Opportunities and Pitfalls: Observations on the Application of Database Software to Bibliographical Activities

T.B. Vincent

Historically, bibliographical work has been characterized by the image of the solitary researcher spending long hours in dimly lit repositories creating lists of materials, using tools no more complicated than the pencil and the file card. The apparent simplicity of this activity is elevated by our profound respect for the patience, the meticulous attention to detail, and the thoroughness required if the work is to be reliable and merit our confidence. Like many solitary pursuits, traditional bibliographic activity lends itself to a kind of nostalgic romanticizing which tends to throw an emphasis on the process as an end in itself rather than on the result. This view should not be surprising. After all, for the bibliographer, the experience does lie in the process, and, at a time when new technology may bring fundamental change to that process, there is a natural tendency to lament the passing of patterns of experience which have defined the nature of a professional activity in an intimate way.

However, by investing established procedures with nostalgic meaning, we often ignore their inherent limitations and play down the frustrations that flow from these deficiencies. Sentimental conservatism can amount to a form of blind stubbornness. When I contemplate the reasons why I was drawn to computer technology, I find that it was not so much a perverse curiosity or the excitement of exploring something new that attracted me; it was the frustration of dealing with the limitations of established structures and patterns that pushed me to look for more effective bibliographic tools. I wanted much greater power of selection and arrangement, I wanted more flexible mechanisms to sort-and-see, and I wanted to spend more time working with the product and less time working on the process. As a researcher and scholar concerned with the history of Canadian literary development, I tend to view bibliographies and indexes as vehicles to get to a scholarly job that needs to be done. In this context, bibliographies and

This paper was given at the 44th Annual Meeting of the Bibliographical Society of Canada held at the University of Alberta, Edmonton, Alberta, on 20 June 1989.

indexes are not ends in themselves. As a bibliographer, however, I also understand the attraction and satisfaction and the importance of order, arrangement, and completeness as legitimate professional values. In short, as a researcher, I want to get to my scholarly point of departure quickly and effectively, but I want to get there in a vehicle that, as a bibliographer, I can trust.

I introduce the term 'trust' here because I think that was the crucial element involved in my coming to accept the applicability of computer technology to bibliographical activity. Ultimately, it was trust, not utility, that was most important. My attraction to the technology began because of frustration with traditional forms; but for me to embrace it, the technology had to pass the test of trust.

When I think about it now, I see that I passed through several stages in reshaping my sense of trust. One of the things I discovered was that the medium in which we work has a great deal to do with our trust in the process with which we are working. I had always worked with cards. I trusted cards because I could see them, I could hold them, I could lovingly hover over them, and most important, because I could control them. But for this comfort I paid a price. Cards can only be arranged in one sequential order at a time. If you need to approach the information on those cards from a variety of different directions, the cards must be resorted and rearranged manually for each type of query. This is perhaps no great problem if you are dealing with a few hundred records. You might think twice if it were a couple of thousands records; but at twenty, fifty, or a hundred thousand records you do not even contemplate the task. Curiosity is sacrificed to practicality and mental health, or more accurately put: the medium to which you have entrusted your information works to constrain and limit the range of scholarly inquiry. I reached a point where I felt that this was too great a price to pay for the comfort of 'card control.' I was frustrated by the knowledge that there was information on my cards that I could not get at without expending considerable time and effort in resorting them, and, as a scholar, I was disturbed at how readily I curbed my curiosity in the face of apparent expediency.

And so I began to look into computer database technology, or more accurately, relational database technology. I should note here that there is an important distinction to be made between relational (or near-relational) database software and textual database software. A textual database is designed to search through big chunks of text very quickly and to locate occurrences of words or parts of words which you have asked it to find. For bibliographical work (certainly for bibliographies of any size), textual databases do not really offer the sophistication, precision, and flexibility that you will ultimately want – at least, that has been my experience. The same

amount of time invested in developing relational database files will pay greater dividends in the end. For the purposes of this paper, the discussion will focus on relational databases unless otherwise specified.

