Clustering Multidisciplinary Chemical Papers To Provide New Tools for Research Management and Trends. Application to Coal and Organic Matter Oxidation

HENRI DOU,* PARINA HASSANALY, LUC QUONIAM, and JACKY KISTER

Centre de Recherche Rêtrospective de Marseille, UA126 CNRS, Centre de St Jêrome, 13397 Marseille Cêdex 13, France

Received June 10, 1988

Chemistry is unique in that it is one of the largest scientific areas covered by one of the most pertinent databases, *Chemical Abstracts*. This unique database provides the information scientist with more than 20 years of materials available online. This prompts us to study and to develop new methods and automatic tools to picture the multidisciplinary relationships existing between various research areas. This paper shows how these tools can be used in science management. They allow us to follow the trends in these areas over many years and, depending on the type of relationship networks, to class the subjects among various categories describing the practices and behavior of the people concerned with this subject analysis.

INTRODUCTION

The *Chemical Abstracts* sections reflect the subject coverage of chemistry by this publisher. Thirty sections were used in the beginning (1907). With the advent of online services, the placement of each paper in a *Chemical Abstracts* section led in fact to section numbers becoming primary and secondary subject codes assigned to each paper.

For the period 1982 to date, 80 sections (see the list in the Appendix) are used. Thus, the identification of subject areas, or rather the degree of their specificity, depends on the *Chemical Abstracts* approach to the grouping of related subjects, somewhat arbitrarily chosen and specifically designed for that publication. Papers analyzed by a Chemical Abstracts Service indexer are assigned to these sections according to the following rules:[1-3] one primary section, with or without subsections, describes the main subject of the paper; one or several cross-reference sections describe other areas of concern; a paper must always have a primary section, but cross-reference sections are not obligatory.

The following example shows one reference obtained from the *Chemical Abstracts* bibliographic database (host ORBIT INFORMATION TECHNOLOGIES[4]). The fields of the reference are indented to provide their meaning to the reader. In this paper, we will be concerned with the Category Code field (CC) that contains the *Chemical Abstracts* sections. From 1982 to date, these sections did not change, and the indexing practices remain constant.

AN	CA05-85846(10)
TI	vacuum microbalance studies on the combustion of Saraji coal
AU	Adams, K. E.; Glasson, D. R.; Jayawera, S. A. A.
SO	*Thermochim. Acta* (THACAS), V 103 (1), p 157-62, 1986, ISSN 00406031
OS	Plymouth Polytech., Dep. Environ. Sc i., Plymouth/Devon, UK, PL4 8AA
DT	J (Journal)
CC	SEC67-3; SEC51; SEC66

To the author's knowledge, the *Chemical Abstracts* sections have been extensively used for bibliography purposes only.[5,6] They have not been statistically analyzed to provide information on the structure of various research networks. Most of these studies have been performed with cooccurrence of terms or cocitations.

MATERIALS AND METHODS

Materials. The materials that will be used are retrieved from the *Chemical Abstracts* database. The reference list is stored on a microcomputer IBM compatible, XT or AT, HDU 20 Mo. The references are obtained by questioning the database on a particular question. This means that any subject can be analyzed; the only limit is the capability of the host to provide the means (software quality) to select the right information from the *Chemical Abstracts* database.

In our case, we decided to analyze the following subjects: coal and lignite oxidation; and organic matter oxidation, including coal and lignite. The following listing shows the history of the search.

SS 1:	ORGANIC (W) MATTER (4556)
SS 2:	ALL KEROGEN# # # (708)
SS 3:	ASPHALTENES (1203)
SS 4:	ALL COAL# (30610)
SS 5:	ALL LIGNITE# (1650)
SS 6:	ALL TAR# (4995)
SS 7:	ALL HEAVY (W) PETROLEUM (W) PRODUCT# (34)
SS 8:	ALL WOOD# (10453)
SS 9:	ALL POLYMER# (117575)
SS 10:	ALL TURF# (239)
SS 11:	1 OR 2 OR 3 OR 4 OR 5 OR 6 OR 7 OR 8 OR 9 OR 10 (164518)
SS 12:	ALL OX#DAT: (49664)
SS 13:	ALL OX#DIZ: (9023)
SS 14:	ALL OX#DIS: (31)
SS 15:	12 OR 13 OR 1 (56795)
SS 16:	ALL OXIDTN (0)
SS 17:	ALL OXIDN (63287)
SS 18:	15 OR 17 (83962)
SS 19:	18 AND 11 (5267)
SS 20:	19 AND 82-82 (729)
SS 21:	19 AND 83-83 (834)
SS 22:	19 AND 84-84 (848)
SS 23:	19 AND 85-85 (885)
SS 24:	19 AND 86-86 (838)
SS 25:	19 AND 87-87 (722)
SS 26:	19 AND 88-88 (38)
SS 27:	26 OR 25 (760)
SS 28:	(4 OR 5) AND 18 AND 87-88 (189)

