J. Toxicol. Pub. Health
Vol. 17, Supplement, pp. 227-236, 2001

Improving International Access to the *IARC Monographs* Database with Linkage to other Sources of Information[1]

Jerry M. Rice[2,*], Michael D. Waters[3] and R. Glenn Wright[4]

[2]*International Agency for Research on Cancer, 69372 Lyon (Cedex 08), France;*
[3]*U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711, USA;*
[4]*GMA Industries, Inc., Annapolis, Maryland 21401, USA*

ABSTRACT *: The IARC Monographs Programme on the Evaluation of Carcinogenic Risks to Humans has reviewed, summarized and evaluated 869 environmental agents and exposures as of June 2000. This large collection includes all relevant published epidemiological data on cancer in exposed humans and results of bioassays for carcinogenicity in experimental animals. Since 1986, cancer data have been systematically supplemented by summaries of other toxicological data that are relevant to assessments of carcinogenic hazard. These include summaries of genetic and related effects of chemicals, which have been prepared as Genetic Activity Profiles (GAP) by the U.S. EPA in collaboration with IARC. As the Monographs have proved increasingly valuable and influential worldwide, they have evolved into an encyclopedia on environmental carcinogenic risks to humans. However, the Monographs have historically been prepared only as printed books with limited distribution, and the Monographs Programme has needed to adjust to expectations of wider availability. Since 1998 the evaluations and summaries have been globally accessible by Internet from IARC (http://www.iarc.fr) and the GAP profiles by Internet from EPA (http://www.epa.gov/gapdb/), with the two websites linked. Improved EPA/IARC GAP database and software, GAP2000, now link GAP profiles directly to the appropriate IARC web pages for summaries of evaluations of a given compound and its overall IARC classification. During the year 2000, by means of optical character recognition (OCR) technology the entire series of IARC Monographs is being converted to an electronic version. The first edition is now available commercially in CD-ROM format and will soon become available on-line at <http://www.gmai.com/IARC>.*

Key Words *: Carcinogenicity, Genetic and related effects, Internet, CD-ROM*

I. INTRODUCTION

The Monographs Programme of the International Agency for Research on Cancer (IARC) is an international, interdisciplinary approach to carcinogenic hazard identification. Its principal product is the English-language book series, the *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, which began publication in 1971. The *IARC Monographs* contain timely, critical reviews of the published scientific literature on the possible carcinogenicity to humans of environmental agents (chemicals, groups of chemicals, complex mixtures, physical or biological agents) or exposure circumstances (occupational exposures, lifestyle and cultural habits), together with authoritative evaluations of the strength of the total evidence for human cancer hazard. The *Monographs* have evolved into what is essentially the World Health Organization's encyclopedia on the roles of environmental agents in human cancer causation, and have proved useful worldwide to scientists, public health authorities, and to the general public. The *Monographs* are widely regarded as authoritative evaluations of carcinogenic hazards to human beings.

Subjects are chosen for evaluation according to two criteria: there must be evidence or suspicion of carci-

[1]This manuscript has been reviewed by the National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency (USEPA) and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the USEPA, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.
Abbreviations : dpi, dots per inch; EPA, United States Environmental Protection Agency; GAP, genetic activity profiles; GRE, genetic and related effects; IARC, International Agency for Research on Cancer; OCR, optical character recognition; PC, personal computer.
*To whom correspondence should be addressed

nogenicity and there must be human environmental exposure. Nominations for evaluations are actively solicited from scientists and public health authorities worldwide, and priorities are established with the aid of international advisors (IARC, 1998). Reviews and evaluations of agents and exposures are carried out by international Working Groups of scientific experts, who are invited to participate on the basis principally of their contributions to the relevant scientific literature. More than 1,000 scientific experts from 43 countries have participated in this Programme since its inception.

Three volumes of the *Monographs* are published annually. In the first 78 volumes, 1971 through June 2000, a total of 869 agents and exposures have been reviewed and the strength of the total evidence for carcinogenic hazard to humans has been evaluated. In 1987, in Supplement 7 to the *Monographs* (IARC, 1987c), a system for formal classification of the strength of the total evidence for carcinogenic hazard to humans was introduced. Since that time, each *Monographs* review has concluded with a formal evaluation that places the agent or exposure that has been reviewed into one of five groups:

Group 1—carcinogenic to humans;

Group 2A—probably carcinogenic to humans;

Group 2B—possibly carcinogenic to humans;

Group 3—cannot be classified as to carcinogenicity to humans; and

Group 4—probably not carcinogenic to humans.

Criteria for inclusion in each group are described in detail in the Preamble to the *Monographs*, together with the procedures followed by Working Groups in preparing the documents and arriving at their evaluations. The degrees of evidence for cancer in humans as a result of exposure to an agent, and for carcinogenicity to experimental animals in bioassays, are separately evaluated first, using predefined criteria. An overall evaluation is then made taking into consideration the human and the animal data, together with other relevant data which vary according to the nature of the exposure under evaluation. For chemicals, these may include pathways of biotransformation in experimental animals and in humans; biomarkers of exposure and of toxic effects; genetic toxicology including patterns of mutation and structural alterations in chromosomes; and other evidence that may contribute to a judgement as to whether a carcinogenic risk to human beings may result from exposure to the agent.

