ChemInform—An Integrated Information System on Chemical Reactions

AXEL PARLOW* and CHRISTIAN WEISKE
Fachinformationszentrum Chemie GmbH, Steinplatz 2, D-1000 Berlin 12, FRG

JOHANN GASTEIGER
Technical University Munich, Lichtenbergstrasse 4, D-8046 Garching, FRG

Received June 1, 1990

ChemInform, an abstracting service issued weekly by FIZ CHEMIE and BAYER, reports recent publications in the field of preparative organic, organometallic, and inorganic chemistry. This information will be made available in a reaction database which can be implemented with the common reaction retrieval systems. The database will grow by 70000 reactions annually. The combination of reaction retrieval with graphic displays of the scheme of a whole reaction "tree" including variations constitutes a unique information system on chemical syntheses.

INTRODUCTION

This paper will present the developments at FIZ CHEMIE, Berlin, and its cooperation partners to build up an integrated system on information on chemical reactions. While there are—or will be in the near future—other reaction databases on the market which collect the information from the published literature completely (CASREACT, BEILSTEIN, ZIC), the ChemInform reaction database contains the most important and interesting new trends in chemistry selected by experts, thereby preventing the user from being flooded with unwanted information. The process of scientifically evaluating and processing the literature also differentiates ChemInform from other in-house databases.

THE SOURCE: *CHEMINFORM*

ChemInform is a weekly published information service containing abstracts of recent papers drawn from about 250 scientific journals which have been recommended to evaluate by an Advisory Board consisting of chemists from universities and companies.

The main aspects for an inclusion in *ChemInform* are the following:

- new reactions and syntheses (including biochemical aspects)
- application of known reactions to the synthesis of new substances or classes of substances
- improved methods, new reagents or catalysts
- syntheses of important natural products
- syntheses of novel organoelement compounds

Reviews of the above-mentioned topics are also considered in the printed issue, but will not be fed into the reaction database, as well as the information from papers touching other areas of high general interest:

- theoretical, physicochemical, and analytical studies
- application-directed work, advanced materials
- structural studies of novel natural products
- interdisciplinary topics

ChemInform abstracts are brief English summaries of the most important aspects of the papers elucidated by structural presentations of reaction diagrams. They concentrate on purpose, results, and conclusions, whereby an increase of knowledge to the chemical community must be obvious. Figure 1 gives a typical example for a *ChemInform* abstract.

In this way *ChemInform* has grown to an up-to-date, easy to grasp, and scientifically sound information service for chemists in chemical research, with emphasis on preparative organic and organometallic chemistry.

The weekly issue of *ChemInform* consists of about 400 abstracts. That means about 20000 abstracts are published annually giving information on about 70000 individual chemical reactions and 145000 compounds per year. Besides its being a rich source of quantity in information on chemical reactions, *ChemInform* also puts strong emphasis on the quality of its data. This builds on its long-standing tradition in providing chemical information and is also a result of the unique structure of its editorial staff. Each issue of *ChemInform* is accompanied by author and classification term indexes. A microfiche version of an annual author index is produced, too.

ChemInform has its roots 20 years ago in the "Schnellreferate" of the "Chemisches Zentralblatt" and in the progress reports of the BAYER AG. In 1970, the West Berlin editorial office of Chemisches Zentralblatt became the division "Chemie-Information und Dokumentation Berlin" (CIDB) of the German Chemical Society. It intertwined parts of its staff with that from BAYER to launch *Chemischer Informationsdienst*, now named *ChemInform*. In 1982, the West Berlin part became the "Fachinformationszentrum (FIZ) Chemie". This long history has provided *ChemInform* with a highly competent staff. At FIZ CHEMIE the abstracts are written by full-time employees which are indexing experts, while at BAYER they are created by chemists working at the lab bench. This well-balanced mixture of personnel with a long-standing experience in indexing and a close contact to the experimental chemistry ensures the high and reliable quality of the abstracts. Critical supervision and consultancy is further provided by a scientific advisory board with members from universities and the chemical industry.