To be able to follow the latest developments in these subjects,

0095-2338/89/1629-0045$01.50/0

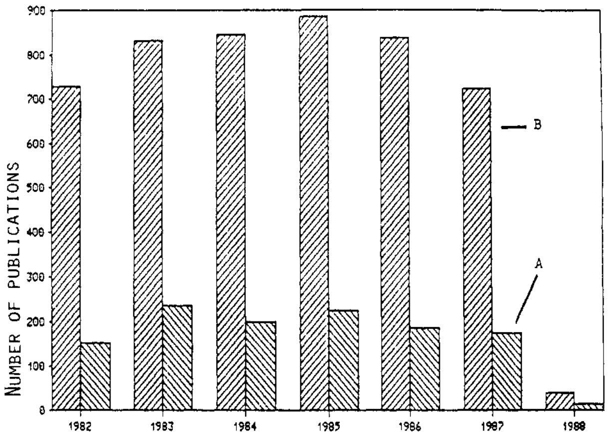

Figure 1. Comparison of the scientific production of coal (A) and organic matter (B) oxidation from 1982 to 1988 (April 21).

we chose to work with years 1987 and 1988. Figure 1 shows the trend in oxidation of organic matter, coal, or lignite. The past two years show fewer references than the others because of the time gap between paper publications and their further appearance in *Chemical Abstracts*.

Methods. The method used for the analysis requires a variety of software that analyze automatically the CC field of each of the two preceding files. The sizes of the two files are 189 references for coal and lignite oxidation and 760 references for organic matter oxidation.

This analysis proceeds in several steps:[7] (i) verification of the downloaded files to see if all the CC fields are really present; (ii) extraction of the CC field to obtain a new file such as

CC SEC51-1; SEC80
CC SEC80
CC SEC22; SEC51
CC SEC61; SEC51
CC SEC22
CC SEC72-4; SEC51; SEC80
CC SEC51-5; SEC79
CC SEC51-2; SEC70; SEC80
CC SEC22-7; SEC51

and (iii) transformation of the above file to get (1) a file containing all the primary sections without their subsections and (2) a file containing all the fields with more than one section and without subsections.

Table I. Pairs and Frequencies

pair	frequency	pair	frequency
51–80	3	51–72	1
51–22	2	51–79	1
51–61	1	72–80	1
51–70	1		

(1) The file containing all the primary sections without their subsections allows one to obtain the graph of the main research poles of the subject. These poles will be represented on a chessboard, where section 1 is at the upper left and section 80 at the lower right. The frequency of a primary section is proportional to the height of the cylinder drawn on the chessboard. Figure 2 shows the background of this representation.

(2) The file containing all the fields with more than one section and without subsections will be used to calculate all the section pairs appearing in each field. For instance

CC SEC72-4; SEC51; SEC80

will lead to pairs 72–51, 72–80, and 51–80. These pairs are representative of potential ties existing between themes 7i and 51, etc.

This process is used for all fields present, and all pairs are cumulated and sorted. For instance, from the above example Table I shows the pairs and frequencies obtained. Once the list of pairs and frequencies is obtained, it is very easy to draw from this table the pluridisciplinary network of themes and its main backbone. Figure 3 shows the network derived from Table I. In this network, the main backbone is 22–51–80, and the subsidiary bonds are 51–61 and 51–70. Poles 51, 80, and 72 are the nodes of the network. We will see that the shape of the network, the node numbers, and the number of clusters according to the frequency threshold used are significant of the research practice of the subject analyzed.

MAIN RESEARCH POLES IN COAL–LIGNITE AND ORGANIC MATTER ANALYZING

Coal and Lignite. Figure 4 represents the main research poles in the field of coal–lignite oxidation. In this figure a unique section (51 Fossil Fuels; Derivatives; and Related

