COMMUNICATION, STORAGE AND RETRIEVAL OF CHEMICAL INFORMATION

COMMUNICATION, STORAGE AND RETRIEVAL OF CHEMICAL INFORMATION

ELLIS HORWOOD SERIES IN INFORMATION SCIENCE

COMMUNICATION, STORAGE AND RETRIEVAL OF CHEMICAL INFORMATION
Editors: J. E. ASH, P. A. CHUBB, S. E. WARD, S. M. WELFORD, P. WILLETT

CHEMICAL INFORMATION SYSTEMS
J. ASH, Information Services Consultant and E. HYDE, ICI Pharmaceuticals Division

CHEMICAL NOMENCLATURE USAGE
R. LEES and A. SMITH, Laboratory of the Government Chemist, London

DESIGN, CONSTRUCTION AND REFURBISHMENT OF LABORATORIES
Editors: R. LEES and A. SMITH, Laboratory of the Government Chemist, London

COMMUNICATION, STORAGE AND RETRIEVAL OF CHEMICAL INFORMATION

JANET E. ASH, B.Sc., M.Sc.
Consultant (Information Services)
PAMELA A. CHUBB, B.Sc.
Consultant, Logica UK Limited
SANDRA E. WARD, B.Sc.Hons., Ph.D.
Head of Information Services, Glaxo Group Research Limited
STEPHEN M. WELFORD, B.Sc., M.Sc., Ph.D.
Research Assistant, University of Sheffield
PETER WILLETT, M.A., M.Sc., Ph.D.
Lecturer in Information Studies, University of Sheffield

ELLIS HORWOOD LIMITED
Publishers · Chichester
Halsted Press: a division of
JOHN WILEY & SONS
Chichester · New York · Ontario · Brisbane

First published in 1985 by
ELLIS HORWOOD LIMITED
Market Cross House, Cooper Street, Chichester, West Sussex, PO19 1EB, England

The publisher's colophon is reproduced from James Gillison's drawing of the ancient Market Cross, Chichester.

Distributors:

Australia, New Zealand, South-east Asia:
Jacaranda-Wiley Ltd., Jacaranda Press,
JOHN WILEY & SONS INC.,
G.P.O. Box 859, Brisbane, Queensland 40001, Australia

Canada:
JOHN WILEY & SONS CANADA LIMITED
22 Worcester Road, Rexdale, Ontario, Canada.

Europe, Africa:
JOHN WILEY & SONS LIMITED
Baffins Lane, Chichester, West Sussex, England.

North and South America and the rest of the world:
Halsted Press: a division of
JOHN WILEY & SONS
605 Third Avenue, New York, N.Y. 10016, U.S.A.

British Library Cataloguing in Publication Data
Communication, storage and retrieval of chemical information. –
(Ellis Horwood series in chemical science)
1. Chemistry – Information services 2. Chemistry – Bibliography
I. Ash, Janet E.
540'.7 QD8.3

Library of Congress Card No. 84-25170

ISBN 0-85312-571-6 (Ellis Horwood Limited)
ISBN 0-470-20145-2 (Halsted Press)

Typeset by Ellis Horwood Limited.
Printed in Great Britain by The Camelot Press, Southampton.

Table of Contents

Preface

The Chemical Structure Association (CSA) Seminar 'The Future of Chemical Documentation' at the University of Exeter in September 1982, from which this book has been derived, had two major objectives: (1) to bring together all those who need or work with chemical information to discuss present systems and to predict and plan future developments; (2) to ensure that the discussions involved those who provide access to published information as well as those who are responsible for the development of internal data handling systems in large companies.

The second objective was stimulated by an awareness that over many years, techniques for handling the two types of information had diverged increasingly; within industry, priority had been given to the development of the databank to store a range of physico-chemical and biological data, accesssible by a variety of sophisticated retrieval techniques based on the chemical structure. In the published literature field, developments in computerized information handling had, until the recent introduction of structure-based searching systems such as CAS ONLINE and DARC, centred on the development of text-based searching techniques for accessing bibliographic information. Because these two development paths were beginning to converge it seemed likely that the two groups would benefit from exposure to each other's activities.

Another reason for bringing these groups together at this time was the enormous growth in information technology which has already taken place and which will certainly continue. Information technology provides not only the background against which future information systems will evolve but also the stimulus for much of the evolution. Since it can also offer the opportunity to develop a single interface for the user to access both his own local data collection and the world's published literature, it was both logical and necessary to use the expertise of 'internal' and 'external' information specialists to consider the possibilities.

