Pure & Appl. Chem., Vol. 49, pp.1793 -1796. Pergamon Press. 1977. Printed in Great Britain.

CAPTURE, EVALUATION AND STORAGE OF DATA
- as seen by CODATA

N. Kurti

Department of Engineering Science, University of Oxford.

Abstract - The paper contains a statement of the aims of the Committee on Data for Science and Technology (CODATA) of the International Council of Scientific Unions (ICSU) and a survey of its activities which range widely from providing critically assessed values of fundamental constants and of certain key values for key substances to the investigation of the methodology for handling space and time dependent data. The establishment - in the near future - of a World Data Referral Centre (WDRC) under CODATA auspices and work on a Directory of Data Sources for Science and Technology are briefly described.

It is obvious that to deal with the whole subject "Capture, Evaluation and Storage of Data" one would need several lectures and not just a 20 minute talk - hence the qualifying sub-title. This will be a very specific treatment, an attempt to explain what CODATA (the Committee on Data for Science and Technology of ICSU, the International Council of Scientific Unions) has done, and is trying to do, in this field. My lecture will contain no hard or particularly useful facts; its rationale could be best characterised by the well known story about the two young Irishmen who one Sunday morning felt very contrite after a boisterous Saturday evening out. So they went to Church, and first one went in and confessed to having committed the mortal sin of fornication. But he would not divulge the name of the person he did it with, even though the priest tried to make it easier for him by mentioning the names of a few young ladies of the town who might have been his willing partners. So, no full confession, no absolution. He went out rather crest-fallen and his pal asked him, "Well, what happened, did you get absolution?" "No, but he gave me three good tips." Similarly I do not expect to get absolution for any past misdeeds or lack of deeds of CODATA but I hope that in the ensuing discussion you will be able to give me and through me CODATA a few good tips on how we can improve our ways and how we can be more helpful to the scientific community in general and to the chemists in particular.

As the chairman mentioned, CODATA was founded at a time when the amount of scientific data being collected was increasing rapidly, and when it was therefore felt that a co-ordination might be necessary. First in the field was the United States, where the National Bureau of Standards was asked to start a National Standards Reference Data System. A large amount of work was also being carried out in the U.S.S.R. and when it was finally decided to set up an international co-ordinating body the United Kingdom entered wholeheartedly into this scheme and so Codata was formed, with France, the Federal Republic of Germany, Japan, UK, USA and USSR as the six "founding" members. CODATA has at present 15 "national" members and they provide most of CODATA's funds. The rest of the membership consists of 12 scientific unions of ICSU and the ICSU Panel on World Data Centres.

What are the aims of CODATA? There is one thing it does not want to do and cannot do, namely to edit and publish large data collections; that is entirely beyond its means. CODATA's raison d'être is expressed in its constitution which says "CODATA working on an inter-disciplinary basis, shall seek to improve the quality, the reliability and accessibility of data of importance to science and technology including not only quantatitive information on the properties and behaviour of matter but also other experimental and observational data". It took incidentally, the better part of 4 hours to draft this passage, especially the last 19 words. The reason was as follows. CODATA was originally set up to improve the quality of data used mainly in chemistry and in physics, that is of data obtained by physical measurements on pure or readily identifiable or readily reproducible substances. That was a fairly straight forward remit and there was no difficulty about expressing it in the constitution. However, in about 1972, 6 years after CODATA's birth, the International Council of Scientific Unions asked CODATA that it should concern itself not only with data important for physics and chemistry, data on pure or readily identifiable substances, but also with data of importance to the earth sciences and to the life sciences. This widening of its scope has somewhat altered the character of CODATA. Although some people regret this departure from CODATA's original philosophy, I believe that this is a good development because CODATA is supposed to work on an inter-disciplinary basis, and by concerning itself

also with earth and life sciences, it can hope to establish in the data field close co-operation between disciplines which deal with "immutable" data and disciplines concerned with data which depend on time and location. So I think from that point of view this was a good move.

What are the specific goals of CODATA? Well, the basic aim is - I quote again from its constitution, "to promote the evaluation, and in general the quality control of data and the methods by which they are acquired". Another aim is to increase awareness amongst all scientists of the importance of data activities and in particular to encourage scientists to participate in them, and connected with it, to promote the improvement of the status and the training of data evaluators and of data compilers. And finally, CODATA wishes to encourage the application of new methods to data handling, storage, and retrieval, as well as to the evaluation, the presentation, and the organised production and dissemination of data. How can CODATA, a relatively small and far from wealthy organisation perform these functions? It does so mainly through small specialist task groups consisting of anything between 4 and maybe 10 scientists all specialists in the particular field the task group is concerned with. These task groups have well defined terms of reference, are established for limited periods at a time and when their task is completed, they are dissolved or put into abeyance. I think that the activities of CODATA can be best explained by listing its Task Groups and describing briefly what each of them has been doing, is doing, or proposing to do in the future.