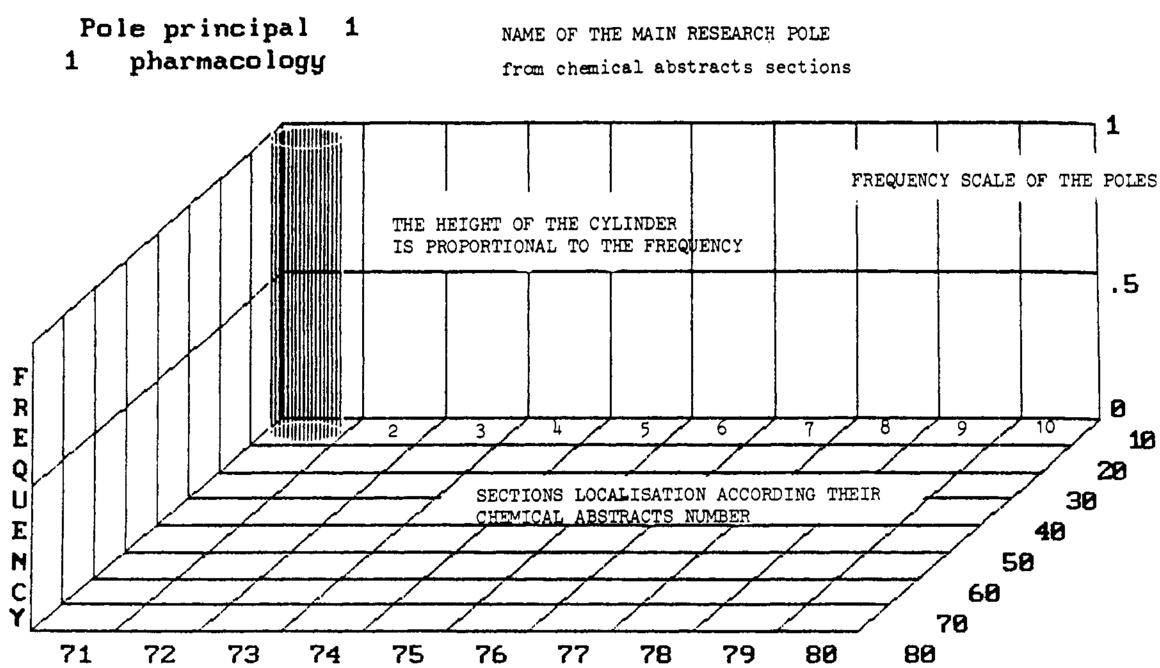

Figure 2. Localization of the main research poles according to their frequency and *Chemical Abstracts* numbers.

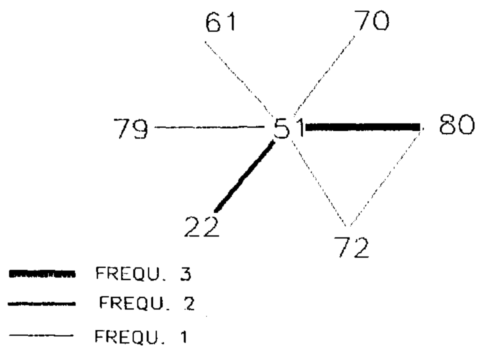

Figure 3. Example of a research network.

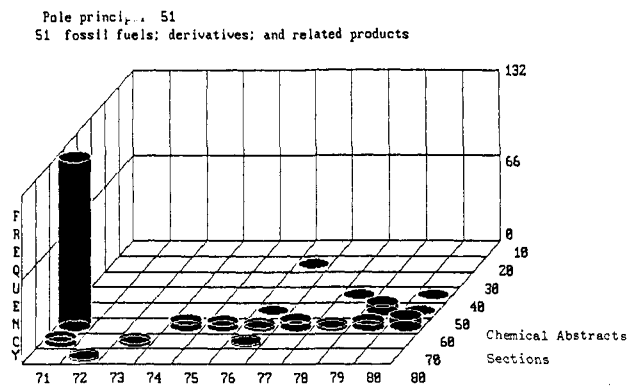

Figure 4. Main research poles of coal oxidation.

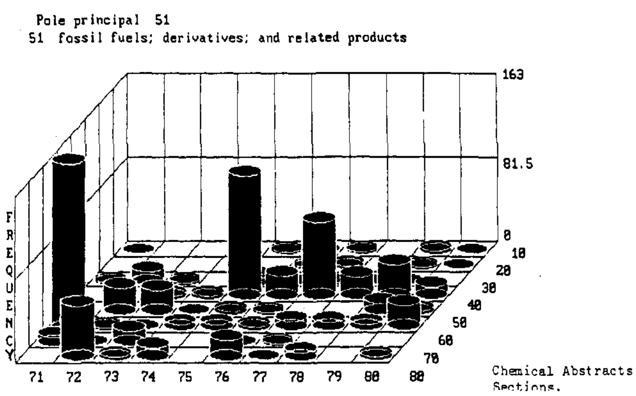

Figure 5. Main research poles of organic matter oxidation.

Products) is overwhelming. This is trivial, since coal and lignite are fossil fuels that are present in section 51. But the lack of other primary sections shows the little concern of coal and lignite in other research areas. This emphasizes the high degree of isolation of this research in the chemical field.

Organic Matter Oxidation, Including Coal and Lignite. Figure 5 represents the main research poles in this field. Because of the query used, section 51 is important, as is section 35 (Chemistry of Synthetic High Polymers). But this also emphasizes that other primary sections are largely concerned with this research. This shows that the subject analyzed is less isolated than coal and lignite oxidation among the chemistry subjects described by the *Chemical Abstracts* sections.

MULTIDISCIPLINARY NETWORK

Oxidation of Coal and Lignite. This network is shown in Figure 6. It shows almost no reticulation and is a perfect representation of a star-type network. This indicates research almost exclusively focused on coal and lignite, with no participation or concern with other areas of chemistry.