New research findings relevant to an evaluation of carcinogenicity, which are published in the scientific literature after an evaluation has taken place, may modify the total evidence for carcinogenicity to such an extent that a new review and evaluation is required. Some especially well-studied agents have therefore been evaluated as many as three or four times in the light of new research findings (e.g., polychlorinated dibenzodioxins: IARC, 1977; 1987d; 1997).

In 1992, the Preamble to the *Monographs* was revised, in the light of scientific advances in understanding the modes of action of various categories of carcinogenic agents, to allow inclusion of information on mechanisms of carcinogenic action in overall evaluations (IARC, 1992a). As a result, certain agents were classified upward, from Group 2B to Group 2A, when the mechanism of carcinogenic action was well understood and there was clear evidence that the mechanism operated both in humans and in animals. Examples include chemically reactive alkylating agents such as diethyl sulfate (IARC, 1992b), for which there were sufficient positive data for carcinogenicity to animals but no data for cancer in humans. More than half the agents currently in Group 2A are in that Group on the basis of sufficient evidence for carcinogenicity in animals, supported by data on genetic and related effects. As of Volume 78 (June 2000):

- 87 agents have been evaluated as carcinogenic to humans (Group 1);
- 63 as probably carcinogenic to humans (Group 2A);
- 235 as possibly carcinogenic to humans (Group 2B);
- 483 as unclassifiable as to carcinogenicity to humans on the basis of data currently available (Group 3); and
- one as probably not carcinogenic to humans (Group 4),

a total of 869 chemicals and other agents that have been evaluated (Table 1).

More recently, evidence has accumulated that some chemicals and chemical mixtures may induce neoplasms in experimental animals by mechanisms that do not predict carcinogenicity to humans (IARC, 1995; Capen *et al.*, 1999). A few chemicals that do

Table 1. 869 Overall evaluations of carcinogenicity from *IARC Monographs* Volumes 1-78 (1972 2000). A downward arrow (↓) indicates evaluations that have been reduced from the next higher level on re-evaluation of additional evidence including data on genetic and related effects and on mechanisms of carcinogenicity. An upward arrow (↑) indicates evaluations that were similarly revised upward.

		Basis for classification		
Group	Definition	Cancer Data	Plus ORD	Total
1	Carcinogenic to humans	82	5↑	87
2A	Probably carcinogenic to humans	25	38↑	63
2B	Possibly carcinogenic to humans	230	5↑	235
3	Not classifiable	478	5↓	483
4	Probably not carcinogenic to humans	1	0	1
	Total			869

cause tumors in experimental animals (e.g., *d*-limonene and saccharin) have been reclassified downward, from Group 2B to Group 3 (IARC, 1999) on the basis of evidence that their carcinogenicity in experimental animals is due entirely to the operation of such mechanisms (Table 1). The classification process is thus a dynamic one that takes into account both the publication of new data and advances in scientific understanding of carcinogenic processes.

II. GENETIC ACTIVITY PROFILES AND COMPUTERIZED DATABASES

The EPA/IARC Genetic Activity Profile (GAP) database was begun in 1983 as a collaboration between the National Health and Environmental Effects Research Laboratory of the U.S. Environmental Protection Agency (EPA) and the IARC Monographs Programme. The methodology was presented to the IARC Working Group for *Monographs* Volume 36 in February 1985 as a means to improve documentation of the evaluations of genetic and related effects (GRE) in the *Monographs*. The Working Group voted to support the use of the GAPs and corresponding data listings by IARC. The first journal publication on GAPs appeared shortly thereafter in *Mutation Research* (Garrett *et al.*, 1984). The profile methodology was used in two additional IARC pilot efforts (in working group meetings for *Monographs* Volumes 39 [June 1985] and 41 [February 1986]). Meanwhile, the methodology was modified to further meet the needs of IARC and to comply with the recommendations of an ad-hoc IARC Advisory Panel that met during the 1985 International Conference on Environmental Mutagens in Stockholm. This committee critically evaluated the methodology and recommended the integration of GAPs with the *IARC Monographs* evaluations of GRE of suspect human carcinogens.

The major effort in creation of the GAP database occurred between 1985 and 1986 with the preliminary review and preparation of the draft GAP database for *IARC Monographs* Supplement 6. For EPA this involved the task of reviewing all the published short-term mutagenicity and other genotoxicity test results for more than 200 compounds, and then preparing quantitative GAPs and data listings for review by the Supplement 6 Working Group. The review and modification of the Supplement 6 database on 195 compounds took place over a period of ten days in December 1985 (Waters *et al.*, 1988b). Nearly one years effort was required to verify all of the data and, by the end of 1987, the *IARC Monographs* Supplement 6 was published (IARC, 1987a). This publication resulted in the creation of the EPA/IARC Genetic Activity Profile (GAP) database.

Following the Supplement 6 meeting, IARC's mode of operation for review of GRE data changed. Henceforth, IARC would perform the primary literature review, and EPA would prepare the quantitative GAPs after the working group meetings. This mechanism has resulted in the addition of 15 to 20 new agents to the GAP database with each new working group meeting where chemicals or chemical mixtures are considered, beginning with *Monographs* Volume 46.

