

A Computer Database System To Store and Display Archival Data on Correspondence of Historical Significance

BRUCE R. WHEATON

FOR THE PAST THREE YEARS the Office for History of Science and Technology at the University of California, Berkeley, has systematically collected data on correspondence preserved in archives and libraries worldwide related to physics and physicists active in this century. The work is part of a larger project supported by the Division of Research Tools of the National Endowment for the Humanities and the Program in History and Philosophy of Science of the National Science Foundation to document interactions among physicists and with representatives of other arts, humanities, and professions.

In 1983 we will publish an inventory of the letters identifying author, recipient, location, and approximate date of each letter. For this purpose we have formulated a computer database system into which data may easily be entered in arbitrary order and from which relevant information may be extracted in a variety of more useful orders. In particular the system will generate a fully cross-

referenced list arranged by correspondent of letters exchanged with each physicist and at the same time will fully identify the archives and collection where the items are presently located.

Our inventory now records half a million letters received or sent by 5,000 physicists. This is about ten times our original maximum estimate of what we would find even though we excluded most collections of institutional records of university physics departments. Before data entry the information filled some 150,000 index cards. A project this big demands a computer database management system. I am indebted to Virginia Aldrich, Paul Deuter, and Stewart Scofield, who wrote initial versions of many of the programs involved in the computerization of our inventory.

A project of this magnitude should not produce or rely on systems peculiar to it; we realized from the start that whatever system we established should be universally applicable to other projects undertaken to characterize collec-

tions of correspondence. Thus our computer system should be of interest to archivists who wish to combine on-line all finding guides to collections under their care, and to persons planning inventories of correspondence collections on any topic or in any time period. Put most plainly, we provide here a complete system to inventory useful descriptive information on correspondence, whatever the subject or period of the exchanges. We provide a systematic approach to all aspects of the inventory procedure: initial tallying of items in an arranged collection, entry of that information to a computer database, manipulation and correction of the on-line data, and formatting for output of data in any of several useful orders. One, but by no means the only, use of the system could be to produce a fully cross-referenced listing of letters from or to a named correspondent in all manuscript collections at a given archives or library.

ISHTCP

The Inventory of Sources for History of Twentieth-Century Physics began in an attempt to organize knowledge about correspondence collections worldwide that would aid historians, sociologists, philosophers, and scientists interested in researching the development of the various subfields of physics since 1900. There are several inventories giving brief descriptions of papers of individual physicists, among them the *National Union Catalog of Manuscript Collec-*

tions, the publications of the Contemporary Scientific Archives Centre in Oxford, Eng., and Schlawe's compendium of German letter collections.¹ However, these and similar sources do not give a complete list of correspondents represented in the collections described. Although very useful to one who wishes to study the particular individual whose papers are preserved, these inventories do not help a researcher searching for letters from others that he or she might expect to find in that collection.

With support from the American Philosophical Society and the American Physical Society, a systematic search begun in 1962 turned up several thousand letters and manuscripts written by physicists active in the development of the quantum theory of the atom.² The fully cross-referenced catalog of these letters markedly aided research on the history of this subfield of physics. Since then the number of relevant collections has increased, and the relatively detailed descriptions that exist individually prompted us to undertake a systematic search for all subfields of pure physics.³ Our goal was and is to publish a cumulative cross-filed listing of all collections by physicist, by correspondent, and by physical location or archives.

Our intent is to provide specific information on the availability of firsthand historical source material. Historians, for example, require not only identification of correspondence of interest but also sufficient indication (apart from

¹Fritz Schlawe, *Die Briefsammlungen des 19. Jahrhunderts*, 2 vols. (Stuttgart: Metzler, 1969). Other useful sources for history of physics collections are: R.M. Macleod & J.R. Friday, *Archives of British Men of Science* (London: Mansell, 1972); Ludwig Denecke, *Die Nachlässe in den Bibliotheken der Bundesrepublik Deutschland* (Boppard am Rhein: Boldt, 1969); *Gelehrten- und Schriftstellernachlässe in den Bibliotheken der Deutschen Demokratischen Republik* (Berlin: Staatsbibliothek, 1959); various volumes of the *Trudy* of the Archives of the Academy of Sciences of the U.S.S.R.; and J. Warnow, *National Catalog of Sources for History of Physics* (New York: American Institute of Physics, 1969).

