obtain permission to use the data? Not all the data archived are public; some can be of limited access or even confidential. So, a convenient method must be devised for the granting of permissions.

(3) How does one query to obtain the data? Since not all variables are likely to be analyzed, the proper way of sub-setting the data must be identified.
(4) How does one join data from disparate sets? To perform such an operation, common index variables that serve to link data records together must be defined, being aware that such an operation can lead to an increased danger of loss of subject confidentiality.
(5) How does one analyze the data? In other words, how can the benefits of secondary analysis be maximized?

On the other hand, the *data archivist's standpoint* implies the appropriate answers to the following questions:

(1) How does one store the data? The format of data storage must be chosen to be useful to a maximum number of researchers and to continue being useful for a good number of years.
(2) How does one store the metadata? Here again, the methods of data collection and coding are to be properly described to ensure their efficient analysis.
(3) How does one provide confidentiality and control access? Here we have two possibilities: either the confidential information can be removed or encrypted, or access must be restricted in some ways.
(4) How does one link to similar data sets? In the case of longitudinal studies, which are quite popular in the necessary study of psychological development, individual identifiers have to be defined, while respecting confidentiality.
(5) How does one advertise availability? The existence of data archives must be made known to researchers inside and outside the field, thus promoting novel and original analyses.

- Data retrieval and data joining issues

Once the problems related to the choice of a data format have been solved, the question of how the data can be retrieved is of paramount importance. Since in most cases, only partial data will be required for analysis, appropriate querying and sub-setting mechanisms must be identified. There are two possibilities: either the entire data set is transferred into the analyst's computer or a specific query can be performed to provide only the data needed. Both systems have advantages and inconveniences, but all things considered, it seems easier and more useful to be able to perform queries on databases. However, local query methods can be costly in time and resources.

Most often, secondary analysis requires the joining of records from different databases. However, data joining can create technical and ethical problems. When using two databases, some method of creating a common randomized index key must be devised. This can be easy from a technical point of view and totally respectful of subject confidentiality, especially if private key encryption methods are used. But as more and more variables become present, the chances of creating unique records increase, and confidentiality can be lost. The linkage of experimental data with demographics always raises the possibility that a specific individual can be identified. Psychologist Boker (1997) identifies "jittering" as a possible way to overcome this problem. It involves "adding a small, normally distributed random value with mean of zero to all fields that might be used to identify an individual by matching against publicly accessible records". Thus, only a small amount of uncertainty is added to the statistical analyses, but the identification of individuals is prevented.

A LOOK AT THE PRESENT AND A PLAN FOR THE FUTURE

Realization that the behavioral sciences need databases now is coming independently to individuals in several sub-disciplines. Thus we have under development or in current operation a database of brain images that is quickly becoming useful to neuroscientists, databases of videotaped and audio-taped material that has already been of great use to developmental psychologists, a database of longitudinal studies that is of use to those studying ageing, a database of intelligence test scores that is of interest to both cognitive scientists and test developers, a cognitive testing database that is invaluable to cognitive scientists, a database of research related to women, and a database of genetic data that will be of use to behavioral geneticists. On the one hand, these efforts are to be hailed and encouraged. But the independence of the efforts is also troubling.

- Coordinated versus piecemeal database development

A classic situation that arose over thirty years ago continues to illustrate the problem. The situation had to do with converting the library classification system for books from the Dewey Decimal System to the Library of Congress system. The story is that the library holdings at a major research university in the United States using the Dewey Decimal System were so extensive that it was impossible to reclassify all the books. The university chose to leave its holdings at the time of conversion in the Dewey Decimal System. New books only were classified under the Library of Congress system. The result was what amounted to two incompatible libraries.

What is troubling about uncoordinated independent efforts to create databases in the behavioral sciences is that a point may be reached where it will become very difficult, and perhaps impossible, to accomplish interoperability among these different archives. They may remain separate because their differences will become so great that they cannot be joined.

That is potentially a tragic misstep for the behavioral sciences. There is a window of opportunity now to build behavioral science archives that can be transformational for the discipline if they are developed as a system. Through systemic development, it may be possible to begin integrating the knowledge that is being generated

within the sub-disciplines but that now remains largely locked within those sub-disciplines.