The fact that this network has no reticulation reinforces the results obtained when the main research poles were examined. Specialists in the field can surely find other applications of

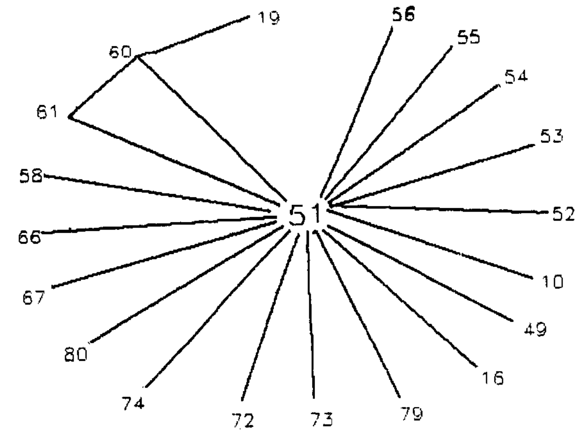

Figure 6. Coal oxidation.

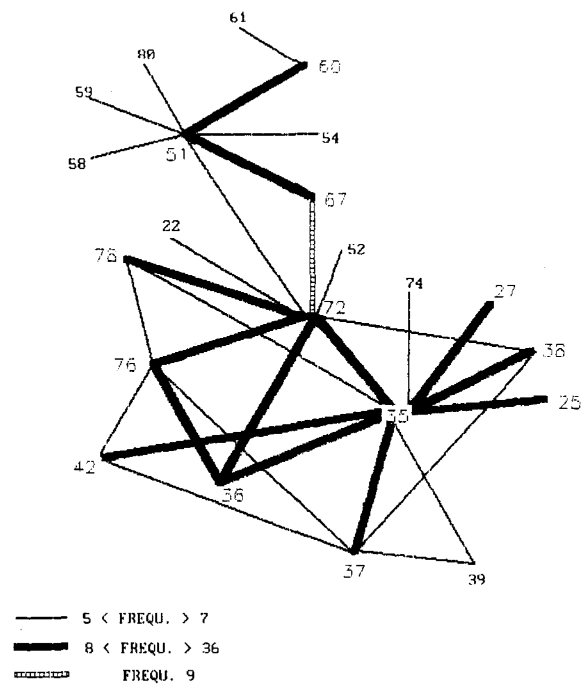

Figure 7. Organic matter oxidation.

the shape of the network. This can be explained in two ways: (1) researchers' habits, i.e., they do not use the techniques and know-how of other disciplines; (2) the impossibility of applying coal or lignite oxidation techniques to other areas.

Oxidation of Organic Matter Including Coal. This network has been drawn in Figure 7. Its pattern is very different from the preceding one. Two main parts must be considered. The first one is focused on section 51 and exhibits only two bonds at low frequency with the second part of the network. The second part is well reticulated, with various connections between different areas of research. This is a good example of a different network shape. This shape indicates a good flux of exchanges between various chemical fields, which surely implies innovative transfers and good relationships between different research groups.

These two patterns are often encountered in scientific papers analysis in pure or mixed forms. They can be used to compare different sets of references, since they give a pattern unique for each set. This "fingerprint" is a different approach to the ways in which the fluxes of production and information are exchanged in a scientific or technical field.

OTHER EXAMPLES OF NETWORKS

To indicate the possibilities of this method, we will briefly show three more applications: the study of French oceanog-

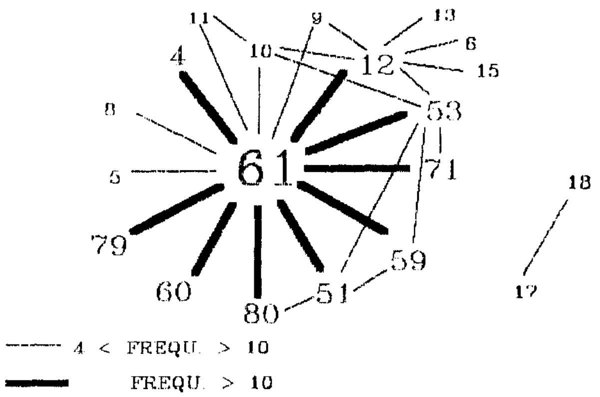

Figure 8. French chemical oceanography from 1982 to April 1988.

raphy from 1982 to date, the study of the research pattern in the universities and research centers located in Marseille (France) for 1986 using *Chemical Abstracts*, and the study of physics in universities and research centers in Marseille using the database INSPEC for 1985.

French Chemical Oceanography. The starting materials have been selected by using various concepts large enough to retrieve from the *Chemical Abstracts* database a set of references that will be a good pattern of French oceanography. For the period extending from 1982 to date, 989 references have been selected.

Figure 8 represents the network of this area of research, using the same technique as above. The network is for part of it (the higher frequencies) of the star type. Reticulation begins to occur at a lower frequency threshold. This reticulation concerns "biochemical oceanography". Discussion with specialists in the field confirms the difficulty of this theme's emergence in French chemical oceanography official bodies.

