Introduction. Symposium on User Reactions to CAS Data and Bibliographic Services

C. H. O'DONOHUE

Philip Morris Research Center, Richmond, Virginia 23261

Received May 28, 1975

In previous years the Division of Chemical Information has held symposia which were critiques of the then available information services and systems.[1] These symposia have covered a range of information services and have given objective evaluations of the systems' advantages and disadvantages. This symposium differs from the earlier ones in that it is limited to the evaluation of one information service, i.e., Chemical Abstracts Service (CAS). The data and bibliographic services offered by CAS have undergone multiple changes in the recent years. These events prompted this symposium which is devoted to the use and evaluation of the CAS files.

The broad objectives of this symposium were: (1) to survey the services that are currently available from CAS; (2) to establish which services are most useful to the information community; (3) to ascertain which features aid in ensuring that chemical information is accessible to those who have a need for it; and (4) to evaluate the disadvantages which the user in his work environment felt affected his use of these services.

The following papers were organized to provide a broad-based evaluation of CAS, from the use of the printed edition of CA to the on-line bibliographic files. The lead paper from CAS[2] gives background information on what CAS has tried to accomplish in recent years and how they view their changes and the effect on providing information to the ultimate user. The remaining papers attempt to bring out the advantages and disadvantages as well as the cost effectiveness of the various CAS services. It is from these experiences and discussions that it is hoped the user of chemical information will benefit.

LITERATURE CITED

(1) Vasta, B. M., "Introduction. Comparative Evaluations of Existing Chemical Information Services Critique Symposium," *J. Chem. Doc.*, **13**, 23 (1973).

(2) O'Dette, R. E., "CAS Data Base Concept," *J. Chem. Inf. Comput. Sci.*, **15**, 165 (1975).

The CAS Data Base Concept†

RALPH E. O'DETTE

Chemical Abstracts Service, The Ohio State University, Columbus, Ohio 43210

Received June 4, 1975

The evolution of the CAS processing system is summarized, stressing the data base concept. Current publications and services are identified in relation to the data base, and some of the potential for their evolution is noted.

I was asked to introduce the symposium by (1) providing a summary and overview of Chemical Abstracts Service (CAS) processing technology with particular emphasis on the data base concept as CAS is applying it; (2) defining current CAS information products, relating them to each other and to the data base; and (3) outlining trends for the future, as we now see them, particularly in terms of kinds of services that users might expect from CAS. Further information on the current CAS production system and the kinds of services available may be obtained by consulting "CAS Today. Facts and Figures about Chemical Abstracts Service."[1]

† Presented in the symposium "User Reactions to CAS Data and Bibliographic Files," 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 7, 1975. The development of many techniques and applications described in this paper were partially supported by the National Science Foundation under Contract NSF-C656, as well as earlier contracts and grants. Chemical Abstracts Service, a division of the American Chemical Society, gratefully acknowledges this support.

COMPUTERIZATION OF CAS

Figure 1, which was created for the 1969 CAS Open Forum in New York,[2] describes the concept of the computerization of CAS as we saw it at that time. The left part of the figure, 1966, is the precomputer system; the right part, 1974–78, is the "ultimate system." The box at the left end of the wiggly line represents receipt of a source document; the box at the other end of the line is the publication of indexed abstracts. The diamonds on the line are processes that must be performed. The stars, which represent transcription, proofreading, and correction steps, are shown with loops because the information does not move forward in these processes; it is simply cleaned up so that it can move forward. An operations analysis which we conducted in the early 1960's described approximately 35 separate steps through which a typical source document was processed between its receipt at CAS and its final indexing.

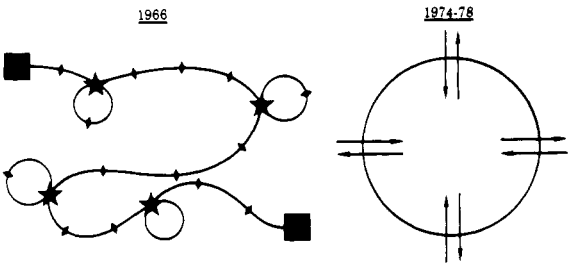

Figure 1. The CAS workflow.