What are the advantages of thinking systemically rather than in a piecemeal fashion about behavioral science archiving? The goal of behavioral science is to understand human behavior. Understanding how the brain works, how social situations shape our actions, how motivation influences behavior—these are all just steps along the path to the goal. If there is any hope of truly understanding human behavior as opposed to understanding components of behavior, that hope lies in putting the pieces together. That is nearly impossible with the narrow specializations that are normal in the field today. We have to build toward a future that enables integration, and relational databases are an important part of the needed empowerment.

- A systemic approach to data sharing

In fact, the behavioral sciences are beginning to examine electronic archiving at a point in the development of electronic storage and communication where we have the luxury of asking what is possible and then being deliberate in the building of a system that takes full advantage of the possibilities. Three years ago, we decided to begin working toward a systemic approach to data sharing. Who is included in that "we?" Currently the "we" is comprised of 23 behavioral science societies, many of them international in their memberships, and the International Union of Psychological Science. And the roster is open for new members as they see fit to join us.

Why is a consortium of societies rather than a university or a set of individuals attempting to undertake this project? The answer goes back to the question of what it is possible to accomplish with today's and tomorrow's anticipated technology. Under ideal circumstances, it should be possible to go well beyond simply the storage and retrieval of raw data. We believe we should think in the terms used by those in the library sciences. That is, we should think in terms of "knowledge management" rather than simply "data management". Knowledge management is the coordination and linking of many knowledge sources. The linkages enable users to move easily within and among the knowledge sources. That ability offers the opportunity to produce new knowledge through the juxtaposition of related pieces of information that had not previously been linked. Meta-analysis is one of the statistical techniques that is being developed to accomplish such knowledge-producing juxtapositions. Common coding procedures are another helpful tool now under development and made possible by the video database we mentioned earlier. Developmental scientists have realized they can standardize their coding procedures using the videotapes. The result is that from now on these scientists will be able to compare results of different studies with greater confidence in the reliability and validity of the comparison. These and other procedures being developed will assure better comparability of data from different studies.

The knowledge sources that we seek to link initially in this knowledge management structure are those one would immediately think of in most of the sciences—the journals, the abstracts, and the raw data. It would be an ideal situation to be able to search back in time and across all abstracts or journals in the field and then to examine and analyze data relevant to those journal entries combining data in ways that they had not previously been combined.

Such a systemic approach would be able to give space to something that is lacking in the published literature, and that is, negative results. Behavioral science journals rarely publish accounts of experiments that yielded negative results. But when such experiments are done well, the fact that results failed to support a particular line of reasoning about a phenomenon is as important to know as are the positive results that are commonly reported in the journals. A system of databases that includes negative results can not only enable new knowledge, but can make the process of discovery more rapid and efficient by saving researchers the trouble of covering the same ground covered by other scientists who to date have not been able to report their results because of the current bias among journal editors against negative results.

- The challenges ahead

Meeting the ideal is challenging to accomplish. Journals and abstracts are owned by both for-profit and not for profit entities. Societies that publish journals often depend on sales for a major portion of their operating revenue. And for-profit companies don't make much profit if they give their products away. Likewise, scientists in the behavioral sciences do not surrender their hard-gained data readily. How can obstacles this formidable be overcome?

In figuring out how to deal with these obstacles, we have been cognizant of literature, again from the library sciences, though the original work was done by economists, on diffusion of innovation. That literature concerns how to successfully introduce something new to any kind of group from a corporation to a culture. One of the most important ingredients in successful diffusion of innovation is something those in the U.S. call "buy-in." That is, the people who are to use the innovation must be accepting enough of the innovation to actually make use of it.

Related to buy-in is the issue of where in an organization innovation begins. If it begins too low in the organizational structure, it will fail because those in the lower parts of the organization lack the authority to spread the innovation through the organization. Typically, there are three groups through which innovation must diffuse to become accepted and used. The groups are early, middle, and late adopters. The early adopters are the eager people who want to try anything that is new. The middle adopters begin to try the innovation after it has shown some sticking power among the early adopters. The late adopters are the most crucial group because their decision either to adopt or reject the innovation makes the difference between whether the innovation becomes part of the organizational culture or fails to be accepted and dies. Late adopters are the people who were sceptical about, or openly hostile to, the innovation from the first.

