BRACERS (III): SYSTEM REVIEW AND QUERYING THE TOPIC_TEXT FIELD

KENNETH BLACKWELL
Russell Archives / McMaster University Library
Hamilton, ON, Canada L8S 4L6

I. BRACERS' PRESENT STATE

The Bertrand Russell Archives Catalogue Entry and Retrieval System has amassed a textual database of 60,400 records describing Bertrand Russell's correspondence.[1] At the time of writing over 59,000 had been proofread, revised and corrected by specialists. Supplementing the formal descriptions of letters are 42,000 lines of text, précis and commentary in the Topic_Text field. It is impossible to say in advance of BRACERS' completion how many letters are preserved in the Bertrand Russell Archives, but my *feeling* for it is that these records represent close to 70 percent of the correspondence.[2]

The initial system, initiated in January 1988 by the Director of Archives and Research Collections and devised by Roland Carter of McMaster's CIS, was finalized in early 1990.[3] A major analysis of the system was undertaken in February 1992.[4] Later that year I published a users' guide to the system, called *Searching BRACERS*. In January 1993

[1] For an early progress report, see my "BRACERS: the Bertrand Russell Archives Catalogue Entry and Retrieval System", *Russell*, n.s. 10 (1990): 158–64.

[2] Such item-level descriptions are not covered in standard archival computer manuals, and much of what follows had to be invented independently (of RAD, for example).

[3] At the same time there was a shift in software that affected table integrity. Up to 1,000 records were lost (the exact number is undetermined, because the old system used different record numbers for additional Topic_Text lines).

[4] "Maintaining Progress in BRACERS" (internal report).

russell: the Journal of the Bertrand Russell Archives
McMaster University Library Press
n.s. 15 (summer 1995): 65–72
ISSN 0036-01631

the system was revised to incorporate enhancements suggested by the staff working on it. A new system manual was written.[5] The previous year, as we reached 40,000 records, the database became, in practical terms, the best place to answer queries on archival holdings.

BRACERS is more than a catalogue. Like many other databases born as library catalogues, with the advent of keyword indexing, or at least the ability to search any and all fields, it has become a research instrument. As a database of all relevant data about Russell's letters, it caters to subject, textual, biographical and bibliographical research as well as archival control. Typical uses are to answer queries on whether certain correspondents are represented (now easily answered by consulting a BRACERS list on the Russell Archives' Gopher and Web sites), what is available in a given time period or on a given date, and whether certain topics are discussed in the letters—besides some letter texts, my notes on the Alys Russell, Morrell, Malleson and Dictation files (to April 1958) are included. Most questions can be answered by using the ready-made queries, and browsing access has been granted to some researchers. These queries are used also by input staff and revising staff to produce daily printouts of input and revision and to check data consistency. Users desire a printed report; less often, an electronic version. I programme reports of any size as delimited ASCII datasets for FTP'ing to the Internet and importation into database or wordprocessing programmes.

Although the opportunity to have BRACERS is invaluable for the Russell Archives in control, bibliographic and topical and textual access, and as a research tool in its own right (see the results of the address field referred to in fn. 5), the present software is short of ideal. All input becomes upper case. No accents or other unusual characters are available. There are two tables, PLBRCO#.RA_RUSSELL_ARCHIVE and .RA_TOPIC_TEXT. The Topic_Text field in the latter table is composed of up to 29 individual lines (actually records) per related record.

The ideal software would be able to retrieve entire individual letter records and create summaries of letter groups. Retrieving the former is at present cumbersome; the latter cannot be done in BRACERS. With these limitations and those affecting searching set out in section III below, BRACERS is invaluable to Russell scholarship and archival control.

[5] [Ruth Toth], "Research Collections System and BRACERS".

II. BRACERS' FIELDS

In-depth users of the BRACERS database should understand the fields. Some of the names are cryptic, and it not always easy to tell how much data a field may hold. A description of each field follows.

COLL_CODE (6 characters): The Collection_Code field originally allowed our records to interface with other records of the Ready Division. Now it records whether a document is in Russell Archives 1, 2, or Recent Acquisitions, identified as RA3.