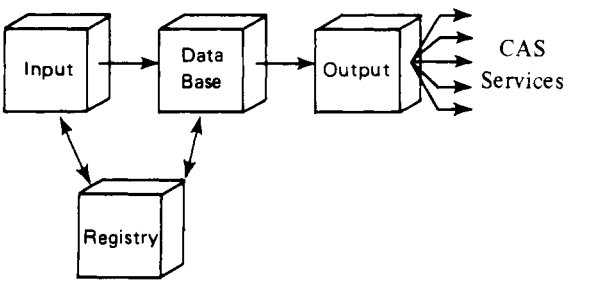

Figure 2. The CAS production system.

Obviously, the process was highly redundant, but redundancy in a manual system is essential for accuracy. The manual process was hierarchical in the sense that each step improved on the work of the steps before. To use modern terminology, it was a sequential batch-oriented system, and the batches were shopping carts full of paper.

The system toward which CAS has been evolving was described in 1968 by the circle and the arrows shown in the right part of Figure 1. This "ultimate system" is an on-line interactive system—not in the sense of subscribers conducting dialogue with the CAS data base, but in a production sense. We projected then that our staff should be on-line to the information being processed so that the sequence of 35 steps in the old system would be reduced to about a half-dozen steps.

There has to be some sequence, of course. A document must be received before it can be identified; then it must be analyzed; then the analyses must be assembled to create a product. The important point is the contrast between the two parts of the figure. In the planned system, the amount of redundancy is vastly reduced; therefore, intellectual staff is used much more efficiently.

Whether we achieve complete on-line processing is no longer viewed as a technological question as it might have been when Figure 1 was first drawn in 1968. It is now an administrative, cost-effectiveness question. We have some on-line subsystems in operation currently and more in development.

CAS PROCESSING SYSTEM

Today, it is most useful to view the processing system at CAS in terms of Figure 2. There are three major subsystems: input, the data base, and output. Input is largely a human function with computer support; output is largely a computer function with very little human intervention. The data base consists of machine-readable, in-process files. At input, we receive source documents, and we select those that are to be included in the data base. We record the bibliographic identification in the computer, and then our staff performs what we call "integrated" or "unified" analysis which provides the abstracts and the index entries which are then recorded in the computer. Bibliographic identifications and analyses are recorded as a stream of data elements—precisely defined small units of information—which, recombined in various ways, constitute bibliographic identifications, abstracts, and index entries as output in products.

The CAS data base is not an archival file, but is, rather, a large number of in-process files. Data base in this sense is a flow-through system. It is where information resides between the time of input and the time of final use in a product or products.

Output includes computer programs and hardware. Each information product is represented by a set of programs which, operating against the data base files, select the correct subset of data elements and combine them in the appropriate manner for that particular product. Products may be printed volumes, microfilm, or microfiche files of computer-readable tapes, and they can be available in a combination of media. Such matters as the typography of a printed product, hyphenation, and formatting are determined by the program, not by any instructions at the keyboard when the information is input. Similarly, tape formats are determined at the output operation.

CAS Chemical Registry System. I do not plan to describe all the CAS processing technology here since processing technology is not the purpose of my remarks. But Figure 2 indicates one extremely important step of processing technology, the CAS Chemical Registry System,[3,4] which must at least be briefly summarized because of its effect on current CAS services and its role as a source of future services. Registry is an automated, molecular structure recognition system which acts to assure precision in the assignment of terminology for indexing substances at CAS. It is an automated vocabulary control system. The CAS Registry System has been operating since the end of 1964 when Chemical-Biological Activities (CBAC) which began publication in 1965 became the first service to publish Registry Numbers.

One aspect of Registry operation has been the building of two sets of files which have considerable potential significance as the basis for new user services. One of these files contains molecular structure representations called connection tables, which include all of the structural details obtainable from the source document. Currently over 3 million different substances are recorded in that file. There is also a file of names, including the CA systematic nomenclature used for indexing (CA Index Names), plus the synonyms—author nomenclature—from the primary literature.

CAS DATA BASE

To examine the CA data base in more detail, it is useful to think of three classes of files: bibliographic information, abstract text, and files containing data elements which are the source of subject matter indexes. Subject index content, not CA subject indexes, reside in the data base. Different, overlapping subsets of these data elements comprise the Formula, General Subject, and Chemical Substance Indexes to *Chemical Abstracts* (CA).