I started with a relatively simple piece of database software and a relatively small number of records. I sat down and explored what the technology could do and what I needed to know to make it do it. I worked with this database until I could make it do what I wanted it to do. I worked with it until I knew I had control over the information I had put in and knew that the information was not going to disappear on its own accord into the dark recesses of some internal computer hell where I could not reach it.

Then an interesting thing began to happen. As I developed a greater trust in the technology, I began to respond to the inherent attractiveness of the machine itself and to the systemic complexities of the software. This is an obnoxious condition to which males seem particularly prone, perhaps because it allows for the development of a great sense of superiority based on very little real knowledge. I had to remind myself that I was a bibliography nerd, not a computer nerd, and that computers and computing software were only important to the extent that they served the genuine needs and requirements of my bibliographical objectives. This is not an idle observation but an important distinction to keep in mind always: the technology must serve the scholarly demands of the profession, never vice versa. It means that, right from the beginning of a project, the bibliographer must be guided by a clear set of bibliographical requirements rooted in the real needs of the researchers that the completed bibliography is intended to assist. If you allow yourself to be guided by what the technology lets you do, then the resulting product may prove to be more a demonstration of the intricacies of the system used than a practical and effective aid for researchers.

But while the use of computers in bibliography presents us with certain intrinsic dangers, it also forces us to re-think some of the bibliographical practices we take for granted. After I had come to trust computers as a storage medium which I could control and which would not do horrible things to my information, I found myself focussing on questions of presentation. I found that traditional patterns of presentation were more important to my confidence in a bibliography than I had previously realized. As far as computers were concerned, the question of form was easily satisfied by demanding that the software I use be able to present the information on screen or on paper in the format I wanted, even though the information itself was not actually stored in that pattern. But it made me realize that the real integrity of a bibliography lies in the accuracy of its information and not in its arrangement. Most of us would accept this observation as a truism, but if we are perfectly honest with ourselves, we should recognize

that we tend to elevate accuracy of form (mechanics) to at least the same level as accuracy of content (substance). Part of that respect comes from the recognition that accuracy of form in a manual medium requires a lot of time, care, and attention, but mostly our respect for form is simply a product of our expectations relating to how a bibliography as an object should look. My experimenting with the computer reminded me that intrinsically a bibliography is not an object. It is a collection of inter-related pieces of information that may be objectified in a variety of forms. We may lament this sharp distinction between bibliography as object and bibliography as information-in-suspension, but we cannot afford to ignore it. Let me give a small example.

I have a great respect for Reg Watters' *Checklist of Canadian Literature*, partly because I knew the man and saw the dedicated effort he put into it, partly because it represents a common-sense marriage of the needs of literary scholarship and the requirements of professional bibliography, and partly because it is the rock on which the contemporary study of Canadian literary culture is founded. But like any fixed-format presentation of information, there are built-in restrictions on the kinds of information you can extract quickly and effectively from the work. Some time ago, in an effort to test some software, I asked my researchers to load the poetry (and subsequently the fiction) sections of the *Checklist* into a database to which we added a number of titles that had been missed by Watters.¹ While the additional titles enhanced the value of the database, its greatest utility came from being able to extract specific subsets of entries or particular selections and arrangements of records as required. If I need to know what books of Canadian poetry and fiction were published in Halifax in 1876, I can go to this database and quickly find out. If I want to see what fiction was published by the McMillan company of Saint John, I no longer have to weigh my desire for that information against the time and effort it would take to get it; the database gives the answer and gives it quickly. If I have part of a title and have no idea who the author is, I can ask the database if the title appears in the data and who authored it. I can use any combination of some fourteen groupings or fields of information to filter out the particular information in which I am interested. In short, the computer gives me all the possible indexes which Reg would have liked to have incorporated into his checklist. It turns that checklist into an incredibly flexible research tool, and opens it to potential search patterns that even the creator of the database may not have thought of.