THE PROCESS: THE BUILDING OF THE SYSTEM

Because the information in *ChemInform* is essentially well-suited for a reaction database, since 1988 work is in progress to implement an integrated production system for both the printed issue and the database.[1-4] The project is funded by the German Minister of Research and Technology (BMFT) and is administered by the Gesellschaft für Mathematik und Datenverarbeitung (GMD). Work is done at several places: at FIZ CHEMIE in Berlin, at BAYER AG in Leverkusen, at CHEMODATA in Munich, and at GTS-GRAL in Darmstadt. The project is coordinated by Prof. J. Gasteiger of the Technical University in Munich.

Data Input. A computerized database and a printed publication have different characteristics and requirements. While a database should contain as much information as possible on the fact which it is describing, a printed abstract has to concentrate on the most important items and has to combine different facts into combination data or even sacrifice some information for the sake of clarity of presentation. This is

0095-2338/90/1630-0400$02.50/0 © 1990 American Chemical Society

Oxazole Derivatives

R 0220

8913-154

Synthesis and Thermolysis of Cycloalkenyl Azides and Iminophosphoranes Aldehydes. A New Pathway to Bisanellated Pyridines. — The 2-chloro-1-formylcycloalkenes (III), obtained from the cycloalkanones (I) by Vilsmeier reaction, are converted to the azides (V) which release nitrogen upon heating to form the fused isoxazoles (VI). Reaction of (V) with triphenylphosphine (VII) gives the iminophosphoranes (VIII) which dimerize at higher temperatures, producing the pyridines (IX). — (TABYAOUI, B.; AUBERT, T.; FARNIER, M.; GUILARD*, R.; Synth. Commun. 18 (1988) 13, 1475—1482; Lab. Synth. Électrosynth. Organométt., Ass. CNRS, Fac. Sci. Gabriel, 21100 Dijon, France; Eng.) — Heß

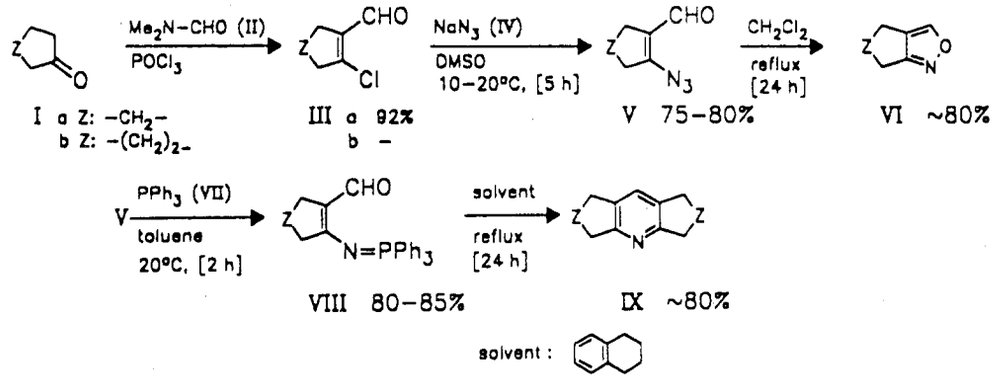

Figure 1. *ChemInform* abstract (Ref. 8913-154).

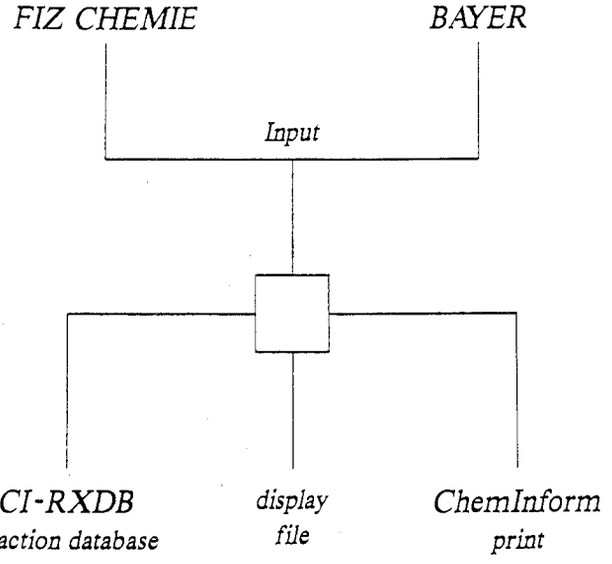

Figure 2. Input data flow.

important to keep in mind when establishing a process to build the two products from one source.