Within the population of possible late adopters, there is a special group. These are the people who have a large investment in the status quo because they are leaders under the current way of doing things. As leaders, they have the power to stop innovation. They also have the power, should they accept an innovation, to propel it into common use. It is important, therefore, in any attempt to bring innovation to a group, to respect the abilities and the needs of those who lead the group.

Following this model, it is clear that if the goal is to spread a technical innovation throughout a science, it will not be adequate to organize the database in a university or even in several universities because they lack the authority across the science to drive the diffusion. The entry point has to be at a level that cuts across universities, research specialty areas, and even national boundaries. The next higher level of organization is the scientific society. But if the goal is to diffuse the technology through the science regardless of where those scientists might be in the world, then it is unlikely that a single association, even if it is large, has the level of authority to accomplish the task. We are left with the conclusion that the best chance to successfully introduce databases across the behavioral sciences is to do it through a consortium of associations both national and international. And that is what we are attempting to do.

- A partnership of scientific societies

An organization based in the United States called the Federation of Behavioral, Psychological and Cognitive Sciences was a ready-made vehicle for starting this long process of necessary cultural change within the behavioral sciences. The Federation is a coalition of 19 scientific societies, many of which are international in their membership. It was representatives of these societies who requested that the project be taken on for the benefit of their societies. They also requested that the project be international in scope from the beginning.

One important aspect of undertaking this project as a partnership of societies is that it is these same societies that publish the major journals of the field. Some have made the switch to electronic publication, but many have not. One of them, the American Psychological Association, invested $100,000 to create software especially suited to electronic publication in the behavioral sciences. APA has agreed to give this software to any other society that wishes to begin publishing electronically. Through this offer, we have the means to bring uniformity to the manner in which electronic journals in the behavioral sciences are published electronically. The software was built with connectivity in mind. It will easily allow an article to be linked with data, related articles, citation indices, or any other kind of electronic information, including raw data.

Even though it was representatives of the societies who originally raised the idea of building databases cooperatively, it is fair to say that those who made the request did not understand all the dimensions of activity their request requires if it is to be fulfilled. An important concern is whether the societies will be able to adjust to a very new way of distributing information. As we mentioned, the hardest and most crucial group to bring to acceptance of innovation is that comprised of those who are the biggest stakeholders in the old way of doing business. With that in mind, we have been thinking that we must involve key individuals who, if left uninvolved in the early planning and implementation of the innovation, would become the most reluctant late adopters. Those individuals are the ones who are currently responsible for the communications outlets of the societies, that is, the journal editors and the heads of publication.

It is convenient that while these people have the most invested in paper forms of communication, they are also the people who care most about communication of science, whatever its form. That is why our current plan is to make the journal editors and the heads of publication of the societies that choose to become members of the database consortium also the directors of the electronic knowledge management system. We expect to bring these communications officials of the behavioral science societies together to test the feasibility of forming a board of directors of the database consortium. In the end, it would be their responsibility to shape the knowledge management structure into the system that will best serve the scientists across their societies. As we said, membership in the consortium would remain open. Each participating society would continue to have the right to appoint a member to the board of the consortium.

- Common protocols for electronic publication of journals

The first thing we would ask of this board would be to establish common protocols for electronic publication of journals. We would present the American Psychological Association software as a possible standard. That software has already been used to make abstracts of all APA journals available online. The abstracts go back 110 years. Full text electronic versions of journal articles from all APA journals are currently available for the past ten years. APA is continuing to add more years of past publications to its database. Most important in the way the software has been designed is that it allows for full search ability within articles, across journals, and across time. It is also ready to enable search ability across archives of data once they come online.

- Science for all

The second early task of this board would be to consider how not to put the societies it represents out of business. That is, electronic knowledge management cannot suddenly make available for free the information their societies sell in order to finance their existence. Long term financing of the databases is likely to be a very difficult problem because of a principle we would establish from the beginning. And that is, in creating a knowledge management system, we do not want to build a device that further divides the information “have” from the information “have not”. In our view, that makes no sense. Our goal is to improve our science by making the knowledge legacy of

our sciences available to all scientists. With such access, faculty in universities and colleges where it would otherwise be difficult to conduct research would be able to do research. These materials would also become generally available for teaching purposes. We want to expand rather than control access.

