INTERNATIONAL COUNCIL OF SCIENTIFIC UNIONS
ISSN 0538-6918

COMMITTEE ON DATA FOR SCIENCE AND TECHNOLOGY

60 CODATA NEWSLETTER

SEPTEMBER 1992

J.-E. Dubois Wins Herman Skolnik Award

Jacques-Emile Dubois, the recipient of the 1992 Herman Skolnik Award, has a rich and distinguished scientific career as well as a preeminent CODATA involvement. The Award was established by the American Chemical Society's (ACS) Division of Chemical Information to recognize outstanding contributions to and achievements in the theory and practice of chemical information science and is named in honor of the first recipient, Herman Skolnik. It was presented on August 25, 1992, at the ACS National Meeting.

Professor Dubois obtained his Ph.D. from the University of Grenoble for his work in infra-red spectroscopy on hydrogen bonding and the regioselectivity of aldolization and has taught at the University of the Saar and, since 1957, at the University of Paris in physical chemistry and molecular spectroscopy.

Jacques-Emile Dubois

Professor Dubois has made significant contributions in many fields of research: kinetics (such as the transfer of rapid protons in prototrophy); the bromination mechanism; the control of aldolization stereoselectivity ("the first significant breakthrough"); the chemistry of crowded structures (ketones); and the first electropolymerized films.

As early as 1960, he developed a keen interest in chemical informatics. He invented the DARC Topological System which led to various applications in

(continued on p. 5)

HIGHLIGHTS

The **Committee on Data for Science and Technology (CODATA)** was established in 1966 by the International Council of Scientific Unions.

Working on an interdisciplinary basis, CODATA seeks to improve the quality, reliability, processing, management, and accessibility of data of importance to science and technology.

Knowledge Discovery in Databases

Many efforts have been invested lately in building new databases and in improving access and man/machine interfaces. However, the amount of available information is increasing so rapidly in some scientific areas (for instance, in astronomy where the earth observation satellites are expected to produce around a terabyte of data a day or in biology with the Human Genome Project), that it is now necessary to develop procedures to automatically generate information of a "higher level," *i.e.*, knowledge, automatically.

The CODATA Task Group on Artificial Intelligence and Computer Graphics (AIGRA) has set as one of its objectives the assessment of new methodologies and products in A.I. in order to transfer them, if possible, to the field of Scientific Complex Data Management (SCDM).[1] In this perspective, we are looking now into the new field of "knowledge discovery in databases" and present below some basic ideas related to this topic.

A recent book[2] published at the end of 1991—a collection of the best papers presented at a specialized workshop—is a good basic reference.

On a very general level, knowledge discovery can be defined as the non-trivial extraction of implicit but unknown information from data. The discovered knowledge is characterized by two aspects (a) its pattern and description in a concise, useful way and (b) its degree of certainty.

Patterns describe simple or complex relationships between subsets of data; for instance, relationships between values in fields of the same record or inter-record patterns. Initial information is seldom complete and consistent, therefore, computed patterns are only approximations. The degree of certainty reflects user confidence towards the newly generated knowledge.

Standard statistical methods can give a first approach for examining the database and the relationships between variables, but these methods are not well suited for structured data types. Their main drawback is that they ignore the domain knowledge.

The general topic of knowledge discovery is not new and has already been addressed through different approaches in machine learning, scientific discovery, and knowledge acquisition in expert systems. However, databases raise new issues due to the fact that they are dynamic (the content is changed by updating and a discovery method should incrementally modify what has already been learned) and that they can be incomplete and noisy. Moreover, a typical real world database is much larger than any set used for machine learning. Previous algorithms and techniques must, therefore, be adapted or modified.

As a typical example, we can name the very popular ID3 induction algorithm of Quinlan which generates a decision tree for classifying the examples given in terms of a set of attributes. This algorithm and its improved versions (ID4, ID5) have been successfully applied to generate expert systems from data, but it is still not quite adequate for databases. First, its implicit assumption is that enough information is available to decide exactly how each data point should be classified. Second, ID3 works only on simple vectors of values and cannot handle complex structured data. Two of the articles of the reference given below present new improvements to this algorithm to solve these problems in the framework of databases.