The process which is implemented now generates first the complete information for the database (Figure 2). An input module (developed by BAYER) collects and verifies the entered data, builds up connection tables for the molecules, and outputs the information in SMD files. SMD is rapidly becoming an international standard for the dissemination of information on chemical structures and reactions. The information flow then splits into two ways: towards the database and the printing process (see below).

In a separate process the textual data are keyboarded and kept in similar files. The textual part contains the abstract text, bibliographic information, location of work, language of the paper, and the abstractor's name. To each abstract, classification numbers are also assigned. The numbers and terms are published in a classification system.[5]

Preparation for Print. A layout module (developed by GTS-GRAL) takes the information from the SMD files and produces a graphic metafile with the complete scheme (see Figure 1) for the subsequent typesetting process. The main topics it has to deal with are the chaining of multistep syntheses, i.e., combining reactions that start up with a reactant synthesized in a previous reaction into one reaction pathway, and the elimination of redundant information, e.g., duplicate or partly overlapping legends. This module performs its task largely automatically, but the prepared scheme can be customized by the user.

Graphic schemes and textual data are combined within the typesetting process. The graphic metafiles are also retained for use as display file in the database.

Data Processing and Enrichment. The programs for further processing the data from the SMD files are being developed at CHEMODATA. These modules enrich the mere structural, factual (yield, reaction conditions, etc.), and textual data from the input process with additional—computer-generated—information on the structures, reactions, and reaction sequences. These include the following:

additional molecule data
- stereochemical descriptors
- tautomer representation (individual and generic)
- physicochemical parameters (partial charges, bond dissociation energies, etc.)
- keywords for the molecules involved

additional reaction data
- balancing of reaction equations
- perception of reaction centers
- classification of the bond rearrangement scheme
- reaction keywords

and data for the overall scheme
- multistep chaining
- separation of parallel reactions

After these enhancements, the data—including the textual information—are kept in an internal workfile.

THE PRODUCT: *CHEMINFORM* PRINT AND *CHEMINFORM* DATABASE

The printed issue of *ChemInform* is produced weekly by the above-mentioned typesetting process.

The data in the workfile are subject to a conversion process into the file format for a reaction retrieval system. Conversion procedures exist for REACCS (by Molecular Design Ltd.)

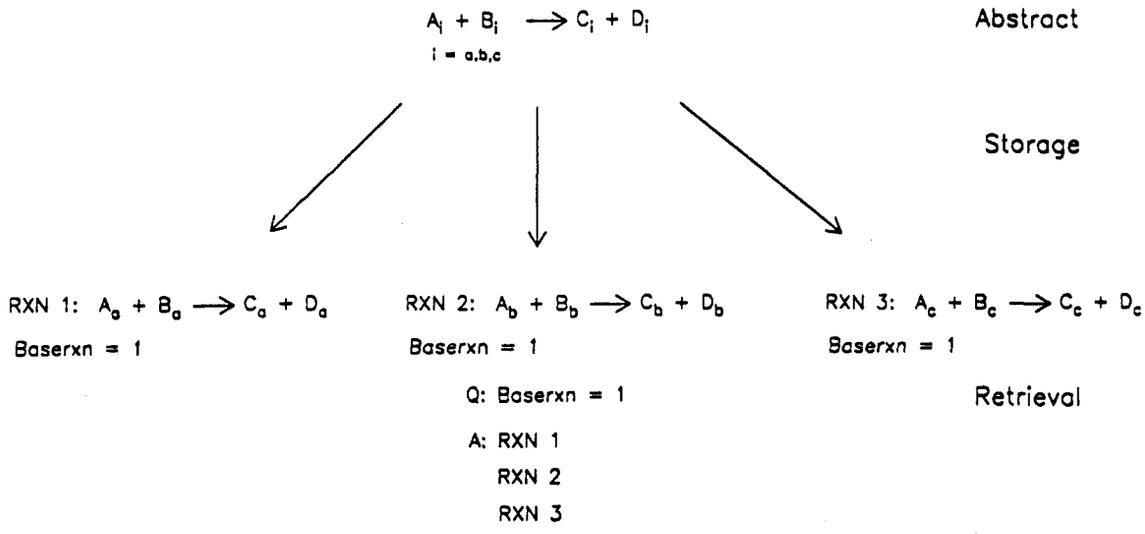

Figure 3. Storing and retrieving reactions with substituent variations.

and ORAC (by ORAC Ltd.), the most common in-house systems on the market, and are developed for Messenger to implement the data online on STN.