- The costs of knowledge management

But knowledge management is not free. There are expenses to create, build, and maintain the system itself. Besides the upkeep of the system, there is the ongoing need of the societies to pay the cost for their operations that are now covered by revenue from the sale of journals. How can we support the databases without hurting our societies while assuring that the income of scientists is not an obstacle to their use of the databases?

The question of how to fairly price electronic information is one of the most hotly debated topics in publishing today. Many experiments in pricing have been undertaken by publishing companies like Elsevier in the for-profit sector and the American Psychological Association among non-profit associations. We cannot claim to have the final answer on the matter. We do, however, have ideas. The one we would present to the board would be to consider the cost of data archiving separately from the cost of journal publication and to price the two by different means. It is our belief at the moment that the best way to make the data archive most open to the widest scope of scientists, students, and scholars is to spread the cost of maintenance across a very large support base. We would propose that the data be considered the commonwealth of the scientific community and that there be a database maintenance surcharge on the dues of members of all member societies. In countries where the scientific infrastructure is relatively undeveloped, this surcharge would be very low, perhaps non-existent. Scientists in wealthier countries would underwrite accessibility for scientists in poor countries until those scientists are more able to take on their share of support. We would also seek government support for the start-up work on the databases. Government support would taper off while support by scientists would increase over a ten year period.

As we mentioned, the pricing of the journal side of the knowledge management system would need to have the possibility of profit built in. The databases archive, on the other hand, would be a break-even enterprise. The bridge between the journal articles and the data is the abstracts. Like the journals, the abstracts are profit makers. As a first approximation of a way to support the abstracts, we are proposing that access to the abstracts be included in the database maintenance surcharge to society members. The abstracts already exist online and on CD-ROM. They are maintained by a for-profit organization. We are suggesting that a portion of the data surcharge would go to that organization as its profit for use of the abstracts by scientists whose societies are members of the consortium. The arrangement would be similar to payment for a site license. Since the company would want to reserve its right to sell access to the databases to libraries and other users who are not part of the consortium, we are suggesting entry procedures to the abstracts for members of the consortium that are similar to those widely used by the societies now to allow members to access "members only" sections of their own web sites. That is, users will need to log on to the abstract service with passwords in order to get beyond a list of titles of articles that might possibly be useful.

From the abstracts, users will be able to determine the articles for which they will need to see full text. At that stage, the system would move the user to the site where a society maintains the full-text of its electronic journals. Users would be charged to download or print the article in much the same way that Medline now charges individual users to access the full text of articles abstracted in its database. In this way, the individual societies will maintain control over pricing of the material each publishes. The abstract and database portions of the system, that is, the part of the system supported by the surcharge on dues would become a gateway to the articles. It is conceivable that offering two online paths to the articles might even create extra revenue for societies that now offer their publications only on paper.

We think that text translators are still in their infancy as far as most natural language material is concerned and certainly as far as electronically available scientific information is concerned. But we expect these translators to improve. Eventually, we expect that our knowledge management system will be able to offer its materials in faithful form in any language a user desires.

- Designing the system architecture

This latter point about adding new capabilities as they are developed brings us to a third responsibility of the governing board. It must exercise its collective wisdom to accomplish the difficult task faced by all archivists using electronic media. That is, it must make choices about the architecture of the system that will help the system adapt to change with a minimum of pain.

It is a problem with which the board and the staff of the database consortium would gain ample experience since the data that would be going into the databases initially will not have complied with any standard. Some of the data are on paper, some on magnetic tape, some on floppy disks in all the sizes and capacities we have experienced over the years. Some were entered on computers that no longer exist using operating systems that were abandoned long ago. Those in the archiving business who have been at this the longest have found that they have needed to find a working machine with functioning software from each era of electronic data storage. They then move the data from one medium and machine to the next in the ascending hierarchy of development until they reach the current generation of technology and software. The initial job of building these databases will be arduous because of the variety of media on which the data are currently stored and the variety of ways in which the data are recorded regardless of medium. We expect that once the older data are in the system, the job of adapting to new technology will not differ in mode from the current method, but will

be more orderly. That is, instead of needing storage devices and software from every era, we should reach a point at which we need only the current and the next generation of device to adapt from one to the other. This would be akin to having, say, both a zip drive and a read-write CD on the same machine.