Figure 3 shows Registry files as separate from but closely related to the data base, simply because at CAS we do not consider the Registry files as an integral part of what we view as "the CAS data base." We view Registry files as a data base in their own right—as an "authority data base"—obviously, very closely related to the CAS "document-oriented" data base because Registry files are used in the control of the index terminology which is used in the subject indexes, synonyms which go into the CA Index Guide, molecular formulas, Registry Numbers, and other information that is part of the data base.

Table I summarizes some characteristics of the information that is currently flowing through the CAS system. To CAS, a source document may be an individual paper, a patent, a technical report, a dissertation, a chapter from a monograph, or an individual paper given at a conference. We consider an issue of a journal to be a document pack-

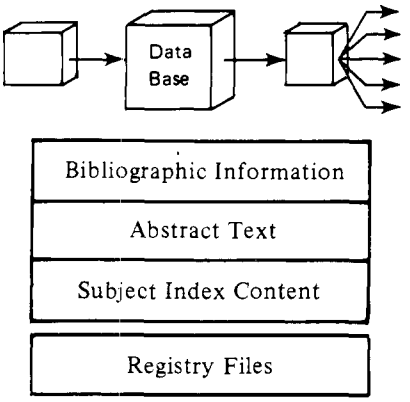

Figure 3. Data base information content.

Table I. Source Documents

	Number of Documents
Examined	2,400,000
Selected	376,000
Abstracted	334,000
Cross-referenced	42,000

Periodicals and other reports - 82%
(from 153 countries in 51 languages)

Patents (from 26 nations) - 18%

age. As shown in Table I, our selection and assignment staff in 1974 examined 2.4 million source documents; these documents were already the product of a selection process in that our acquisition group, surveying worldwide scientific and technical literature, had selected certain packages as candidates for CAS coverage. From the 2.4 million documents examined, 376,000 were selected because they were pertinent to CAS in terms of subject content. Of these, 334,000 were abstracted because they were, in fact, documents which we had not abstracted before; they were new documents. The remaining 42,000 (376,000–334,000) were equivalent patents. CAS policy is to abstract the first patent that we receive on an invention, and when that invention is subsequently patented in other countries, we recognize the subsequent patents as unique documents but do not again abstract them—we cross-reference duplicate patents in the Patent Concordance.

As Table I shows, 82% of the 334,000 abstracts came from nonpatent literature; the serial literature component—a very large part of the total 82%—consisted of papers reporting research, development, and applied technology performed in 153 different countries, and published in 51 languages. Those numbers fluctuate from volume to volume of CA, but between 145 and 155 countries and between 45 and 55 languages are generally represented. The remaining 18% of the 1974 abstracts are of patents from the 26 countries whose patents CA routinely covers.

The CAS definition of subject coverage is treated in detail in a CAS publication, "Subject Coverage and Arrangement of Abstracts by Sections in *Chemical Abstracts*."[5] This publication is an outgrowth of a staff operations manual. Each of the 80 CA sections is defined in detail, including identification of areas at the periphery of our coverage which are not to be included in CA.

Figures 4–9 describe the major information products of CAS in terms of what portions of the data base they contain. As Figure 4 shows, CA includes bibliographic identifications, abstracts, issue indexes, and volume indexes. The

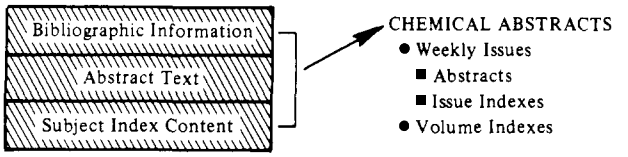

Figure 4. Data base information content.

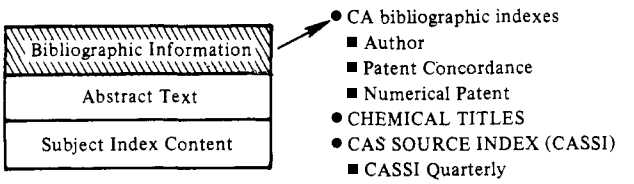

Figure 5. Services from bibliographic data.

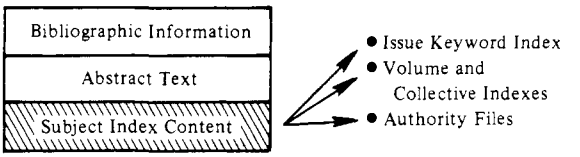

Figure 6. CA subject content.

weekly CA issues plus the volume indexes contain all the information in the CAS data base.