To begin to appreciate the possibilities that a bibliography in computerized form offers the bibliographer and the researcher, it is helpful to imagine that bibliography, not as a sequential list, but as a fluid suspension of interrelated pieces of information, waiting for you to give it the particular

genetic code that will dictate the specific shape and pattern of the desired body of information that you are seeking in any given instance. If you wish to create a certain kind of list and to print it out in a desired format, you can do so quickly and easily. But in fact it is more practical to leave the information in its suspended form and let each individual user create or design the particular listing that he or she desires. This is not unlike what we do when we use the computerized catalogue of holdings in an automated library. The main difference is that the library software prepackages the information you get and limits the range of the queries you can make. While this simplification may be desirable for the unsophisticated user, it is regrettable that it should restrain or preclude more experienced users from exploiting the full potential of a system whose greatest virtue lies in the flexibility of choice and the power of manipulation that it places in the hands of the individual.

Although it is the enhanced powers of choice and manipulation that impress me most when relational database software is applied to bibliographical concerns, there are other benefits worth noting. Bibliographies stored in databases are open-ended; they can be added to, edited, or amended at any time, quickly, easily, and inexpensively. And the software places any additions in an appropriate relationship to what is already there so that they appear where they should in subsequent listings. Granted, there are logistical problems in updating multiple copies of database files if they are widely distributed, but these problems are not unsolvable.

Another interesting possibility is the merging or linking of database files from different sources. For example, if I had a database file containing a bibliography of Canadian war literature and you had one containing a bibliography of Canadian military history, it might well be advantageous to link or merge the two files so that a researcher would be able to query both simultaneously. This is not as simple a process as I make it sound, but it is far from being impossibly difficult. I raise the idea because, as time goes by and the importance of specialized bibliographies in computerized form becomes more apparent, we will need to think seriously about swapping, linking, and merging our bibliographical efforts with one another. My point is that the technology not only facilitates the creation of bibliographies but it also provides a means of effectively consolidating bibliographical information in a wide range of pertinent packages.

A related activity would be to transfer some of the data from an existing database to use as a foundation on which to build a somewhat different database. Here is an example. In the largest project that my researchers are working on, we have (in lieu of a better term) 'indexed' a number of nineteenth-century Canadian university periodicals.² Some university archives have shown an interest in acquiring the data relating to their institution's

publications, and I have had to point out that, in the context of our project, the entries are not annotated to the degree that their users would find most useful. However, if our data were used as a foundation to which a more refined set of annotations was added, anywhere from one third to one half of the time and cost of creating the desired index could be saved. I believe that bibliographers and indexers must build on each other's work, and storing our information in computerized form will facilitate the process of give-and-take.

In the case of very large bibliographies, computerization becomes essential, both in the creation and in the using of that information. The technology allows you to expand the cross-referencing structure of the data while retaining full control of the interrelationships, and when it comes to using that cross-referenced network, it lets you do so at remarkable speeds. I am currently working on a database that at this point contains more than 110,000 records, each of which is broken into fourteen fields.³ This means that there are over a million and a half pieces of information, all interrelated, in the body of this database, representing a rather detailed table of contents of all the extant issues of 140 nineteenth-century Canadian periodicals. If I ask the database for a typical set of information relating to a particular author and his publications, generally the computer will locate, retrieve, arrange, and present all of the requested information in less than five seconds. Even when the queries are complex and require that all or part of each record be searched, the information is usually delivered in under eight minutes. That same body of information in any other medium would literally be too awkward, if not physically impossible, to handle as a whole, requiring that it be broken down or subdivided into separate usable portions. The computerized database, however, allows us to search all issues of all 140 periodicals all at once, and immediately thereafter permits the researcher to pursue a wide variety of other types of searches simply by altering the search criteria. For example, if I had located citations for all of Bliss Carman's periodical poems, noted a particular title, and wondered what other poetry by other poets was published in the same issue, I can find that information in a matter of seconds. If at that point I wondered what other poetry was published in other journals that same month and year, I can find out in seconds. The technology allows me to control and manipulate an enormous amount of bibliographical information as if I were simply flipping a few pages. To have that kind of propinquity between wondering and knowing, between question and answer, is extremely important to researchers. It frees them to pursue questions that they have not considered pursuing before.