²T.S. Kuhn, J.L. Heilbron, P. Forman, and L. Allen, *Sources for History of Quantum Physics. An Inventory and Report* (Philadelphia: American Philosophical Society, 1967).

³For the limits of the search see the 7,000-item bibliography by J.L. Heilbron and Bruce R. Wheaton, *Literature on the History of Physics in the 20th Century*. Berkeley Papers in History of Science, vol. 5. (Berkeley: OHST, 1981).

specifying content) to judge from a distance whether the items should be procured or consulted. A historian usually is familiar with names of individuals and institutions, and knowing the names of the two parties of an exchange will indicate potential relevance. We decided therefore that it was impractical and unnecessary to indicate content of letters; indeed we discovered very quickly that to attempt to do so was extraordinarily time-consuming (multiplying by a factor 3 or 4 the time needed to characterize an exchange) and was subject to grave ambiguity. Numerous categories had to be established *a priori*, and subject classification of the same letters varied widely depending on the person who classified them and the particular interests he or she brought to the task.

Recognizing that users familiar with the general course of development of physics in our period of concern could infer general subject from the date and names of the exchange, we limited our specification to: 1) the correspondent's full name; 2) the date(s) of letters; 3) whether the letters are from or to the correspondent; 4) the number of pages involved if a photocopy or microfilm record were to be requested; 5) the name of the collection in which the letters reside; and 6) the name and address of the library or archives to contact. In special cases, where the archives does not maintain an amalgamated catalog of letters and the collection is not arranged by correspondent, an additional reference to box and folder number within the cited collection is needed.

It became clear to us from the start that the number of items to be described, even in our relatively limited domain of modern physics, is too great to allow individual dates to be included. Nor is it necessary to provide them. A user sufficiently familiar with the subject can

deduce from correspondents' names and approximate dates of letters whether their correspondence might be of interest. Accordingly we divided the period of our concern, 1896–1952, into nine year-intervals, each approximately five years long and demarcated by events of major conceptual importance in the development of physics. We report the number of letters sent by each physicist and, separately, those sent to him, in each of these year-intervals.

Data Collection

To facilitate accumulation of data in a consistent form we printed index cards with spaces for the data we seek. A system was established to ensure that needed conventions were followed in filling out the cards. The success of the data collection system in maintaining consistency allowed us to use the completed cards as data-sheets for key-entry of the information. Our basic inventory card is reproduced as Figure 1. Each card records all information concerning an exchange of letters between two individuals that is contained in a single collection at an archival repository. This makes it possible for us to intermesh information on all correspondence with a given physicist regardless of the collection in which it is found.

Referring now to Figure 1: at position A we give the physicist's name and at B the name of his correspondent. Letters written by the physicist are counted in the first vertical column, C; the approximate number of pages counted appears in column D. Note that both numbers are under the right-pointing arrow that indicates the direction of the letter. Similarly, letters from the correspondent to the physicist are counted in columns G and H. Columns E and F are used only in the event that the library has cataloged the documents without specifying who is author and who recipient,

<u>Einstein, Albert</u> ^(A)		<u>Freud, Sigmund</u> ^(B)	
Physicist		Correspondent (FULL name)	
pre 96	(C)(D)	(E)(F)	(G)(H)
96-99	/	/	/
00-06	/	/	/
07-13	/	/	/
14-21	/	/	/
22-25	/	/	/
26-31	3/4	/	5/12
32-38	/	/	/
39-45	/	/	/
46-52	/	/	/
post 52	/	/	/

Archive code:
(I) USC a Ø4

Collection: P C N
(K) Planck papers

Microfilm copy:
(L) 4, 3

Add'l reference:
(M) See Pauli correspondence, box 10, folder 3
[There are more letters than these]

SPECIMEN

Figure 1. Data entry card

e.g. "five letters exchanged in 1941." Notice that in addition to the nine-year intervals of the inventory, there are general intervals (pre-1896 and post-1952) for the period before and after those of direct concern. Undated letters are entered in an unused line relabeled "NO-DA."