The impact of the GAP methodology on the IARC review process was significant and immediate. The review of GRE in Supplement 6 in December 1986 preceded the evaluation of the animal and human carcinogenicity of the same group of agents in the working group meeting for Supplement 7 in March 1987. That working group was provided with an organized data set that clearly described the GRE of each compound and mixture. These data were used in the first-ever IARC overall evaluations of carcinogenicity (IARC, 1987b).

During 1987 programming for the personal computer (PC) version of the GAP database was begun by

W.J.A. Lohman and P.H.M. Lohman, who have continued to refine the software. The first version was completed in August, 1987, and was demonstrated at the annual (North American) Environmental Mutagen Society meeting in Charleston, South Carolina in March 1988.

A general description of the GAP database (Waters *et al.*, 1991) was published in the report of a meeting on "Databases of Genotoxicity and Carcinogenicity and their Usefulness in Hazard Evaluation" (Parodi and Waters, 1991). Critical examination and evaluation of the GAP database has been undertaken in a series of assessment documents. The first of these reports (Jackson *et al.*, 1993) evaluated the genetic toxicology of substances considered to be nongenotoxic carcinogens, including the carcinogenicity data as well as the mutagenicity data on these chemicals. One conclusion from this survey was that there may be relatively few truly nongenotoxic carcinogens, once such presumed nongenotoxic agents have been adequately tested for their ability to induce gene mutation, chromosomal aberrations and aneuploidy. The second paper (Bridges *et al.*, 1993) investigated the sensitivity of several short-term tests to detect germ cell mutagens, and a third paper (Waters *et al.*, 1994) addressed the specificity, predictivity and accuracy of the same short-term tests applied to germ cell mutagens and nonmutagens, as well as the quantitative performance characteristics of the tests. Tice *et al.* (1996) assessed the utility of the database with reference to human exposures to environmental mutagens. More recently the entire database has been reviewed with regard to its utility in the classification of putative human carcinogens (Waters *et al.*, 1999). These assessments are expected to help shape the way in which short-term tests are used and interpreted in the future.

Current Status. Data abstracted from approximately 8000 references for about 700 agents have been compiled in the current version (GAP2000) of the EPA/IARC Genetic Activity Profile (GAP) database, which includes volumes 1-76 of the *IARC Monographs* as well as several EPA studies (e.g., pesticides, Superfund waste-site chemicals, or hazardous air pollutants). The EPA/IARC GAP database is now distributed internationally via the Internet and is in use in national and international governmental organizations, U.S. federal and state agencies and many private companies. The database and software used to display histogram plots (profiles) and data listings for individual chemical agents have recently been upgraded to a 32-bit version in GAP2000, and are now available for downloading without charge at *http://www.epa.gov/gapdb/*.

The profiles provide a visual overview of the doses and results from original studies reported in the open literature for multiple tests used to evaluate the genetic and related effects of chemical agents. Either the lowest effective dose or highest ineffective dose is recorded for each study (Fig. 1). These values are plotted on the y-axis for each test result: positive results appear above the x-axis and negative results appear below the x-axis. Up to 200 different short-term tests, identified by three-letter codes, are represented across the x-axis of the profile. Tests are presented sequentially according to the phylogeny of the test organisms and the end points of genetic activity. A unique reference to the published data is cited for each entry in the database. A complete tabular listing of the data shown in a profile can also be produced.

GAP2000 uses Windows™ features, including mouse functions, icon tool bars, radio buttons, clipboard copying and conventions to organize multiple windows (i.e., cascade, tile, etc.). The cascade and tile features enable the viewing of multiple profiles and/or data listings. Other features include options to obtain complete bibliographies of cited references for each chemical or for the entire database, the capability for searching the text within a given bibliography, a display of the IARC evaluation of carcinogenicity on the profiles and data listings, and an option for viewing a table that shows details of the IARC evaluation for all chemicals that have been evaluated by the Monographs Program. This table is accessed by clicking the [IARC Evaluation] button (Fig. 1). GAP2000 also provides hyperlinks from the chemical name to the IARC web page that contains the written summary of the *Monographs* evaluation. A chemical structure database for GAP (GAP ChemFolder) is also available without charge. The GAP ChemFolder chemical structure database is searchable by chemical formula, molecular weight, or other data such as Chemical Abstracts Service registry number, chemical class, IUPAC name, or chemical properties using commer-

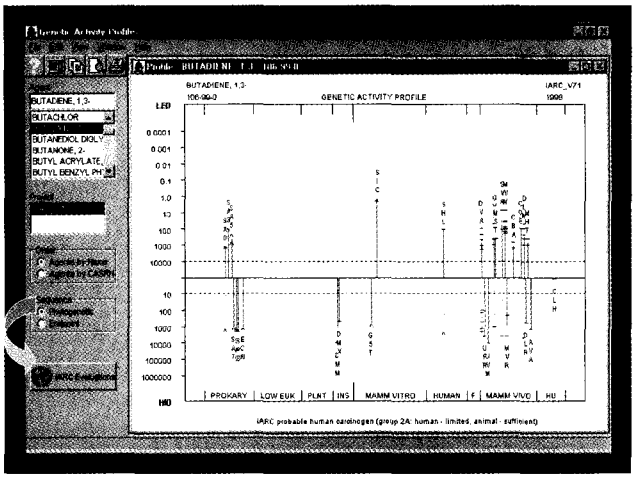


Fig. 1. GAP2000 profile for 1,3-butadiene, shown as the on-line presentation with direct linkage to the IARC website. This provides direct access to the most recent IARC evaluation of the compound for carcinogenicity to humans and the narrative summary of the evidence considered by the IARC working group in making the carcinogenicity evaluation.

cially available software. A trial version of this software can be used 25 times or up to one year from the date of installation. Chemical structure or substructure searches can also be used to identify structural analogs within the database[5].