Figure 5 shows the series of CA bibliographic indexes which are derived by computer programs from the bibliographic data component of the data base. These indexes include both the Author Index for each issue and the rather different Author Index for each volume of CA; both are drawn from the same stream of bibliographic data. The Patent Concordance and the Numerical Patent Index are also considered bibliographic indexes. *Chemical Titles* and the *CAS Source Index* (CASSI) are also derived from the bibliographic information component of the data base.

From Figure 6, we see that three classes of products are derived from the subject matter index component of the data base: the Keyword Index for each CA issue, the subject indexes for each volume and collective period, and what are variously called authority files, dictionary files, or "indexes to indexes." The volume and collective subject indexes include the General Subject Index, the Chemical Substance Index, and the Formula Index. The most important current illustration of a published authority file is the CA Index Guide with its supplements.[6] The Guide contains cross-references, synonyms, indexing policy notes, and instructions for using CAS products. A search of CA subject indexes since Volume 76 cannot be conducted rationally without the use of the Index Guide, and in that sense the use of *guide* in the title has an incorrect semantic implication for some users. Unless Ninth Collective Index terminology is known both for substances and for concepts, failure to use the Index Guide and its supplements virtually guarantees an incomplete or even zero result search. The Index of Ring Systems and the Registry Handbook are other examples of authority files which are produced by combining selected data elements from the subject matter index component of the data base.

Figure 7 depicts services which are best viewed as vertical slices of the data base; that is, they contain the bibliographic identifications, abstracts, and subject index entries pertinent to some subset of sections of CA. When CBAC and POST were created (CBAC in 1965 and POST in 1967), the state of CAS processing technology required any such special abstract publication to consist of a group of adjacent sections of CA. Actually, at first, CBAC and POST were produced virtually independently of CA. When CBAC and POST were later integrated with CA and made

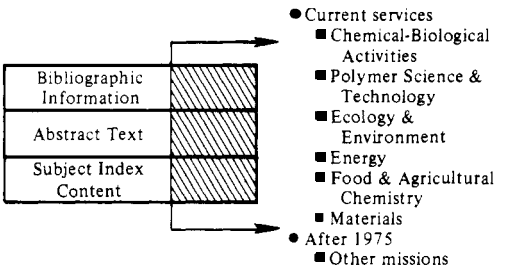

Figure 7. Computer-readable services.

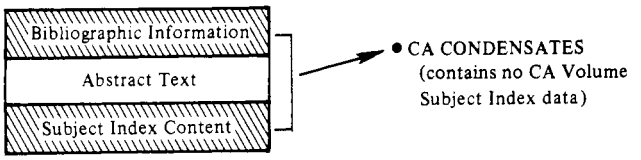

Figure 8. Computer-readable services.

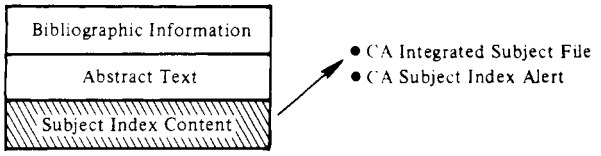

Figure 9. Computer-readable services.

up of CA sections, the sections had to be contiguous. With the new services of that type offered for 1975 (Ecology and Environment, Energy, Food and Agricultural Chemistry, and Materials), that concession is no longer necessary, nor is it observed in the 1975 version of CBAC. CAS processing technology has matured to the point where we can select and assemble any combination of sections. The sections do not have to be adjacent to each other, and it is possible to duplicate a section in more than one service; for example, at present, Section 4, Toxicology, is included in CBAC, in Ecology and Environment, and in Food and Agricultural Chemistry. We think that these capabilities have important implications for deriving information services for missions which are entirely or to a significant extent within the subject scope of the CAS data base. We now have the technology to assemble that subset of sections which is appropriate to a given mission. However, these services are available in computer-readable form only. The economics of printing are a different matter from the technology of assembling a data file.

Probably the most widely known computer-readable file from CAS is *CA Condensates* (see Figure 8) which may be thought of as a computer-readable issue of CA without the abstracts. *CA Condensates* contains all of the bibliographic identifications and the Keyword Index phrase material, issue-by-issue. *CA Condensates* does not contain volume index material.

