UNISIST

Study on the problems of accessibility and dissemination of data for science and technology

SC.74/WS/16

United Nations Educational, Scientific and Cultural Organization

UNITED NATIONS EDUCATIONAL,
SCIENTIFIC AND CULTURAL ORGANIZATION

STUDY ON THE PROBLEMS OF ACCESSIBILITY AND DISSEMINATION OF DATA FOR SCIENCE AND TECHNOLOGY

prepared under UNESCO contract by
Accessibility and Dissemination of
Data (ADD) Task Group of the
Committee on Data for Science and
Technology (CODATA) of ICSU

P R E F A C E

Numeric and quantitative data may be regarded as a "crystallized" presentation of the essence of scientific knowledge in the most accurate form and play as such an ever increasing rôle in science and technology.

UNISIST Recommendation No.10 reads : "The collection, critical evaluation, organization and dissemination of numerical data is functionally closely related to the processing of published literature and must be provided for in any future network of information services. Special attention should be paid to the development of networking capability among numerical data centres and to the functional interrelationship of such centres with the bibliographically orientated network."

Bearing in mind that reliable data is essentially useless if it cannot reach the user in good time and proper form in response to his request, the UNISIST Secretariat charged the Committee on Data for Science and Technology (CODATA) of ICSU to carry out a study on the problems of accessibility and dissemination of data for science and technology.

The goal of the study was to give a realistic picture of the situation in the field of the evaluation, compilation and dissemination of reliable data, to analyze existing deficiencies and gaps and to develop recommendations for future action, paying special attention to the needs of developing countries.

The report defines three functions needed to provide accessibility to scientific and technical data :

- data evaluation and compilation service
- data dissemination service
- data referral service

In accordance with this subdivision of functions a global scheme or network for dissemination of S & T data is proposed including :

- data evaluation centres
- data dissemination centres
- data referral centre

This does not, however, necessarily mean the establishment of three new organizations for each scientific or technical discipline or subject, but calls for the use and expansion of all present facilities.

Within this scheme,

1) Data evaluation centres are differentiated by disciplines or problems and only one (or a limited number) of them is active in a given scientific area. This means that they can be the sources for the supply of data of high quality on a world-wide basis. It is therefore suggested that no country, however big it is, will possess data evaluation centres in all disciplines.

2) Data dissemination centres have the responsibility for handling a broad range of scientific and technical disciplines and use the products of data evaluation centres as input.

The output from data evaluation centres may be often used directly by the users, but, in addition, such a centre should supply data to a wider range of users through data dissemination centres.

Every country should have, in principle, a data dissemination centre, so that a domestic user can communicate with it easily, but it may, however, be appropriate to create a regional data dissemination centre. In contrast to data evaluation centres, data dissemination centres are not restricted in specialization, but cover a wide area.

A data dissemination centre may offer the following services :

(a) to collect published compilations of evaluated data and to offer them for reference use;

(b) to search particular data on requests from collections of data centres and other data dissemination services;

(c) to provide current awareness services;

(d) to assist users in identifying where to find required data.

3) The Global Referral Centre directs users' enquiries for data to the sources capable of supplying the needed material.

The basic problem, leading to the necessity of the creation of the Global Referral Centre, is that reliable data compilations, prepared at substantial cost, are in many cases not accessible to users promptly enough to permit their application where needed. In some instances this lack of accessibility is due to the user's ignorance about the existence of the needed compilation; in others, he may be aware of its existence, but does not know how to get it.

Apart from the function of guiding users to the location of data, the Global Referral Centre supplies appropriate information to the local data dissemination centres, so that the latter can provide referral services for the disciplines in their charge.

Recommendations 4 - 10 of the Report deal with the problems mentioned above.

Other recommendations relate to :

(1,2) Governmental support for data compilation, evaluation and dissemination projects.

(3) Manpower - necessity of specialized training for professional "data specialists".

(14) Standardization of data presentation (tabular and graphical formats for uniform presentation and transmission of data). In connection with this recommendation it would be interesting to note that "a guide for the presentation in the primary literature of numerical data derived from experiences" has been already developed by CODATA in cooperation with UNESCO and will be soon largely distributed as a UNISIST document.

(15,16) Standardization of data exchange media and of computerized bibliographic services.

(17) Data descriptive records suitable for primary journals and abstracting services. The main concern is to provide :

(a) abbreviated methods for identifying in the body of the abstract, data contained in the paper;

(b) flags for abstracts of papers, containing data, by using special keywords that modify the abstracts and are applicable for computer search;

with the aim of identifying every primary publication that contains data and notifying physical, biological or other quantities involved in data.

(18-29) Assistance to developing countries.

The spectrum of the recommendations related to developing countries is quite large - starting with the proposal to publish a general booklet on the rôle of data in modern science and technology and finishing with appropriate financial and technical assistance to developing countries to encourage their efforts towards establishing effective mechanisms of data dissemination in their countries.

The study on the problems of accessibility and dissemination of data for science and technology was presented at the first session of the UNISIST Advisory Committee in February 1974 and, in general, met with the approval of the Committee. The Committee's recommendations in relation to the study may be summarized as follows :

1. High priority to recommendation 10 - the establishment of a Global Referral Service.

2. Awarding high priority to recommendation 22 - publication of a general booklet on the rôle of data.

3. Supporting strongly the series of recommendations 18-29 as providing support to developing countries, but with low priority accorded to recommendation 29 (computer linkage via satellite).

While noting that the study did not necessarily reflect its point of view, the Committee supported the proposal of the UNISIST Secretariat for a large distribution of the study having in view the importance of stimulating governments and scientific organizations to endorse the programmes for evaluation and dissemination of data.

MEMBERSHIP OF CODATA/ADD TASK GROUP

Chairman: Professor M.Kotani, Science University, Kagurazaka 1-3, Shinjuku-ku, Tokyo, JAPAN.

Members: Mme A.David, Institut Français des Combustibles et de l'Energie, 3 rue Henri Heine, 75016 Paris, FRANCE.

Professor I.Eliezer, Department of Chemical Physics, Weizmann Institute, Rehovoth, ISRAEL.

Dr.H.W.Koch, American Institute of Physics, 335 East 45th Street, New York, NY 10017, USA

Professor C.N.Rao, Department of Chemistry, Indian Instute of Technology, Kanpur 16, INDIA.

Dr.S.A.Rossmassler, Office of Standard Reference Data, National Bureau of Standards, Washington D.C.20234, USA. (Deputy:Dr.D.R.Lide, same address).

Dr.D.G.Watson, Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge, England CB2 1EW, UK.

Secretary: Professor M.Kizawa, Faculty of Engineering Science, Osaka University, Toyonaka, Osaka, JAPAN.

Ex-officio: Dr.C.Schäfer, CODATA Central Office, Westendstrasse 19, 6 Frankfurt/Main, FEDERAL REPUBLIC OF GERMANY.

Observer: Dr.A.Wysocki, UNISIST Office, UNESCO, Place de Fontenoy, Paris 7, FRANCE. (Deputy:Mr.V.Rybatchenkov, same address).

NOTE ON THE PREPARATION AND PRESENTATION OF THE REPORT

This Report has been prepared by the CODATA Task Group on Accessibility and Dissemination of Data, CODATA/ADD, established in September, 1972.

Since most of the work had to be done by correspondence preliminary drafts were prepared in Japan and distributed to members for their critical comments. For this purpose a Japanese Working Group was formed consisting of the chairman, Professor M.Kizawa (secretary), Professor T.Shimanouchi, Dr.Y.Mashiko and Mrs.E.Ushijima. Drafts were written by the chairman in collaboration with Professor Kizawa and Mrs.Ushijima, Professor Shimanouchi and Dr.Mashiko acting as consultants. In March 1973 questionnaires drafted by Professor T.Shimozawa and Mrs.Ushijima, and authorized by the Task Group, were sent to ICSU adhering organizations and national CODATA committees of some 55 countries to collect information to be used in the Report. Unfortunately the response to the questionnaire was very low.

The first meeting of the Task Group took place at UNESCO House in Paris on July 17-18, 1973. At this meeting Mr.V.Rybatchenkov deputized for Dr.A.Wysocki. The overall content of the Report was discussed but the main concern was with the details of Chapter I (Recommendations).

The second meeting was held at the Swedish Royal Academy of Sciences in Stockholm on September 8-9, 1973. In the chairman's absence, this meeting was conducted by Professor Kizawa. Dr.D.R.Lide of NBS deputized for Dr.Rossmassler and the Task Group is most appreciative of the substantial contribution made by Dr.Lide. At this second meeting the content of the Report was essentially agreed and the remaining work of an editorial character was entrusted to an editing committee, consisting of the chairman and Dr.Watson. The committee met in London on November 3-4,1973 and decisions were made on some reorganization of chapters, sections and recommendations. Production of the Report in its present

form, including the 'polishing' of the English language, has been carried out mainly at Dr.Watson's office.

The Task Group further acknowledges with thanks the contributions and suggestions made by Professor J.E.Dubois and Dr.van der Heide.

Due to lack of time it has been impossible to examine many of the problems in as great depth as we would have liked. In particular, various aspects of earth and life sciences have been only partly, and rather fragmentally, incorporated. The Task Group hopes that these points will be remedied and improved with the cooperation of a wider group of scientists. It is hoped that, in spite of these deficiencies, the present Report will be useful to many people who are interested in the problems of dissemination of data and that it will provide a substantial basis for future studies.

Since the Report is rather large the Task Group feels that it will help readers if it is presented as two documents:-

(a) Main Report (Chapters A-H)

(b) Recommendations (Chapter I).

Footnotes to the recommendations are provided, where applicable, referring the reader to the appropriate sections of the Main Report. It should also be noted that, although recommendations 18-29 are specifically directed to the needs of developing countries, nevertheless the other recommendations 1-17 are also relevant to these countries. These latter recommendations, of course, are of interest also to developed nations.

The reader should note that, in many places throughout the Report, we use the word 'science' when, in fact, we also mean 'engineering', 'medicine' etc. This abridgement has been used simply to make sentences less cumbersome.

EXECUTIVE SUMMARY

Both science and technology gain much effectiveness because they can utilize, in each new undertaking, the results of past efforts. Such useful and important results are to be found, in concentrated form, in compilations of reliable data.

In developing its plans for UNISIST, UNESCO recognized that valuable progress could be stimulated in both developed and developing countries by assuring that compilations of reliable data would be readily accessible to all scientists and engineers. Accordingly, UNESCO engaged the CODATA Task Group on Accessibility and Dissemination of Data (CODATA/ADD) to provide advice on the availability of reliable data and recommend improvements where needed.

CODATA/ADD began its study in September 1972. Earlier investigations had already established that scientists and engineers find reliable data essential in all aspects of their work; major private, national and international programs are actively seeking to increase the supply of reliable data compilations in recognition of this fact. CODATA/ADD concluded, however, that current data evaluation programs are not capable of filling all of the needs of users. Further, the existing resources of evaluated data are less useful than they should be because a variety of barriers (unawareness, language and vocabulary confusion, failure of communication channels, differing needs in different disciplines, geographical separation, currency problems, lack of coordination etc.) often impose restrictions on the ability of users to obtain the data they need within the time available.

In making its study, CODATA/ADD gained understanding of the nature and varying characteristics of scientific and technical data. It learned that different disciplines acquire, store, evaluate and apply data in divergent ways. It learned that many separate organizations are involved in the matter of data accessibility and

dissemination, with no clear definition of responsibilities among those taking part. CODATA/ADD has attempted to summarize in the attached Report what it learned, so that UNESCO might have a full picture of this vital topic.

Finally, CODATA/ADD has analyzed its findings and prepared a list of recommendations (chapter I) which it has attempted to present in specific and clearly defined terms. For summary purposes, the recommendations may be placed in five groups, as follows:

Group I (1-3)

Areas where governments and scientific organizations should give greater financial support and more explicit endorsement to existing programs for evaluation and dissemination of data.

Group II (4-10)

Description and proposals for implementation of a three-element organizational structure for providing data services for users in developed and developing countries.

Group III (11-13)

Areas where CODATA, scientific organizations and private publishers should provide specific working tools (such as directories, handbooks and announcements) to increase accessibility of data.

Group IV (14-17)

Areas where standardizing bodies and international organizations should take specific steps to improve compatibility and/or standardization of data outputs and data-handling procedures.

Group V (18-29)

Specific steps to be taken in regard to data accessibility problems in developing countries.

TABLE OF CONTENTS

SEPARATE DOCUMENT

I. RECOMMENDATIONS

ASSISTANCE TO DEVELOPING COUNTRIES

A. INTRODUCTION

(1) Important Roles of Data in Science and Technology

In modern civilization scientific knowledge constitutes the most important objective knowledge concerning Nature. It has been acquired by the strenuous efforts of numerous research workers during the most recent centuries. The remarkable progress of natural science has been possible by virtue of the systematic structure of scientific knowledge: namely, new research results are built upon the achievements of predecessors, and this successive accumulation of knowledge has led to the construction of the huge edifice of modern science. The essential factor in this construction process of science is the smooth communication of research results, as scientific information, among research workers, so that scientists are well-informed of the achievements of their predecessors and can be rid of research duplication. Thus scientific information serves as the key link connecting research workers. The applicability of the achievements of science to technology also depends on this feature of scientific knowledge.

Data with which we are concerned in this Report may be regarded as a "crystallized" presentation of the essence of scientific knowledge in the most accurate form. Data, as usually understood in physics and chemistry, are numerical data representing the magnitudes of various quantities and, until recently, CODATA was mainly concerned with this kind of data. If we further include basic qualitative data such as the chemical structures of molecules, decay schemes of unstable nuclides, sequences of genes on chromosomes, etc., it may not be unrealistic to say that data constitute the reliable essence of scientific knowledge.

Data can not play their proper, significant roles unless their degree of accuracy and dependability is guaranteed. For this reason

evaluation is indispensable to data. Evaluated data should pass freely within the scientific community, independent of the research workers who contributed to the generation and evaluation of these data. With the enhanced growth of this flow of information, science can, not only make its own progress, but also contribute effectively to the benefit of human society.

There are, of course, varying degrees of evaluation; furthermore, any selection of recommended values is subject to change in the light of improved measurements. Even in the case of the fundamental physical constants, the most thorough evaluation cannot be regarded as giving a "final" set of values. Therefore, all evaluated data should be understood as the most reliable values at the time of evaluation. Utilization of such data can be expected to continue in parallel with efforts to establish more definitive values. The situation is different for time-dependent data that cannot be remeasured; see B(1).

(2) Characteristic Features of Data Dissemination Services among Scientific Information Services in General

Until recent years the publication of data compilations served as the principal means of dissemination of evaluated scientific data. Recently, however, data evaluation centres have proliferated and it has become exceedingly difficult for most institutions to acquire and store all the data compilations published in many countries. Furthermore, a significant portion of scientific data is not published at all, but stored in various depositories. The International Compendium of Numerical Data Projects, edited by CODATA, is an important tool for identifying various data sources throughout the world but, for geographical and other reasons, the need has emerged for more direct mechanisms to improve access to and dissemination of evaluated data. This constitutes the background to which our Task Group, CODATA/ADD,

was created by CODATA in 1972.

In this section some characteristic features of data dissemination services will be briefly discussed.

In most cases the task of scientific information services is to help users identify and obtain scientific papers which may contain the information they seek. Thus the ultimate units of service are individual papers. Abstracts and indexing journals are intended to serve as useful tools for identifying the required papers. In SDI and retrospective search services users usually receive lists of titles of scientific papers prepared by the information services.