Chemistry at Marseille in 1986. The same technique has been used to draw the network and to determine the main research poles. The references have been selected by using a limitation by town and date. The network, at high and low frequencies, exhibits various clusters that are not related to each other. The first one is related to life sciences chemistry. The second deals with thermodynamic chemistry and the third with inorganic and surface chemistry. It is easy to see that when the frequency threshold decreases, new clusters and bonds appear. This reveals new clusters such as marine chemistry and organometallic chemistry. All the results are indicated in Figure 9.

It is interesting to note that this network condenses 556 references and indicates at a glance the structure of the chemical research at Marseille and the relative proportion of the scientific production.

In this study, we can see that various directions of research are examined all together and the different clusters emerge. These clusters, depending on the frequencies at which they bond (or do not bond), are a good indication of the fluxes of exchanges in the geographic area analyzed.

Physics at Marseille in 1985. The INSPEC classification (see the Appendix) was used to analyze the structure of the network of physics at Marseille in 1985. The file size was 356 papers. From Figure 10, it is easy to see that three clusters emerge: optic and optoelectronic, theoretical physics, and astronomy. This example shows that the method is general and can be applied to all databases that provide in their references meaningful codes, such as WPI, WPIL (Derwent codes), BIOSIS (Biocodes), Management, and Predicast.

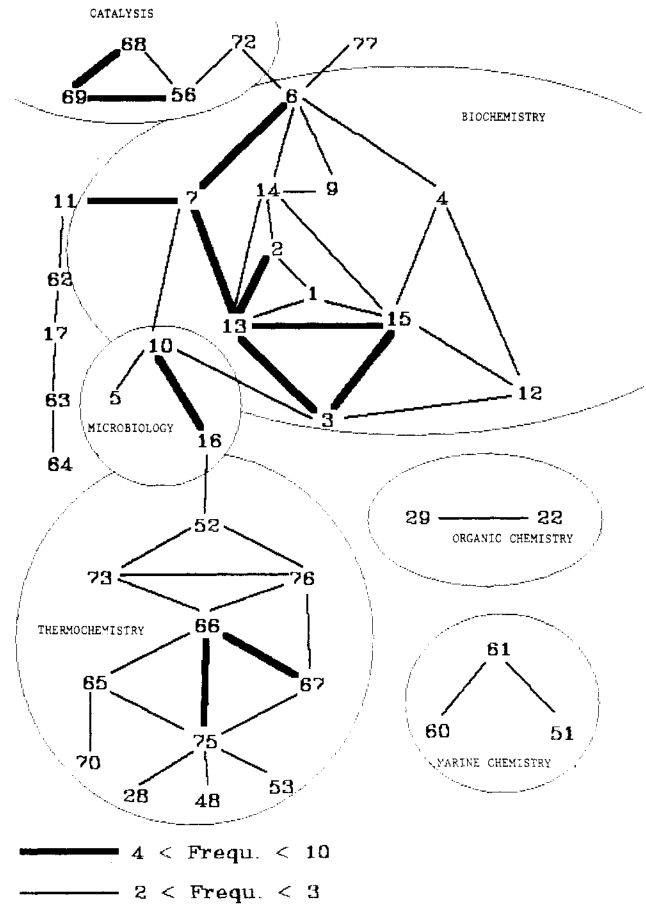

Figure 9. Chemistry in Marseille (source: *Chemical Abstracts* 1986).

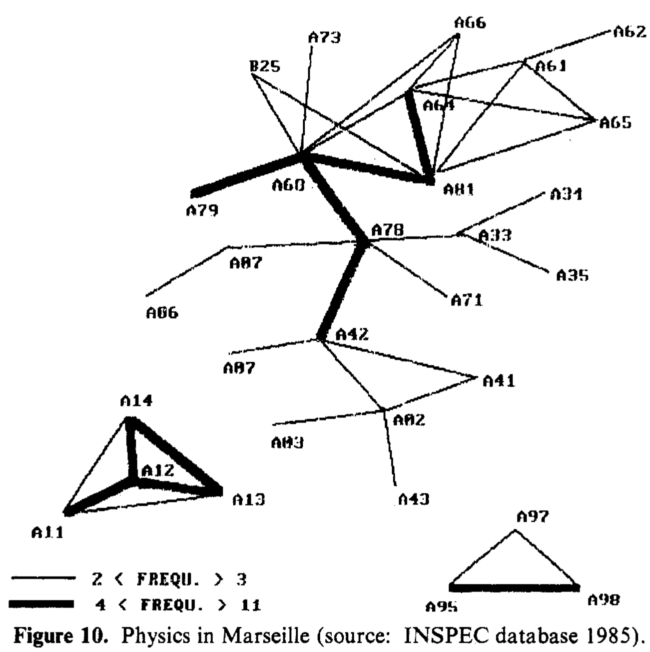

Figure 10. Physics in Marseille (source: INSPEC database 1985).