To identify the archives or library we use an eight-character code, the central six characters of which are shown at entry I: a leading symbol identifies privately-held papers, two identify the country, two the city, a two-digit number specifies the archives, and a final character identifies the letter collection in the event that the archives separates its collections by name of individual. Because the collection is frequently that of either the physicist or the correspondent, we make provision in section K for either to be specified by circling "P" or "C". "N" indicates that the archives catalogs all collections together, so that no specification of col-

lection is needed. Otherwise the collection name is entered in space K.

We keep a record of microfilm copies of the exchanges we report. This is frequently helpful to a potential user who otherwise might have to travel to consult letters. Reel and section number(s) are entered at L. If an additional reference to box and folder number is necessary, it is put in M.

The Database

After conducting tests of several computer systems available to us we selected INFOS, developed originally by the Data General Corporation to handle business inventory and billing records. Like any database system, data records that contain the desired information are each associated with one or more keys. A key is a characteristic address that leads directly to the associated data record; in INFOS the keys may be of any pre-established length. The same record

may be addressed by several different keys, each occupying its place in a different sequence of keys. The usefulness of this way of proceeding will become clearer as we discuss our particular system.

INFOS provides multilevel key definition of fixed-length or variable-length records and will access records either randomly or sequentially, or will begin a sequential read at a random location. Furthermore, INFOS is fully compatible with a widely known programming language, COBOL. All of our procedural programs for data entry, for manipulation and correction of records, and for producing backup files of records are written in COBOL. All of our records are of fixed length (87 characters). Our experience has shown that having all records of fixed length greatly simplifies their manipulation by available sorting and merging utility systems. Each different record type is identified by its leading digit and each is accessible by one or more index keys.

We create several different types of records. Central to the inventory are the three record types that contain the data from the index cards:

—The “line” record contains the basic information on numbers of letters and numbers of pages within one of our selected year-intervals, that is, one data line (entries C through H) on an inventory card. The record consists of: part of the physicist’s name and part of the correspondent’s name, identification of the year line and year interval that this record reports, six three-digit numbers specifying the numbers of letters and pages, and the date and time that the record was entered.

—The “card” record encodes the identity of the archives and collection and contains: part of the physicist’s name and correspondent’s name, full archives code with collection code assignment, 18

characters of microfilm identification, a single character that indicates whether there is or is not an additional reference record for this card, and the date-time of entry.

—The “additional reference” record contains the physicist’s name, date-time, and up to 64 characters of additional reference for the card.

These three record types are linked to one another by having virtually identical keys, each containing part of the physicist’s name and the date and time that the records were created. Therefore, all records that together comprise an inventory card—which consists of at least two and as many as fourteen 87-character records—may be fully retrieved using the key to any one of the records. In practice we use the “card” record for this purpose, the one that identifies physicist, correspondent, and archives code. There is one of these records for each card. Having found it, one has all the information needed to define the keys of, hence to find, all associated “line” records—each of which contains the data specifying numbers of letters and numbers of pages exchanged in one of the year-intervals. The card record also leads directly to any associated additional reference record.

The interactive format used for data entry requires other types of records. We started with a sub-index of 5,500 records, each containing a physicist’s name and biographical data. This is an authority list to standardize spelling and form. These records, which allow a full 40-character field for each name, were entered before any entry of data cards began. When entering inventory cards, the key-entry operator (KEO) only selects the physicist he or she wishes to use from the on-line authority list; physicists’ names are never entered directly by the KEO. On the other hand, the correspondent’s name is entered by

the KEO. Each newly added correspondent's name produces its own 40-character field within a new 87-character correspondent record, and every subsequent name is compared to the growing correspondent sub-index to see if it is already there. The operator then has the option to use the old name without creating a new correspondent record.

Each collection code established for a given archive produces an entry in the sub-index of archives and collection code assignments. The code for archives and collection on all subsequent cards is compared to this sub-index to see if the proper assignment already has been made or whether it needs to be made. The interactive searches of previously entered data maintain consistency and prevent, for example, creating an unnecessary duplication of correspondent name records. The ability to make these searches during data entry is one of the advantages of an interactive database entry system.