On the other hand, the ultimate units of data services are individual pieces of data, which are units of scientific knowledge themselves. Evaluated data can be isolated from many original sources from which data were collected. Although, sometimes, users may wish to refer to source papers to find out how the data were measured, the main function of data dissemination services consists in providing users with the actual data, together with information pertinent to those data.

This contrast between documentary information services and data dissemination services leads to the following differences in their characters.

Firstly, characterization of scientific papers by abstracts and/or by key-words or descriptors is considerably more difficult and incomplete compared with the characterization of a piece of data in terms of the kind of quantity, substance and its state, and other conditions of measurement. Classification and indexing of data can be more precise than for scientific papers.

Secondly, the personnel engaged in data dissemination services must have deeper scientific knowledge than those in documentation services such as libraries, etc. Identification of data requires

the detailed scientific description of properties, substances, phenomena and measuring conditions. Also, various processes in data activities, such as evaluation and indexing, can be handled only by persons with the appropriate scientific background.

A data dissemination service may not necessarily be provided by an independent agency; it may be organized jointly with a documentary information service, or in association with a data evaluation centre. This problem will be discussed later in Chapter G.

B. CATEGORIES OF DATA

For several years after its establishment CODATA was almost exclusively concerned with data in physics and chemistry. Quite recently, however, CODATA has decided to expand its scope to cover data in geosciences and biosciences, and some data primarily of engineering interest may be included. Even if we confine ourselves to data in the natural sciences, the variety of data to be classified in a systematic way is almost as extensive as the variety of research. Because of this wide range, it is necessary to specify the subdiscipline of science in which one has developed the data, and then to use standard categories for organizing the data developed by the research workers in those subdisciplines.

By way of general introduction to the description of categories of data involved in this Report, Table 1 lists varieties of data under categories that are independent of the disciplinary designations and gives examples of data in each category of the general disciplines in the natural sciences. An explanation is now given of these categories of data.

(1) Time-Independent and Time-Dependent Data

a_1: Data which can be measured repeatedly

a_2: Data which can be measured only once

Data in pure physics and chemistry are normally of type a_1. They may be of type a_2 in some rather rare cases where purely physical data are obtained by utilizing geophysical or cosmological phenomena, such as data of high energy physics obtained from cosmic ray observations, curvature of light paths due to a gravitational field observed on the occasion of a total eclipse of the Sun.

TABLE 1 - VARIETIES OF CATEGORIES OF DATA

	CATEGORIES OF DATA	CHEMISTRY/PHYSICS	GEO-/ASTRO-SCIENCES	BIOSCIENCES
a_1	Data which can be measured repeatedly	Most data	Geol.structures, rocks Accel. due to gravity Fixed stars	Most data
a_2	Data which can be measured only once		Volcanic eruptions Solar flares, novae	Rare specimens Fossils
b_1	Location-independent	Most data	Minerals	Most data, excl. extraterrestrial
b_2	Location-dependent		Rocks, fossils Astronomical data Meteorological data	Rare specimens Fossils
c_1	Primary observational or experimental data	Optical spectra Crystallographic F-values	Seismographic records Weather charts	
c_2	Combinations of primary data with the aid of a theoretical model	Fundamental constants Crystal structures	Shape of Earth Temp.distribution in Sun	Genetic code
c_3	Data derived by theor.calculation	Molecular properties calc.by quantum mechanics	Solar eclipses predicted by celestial mechanics	

TABLE 1 (contd.) - VARIETIES OF CATEGORIES OF DATA

	CATEGORIES OF DATA	CHEMISTRY/PHYSICS	GEO-/ASTRO-SCIENCES	BIOSCIENCES
d_1	Determinable data	Most macroscopic data		
d_2	Stochastic data	Polymer data Structure-sensitive properties	Soil composition Solar flares	Most data Metrology
e_1	Quantitative data	Most data	Seismic data Meteorological data	
e_2	Qualitative data	Chemical struct.formulae Properties of nuclides	Rock classification	Amino-acid sequences Classification of strains
f_1	Data presented as numerical values		Meteorological data	
f_2	Data presented as graphs or models	Phase diagrams Stereoscopic molecular diagrams Molecular models	Geological maps Weather maps	Genetic pathways

Note: A given group of data can be categorized simultaneously by several 'facets' a,b,c etc.;

for instance, the nature of meteorological data characterized as a_2,b_2,c_1,d_2,e_1 and f_1 (or f_2).

There are type a_1 data as well as type a_2 data in geosciences and astronomy. Data concerning geological structures, rocks, acceleration due to gravity and fixed stars are normally of type a_1, while those concerning volcanic eruptions, atmospheric conditions, solar flares and novae are examples of data of type a_2. Data which are ordinarily considered as type a_1 may have the character of type a_2, depending on the accuracy of measurements and length of time scales concerned; examples are found in geodesic data, chemical composition of the atmosphere, etc. Most biological data are of type a_1, although records of abnormal individuals, of extremely rare species and of rare fossils have type a_2 character. Furthermore, in cases when remeasurement is very difficult, for technical or financial reasons, data which are, in principle, of type a_1 become practically type a_2: for example, data concerning the Earth's interior and deep ocean beds, data concerning the Moon and the planets observed by manned or unmanned space ships, belong to this category.

Concepts and methods of evaluation are different for data of these two types. Since, until now, CODATA has been mainly concerned with data type a_1, the main scope of the present Report will be data of this type.

(2) Location-Independent and Location-Dependent Data

b_1: Data independent of location of objects measured

b_2: Data dependent on location of objects measured

Data in pure physics and chemistry belong to type b_1, while data in earth sciences and astronomy normally belong to type b_2. It is important to note that data on minerals are usually of type b_1, while data on rocks are of type b_2. Most data in biology are of type b_1, if extraterrestrial life is excluded, but some of them will have the character of type b_2 when ecological and biogeographical aspects are introduced. Rare data in biology, such as mentioned

under (a), are of type b_2.

(3) Primary, Derived and Theoretical Data

c_1: Data obtained by experiment or observation designed for measurement of this particular quantity (primary data)

c_2: Data derived by combining several primary data with the aid of a theoretical model

c_3: Data derived by theoretical calculations

Distinctions between c_1 and c_2, and c_2 and c_3 are not clearcut. We understand c_1 in a rather broad sense; for instance, values of velocity derived by measuring length and time, or by measuring the Doppler shift in the frequency of emitted or reflected light, may be considered as data of type c_1. Very often experimental determination of a single quantity q is done by using values of other quantities, but as long as the main experimental efforts are made in measuring q, the values obtained are considered as data of type c_1. Thus, Millikan's oil drop experiment provided a datum of electronic charge of type c_1.

It should be noted that data of type c_1 are generally processed in some manner before they are reported. The degree of processing may vary over a large range. While appropriate processing can make the data more convenient for a wider circle of users, some information of interest to others may be lost. If this is the case, it is advisable to provide for storage of the data in its form prior to the stage of processing at which the information is lost. This is especially important with type a_2 data or other data that are difficult to remeasure.

Well-known examples of data of type c_2 are the values of fundamental constants, atomic weights, etc. Data on electronegativity of atoms, Fermi surfaces in solids, the shape of the Earth and the temperature distribution in the Sun are other examples of data of

type c_2. Data on crystal structures derived from X-ray analysis may be regarded as of type c_2 if F(hkl)-values are considered as primary data. The genetic code in molecular biology is an example of type c_2 data in biology.

Examples of type c_3 data are data concerning solar eclipses predicted with the use of celestial mechanics and data on molecular properties calculated with the use of quantum mechanics. It is true that basic data such as fundamental constants are used in such work, but the main scientific efforts in these cases are theoretical calculations.

(4) Determinable and Stochastic Data

d_1: Data on a quantity which can be assumed to take a definite value under a given condition (determinable data)

d_2: Data on a quantity which takes fluctuating values from one sample to another, from one measurement to another, etc., even under a given condition (stochastic data)

Most data in macroscopic physics and chemistry may be regarded as determinable data. In principle, minute fluctuations due to Brownian motion are often present but usually they can be safely ignored. However, some properties which are called structure-sensitive may give stochastic data, such as properties concerning the fracture of solids. Stochastic nature becomes apparent in the intermediate region between macroscopic and microscopic, e.g. in studies on polymers which usually take different shapes in solution and in Brownian motions in general.

At the microscopic level the stochastic nature of data is related to the quantum-mechanical uncertainty. If one takes a sample containing N radium atoms and measures the number of α-particles emitted in one second, data will be stochastic and,

only in the case of very large N, do the data become determinable, corresponding to the situation that the probability of disintegration is definite.

In biology, most data on individual animals and plants of a given species are different from one specimen to another. Such data may be regarded as stochastic in our sense, if the "given condition" is understood to be the specification of the species. In radiation biology and pharmacology (including toxicology) the effects on animals of physical and chemical actions are expressed by data of this type. Lethal doses of toxic substances are examples of such data.

Time-dependent data are not necessarily stochastic; they are usually determinable if the "given condition" is understood to include the specification of time. If the "given condition" does not include the specification of time, or the time is only broadly specified, time-dependent phenomena may give stochastic data. In this sense some meteorological data may be considered as stochastic.

In the case of stochastic quantities, statistical data obtained from a number of individual measurements are of practical importance. LD_{50} (lethal dose for 50% deaths) is the statistical median (middle-value) of lethal doses. Population statistics, such as statistics of age distribution give another example of statistical data. Statistical data in metrology are familiar to us.

(5) Quantitative and Qualitative Data

e_1: Quantitative data

e_2: Qualitative data

Quantitative data in the strict sense are measures of scientific quantities expressed in terms of well-defined units. Scientific quantities should be defined by logic intrinsic to science, and the sum and difference of two quantities of the same kind should have

definite meanings. In this case the definition of a unit suffices to reduce the magnitude of a quantity to a numerical value. Thus quantitative data are given as numerical data. Most data in the "exact sciences", including physics and chemistry, are quantitative data in this sense.

There are some "quantities" employed in science which can be measured only by using arbitrary scales. Until thermodynamical temperature was established, temperature was measured by using more or less arbitrary scales and the sum of two temperatures had no definite meaning. There is no absolute definition of the hardness of solids independent of the method of measurement. In meteorology the force of winds is commonly expressed by the so-called wind-force scale. Such data must be placed in a different category from data expressible in terms of the accepted SI base units.

Qualitative data, taken broadly, may include any scientific definitive statement concerning scientific objects. Qualitative data in this broad sense are almost equivalent to established scientific knowledge. It is almost certain, however, that CODATA will not broaden its scope to such an extent. For our present purpose it seems reasonable to include basic qualitative data concerning objects of which numerous varieties exist in the same category. For instance, structures of molecules as expressed in terms of structural formulae are basic qualitative data in chemistry and the objects in this case are molecules whose varieties are enormous. The number of different nuclides is limited but, even so, the basic properties of nuclides, such as stable, α-radioactive etc., are important qualitative data.

In biochemistry and molecular biology the sequences of amino-acids in proteins and of nucleotides in DNA and RNA constitute basic qualitative data, of which compilations are actually being published. In microbiology, classification of strains of micro-organisms is made in terms of a set of answers (affirmative or negative) to posed questions concerning morphological, metabolic and other characteristics of each strain. Such data belong to type e_2.

In the practical handling of data, qualitative data can be treated as numerical. For instance, "yes" and "no" in the above example can be coded by 1 and 0, and then the classification data of micro-organisms can be represented by numbers on the binary scale.

In geology and biology it is often very important to conserve objects from which data have been derived and make them available for further studies. Fossils and meteorites are examples of such objects. These objects may be regarded as rich stores of valuable data yet unmeasured. However, discussion of this problem may be irrelevant to the proper content of this Report.

(6) Data as Numerical Values, Graphs or Models

f_1: Data presented as isolated numerical values

f_2: Data presented in graphical form or as models

This distinction refers to presentation but, even in the course of generation, data are often obtained in graphical form when the quantity measured depends on a parameter. The parameter may be spatial distance (e.g. in paper chromatography), time (in many recording instruments), wavelength (in spectrography), voltage (e.g. in measurement of I-V characteristics of semi-conductors) and so on.

In some other cases, graphs are constructed for the sake of helping users grasp a mass of data by visual perception. Phase diagrams of binary mixtures represent one example. Charts and maps (such as geological maps) also belong to this category.

Stereoscopic vision is used to visualize three-dimensional structures, such as "tertiary" structures of biological molecules.

Transformation of discrete sets of numerical data to graphs and charts and the reverse transformation are required in the course of data processing and data presentation. In the latter transformation (pattern coding), use of Fourier transforms, derivation of some numerical values which characterize given graphs and other methods are used in addition to the conventional method of choosing discrete values of parameters and giving numerical data corresponding to these parameter values.

In addition to these aspects, data may be classified according to the nature and size of the user communities. For instance, data may be classified into those of some interest to engineers and those of little interest to them. It is important to bear in mind, however, that distinctions based on expected users' needs may change with the progress of science and technology, and even with the advancement of education in general. Due consideration of the classification of data based on users' needs is very important for the purpose of the present Report and this problem will be discussed in chapter D.

C. DATA COMPILATION AND DATA SERVICE FROM THE USERS' VIEWPOINT

(1) Historical Background

Reflecting the attention of scientists to the important roles of numerical data in science and technology, there have been various landmarks in data compilation activities. These were the Landolt-Börnstein Tabellen, the first edition of which appeared in 1883 in Germany, Tables Annuelles de Constantes et Données Numériques appearing in the years 1910 to 1930 in 10 volumes in France, the International Critical Tables of Numerical Data of Physics, Chemistry and Technology, in 8 volumes in the years 1926 to 1933, etc.

As shown by the fact that the Landolt-Börnstein Tabellen increased its pages from 281 in the 1st edition in 1910 to 20,000 pages in 26 volumes of the 6th edition in 1950-1959, the volume of numerical data has exceeded the capability of these data compiling activities to cover all fields of sciences. The International Critical Tables is considered to be the great general comprehensive data compilation work and copies are being sold even today, after 40 years of publication. According to the survey in 1965 on the needs of American Chemical Society members for property data, this Table was mentioned by half of its respondents (801 members) as the most frequently consulted data compilation, even though most of them commented that present data compilations satisfy their requirements poorly, or at best moderately. Users throughout the world have always hoped that revisions and supplements of the Table would be produced. However, the U.S.National Research Council Committee on Tables of Constants, which took the financial and editorial responsibility for this international compilation work, concluded in 1955 that it was impossible to carry out the work for a complete revision and

extension of the International Critical Tables.

On the other hand, there is always a demand for handy, time-saving books of tables for the everyday use of research workers. These are selective listings matched to their needs and they are easier to keep up-to-date than large exhaustive compilations like the International Critical Tables. The Kaye and Laby Tables of Physical and Chemical Constants, compiled in Britain, came out as a single volume and ran through 13 editions from 1911 to 1967, being revised every 2 to 4 years.

The tendency arose to publish not a complete general data compilation, but only separate volumes covering specialized topics, such as a new series of volumes of the Landolt Börnstein tables which appeared after 1959, in nuclear physics and technology, magnetic properties, astronomy and astrophysics, atomic and molecular physics, etc. Similarly, the Tables Annuelles de Constantes et Données Numériques became Tables de Constantes Sélectionnées.