One of the most popular patterns of knowledge is represented by *if-then* rules which relate two sets. Generally three types of rules are considered: exact, strong, and weak. An exact rule is always correct. In databases these usually represent domain dependency and are normally known to the database user. The strong and weak rules can be represented as probabilistic rules or rules with plausibility factors.

In a relational database, according to Y. Cai *et al.*,[2] we consider that two types of rules can be derived: a characteristic rule and a classification rule. In the former case, the rule is an assertion that characterizes the concept satisfied by all the data in the database. In the latter case, the rule is an assertion that discriminates the concepts of one class from those of others. Another way to look at rules associated with relational databases is given by Piatetsky-Shapiro:[2] a database rule is defined as an implication (C1 $\rightarrow$ C2) where C1 and C2 are conditions stated on the database universal relation (the universal relation is a logical file which combines all fields of interest). To each A $\rightarrow$ B rule is associated an interest-function which is computed as a function of p(A), p(B), p(A&B) (p being the probability of A, respectively, of B and A&B), rule complexity and other parameters such as the mutual distribution of A and B or the domain sizes of A and B. Sometimes user-specified *interest weights* for different fields can also be introduced. The rule discovery task is then to find N rules with the highest rule-interest function.

Scientists are well aware of techniques used in scientific discovery, but scientific data and data in databases are quite different. Scientists generally work with limited data of good quality and not with sparse and noisy data. However, some previous systems have been adapted to discover weak regularities (as opposed to deterministic scientific regularity, typically expressed as an equation) in databases.

Instead of looking at pattern forms, research areas can be analyzed in terms of specific objectives such as: the discovery of numeric or qualitative laws, discovery using domain knowledge, intra-record *vs.* inter-record relationships, etc.

A last problem, analysed in the article of D. E. O'Leary[2] is concerned with database security. Knowledge discovery techniques can lead to *unauthorized knowledge*, that is knowledge which the owners of the database consider to be proprietary or secret. Therefore, in some environments it would be necessary to introduce specific controls.

In conclusion, one can quote J. R. Quinlan's *Foreword*: "the 1990s should see the widespread exploitation of knowledge discovery as an aid to assembling knowledge bases."

For the AIGRA group, studies on *knowledge discovery in databases* are a necessary preliminary step in solving the more general problem of "Data and knowledge based systems for decision making in science and technology."

The AIGRA group intends to organize a workshop on this topic in 1993, and any comments and proposals will be welcomed.

(continued on p. 3)

ICSU Statement on Gene Patenting

The International Council of Scientific Unions (ICSU) is an international non-governmental organization whose mandate includes the promotion of cooperation in the basic sciences and the safeguarding of the principle of the universality of science and the free flow of scientific knowledge.

The Council is aware of the tremendous potential benefit of genetic research for humanity and realizes that new ethical and social dimensions arise from this. Accordingly, ICSU strongly believes that efforts to patent genetic information should not jeopardize either progress in the basic sciences or access to the information which is necessary for such progress to continue.

ICSU asserts its view that information about nucleic acid sequences cannot be patented *per se*. Such sequences should be patentable solely within the context of their demonstrated significance and/or application (*e.g.*, regulatory signals, antisense RNAs, probes, etc....)—and not of their **potential** products (*e.g.*, proteins)—and provided that this can be shown to be "novel," "non-obvious," and "useful."

Under such circumstances, patenting of complementary DNA sequences (cDNAs) would distort the patent process, which is designed to protect applications, methods, and products, on the basis of proven facts and not mere expectations, and normally serves society by stimulating the investments and developments necessary to provide useful products and services. Any deviation from such patenting principles would run counter to the best interests of science and hinder international collaboration in such endeavours. ICSU therefore cautions against decisions which may be irreversible, such as those possibly emerging as a result of the recent patent requests concerning complementary DNA (cDNA) sequences corresponding to portions of unknown messenger RNAs (mRNA).