When CAS first began to use computers, roughly 15 years ago, one of the most important needs that users voiced was for the ability to search the Subject Index by computer. This ability was not achievable overnight, but it is now a two-faceted reality (Figure 9): the CA Integrated Subject File (CAISF) is the computer-readable equivalent of the volume indexes; that is, it is issued roughly every six months and is virtually a computer-readable image of the printed index with the hierarchical structure of the index and other form-related characteristics. The CA Subject Index Alert (CASIA) is a completely different approach to the same component of the data base. CASIA is a biweekly computer-readable issuance of the Subject Index entries, arranged in abstract number order, for each document for which index entries were released to the data base in the previous two weeks. CASIA contains the same subject index entries that will later appear in the printed volume subject index and in the Integrated Subject File, but in CASIA they appear, on the average, about nine months earlier; thus CASIA provides both current awareness and subject-index-in-depth. Because of the large number of entries that contain CAS systematic nomenclature for substances, CASIA also lends itself to searching for molecular substructures by means of nomenclature fragments.

Table II. Help for Users

- Subject Coverage Manual
- Introduction in publications
- Index Guide
- SDF Specifications Manual
- Special user aids
- User education program

CONCLUSION

Two topics will serve to complete this introduction to the CAS data base: aids to the use of CAS products and a brief glimpse of the future.

User Aids. Table II lists the main areas of current development at CAS. The subject coverage manual was discussed earlier. The introductions to all of the printed CAS publications have been considerably upgraded over the last several years. Now they contain enough detailed information to serve as a guide for using the publications they accompany. The CA Index Guide, which was also mentioned previously, is an integral part of the indexes to CA, and the Introduction to the Index Guide is especially informative, discussing the structure of the CA Subject Indexes and including an extended introduction to use of systematic chemical nomenclature in CA.

For users of CAS computer-readable files, there is the three-volume loose-leaf Standard Distribution Format Specifications Manual.[7] All CAS computer-readable products are distributed in the CAS Standard Distribution Format. A set of 13 experimental user aids is now available. These aids are intended to support use of CAS computer-readable files. They are called "experimental" because we are not as yet certain which kinds of aids will prove most useful and still consider this activity to be in the development stage. A subset of those aids specifically dedicated to CA *Condensates* has been announced.[8,9]

CAS has also formally established a User Education Program in which we will coordinate the capabilities we have and develop new capabilities, including a program of seminars, tutorials, and so on to help users to use CAS information products more effectively.

Future. One facet of future CAS service not so far touched on is the use of CAS files, processing technology, or some combination of files and processes to perform substructure searches. Until recently the emphasis in the content of CAS computer-readable files has been on other than molecular structure information. However, a number of different CAS development paths have converged to make accessible current data streams that permit searches for molecular structure information (substructure search) coordinated with searches for information about those substances and with bibliographic information identifying the source documents.[10]

Several organizations with whom CAS has been cooperating have in operation or in advanced development various approaches to the practical application of their techniques.

LITERATURE CITED

(1) "CAS Today. Facts and Figures about Chemical Abstracts Service," Chemical Abstracts Service, Columbus, Ohio, 1974, 32 pp.
(2) "Report on the Twelfth Chemical Abstracts Service Open Forum" (New York, N.Y., Sept 7, 1969), Chemical Abstracts Service, Columbus, Ohio, 1970, 22 pp.
(3) Dittmar, P. G., "Derivation of the Registry III Structure Record," Chemical Abstracts Service, Columbus, Ohio, Jan 1974 (available through the National Technical Information Service, Springfield, Va., PB-232-928/AS).
(4) "Progress in Building a Chemical Registry System," Chemical Abstracts Service, Columbus, Ohio, Aug 1974 (available through the National Technical Information Service, Springfield, Va., PB-235-125/AS).
(5) "Subject Coverage and Arrangement of Abstracts by Sections in *Chemical Abstracts*," 1975 ed, Chemical Abstracts Service, Columbus, Ohio, 180 pp.
(6) "CA Index Guide," Chemical Abstracts Service, 1975, LCCC 9-4698.
(7) "Chemical Abstracts Service Specifications Manual for Computer-Readable Files in Standard Distribution Format," Chemical Abstracts Service, Columbus, Ohio, Dec 1974, 1300+ pp, ISBN 8412-0150-1, LCN 72-87124.
(8) "Chemical Abstracts Service Search Aids for the 9th Collective Index Period (1972–1976)," Chemical Abstracts Service, Columbus, Ohio, June 1974, 109 pp, ISBN 8412-0198-6, LCN 74-80986.
(9) "Substructure Searching of Computer-Readable CAS 9CI Nomenclature Files (based on Nomenclature in the Ninth Collective Index of *Chemical Abstracts)* (1972–1976), Chemical Abstracts Service, Columbus, Ohio, Aug 1974, 128 pp, ISBN 8412-0204-4, LCN 74-14778.
(10) Fisanick, W., Mitchell, L. D., Scott, J. A., and Vander Stouw, G. G., "Substructure Searching of Computer-Readable Chemical Abstracts Service Ninth Collective Index Chemical Nomenclature Files," *J. Chem. Inf. Compt. Sci.*, **15**, 73 (1975).