In that period a number of continuing data compiling activities, limited to particular disciplines or with separate publications devoted to particular sets of quantities, had come into existence in the United States. These are the American Petroleum Institute Research Project 44 on the physical, thermodynamic and spectral properties of hydrocarbons and related compounds, the Manufacturing Chemists Association's Projects, the U.S.Atomic Energy Commission Projects on nuclear data, etc. The same kinds of data activities were also being carried out in several countries to meet the urgent needs of the scientific communities of their respective countries for up-to-date, evaluated numerical data in various disciplines.

In 1957 the Office of Critical Tables was established in the United States by the National Academy of Sciences-National Research Council for the coordination and encouragement of data compiling activities and for better accessibility of necessary data. This Office was the first administering organization which was established solely for data activities. With the support of the Office of Critical Tables, and resulting from the wide recognition of the importance of data activities and the necessity for strong and systematic funding of these activities, the National Bureau of Standards was assigned the responsibility for the establishment of the National Standard Reference Data System and the Office of Standard Reference Data was created for administering the system in 1963. The general functions of the system are to coordinate and integrate existing data activities into a systematic, comprehensive programme, supplementing and expanding technical coverage when necessary, establishing and maintaining standards for the output of participating groups, and providing mechanisms for the dissemination of the output as required.

Reflecting this American development, the Department of Scientific and Industrial Research in Britain established a governmental programme for data compilation in 1964. After the reorganization of the Government civil service agencies in 1965, the Office for Scientific and Technical Information (OSTI) of the Department of Education and Science, together with the Department for Trade and Industry (DTI), took over this function and now carry the responsibility for coordinating and promoting data activities in Britain. A survey was made on the status of British data activities and two lists entitled Critical Data in Britain were issued. The list, in its newest edition, under a revised title, Data Activities in Britain, identified about 100 projects

in progress by their subject fields with information on organization, coverage, analysis and publications. Among these are current data projects in X-ray crystallography, mass spectroscopy, thermodynamic properties of gases, etc.

In the Soviet Union, the Academy of Sciences supports several data evaluation projects in the physical sciences and the State Service for Standard and Reference Data (GSSSD) has broad responsibility for scientific and technical data.

As shown by the fact that revisions and regular extensions of comprehensive data compilation works have become difficult according to the experience in the United States of the NRC, it is widely recognized that no one country could supply all of the financial, technical and manpower resources needed for the production of data compilations to cover broad and diverse fields. Canada, France, Germany, Japan and Poland were also making efforts in coordinating data activities within each country and moving toward participation in world-wide projects. With the purpose of providing the needed international cooperation and guidance, the Committee on Data for Science and Technology (CODATA) was established in 1966 by the International Council of Scientific Unions. Henceforth, with representation from 11 International Scientific Unions and 14 countries (as of 1973), CODATA has been actively encouraging, on a world-wide basis, the production and dissemination of critically evaluated numerical data. Each member nation has organized a National Committee for CODATA, attached to the ICSU adhering body. Each committee acts as a central body for promotion and coordination of data compilation activities and better dissemination of data within the country, as well as being the national link to CODATA.

In the United States, the Numerical Data Advisory Board was established in 1969, replacing the Office of Critical Tables, to cope with the growth of the National Standard Reference Data System at the National Bureau of Standards and the emergence of CODATA. It provides a focal point in the National Academy of Sciences-National Academy of Engineering-National Research Council in all matters pertaining to the compilation and evaluation of numerical data for science and technology. It has responsibility for the U.S.National Committee for CODATA and provides liaison between the U.S.scientific community and CODATA.

(2) Present Situation and Recent Trends in Data Dissemination

The first and major task of CODATA was to survey, on a world-wide basis, what work on the critical compilation of evaluated numerical data is being carried out in each country and in each Union and what output is available from each project. The results of the survey were published in book-form as the International Compendium of Numerical Data Projects (Springer-Verlag, New York - Heidelberg - Berlin, 1969, 295 pp). This book lists identifiable data compilation projects throughout the world with information on their scope, mode of operation and form of dissemination of output. It enables users to determine what compilations containing critically evaluated data are now available and what centres or organizations produce or aid the production of such data for publication on a continuing basis. More than 150 projects and data centres in 26 countries are described in detail. Approximately 120 handbooks and other sources of useful tabular data are listed with full bibliographic notes. The listing is not complete and more comprehensive coverage is hoped for future editions to make the Compendium function as a useful guidebook for better access to data.

The CODATA Newsletter announces, in every issue, current publications of data compilations with comments on their contents and accessibility. The CODATA Task Group on Data for Chemical Kinetics published a report (CODATA Bulletin, No. 3,1-28, 1971) which lists compilation and data evaluation activities in chemical kinetics, photochemistry and radiation chemistry. The major part of the report is a catalogue of current review articles in primary journals, with bibliographic information and abstracts which emphasize their data contents. Of about 230 sources listed in the catalogue, 125 have been published and others are 'in press', 'in preparation' or 'planned'.

As output of the NSRDS activities in the United States, 50 titles in the NSRDS-NBS series and about 70 other data compilations, bibliographies and translations have been published. Over 100,000 copies of these documents have been distributed since this programme began and this total is augumented by wide-spread secondary distribution of data in various handbooks.

In March 1972, a new powerful tool for wider, prompter and more user-oriented dissemination of standard reference data appeared as a quarterly journal publication, named Journal of Physical and Chemical Reference Data. The journal acts as a new channel for the primary dissemination of the output of the NSRDS activities and is expected to make a significant contribution, with the joint collaboration of the American Chemical Society, the American Institute of Physics and the National Bureau of Standards, to improving the accessibility of reliable reference data to the scientific and technical community.

Nuclear Data is a compilation journal in the field of nuclear-structure physics which solely provides collections and evaluations of data and proves that journal publication is a pertinent format for better dissemination of data. As early as 1956, the American Chemical Society started the publication of the Journal of Chemical and Engineering Data to meet data needs. Atomic Data, Organic Magnetic Resonance, International Journal of Chemical Kinetics, Journal of Chemical Thermodynamics, Mass Spectrometry Bulletin, etc., are new journals of international scope, specially concerned with data. Some of them have supplement issues which are devoted to longer data compilations, e.g. spectra of compounds in standard format with experimental details.

It is said that in recent years more than one million papers on science and technology are published each year in primary journals. Many of them contain numerical and quantitative data. Because of the startling growth of scientific papers a practice has recently arisen that authors are recommended to minimize data in papers thereby saving space and increasing the communication capacity of the journals. In order to publish research results as soon as possible, communication formats such as letters, short communications etc., which omit data, have become of great importance. Abstracts, which contain few or no data, have become significant tools in searching for information among voluminous scientific publications. Reflecting these circumstances, some situations have evolved whereby the numeric and quantitative data are dissociated from the non-numeric information. These are exemplified by the emergence of the data journals, mentioned previously as dissemination tools exclusively for data, the publication of sets of data sheets and the establishment, in many countries, of data banks or data centres where the data are generated or acquired, evaluated, stored for documentation and

prepared for dissemination upon request.

Since the end of the 1960's new techniques, media and systems for making organized data available have been introduced. The development of microphotography techniques has brought various micro-forms, including microfilm, microfiche, film strips and many others, as storage and distribution media for voluminous data. As a result of the progress of computerized techniques in the handling of numerical data, magnetic tape files of data are established or are being established in some data centres, as reported in CODATA Bulletin, No. 4(1971) by the CODATA Task Group on Computer Use. This markedly raises the efficiency of data retrieval and dissemination. In some cases, all printed output from these files is obtained by computer-based photocomposition.

Beside being a convenient medium for data storage, magnetic tapes are used for data dissemination. Copies of data files are now available on magnetic tape from such organizations as the American Society for Testing and Materials (infrared spectral data), the American Chemical Society (physical and thermodynamic properties), the National Neutron Cross Section Center (USA), the Crystallographic Data Centre (UK), the Mass Spectrometry Data Centre (UK), Centre d'Information de Thermodynamique Chimique Minerale (France) and the Banque IMA de Données Toxicologiques (France). In accord with the increasing use of computers by both data centres and users, data dissemination through magnetic tapes and disk packs is expected to be more widely practised in future.

References

Hall, R.M.S. : The Development of the United Kingdom Data Program. J.Chem.Doc. 7,18-20,1967.

Weisman, H.M. : Needs of American Chemical Society Members for Property Data. J.Chem.Doc. 7,9-14,1967.

Rossini, F.D. : Data for Science and Technology; from the Past into the Future. CODATA Newsletter 1, 2-4, 1968.

COSATI Directory of Federally Supported Information Analysis Centers; COSATI Report No. 70-1, 71 pp, 1970.

Critical Evaluation of Data in the Physical Sciences - A Status Report on the National Standard Reference Data System; NBS Technical Note 747, 72 pp, 1972.

D. PROBLEMS CONCERNING DATA COMPILATION AND EVALUATION, PARTICULARLY FROM THE USERS' VIEWPOINT

Concern for the needs of the user must be a first principle for all those who handle, evaluate and disseminate scientific and technological data. Reliable data may be stored in different organizations, but unless these data can reach the users in good time and proper form in response to his request, they are essentially useless.

On the other hand, to utilize data services effectively, users must have a good overall understanding of the collection, evaluation and organization of data, including the common usage of terminology, symbols and units in the disciplines concerned.

In the following discussion categories of data based on user characteristics are given and an outline of the generation, collection, evaluation and organization of data is described.

(1) Categories of Data Based on the Character of User Needs

In Chapter B it has been shown that data can be categorized from various points of view. It will, however, be more significant, from the viewpoint of data dissemination, to consider a categorization in terms of the population and character of data users as well as by the inherent character of the data themselves. This is very difficult to do in a rigorous way. Furthermore, any such classification is valid only for a limited time, since the progress of science and technology naturally leads to changes in the pattern of data use. However, it may be convenient to recognize the following broad classes of data:

g_1: data generated in a specific discipline and used almost exclusively by specialists in the same discipline

(Examples)

- Para- or diamagnetic susceptibility of compounds
- Electric quadrupole moments of atomic nuclei
- Constants representing the anharmonicity of normal modes of vibration of molecules
- Density of frequency distribution of normal vibrations of crystals
- High resolution infrared spectra
- Crystal structure factors of diffracted rays in crystal structure analysis: F(h,k,l)
- Seismographic records of earthquakes

g_2: data that are also used by research workers in a limited number of related disciplines

(Examples)

- Characteristics of ferromagnetic materials (initial susceptibility, saturated magnetization, coercive force, hysteresis and their temperature variation, Curie points, etc.)
- Binding energy of atomic nuclei from protons and neutrons
- High resolution NMR spectra
- Infrared and Raman spectra
- Optical transition probability
- Rate constants of chemical reactions
- Steam tables
- Number of chromosomes in cells for biological species or strains

g_3: data used more widely

(Examples)

- Fundamental physical constants
- Physical properties of materials
- Physico-chemical properties of organic and inorganic compounds, including fundamental thermodynamic data
- Electronic structures of atoms
- Atomic structures of common molecules
- Geological structures (geological maps)
- Toxicity of chemicals
- Human visual sensitivity to colours

It should be emphasized that the distinctions between these categories are not clearcut. As science and technology progress, some data that have been used mainly within a single discipline become important in other disciplines. A typical example will be seen in the case of the electric quadrupole moment (e.q.m.) of an atomic nucleus, categorized in g_1 above; data on e.q.m. are not only important in nuclear physics but are also useful for finding the electron distributions near the nuclei in atoms, molecules and solids since the products of e.q.m. with the electric field gradient at the position of the nuclei are measured by the hyperfine structure of spectra, electric quadrupole resonance, etc. To mention another example, data on microwave spectra, which are mainly used for the study of the structure of molecules, have proved to be useful for identifying molecules existing in interstellar space.

(2) Generation and Presentation of Data

Data in science and technology are usually obtained by experiments or observations which are carried out, in most cases, by individual or small groups of research workers. In many cases data are obtained, as the need arises, as values of quantities measured in the process of performing research although, in some cases, the object of the research is to obtain the data. In the former case, the data are included in the original papers and are presented in primary literature such as scientific journals, only if the author deems it necessary in describing the main results of the research. Some of the data are not published at all when they are not essential for the description of the main results of the research or when they have been obtained solely as a by-product of the research. For example, data on specific heat as a function of

temperature may remain unpublished, even though the temperature dependence of the specific heat has been precisely measured, when the purpose of the study is to find phase transitions in solids and the result is negative. It is, however, desirable that such data are submitted to and stored in appropriate data banks and data depositories, thus facilitating their potential utilization.

Raw data obtained directly by experiments and observations, such as readings of various instruments and curves plotted with automatic recorders, are also of significance for further use and are sometimes published. Optical absoprtion spectra obtained with automatic recording spectrographs and Mössbauer spectral data plotted with automatic Mössbauer instruments are examples of these. In some cases only the processed data which are regarded as having scientific significance are published. For example, in light absoprtion spectroscopy the maximum absorption wave lengths and the half-width values of the main absorption bands are often published but not the spectra themselves. In ESR, NMR and ENDOR experiments, where the main interest concerns paramagnetic relaxation, the published papers often exclude the raw data and include only the values of the relaxation times T_1 and T_2 with some description of the method of measurement. In the analysis of crystal structures by X-ray diffraction, it is often the case that the F-values indicating the intensity and phase of each diffracted wave are not reported in the paper but the electron density distribution or the locations of atoms which have been deduced are presented. Such "omissions of raw or intermediate data" are encouraged by the recent tendency to limit the increase in the size of scientific journals. It is, however, recommended that such intermediate data are made available to research workers in the same discipline, especially when they are difficult to obtain. They should be stored in special data banks or kept by the authors, even though they belong to category g_1.

Data which are used by many scientists, such as those in categories g_2 and g_3, are often generated systematically as well as being obtained in the process of research. In this case, systematic production of high-quality data satisfying certain criteria is the direct purpose of the study. Such activities will be further categorized as follows:

(i) Data generation with standard instruments and methods by a research organization; for example, data generation of infrared absoprtion spectra. In this case the research organization carries out data compilation automatically and may be called a systematic data generating project.

(ii) Generation of high-quality data by world-wide organizations and laboratories which collaborate in the sharing of responsibilities. Examples of this category are seen in the "steam-tables project", determining the thermodynamic and physical characteristics of water with high precision to high temperatures (an international organization IAPS, International Association on the Properties of Steam, has been set up for this purpose); also the project of IUPAC to determine with high precision the thermodynamic properties of industrially important gases.

In the earth sciences and astronomy, there are a number of temporary and permanent organizations for observation. Acquisition of geophysical and solar data in the IGY (International Geophysical Year) and the IQSY (International Quiet Sun Year) was carried out by a cooperative organization of observation over a limited period, and the data thus obtained are gathered at several World Data Centres. FAGS (Federation of Astronomical and Geophysical Services) is one of the permanent cooperative organizations concerned with the assignment of observation tasks.

Meteorological data are obtained through systematic observation

in each country by a network of meteorological stations and observatories, and global cooperation has been established under the auspices of WMO (World Meteorological Organization). In recent years observation of cloud amount, etc. with meteorological satellites is being carried out in a global system. In these cases voluminous raw data have been generated and most of them are transformed by specialists to category g_3 where they are widely utilized. It is, however, necessary to establish a well-organized storage system for raw meteorological observation data, because these data will be used for many purposes other than for meteorological research. It has also been recognized that photographs of the Earth's surface taken from satellites tend to be widely used for the study of geological structures and environmental problems.