In addition to its role in *Monographs* documentation, the graphic display of the multi-test information in GAPs has proved useful for comparative assessments of both qualitative and quantitative results across several dimensions (e.g., concordance across species and endpoints, identifying data gaps, and evaluating relative potencies of chemicals). Structurally similar compounds frequently display qualitatively and quantitatively similar profiles (Garrett *et al.*, 1986; Waters *et al.*, 1993a). By examining the patterns of GAPs of pairs and groups of chemicals, it is possible to make more informed decisions regarding the selection of test batteries to be used in evaluating chemical analogs (Waters *et al.*, 1988a). GAPs have provided useful data for the development of weight-of-evidence hazard ranking schemes (Brusick *et al.*, 1992). In addition, some knowledge of the potential genetic activity of complex environmental mixtures may be gained from assessing the GAPs of component chemicals (Waters, 1990a).

[5]The software used for managing the database, GAP ChemFolder, is provided by Advanced Chemistry Development, Inc. (ACD). For more information visit the ACD website at *http://www.acdlabs.com.*

III. *IARC MONOGRAPHS* WEBSITE

Establishment of a Monographs Programme webserver (Fig. 2) within the IARC website at http://www.iarc.fr has compensated for a number of weaknesses that were unavoidable in dissemination of the *Monographs* as books alone:

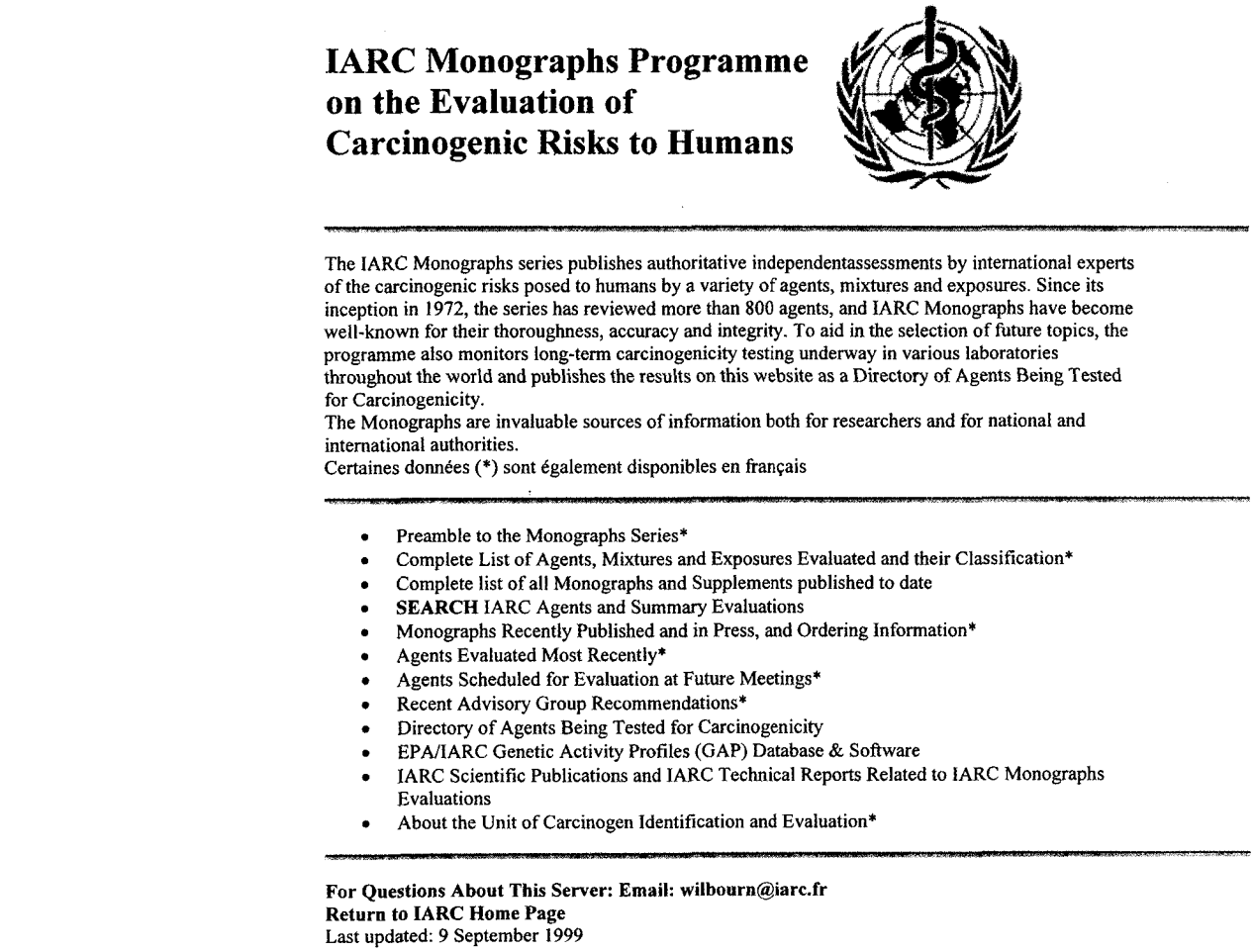
- Limited press run and a single printing, therefore limited accessibility;
- Each book prepared as an independent entity;
- Lack of subject index.