Any gathering held to discuss chemical information is sure to attract large numbers of information scientists. In addition, the importance of two further

groups to any real discussion was recognized – the database vendor and the chemist. The former needed to be involved because, in general, vendors have not appeared to be particularly susceptible to comments on their existing services and suggestions for improvement, nor do they seem to survey user needs actively. Also, vendors are slow to adopt the results of information research. The chemist is notoriously difficult to attract to discussions on information services yet as the end user, the chemist should be the prime focus of any discussion on chemical information requirements. This is particularly important at a time when new technology is laying the foundations on which major changes can be constructed.

This book is therefore directed towards those who generate and those who use chemical information, towards those involved in its organization and distribution. Covering the three distinct areas of chemical information handling – bibliographic, numeric, and structural – it aims to review fully all current techniques as well as the latest developments in research and relates them to user requirements. The authors have based their chapters in part on the material presented at the conference and on the contributions of those who participated in formal and informal discussion. The conference material has been heavily augmented to ensure that the information presented here gives a more thorough overview of the selected topics. The authors are aware, however, that the book's content is not absolutely comprehensive and that information, while current at the time of going to press, dates extremely quickly in such a fast-moving area.

The book begins by considering the information needs of chemists, looking at their working environment and commenting on the limitations of existing techniques and services. Following this, the book takes an historical perspective beginning with the longer established methods of handling textual information in Chapter 2. This reviews both primary and secondary publication and also includes sections on patents and the major printed reference books. Chapter 3 describes the computerization of bibliographic information pioneered by the growing online industry and includes references to the development of the electronic journal. The growing interest in the availability of databanks which contain directly usable numeric or textual information is reflected in Chapter 4 which describes the major databanks so far developed and problems and factors affecting the development of databanks.

The most important aspects of chemical information handling are obviously those techniques developed for storage and retrieval of chemical structures. They are extensively examined in Chapters 5 to 9 which cover existing methods for the representation and registration of structures in computer systems, the techniques developed for substructure searching, and software for the storage and retrieval of chemical structures, both commercially available packages and those developed for use with particular databases. The various approaches so far taken in attempts to develop generally applicable reaction indexing systems which are of prime interest to the chemist are described. Techniques used to exploit stored chemical structural information are also considered and include

structure–activity methods, molecular modelling and computer-aided chemical synthesis.

Since all of the topics covered in these chapters feature computerized information handling techniques, Chapter 10 examines current developments in computer and telecommunications technology. The general principles of graphics hardware and software are also discussed.

Chapter 11 highlights the important themes identified in previous chapters, indicating the main growth points for future developments and highlighting possible problems. But reviewing the most significant areas of change proves enormously difficult when so many influential developments will come undoubtedly from outside the traditional information world.

Work on individual chapters was undertaken as follows:

J. E. Ash – Chapters 2 and 3
P. A. Chubb – Chapters 4 and 11
S. E. Ward – Preface and Chapter 1
S. M. Welford – Chapters 5, 6, 7 and 9
P. Willett – Chapters 8, 9, 10 and 11

Particular acknowledgement is made to Dr Peter Murray-Rust of Glaxo Group Research Ltd who made a considerable contribution to the sections on computer graphics and Dr Wendy Warr of ICI plc Pharmaceuticals Division on whose presentation Chapter 7 is based. Also to Dr Paul Rhyner of Ciba-Geigy plc, many of whose ideas on the future development of chemical information systems are included in Chapter 1. Mr Ernie Hyde of Fraser Williams (Scientific Systems) Ltd and Dr David Bawden of Pfizer Central Research Ltd are also thanked for their helpful comments. The conference was organized by Jane Whittall (then Gaworska) of Beecham Pharmaceuticals Ltd and her considerable contribution to the conference programme must be acknowledged here.

Finally, we should like to thank all the following contributors to the conference for permission to use material from their lectures in the preparation of this book.