CLASS_NO (5): The three-digit classification number assigned to all record series in RA1 and 2. For RA3, the Recent Acquisition number is used. It goes as high as a five-digit number when the alpha suffix is included. Leading zeros are used for the lower numbers in RA3 and all numbers in RA1 and 2.

DOCUMENT_NO (8): The six-digit, unique, incrementing number assigned to every document in RA1 and 2. Some document numbers in RA2 are also followed by one or two letters of the alphabet. Thus the total field is eight characters. For documents in RA1 whose folders do not contain their numbers, this field is left blank for the time being. For RA3 this field is filled in only if document numbers have been assigned (e.g. the Constance Malleson papers) or letter numbers assigned (e.g. Ottoline Morrell letters). The first six characters must be filled in when the field is used.

BOX_NO (4): Used only as a guide to retrieve documents, e.g. oversize or overlength RA3 items, where boxes have been numbered. Otherwise the field is left blank.

SOURCE_TEXT (21): Used in conjunction with RA3, it indicates an item's provenance. If the source is an institution, we aim to use the ALA list of institutional abbreviations. If it is an individual, the name is entered, surname first. It does not allow for indicating whether the item was purchased or donated or other details, and therefore is not intended to replace the Recent Acquisitions accession sheet.

TO/FROM_A_NAME, TO/FROM_B_NAME (35 each): These four fields allow us to record correspondent and recipient as well as (*a*) a second correspondent and recipient, or (*b*) a qualifier, e.g. the name of an institution or company of the correspondent and recipient. If the letter is catalogued under the corporate name, that name appears in the first field. If a person is writing for an organization, the organization name goes in the first field. If there are more than two correspondents for one letter, the extra names are added in the Topic_Text field.

Names were originally to be entered conforming to the general stan-

dards set down by Research Collections, modified by us and communicated to them. Names were to be asterisked at their end if they were inferred. There must be no doubt as to the identity of the correspondent before the asterisk can be omitted. If there is doubt the name is followed by "(?)". Even if some form of the name appears, the asterisk is necessary except in special cases, like Stanley Unwin. The use of asterisks interfered with searching and the issuing of reports and was discontinued.

The standard form of the name will be used in all cases; it is not necessarily what appears on the letter. A letter with a typed signature is not a signed letter. Although it would be ideal to use the University Library's MORRIS catalogue authority file, the number of hits would be small as most letter-writers are not book authors. An authority (or consistency) file to consult for standardizing the individual form of entries is available from data inputted; the file derives from the Distinct Names Report. This file is available on the Gopher and Web entries (http://www.mcmaster.ca/russdocs/distinct.htm) for the Russell Archives. A smaller, controlled list of difficult names is maintained in the manual.

DATE_TEXT (13): Here we record the date on the document. There is room to indicate whether the date comes from a postmark, indicated by an "*". Supplied dates and partial dates will be indicated by the use of an asterisk at the end of the date. If there is no date, we indicate the fact by "ND".

ENCL_REF_TEXT (30): This field describes any enclosures with the letter being catalogued. They are briefly indentified, e.g. essay, clipping, photo. The number of sheets is not included in the Pieces count. The original use of this field for linking a reply to the letter it answered is dormant. There is provision for referencing other copies of the same document, as with Dictation, typed copies made at Russell's behest, RA2 multiple carbons, and letters enclosed with other letters.

FORM_TEXT (12): Records the physical form of the document. We use the abbreviation list provided by Research Collections with many additions by us. If the document is not an original, a parenthetical modifier must be added to the description indicating whether it is a photocopy, "(X)", or a microfilm, "(M)". An approved list of abbreviations is maintained in the manual.

PIECES_NO (3): This field records the number of pieces of paper that comprise each document. This the count that matters most for security. It is usually equivalent to "sheets" on the archival folders. Since the purpose is to reflect the document at hand, whether original or photocopy, a single photocopy containing images of two separate pieces of paper counts as one, and photocopies of the recto and verso of the same piece of paper count as two. Envelopes, if they belong to the letter being described, are noted separately, as "E" following the number of pieces, as in "1E". Enclosures are never included in the count.

