

# Enhanced access to information via personal bibliographic databases

Thomas E. Wolff

*Amoco Research Center, Naperville, Illinois, USA*

---

The Information Research and Analysis group provides scientific, technical and business information to customers within Amoco Corporation. In general, this information is retrieved from databases on the major online systems. Since members of our group have considerable and varied experience working in Research and Development within Amoco and elsewhere, we understand our customers' businesses and concerns and are well situated to carry out searches in accordance with their needs. We add value to our services by analysing the search results for relevance and editing the output for clarity and improved accessibility. Each report includes a cover letter describing search strategies used and results obtained. On request, additional technology assessment will be provided. However, our reports are generally collections of database citations with abstracts. The customer can then order the full documents as they deem necessary, doing his or her own assessment of the subject area.

We place considerable emphasis on providing search reports which can be most effectively used by our customers, and we look to add this value at minimal cost. Traditionally, our efforts have focused on printing reports with only the most appropriate amount of bibliographic or indexing information, or on organising reports into useful subject categories, occasionally adding a Table of Contents. But no matter how well the information is presented, reports are almost always used just once and just by the original requestor. It is clear to us that this is not an effective use of the company's resources. Information is too expensive to be used in such a short-term and short-sighted manner. Therefore, we have begun to provide our search results as bibliographic databases using selected personal computer programs. As will be discussed further, these databases provide our customers with a convenient means to evaluate the search results initially, to revisit these reports, and in some cases, to share that information with colleagues.

The development and usage of these bibliographic databases involves numerous process steps of varying complexity. As these are discussed, reference will be made to Figures 1 and 2 at the end of this article, which show patent and journal article citations, respectively. Each figure is divided into: (a) information downloaded from the online system; (b) personal computer 'screen' images of the information in a Library Master database; and (c) the citation information reformatted and output

## Enhanced access to information via personal bibliographic databases

from the Library Master database and automatically converted to a word processor document.

### What are personal bibliographic databases?

Two years ago, the moniker Personal Bibliographic Databases seemed perfectly appropriate for the products we were developing. Individuals would request a search and the information was provided as a personal computer database along with a hard copy report if requested. So 'personal' referred to the personal requestor as well as to the PC. The PC was the medium of choice because no appropriate mainframe (IBM or VAX) database programs were available at Amoco and the purchase price for new ones was generally over \$10,000, well beyond the means of any individual users. Now we find that research groups are requesting searches for the whole team to use. An individual's PC is too restrictive. We are again looking at mainframe programs, but more likely we will follow the distributed computing path and load the databases on local area networks. At that point, these may well be 'Team Bibliographic Databases' or 'Research Project Bibliographic Databases.'

The designation 'bibliographic' may not be appropriate in the future either. Right now we have difficulty incorporating chemical structures or numerical data, as from Beilstein, in our personal bibliographic databases. In addition, we look forward to adding graphics, either from CD-ROM records or from the online patent databases of the future. Finally, the addition of hypertext capability would further enhance the value of the customers' databases. We anticipate that software packages will improve considerably and require that any data imported into today's databases must be exportable in formats appropriate for tomorrow's improved programs.

The personal bibliographic databases of today are requested by individual researchers to augment or replace 'traditional' searches. Databases are particularly appropriate for broad searches of hundreds or thousands of citations or for periodically updated searches which gradually build up the database. Since the database software is well suited for citation categorisation and sorting, and for putting out reports in word processed formats, formation of the database can be a means to produce a better hardcopy report. In this case, the database may be a side benefit.

Search results from numerous databases and online systems are often combined into a single personal bibliographic database. Output from ORBIT, QUESTEL and STN is well-suited for importing because of the field-delimited formats with unique field tags and well-defined text formats. Information from Dialog Information Services may also be used, either in field-delimited 'tagged' or in native, untagged format. However, problems arise in either case. Tagged format is preferred, although one must deal with non-unique field designators and inconsistent formatting. Untagged information is probably best suited for importing into structureless personal databases, which are not further considered here, although it may be imported into structured databases using the special untagged-format Biblio-Links import/conversion program used with Pro-Cite bibliographic software.