(3) Collection and Evaluation of Data

As described in Chapter A (1), data cannot be fully utilized without being evaluated to some extent. If each research worker who generates data through measurement and/or observation checks the measurement sample thoroughly, pays careful attention to minimizing systematic and random errors, and endeavours to use international standards in presenting numerical values and graphs, such efforts would be highly appreciated as the first step to evaluation. More systematic evaluation, however, is carried out by data evaluation centres. Those various data activities mentioned in (2) are closely concerned with the generation of data, and such systematic data generating projects in particular evaluate data more rigorously than is customary for other data generators. Most other data evaluation centres extract data from papers published in journals and reports, and evaluate them. In this case the collection of data is an important aspect of data activities. The word "collection" in the title of this section is used with this meaning.

The task of extracting data from papers is very difficult because data do not usually appear in the titles nor the abstracts of papers and are not easily indexed with keywords. A retrospective data search in the older papers is even more difficult and, as a result, published data are sometimes completely overlooked.

To avoid this, it is recommended that the important data should be indicated in the abstracts. It may, however, be limited to the data which the author set out to collect in his study. A more comprehensive method would be for bibliographic information centres to extract data at the time of abstracting and indexing and refer this information to the data evaluation centres. In order to do this, it is necessary that data evaluation centres and bibliographic information centres act in cooperation. It is also of significance that systematic cooperation is planned between the editors of journals and data evaluation centres so that important data which appear in the papers may be indicated by the authors to the editors and this information communicated to the data evaluation centres. Data descriptive records used in abstracts in nuclear physics and a flagging system recommended by an IUPAC committee will be very helpful in this respect. The realization of such cooperation is worthy of full examination so that unnecessary duplication of intellectual effort can be avoided.

As to the collection of data which do not appear in papers, cooperation between the data depositories and data banks described in (2) is also necessary. This is particularly important for engineering data. In some cases, it will be profitable to promote active cooperation with individual research workers or research organizations involved in generating the data of interest to data evaluation centres.

There follow some comments on data evaluation:

(i) Evaluation of type g_1 data owes very much to careful examination by the generator because it is difficult to establish data evaluation centres for such data which are used by a limited number of scientists. Sometimes, however, when the number of research workers in a special field is fairly large, although the field is almost independent of others, data are collected and evaluated by the spontaneous efforts of some of the research workers and are presented in specialized journals. It is recommended that such an activity be listed in the CODATA International Compendium as a data evaluation project irrespective of its scale, provided that it is carried out on a continuing basis.

(ii) For the voluminous intermediate raw data of type g_1, organizations have been mostly established where such data are processed and data useful to a wider range of users are made available. This is seen in time-dependent data (type a_2) such as in the case of meteorological data, or in location-dependent data (type b_2) such as data in geosciences. In these cases the most important factor in the evaluation is the elimination of systematic errors in the measurements or observations and raising of the precision of the measurements by standardizing the instruments and their mode of use. In processing raw data, some data conversion may take place which appears inconsistent with normal evaluation procedures. For example, when weather and climate maps are plotted small variations of data in a localized area or over a short time period are smoothed out. The task of making reliable weather maps, employing the meteorological elements in the upper atmosphere, results of radar observation, etc., in addition to the ordinary observed meteorological elements, can be regarded as a data evaluation process which includes the generation of data of type c_2.

(iii) In the organized generation of data, in particular of type g_2 and sometimes of type g_3, evaluation of the data consists of a systematic generator's evaluation. In the case of infrared spectra, data are compiled which are obtained under certain measuring conditions specified according to Classes I, II and III of the Coblenz Society. In this case, measurements will proceed almost automatically when the standard sample is obtained, and evaluation mainly consists in the checking of the purity of the standard sample and the measuring conditions. In the case of NMR spectral work a similar specification of the measuring conditions has already been prepared and published.

(iv) Theoretical considerations are often necessary in the evaluation of data in the categories of g_1, g_2 and g_3. A typical example can be seen in the consistency of theoretically related data among the values of enthaply, entropy, internal energy, free energies, specific heat, and vibrational and other excitation energies in molecules and solids. When the number of variables measured is larger than that of the theoretically independent variables, the primary condition for evaluation will be to obtain the most probable or the most reliable set of values which, as a whole, are theoretically consistent with the variables concerned.

(v) It sometimes happens that several fundamental values are not measured independently but derived from measured values of various combinations of them. A most typical example of this is seen in the determination of fundamental physical constants. For the values of the velocity of light in vacuum (c), Planck's constant (h), elementary charge (e) and electron mass (m), such combinations as e/m, h/mc, e/h, e^2/hc, eh/mc and me^4/ch^3 are measured in addition to the direct measurements of c and e. It should further be noted that the quantum electrodynamic fine structure of hydrogen atoms (Lamb shift) can be expressed as a series of these fundamental constants.

Over a long period work has been caried out on the critical examination of each measurement method, for the checking of probable errors for each of these various measured data and for the determination of a consistent set of values of the fundamental constants with their associated probable errors. The values of the fundamental constants were presented in 1963 by Birge and others. Later, there were some activities by the IUPAP Commission on Fundamental Constants and Related Constants. In recent years the CODATA Task Group on Fundamental Constants, chaired by Dr.E.Richard Cohen, was established and a new set of recommendable values has been determined. These activities can be regarded as sophisticated examples of data evaluation.

(vi) A scientific quantity is very often a function of one or more parameters. Many physical quantities, for example thermodynamic quantities, are usually functions of temperature and pressure, and the acceleration due to gravity is a function of altitude. Independent measurements of these values do not always correspond to the same values of the parameters. A complicated evaluation will be required when the determination of values as a function of parameters is undertaken from the numerous data concerning a quantity of this kind. When a theoretical formula can be obtained for the function, this task will be concerned with the determination of parameters included in the formula, but sometimes sytematic deviation from the theoretical formula has to be considered. When the theoretical formula is not available, evaluation is often made by using a least squares fit of the measured values to a suitable experimental formula including a few parameters. By these methods original data can be evaluated indirectly. This is regarded as a kind of smoothing but careful attention must be paid when this method is employed because the data which are supposed to show an intrinsic anomalous behaviour over some range of the parameter may be missed; for example, the

possible non-detection of delicate phase transitions in solids, which could be found by careful measurement of the temperature variation of specific heat.

(vii) As a result of the above considerations, in order to evaluate the data completely from the collection of published data, it will be necessary to obtain more detailed information on the measurement from the original author and/or to remeasure the same quantity by a new experiment. Quite a few data centres, e.g., Thermophysical Properties Research Center, are provided with such experimental facilities.

(4) Units, Symbols and Nomenclature

According to its Constitution the functions of CODATA include "To encourage the use of the nomenclature, terminology, symbols, constants and units advoated by the responsible Unions, and to promote the establishment of guidelines for the presentation of scientific data".

The most important requirement in expressing data is a clear indication of units. Since the same data may be used in many disciplines and in many countries, conformity to internationally accepted units, as well as recommended symbols and nomenclature, is highly desirable. This is particularly true of data in categories g_2 and g_3, but also applies to category g_1, since such data may achieve wider use at a later time.

(i) Use of SI Units

The International System of Units (SI) has been adopted by the General Conference of Weights and Measures (CGPM) and endorsed by most scientific Unions. The details of the SI and of the non-SI units that are currently considered acceptable by the CGPM are given in Le Système International d'Unités (SI) (English translations published as National Bureau of Standards Special Publication 330, 1970 and by Her Majesty's Stationery Office, London, 1970) as well as in

documents of various Unions. These units should be used unless there are strong reasons to do otherwise. If non-recognized units must be used, they should be clearly defined in terms of the SI in the document where the data are reported.

(ii) Terminology and Symbols for Scientific Quantities

In reporting data, both quantitative and qualitative, due consideration should be given to the terminology, nomenclature and symbols for scientific concepts and quantities. Most of the Unions in ICSU have committeees dealing with terminology, nomenclature and symbols in their disciplines, and many have published manuals such as Symbols, Units and Nomenclature in Physics (SUN Commission, IUPAP, 1965) and Manual of Symbols and Terminology for Physico-chemical Quantities and Units (Commission on Symbols, Terminology and Units, IUPAC, 1969). The recommendations of the Unions should be followed whenever possible. In any event, each symbol for a quantity should be defined clearly upon its first use in a paper.

As science progresses, new concepts and theories may require the introduction of new quantities and symbols. A period of confusion sometimes results in which widely varying practices are followed. In such cases the Unions concerned should move to standardize terminology as soon as intelligent decisions can be made.

(iii) Nomenclature and Encoding of Chemical Substances

There is a real need for a unique definition of chemical structure in natural language which is understandable on the printed page and yet logical and unambiguous to a computer programme. The natural language representation of each structure must have a unique positioning in any organized list. The original system must be well designed so that it will not be necessary to make continuous changes. This is the agreement at the 1971 meeting of IUPAC Interdivisional Committee on Machine Documentation in the Chemical Field.[1]

(1) IUPAC Information Bulletin, 42/43, p.7, 1972.

IUPAC recommends definitive rules for nomenclature of chemical substances.[1] However, the rules define the principle of naming rather than the unique name of a compound. Existing IUPAC nomenclature recommendations are certainly not all suitable for machine handling. At present, we have no completely systematic nomenclature of chemical compounds, though many efforts are directed towards the machine handling of chemical structures and the computer generation of nomenclature.

However, simple referencing of compounds in a computer can be performed through an arbitrary code number instead of the complete name. Such a number is provided by the Chemical Abstracts Service with the 8-digit Registry Number. Since referencing is not the only operation that computers must perform, other kinds of chemical compound representations are needed i.e. chemical codes. The latter have advanced in the last ten years, partly due to the improvements of computer technology. Earlier computers were slow, programming was difficult and storage capacities were poor. Also communication with them was restricted by the punched card technique. Now modern computer systems offer ever increasing processing capabilities, storage capacities and input-output communication devices.

As a result of technlogical progress, it is now possible:

- to register large numbers of chemical compounds and, for every compound, large amounts of data, including structural data,
- to use sophisticiated software either for mere documentary tasks or for computer-aided design projects,
- to handle directly chemical diagrams via graphical input and output.

In order to benefit from these improvements, a computer representation of chemical compounds should fulfill the following

---

(2) See eg. Nomenclature of Organic Chemistry, issued by the Commission on the Nomenclature of Organic Chemistry; Butterworths, London, 1969.

requirements:

R 1: it should be easily generated by a computer program, according to a logical set of rules, from a graphical input,

R 2: it should be able to generate a two-dimensional or three-dimensional display of molecules for editing and design purposes,

R 3: it should provide easy access to search elements for documentary processing, particularly for flexible substructure search,

R 4: it should make provision for computer-aided design processing,

R 5: it should be open enough to cope with every single bit of structural information to be registered, for present or future needs.

In addition to these intrinsic qualities, an encoding process should also be interconvertible with other codes. Interconversion is a necessity for considerations of compabitility with existing files, increasing costs resulting from the growth of files, increased user needs for sophistication and flexibility of working software systems.[3] From a more fundamental point of view, interconversion between two codes allows for the evaluation and improvement of the numerous conventions adopted for each code. The early fragment codes and line notations were designed with a view to quick manual indexing, fast retrieval and short-term storage. These features have led to the development of numerous data-bases involving such codes as the Ring Code,[4] the IUPAC Notation for organic compounds,[5] or the Wiswesser Line Notation.[6]

---

(3) C.E.Granito, J.Chem.Doc.,13,72,1973.

(4) W.Steidle, Pharm.Ind.,19,88,1957.

(5) Rules for IUPAC Notation for Organic Compounds, issued by the Commission on Codification, Ciphering and Punched Card Techniques; Longmans,London,1961.

(6) The Wiswesser Line-Formula Chemical Notation, revised by E.Smith; McGraw-Hill,New York,1968.

However, neither fragment codes nor line notations can cope with all the above-mentioned requirements. Furthermore, only the codes which ensure no loss of information, i.e. topological codes, can fulfill those requirements. This justifies the use of a topological descriptor combined with a fragment type of approach in the GREMAS System.[7] In the DARC System[8] a unique topological code is being used to provide all the elements necessary for structure handling procedures. The Chemical Abstracts Service which is the largest supplier of chemical documentation has developed a topological code originally aimed at checking its CA index names. The CAS code now provides, in addition, a useful tool for international cooperation in literature indexing and a starting material for structure processing in the framework of users' systems. Thus, several retrieval systems have been developed, using the CAS Registry System. Systems have been implemented in the U.S.A. (National Institutes of Health at Bethesda, Md, and University of Georgia), in Switzerland (DCA - Basel) and in France (DARC System). Substructure search is mostly based on topological screening techniques such as the augmented atoms,[9] the TSS[10] and the FREL subcode.[11]

---

(7) R.Fugman,W.Braun,W.Vaupel,Nachr.für Dok.14,179,1963

(8) J.E.Dubois,H.Viellard,Bull.Soc.Chim.Fr.,839,1971.

(9) G.W.Adamson,J.Cowell,M.F.Lynch,W.G.Town,A.M.Yapp,J.Chem.Soc.C,3702,1971

(10)M.Milne,D.Lefkovitz,H.Hill,R.Powers,J.Chem.Doc.12,183,1972.

As has been previously stated, the ability to interconvert is an important feature for codes. The generation of fragment codes is feasible starting from topological codes or line notations (the opposite is untrue). Generation of the Ring Code and the GREMAS starting from the WLN or a topological input has been reported. Interconversion between WLN and topological codes such as the CAS or the DARC has also been developed.[3][17][18] However, economical interconversion with a topological code requires that the second code include the same amount of logic, i.e. be also topological. This is the case for the DARC code which is now being generated, on an experimental basis, from the CAS Registry Files on a large scale.

(5) Facets for Describing Data

A given piece of data generally refers to the magnitude of some quantity characterizing some property or phenomenon of a certain system, measured under a certain condition. This is illustrated in Table 2.

Accordingly, in order to describe data fully it is necessary to give an adequate description of each facet outlined above. It would be desirable for these descriptions to be standardized in machine-readable form so as to be useful for retrieval of data. However, since the items to be included in each facet for different disciplines are very different, uniformity of description covering all disciplines will be very hard to obtain.