CONCLUSION

The automatic analysis of *Chemical Abstracts* sections, using downloaded data, is a strong and powerful tool to picture the research activities of a field. Since all *Chemical Abstracts* references give sections, all the subjects leading to references in this database can be worked out with this method. Since the network shapes can be very different (star, reticulated, with various clusters, mixed) and since the shape also depends on the frequency threshold used, indicators can be derived from such a study.

The comparison of networks obtained from year to year, using the same query, is a good way to picture research trends and to evaluate the degree of multidisciplinarity and exchanges.

The comparison with the network obtained when the oxidation of organic matter was analyzed suggests that oxidation of coal and lignite stands alone, as a subject by itself, different from the techniques and research developed during the study of the oxidation of other chemical compounds.

All these possibilities suggest that these networks may be used in research policy and management and also by individuals who wish to examine the place of their work among the scientific production dealing with their research subject.

Other clustering techniques such as citation counts[8] and co-word analysis[9,10] have been developed during recent years. These methods, which also lead to scientific networks, are generally more difficult to use and necessitate larger computers.

We believe that because of their simplicity the use of networks grounded on scientific codes is a valuable tool that quickly obtains key information from various databases.[11]

ACKNOWLEDGMENT

We thank the DBMIST and ORBIT INFORMATION TECHNOLOGIES for their help. This work has been made under the French PARUSI program to increase research in information sciences in France.

APPENDIX

List of *Chemical Abstracts* Sections

1 Pharmacology
2 Mammalian Hormones
3 Biochemical Genetics
4 Toxicology
5 Agrochemical Bioregulators
6 General Biochemistry
7 Enzymes
8 Radiation Biochemistry
9 Biochemical Methods
10 Microbial Biochemistry
11 Plant Biochemistry
12 Nonmammalian Biochemistry
13 Mammalian Biochemistry
14 Mammalian Pathological Biochemistry
15 Immunochemistry
16 Fermentation and Bioindustrial Chemistry
17 Food and Feed Chemistry
18 Animal Nutrition
19 Fertilizers, Soils, and Plant Nutrition
20 History, Education, and Documentation
21 General Organic Chemistry
22 Physical Organic Chemistry
23 Aliphatic Compounds
24 Alicyclic Compounds
25 Benzene, Its Derivatives, and Condensed Benzenoid Compounds
26 Biomolecules and Their Synthetic Analogs
27 Heterocyclic compounds (One Hetero Atom)
28 Heterocyclic Compounds (More Than One Hetero Atom)
29 Organometallic and Organometalloidal Compounds
30 Terpenes and Terpenoids
31 Alkaloids
32 Steroids
33 Carbohydrates
34 Amino Acids, Peptides, and Proteins
35 Chemistry of Synthetic High Polymers
36 Physical Properties of Synthetic High Polymers
37 Plastics Manufacture and Processing
38 Plastics Fabrication and Uses
39 Synthetic Elastomers and Natural Rubber
40 Textiles
41 Dyes, Organic Pigments, Fluorescent Brighteners, and Photographic Sensitizers
42 Coating, Inks, and Related Products
43 Cellulose, Lignin, Paper, and Other Wood Products
44 Industrial Carbohydrates
45 Industrial Organic Chemicals, Leather, Fats, and Waxes
46 Surface-Active Agents and Detergents
47 Apparatus and Plant Equipment
48 Unit Operations and Processes
49 Industrial Inorganic Chemicals
50 Propellants and Explosives
51 Fossil Fuels, Derivatives, and Related Products
52 Electrochemical, Radiational, and Thermal Energy Technology
53 Mineralogical and Geological Chemistry
54 Extractive Metallurgy
55 Ferrous Metals and Alloys
56 Nonferrous Metals and Alloys
57 Ceramics
58 Cement, Concrete, and Related Building Materials
59 Air Pollution and Industrial Hygiene
60 Waste Treatment and Disposal
61 Water
62 Essential Oils and Cosmetics
63 Pharmaceuticals
64 Pharmaceutical Analysis
65 General Physical Chemistry
66 Surface Chemistry and Colloids
67 Catalysis, Reaction Kinetics, and Inorganic Reaction Mechanisms
68 Phase Equilibrium, Chemical Equilibriums, and Solutions
69 Thermodynamics, Thermochemistry, and Thermal Properties
70 Nuclear Phenomena
71 Nuclear Technology
72 Electrochemistry
73 Optical, Electron, and Mass Spectroscopy and Other Related Properties
74 Radiation Chemistry, Photochemistry, and Photographic and Other Reprographic Processes
75 Crystallography and Liquid Crystals
76 Electric Phenomena
77 Magnetic Phenomena
78 Inorganic Chemicals and Reactions
79 Inorganic Analytical Chemistry
80 Organic Analytical Chemistry

INSPEC Classification

01 Physics, General A00
02 Communication, education, history and philosophy A01
03 Mathematical methods in physics A02
04 Classical quantum physics; mechanics and fields A03
05 Relativity and gravitation A04
06 Statistical physics and thermodynamics A05
07 Measurement science, general laboratory techniques, and instrumentation systems A06
08 Specific instrumentation and techniques of general use in physics A07
09 Physics of Elementary particles and fields A10
10 General theory of fields and particles A11
11 Specific theories and interaction models; particle systematics A12
12 Specific reactions and phenomenology A13
13 Properties of specific particles and resonances A14

