[Radiocarbon, Vol 25, No. 2, 1983, P 661-666]

RADIOCARBON DATABASE: A PILOT PROJECT

STEINAR GULLIKSEN

Radiological Dating Laboratory, The Norwegian Institute of Technology, Trondheim, Norway

ABSTRACT. Computer storage and surveys of large sets of data should be an attractive technique for users of ^{14}C dates. Our pilot project demonstrates the effectiveness of a text retrieval system, NOVA STATUS. A small database comprising ca 100 dates, selected from results of the Trondheim ^{14}C laboratory, is generated. Data entry to the computer is made by feeding typewritten forms through a document reader capable of optical character recognition. A text retrieval system allows data input to be in a flexible format. Program systems for text retrieval are in common use and easily implemented for a ^{14}C database.

INTRODUCTION

The value of information is, of course, dependent on its accessibility to users. This basic principle is obviously valid for the information contained in the rapidly increasing multitude of ^{14}C dates. Global accumulation rate of dates is probably > 20,000 a year. For scientists using ^{14}C dates, keeping "up-to-date" can be quite laborious.

A gallup poll of the majority of Norwegian scientists using ^{14}C dates revealed that more than 90% would like to see the establishment of a ^{14}C database, for efficient and thorough information retrieval. At Ninth International Radiocarbon Conference a recommendation was made to set up a format for reporting dates compatible with computer-assisted information retrieval (Berger and Suess, 1979). Otlet and Walker (1981) discussed the possible development of their system for filing and reporting dates into a database suitable for information retrieval. Polach (1980) tackles the problems of retrieving ^{14}C data through a bibliographic approach. Expensive searching through eight relevant databases yields unsatisfactory recovery due to the multidisciplinary nature of radiocarbon literature; consequently, a specialized bibliography is necessary for efficient retrieval.

DATABASE DESIGN

A pilot project was initiated to evaluate database design for storage and retrieval of ^{14}C dates. Information contained

in a database may vary from pure text to sets of numerical or numerically-coded data. Typical for the data associated with a ^{14}C date is the combination of these types of data. The text that describes the environmental background of a sample, the general aspects of the research project of which the sample is a part, and the specific chronologic goal of the measurement, could all be coded by using selected descriptive categories. An information retrieval system that provides for a search in free text will permit the entry of such data with restrictions only on space and relevant terminology and abbreviations.

Users of dates produced by the Trondheim ^{14}C laboratory were asked to suggest parameters that should be available for searching the database. Such a variety of terms appeared that we decided to consider a system able to search in free text, ie, specific dates could be recovered by searching for any term or combination of terms that appears in the description of the date. The output from this type of database would be less cryptic.

We consulted the Norwegian Computing Center for the Humanities, at The University of Bergen. Their system, NOVA STATUS, is a redesigned and extended version of the STATUS I system developed at AERE Harwell, England. We considered this system suitable for our project.

DATA INPUT FORMAT

All information associated with a ^{14}C date is contained in a document assigned to this date. Descriptive parts of the document will be in free text, but a relatively large part of the information is formalized, ie, given in standardized format, eg, numerically arranged. To obtain high retrieval efficiency for such mixture of free text and formalized data, we must structure the documents. This is done by reserving fields in the document for formalized information with assigned prefixes.

We try to conform to the extent of information given in RADIOCARBON date lists, assuming that the optimum compromise between precision and space requirements is represented by the date list layout.

Data entry can be made in several modes; the choice will mainly depend on the availability of computer peripherals. University mainframe computers are preferred for housing the database, with access through a local terminal, which is also convenient for updating the file. The documents may be prepared gradually along sample routing through the dating process, appearing as an extract from computerized laboratory household routines. Printouts of date reports, date lists, laboratory status reports, etc (Otlet and Walker, 1981) can be

01 Sample ref.no.	T-3919	02 Journ. ref.	1111	03 Subject	Zoology, ethology, geology							
04 Submitter	O. Salvigsen											
05 Institute, address	Norwegian Polar Research Institute, Oslo, Norway											
06 Submitters sample ref.	-76, Sa.no. 38			07 Material	bone							
08 Locality	Svartknausflya			09 Spec., fract.	walrus, collagen							
10 Town, municipal.	Nordaustlandet	11 Province	Svalbard			12 Country	Norway					
13 UTM coord.		14 Lat.	N 79 25			15 Long.	E 22 00					
16 Context, stratigr.	From surface of raised beach at 65 m a.s.l. 5 km from present shore											
17 Project descr.	Glacial history of Arctic regions. Animal behaviour at end of life											
18 Sample objective	Dates raised beach or age of walrus wandering inland before dying											
19 ^{14}C-age	1040*50	20 Report date	811106	21 Access date	820601	22 R.carb. ref.						
23 Calibr. age	AD930*70	24 Calib. type	MASCA	25 R.voir age	440	26 Incl. (I)	I	27 δ^{13}C	-15,5	28 Estim. (E)		
29 Agreement		30 Div. cause										
31 Comment Lit.ref.	Animal has been able to move an impressive distance on land											

Fig 1. Sample form for optical document reading - all the information is for one ^{14}C date.

incorporated into the system. A microcomputer employed as a terminal should be useful for this purpose.

For the pilot project, dates previously produced by the Trondheim laboratory were selected. Optical document reading is preferred for the data entry. The information is typewritten, using the IBM OCR-B font designed for optical character recognition, on forms with guidelines for document structure printed in a non-reproducible color. Documents can be prepared off-terminal by personnel without computer experience and at convenient periods. Figure 1 shows a completed form.