Let us briefly discuss the description of "system". When the system is simply a pure chemical substance, the description reduces

TABLE 2 - FACETS FOR DESCRIBING DATA

Example 1	Boiling point of ethyl alcohol
System:	Chemical substance viz. ethyl alcohol, C_2H_5OH.
Property or Phenomenon:	Coexistence of liquid and gaseous phases in equilibrium.
Quantity:	Temperature.
Measurement Condition and Environment:	Under standard atmospheric pressure.
Example 2	Dissociation of hydrogen molecule by electron impact
System:	Hydrogen molecule H_2 and electron beam incident on the molecule.
Property or Phenomenon:	Dissociation of hydrogen molecule $H_2 \rightarrow H+H$.
Quantity:	Collision cross section effective to dissociation and directions of motion and kinetic energies of dissociated protons, as functions of energy of incident electrons.
Measurement Condition and Environment:	In vacuum of 10^{-5} Pa.
Example 3	Electron paramagnetic resonance absorption (EPR) by myoglobin
System:	Single crystal myoglobin from sperm whale prepared from aqueous solution with phosphate buffer, at pH=6.5.
Property or Phenomenon:	Absorption of 10GHz microwave, under applied static magnetic field of varying intensity.
Quantity:	Rate of absorption of microwave as function of crystal and of intensity of static magnetic field.
Measurement Condition and Environment:	Temperature 4.2K Humidity

TABLE 2 (CONTD.) - FACETS FOR DESCRIBING DATA

Example 4	Movement of the Earth's crust by earthquake
System:	A point on Earth's surface (latitude N x^{o} and longitude E y^{o}). Earthquake starts at Greenwich time oh om os, on day, month, year.
Property or Phenomenon:	Vibratory motion caused by earthquake.
Quantity:	(e.g.) Horizontal (NS and EW) components of acceleration of the point.
Measurement Condition and Environment:	The point is on soil made of volcanic ashes. No building within 50m.

to nomenclature and notation of chemical substances discussed in (4)(iii). But even in this case extremely accurate measurements may depend on isotopic composition, in which case a pure substance must be regarded as a mixture. A substance may be in the solid, liquid or gaseous state. In the solid and liquid states, it is often necessary to indicate whether the substance is crystalline or amorphous, since the glassy state and liquid crystals are extensively studied. The degree of order in the atomic arrangement of alloys, the degree and nature of imperfections in crystals and the amount of trace impurities must be included in descriptions of systems when data refer to structure-sensitive properties. For data concerning surface phenomena such as adsorption, heterogeneous catalysis, electron emission and contact potential, a description of the state of surfaces or interphase boundaries is required. In life sciences systems are usually related to some part (biomolecule, cell, organ, etc) of individual animals or plants of a certain species, as, for instance, myoglobin from sperm whale, membrane of giant axons of a squid. In earth sciences systems are primarily defined in terms of the location on the Earth. Even in petrology rocks must be defined by giving the geographical site and the geological structure of the site from which the rock came.

This brief examination reveals a very complicated and heterogeneous situation in describing "systems". Similar complications exist in descriptions of the other facets. The crucial point is to describe clearly each of those facets that is relevant to a given set of data.

(6) Classification and Indexing of Data

When the facet structure for describing data is taken into account, classification and indexing of data may, in principle, be reduced to consideration of each facet. In restricted areas such as property data of pure chemical substances this is expected to be feasible and further studies in this direction are very desirable.

Since, however, items needed for description of data are very different in each subject-area, it may be more practical at present to adopt less formal classification schemes developed in various disciplines. In this case, some aids may be needed for precise specification of data, for example, the addition of natural words, descriptors, keywords, etc. One merit of this is that data and original papers are classified by the same scheme in many cases. From the users' standpoint, it is desirable that classification schemes in use in different disciplines are made easily known.* Collection and referral services for classification schemes are offered by several organizations, such as the Special Libraries Association (SLA) in U.S.A., ASLIB in U.K., etc.

A more or less similar consideration applies to indexing of data. It is to be noted, however, that use of sets of descriptors or keywords for characterizing scientific papers may not be an effective means for identifying data. Indexes based on facets such as substance, property, phenomenon, quantity may be useful in some subject areas. Further studies on indexing of data should be made by CODATA.

---

* An example of a classification scheme for physics and astronomy is described in Physics Today, July 1973, p.64.

E. USE OF COMPUTER AND TELECOMMUNICATION TECHNOLOGIES IN DATA SERVICES

(1) Present Situation

It has been widely recognized that an electronic digital computer, if it is properly used, can be a powerful tool in data handling. High-speed processing is one of its most important features which will enable us to handle a large volume of data, sometimes even exhaustively, within a practicable short time. The short turn-around time realized by computers will stimulate the scientists' eagerness for the utilization of data. Using computers, we will also be able to expect the availability of fine-grained services which have as yet been unrealized owing to the prohibitive volume of manpower needed.

One need hardly state that numerical calculations, sometimes very complicated, are processed quickly by computers. Calculations of most probable values and probable errors, derivation of values from related observed data, interpolation of parameter-dependent data in tabulations and determination of crystal structures from X-ray diffraction data are examples of this application. These are, however, not the only applications of computers in data handling. Generation, storage and retrieval of data, and the associated processing capability are the principal functions which we can expect to use in the computer processing of data, and especially in the accessing and dissemination of data.

Together with computer techniques, data transmission is also of concern to us. It effectively shortens the geographical distances between the locations where data are generated, processed and/or used. In addition to private or leased circuits, conventional telephone lines can also be used for this purpose. In some advanced countries high-speed transmission lines and even satellites are

available.

The CODATA Task Group on Computer Use is responsible for this subject and much can be expected from its activities. Its state-of-arts reports have been published in CODATA Bulletins, No. 1 and No.4. Four oral reports, programmed by the Task Group, were given in Session VI of the Third CODATA International Conference held at Le Creusot, France in June 1972. A symposium entitled "Man-Machine Communication for Scientific Data Handling" was held, under the auspices of this Task Group, at Freiburg im Breisgau, Fed.Rep. of Germany, in July 1973, in which about one hundred scientists participated.

(2) Estimation of Developments in the Near Future

Among recent developments in computer technology, the time-sharing system, supported by the data transmission technique, is the most important, enabling us to use large computers of high performance for a relatively low cost. Furthermore, we can share a voluminous file which stores data and other information for science and technology. When this system is available, the installation of a smaller and less capable computer at each academic institution may be less convenient and less efficient.

The time-sharing system is a promising tool for the access of data and before long it may be possible for us to have a world-wide centralized data retrieval system using the time-shared processor unit with multi-terminals installed in a number of user countries. A data retrieval experiment through satellite telecommunication, demonstrated successfully at the Third CODATA Conference held at Le Creusot, France, is a milestone toward this objective.

Another important problem to which we must pay attention is the development of peripherals, input devices in particular, because they can give rise to the bottlenecks in operational computer systems. Wide use of character readers with the versatility of human eyes will

hardly be realized in the near future and, as a result, we will have to depend on some less convenient alternatives. Another difficulty is encountered when dealing with some types of data, such as chemical formulae, which are not described by regular alphanumeric characters. Practial solutions of these problems, including computer graphics using cathode-ray tube displays, are of great concern when computers are used for data handling.

The capacity and speed of memories are also of particular interest in data handling. Although, in practice, there are mutual conflicts among a larger capacity, a shorter access time, a lower cost and smaller physical dimensions, the development of a "data-oriented computer" which optimizes these factors should be encouraged.

Financial problems are always the primary factor hindering the use of computers and economic evaluation is necessary when a computer system is to be employed. Sufficient income can usually not be expected when a computer system is used solely for data handling. As a result, some other profitable applications should be planned for the sharing of the system, unless the academic performance of the system is recognized to be worthy of the investment and expenditure.

In some of the developing countries computerization of data handling may still be premature because the basic usage of computers is not yet sufficiently widespread. In such countries, the general level of computerization should be raised before planning to apply computers to data handling.

F. ACCESS TO DATA BY TRADITIONAL CHANNELS

(1) Personal Contacts with Professional Colleagues

Scientists and engineers never work in a position of complete isolation, but they are usually in close contact with others in the same subject fields or related areas, within or outside their own institutions. They obtain information and knowledge through direct conversations on the occasions of lectures, conferences or meetings of learned societies, by exchange of letters and use of telephones. The group of research workers, often called the 'invisible college', plays an important role not only in various phases of research progress, but also in the exchange of information.

In order to obtain the data they need, scientists usually try to communicate not only with colleagues of their own acquaintance, but also with other persons or institutions in the same areas, directly by visits, telephone calls or letters. However, it frequently occurs that in asking for the particular data they experience difficulty in directing their inquiries to the correct persons or places. In certain countries, listings of the titles of research projects of scientists and scientific research institutions are prepared and distributed, and referral services exist to direct inquiries to the appropriate sources. It is very desirable that such inventory and referral services for access to data are made widely available.

(2) Primary Publications

A very common method of searching for data is through papers or other reports appearing in regular primary journals. This is a significant channel in the sense that scientists can discover how the data were measured and can judge their reliabilities or errors.

However, difficulties exist in how to search for the paper containing the required data. Abstracts and indexes of papers are usually prepared without much attention to the data content. Since there has been no other effective tool, it is now a common practice that one consults cumulative indexes to primary or secondary journals, searching for papers likely to contain the required data and then scans them one by one to identify the data.

To help users obtain data from primary papers, it is highly desirable that abstracts clearly indicate what kinds of new data are contained in the respective papers, and that indexes are compiled with due attention to the provision of means of data retrieval. A new data-oriented abstract, called a data descriptive record, is used in two journals, viz. Nuclear Physics and the Physical Review C.

Another system, data flagging, proposed by a committee of IUPAC, is now being examined by an ICSU/AB - CODATA joint working group. In this system, to each abstract is attached short symbols (flags) indicating the kinds of new data contained in the paper. For example, AD, UVS might signify that the paper contains new data concerning adsorption and ultraviolet spectra. Flagging may be done by authors or editors in primary journals, or by abstract services if this additional work is economically feasible.

(3) Handbooks

The publication of handbooks is very helpful for data dissemination, provided that the scope of subject-oriented or mission-oriented coverage is clearly stated and that the users' level (such as for specialists, for teachers or for students) is taken into consideration in their editing. The sources of data in every table in these handbooks must be clearly described, as they are frequently edited by partial or abbreviated reproduction

of original compilations. A handbook for research workers must be edited so that users can obtain more detailed information on the data, as necessary, from the reported sources. For this purpose, it must be clearly indicated whether a table in the handbook is the complete reproduction of the one in the source or a reproduction with abbreviation of some kind. When a table in the handbook itself is an original compilation, this should be clearly indicated and the data sources should also be given. In presenting numeric values of data, it is desirable to indicate the probable errors or significant digits wherever possible. A table of data should be so arranged that it can be used without the barrier of natural languages; it is desirable that explanatory notes be provided in other major languages. In the CODATA Compendium it would be useful to list the names of handbooks which are above a certain quality level (see Recommendation 13 of chapter I).

(4) Data Compilations

Compilations published in book-form, or distributed in cards and magnetic tapes, are major output products from the data collection and evaluation activities of 'data centres'. There have been published standard data compilations covering broad subject-areas, such as Landolt-Börnstein Tabellen and International Critical Tables, but in recent years many compilations are issued by various data centres with comparatively limited subject-area coverage. A data compilation in a very specific area can appear as a paper in a primary journal, a review journal or a journal solely devoted to data. In these circumstances, users have difficulty in finding out what kinds of data compilations are published in each subject-area.

The publication of the International Compendium of Numerical Data Projects was the first task of CODATA and makes a meaningful contribution in giving a clear view of the world-wide situation with

respect to data evaluation centres, including centres for the generation of critically evaluted data. It lists data evaluation projects which are continuous in the fields of physics and chemistry, together with their output products.

However, the Compendium could, from the users' viewpoint, be improved in several ways. In accord with the extension of the subject scope of CODATA, the Compendium should cover data projects in life sciences and earth sciences, and the feasibility of publishing separate volumes for each subject field should also be examined. There has been no revision since its first publication in 1969. It should be revised at regular intervals from four to eight years and between revisions it should be updated by supplements which announce changes in existing projects and newly-started projects. Due attention should be paid to adequate irdexing in future editions.

The Compendium was originally designed to be a book of common use in every laboratory throughout the world but, in practice, it is not widely distributed. The publication of cheap paperback editions would help to increase the circulation among individual research workers. For libraries, information centres and data dissemination services, distribution in cards, loose-leaf forms, microforms, etc. should be considered. A subscription scheme whereby supplements are provided would be desirable. The CODATA Central Office is responsible for the editorial work of the Compendium and the CODATA National Committee in each country is requested to cooperate more actively in the acquisition of source information.

In addition to the Compendium, some channels are necessary to convey to research workers rapid and extensive information on activities relating to both existing and newly-started data

evaluation centres with details of their output products and means of access. The CODATA Newsletter carries information on newly-started data centre activities but often it is not available to those who really need it. It is desirable that journals, Union circulars distributed to national members and other society publications should publish, in agreement with data evaluation centres, announcements or articles giving this kind of information in the fields of their respective interests. Journals designed for data problems may also be useful media for providing this kind of information. Further, the organization of data referral services, described in G(6), is expected to provide an efficient method of helping users locate the necessary data.

It is desirable that the significant data compilations are available in the near vicinity of research workers. Each country should have at least one comprehensive collection of data compilations, in university libraries, special libraries or the so-called national library. To utilize data compilations stored on magnetic tapes, bilateral agreements for searching the magnetic tapes must be established between user organizations and the data centres which have prepared the magnetic tape files. This topic is related to problems concerning the anticipated functions of data dissemination services and will be discussed in Chapter G.

In the present (and future) situation where data are compiled by many data evaluation centres and organizations for critical data generation, it is suggested, from the users' standpoint, that a tool be developed for data retrieval from several data compilations in the same or related disciplines. For example, although compilations concerning thermodynamical properties and optical spectra of various chemical substances have been published by many centres, a comprehensive index for access to the compilation in which the data

on a given property of a given substance will be found would be very useful. It would be unrealistic to compile a substance index which includes all kinds of compounds because the number of compounds is more than 10^6, even if limited to pure materials. However, it would be possible, if the necessary human resources and funds are provided, to prepare an index for 10^3 to 10^4 commonly used compounds indicating the compilations in which any of twenty kinds of properties for each compound are given. It would also be possible to list similar compounds in one entry. The publication form could be either a book, a magnetic tape which will be used in data dissemination services, etc. The feasibility of providing such a tool for data retrieval within the frame of UNISIST would be worthy of consideration.

Two typical data search problems are as follows: (i) the user wishes to know the value of the data concerning a specified substance (or other object) and (ii) the user wishes to know the name of the substance and the condition which gives the known value or graph of the data. For the latter case special tools would be required, examples of which are: tables in which the melting and boiling points of many compounds at atmospheric pressure are listed in order of temperature, and tables in which atoms and ions are listed in order of the wave lengths of the absorption lines of their arc and spark spectra. It would also be possible to make a similar index for the identification of alloys and industrial materials from their property data. The necessity for more sophisticated tools for the identification of compounds should be studied by analytical chemists.

(5) General Scientific Information Centres

A large number of general information services for science and technology have been developed in recent years. Many offer flexibility, up-to-date coverage of the current literature and convenience in identification of the titles of papers and books most relevant to a user's interest. The majority of these services do not make any provision for handling or specifically identifying quantitative data. However, there are some specialized subject-areas (toxicology, drug information and some aspects of environmental science, for example) where general information centres and documentation centres do give particular attention to data collection and data dissemination. Since general scientific information centres have the necessary organization for the dissemination of bibliographic information it may be appropriate for some of them to handle, in addition, scientific and engineering data. This arrangement would certainly be useful until data dissemination centres are established and function effectively. To this end national governments should make the necessary financial provisions so that general scientific information centres can cope with this additional service.

G. DATA EVALUATION, DISSEMINATION AND REFERRAL SERVICES

(1) Historical Background

As already described, the principal data compilations were extremely limited in number in and before the first half of this century, and each of them had a fairly wide coverage. It was, therefore, common for users to obtain the data they wanted using one or several of these data compilations. In the third quarter of this century, however, the amount of data in science and technology grew rapidly and, as the techniques for data evaluation became more sophisticated, time-consuming and expensive, these data compilations with very wide coverage could not be maintained. As a result, data evaluation centres which cover narrow disciplines were gradually established and, moreover, there appeared several organizations which measure and generate critical data systematically for the disciplines in which data needs were high. Thus data compilations became more numerous and varied. National coordinating and administrative offices were established in several countries and, with this background, CODATA was established. The urgent business for CODATA was to take the necessary measures to help those various data projects provide a better service to the scientific and technological communities and, in particular, to enable users to find out the present status of those projects. The International Compendium of Numerical Data Projects was published for this latter purpose. In the early days, CODATA considered that a user could obtain the data he needed if he could identify the relevant data evaluation centre or data project, write a letter and send some money as necessary. In the plan of the World Centres of Numerial Data for Science and Technology presented in 1969 by Prof.F.D.Rossini, First President of CODATA, the CODATA Central Office was expected to serve as a referral centre or clearinghouse. This plan may

be summarized as follows:

(a) There should be only one data evaluation centre in the world for each narrow discipline in science. In the event that more than one data evaluation centre of equal competence would appear, they should coordinate and share their responsibilities to avoid unnecessary duplication in data activities.