We are proposing to use the American Psychological Association computing facility as the first site of the databases because the APA has built a system especially for behavioral scientists. It has successfully transitioned across nearly twenty years of innovation in storage media. And the system is modular in its construction with a variety of storage devices already operating within the system, and with the ability to add new devices as they become available. We think this will give our databases the best hedge against obsolescence.

- A proposal for international development

An interesting aspect of our effort to develop these databases internationally is that we must be understanding of the data sharing constraints of those in each country wishing to participate. That understanding is complex in the sense that it must include understanding of national legal constraints on data sharing, a matter that is of particular importance with respect to European Union participants since the current preference in the EU seems to be toward respecting the privacy of individuals at the expense of free scientific communication and toward valuing copyright protection over the free flow of information for scientific purposes.

We must also understand cultural sensitivity about sharing data with those in other nations, particularly the United States. For better or worse, the United States is not the most trusted of countries around the world. A considerable number of non-U.S. behavioral scientists, over and above the general sensitivity present throughout the discipline about sharing data at all, are even more reluctant to turn their data over to a facility located in the United States.

To be understanding of that reluctance, the board will contain representatives from participating societies around the world. Moreover, our proposal is to have satellite or mirror sites at least in Europe and Asia, and in all likelihood in Africa and Latin America as well. For efficiency, it is likely that most processing of data to prepare it for distribution would take place in the central office in the U.S. The satellite sites would collect data sets in their own region, collaborate in assuring that the data sets are arranged in ways that facilitate use. And the mirror sites would also contain full sets of the material in the database system.

As we mentioned earlier, a number of databases already exist in the behavioral sciences. We would wish to have these databases become part of a system that appears seamless to the user. It is likely that these existing sites would continue to collect data of the kind that is their specialty. So in some cases, the central facility for the database consortium would be a satellite site with respect to these pre-existing databases. The ability to sit at one's desk and move from one set of data to another and see some uniformity in how those data look is more important to us than where the data reside. But understanding that the question of physical presence of the data is not unimportant to everyone, we will make the accommodations we must in order to achieve the usability that is our goal.

As we have mentioned, the data of the behavioral sciences come in a broad variety of forms from brain scans to questionnaires. We also said that this variety has been a factor impeding the integration of knowledge produced by our scientists. Merely making it possible for a social psychologist to look at a brain scan, or a psycho-biologist to look at a video of behavior is not going to be enough to foster efforts to integrate knowledge. On a surface level, but still a very important level, we think that there must be a common user interface to access the data sets. This is the same idea that dominates the development of computer operating systems today. One uses the same techniques to carry out a variety of functions in a variety of kinds of software. We all know how to use pull down menus, how to cut, to paste, etc. It is this kind of surface comparability that we want to strive for in the common user interface. Our rationale is that we must make the use of each database as similar as we possibly can in order to encourage scientists to explore beyond the bounds of their particular area of specialty.

The surface similarity will carry users only through the door, so to speak, because the databases will be very different in their content. How do we encourage users to plunge into data of a kind with which they lack familiarity? Drawing on the experience of those who are working on behavioral science databases right now, it seems that the most important tools to encourage use are a clear description of the data and what could be called a clear user's manual. One of the objections we hear regularly from scientists about sharing data is that no one will understand the data sufficiently to use it properly. That is a very real possibility. So in order to begin using a data set, we expect to require of potential users that they pass through a process. That process would require them to read a description of the data that is extensive enough to satisfy the producer of the original data that a potential user will know the kind of data collected, how the data were collected, what constraints must be observed in use of the data, how the data are coded, what analyses were originally used on the data, and at what level of aggregation, if any, the data accessible to the potential user will be found. Basically, this combination data description and user's manual will prepare the potential user to use the data properly.

- The methodological database

If after gaining this grasp of the data, the user is ready to proceed, then he or she may do so. Many users who are looking at a particular kind of data for the first time, will not be ready to proceed. To anticipate such cases, we would propose to cross-reference the user's manual with a special methodological database that will be part of the database system.