Mr L. S. Adler, Chemical Industries Association
Notification of Chemical Data and Questions of Confidentiality

Dr F. H. Allen, Crystallographic Data Centre, Cambridge
Molecular Structures Rejuvenated – the Role and Utilization of the Cambridge Structural Database

Mrs F. H. Barker, Royal Society of Chemistry, Nottingham
Development of Chemical Databanks

Dr K. P. Barr, The British Library Lending Division
Obtaining Chemical Information

Dr J.-C. Bonnet, Télésystèmes-Questel
Chemical Databases

Professor R. T. Bottle, The City University
Present Methods of Communicating Chemical Knowledge

Dr J. Brandt, Technical University of Munich
A Systematic Classification of Reactions by Electron Shift Patterns

Dr H. D. Brown, Merck, Sharpe & Dohme, USA
Information Needs of Medicinal Chemists in the Pharmaceutical Industry

Dr J. Buckingham, Associated Book Publishers Ltd
Heilbron's Dictionary of Organic Compounds DOC 5

Mr P. T. Bysouth, Glaxo Group Research Ltd, Greenford
Retrieval of Substance Information from the Biochemical Literature

Dr R. Coleman, The Government Chemist
Introductory Speech

Dr C. Cundy, ICI plc, Runcorn
A Personal View of the Generation and Use of Information in Inorganic Chemistry Research

Miss J. Dalton, Beecham Pharmaceuticals Ltd
Compound Registration

Dr J. B. Davis, Health and Safety Executive
The New Substances Notification Scheme – Scientific and Operational Aspects

Mr R. Dean, Excerpta Medica, Amsterdam
DIOL – Drug Information Online

Dr A. Everett, Wellcome Research Labs., Beckenham, Kent
The Information Needs of Physical Chemists

Dr E. Garfield, ISI
Chemical Information Processing at ISI

Dr M. Hann, G. D. Searle
How does a Bench Chemist's Work Benefit from New Substructure Technology?

Mr. A. Haywood, Exxon Office Systems
Trends in Telecommunications, Communications and Hardware

Dr S. R. Heller, Environmental Protection Agency, Washington DC
Experience in the Development of CIS

Dr P. J. Hills, University of Leicester
Communicating at Conferences: The Presentation of Chemical Papers

Dr P. L. Holmes, Blackwell, Technical Services, London
Document Delivery Developments

Dr R. Hyde, Wellcome Research Laboratories, Beckenham
Data Requirements for QSAR Studies

Dr R. A. Y. Jones, University of East Anglia
Naming Organic Transformations

Dr H. Kaindl, Sandoz Ltd, Basle, Switzerland
The Function of the Internal Database – Burial Ground or Intelligence Service?

Dr A. Kolb, IDC, Frankfurt, West Germany
Are Fragmentation Codes Obsolete?

Dr R. Langridge, University of California, USA
The Opportunities for Graphics Techniques in Chemistry

Dr R. Linford, Leicester Polytechnic
Information Needs – the Academic Viewpoint

Dr J. S. Littler, University of Bristol
The Nomenclature of Transformations and Mechanisms – A View from IUPAC

Dr R. Luckenbach, Beilstein Institute, Germany
The Beilstein Handbook – After the First Centennial

Dr A. Mackay, Birbeck College, University of London
Information and Dimensionality – the Communication of Chemical Knowledge

Dr D. S. Magrill, Fisons Pharmaceuticals Ltd, Loughborough
A Registry Number for Today

Dr S. Marson, Molecular Design Ltd
Interactive Graphics Systems for Chemical Research

Dr G. Moreau, Roussel-Uclaf
Substructure Search System

Mr A. Negus, Consultant
Software Trends

Mr P. Nichols, Pergamon InfoLine Ltd
Specialized Chemical Files from InfoLine

Mr P. Norton, Derwent Publications Ltd
Coding and Retrieval of Markush Structures in the Derwent Central Patents Index – Past, Present and Future

Miss M. O'Hare, The British Library, R&D Division
Translating Research Results in Chemical Information into Practice

Dr E. Onerato, ESA/IRS
The Chemical Abstracts Service Bibliographic File on ESA/IRS

Dr D. P. J. Pearson, ICI Plant Protection Ltd
Problems Associated with the Use of Computers as Aids to Chemical Synthesis

Mr J. Revill, Beecham Pharmaceuticals Ltd
European Inventory of Existing Chemical Substances EINECS

Mr J. F. B. Rowland, The Royal Society of Chemistry
Do we need the Scientific Paper?