ADDRESS_CODE (3): This field records only Russell's address and is taken from letters written by him. No other addresses are recorded. Fresh alphabetized lists of all abbreviations can be printed out using the Address and Date Range query. The list is also in a WordPerfect table, connecting the three-digit abbreviation to the full address and sorting on addresses.[6]

When inputting records, the operator will use only address codes from the list. If a new address code is needed, it will be made up (following procedures listed below) and entered on the WordPerfect list. In the case of London, "L" will represent London followed by an abbreviation representing a specific address in that city. No other abbreviation code should begin with "L" so that all London addresses may be culled from the system. If a specific address is not known, then the plain "LON" abbreviation will be used. Outside of London but still in the UK, addresses can be given as either houses, hotels or cities. When Russell is writing on a train ("TRN") or a ship ("SHP"), the address will reflect that. In countries other than England, we have the following options for his address: the country, state/province, city, hotel, name of house. We do not ordinarily make a hotel–specific address for letters written on his lecture tours.

PUBLISHED_TEXT (30): This field will record whether the item has been published and, if so, where. If the answer is definitely negative, the field should be filled in with an "N". If it is affirmative, and the letter was published in the *Autobiography* or another of Russell's books, the Blackwell–Ruja *Bibliography of Bertrand Russell* entry number for the item is used. Only the first instance of publication will be shown in this field. If the letter has been published by another author, the title of the work will be given. If the title is too long for the field, an alphabetic code and list will have to be prepared. In these cases some bibliographic details must be given in order that the published version can be located. Only letters published in full will be represented in this field.

LAST_CHANGED_DATE (6): Filled in by the DB2 programme, it is useful in "edit reports" for checking entries made over a given period.

RECORD_NO (7): It is unique and unchanging. Records that are deleted mean permanently deleted record numbers. The unique record number provides a means of making cross-references in the Encl_Ref_Text field.

TOPIC_TEXT (1,885, or 65 per line): Each line is actually a different record in the RA_Topic_Text table. As a result, lines don't wrap during input. Words should not be hyphenated or otherwise broken at the end of lines. The field may contain text from the letter, précis and commentary.

[6] See Sheila Turcon and myself, "BRACERS (II): Russell's Addresses", *Russell*, n.s. 14 (1994): 179–92. For those following the growth in our knowledge of the addresses from which Russell wrote letters, the total at the end of April 1996 was 392.

III. SEARCHING THE TOPIC_TEXT FIELD

At the beginning we tried a controlled vocabulary, or thesaurus, for the indication of topics, but the available skilled staff was insufficient. The text of some letters to Lady Ottoline Morrell was entered (sometimes running over to one or one more full records), along with the above-mentioned notes on various correspondence files. The first sentence of each letter to Lady Ottoline was entered to provide identification when other features are ambiguous, as they often are. The cataloguer uses the field to write notes about unusual features of the letters. There is, in fact, a great deal of information in the Topic_Text field. The DB2 table is 20 percent as large as the table for the rest of the fields. Because of its space-saving structure, the Topic_Text table does not include blank lines, or rows, whereas the main table includes blank fields.

Once the limitations are understood, searching the Topic_Text field is straightforward. Essentially you are looking for a string of characters on an isolated 65-character line. You may also look for a pattern of characters, i.e. wild-card characters are accepted. The system returns, however, only one line at a time, so context may be severely limited until the Topic_Text for the entire record is retrieved. But the system does return as many lines in the record as contain the string being searched for. Upper vs. lower case doesn't matter.