(b) A user should be able to discover which data evaluation centre will provide him with the data he needs by asking the CODATA Central Office. Payment for the output will be covered by the money which he sends to the Central Office. The Central Office will switch the user's request to the appropriate data evaluation centre, which will directly mail the reply to the user.

Although this plan of Rossini, (b) in particular, did not come into existence for fear that too much burden might be placed upon the CODATA Central Office, it should be recognized as a first step in the design of data dissemination services.

In making plans for better data dissemination, it must be recognized that information services of the type needed are expensive, that they take much time to develop the ability to provide for user's need and, most important, that users take much time to become aware of new services. Because of all three of these factors, it is essential to make full use of existing data centres, expanding their capability to serve users whenever feasible, rather than starting new services.

In the text which follows, three aspects of data accesibility and dissemination are described as separate functions. The separation is intended to aid in the description and understanding

of the problem. It is not intended to suggest that the centres which perform the functions must be separated physically or intellectually from one another. Rather, the Task Group sees, for the last quarter of this century, a system of data dissemination services based on present centres, with expansion and greater support for some, and with augmentation through the establishment of new centres where needed. The latter action appears to be most likely in order to bring better user-oriented data service to countries and regions where data dissemination is now especially handicapped.

(2) Categories of Data Services

From the user's point of view, there are three functions needed to provide accessibility to scientific and technical data:

data evaluation and compilation service

data dissemination service

data referral service

For each of these functions, an organization must exist which has appropriate technical competence and adequate financial and manpower resources. In some cases, the functions will be performed by centres especially established to do that specific task. Thus there will be, as components of the system:

data evaluation centres,

data dissemination centres, and at least one

data referral centre.

This does not necessarily mean, however, the establishment of three new organizations for each scientific or technical subject. Rather, it will be advantageous to utilize and expand all present competences. In some cases, it may be most efficient to have both dissemination and evaluation done by the same group. Wherever such a pattern has been successfully established, the system should

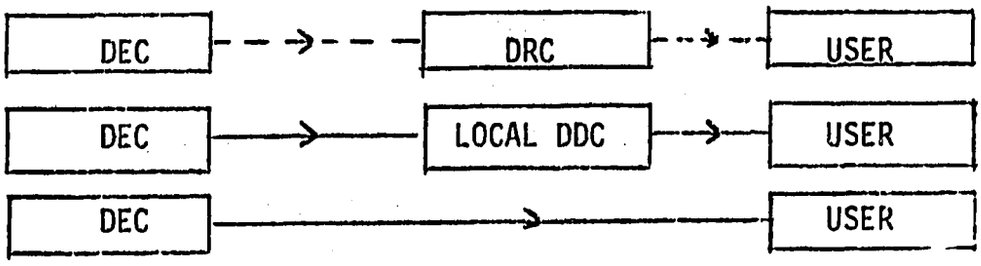
utilize it fully. In other cases, it will be most efficient to set up separate dissemination centres to serve one country or one region. In general, whenever this latter action is taken, the separate dissemination centre should have responsibility for handling a broad range of scientific and technical subject matter, so that it can be of greatest usefulness to the community which it is intended to serve.

The important point is that the function of data dissemination service should be undertaken by appropriate organizations, either existing organizations or new ones. The organization which generates critical data systematically is included in the data evaluation centre for the convenience of discussion.

(3) Data Evaluation Centres

The basic function of a data evaluation centre (DEC) is to compile and evaluate data in a specialized area and there is normally only one (or a very few) data evaluation centre in the world for each subject-area. Such centres can thus provide data of high quality to the scientific and technological community. It is unlikely that one country, however large, would have data evaluation centres for all scientific fields within its territory.

The user should be able to get information regarding the activities of a data evaluation centre from the global data referral centre (DRC). The output of the data evaluation centre will be available to the user through a local data dissemination centre (DDC) or directly from the evaluation centre. This latter option will be discussed in (5).



- - - - → represents the flow of information about data sources, services etc.

——→ represents the flow of data themselves

Figure 1

(4) Local Data Dissemination Centres

In contrast to the data evaluation centre, a data dissemination centre need not be specialized in a narrow scientific field but should cover a wide area, such as chemistry in general. In principle, every country should have a data dissemination centre so that a user can easily communicate with it, in his own language, by domestic telephone calls and visits, if necessary. It may, however, be appropriate to provide a data dissemination centre for a region of several countries or for part of a country, depending on the scale and characteristics of the users. We refer to such a centre, whether national or regional, as a local data dissemination centre to distinguish it from the global data dissemination centre which is more specialized and which will be discussed in (5).

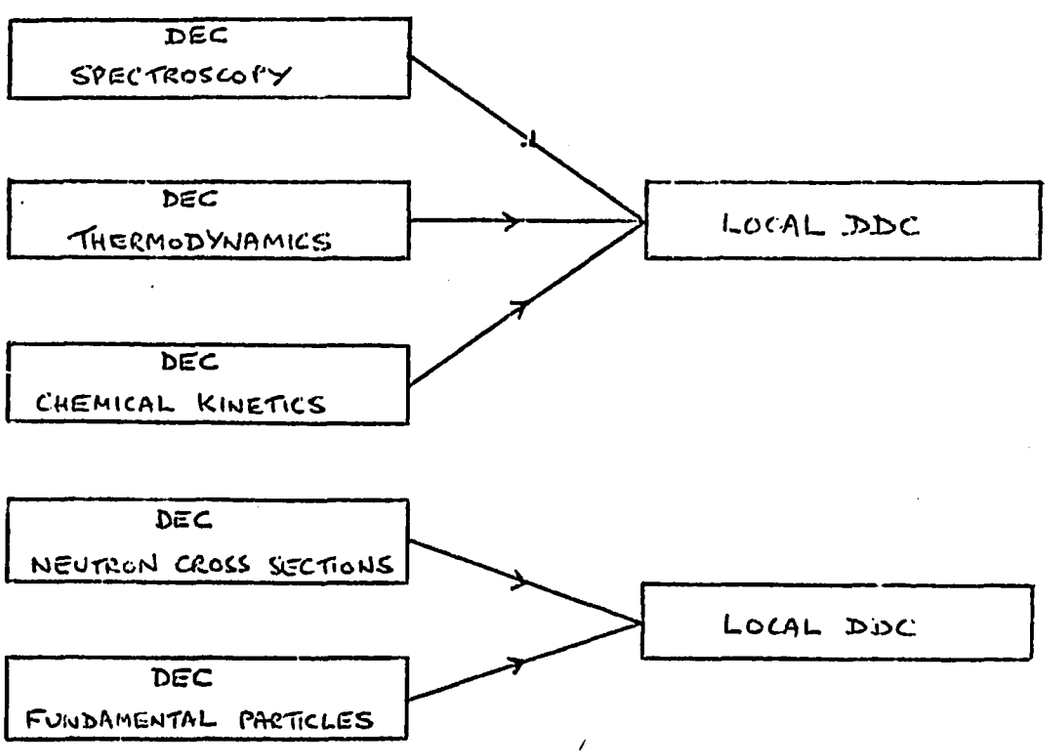


Figure 2

A local data dissemination centre, as an organization, may be attached to a university, research institute, data evaluation centre or a scientific literature information centre. In the latter case it must be borne in mind that personnel well-educated and trained in a scientific discipline are indispensable to the operation of data dissemination centre.

A local data dissemination centre should offer some or all of the following dissemination services:

(a) to collect published compilations of evaluated data and to offer them for reference use (library-type service);

(b) to collect and store other media, such as magnetic tapes, carrying data and provide data from them to users on request;

(c) to assist users in identifying where to find required data;

(d) to search particular data on request from collections of data centres and other data dissemination centres, in and out of the country, by mail or telecommunication;

(e) to provide SDI services and other services for current awareness, regarding data concerning specified subject, at reasonable cost;

(f) to carry out retrospective search of data on users' request;

(g) to issue secondary publications on data.

In this list (c) represents a data referral service i.e. a local data dissemination centre may provide data referral services to local users, thus complementing the function of the global data referral centre.

The services of a data dissemination centre may be expanded to cover activities designed to help the user obtain data from the primary literature and data depositories. Thus, when a system of data descriptive records or flagging becomes widely used in primary papers or abstracts, data dissemination centres may offer SDI and other services based on these records or flags (See Recommendation 17 of Chapter I). These data have not yet been evaluated by data evaluation centres but a qualified user can reasonably judge the character of the data, including the measuring method described in the paper, and such data are often very useful.

A local data dissemination centre is advised to instal terminals in the districts or institutions having high demands for data use. Communication between terminals and the centre is possible in various ways according to the progress of data transmission techniques: such as by using a telex system or by using, on a time-sharing basis, computers of the centre which may be equipped with graphical display units, etc. If on-line communication could be effectively realized, terminals could be linked with a data evaluation centre in a remote location, and the need for having a data dissemination centre in close proximity might decrease to some extent. In reality, the local data dissemination centre seems to be needed for economic reasons and as a constituent unit of a world network of data dissemination service.

The establishment and smooth operation of a local data dissemination centre will make a great contribution to the advancement of science and technology in the country or region. In recognition of this, every national government should make the utmost efforts for its realization and encouragement.

(5) Global Data Dissemination Centres

As indicated in (3), it may be helpful in certain cases for the user to obtain data directly from a data evaluation centre. This implies that the data evaluation centre possesses a dissemination mechanism. The functions of the data evaluation centre have thereby been extended and it is convenient to describe such a centre as a global data dissemination centre. In certain circumstances, however, a new organization for global dissemination is established, separate from the existing data evaluation centre. The important distinction between global and local data dissemination centres is that the global centre is more highly specialized in subject matter and, for that

speciazliced subject, seeks to respond to the world community of users. Thus several global centres will be linked to each local centre in the network.

The functions of a global data dissemination centre are:

(a) To store all output products (compilations, magnetic tapes, etc.) of a data evaluation centre in the respective field

(b) To comprehensively collect and store other significant and qualified data

(c) To disseminate data upon users' request

(d) To constantly exchange necessary information through the network among the local data dissemination centres in related subject fields

(e) To provide necessary information to a data referral centre.

A global data dissemination centre should be equipped with a large-memory computer, store data in a computer file adopting appropriate data classification and indexing schemes, and be prepared to retrieve data according to the variety of users' needs by automated searching of the files. The communication between a global data dissemination centre and a local data dissemination centre may be made at three levels, according to frequency: eventually by computer-based on-line telecommunication, more likely in the immediate future by telex system or by postal service. When telecommunication circuits, including satellites, are to be employed in the network of data dissemination centres, it is desirable that the countries concerned should financially assist the effective utilization of the network by taking necessary measures such as defraying the cost of the trunk lines. A global data dissemination centre is also expected to provide local data dissemination centres with services for the distribution, on loan, of output products of a data evaluation centre, as well as various

compilations issued by a data dissemination centre itself. In this case, an agreement should be made concerning legal and economic restrictions, etc.

In addition to cooperation among data dissemination centres, global and local data dissemination centres should set up channels for exchange of information with the data referral centre. It is essential for a global data dissemination centre in a specialized field to be closely linked with a data evaluation centre in that field, and local dissemination centres should also be linked with data evaluation centres in the related fields, either directly or through global data dissemination centres.

(6) Global Data Referral Centre

It is desirable that a single data referral centre is established which serves users throughout the world. It might be realised, for example, as a programme of UNISIST. The functions of a data referral centre would include the following:-

(a) to collect, on a world-wide basis, information on data resources relating to data generation, evaluation and compilation;

(b) to prepare a comprehensive file on the kinds of data available from these resources with a detailed subject index for data access (inventory service);

(c) to guide users to the appropriate resources where they may find the required data (referral service).

Functions (a) and (b) constitute the essential activities involved in the updating of the CODATA Compendium.

Appropriate information should be supplied from the global data referral centre to data dissemination centres in various countries so that the latter can provide referral services for the disciplines in their charge.

The functional relationships provided by the various organizations described in this chapter are summarized in Figure 3.

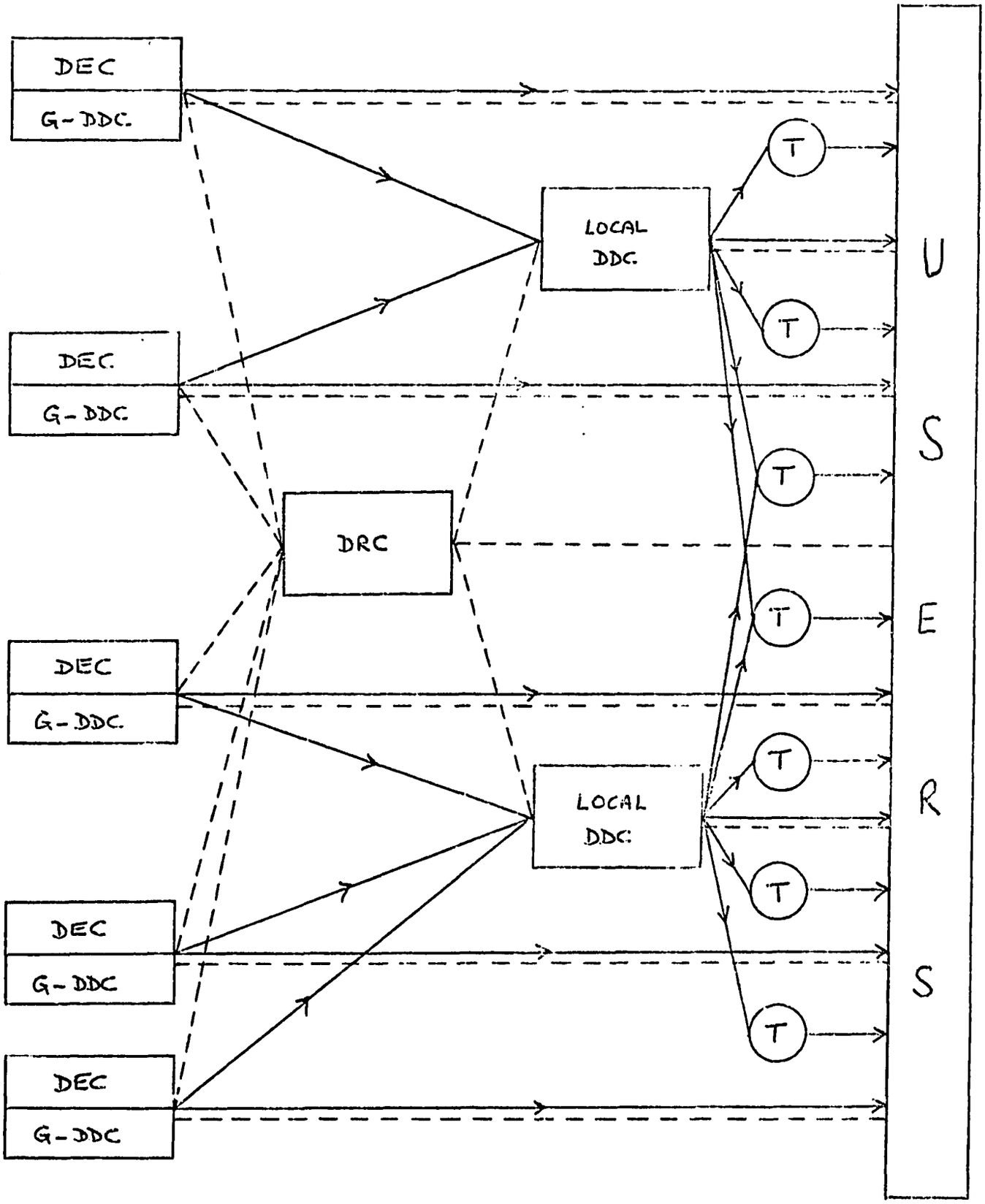

DEC
G-DDC.
DEC.
G-DDC
DEC
G-DDC.
DEC
G-DDC
DEC
G-DDC
LOCAL
DDC.
LOCAL
DDC:
DRC
T
T
T
T
T
T
U
S
E
R
S

Figure 3

H. PREREQUISITES FOR DESIGN AND IMPLEMENTATION OF DATA DISSEMINATION SERVICES, WITH SPECIAL REFERENCE TO DEVELOPING COUNTRIES

The Task Group has not been able, within the time allotted under the present contract, to make a detailed study of the problems associated with the design of data dissemination services. However, in this chapter, some of these problems are identified and preliminary comments made. It is hoped that the Task Group may be able, in the future, to study the problems in greater depth.