ICSU urges the relevant authorities, particularly in countries where patent applications in this field have been or are soon to be filed, to consider such applications taking due account of the possible implications and to ensure a strict application of established patenting principles, thereby setting an example for other countries in which similar cases may arise in the future.

ICSU would welcome a formal international agreement on this subject.

--Paris, June 1992

Knowledge Discovery (continued)

1. New Perspectives in Scientific Complex Data Management, H. Bestougeff and J. E. Dubois (Editors), *CODATA Bulletin*, **Vol. 22**, No. 4, 1990.

2. Knowledge Discovery in Databases, Gregory Piatetsky-Shapiro and William J. Frawley (Editors), AAAI Press/ The MIT Press, 1991.

--Jacques-Emile Dubois
--Helen H. Bestougeff

CODATA Calendar

1992

October

17-18 Asian-Oceanic Data Sources Task Group. Beijing, China

19-22 International CODATA Conference. Beijing, China

23-24 CODATA General Assembly. Beijing, China

December

2-4 CODATA Task Group on Geothermodynamic Tables. Paris, France

Advanced Materials Workshop Held in Italy

The international workshop "Regularities, Classification, and Prediction of Advanced Materials" was held on April 13-15, 1992, at Como, Italy. An impressive array of speakers from all over the world was assembled by the workshop's organizer, John Rodgers, Research and Development Coordinator for CISTI's Scientific Numeric Database Service, CAN/SND.

The need for new materials with special properties, tailored to specific applications in key technological areas, is becoming increasingly important. The development of a systematic approach complete with improved techniques for materials classification is long overdue. The workshop focussed on an interdisciplinary approach to materials prediction. Disciplines involved ranged from quantum mechanics, solid state physics, crystallography, materials science, and computer science. The workshop covered theories and techniques pertaining to materials regularities and prediction, pattern recognition and materials classification schemes.

Co-sponsors were the U.S. Office of Naval Research European Office, IBM-Italy, the European Institute of Technology, and CODATA, among others. The proceedings will be published in the *Journal of Alloys and Compounds*. The workshop's success led to a biennial scheduling.

Thought.....

"The major source of missed items in information retrieval comes from the inability of electronic retrieval systems to establish connections between two different formulations of the same concept and between two different descriptions of the same idea."

Alain Bonnet in the "Journal of the Association for Global Strategic Information," March 1992 (ISSN 0965-4380)

Task Group on Materials Database Management

Materials Database Newsletter **August 1992, Number 16**

DATABASES

Magnesium Elektron Ltd, in association with Norsk Hydro, has produced a database diskette containing extensive property information on a range of commercial and specialty **magnesium alloys**. The project has been sponsored through the UK Department of Trade and Industry's Materials Matters programme and uses the MATUS database from Engineering Information Company. Wrought and cast magnesium alloys are covered. FURTHER INFORMATION: Magnesium Elektron Ltd, Regal House, London Road, Twickenham, TW1 3QA, England.

The Cambridge Materials Selector is a materials selection system based on a methodology that can be used to optimise the performance of engineering components. It contains a database on the properties of almost all the main types of engineering materials, and a sophisticated data management system. A demonstration version is available. FURTHER INFORMATION: Mrs S J McKay, Lynxvale Limited, 20 Trumpington Street, Cambridge CB2 1AQ, England. Tel: +44 223 334755; Fax: +44 223 332797.

STN International has published information sheets for the factual database **COPPERDATA** (Copper Development Association Standards and Data) which is available through the Materials Property Data Network (MPDN). The database contains up to date numeric data for 166 wrought and 95 cast US coppers and copper alloys, including the commercial bronzes, brasses, copper nickels, nickel silvers and high-copper alloys. Data is available on compositions, specifications, properties, applications, fabrication and suppliers. FURTHER INFORMATION: MPDN Inc., 2540 Olentangy River Road, P. O. Box 02224, Columbus, OH 43202, USA. Tel: +1 614 447 3706.