Keys

It is in the structure of keyed sub-indexes that the value of the database approach is realized. Each different type of key defines a separate sub-index. And each sub-index defines, in effect, another copy of the entire set of records. Each record is accessible through one or more keys; by suitably defining the key, the order in which records are read may be changed. There is no requirement that the keys contain the same data as the record, but this is frequently the most useful way to assign them. Our keys are of differing length, from 8 to 51 characters; the key length within one sub-index is constant. Given a full or partial key, the search for its associated record is optimized for efficiency by the database management software. It uses a binary search, meaning that it starts

with the central record in the sub-index and successively eliminates half of the remaining index per attempt. The search takes no more than seconds even for the large files we now have assembled. Our database, the Inventory of Sources for History of Twentieth-Century Physics, occupies 30 million characters (bytes), the associated index structure 60 million.

Four independent keys lead to each card record, and each card record then leads through common keys to the associated line and additional reference records:

—The "PP" key reads records by physicist's name, correspondent's name, and archives code, in that order. This key specifies the order in which the final inventory itself will appear: arranged alphabetically by physicist's name, with data filed under each name alphabetically by name of correspondent, and for each correspondent chronologically by date interval.

—For diagnostic and backup purposes it is necessary also to key the records by date and time of entry. The "P2" key reads records in order by date and time of entry, physicist's name, and correspondent's name.

—In order to find records by name and to cross-reference exchanges between two physicists, it is necessary to be able to read records according to correspondent's name wherever in the alphabetical sequence of physicists the records belong. There is accordingly a key called "PC" that accesses records according to the sequence correspondent's name, physicist's name, archives code.

—Finally, to allow generation of lists of material reported from a given collection or from a given archive, there is a key called "PA" that uses the sequence archives code, physicist's name, correspondent's name. Figure 2 shows schematically the ways that data may be recovered using the major keyed sub-indexes.

In all four of these modes of access

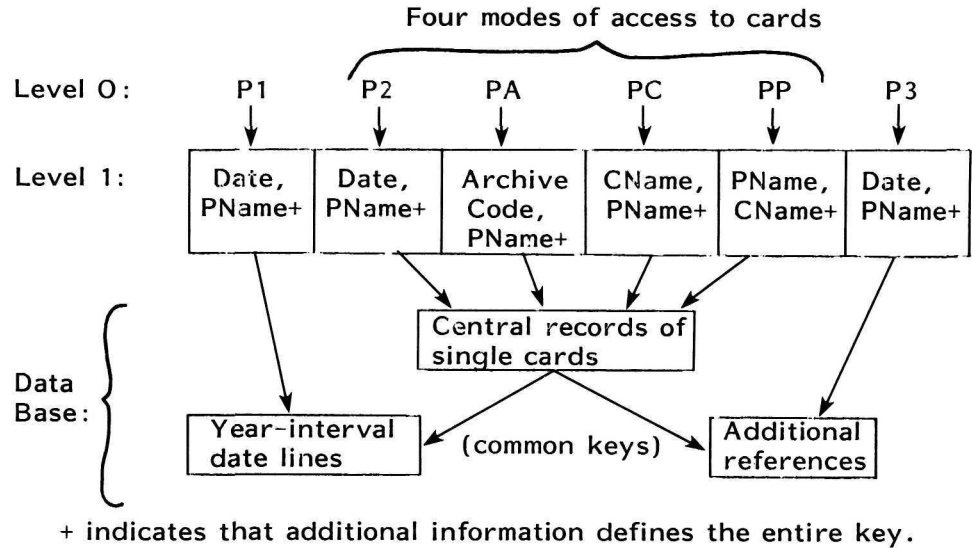


Figure 2. Logical hierarchy of part of inventory database

one may specify all names and codes and go directly to the card of interest. Or, if some categories are left undefined, the database may be read sequentially from the starting point in the order defined by the key used. It is here that the greatest power of the database approach is exploited; a search of the index cards themselves for all entries listing a given *correspondent* or *archives* is simply not practical by hand. Even a search by physicist's name can only be conducted if all cards are carefully filed in order. The computer database is, on the contrary, oblivious to the order in which cards are entered.