BRACERS offers the following canned queries:

BRACERS REPORTS

ENTER OPTION FROM LIST BELOW ===>

1. Input proofs
2. Daily input proofs
3. Select by topic text
4. Select by name (to, from)
5. Select by name (to, from) and topic text
6. Select by to name and date and date range
7. Select by from name and date and date range
8. Select by date and date range
9. Select by collection and class
10. Select by address and date range
11. Select by name (Russell & Research Coll)
12. Distinct names report

The third and the fifth ones search the Topic_Text field. The fifth query is chiefly for searching for a name that may appear in any of the four fields for correspondents' names or in the Topic_Text field. The third query is used when the name fields are to be excluded from the search. A reason could be to limit the size of the report.

Selecting option 3 brings up this panel:

SELECT BY TOPIC TEXT

Key a percent sign before and after the search text:

TOPIC ===>

The percent sign tells the system to search for the string anywhere on each line of the Topic_Text table. That is, there may be any number of any characters before or after the string. The underline character ("_") allows a search for a pattern that is missing a single character at the position designated. Thus "%h_bomb%" will be matched by Russell's standard spelling (with a hyphen) as well as by occurrences lacking the hyphen, as may happen in the titles of organizations or by input error.

Because of the limitation of a search to hits on single lines—as opposed to the overall Topic_Text for a given record for a letter—it makes little sense to try to construct Boolean variations in searches. When I have tried to do so the load on the system is so great the "governor" kicks in and the query is prevented from running. The system will not compare the results of hits on one line with hits on another, or the lack thereof. My way around this limitation is to run two searches. In the first I will ask for the hits on (say) "%mathematics%" and in the second those on "%beaut%". When I have the results, I look to see if any have record numbers in common. To search with the operator "not" is not fruitful. In SQL, the querying language being used, the operator in this context is "not like". In running such a query, you are looking for single lines that do not contain a given string. Chances are you'll find thousands of them. "Not like" has to be entered as an alteration to a canned query after the query has been run. Facility in customizing canned queries is highly recommended to in-depth users of BRACERS, even in the case of seeing the full Topic_Text for the context of a search hit. If the Last_Changed_Date for a record as well as its number are to hand, the record number can be requested by the second in the table of

options above. If you have only the record number, you will need to customize an SQL query (LQQMFO16, to be precise).

The system as it runs on the aging IBM mainframe computer is quick and capacious and will easily print out long reports. I have experimented with a copy of the tables imported into dBase V on a personal computer. One can easily imagine the Topic_Text data in a field like a "memo" or "longtext" field complete with word-wrap and searching with Boolean operators and proximity qualifiers. Perhaps BRACERS will one day be transferred to such a database management programme. I have done a study[7] to assess transferring BRACERS to a large library database programme (Horizon), but although the fields can be matched with MARC fields there would be considerable obstacles in retrieval operations on the date field. For authorities files and keyword indexing, the database would have to be separate from the online public access catalogue of the Library. And the indexing and display of the new Topic_Text field would need to extend further, i.e. to the end of the field; it could be made larger by repeating it. Copyright considerations for the material in the Topic_Text field would militate against unrestricted public access.

In the meantime, the task is make the Topic_Text field more comprehensive. CIS did not devise a programme to upload letter texts made electronically available to the Russell Archives by editors. The operation could be done through SQL's "update" command, although preparation of already transcribed texts would be much more feasible to do in a "longtext" field with word wrap. Some 2,000 letter texts are available to be added to the database at this time, thanks to Nicholas Griffin, who has transcribed them in his research for the two-volume *Selected Letters of Bertrand Russell.* It would be invaluable to be able to search on so many letters at once. In the digital age, researchers are coming to expect online images of documents, for which we do not yet make provision.[8] Ideally, they would have a searchable corpus of transcripts as well.[9]

[7] "Assessment of Horizon as a Migratory Possibility for BRACERS Fields", Feb. 1995.

[8] See, for example, Ramesh S. Krishnamurthy and Clifford S. Mead, "An Overview of the Project on the Imaging and Full-Text Retrieval of the Ava Helen and Linus Pauling Papers at Oregon State University Libraries", *Microform Review,* 24, no. 1 (winter 1994–95): 12–15.

[9] I want to thank Sheila Turcon for commenting on this paper in proof and, indeed, for maintaining much of the BRACERS system over many years.