STN International has also provided data sheets for **CRYSTMET** (NRCC Crystallographic Data File), a substance-based database containing critically evaluated data for metals and intermetallic phases. The compounds are identified by name and molecular formula. Structural data such as crystal symmetry and crystal lattice parameters, as well as density and temperature factors, are given and the data generation program "DIST" for calculating and displaying interatomic distances and angles is included. FURTHER INFORMATION: FIZ Karlsruhe, P. O. Box 2465, W-7500 Karlsruhe 1, Germany.

RECENT PUBLICATIONS

The second edition of UK **Materials Information Sources**, compiled by Keith W Reynard, is now available. Published by the Design Council, London, this comprehensive directory of information sources and advice on all types of engineering materials available in the UK has been greatly enlarged, with increased coverage of advanced materials. The book includes an extended A-Z listing of over 2,300 industrial, commercial and educational sources of information. The publication is available for £25.00 plus £3.00 p & p from: The Design Council Bookshop, the Design Council, 28 Haymarket, London, England, SW1Y 4SU.

Also available from the Materials Information Service at the Design Council is a new quarterly publication MIS Bulletin which is designed to help develop and maintain a knowledge of UK engineering materials information sources. The bulletin is available on subscription for £60.00 a year. Details from the Design Council at the address above.

FORTHCOMING EVENTS

The CODATA-CEC Joint **Workshop on Materials Data for Computer Aided Engineering**, originally planned for the winter of 1991, will now be held 1-3 February 1993 in Frankfurt am Main, Germany. The principal objectives of the workshop are to define the actions needed for a more effective integration of materials information into computer aided engineering systems and to identify those who will be capable of contributing to these actions.

The fourth International Symposium on **Computerization and Use of Materials Property Data** will be held in Gaithersburg, MD, USA, 6-8 October 1993. The symposium will focus on a number of topics, including: end user participation in systems development; applications of advanced computing technologies; user interface design; integration of computerised data into design and manufacturing; data analysis for quality and reliability; economic impact of systems in use; and legal liability issues.

CALENDAR

1-3 February 1993: Frankfurt am Main, GERMANY

CODATA-CEC Joint Workshop on **MATERIAL DATA FOR COMPUTER AIDED ENGINEERING**. CONTACT: DECHEMA, Abt. I&D, Theodor-Heuss-Allee 25, D-6000 Frankfurt/Main 15, GERMANY.

6-8 October 1993: Gaithersburg, Maryland, USA

Fourth International Symposium on **COMPUTERIZATION AND USE OF MATERIALS PROPERTY DATA**. CONTACT: ASTM, 1916 Race Street, Philadelphia, PA 19103-1187, USA.

Future meetings of **ASTM COMMITTEE E49** have been scheduled as follows:

16-18 November 1992:	Miami, FL, USA
May 1993:	Atlanta, GA, USA
November 1993:	Fort Worth, TX, USA

EDITOR: W. G. Jackson. EDITORIAL OFFICE: The Institute of Metals, 1 Carlton House Terrace, London, UK, SW1Y 5DB. Tel: +44 1 839 4071; Telex: 8814813; Telefax: +44 1 839 2289. There are no restrictions on the reproduction and distribution of the contents of this Newsletter.

(continued from p. 1

artificial intelligence. This work evolved towards conceptual imagery and the use of computer graphics as a practical and theoretical tool for chemists. His DARC software is used worldwide on the CAS and DERWENT large data banks. He has received two graphics awards: the FISA d'Or 1986 and the Palaiseau Scientific Prize.

Founder of the Institute of Topology and Systems Dynamics (ITODYS) at the University of Paris (VII), Professor Dubois was Chairman of the IUPAC Committee on Machine Documentation and of the CODATA Computer Task Group. He is currently Chairman of the CODATA Task Group on Artificial Intelligence and Graphics (AIGRA).