Replacement of an In-House Current Awareness Bulletin by *Chemical Abstracts* Section Groupings†

JEAN S. PETERSON

Contribution No. 2263 from the Central Research and Development Department, E. I. du Pont de Nemours and Company, Experimental Station, Wilmington, Delaware 19898

Received March 31, 1975

An in-house current awareness bulletin that had been published for more than 50 years was replaced with *Chemical Abstracts* Section Groupings. Net cost savings are about $150,000 a year. Most of the clients are pleased or at least satisfied with the change.

The Central Research Department of the Du Pont Company published an in-house current awareness bulletin for more than 50 years. This paper reports the results of a study carried out in 1971–1972 on the feasibility of replacing the bulletin with a commercially available service, *Chemical Abstracts* Section Groupings.

CURRENT JOURNAL BULLETIN

The *Current Journal Bulletin* (or CJB) was a weekly publication of abstracts of articles in current issues of scientific periodicals. Because the interests of the Du Pont Company are broad, the bulletin covered chemistry, physics, biology, and engineering. To facilitate use of the bulletin, abstracts were assigned to one or more of 20 classes (Table I). A detailed subject guide (Table II) was included with each issue to guide users to specific topics of interest.

A very popular section of the bulletin was the Highlights Section which selected items closely related to Du Pont research interests or of particular scientific interest, such as the first reports of the helical structure of DNA and the total synthesis of chlorophyll.

The abstracts (Figure 1) were either edited author abstracts or abstracts prepared directly from the articles. Original abstracts were written for articles from foreign language journals and for articles that did not have author abstracts such as the communications in the *Journal of the American Chemical Society*. At various times we covered journals in the German, French, Italian, Dutch, and Russian languages. The bulletin was designed to be easy to scan with short abstracts, liberal use of chemical structures, short underlined headings in capital letters, and citations following the abstracts. The final version of each issue was typed on large pages and inexpensively reproduced by a photoreduction technique. The bulletin was very popular with research people. In all surveys to evaluate information sources, the CJB was rated as the most useful source.

Over the years, like the literature itself, the bulletin grew. In 1959, the bulletin covered about 800 journals and published about 50,000 abstracts a year, about 1000 per issue. Coverage included nearly all the scientific periodicals subscribed to by the Lavoisier Library, Du Pont's largest collection of technical literature. By 1972, as journals increased rapidly in size and frequency of issue, coverage was reduced to 275 journals but even then we were publishing 60,000 abstracts a year, an average of 1200 per issue.

These 275 journals were selected by the CJB staff as being most representative of research interests. We tried to choose broad, basic journals rather than narrowly specialized ones, journals that would report work of interest to many of our scientists rather than to three or four on a specific project. It was a hard job, a very unsatisfactory one. We had some help from our clients who would send recommendations for adding or deleting journals, mostly adding. The staff had grown to a supervisor, seven full-time technically trained abstractors, and ten clerical workers. The Library was subscribing to about 1300 journals, so the bulletin was fighting a losing battle trying to handle the current literature and furnish current awareness to many scientists in numerous areas of research.

† Presented in the Symposium on "User Reactions to CAS Data and Bibliographic Services," 169th National Meeting of the American Chemical Society, Philadelphia, Pa., April 7, 1975.