14 Nuclear Physics A20
15 Nuclear structure A21
16 Radioactivity and electromagnetic transitions A23
17 Nuclear reactions and scattering: general A24
18 Nuclear reactions and scattering: specific reactions A25
19 Properties of specific nuclei listed by mass ranges A27
20 Nuclear engineering and nuclear power studies A28
21 Experimental methods and instrumentation for elementary particle and nuclear physics A29
22 Atomic and Molecular Physics A30
23 Theory of atoms and molecules A31
24 Atomic spectra and interactions with photons A32
25 Molecular spectra and interactions with photons A33
26 Atomic and molecular collision processes and interactions A34
27 Properties of atoms and molecules; instruments and techniques A35
28 Studies of special atoms and molecules A36
29 Classical Areas of Phenomenology A40
30 Electricity and magnetism; fields and charged particles A41
31 Optics A42
32 Acoustics A43
33 Heat flow, thermal and thermodynamic processes A44
34 Mechanics, elasticity, rheology A46
35 Fluid dynamics A47
36 Fluids, Plasmas and Electric Discharges A50
37 Kinetic and transport theory of fluids; physical properties of gases A51
38 The physics of plasmas and electricity discharges A52
39 Condensed Matter: structure, thermal and mechanical properties A60
40 Structure of liquids and solids; crystallography A61
41 Mechanical and acoustic properties of condensed matter A62
42 Lattice dynamics and crystal statistics A63
43 Equations of state, phase equilibria, and phase transition A64
44 Thermal properties of condensed matter A65
45 Transport properties of condensed matter (nonelectronic) A66
46 Quantum fluids and solids; liquid and solid helium A67
47 Surfaces and interfaces; thin films and whiskers A68
48 Condensed Matter: Electronic Structure, Electrical, Magnetic and Optical Properties A70
49 Electron states A71
50 Electronic transport in condensed matter A72
51 Electronic structure and electrical properties of surfaces, interfaces, and thin films A73
52 Superconductivity A74
53 Magnetic properties and materials A75
54 Magnetic resonances and relaxation in condensed matter; Mössbauer effect A76
55 Dielectric properties and materials A77
56 Optical properties, condensed matter spectroscopy and other interactions of matter with particles and radiation A78
57 Electron and ion emission by liquids and solids; impact phenomena A79
58 Cross-Disciplinary Physics and Related Areas of Science and technology A80
59 Materials science A81
60 Physical chemistry A82
61 Energy research and environmental science A86
62 Biophysics, medical physics, and biomedical engineering A87
63 Geophysics, Astronomy and Astrophysics A90
64 Solid Earth geophysics A91
65 Hydrospheric and atmospheric geophysics A92
66 Geophysical observations, instrumentation, and techniques A93
67 Aeronomy and space physics A94
68 Fundamental astronomy and astrophysics, instrumentation and techniques and astronomical observations A95
69 Solar system A96
70 Stars A97
71 Stellar systems; galactic and extragalactic objects and systems; The Universes A98
72 Electrical. General topics, Engineering Mathematics and Materials Science B00
73 General electrical engineering topics B01
74 Engineering mathematics and mathematical techniques B02
75 Materials science for electrical and electronic engineering B05
76 Electrical. Circuit theory and Circuits B10
77 Circuit theory B11
78 Electronic circuits B12
79 Microwave technology B13
80 Electrical, Components, Electron Devices and Materials B20
81 Passive circuit components, cables, switches and connectors B21
82 Printed circuits, thin film, thick film and hybrid integrated circuits B22
83 Electron tubes B23
84 Semiconductor materials and devices B25
85 Dielectric materials and devices B28
86 Electrical. Magnetic and Superconducting Materials and devices B30
87 Electrical. Magnetic materials and devices B31
88 Electrical. Superconducting materials and devices B32
89 Optical Materials and Applications, Electro-optics and Optoelectronics B4
90 Optical materials and devices B41
91 Optoelectric materials and devices B42
92 Lasers and masers B43
93 Electromagnetic Fields B50
94 Electric magnetic fields B51
95 Electromagnetic waves, antennas and propagation B52
96 Communications B60
97 Information and communication theory B61
98 Telecommunication B62
99 Radar and radionavigation B63
100 Radio, television and audio A64
101 Electricals, Instrumentations and Special Applications B70
102 Measurement science B71
103 Measurement equipment and instrumentation systems B72
104 Measurement of specific variables B73
105 Elementary particle and nuclear instrumentation B74
106 Medical Physics and biomedical engineering B75