In addition, the Internet provides the most efficient way to make available the complete list of evaluations, which changes several times each yearafter every Working Group meetingand is therefore unsuited to distribution in printed form. Such lists have never been included in individual *Monographs* volumes, but are among the most frequently requested Monographs Programme documents.

Perhaps the most important added value of the electronic database is its search engine. The search engine compensates for the lack of subject indexes in individual volumes and effectively integrates the entire series by providing access to all summaries of reviews as well as to the list of evaluations. The search engine also allows identification of subjects by alternative names, using lists of synonyms and variant spellings. Although the *Monographs* are prepared in English only (a practice that avoids translation errors and minimizes the number of alphabetic letters and characters used), variant spellings do exist, reflecting principally differences between British and American usage. For example, "estrogen" in the USA becomes "oestrogen" in the UK (and in the *IARC Monographs*), with a significant effect on the expected placement of such an entry in an alphabetic listing. Search engines can also compensate effectively for multiple names of compounds. The Internet database is therefore in im- portant ways more versatile than the original printed documents.

IV. CONVERSION OF FULL-TEXT *IARC MONOGRAPHS* VOLUMES INTO ELECTRONIC FORMAT

Access to the complete text of the *IARC Monographs*



IARC Monographs Programme on the Evaluation of Carcinogenic Risks to Humans

The IARC Monographs series publishes authoritative independentassessments by international experts of the carcinogenic risks posed to humans by a variety of agents, mixtures and exposures. Since its inception in 1972, the series has reviewed more than 800 agents, and IARC Monographs have become well-known for their thoroughness, accuracy and integrity. To aid in the selection of future topics, the programme also monitors long-term carcinogenicity testing underway in various laboratories throughout the world and publishes the results on this website as a Directory of Agents Being Tested for Carcinogenicity.
The Monographs are invaluable sources of information both for researchers and for national and international authorities.
Certaines données (*) sont également disponibles en français

- Preamble to the Monographs Series*
- Complete List of Agents, Mixtures and Exposures Evaluated and their Classification*
- Complete list of all Monographs and Supplements published to date
- **SEARCH** IARC Agents and Summary Evaluations
- Monographs Recently Published and in Press, and Ordering Information*
- Agents Evaluated Most Recently*
- Agents Scheduled for Evaluation at Future Meetings*
- Recent Advisory Group Recommendations*
- Directory of Agents Being Tested for Carcinogenicity
- EPA/IARC Genetic Activity Profiles (GAP) Database & Software
- IARC Scientific Publications and IARC Technical Reports Related to IARC Monographs Evaluations
- About the Unit of Carcinogen Identification and Evaluation*

For Questions About This Server: Email: wilbourn@iarc.fr
Return to IARC Home Page
Last updated: 9 September 1999

Fig. 2. IARC Monographs home page (*http://www.iarc.fr; → Publications; → Monographs Programme*).

is currently being improved through the conversion of the existing text documents into an electronic format. The primary benefit of this process is the increased and global availability of the *Monographs* through the use of CD-ROM media as well as the Internet. The complete set of *Monographs* will be available, including all of the early volumes that are now out of print. This is important, because of the historical overviews that are included in many volumes of the *Monographs* and that serve as a unique guide to the early scientific literature. The *Monographs* availability in electronic form also provides previously unavailable opportunities to search document content, including keyword and relationship search.

Approximately three-fourths of the *Monographs* volumes, comprising volumes 1 through 55, exist only in the original paper form. The remaining volumes exist in several different electronic forms, ranging from simple ASCII text, to Word and Quark documents. The task, therefore, involves the conversion of four different types of documents into one common format: Adobe.pdf. Word and Quark document conversion primarily requires simple format adjustments to ensure that pagination and table positioning remains true to the original printed volume, while ASCII text conversion requires significant formatting effort to reflect the original document. However, the volumes that must be converted from hardcopy into electronic form pose significant challenges regarding not only formatting to retain their correspondence with the original document, but in the accuracy of the text conversion as well. Thus, the conversion process as well as the tools used form equally important parts of the solution (Bunke, 1997). Because this process is essential to the rescue of any printed-text-only documents, it is presented here in some detail.

Document Scanning. Capture of the original text is accomplished by combining optical scanning with optical character recognition (OCR) techniques. Optical scanning uses a feed-through scanner at a resolution of 300 dots per inch (dpi). This resolution setting is a key factor in providing an image of the document that contains well-formed characters that contribute to the success of subsequent OCR. The use of a lower image resolution setting provides the ability to save more images of entire pages in a memory bank of a specific size. However, the quality of the text characters is likely to suffer degradation through a lack of sharpness on edges and curves. This can pose significant difficulty when attempting to use OCR techniques to read these characters, and is likely to result in a higher error rate by making it harder to distinguish one character from another. The use of a scanner resolution above 300 dpi, however, provides only marginal improvement to the document image that does not translate to improved OCR accuracy, while significantly increasing memory requirements. Figure 3 illustrates the differences obtained in text images using different resolution settings.