With these four keys to define access paths in the database, we can extract information in several useful ways. Simple utility software for sorting and combining the data record files that result allow us to format the desired data in virtually any manner we wish. The large published inventory will be produced in this manner, the resulting tape of line-printer output being directed to existing hardware that produces 250-page microfiche records. We will then publish conven-

tionally an accompanying guide and name index; the index itself will derive from the machine-readable sub-index of correspondents' names created during data entry.

Key Entry

One of the most difficult problems of such an undertaking as ours is to ensure consistency of data entry format while maintaining accuracy. To achieve this goal we created a screen-formatted COBOL program that takes most of the ambiguity out of the key entry operation. In this way we have been able to use relatively unskilled part-time student employees to input data directly from our basic index card records. Data entry is accomplished on a CRT terminal with sufficient internal logic to control screen formatting command characters.

The format of the screen is roughly that of the data card itself; the congruence in form allows quick and accurate corroboration by the operator of entered data. At several places in the main entry loop, the operator is asked to verify that all entered data are correct

Inventory Card

Physicist	<->	Correspondent
Einstein, Albert		Freud
Right Physicist? <u>y</u>	Closest Corr.:	Freud, Sigmund
(Y,N,F,B,L=accept)	Lower or U=add upper name to CAL)	Which correspondent? <u>L</u>
		So far so good? <u>Y</u>

A
B
C

Figure 3. Interactive screen display for names. "CAL" is the correspondent authority list.

before continuing on to the next section of a card. The main loop is subdivided into three sub-loops: one to ensure correct physicist's and correspondent's names, one to verify that numbers of letters and pages are accurate, and one to verify that the proper archives and collection identification has been given. Figures 3, 4, and 5 show the stages in screen orientation for an entered card; Figure 5 comprises the full screen and should be compared to Figure 1, the data card itself. All underlined entries in the figures are additions by the KEO; the rest are items provided by the program.

The typical sequence of events for entry is as follows. After completing a card, the cursor waits in position A (Figure 3) for a response by the KEO. If, as is usually the case, the next card is to be for the same physicist, the response "Y" accepts the previous name and moves the cursor to the field for entry of the correspondent's name. But the physicist can be changed by typing "N" and then overwriting the physicist's name. A new name is never entered on records but is used to search in the physicist authority list, and the closest match is displayed. The KEO may then search forward or backward in this sub-index by responding "F" or "B" until the proper name is found. "Y" then moves the cursor on to the correspondent name field.


After entering the correspondent's name at B, the existing correspondent

name sub-index is searched for the closest match and that is displayed under the typed-in name. The KEO may then choose either upper (new) or lower (old) name for the records. A first sub-loop check at position C allows any part of the loop to be repeated. Once the check has been passed by a "Y" response, the names are written to the corresponding parts of the database records.

At this point the cursor enters the field of year-interval lines. The standard intervals are given as default values, and we try not to deviate from them. The dates may be changed if necessary by rewriting them. Of course, were the system used for some other subject or period, the default year-intervals would be modified permanently within the program. To save time the KEO is given the option at the beginning of this sub-loop to skip to a line below as specified by the first command. Otherwise each relevant line is completed with up to six three-digit numbers. See Figure 4.

As each line is completed and accepted, the numbers are automatically right-justified. The bottom line is for undated letters. The "+" in the exchange column of the "NO-DA" line indicates that additional letters are known to exist in this exchange but that we lack specific information to classify them. Following completion of numeric entry, or after use of the special key command to leave numeric entry mode, the KEO is again prompted for acceptance of all

Physicist	Inventory Card	Correspondent
Einstein, Albert	↔	Freud
Right Physicist? <u>y</u> Closest Corr.: Freud, Sigmund (Y,N,F,B,L=accept lower or U=add upper name to CAL) Which correspondent? <u>L</u>		
		So far so good? <u>Y</u>
-----> Exchange <-----		
---95	/	/
96-99	/	/
00-06	/	/
07-13	/	/
14-21	/	/
22-25	/	/
26-31	<u>3/</u> <u>4</u>	<u>5/</u> <u>12</u>
32-38	/	/
39-45	/	/
46-52	/	/
53-++	/	/
NO-DA	/	+/