Personal databases will soon be drawing upon other sources of information as well. Already, diskettes and CD-ROM are replacing many hardcopy sources, such as the Science Citation Index from the Institute for Scientific Information, which is

available on CD-ROM and includes abstracts.[1] The CABI (Commonwealth Agricultural Bureaux International) now provides 47 printed abstract journals on diskette.[2] Many 'full-text' journals are available online, such as the extensive Chemical Journals Online (CJO) on STN. These online journals suffer from lack of tabular and image information, which both the online services and personal bibliographic database software will have to address. Eventually, many journals will be published exclusively online. While this may seem unlikely, The Online Journal of Current Clinical Trials, a peer-reviewed journal, is the first of a series of anticipated publications from Primary Journals Online. Although many issues must be dealt with, including fees, copyright and distribution, a level of acceptance has been reached with BIOSIS now indexing articles from the Online Journal of Current Clinical Trials.[3]

### **The value of personal bibliographic databases**

The use of personal databases may be characterised by the following: convenience, communication, idea generation and cost.

#### *Convenience*

Search results are in an accessible format with many of the search techniques available online and with enhanced sort and output capability. Information has extended usable lifetime, well beyond the 'read-once-and-file-away' lives of most hardcopy reports. Multiple searches may be combined into larger databases, or subsets may be conveniently separated into narrow-focus databases. The value of individual records may be enhanced by annotation, cross-referencing or combination of related records.

#### *Communication*

Databases can be used to create custom bibliographies and topical reports. Shared databases can be annotated by individual users to communicate key features of the cited references.

#### *Idea generation*

Convenient searching and browsing through information should lead to generation of new ideas. One's perspectives change as projects progress, so each time questions are asked and the data evaluated, there are new opportunities to learn and generate new search questions. When information in a personal database is found to be incomplete, or when relevant information cannot be found because of limitations in the database indexing or search software, new queries are made in the source online databases, which may then lead to an expanded personal database.

#### *Cost*

Some of the value in maintaining information in an accessible personal database may lead to actual cost savings. Certainly, many 'trivial' questions could then be answered on one's own PC. In addition, some queries of the source database may become 'unnecessary' because the answers have already been obtained in prior searches. On the other hand, the more 'information-aware' researchers should continue to generate more search questions as their knowledge of a field increases.

## Enhanced access to information via personal bibliographic databases

The net search costs may decrease or increase, but the savings in doing better research should be substantial when search information is well used.

### Database packages

Many database software packages appropriate for developing personal bibliographic databases are available and have been reviewed [4-6]. Our experience is with IBM-compatible personal computers, for which we have evaluated four programs in detail:[7] Library Master; Notebook II; Papyrus; and Pro-Cite with Biblio-Links. Two of these have recently been upgraded, Library Master to a local area network version with improved functionality, and Pro-Cite to a more efficient version. Lately, we have also considered three other programs: EndNote, recently translated from the Macintosh version[8]; ideaList[9]; and STN Personal File System, essentially STN Messenger language for the PC with very efficient importing for downloaded STN records. On the basis of our evaluations and our customer acceptance, we have settled on Library Master and Pro-Cite as our recommended software. However, as stated earlier, we will be continually looking for improved ways to bring this information to our customers. For example, our department is evaluating document management software, such as Verity or Excalibur, for managing Amoco's internal documents. Many of these have sophisticated search capabilities. When Amoco moves from the mainframe environment and adopts one of these programs, we may find that they will also be well suited to our bibliographic databases.

### Principal concerns regarding personal database development

#### *User-definable Record Formats*

The patent user community is not a significant constituent of the bibliographic software market. This is most clearly shown by the patent record type available with the software. Some programs list patent types as standard but the format is so limited as to be almost worthless. For example, Papyrus has a patent record type with only nine simple fields, and it is impossible to search or sort on assignee. In Pro-Cite, the patent record format, which is available as a supplemental workform, is acceptable in its simplicity and, even better, modifiable through standard Pro-Cite procedures. However, multiple paragraph abstracts, multiple entry fields (e.g., abstracts from multiple equivalents), and lists of applications or patent family equivalents cannot be easily imported or handled properly in Pro-Cite. These specially formatted fields are all 'word-wrapped' into single, continuous, difficult-to-read paragraph fields. Other database program producers have not considered patents at all. However, for some, the program's flexibility and power can enable creation of a useful patent record type.

The first two screens of a patent record as created by us Library Master in are shown in Figure 1b. These 'data input forms,' as the screens for editing and browsing are called, are fully user-customisable. All bibliographic information fields are shown on the first screen including the first eight lines of the abstracts. Multiple-element list fields, such as author, equivalent patent and applications, are fully viewable by moving the cursor to them, but only the first two lines are shown initially. The remainder of the abstract is also available through the expand-field option. The

second and subsequent screens in our Library Master patent records include a comment field, indexing, other abstracts, and patent claims. The first screen of a journal article record in Library Master (Figure 2b) shows the similarity we have maintained between the layouts of patent and article screens. For example, in each record type the first screen is headed by the title, followed by bibliographic information and then the abstract.