Data usually begin in the laboratory but become public knowledge when they appear in the primary literature. Alas, all too often the way data are presented or the experimental methods described, make subsequent evaluation of the results, or a repetition of the experiment very difficult. Therefore a task group was set up to study quite generally the representation of data in the primary literature, mainly in physics and chemistry. Their report was published as a Unesco-UNISIST Guide and in the CODATA Bulletin (1). In addition, it appeared in several scientific journals and magazines in 4 or 5 languages. To supplement this general Guide the CODATA Task Group on Data for Chemical Kinetics was asked to prepare a more specific guide covering its field of interest. Their report was also published in the CODATA Bulletin (2).

Two Task Groups are preparing Guides on the PRESENTATION OF DATA IN THE BIOLOGICAL SCIENCES and PRESENTATION OF DATA IN THE EARTH SCIENCES and it is hoped that their reports will be published during 1977.

Finally, CODATA Bulletin has also published a Report of the ICSU Inter-union Commission on Biothermodynamics on biochemical equilibrium data (3).

The rapid increase in the volume of scientific literature makes the "capture" of data more and more difficult - there are too many places in which to look for them. The task would be rendered easier if an internationally agreed system of "flagging" and "tagging" were generally adopted by both primary and secondary publication. A data "flag" indicates that the article in question contains numerical data and may also identify its general nature, e.g. spectroscopic data or results of geological observations, while a data "tag" is usually more specific. CODATA, jointly with the ICSU Abstracting Board, established a WORKING GROUP ON TAGGING AND FLAGGING which published its report in June 1976 in the CODATA Bulletin (4). I am very worried that after so many years we still do not seem to be nearer a generally accepted scheme and can only hope that the recommendations of the CODATA-ICSU/AB Working Group will be acted upon.

In the logical sequence of events we come next to the critical evaluation of the data, their testing for internal consistency and for compatibility with other results. Believing that these skills - as well as a proficiency on treating experimental data - are not as common as would seem desirable, CODATA has set up a Task Group to organize appropriate training courses on an international scale. As a result the first CODATA/UNISIST Training Course in the Handling of Experimental Data was held in Varazdin, Yugoslavia, during August 1976. Two one-week courses, run consecutively, were attended each by about 20 students from a dozen countries. They were intensive courses consisting of lectures, seminars and practical work, given or conducted by experts from several countries. A detailed description of the course and the lessons learned from it will be published in the CODATA Bulletin and a similar course to be held in Poland is being planned for 1977.

As to the results of critical evaluations, CODATA, while not producing great compilations, does try to establish critically evaluated lists for selected categories of substances and/or properties. Five Task Groups have been set up for this purpose:

Key Values for Thermodynamics
Critically evaluated data for the enthalpies of formation, the absolute entropies (at 298.15K) and the incremental enthalpies between $0^{o}K$ and 298.15K have been established for about 100 substances. The results appeared in six CODATA Bulletins (5).

Fundamental Physical Constants
This Task Group reviews continuously the values of about 50 fundamental constants and has published a consistent set of values in CODATA Bulletin (6).

Transport Properties
This Task Group is concerned chiefly with the thermal and electrical conductivities of metallic and dielectric solids, and in particular with the establishment of standard measuring techniques and of standard values for a few well-defined test substances. The Task Group has recently widened its activities to include data of interest to geologists and geophysicists.

Data for Chemical Kinetics
The first report of this Task Group reviewed the data activities in this field (7), while its second report dealt with the presentation of data (2). The Task Group has recently been reactivated with more restricted terms of reference and it will concern itself mainly with rate constants and photo-chemical yields for processes relevant to stratospheric pollution. It is true that a number of national and international organizations are already active in this field but it was felt that since this is a politically sensitive area an international, but non-governmental, body like CODATA could be of some use.

Finally, a Task Group for Physical Property Data for the Chemical Industry has been set up recently which will, in the first instance, try to establish rules for the estimation of properties which as yet have not been measured - e.g. the evaluation of the properties of mixtures of fluids from the known data for the constituents. Unlike in most other CODATA endeavours the emphasis here will not be on high precision and accuracy but rather on a reliability sufficient for the design of chemical plant.

Since computers are used widely in all data activities one of the first CODATA Task Groups was on Computer Use. It published two editions of a review of computer use in data centres (8) and more recently it ran a very successful symposium on the use of computers in scientific data handling (9).