Miss K. Shenton, SDC Search Service
You and ORBIT: the Chemist's Right

Dr J. Sibley, Shell International Petroleum, London
Patents, the Undervalued Resource

Mr B. Stanford-Smith, National Computing Centre Ltd, Manchester
Technology and the Availability of Information

Dr H. W. D. Stubbs, The Royal Society of Chemistry, Nottingham
The Future of Chemical Documentation – Economic Aspects

Professor R. L. M. Synge, University of East Anglia
Some Experience with the Secondary and Tertiary Services

Dr S. Terrant, American Chemical Society
Online Access to Full Text ACS Primary Journals

Mr S. Vogt, Lockheed Dialog
The Future of Chemical Information on Dialog

Dr W. A. Warr, ICI Pharmaceuticals Ltd
Software

Dr S. M. Welford, University of Sheffield
Towards Simplified Access to Chemical Structure Information in the Patent Literature

Dr P. Willett, University of Sheffield
Some Automatic Approaches to the Indexing of Chemical Reactions

Dr P. W. Williams, Userlink Systems Ltd
Improving Online Access by using Microtechnology

Professor W. T. Wipke, University of California
Reaction Storage and Retrieval

S. E. WARD
Ware
May 1984

1

The information needs of chemists

1.1 INTRODUCTION

Chemistry, since its liberation from the secretive practices of the alchemists, has always been that branch of human endeavour where information is most effectively communicated and best organized for retrieval. This is partly because records of experiments in chemistry have lasting practical value – it is not uncommon to find a chemist following preparative details which were published more than a century ago or confirming the appearance of a substance with a physical description of similar antiquity. Also, chemical substances, by their nature, lend themselves to systematic documentation and indexing, based on concepts of chemical structure which have remained constant over a relatively long period and which are largely independent of language. Most of all, it is because the chemical task has bred a co-operative spirit among chemists which has fostered the exchange of information and encouraged the evolution of formal documentation procedures.

Development of systematic information handling in chemistry is very firmly rooted in the efforts of chemists themselves. The chemist elucidated the principles of chemical structure, and, following this, invented nomenclature systems and other techniques for describing structures. The chemist, largely through the various national professional chemical societies, has been and continues to be responsible for the development and production of most chemical literature. Cooperation between groups of chemists has also led to the development of certain key data collections.

Although the chemist still produces chemical information, in recent years he has become divorced from overall control of its communication and retrieval. The two major factors contributing to this separation are, firstly, the huge increase in the volume of the chemical literature and, secondly, the growth in the use of computers for information handling. Over the past forty years the annual volume of chemical papers has increased roughly five-fold, while novel experimental techniques have greatly increased the information content of these

documents. As a result, at least in industry, much of the chemist's role in information gathering and analysis has been seconded to the information scientist for convenience and time saving. The centralization of searching expertise, initially required for the most efficient exploitation of online databases, has persisted in many organizations. Thus many chemists are still second-class information citizens making do with the traditional methods of searching the chemical literature, and frequently only the librarians and information scientists have the full spectrum of modern techniques at their disposal. The development of improved access methods to online chemical databases is reversing this trend although chemists are only slowly adopting online techniques.

Not only has the chemist become separated from the published literature but, in industry, this separation extends to the chemist's own research results. The development during the 1960s and 1970s of centralized industrial databanks maintained and accessed by information scientists is only now being succeeded by the introduction of online systems which are searchable by chemists as part of their daily routine and to which chemists themselves contribute data directly.

As the rest of this book will show, chemical information techniques are evolving extremely rapidly. The systems and services available to the chemist are already among the most sophisticated of any, yet there are still a large number of outstanding problems and opportunities. It is therefore appropriate to focus on the information needs of the chemist in this opening chapter.

There has been little or no formal documentation of the chemist's requirements for information. A complete survey of the enormous range of chemical activity and the variety of working environments is beyond the scope of this book. Likewise the considerable need of the non-chemist and non-scientist for chemical information cannot be considered. Instead, two types of chemists are selected for review, the teacher, and the research worker, particularly the research and development chemist in the pharmaceutical industry, an industry which has long recognized the importance of effective information provision for research and development. Both groups have exerted substantial influence on the evolution of chemical information services and the role and scientific information needs of both are examined. Recommendations are made for improvements in chemical information handling based on limitations noted in existing systems and services. Requirements for training chemists in information retrieval techniques are also included since the lack of training is recognized as being one of the causes of lack of acceptance of modern information methods by the chemist [1]. For completeness the changing role of the chemical information scientist is also considered since the information scientist will continue to play a crucial part in chemical information for some considerable time.

1.2 THE RANGE OF INFORMATION REQUIRED BY CHEMISTS

The data and text generated by chemists for use by chemists comprises both