THE SEARCH

Entering a search session, the system requests identification. Several security levels can be defined for different sections, allowing retrieval only for password-holders at the appropriate priority level. The search is activated by queries listing search words, phrases, or combinations defined by Boolean operators. As the free text information is based on descriptions supplied by submitters, terminology varies, and the search must include several synonyms to ensure complete recovery of relevant information. For this purpose, it is useful

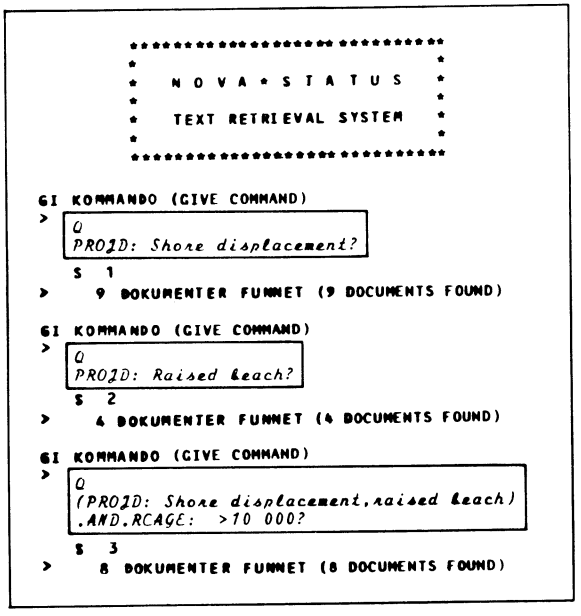

Fig 2. Search for documents containing expressions "shore displacement" or "raised beach" with ^{14}C age > 10,000 yr.

```
GI KOMMANDO (GIVE COMMAND)
>
  READ 1
      7
01  LREF:T-2634
02  JREF:958
03  SUBJ:Geology
04  SUBM:B. Stabell
05  INST:Department of Geology, University of Oslo, Norway
06  SREF:10709
07  MATL:silty clay
08   LOC:Hamravatn, Steinsland
10  TOWN:Sund
11 PROVC:Hordaland
12 CNTRY:Norway
14   LAT:N 60 12
15  LONG:E 05 05
16 CTEXT:Lacustrine sediment immediately overlaying earliest isolation in
         basin with threshold level orig. at 29 m a.s.l.Depth 1345-1340 cm
17 PROJ#:Shore displacement
18 SAOBJ:Earliest isolation from the sea during regression in Bølling
19 RCAGE:12650 # 110
20  REPD:780217
21  ACCN:820601
27  DC13:-24.8
29  AGRM:Good
```

Fig 3. Document retrieved in search shown in figure 2.

to define macros, which are groups of synonyms that can be used to retrieve documents in which at least one of the words appear.

Structuring of documents implies that either the whole document or parts of it may be searched. Searching in a field is executed by addressing a prefix assigned to the field. This function is vital, especially when retrieval is based on specific values of numerical data. Relation operators are applied when values within a range are searched. Only numerical data in a standard format can be compared by using relation operators.

The output of a search session may be a complete formatted printout of all documents retrieved, or a short-form presentation defined by users to consist of any desired combination of document fields. Figure 2 shows how the pilot database responds to a search for the expression "shore displacement" or "raised beach" in the project description field.

The query about documents containing either of these expressions and a ^{14}C age > 10,000 resulted in a recovery of eight documents (fig 3).

PROSPECTIVES

Many university computers are linked together by international datanets. If local databases are generated with

standardized information formats, simultaneous searches in several databases should be possible. Although conversion programs may enable conversation between different database program systems, they should preferably be of the same type with respect to handling capacity of mixed free text and formalized information. The standardization of an information format and eventually a database system should probably be evaluated by the European Study Group on Physical, Chemical and Mathematical Techniques Applied to Archaeology (PACT).

A new general-purpose information retrieval system has been developed jointly by several Norwegian institutions. It is based on experiences with NOVA STATUS and will be portable, ie, adaptable to any computer equipped with a FORTRAN 77 compiler. The SIFT (Search In Free Text) information retrieval system will be made available free of charge (in English). The Council of Europe will implement the system for retrieving judicial information.

CONCLUSIONS

Retrieval of selected data by searching a database for radiocarbon dates should be of great value to scientists utilizing such data. Standardization of information format is desirable for searches through multinational datanets. Database systems capable of searching in free text are well suited for radiocarbon data.

ACKNOWLEDGMENTS

The pilot database was generated and searched by Sigbjørn Århus at the Norwegian Computing Center for the Humanities. The work was financially supported by the Norwegian Research Council for Science and the Humanities (NAVF).

REFERENCES

Berger, R and Suess, HE, eds, 1979, Radiocarbon dating, Internatl radiocarbon conf, 9th, Proc: Berkeley, Univ California Press, p xii.

Otlet, RL and Walker, AJ, in press, The computer writing of radiocarbon reports and further developments in the storage and retrieval of archaeological data, in Mook, WG and Waterbolk, HT, eds, Internatl symp on ^{14}C and archaeology, 1st, Proc: Strasbourg, PACT, in press.

Polach, D, 1980, The first 20 years of radiocarbon dating, An annotated bibliography, 1948-68; a pilot study, in Stuiver, M and Kra, R, eds, Internatl radiocarbon conf, 10th, Proc: Radiocarbon, v 22, no. 3, p 997-1004.