These observations on the effect of relational databases in opening up new or enhanced bibliographical opportunities leads me to comment on

the need for us to explore what new kinds of bibliographical tools are made possible by the advent of these relational structures. I believe it is time we began to rethink what kinds of information are pertinent to a bibliographical entry in a research-oriented database. As I have tried to indicate, the flexibility of these structures allows for a variety of different kinds of operations or searches and the range of that variety depends upon the kinds of information that are included in the first place. What I am suggesting here is that we should not simply apply this technology to the patterns we have always employed, but attempt to discover what kinds of patterns it allows us to pursue that we have not thought of before. This is an area of bibliographical research that I find very interesting and in which I feel there should be more activity. My own approach has not been particularly radical. I have generally attempted to create the potential for more than one operation in a database by combining the information needed for two somewhat different but overlapping bibliographical structures. In one case, I experimented with trying to integrate literary chronology and bibliography – with mixed results. The idea that has occupied most of my time, however, involves combining a form of contents reporting with a level of subject indexing, and applying that to eighteenth- and nineteenth-century Canadian cultural journals.⁴

That project began with a desire to find a way in which one could both bibliographically locate published items and understand something of the context in which they were published. The reason for that approach is rooted in my research experience: when information is taken out of context or placed in an artificial context (such as a sequential index), it is easily misinterpreted. By experimenting with expanded forms of bibliographical citation, I tried to find a way of creating an efficient and orderly guide to the elements that make up the larger whole, while still retaining something of the inherent structure of that entity in order to preserve an element of context. In the first instance (before becoming involved with computers), I produced printed bibliographical skeletons of journal contents with supporting indexes.⁵ This format threw the focus on content rather than index, creating some confusion among indexers, who found the work disconcerting because it functioned quite differently from a structure like *Canadian Periodicals Index* for example. I was not completely happy with it either because it forced me to set a focus, almost arbitrarily, and I would have preferred to let the user set the focus at the time of use. That is where computers and relational database software in particular made a major contribution to my work. They did not simply give me the speed and control to handle a large volume of data; they also allowed me to structure and present my work to the user in the manner that I felt most appropriately reflected its intentions.

A typical entry consists of three kinds of information dispersed under various headings or fields. There is the basic bibliographical citation necessary for locating the item in the original journal. Then, there are annotations relating to genre, subject, and Canadian authorship, and a record of the first line for poetry. Finally, there is a sequential coding system that allows for reconstructing all or part of a journal's contents in skeleton form. Linked to this file is a separate body of descriptive information relating to the particular journals that have been entered. Also there is a database of known pseudonyms which may eventually be linked as well. In its simplest function, this collection of information may be approached as an index, ordered under any of its fields. But as a researcher, what I find most exciting is the opportunity to use this tool as a scanner or viewer of selected information. It can be used to scan a cross-section of journalistic, intellectual, and literary activity at different points in time, across various journals, and in selected locations. Such views allow the researcher to get a quick insight into intellectual and literary contexts and to make a number of comparisons rapidly.