107 Aerospace facilities and techniques B76
108 Earth sciences B77
109 Sonics and ultrasonics B78
110 Electrical. Power Systems and Applications B80
111 Electrical. Power networks and systems B81
112 Electrical. Generating stations and plants B82
113 Power apparatus and electric machines B83
114 Direct energy conversion and energy storage B84
115 Electrical. Power utilisation B85
116 Electrical. Industrial application of power B86
117 Computer and Control. General and Management Topics C00
118 General control topics C01
119 General computer topics C02
120 Management topics C03
121 Computer and Control. Systems and Control Theory C10
122 Systems and control theory. Mathematical techniques C11
123 Systems theory and cybernetics C12
124 Control theory C13
125 Computer and Control. Control Technology C30
126 Control and measurement of specific variables C31
127 Control equipment and instrumentation C32
128 Control application C33
129 Numerical Analysis and Theoretical Computer Topics C40
130 Numerical analysis C41
131 Computer metatheory and switching theory C42
132 Computer Hardware C50
133 Computer Hardware. Circuits and devices C51
134 Computer Hardware. Logic design and digital techniques C52
135 Computer Hardware. Storage devices and techniques C53
136 Analogue and digital computers and systems C54
137 Computer peripheral equipment C55
138 Computer Software C60
139 Software techniques and systems C61
140 Computer Applications C70
141 Computer applications. Administrative data processing C71
142 Computer applications. Information science and documentation C72
143 Computer applications. Natural sciences C73
144 Computer applications. Engineering C74
145 Other computer applications C75

BIBLIOGRAPHY

(1) Dickman, J. T.; O'Hara, M. P.; Ramsay, O. B. *Chemical Abstracts. An Introduction to Its Effective Use*; ACS Audio Course No. 52; American Chemical Society: Washington, DC.
(2) Subject coverage and arrangement of abstracts by sections in Chemical Abstracts. Editor American Chemical Society, 1982.
(3) Shinichiro, K. Changes in sections of Chemical Abstracts. *Shikoku Kokenkaiho* **1982**, *33*, 2–6.
(4) Orbit Information Technologies is a Pergamon Infoline Co., which offers online most of the largest scientific databases and patent databases.
(5) Inge Berg, H. Subject compatibility between Chemical Abstracts Subject sections and search profiles used for computerized information retrieval. *J. Chem. Doc.* **1972**, *12*, 110–113.
(6) Peterson, J. S. Replacement of an in-house current awareness bulletin by Chemical Abstracts Section Groupings. *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 169–172.
(7) Dou, H.; Hassanaly, P. Mapping the Scientific network of patent and non-patent documents from chemical abstracts for a fast scientometric analysis. *World Pat. Inf.* **1988**, *2*.
(8) Ganz, C. Bibliometric models for international Science and Technology. *Rev. Fr. Bibl.* **1987**, *2*, 2.
(9) Callon, M. C.; Courtial, J.; Turner, W.; Bauin, S. Problematic networks: an introduction to C-word analysis. *Inf. Sci. Soc.* **1983**, 191.
(10) Turner, W. A.; Chartron, B.; Laville, F.; Michelet, B. Packaging information for peer review: new co-word analysis techniques. *Scientometrics* (submitted for publication).
(11) Jakobiak, F. Maitriser l'information critique. *Les editions d'organisation*; 1988; ISBN 2708108743.

Representation and Matching of Chemical Structures by a Prolog Program

JOSEPH L. ARMSTRONG[‡] and D. BRYNN HIBBERT*

Department of Analytical Chemistry, University of New South Wales, P.O. Box 1, Kensington, NSW 2033, Australia

Received September 8, 1988

A notation for describing chemical structures based on the connectivity of entities in the structure is presented in a Prolog program. Following simple rules the chemical structure is encoded as a Prolog data structure that is called the symbolic name (s-name) of the structure. An algorithm to decide if two different encodings of the same structure represent the same chemical can be used as the basis of a decision procedure to determine if an unknown chemical already exists in a large data base of s-name encoded chemical structures. A transformation on the s-name, the generic or g-name, is described that makes searching in a large data base of unknown chemicals very efficient. The s-name notation has the property that the level of abstraction used in the description may be changed, with common substructures named and descriptions nested to any level. The 35 isomers of C_9H_{20} are used as an example of this method.

INTRODUCTION

The traditional method of finding out about an unknown structure is to transform the structure into a unique name and look up this name in a chemical data base such as *Chemical Abstracts*. The fact that a structure can be converted into a unique name (which is an alphabetic string) means that the abstracts can be ordered with respect to this name. The method breaks down if the name is not unique (i.e., we assume that applying the standard naming rules to a given structure results in a unique name—expressed another way we assume that a given name has exactly one derivation using the rules of naming). There are no correctness proofs for the naming rules. Conventional methods of naming chemicals are complex and error prone. As structures become large and more complex, the corresponding systematic names become more and more difficult to understand. This problem led Silk[1] in 1981

* Author to whom correspondence should be addressed.
[‡] Present address: Computer Science Laboratory, Ericsson Telecom, Stockholm, Sweden S-126 25.

0095-2338/89/1629-0051$01.50/0