Two very recent Task Groups will address themselves to more general questions, namely the Internationalization and Standardization of Thermodynamic Data and the Methodology of Handling Space and Time Dependent Data, the latter being of particular interest to geologists and geographers.

But it is not enough to have good, critically evaluated data. It is important that they are widely disseminated and readily accessible. The CODATA Task Group on the Accessibility and Dissemination of Data has been grappling with this question for over 4 years with the generous support of UNESCO-UNISIST. After the publication of its report (10) it turned its attention to the more specific question of Data Referral Centres. Many of the developed countries have such centres - national or local - to which people can turn for information about where data in a certain field may be obtainable. The Task Group studied the feasibility of a World Data Referral Centre which would link the national and local centres and recommended that such a centre be set up. Preparations to establish such a World Data Referral Centre (WDRC) in Paris, in the same building where the CODATA Secretariat (together with ICSU and a few other ICSU bodies) are housed, is in hand and it is hoped that the WDRC can be in operation sometime in 1978.

One of the first tasks of CODATA was the publication in 1969 of an International Compendium of Numerical Data Projects (11). While this is still a fairly useful book, some of the information is out of date and - more important - a very large number of projects are missing from it, either because they were started since publication of the Compendium or because they were concerned with the Life and Earth sciences which in the early days were outside CODATA's scope. It has therefore been decided to publish an entirely new Directory of Data Sources for Science and Technology. This will be published initially in parts - each covering one more or less wide area - as separate issues of the CODATA Bulletin, but the sections will eventually be consolidated into a single volume. The part on crystallographic data is due to be published in spring or early summer 1977. Sections on biological sciences and on astronomy - astrophysics will follow soon after.

This brings me to the end of my review of CODATA's role in the capture, evaluation, storage and dissemination of data, and this is the right moment to pay tribute and give thanks to those scores of scientists who give their time and services freely in helping this good cause. Although for many who take part in these activities a considerable financial sacrifice is involved, CODATA's work still costs money - our annual budget is about $110,000 - and for most of this we rely on the contributions from our National Members and from ICSU, and on contracts - mainly from UNESCO-UNISIST.

In order to expand our work we shall need more funds and this is where industrial firms, national or international data organizations, etc. could help. By becoming "Associate Organizations" they would become involved in various CODATA activities, would have access to unpublished reports and memoranda, and would help CODATA financially. So, to echo the young Irishman's remark, "I'm waiting for a few good tips!"

REFERENCES

1. Guide for the Presentation in the Primary Literature of Numerical Data Derived from experiments. CODATA Bulletin No.9, Dec.1973

2. The Presentation of Chemical Kinetics Data in the Primary Literature. CODATA Bulletin No.13, Dec.1974

3. Recommendations for Measurement and Presentation of Biochemical Equilibrium Data. CODATA Bulletin No.20, Sept.1976

4. Flagging and Tagging Data - to indicate its presence and facilitate its retrieval. CODATA Bulletin No.19, June 1976

5. Key Values for Thermodynamics:

   Tentative Set - Part I. CODATA Bulletin No.2, Nov.1970
   Final Set - Part I. CODATA Bulletin No.5, Dec.1971
   Tentative Set - Part II. CODATA Bulletin No.6, Dec.1971
   Tentative Set - Part III. CODATA Bulletin No.7, Aug.1972
   Recommended Values. CODATA Bulletin No.10, Dec.1973
   Recommended Values. CODATA Bulletin No.17, Jan.1976

6. Recommended Consistent Values of the Fundamental Physical Constants, 1973. CODATA Bulletin No.11, Dec.1973

7. A Catalogue of Compilation and Data Evaluation Activities in Chemical Kinetics, Photochemistry and Radiation Chemistry. CODATA Bulletin No.3, Dec.1971

8. Automated Information Handling in Data Centres. 1st Edition, CODATA Bulletin No.1, Oct.1969; 2nd Edition, CODATA Bulletin No.4, Nov.1971

9. Man-Machine Communication in Scientific Data Handling. CODATA Bulletin No.15, March 1975

10. Study on the Problems of Accessibility and Dissemination of Data for Science and Technology. CODATA Bulletin No.16, Oct.1975

11. CODATA International Compendium of Numerical Data Projects. Springer-Verlag, Berlin-Heidelberg, 1969

All the above publications can be obtained from the CODATA Secretariat (51 boulevard de Montmorency, F-75016 Paris, France; Tel.525-0496, Telex 63553F), to which all enquiries concerning CODATA should be addressed.