(1) Size and Variety of User Groups in a Country or in a Region

(i) Importance of surveys on user needs

An estimation of the number and variety of data users should be made as a basic requirement in designing a data dissemination service. This estimation is generally very difficult. The demand for data depends on the accessibility of data including procedures and costs, and it must be taken into account that an unexpected demand will increase when data become easily accessible. Surveys on the demand for computer user usually indicate very low demand when computers are not yet easily available to users, and an increasing demand as they become more available. The same situation might occur in surveys on data needs.

Attention should be paid to the fact that the needs for obtaining data are varied with respect to urgency, quality, quantity, etc., depending on the group of users and the kind of data concerned.

Basic data are indispensable in designing governmental projects for research and development, and particular data are urgently needed by research workers in their research planning and by engineers in their engineering design projects. In these cases it may well be that a considerable sum of money can be expended in acquiring the necessary data. On the other hand, laymen's needs of data in

relation to everyday life and health, and teachers' needs for data concerning teaching materals are, in general, not so serious. In these cases, the number of users may be very large but they cannot be expected to pay large sums to obtain the data.

In Chapter D, data are categorized, from the user's viewpoint, in three groups, viz. g_1, g_2 and g_3. The needs of users for each group of data will be discussed here:

(ii) User needs for data of category g_1

Users of data g_1 are mainly research workers in universities and research institutions. Data of this group are generated and evaluated in a specific subject discipline to which the users belong, so users are usually familiar with sources, evaluation centres, compilations etc. Data g_1 are very varied in nature and subject knowledge at a high level is required for their handling. It is true, in principle, that the dissemination of data g_1 can be entrusted to the community of the subject discipline to which they belong. Therefore, in designing dissemination services for data g_1 it is necessary to investigate why the existing dissemination system within the subject-area is considered insufficient, as well as to simply survey the size of the user population.

Researches are freely conducted beyond the fixed frames of scientific disciplines and research workers are frequently interested in problems of other subject disciplines with which they have had little contact in the past. Accordingly, data g_1 which have been used within a single subject-area are also needed in other subject-areas. With the growth of needs of this kind, a new relation between two subject-areas becomes established and data g_1 will change to data g_2. For the progress of science due attention should be paid to the interdisciplinary need for data which occurs

sporadically at the early stage of changes in subject relationships. Some mechanisms should be prepared for access to data g_1 as needs arise in other subject-areas.

In developing countries the demand for data g_1 is not so large because there are not so many universities and research institutions. However, this means that those scientists who are engaged in advanced research are scattered within the country and they face difficulties in effectively joining the world scientific community and enjoying the merits of the so-called 'invisible college'. On the other hand, we often find a group of advanced research workers in a developing country, in the case where an institution for observation or research, which is significant from the world-wide view, is established with international cooperation. Especially in the fields of astronomy, geology, biology, etc., there are academic necessities for having observatories or research institutions in tropical regions and the southern hemisphere. Examples are the European Southern Observatory in Chile, the International Entomology Research Centre in East Africa, etc. Efforts should be made within each subject-area to eliminate the isolation of these research workers or research groups, and this point should also be taken into account in designing the data dissemination service.

(iii) User needs for data of categories g_2 and g_3

Data g_2 and g_3 present many problems from the viewpoint of dissemination services. These data are generated in one specialized discipline and are used in other disciplines which may belong to the basic sciences, applied sciences or technical fields. Users are scientific research workers in universities and research institutions, employees of the government or private enterprise who work for mission-oriented research and development and, in addition, medical practitioners, technical consultants, school teachers, etc.

Careful investigations are necessary to ascertain the quality and quantity of data demand according to the variety of users. In some countries, such surveys are now being conducted for limited subject disciplines. For instance, in West Germany a survey is now in progress on the data demand in chemistry and physics.

In developing countries, dissemination services should be designed especially for data g_2 and g_3 in the subject fields relating to:

(a) the industry particularly developed in the country or the region,

(b) the governmental projects on economic and social development of the country.

According to the natural resources and availability of suitable sites for plant building in the region, industries utilizing mineral resources like oil, uranium, copper, tin, nickel etc., plant resources like natural rubber, natural fibre, etc. and animal resources in relation to fisheries, dairy farming, have made progress in developing countries and research and information activities have also become active. This exemplifies case (a).

The following projects are possible examples of case (b):

- Development and utilization of water resources
- Afforestation of arid zones
- Desalting of sea water
- Generation of electricity by nuclear power
- Social development concerning transportation, telecommunication, health, welfare, education, etc.

For these projects, the needs for scientific and technical data mainly come from those working in research and development in governmental or industrial institutions, from planners or managers

of these projects and also from practitioners in applied science and technology.

(2) Procurement of Personnel Required in Data Centre Activities and Training of Users

(i) Introduction

Data are generated and evaluated by scientists, sometimes assisted by technicians for routine measurements. Data are utilized by scientists as the basic and indispensable elements for further research. From these simple facts it is clear that scientific communities should pay due attention to data problems and make the necessary efforts for facilitating the effective utilization of data of high quality. In many of the tasks involved in data activities participation of research scientists is indispensable; in fact data activities in the early days were almost completely borne by volunteer scientists.

However, as various data services emerge involving documentary, computing, scientific, administrative and clerical work, the need for recruiting data specialists of various backgrounds has become apparent.

Data services can be effectively and smoothly performed only by the close cooperation of data specialists of various backgrounds and scientists. The problem of procurement of such personnel required for data activities should be considered, within the framework of the overall scientific manpower policy, by national governments, and also by UNESCO, CODATA and other international organizations in science and information.

(ii) Participation of research scientists

Among the tasks involved, the evaluation of data is most dependent on the work of active scientists, or at least of those who have considerable experience of scientific research. Scientists'

participation is also required in the collection of significant data from primary papers and in editing the compilation of data. In addition to these tasks, the responsibility for making decisions regarding problems of standardization, form of required services, etc., should be shared by scientists, e.g. through their participation in appropriate committees or advisory bodies. Generally speaking, however, it may not be advisable for a research scientist to be engaged exclusively in data work on a permanent basis, because this is hardly compatible with his essential research activity. Rather, a scientist should take part in data activity on a part-time basis, or on a full-time basis for a limited period. By participating in such a manner scientists may not only contribute to data services but also derive benefits for their own research work, without prejudicing their ability for creative scientific research.

(iii) Training and education of "data specialists"

Various tasks in data activities, such as generation, identification, evaluation, compilation and retrieval of data, require both sufficiently high scientific knowledge and specialized competence in modern data handling techniques. There is a considerable need for capable "data specialists", professionals in various data activities, especially for those in mechanized or computerized data service systems. Some measures should be urgently taken for the procurement of personnel for data activities and for the training and education of data specialists.

People whose background is science, librarianship or information science should be encouraged to enter this profession, being given the pertinent supplementary training or education needed for execution of specialized tasks in data activities.

Data specialists are required to have a good basic understanding of the problems and tasks relating to data and data handling. On the

scientific side, a basic knowledge is required regarding the definitions and natures of quantities in different disciplines, possible theoretical relations among them, measurement and reduction methods, as well as the theory of errors, statistics, etc. On the information and documentation side, data specialists must be competent in data handling including computer use and in the use of existing data compilations and services of various data centres. A knowledge is also needed of units, symbols and terminology, together with related methods and practices of classification and indexing pertinent to data. Skills should be developed in assigning descriptive records and flags to data, in identifying and editing of data, etc.

Appropriate training courses should be designed for those who wish to be data specialists. Courses for scientists may be different from those for librarians and information scientists but some others may be common to both categories. Besides training courses, much may be learned from actual experience of data service.

Research scientists usually tend to be very specialized in certain subject-areas and to have little concern with other areas. However data specialists must have an interest in and a knowledge of broader scientific fields. For this and other reasons, supplementary scientific training may also be required for those who have been educated in one discipline of science and wish to be data specialists. To encourage competent people to enter this career, a wider recognition of the importance of data activities and of giving higher status to data specialists are indispensable. This applies particularly to those who have graduated from science courses.

At present, there is, in every country, a shortage of teachers and lecturers for such training and education. Appropriate manuals and textbooks for these courses should be compiled to provide guide-

lines for instruction. It is very desirable that specialized lectures, meetings and seminars be organized for the benefit of the teachers and lecturers.

In the long run, as is now familiar in librarianship and information science, the establishment of curricula for this vocational education, leading to degrees at science colleges and universities may be the most fundamental way of procuring data specialists. The curricula must be designed to include the basic courses described above. Higher degrees may be required for teachers and lecturers.

In addition to training and education, research and development activities on data handling techniques should be promoted. Research by data specialists on easier and more efficient ways of description, classification, indexing and flagging of data, design of more efficient data handling systems, etc., should be encouraged. Many of these problems are related to standardization. In fact, the interchange of information on data among different data centres in a data dissemination network can be effectively made only when the descriptions of various elements of data are compatible, or at least convertible. In the description of data units, symbols and terminology constitute the essential parts, so that efforts for achieving standardization on these matters should be strengthened. This is also relevant to the interface between services and users, since uncontrolled use of these terms may cause considerable misunderstanding and trouble.

(iv) Training and education of scientists in general as generators and users of data

Scientists should know how scientific and engineering data are to be presented, published, retrieved and disseminated so that, as authors, they may present their data well and also be efficient users of data. In order to make the best use of data and the existing data handling systems, scientists must become familiar with existing data

compilations and services and should have a good understanding of the basic features of data handling techniques. It is desirable that colleges and universities in science and engineering offer, as part of their regular scientific curricula, courses on data handling and data use, which may be either independent or combined with courses on scientific information in general. Scientific societies and associations should also be encouraged to hold series of lectures, meetings or training seminars on these subjects at appropriate intervals.

(v) Special problems in developing countries

In developing countries, data availability and dissemination are, in general, handicapped by their isolated locations, low productivity of data and for other reasons. At present, most scientists are obliged to search personally for the data they need and spend considerable time and effort in so doing.

On the other hand, in some developing countries the potential productivity of useful data is high, particularly in some fields of life and earth sciences, but such "reservoirs" of data are undeveloped. These data, if available, would be useful to their own countries as well as of value to scientists throughout the world. In such cases, it is desirable for these countries to launch programmes for systematic observation, collection and dissemination of the relevant data; by this scientific endeavour it is expected that these countries would establish closer contact with the world scientific community and receive the necessary impetus to activate local scientific research. In order to realize this, however, it is important to note that some appropriate financial and scientific assistance should be provided from UNESCO and/or other international or national organizations.

In any case, it is very important to improve accessibility and dissemination of data for scientists and other users of data in developing countries. Information about the global data referral

centre must be provided to groups of users in these regions. A next step would be to collect important data compilations in appropriate organizations such as information centres, science libraries, etc. With two or three personnel, who may be librarians with some basic training in scientific data, a modest data dissemination service based on data compilations may be started. Part-time assistance by scientists from nearby universities or research institutes would be very helpful. As needs grow, data specialists will have to be recruited and more advanced services provided, maintaining close contact with other data evaluation and dissemination centres.

In the design of data dissemination services in developing regions it is particularly important to estimate user needs, as described in (1), and also data productivity in the region. Surveys should be conducted to assess the present situation. The size and scope of data dissemination centres may be determined by criteria which, it is hoped, will be formulated by UNESCO in consultation with CODATA.

To realize this, appropriate assistance from developed nations, either individually or through international organizations, are indispensable. Assistance should be directed to help these countries start needed data services and gradually build up systems to educate and train necessary personnel in these countries. At least in the early stages, it may be necessary for developed countries to send instructors to and accept trainees from these countries. Sometimes it may be necessary, for a certain period, to operate the data service systems by personnel sent out from developed countries. In order to make such assistance most effective, UNESCO is urged to take the initiative, based on the results of necessary pilot studies and in consultation with UNDP and other international organizations and national governments, in launching a coordinated and harmonized

programme for the improvement of data accessibility and dissemination in developing areas of the world.

(3) Financial and Legal Problems

Our Task Group has been unable to make a detailed investigation of the financial and legal problems involved in the establishment and operation of a data dissemination centre. Various points can be identified for further study and exmination:

(i) Basic philosophy of financing

The expenses for the establishment and operation of a data dissemination centre can be defrayed from the following sources:

(a) Governmental subsidies

(b) Assistance from international organizations

(c) Assistance from foundations and other private organizations in the country

(d) Payments from users

(e) Payments and donations from groups having high demands for data (e.g. industrial community, etc.)

It is strongly hoped that the government should bear the basic financial burden of running a data dissemination centre in partial execution of the national science policy, because activities of a data dissemination centre greatly contribute to national interests in raising the scientific level of the nation and promoting economic and social development. For the same reason, foundations and organizations which are seriously concerned with the encouragement of science and the utilization of the results, are requested to give appropriate support to a data dissemination centre. To achieve this, the efforts of academic and research institutions and learned societies will be necessary.

It is necessary to obtain wide recognition of the cost of information, including data, since information requires, in general, costly intellectual work for its generation and also requires advanced

techniques in its processing and dissemination. On the other hand, the traditional philosophy, that scientific achievements should be made freely accessible to everyone, has lent enormous support to the progress and spread of science. It is improper that evaluated data be marketed at much higher prices than document information. It may be impossible for university research workers to use high-priced services for data. Just as nobody expects the price of scientific journals to cover the cost of the research from which the papers resulted, the prices for data dissemination services must consist largely of the direct cost and should not cover the cost for data acquisition and evaluation, the cost for input to data files, the initial investment cost for computers, etc. However, the data demand of industry can be considered as a means for gaining profits for these enterprises and extra charges or conditions may be levied on industry.

Government subsidies and service income may be very limited when a data dissemination centre is established in a developing country or a developing region. Direct or indirect support, such as training of personnel, advisory services, etc., from UNESCO together with other UN agencies and international organizations, will be needed.

UNESCO should urge governments of adhering countries and international organizations to recognize the necessity for worldwide implementation of data dissemination centres and, at the same time, should start an appropriate programme for this purpose within the frame of the UNISIST plan.