Figure 4. Screen display for names and numeric entry

Physicist	<-->	Correspondent
Einstein, Albert		Freud
<hr/>		
Right Physicist? <u>Y</u>	Closest Corr.:	Freud, Sigmund
(Y,N,F,B,L=accept lower or U=add upper name to CAL) Which correspondent? <u>L</u>		
So far so good? <u>Y</u>		
<hr/>		
----->	Exchange	<-----
---95	/	/
96-99	/	/
00-06	/	/
07-13	/	/
14-21	/	/
22-25	/	/
26-31	3/ 4	5/ 12
32-38	/	/
39-45	/	/
46-52	/	/
53-++	/	/
NO-DA	+/-	/
all values OK? <u>Y</u>		
Same Archive data as on last card? <u>N</u> Archive Code: <u>USCa04</u> NOT IN DATABASE Collection code: <u>a</u> Add to DB? <u>Y</u> Name of Coll.: <u>Planck papers</u> Microfilm: <u>Reel 3, section 4</u> Want to enter additional reference? <u>Y</u> Ref: See Pauli correspondence, box 10 folder 3 OK? <u>Y</u>		

Figure 5. The complete screen display for an inventory card

displayed data at position D (Figure 4) and given the opportunity to loop back through the field to make corrections. If there are no corrections to be made the line records are written, one for each line that contains numbers.

The final sub-loop in data entry constructs the basic card record. The archives code and collection identifier from the previous card remain on the screen as default values at position E (Figure 5) because cards are frequently entered from the same source in succession. The KEO has the option to copy all of the data needed to define a card from the previous one or to rewrite this information. Given a new archives code, the sub-index of previously assigned collection identifiers is searched and each identifier for collections at that archives is displayed in sequence at position F on operator request. If the needed code is found, that can be used; if it is not, the KEO has learned the next sequential letter to assign. The KEO may then add that letter at position G, suitably identified at position H, and the code with identifier is added to the sub-index.

The KEO then has the option to enter at position I a reference to a microfilm copy of the documents described. This information is entered in the basic card record. The KEO also has the option either to create or not to create a record to hold the additional reference. Usually it is not necessary, but if it is, up to 64 characters may be included at K. We have not imposed a fixed format on the additional reference, as it varies greatly in form from one collection and archives to another.

Finally the KEO is asked if the entire card is correct. Possible answers are "Y," in which case all records are written and the cursor moves back to the top of the screen, and "N," in which case the entire third sub-loop is begun again. If the KEO wishes at this point to discard

the entire card he or she may answer "A" and all card-related records stored in the buffer are erased without being added to the database. The correspondent-name and collection-code assignments are written to their respective sub-indexes in any event.

Backup Systems

Because the entry procedures require that data be written to the growing database, there is a small but not negligible probability that a system failure will destroy the integrity of the entire database and index structure. For this reason an extensive system of backup procedures has been established. The entry program itself makes a second copy of all entered data even in the event of computer malfunction that damages the database. A provision is also made to create a backup file of each day's entries; for this purpose, as well as for diagnoses of the database, all entered records include the day and time of entry. Records added on a specified day are easily and swiftly retrieved using their sub-index keys.

Backup files are retained until the next system dump to tape. Thus the worst possible case in the event of a system failure requires no more than loading the most recent tape copy and restoring the more recent data from the relevant backup sequential files. In addition, the backup files can be used to retain a permanent copy of the data that does not require the memory overhead of the database structure and indexes. We can, for example, send a tape of the sequential files to another computer system where it can be used in a variety of ways to provide inventory data. It can also be used as data input to a database management system entirely different from INFOS if this is more convenient at the remote site.

Our use of INFOS allows us to search the database while we are entering new data; this maintains consistency and accuracy. But use of the resulting backup sequential files is in no way dependent on INFOS or, indeed, on a database management system at all.

Several other systems are designed to manipulate the database in one way or another. It is occasionally necessary, for example, to change a designated archives or collection code. There is a COBOL program capable of finding all instances of the erroneous code in records and keys and of changing them to a specified correct version. We also provide means to alter the spelling of a correspondent's name in all relevant records and keys. And we have devised means to invert certain records when the correspondent is recognized to be a physicist. Any single data card reporting correspondence between two physicists will exist in two copies in the database and will be reported under both names in the printed inventory.