While I am excited about the possibilities presented to researchers by information held in electronic form, in the last few years I have become increasingly worried about the effects that an indiscriminate application of computer technology can have on research inquiry patterns. I now feel that one can be too enthusiastic about the application of computers to bibliographical activity. The computer has the capacity to perform tasks that perhaps it should not do, either because the information it produces is meaningless or because the doing of that task will have a negative or diluting effect on the research that flows from overly simplified patterns of intellectual exploration. The concern raised here relates to the experiential importance of process in the intellectual pursuit of concept and understanding. I am speaking here as a researcher in the humanities and arguing for a sensitivity to the complex process involved in the search for and acquisition of understanding. Data collection is only part of that process; the search patterns and thought patterns that assist and enhance the formulation of ideas must be protected. I ran into this problem in trying to decide how to handle subject annotation in my work. I was concerned that the proliferation of subject annotation (which the technology would permit) would allow the system to refine procedures of exclusion to the point where they would overly simplify search patterns and impoverish the overall research experience. Research is a learning process and not just the mechanical accumulation and arrangement of information. In the humanities, at least, the structure of developing an understanding of a subject is not a linear progression; it just appears that way because it unfolds in time. As you begin, there is obviously a large body of information that clearly is

not pertinent to your concern and there may well be an easily identifiable body of information that clearly is pertinent to your concern. But in addition, there is a significant body of information that 'may be' pertinent to your subject and only you as the writer / thinker / researcher can decide the nature and degree of pertinence in each individual case. The process of evaluation and discrimination that takes place here is not simply a research requirement or scholarly necessity; it is an integral part of the growth of the learning process that underpins the intrinsic value and the success of that project as a whole. I believe that anything that minimizes scholarly involvement in assessing the 'may-be-significant' area of information-gathering is detrimental to the intellectual quality of the final result, and insidiously so. Mediocrity is hard to diagnose. Scholars are human and tend to take the line of least resistance without thinking too much about how their choices affect final outcome. They adapt themselves and their work to the circumstances of information accessibility as a kind of objective 'given' shaping the conditions under which they pursue their scholarly work. I am sure that there are scholars who wish in their hearts that more documents and information could be arbitrarily destroyed, lost, or made inaccessible, because it would simplify their work enormously. The cynical pragmatism that fosters that wish (and we are all touched by it) sees computerized searching as a release from some of the onerous but necessary burdens of scholarly research rather than as an instrument that allows for wider and more thorough searching. There is a human tendency to relegate an inordinate part of the responsibility for the job you are doing to the tool that you are using, especially if that tool is new and apparently complex. Such thinking inevitably enhances the apparent importance of process as an end in itself and diminishes the intrinsic value of result.

When I apply this observation to the use of computers for bibliographical searches, I begin to worry that the process itself will alter and oversimplify important features of humanities search-patterns. Such simplification could enhance the probability of producing more scholarship, but scholarship of reduced or inherently mediocre intellectual quality. Because of this concern, I was reluctant to multiply or refine the subject annotations in the periodical database. I felt it was very important that the scholar be required to throw a wide net so as to gather in, not only the pertinent entries, but also the widest range of entries that 'may-be-pertinent.' The responsibility for analysing, evaluating, and deciding is then fixed where it should be – in the hands of the researcher. If, in our enthusiasm for the systemic process, we cater to the weaknesses of the user, we may well contribute to the decline of the quality of the intellectual insights we wish to nurture and serve. When integrating computers into your work,

keep in mind that as bibliographers and indexers your first responsibility is to the quality and character of the results of your work, not to the efficiency of the systems you employ. If you make a bomb, you must take responsibility for the results no matter who drops it. If you make a bibliography or an index, you must to some degree take responsibility for the results that flow from it. Never underestimate the importance of your work and its effect on the character and quality of Canadian scholarship.

The mundane realities of the profession, however, are fraught with frustration and you should be aware of some of the irritants associated with developing computerized databases. After you have struggled to master the intricacies of your computer tool, after you have carefully designed your bibliographical structure, after you have gathered and loaded all the information and got it up and running, you will find that there is a serious logistical problem relating to the dissemination of your work, if you wish to keep it in an electronic medium rather than transfer it to a printed medium. Ideally, in order to retain its full utility, you would like to set up your database in a central automated repository which could be reached by potential users over toll-free lines. However, unless your information happens to have commercial or scientific value, you will not find a public or private agency interested in it. There is no central repository in this country for databases, such as those in the humanities, which require some form of subsidy for their establishment and continuance. There are some other, less-than-ideal, options, but none of these has been organized for the effective dissemination of this kind of material. In fact, there is not even a reliable clearing-house where you might find out what has been done, what is available, and what is currently being worked on. In short, these are significant logistical obstacles inhibiting the development of computerized bibliography in fields where traffic in information-access is not commercially viable.