This particular database, the methodological database, is important to our fields for a number of reasons. In the United States in recent years, the number of universities offering a firm grounding in methodology has shrunk. It is becoming more common for graduate students to take their statistics and methodology courses outside their departments. At the same time, the number of people whose primary area of specialization is methodology has also shrunk. There are not enough methodologists to go around anymore, and their average age is rising. So we have a growing methodological crisis, at least in some countries, that can be partially addressed through this database project. For each data set, we expect to build a unit in the methodological database on how to treat data of that kind. We envision that faculty will be able to use this database to improve the teaching of methodology to graduate and undergraduate students alike. We also expect it to be a means for those already engaged in research to sharpen their skills or to pick up an analytic technique with which they are unfamiliar.

The need to do this arises because of one other wrinkle in our methodological personnel crisis. Those who are methodologists used to publish in the same journals as those who use the methodologies. But now it is more common for methodologists to publish in methodological journals which other scientists do not read. The result is that while methodology is developing constantly, the number of those who can use the old methodologies well and who have any inkling of powerful new methods of analysis is going down. So we want the products that are produced by the database consortium not only to make data itself more available to scientists but also to make the tools for using those data properly more easily accessible as well.

CONCLUSION

What we have been outlining is a complex and unprecedented undertaking for our sciences. While we can talk about ideas, to be practical, we have to have a way to implement those ideas. We are talking about putting together a set of very different databases in such a way as to make them reasonably useable to scientists unaccustomed to databases at all and equally unaccustomed to analyzing data outside their areas of specialization. Just as no single university or society can manage such an undertaking, no single group of database builders can do an equally good job with each database. That means that we must have an overall management team and then a set of teams each with expertise necessary to build the database that makes use of data in their specialty area. Moreover, the undertaking is large enough that it cannot be accomplished in a short period of time. We are estimating that we will need ten years to put together the basic skeleton of the system. There is still some debate over the best way to accomplish this since we need the undivided work of team members for at least a year at a time just to bring a database online at a basic level. And getting a year of a scientist's time is not an easy thing to do.

Here is our thinking right now about how to accomplish it. We believe we need three layers of workers on the project: A management layer, a layer of permanent professional staff, and a layer of rotating topical specialists. We would seek government funding to sustain the project through its start up years. The management layer would consist of three people: A professional archivist who would oversee the work of the rotating topical specialists, a technical manager who would oversee equipment, software, and the professional staff. And an executive secretary for the board who would also be responsible for overseeing efforts to promote use of the database by scientists.

The professional permanent staff layer would consist of a specialist in information system architecture and perhaps three data management specialists who would take the data collected by the rotating topical specialists and actually put that material into the online database system.

The rotating staff would consist of scientists who would compete for fellowships to begin the gathering of data in their field and to arrange these data with the guidance of the database architect into an order that will make them useful to fellow scientists. We envision rotators spending a year in their fellowship. Their offices would be located at the American Psychological Association building in Washington, D.C. We expect that these fellows would enlist the help of others in their field and that they would help to build a network of people in their field who would aid them in obtaining the data. Under this plan, one sub-disciplinary database would be added to the system each year together with the methodology modules that would correspond to the data collected by the fellows.

An obligation of the fellows as they near the end of their fellowship year would be to go out and preach the gospel. That is, they would visit the major university departments in their specialty area and show faculty and graduate students how to use the system. We cannot tell at this point whether this method of attempting to cause the culture of the science to evolve will work. We would expect to try a variety of other advertising mechanisms to complement the method of direct contact. We consider the projected ten years of initial development time for the project to also be the most crucial period to work at making the databases an important part of the way our scientists do business. That, in a rather large nutshell, is our proposal. The authors are anxious to know the views of others, many of whom, it is acknowledged, will have practical experience in these matters that could prove invaluable to the undertaking.

REFERENCES

Boker, S.D. (1997). *Using the Internet and the World Wide Web for archiving psychological data*. Paper presented during a Symposium on Data archiving in Psychology, Annual Meeting of the American Psychological Association, 16 August, Chicago, Illinois.

Sales, B.D., & Folkman, S. (Eds.) (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association

Sieber, J.E. (1992). *Planning ethically responsible research: A guide for students and Internal Review Boards*. Newbury Park, CA: Sage Publications.