There is a modified version of the entry program that calls individual cards to the screen and allows corrections to be made on all records that comprise that card. We also have programs that will extract all cards for a given physicist, or for a given correspondent, or all cards in a given collection or at a given archives. Simple sorting procedures may then be applied to the resulting files to put the information in any desired format.

Other Applications

We wished to design the computer-system as generally as possible so that it might be applied to other purposes. That we developed it for 20th-century physics is in no way limiting. With no essential modification the inventory system can be used to catalog and retrieve informa-

tion on published and unpublished correspondence from any time period and in any subject. Even the present restriction to reporting numbers of letters in short year-intervals can be removed. For the purpose of compiling an inventory of physicists' letters that have been published, we have developed a modified entry and database system that records dates of individual letters.⁴ All of the capabilities of the original system are retained; we use the same backup program, can alter spellings of names at will, can change bibliographic references (which here take the place of archives-collection codes), and can access the data in identical ways. Of course the size of the resulting files is, for the same number of documents, some three to four times greater. The output of this aspect of our project—the inventory of published letters—can be encoded so that it may be directly and electronically set in type and printed as a book.

Nor do potential applications need be limited to records of correspondence. The binomial specification of records allows its use for systemization of any data by name of individual and subject, for example to classify records related to a teacher and his or her students. With abbreviated titles, or with a small change in the size of the records, one might use the system to inventory authors with their book titles, writers or composers with manuscript titles, or painters with paintings.

The system as it stands, whether to record numbers of letters in time intervals or to specify individual dates, might usefully be employed by a library wishing to compile an amalgamated catalog of the correspondence it holds in its manuscript collections. It is not necessary to complete the index cards as key entry records if one can work from

⁴The 25,000 references to quotations of letters appear in Bruce R. Wheaton and J. L. Heilbron, *An Inventory of Published Letters To and From Physicists, 1900–1950*. Berkeley Papers in History of Science, vol. 6. (Berkeley: OHST, 1982).

existing inventories of collections. The input catalog need only include correspondent names and letter dates in a compatible arrangement. We require the cards for two reasons, neither of which need affect work at a library already possessing finding guides: first, our data is assembled from more than 1,500 libraries and archives and there is no standard in cataloging letter collections. Furthermore much of our reporting is done in the field on uncataloged collections. Our cards were initially designed as a swift means of cataloging collections; they only subsequently became key-entry forms.

For our uses to record letter collections we have found that we are able to enter about 100 complete cards (that is, all data on 100 exchanges of letters) per operator hour. The fixed costs are key-entry operator time and equivalent computer time-sharing costs. If the relevant information concerning a manuscript collection is already available in key-entered form, whatever its format, a conversion program can be designed to rewrite that information into a format that would allow it to be directly added to the inventory database. If a library has already invested in the entry of data, use of our system would not require it to be reentered. Each entered card generates on the average 300 bytes of record data and another 1,200 bytes of index structure. If the database approaches the size of ours (150,000 exchanges) it will occupy a sizeable frac-

tion of a 200-megabyte disk-pack; it would be necessary to ensure in advance availability of sufficient computing space to undertake projects of great magnitude.

But the advantage of such a system, even if used only to record exchanges of correspondence, can hardly be overemphasized. A library that had key-entered inventory finding aids from all of its manuscript collections would have an invaluable resource on-line for its staff and users. One can specify locations of documents within collections to any degree of accuracy down to box, folder, file, even sheet number. Unavoidable cross-references may be usefully retained; a letter from A to B that is filed under name C in the papers of D is simply and directly recorded.

Library staff needing to know whether any collection at the library contains letters from or to any specified individual may have that data immediately. Users wishing to know about all material complementary to that in one collection may in this way search effortlessly in all. Better control over the content of manuscript collections may be maintained. And best of all, because the database system is oblivious to the order in which data is entered, new data may always be added in sequence without reentering any of the old. With proper systems management the initial investment in putting data on-line is never lost, and over the years of expanding use will repay itself many times over.