(ii) Legal problems

In the case where a local data dissemination centre is designed to offer services using the output products of global data dissemination centres or data evaluation centres, agreement should be made concerning the conditions governing these services. Arrangements should be agreed upon by both parties for the regular purchase of all documentary output products from global data dissemination centres, possibly at reduced prices, for providing magnetic tapes on a lease or purchase basis, for the regular exchange of information necessary for the service, for the conditions governing various service modes etc. One condition might be that a local data dissemination centre would input data generated in the region to the files of data evaluation centres. These matters are concerned with the policies of each centre and can not be discussed in generalised terms. It would be desirable, however, to suggest several model patterns by examining the nature of existing agreements.

Concerning the agreement on establishing a data dissemination centre in a region comprising two or more developing countries, a model pattern should be drawn up by UNESCO.

There will be problems related to the copyright of each country in designing data dissemination services, as well as document information services. Careful examination is necessary to determine whether it is permitted, under the copyright law, to provide users with photocopied parts of printed data compilations, magnetic tapes, disk packs, etc. To encourage the wide dissemination of scientific information, it is hoped to establish an international agreement on copyright practice to reduce unnecessary restrictions and, if necessary, to stimulate revisions of the existing copyright laws. (Refer to the UNISIST Recommendation 19).

I. RECOMMENDATIONS

(1) Governmental Support for Data Compilation and Evaluation

Evaluated data constitute the essence of scientific achievements and are indispensable for the further advancement of science and technology. Furthermore, there is an established tradition according to which scientific knowledge is made easily available to all scientific and technological user groups. In recognition of these principles national governments and agencies responsible for scientific research should encourage and extend financial support to internationally-qualified projects for the compilation and evaluation of scientific data. There should also be recognition of the national benefits to be gained by the development and utilization of internal data-evaluating competence.

While the profits obtained by marketing the outputs of data evaluation projects can be expected to defray some part of the production costs, the cost of the intellectual effort required for data evaluation activities and the basic investment in information-handling systems (input operations) should not be included in calculating output prices.

See: A(1), F(4), H(3)

(2) Governmental Support for Data Dissemination

Although scientific information services in general have made substantial growth and progress in recent years, there are very few services designed for dissemination of scientific and engineering data. Recognizing that easy access to numerical data of high quality is of vital importance in the development of science and engineering and consequently is essential to national interests, national governments are urged to establish or help establish a system of national data dissemination centres for those subject-areas in which users' needs are high in the country.

See: A(2), G(4)

(3) Manpower

Recognizing the importance of data compilation, evaluation and dissemination, CODATA, UNESCO, national governments and scientific communities should consider ways to encourage scientists and engineers to participate in various data activities. Further, means should be studied for providing recognition and specialized training for professional "data-specialists", people with good understanding of one or several fields of science together with advanced capability in modern data-handling techniques.

See: A(2), H(2)

(4) Organizational Structure for Data Services

A number of data service functions are needed to ensure that reliable data are available promptly and conveniently to the user. The necessary functions include:

- data evaluation and compilation
- data dissemination
- referral of the user to an appropriate source for the data he needs.

It is recommended, therefore, that CODATA, UNESCO, the various scientific Unions and the interested national governments recognize the nature of the three functions identified here, and that they each work (within the boundaries of their own competence) to develop an interconnecting structure of centres to better meet users' needs. The recommended structure would consist of:

- data evaluation centres
- data dissemination centres (local and global)
- a single referral centre for data sources.

Specific recommendations, concerned separately with the three types of centres, are given below.

See: G(2)

(5) Need for Additional Data Evaluation Centres

There are many important scientific and technical subject-areas for which no centre is at present performing data evaluation functions. It is recommended, therefore, that interested organizations, especially scientific bodies and national governments, sponsor new data evaluation centres working in these subject-areas.

See: G(3)

(6) Possible Role of Data Evaluation Centres in Global Data Dissemination

It is recommended that some data evaluation centres, working in specialized fields where there is a high demand for data, should be encouraged to evolve and expand into global data dissemination centres in the respective fields. Such an evolutionary process, it is noted, can occur only on the basis of global need, plus established high scientific competence in an existing centre, plus financial resources adequate to perform world-wide dissemination as well as evaluation services.

See: G(5)

(7) Geographical Distribution of Data Dissemination Centres

A data dissemination centre may cover a broad field of science or may specialize in a scientific discipline, in a mission-oriented field of study, etc. A single country may have many data dissemination centres in different specialized subject-areas, or in some cases serving assigned parts of the country. However, it may be advisable in some subject-areas to establish a data dissemination centre for a region comprising two or more countries, under the joint sponsorship of the governments concerned, possibly with the assistance of UNESCO.

See: G(5)

(8) International Collaboration in Data Dissemination Centres

Although the services of a data dissemination centre are to be provided mainly to users within the country, operations on the input side of a data dissemination centre are essentially international.

UNESCO, CODATA and international organizations in science and in scientific information should study the various problems involved in creating data dissemination centres, including funding and charges to the users, staffing and connecting centres in a global network, and give advice to governments and other appropriate bodies in order to promote the realization of data dissemination centres and their global network.

See: G(4), G(5)

(9) Temporary Dissemination Services by General Scientific Information Centres

Each national government should take appropriate steps to enable existing scientific information centres to handle scientific and engineering data additionally, at least temporarily, until data dissemination centres are established and function effectively.

See: A(2), F(5)

(10) Referral Service for Data Sources

CODATA should, in close cooperation with UNESCO, work toward the establishment of a central referral service on sources for scientific and technical data. The goal should be a service capable of providing guidance to users on the availability of data from data evaluation centres and data dissemination centres throughout the whole world. Requests for specific data would be handled either by sending the request on to the appropriate centre, or by instructing the requester to contact a particular centre himself. This service should be based (at the outset) on information gathered for the CODATA Compendium. National experience to date indicates that a service of this type should not charge users for its assistance. Accordingly, central or cooperative support would appear appropriate. UNESCO should consider the desirability of including this service in the UNISIST programme.

See: G(6)

(11) Revision and New Publication Forms of the CODATA Compendium.

Revision of the International Compendium of Numerical Data Projects should be completed promptly. The title "Directory" rather than "Compendium" is recommended, as providing better indication of the contents of this work. The revision should cover not only projects in physico-chemical data but also those concerned with data in life sciences, earth sciences and engineering sciences. The increased scope may require a multi-volume edition. A systematic and comprehensive subject index to the Directory should be prepared with the participation of subject-matter experts and taking due note of indexing studies currently being completed by several organizations.

CODATA should plan to revise the Directory at regular intervals, with the possibility of supplements between revisions. CODATA should also examine alternative forms of publication, e.g. cards, loose-leaf forms, microforms, etc. UNESCO, ICSU, WFEO, and others should call attention to this important source guide and promote its utilization.

See: C(2), F(4)

(12) News Announcements of Data Activities

CODATA and others concerned should inform data user communities of newly emerging data evaluation and compilation projects, as well as changes in the activities of existing projects, as widely and effectively as possible. The CODATA Newsletter can provide one channel to serve for this purpose. CODATA should also encourage the publication of such announcements by governmental agencies, professional societies and Unions.

See: C(2), F(4)

(13) Handy Compilations and Tables of Data

Handbooks providing collections of scientific and engineering data relevant to various fields are valuable sources of information. CODATA should include a list of handbooks in its Directory of Numerical Data Projects. The traditional participation of commercial publishers in the production of handbooks of reliable scientific data should be recognized and encouraged.

See: C(1), C(2), F(3)

(14) Standardization of Data Presentation

CODATA, in close cooperation with other interested international scientific, engineering and standards organizations, should promote standardization of tabular and graphical formats to be used for the uniform presentation, transmission and tabulation of data. The method and format of describing data should be specified for the various subject-areas and should take due cognizance of the report being prepared by the CODATA Task Group on Publication in the Primary Literature.

See: H(2)

(15) Standardization of Data Exchange Media

CODATA, in close cooperation with ISO, should stimulate and promote standardization of the recording media used for data exchange in machine-readable form, such as magnetic tapes, magnetic disk packs, etc. In addition to the physical characteristics of these media, formats and codes for recording information on the media, and other factors concerned with interchangeability should be standardized. It is advisable to adopt the relevant ISO standards in principle and to formulate supplementary standards as necessary.

See: H(2)

(16) Standardization of Computerized Bibliographic Services

ICSU/AB and CODATA should encourage abstracting services and data evaluation and compilation centres to develop compatible and interchangeable computerized services. The interchange could help ensure that both abstracting services and the centres attain their common goal of comprehensive coverage of the world literature containing data.

See: D(3), H(2)

(17) Data Descriptive Records Suitable for Primary Journals and for Abstracting Services

UNESCO, ICSU/AB, and CODATA should establish a few selected working groups of representatives of primary journals, data evaluation and compilation centres, and abstracting services in order to develop data descriptive records appropriate for various subdisciplines of science and engineering and suitable for use in primary journals and abstracting services.

The standardized records should provide (a) abbreviated methods for identifying, in the body of the abstract, new data contained in the paper; and (b) flags for abstracts of papers containing data by means of special keywords that modify the abstract and that are especially designed for computer searching. The minium goal of such an effort would be to identify every abstract of a report or article that contains data together with the designations of the physical or biological quantities involved in the data.

See: D(3), F(2)

(18) Pilot Study on Data Productivity

Pilot studies should be carried out on the actual production and potential productivity of useful scientific data in developing countries, taking into account the national needs and competences of the individual countries, and the local availability of unique opportunities for acquisition of data associated with natural phenomena (earthquakes, eclipses, etc.). When it becomes clear that collection, production and evaluation could be made of valuable data in certain fields of science by modest investment, UNESCO should encourage the government of the country concerned to launch such data projects, providing financial and technical assistance. In this connection, UN agencies such as FAO, WHO, WMO should be consulted, as well as ICSU and its constituent Unions. This will contribute to the promotion of participation by developing countries in scientific activities of global significance and to due recognition of the importance of scientific data by the governments and scientific communities of those countries.

See: H(1), H(2)

(19) Pilot Study on Data Needs

Pilot studies should be carried out on the needs of data existing in various circles in developing countries, in collaboration with scientists, engineers and governmental agencies of the countries concerned. The quantity and quality of data needs in the various subject disciplines should be estimated. Data needs related to development projects should be specially investigated in developing countries. The nature and magnitude of data needs will provide basic criteria for establishing data dissemination centres in these countries.

See: D(1), H(1)

(20) Data Dissemination Services

In principle, any scientist, irrespective of the geographical location of his working site, should have easy access to the data he needs. National governments and scientific organizations of developing countries should pay particular attention to the methods of facilitating access to scientific data for scientists and other persons who need those data.

The first step to be taken in this connection may be to collect important data compilations issued by many data centres in the world and to make them available for scientists and others. If the number of users of data in the country is limited, two or three staff members trained in library and information services with a general understanding of science may be appointed to provide limited data services to users. This may be regarded as the early stage of a data dissemination centre. As the number of users increases and more advanced services become necessary, staff members with further qualifications should be included and more sophisticated information handling systems may be introduced. Cooperation with other data dissemination centres, global, national, etc., should be established so that the centre may be linked to the global network for data dissemination.

CODATA, with the support of UNESCO, should formulate and publicize guidelines for the design of data dissemination services. The guidelines should include, among others, estimates of operating and manpower requirements for the setting up of data dissemination centres in various stages, and criteria by which an appropriate stage can be determined for a given country according to the size and nature of the expected user group, available resources, etc. These guidelines would, of course, be useful to developed as well as developing countries.

See: A(2), C(1), C(2), F(5), G(2), G(4), G(5), H(1), H(2), H(3)

(21) Financial and Technical Assistance

Recognizing that the growing gap in science and technology between developing countries and highly industrialized countries is a matter of grave concern for mankind from the viewpoint of future world stability, and considering that easy access to scientific data represents a basic element indispensable to the progress of science and technology, UNESCO, together with other UN agencies and ICSU, should extend appropriate financial and technical assistance to developing countries to encourage and facilitate their efforts toward establishing effective mechanisms of data dissemination in their countries.

See: H(2), H(3)

(22) General Booklet on Role of Data

UNESCO, in collaboration with CODATA, should publish a booklet in which the important role of scientific data in modern science, including its applications, and current activities in many countries concerning the generation, collection, evaluation and compilation of data are explained in ordinary words. Such a booklet should be distributed to national governments, scientific organizations, information centres, etc., in developing countries.

See: F(3)

(23) Data Problems as Agenda Items at UNISIST Meetings

During the meetings of the UNISIST Steering and Advisory Committees attention should be paid to the problems of scientific data evaluation and compilation, with special reference to access and dissemination in developing countries. Representatives of CODATA, scientific Unions and information services, as well as a few data centres, should be invited to participate in these discussions.

See: H(3)

(24) Visits by Experts

Individual experts or groups of experts should be sent to developing countries upon request to examine the actual situations and to advise governments or other organizations in the respective countries on the planning of data dissemination services. UNESCO should also send experts for certain periods to countries where data dissemination centres or related services have been newly established, to help with the administration and implementation of their projects upon request.

See: H(2)

(25) Training Courses

Assistance should be provided in the training of personnel required for data dissemination services. It is desirable to organize training courses in developed countries, inviting trainees from developing countries with fellowships offered by UNESCO or other appropriate organizations. It is further effective to hold training seminars at appropriate places in developing countries by sending expert instructors with UNESCO's assistance. Manuals for training should be prepared, by UNESCO or other appropriate organizations.

See: H(2)

(26) Involvement of Regional Documentation Centres

UNESCO should examine the feasibility of giving to national and regional documentation centres, established with UNESCO's help, responsibility for data dissemination services for the region accompanied by a modest increase in staff and facilities. If this is successfully done, the regional data dissemination centre may be considered as attached to the documentation centre of the same region. It is important, however, to pay due attention to the fact that data dissemination services and documentation services have different requirements concerning the background training of personnel, so that adequate supply of additional staff is necessary. Cooperation of some staff members of nearby universities will be very useful.

See: A(2), F(5)

(27) National and Regional Data Dissemination Centres

Although certain countries are officially designated as developing countries, nevertheless they may have a considerable tradition in science and engineering. Thus it may be appropriate to establish national data dissemination centres in these countries. On the other hand, if there is a smaller scientific activity it may be more effective to establish a data dissemination centre for a region comprising two or more countries.

UNESCO should study this problem in cooperation with national governments and they should jointly decide which of the two alternatives is more appropriate. In the case of regional data dissemination centres, UNESCO should promote their establishment and should act as mediator eg. by convening meetings of representatives of the countries concerned to reach multilateral agreements.

See: G(4)

(28) Legal and Currency Problems

UNESCO and CODATA should find ways and means of facilitating the access of developing countries to scientific data with particular reference to problems of copyright and currency. The latter could, for example, be tackled by extending the use of UNESCO coupons.

See: H(3)

(29) Adoption of Computerized Techniques

With the aim of connecting data dissemination centres in developing countries to a global data dissemination network, financial and administrative problems involved in the introduction of computerization and in the use of telecommunication and satellite communication should be studied in depth by UNESCO. Assistance should be extended by UNESCO, UNDP and other organizations in order to enable data dissemination centres in developing countries to improve their services step by step, using storage of data on various media such as magnetic tapes and modern telecommunication facilities. It may be reasonable to consider these problems together with the corresponding problems in handling bibliographic information in general.

See: E(2), G(5)