Testing of OCR accuracy at different scanner resolutions has shown that images captured at 350, 400 and 600 dpi do not yield improvement in overall OCR accuracy as resolution increases above an inherent error rate of 5~10%, independent of the OCR engine used (Blando, 1994; Dickey, 1991). OCR accuracy is measured in terms of its character accuracy. To define character accuracy, the number of insertions, substitutions, and deletions required to correct the OCR output to agree with the "correct" text are measured (Rice, 1992).

Different kinds of errors are encountered at different resolutions. For example, the character "O" may be interpreted as "C" more often at 300 dpi than at 600 dpi. However, the character "&" is more often misinterpreted as "6" at the higher resolution. This may be attributable to the smaller pixel sizes at the higher resolution. For example, the 300 dpi "P" in Fig. 3 has a very sharp edge as all variations of this edge fall within the same pixel. However, the same variations of the edge may fall across the boundary of two or more pixels in the 600 dpi image, thus creating a rough or serrated edge. The choice of resolution setting is thus based upon identifying the types of errors likely to be encountered, rather than reducing the number of errors. There is no inherent advantage to preferring one type of error to another since as many errors as possible must be detected and corrected. However, knowing which errors to expect assists the creation of automated tools to check specifically for these errors while editing the documents.

Optical Character Recognition (OCR). Upon scanning a document, OCR techniques are employed to translate the character images into recognizable text. This is accomplished by converting the character image into a bitmap, illustrated in Fig. 4, which is a matrix of pixels that are turned on or off in relation to the image. This provides for another source of error that corresponds to pixel boundary variation. This bitmap representation forms the basis for the process of determining what the image represents in the form of text.

Both edge and corner detection algorithms are used to help discern character location and orientation, and adjustments are made for crooked and/or uneven edges that may be attributed to skew of the page through the scanner. Once the edges are detected and compensation for skew has been performed, efforts

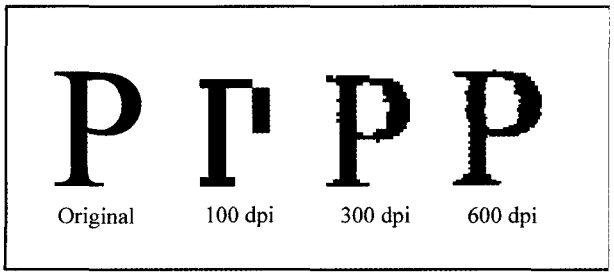


Fig. 3. Optical character recognition: differences in character images at different resolutions.

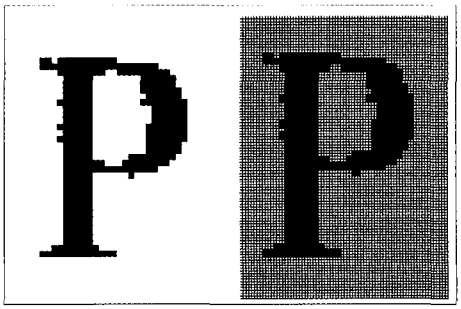


Fig. 4. Optical character recognition: character image conversion into bitmap representation.

are made to identify the image as a specific character.

The actual recognition of characters from images can be accomplished in a several ways. For example, dictionary methods may be used to determine a best match where a small lexicon of distinct characters exists. Practical considerations, however, eliminate the use of such an approach. For example, English language text exhibits 26 alphabetic characters, ten numerals, and fifteen or more special characters associated with punctuation, etc. Multiply this by 150 or so possible different fonts, plus bold and italic styles, and over 23,000 different characters can result. Such an abundance of possible characters requires that more innovative approaches be used, including artificial intelligence and neural network-based techniques for pattern recognition (Pavlidis, 1993). One such method includes the use of techniques that attempt to distinguish the properties of each character by mapping bitmap coordinates using quantization techniques to map the white space of letters into matrices (Pagurek, 1990). These matrices correspond to regularities in absolute and relative positioning of text elements that are used to identify characters regardless of size or orientation. Other techniques involve the training of algorithms for estimating character widths, character locations in a word, and match/nonmatch probabilities from unsegmented text (Xu, 1999). Indeed, numerous other methods have been tried, including document zoning and word usage within a document, with varying degrees of success.

Most errors that occur through the OCR process are generally associated with the misidentification of one character as another. Typical of these efforts are: O appears as C, F ⇒ P, P ⇒ R, R ⇒ F, E ⇒ F, 6 ⇒ &, 5 ⇒ S, M ⇒ AA, and "The" ⇒ M E. In some cases the initial syllable of a word may be truncated. One particularly embarrassing example of this is the tendency for the word "woman" to be replaced by "man". Knowing the specific types of errors likely to be encountered makes it possible to train special dictionaries to detect potential instances of their occurrence. The combined use of the special, English and medical dictionaries makes it possible to detect the vast majority of OCR errors that are likely to occur. The rest are best found through the efforts of a good human editor. Indeed, this effort has resulted in the improvement of some of the documents from their original form.