But even as you begin, there are frustrations that should be recognized, and you should be prepared to exercise patience and to invest a significant amount of time in coming to terms with this new medium. In the first place, you are faced with a confusing and intimidating array of various packages of database software that you might employ in your work. It seems unfair: after just having made the emotional leap of faith to try this stuff, you are confronted with an important decision which seems to require a fair amount of technical experience. Consider it a test of your commitment: talk to people, whine, complain, get mad. It is important to develop a healthy disrespect so that the technology is always proving itself to you and not you to the technology. Have a very good idea of the job you plan to do, and judge the capabilities of the software in the light of what you want it to do. Do not let the software define your job; require the soft-

ware to meet your conditions. Be true to your bibliographical instincts.

While it is not my place to recommend specific software, I would like to point out some things you might keep in mind if you are shopping around. Be sure that the software you choose allows you to dump or remove the information you put into it in a form that allows you to transfer it to another brand or type of database. In our business, portability and transferability of data are extremely important. If we are going to work together, our systems must be able to use each other's information. Most respectable software packages have this capacity, but ask about it. Second, look at what is involved in defining the field structure for the databases you are interested in building. In some cases, the field structure is already defined and you just fill in the information. In other cases, you define the structure yourself to suit your needs. I prefer the latter, but I demand that the operation of defining the field structure be simple and straightforward. You should not have to take a course in database theory to build a database. Third, I expect the procedure for entering information to be even simpler and more straightforward. You should be able to assign the task of 'inputting' to someone who knows nothing about how databases work beyond the act of entering data. Fourth, if you expect your database to grow to a significant size, you should think about response-time, that is, how long it takes the software to retrieve answers to your queries. Speed has a lot to do with the power of the actual computing system you are using (the hardware), but it is also related to the way your database software programme is built. If the number of records in your database will likely exceed one or two thousand, you should look seriously at the question of response-time. Finally, but in many ways most importantly, you should be sure that the software you use allows you to print out (or put up on your screen) not simply the information you want but the information you want in the form that you want it. The software should have a very flexible report-generating capacity, one which allows you to present information in appropriate and recognized bibliographical formats. It is a matter of bibliographical professionalism and is related to the intangibles of confidence and trust in your work. Sometimes, those who write database programmes do not appreciate the importance of this requirement, so you must demand it of them.

The final thought I would leave with you is this: the computer is simply a tool designed for certain kinds of work. It bears the same relationship to human endeavour as the chisel, the pencil, the hair brush, the chainsaw, the typewriter, and the blender. Like all tools, you learn how to do things with it, then you take it for granted and get on with the business of creating what you intended to create in the first place. Like all tools, its sole purpose is to facilitate human creative intention in turning idea into object. If

I felt for a moment that computers restrained or constricted human creative endeavour, I would urge you to kick their little screens in.

NOTES

- 1 Further information on any books of poetry or fiction which are 'not in Watters' would be greatly appreciated.
- 2 This group forms part of the Early Canadian Cultural Journals Database project at the Royal Military College, Kingston, Ontario.
- 3 ECCJ Database project.
- 4 ECCJ Database project.
- 5 Five of these contents-reports-cum-indexes have been published as departmental monographs by the Department of English and Philosophy at The Royal Military College. They include *The Nova-Scotia Magazine, 1789-1792* (Kingston, 1982), *The Acadian Magazine, 1826-1828* (Kingston, 1982), *The Provincial, or Halifax Monthly Magazine, 1852-1853* (Kingston, 1982), *The Amaranth, 1841-1843* (Kingston, 1984), and *The Canadian Magazine and Literary Repository, 1823-1825* (Kingston, 1984).