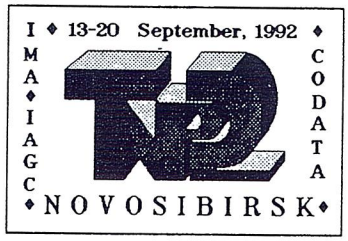
Conceptual creativity and originality characterize the scientific work of Professor Dubois. In his fruitful approach to chemical informatics, he stressed the notion of ordered spherical environments in molecular sites, which is essential for handling compounds and controlling their behavior. This concept breaks with the traditional view of function, structure, and family. The sequence of substructure (SS), structure (S) and hyperstructure (HS) locate chemical entities in their structural context and evaluate their local or global properties according to topological and topographical information. As a structural unitary theory, the DARC System offers a new logic and new languages to chemistry, enabling it to make maximal use of computer-age resources.

These original concepts endow chemistry with immense possibilities. Professor Dubois has done pioneering work in both the theory and the development of avant-garde expert systems, especially in spectroscopic elucidation systems and computer-aided drug design with topological assessment of similarities. His recent work deals with the statistical assessment of knowledge in large chemical files and with molecular modelling of carbohydrates and strained systems.

By this Award, the Division hopes to encourage the continuing advancement of chemical information science in areas such as design of new and unique computerized information systems, preparation and dissemination of chemical information, design of new indexing, classification, and notation systems, chemical nomenclature and structure-activity relationships, and numerical data correlation and evaluation.

--Bettijoyce Lide

Thermodynamics of Natural Processes



This conference sponsored by the International Association of Geochemistry and Cosmochemistry, CODATA, Unesco, and the International Mineralogical Association was achieved by organizing committees under the direction of a committee headed by Professors N. Sobolev, G. Kolonin, G. Kuskov, and Dr. G. Shverdenkov. Sessions were held in Akademogorodok (Novosibirsk) with many participants from Europe and abroad. Approximately 150 papers and posters were presented and visits were made to relevant sites.

Geographic Information Systems

The Fifth International Conference on Geographic Information systems will be held in Ottawa from March 21-25, 1993. This annual event is organized by the Surveys, Mapping and Remote Sensing Sector (SMRSS), EMR Canada, in cooperation with the Canadian Institute of Surveying and Mapping and the Inter-Agency Committee on Geomatics.

The Conference Director is Mr. L. Aubrey who can be reached at the following address: Canadian Conference on GIS, c/o SMRSS, EMR Canada, 615 Booth Street, Ottawa, Ontario, K1A 0E9. Tel.: 1 (613) 995-0266; FAX: 1 (613) 995-6001.

Directory of Numeric Databases

A *Directory of Numeric Databases*, covering all fields of natural sciences, has been issued by the International Council for Scientific and Technical Information (ICSTI). Requests for complementary copies of the page document should be addressed to Marthe Orfus, ICSTI Executive Secretary, 51 Boulevard de Montmorency, 75016 Paris, France. ICSTI would welcome any contributions that would make the next issue of the directory more complete.

Donald Wagman Dies

Donald Wagman died from coronary problems while playing tennis on September 2, 1992. He has had a long and distinguished CODATA career, particularly in Key Values for Chemical Thermodynamics and Chemical Thermodynamic Tables Task Groups and at the National Bureau of Standards.

Editor : Edgar F. Westrum, Jr.
Department of Chemistry, University of Michigan,
Ann Arbor, MI 48109
Telephone: (313) 764-7357 / Telex: 8102236056
FAX: 1-313-747-4865

Published four times per year (February, May, August, November)

Associate Editor: Phyllis Glaeser
CODATA Secretariat, 51 Blvd. de Montmorency,
75016 Paris, France
Telephone: 33 1 45250496 / Telex: 645554F
FAX: +33 1 42881466 / Cables: ICSU Paris 016

Assistance in dissemination provided by National Committees

INTERNATIONAL COUNCIL OF SCIENTIFIC UNIONS
COMMITTEE ON DATA FOR SCIENCE AND TECHNOLOGY

CODATA / NEWSLETTER

CODATA, 51 bd. de Montmorency, 75016 Paris