As a method of accurately and efficiently capturing *Monographs* text data, OCR remains preferred to the most likely alternative, which is retyping the volumes by hand using dual typists and comparison between texts to identify errors. This brute-force approach is time proven, but is labor intensive and is subject to transcription and other errors. Such errors tend to unsystematic and unpredictable, and make it difficult to engineer automated tools for their detection and correction.

V. DISCUSSION

This paper chronicles the evolution of one internationally important database, the *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*, from its origins as a series of hand-typed volumes printed by photo-offset methods in 1972. The *Monographs* have become an integrated system of computer-based printed text and multiple searchable electronic databases that are available both as books and as CD-ROM and on-line via the Internet in the year 2000.

This evolution has been accomplished because the *Monographs* continue to be needed, by scientists, by regulatory authorities, and by the general public. They are the only comprehensive, international, interdisciplinary catalog and authoritative review of agents in the human environmentchemicals and mixtures, infectious agents, various forms of radiationthat present a cancer risk to humans, or that have been studied to investigate the possibility of such a risk. Their authority derives entirely from the credentials of the research scientists worldwide who agree to participate in IARC working groups, and who in effect subject the published scientific literature to a second cycle of peer review in the process of preparing the *Monographs*.

It seems clear that in the future, information resources like the *IARC Monographs* database will become increasingly dependent on electronic methods for publication and on the Internet for distribution to readers. We do anticipate however that printed books will need to be produced in parallel with electronic publications for the foreseeable future. This is because books are far more practical than computer-based

technology for detailed studies, and also because the Internet is not yet universally, economically and reliably accessible.

The growth of electronic databases over time presents certain problems that result from sheer magnitude. The most significant problem is the challenge of properly searching a single large database for specific data. Text-search, for the simple mention of a term or of several spatially linked terms, may yield a very large number of positive identifications and a high proportion of these may not be useful, or there may simply be too many to examine for possible usefulness. Sequential subject searches of indexed terms in a series of smaller, limited databases may yield more useful information. There appears to be a practical need for linked, rather than fully integrated, databases. Also, there need to be multiple and redundant linkages among databases, to improve the likelihood of successful retrieval of needed information in an efficient way.

A problem for electronic databases that does not exist for the print (hard copy) versions is their dependence on word processing and "desk top publishing" software. Such software has an astonishingly rapid rate of obsolescence. Decisions regarding upgrades to a more advanced version of a software package, or conversion to a different one, must be made from time to time to preserve the database. Inevitably such decisions have an adverse impact on linkages to other databases which may no longer be compatible with the new software. Some systematic decision-making process will be required to keep linked databases usable.

There is potentially a useful role here for scientific societies in maintaining the websites that serve as the gateways to linked on-line databases. Scientific societies can in principle also play a useful role in assuring the quality of databases considered for inclusion in a linked library of such databases: does each, for example, include a peer review process for validating newly added data. International scientific societies may in the future provide centralized, on-line linkages among many websites that offer various kinds of toxicological information, as is now done for genetic and related effects by the International Association of Environmental Mutagen Societies (*http://www.iaems.nl*).

REFERENCES

Blando, L.R. (1994): Evaluation of page quality using simple features. Master's Thesis, Department of Computer Science, University of Nevada, Las Vegas.

Bridges, B.A., Stack, H.F. and Waters, M.D. (1993): The performance of *in vitro* mutagenicity tests in identifying germ cell mutagens in Current Issues In Toxicology (Madle, S. and Muller, L. eds.) MMV Medizin-Verlag, Munich, pp 17-23.

Brusick, D.J., Ashby, J., de Serres, F.J., Lohman, P.H.M., Matsushima, T., Matter, B.E., Mendelsohn, M.L., Moore, D.H. II, Nesnow, S. and Waters, M.D. (1992): A method for combining and comparing short-term genotoxicity test data: Preface. A report from ICPEMC Committee 1. *Mutat. Res.*, **266**, 1-6.

Bunke, H. and Wang, P.S.P. (1997): Handbook of Character Recognition and Document Image Analysis. World Scientific Pub Co., River Edge, NJ, pp 17-23.

Capen, C.C., Dybing, E., Rice, J.M. and Wilbourn, J.D., eds. (1999): *Species Differences in Thyroid, Kidney and Urinary Bladder Carcinogenesis*, IARC Scientific Publication No. 147. International Agency for Research on Cancer, Lyon.

Dickey, L.A. (1991): Operational Factors in the Creation of Large Full-Text Databases, DOE Infotech Conference, Oak Ridge, TN.

Garrett, N.E., Stack, H.F., Gross, M.R. and Waters, M.D. (1984): An analysis of the spectra of genetic activity produced by known or suspected human carcinogens. *Mutat. Res.*, **134**, 89-111.

Garrett, N.E., Stack, H.F. and Waters, M.D. (1986): Evaluation of the genetic activity profiles of 65 pesticides. *Mutat. Res.*, **168**, 301-325.

IARC (1977): IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man, Volume 15. Some fumigants, the herbicides 2,4-D and 2,4,5-T, chlorinated dibenzodioxins and miscellaneous industrial chemicals. International Agency for Research on Cancer, Lyon.

IARC (1987a): IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Supplement 6. Genetic and Related Effects: An Updating of Selected IARC Monographs from Volumes 1 to 42. International Agency for Research on Cancer, Lyon.