While the workfile contains as much information as possible, the conversion routines must skip the data the target system cannot handle at present. As the systems are enhanced, more information will be available.

To make full use of all the information contained in the database, the retrieval system should handle and make searchable the following items:

Molecules:

- full structure and substructure including stereochemistry
- tautomer form, class of tautomers
- alphanumeric data (keywords, formula)

Reactions:

- all reaction participants structurally (also solvents, catalysts)
- reaction centers, bond rearrangement scheme
- yield or conversion, stereoselectivity, enantiomeric and diastereomeric excess
- conditions (temperature, time, etc.)
- other numeric data: bond dissociation energies, etc.
- alphanumeric data: keywords (e.g., mechanism)

Document-based information:

- bibliographic data (authors, publication, etc.)

UNIQUE FEATURES NOT OR ONLY PARTIALLY AVAILABLE IN OTHER REACTION DATABASES

The ChemInform reaction database gives state-of-the-art information on chemical reactions and can be seen as a window into the current chemical literature. Its special features are

- full stereochemistry up to coordination number eight
- explicit coding of organometallic compounds
- individual and generalized coding of tautomers
- reaction site coding and bond rearrangement scheme
- coding of electronic and energetic parameters
- treatment of generic and multistep reaction schemes

From the data contents, a chemical reaction in the *ChemInform* database is not an isolated event but treated within broader perspectives. Due to the fact that the whole *ChemInform* abstract, i.e., the excerpt from a publication, is the basic information unit, reactions that are described within one article are tied together very closely.

This, for instance, makes it easy for users to get an impression of any other substituents used in an analogous manner. As soon as a (generic) reaction with several variations is assembled into the database by storing the individual reactions, a base reaction ID number is provided, and all variations (which have a pointer to that base reaction) are retrieved easily. Figure 3 illustrates this point.

An individual reaction can also be viewed in its chemical environment by looking at the entire sequence of a multistep synthesis tree. Pointers for preceding and succeeding reactions are provided for this access.

The most impressive way, however, to grasp the whole environment at a glance is to access the graphic metafiles, which are also available with the ChemInform database. A retrieval system that makes provision for external graphics can show the entire scheme of an abstract (see Figure 1). Vendors of such systems are encouraged to enhance their software in this point. While a PC screen is suitable for schemes containing one or two lines of reactions, it is insufficient for three or more lines. Either a high-resolution workstation or a graphics display environment with zooming or scrolling functionality would be necessary.

ACKNOWLEDGMENT

We thank the German Minister of Research and Technology (BMFT) and the GMD-PTF for supporting and handling the project. All people involved in discussions are thanked for their suggestions. Our thanks also to the staff of the development teams, who are doing a great job.

REFERENCES AND NOTES

(1) Christoph, B. Graphische Eingabe von chemischen Reaktionen unter Verwendung von GKS auf Mikrocomputern. In *Software-Entwicklung in der Chemie 1*; Gasteiger, J., Ed.; Springer: Berlin, 1987 pp 13–16.
(2) Parlow, A. *ChemInform*—Printed Issue and Database. In *Chemical Structures*; Warr, W. A., Ed.; Springer: Berlin, 1988; pp 409–411.
(3) Blücher, I.; Christoph, B.; Ehrhardt, F.; Parlow, A. Elektronisches Publizieren von Text und chemischen Strukturen am Beispiel des ChemInform. In *Software-Entwicklung in der Chemie 2*; Gasteiger, J., Ed.; Springer: Berlin, 1988; pp 395–398.
(4) Gasteiger, J.; Weiske, C. *ChemInform*—An Integrated Information System on Chemical Reactions. In *Proceedings of the 13th International Online Information Meeting*; Learned Information: Oxford, 1989; pp 147–154.
(5) Weiske, Ch. *Das Klassifikationssystem des ChemInform*, DK-Mitteilungen 21; DIN Deutsches Institut fuer Normung: Berlin, 1977; pp 11–15.