The patent record screen also shows new fields generated by parsing fields in the imported data or by modification of the downloaded information prior to importation. One useful conversion is the separation of patent number from its date, required since most patent files combine them in one information field. Another is the creation of the earliest priority number and date fields, information that is usually buried in other application fields in online records. These separate, searchable earliest priority application fields are valuable to simplify the identification of patents from the same family. We make substantial use of KEDIT macros to reorganise downloaded records, to make the information both more accessible in the personal bibliographic database and more consistent from source to source. For example, the information to be imported is reorganised to take advantage of the Library Master feature which allows any field to be designated as a date-format field. These fields can also have searchable text following the date. The application and patent number fields have dates first, which allows both dates and numbers to be searched appropriately.

### *Importing*

As in most computer software applications, a balance must be struck between program flexibility and ease-of-use. The broad variety of information sources makes critical the program's flexibility in importing process. At the same time, this process should be straightforward and efficient, since users generally wish to pass quickly through the importing stage and on to using the information. Software producers have met the importing challenge various ways. Most create separate importing modules or programs. For our purposes, EndLink, the importing program for EndNote, errs on the side of simplicity because it is a 'black box,' as described by a sales representative, which allows for no alteration of the downloaded information. At the other end of the scale, Papyrus requires development of a complicated import template which must account for all possible variations in the imported information. Although the Papyrus producer will create customised importing programs for the users, we have found them to be generally ineffective, as for importing Chemical Abstracts information from STN, for example. Developing import templates can be much more straightforward. The Convert program with Notebook II is extremely simple to customise, and the import facility for Library Master is nearly as clear-cut. The balance is probably best struck by Pro-Cite, which provides separate Biblio-Links for each online systems and many CD-ROM products. The Biblio-Links can also be modified readily. However, until recently, the pricing policy for the Biblio-Links made them considerably more expensive than the database software Pro-Cite itself; with the newly created package sets, the import and database software are more nearly matched in cost.

## Enhanced access to information via personal bibliographic databases

Responsibility for importing difficulties also lies with the database producers. Personal computer importing programs must be powerful because online information is so complex and inconsistent. Some databases have dozens of document types. For example, at last count, Compendex has 26 document types on STN. Of more concern, fields frequently contain more than one type of information, necessitating the use of sophisticated field parsing or editor programs. For example, in APILIT, the source field has almost no consistent format at all and can contain CODEN or ISSN numbers or references to other sources such as Petroleum Abstracts. Similarly, the source field in CA file article citations contain all the bibliographic information except year of publication (see Figure 2a). CA file patent record source (SO) field does not contain the application or patent numbers or issue date, as might be expected, but rather the 'country' information, number of pages, and, for Polish patents only, a phrase which describes that the abstract was taken from the application. Rather, the actual 'patent information' is found in the PI field.

Online database information can also be presented badly. For example, the tagged formats for Dialog Information Services appear to be just an afterthought created out of the 'normal' format. Dialog field tags frequently are not unique, and they may contain subfield descriptors in angle brackets at the beginning of the field's text area (see Figure 1a of a Derwent World Patent Index record). Dialog also uses the virgule or vertical bar to designate end-of-field, but this mark is not applied consistently throughout records or between files. Our KEDIT macros make major changes to Dialog tagged output.[7] In addition, we have recently written macros which convert untagged Derwent output (e.g., Dialog format 7) to the equivalent tagged format. This allows us to use Derwent citations for importing, even if we had not planned on it when the search was originally downloaded. The whole issue of having to remember to use tagged format does not exist when downloading information from ORBIT, Questel, or STN. These online systems only produce tagged output, which is generally more consistent than that from Dialog.

The customer-oriented approach of the CABI seems particularly enlightened.[2] Subscribers to CAB Abstract Journals on diskette requested search and retrieval software. Other database providers have responded by including specialised retrieval software with their data, ISI (Science Citation Index on diskette) or Ziff-Davis (Computer Library on CD-ROM). However, CABI recognised that subscribers had their own preferred database software which could be used for information from varied sources. To accommodate this need of the majority of their subscribers, CABI publishes their abstracts journals in both comma-delimited and Pro-Cite proprietary formats and provides information on potentially useful database software. The lesson to be drawn is that information should be formatted consistently and thoughtfully in consideration of importation into the personal databases of the customer's choice.

### *Browsing, searching and general ease of use*

The forecasted demise of hardcopy information is often regretted because of the value and pleasure in browsing and the serendipity in finding unsought-for information of interest. This ability to browse information conveniently is critical to idea