IARC (1987b): IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Supplement 7. Overall Evaluations of Carcinogenicity: An Updating of IARC Monographs Volumes 1 to 42. International Agency for Research on Cancer, Lyon.

IARC (1987c): Preamble in IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Supplement

7: Overall evaluations of carcinogenicity an updating of IARC Monographs Volumes 1 to 42. International Agency for Research on Cancer, Lyon, pp 17-32.

IARC (1987d): 2,3,7,8-Tetrachlorodibenzo-*para*-dioxin (TCDD) in IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Supplement 7: Overall evaluations of carcinogenicity an updating of IARC Monographs Volumes 1 to 42. International Agency for Research on Cancer, Lyon, pp 350-354.

IARC (1992a): Preamble in IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 54. *Occupational Exposures to mists and vapours from strong inorganic acids; and other industrial chemicals.* International Agency for Research on Cancer, Lyon, pp 13-32.

IARC (1992b): Diethyl sulfate in IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 54. Occupational Exposures to mists and vapours from strong inorganic acids; and other industrial chemicals. International Agency for Research on Cancer, Lyon, pp 213-238.

IARC (1995): *Peroxisome Proliferation and its role in Carcinogenesis.* IARC Technical Report No. 24. International Agency for Research on Cancer, Lyon.

IARC (1997): IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 69. Polychlorinated dibenzo-*para*-dioxins and polychlorinated dibenzofurans. International Agency for Research on Cancer, Lyon.

IARC (1998): Report of an ad-hoc *IARC Monographs* advisory group on priorities for future evaluations. IARC Internal Report No. 98/004. International Agency for Research on Cancer, Lyon.

IARC (1999): IARC Monographs on the Evaluation of Carcinogenic Risks to Humans, Volume 73. Some chemicals that cause kidney or urinary bladder tumours in rodents, and some other chemicals. International Agency for Research on Cancer, Lyon.

Jackson, M.A., Stack, H.F. and Waters, M.D. (1993): The genetic toxicology of the putative nongenotoxic carcinogens. *Mutat. Res.*, **296**, 241-277.

Pagurek, B., Dawes, N., Bourassa, G., Evans, G. and Smithers, P. (1990): Letter Pattern Recognition, Proc. Sixth Conf. on Artificial intelligence Applications, IEEE, pp 313-319.

Parodi, S. and Waters, M.D. (1991): Introduction and summary. Genotoxicity and carcinogenicity databases: An assessment of the present situation. *Environ. Health Perspect.*, **96**, 3-4.

Pavlidis, T. (1993): Recognition of Printed Text under Realistic Conditions, Pattern Recognition Letters, **14**, 317-326.

Rice, S., Kanai, J. and Nartker, T. (1992): A Report on the Accuracy of OCR Devices, Technical Report ISRI TR-92-02, University of Nevada, Las Vegas.

Tice, R.R., Stack, H.F. and Waters, M.D. (1996): Human exposure to mutagens - an analysis using the Genetic Activity Profile Database. *Environ. Health Perspect.*, **104**, 585-589.

Waters, M.D., Stack, H.F., Rabinowitz, J.R. and Garrett, N.E. (1988a): Genetic activity profiles and pattern recognition in test battery selection. *Mutat. Res.*, **205**, 119-138.

Waters, M.D., Stack, H.F., Brady, A.L., Lohman, P.H.M., Haroun, L. and Vainio, H. (1988b): Use of computerized data listings and activity profiles of genetic and related effects in the review of 195 compounds. *Mutat. Res.*, **205**, 295-312.

Waters, M.D., Claxton, L.D., Stack, H.F., Brady, A.L. and Graedel, T.E. (1990a): Genetic activity profiles in the testing and evaluation of chemical mixtures. Teratogen. *Carcinog. Mutagen.*, **10**, 147-164.

Waters, M.D., Stack, H.F., Garrett, N.E. and Jackson, M.A. (1991): The genetic activity profile database. *Environ. Health Perspect.*, **96**, 41-45.

Waters, M.D., Richard, A.M., Rabinowitz, H.F., Stack, H.F., Garrett, N.E., Lohman, P.H.M. and Rosenkranz, H.S. (1993a): Structure-activity relationships--computerized systems in Proceedings of the Scientific Group on Methodologies for the Safety Evaluation of Chemicals (SGOMSEC) -8 Workshop on Cross-Species Differences in DNA Damage and its Consequences. Research Triangle Park, NC.

Waters, M.D., Stack, H.F., Jackson, M.A., Bridges, B.A. and Adler, I.-D. (1994): The performance of short-term tests in identifying potential germ cell mutagens: a qualitative and quantitative analysis. *Mutat. Res.*, **341**, 109-131.

Waters, M.D., Stack, H.F. and Jackson, M.A., (1999): Short-term tests for defining mutagenic carcinogens in *Results of Short-and Medium-term Tests for Carcinogens and Data on Genetic and Related Effects in Carcinogenic Hazard Evaluation*, IARC Scientific Publication No. 146 (D. McGregor, J.M. Rice and S. Venitt eds.). International Agency for Research on Cancer, Lyon, pp 499-526.

Xu, Y. and Nagy, G. (1999): Prototype Extraction and Adaptive OCR, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21, pp 1280-1296.