UCID-3792

APPENDIX I

STANDARDS FOR MULTILATERAL AND WORLDWIDE EXCHANGE OF GEOTHERMAL DATA

by

J. J. Herr, S. L. Phillips, S. R. Schwartz, and T. G. Trippe

Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

ABSTRACT

Standards for the exchange of machine-readable data are discussed at five levels covering (1) physical characteristics of the medium used in the exchange, (2) overall structure of the data file, (3) format of individual records, (4) types of information contained in a record, (5) authority files for information definition. The standards are used by the National Geothermal Information Resource for compilations of geothermal energy data.

Sponsored by the U. S. Energy Reserach and Development Administration.

DISCLAIMER

Introduction

The need to utilize domestic energy and mineral sources which are so essential to our health and economic well-being cannot be questioned. Based on the availability of energy sources, both developed and developing nations must modify plans for expanding or building industrial capacity, and improving transportation systems. A very delicate balance exists between acceptable energy sources and those with imposed conditions which stifle and impede rational progress. The need to obtain sound energy and mineral data is obvious.

Many developed and developing nations are resolved to become energy independent of foreign sources to the extent possible. A most promising new domestic energy source -- geothermal -- while not yet well understood will surely play a major role in the energy economy of the future. There exists a worldwide research and development activity in geothermal energy with the objective to stimulate and facilitate the early and significant utilization of geothermal energy for power production.

Just as important as research and development of geothermal energy is the collection and dissemination of current information about geothermal science and technology. It is the case in the geothermal field -- and understandably so because the field is undergoing so rapid a growth -- that there are many gaps in what is termed "public knowledge" about geothermal energy. New techniques have been developed for research that should be widely applied for exploration and utilization; basic scientific data may be applied to predictive modeling thereby saving much time and effort; some geothermal resources may not be made productive by any current technology; other resources may be utilized quickly without the necessity for further development work. To fill these information gaps is necessary and urgent; if the gaps are not filled, there is the penalty of unnecessary duplication and expense in research, and the additional expense, not often measurable in monetary terms, that needed energy sources cannot be utilized.

The Lawrence Berkeley Laboratory is sponsored by the Energy Research and Development Administration to establish a National Geothermal Information Resource (GRID). The objective of GRID is mainly to compile and disseminate evaluated data and bibliographic information on the following six major categories of geothermal science and technology: (1) physical chemistry; (2) exploration; (3) utilization; (4) institutional considerations; (5) environmental aspects; and (6) reservoir characteristics. See Figure 1.

GRID is involved in a number of cooperative agreements for the interchange of information and data with other organizations on a worldwide basis. This lead to an active interest in the question of standards for bibliographic, numerical, and other types of data (e.g., maps). Standards for the interchange of bibliographic data are currently well established, and the utilization of these standards is increasing. GRID's bibliographic file is computer-based and patterned after that of the International Atomic Energy Agency's International Nuclear Information System (INIS). Utilization of the INIS format and file definition ensures that GRID's bibliographic data will be compatible with and usable in other bibliographic systems. Standards for the interchange of data, in general, and bibliographic information, in particular, are discussed below. Appendix II is the GEODOC Reference Manual, which defines GRID's bibliographic data structure and illustrates this discussion.

Standards for the Interchange of Machine-Readable Data

In the exchange of machine-readable information, there are several levels at which standardization can be discussed. These are: (1) physical characteristics of the medium of exchange (usually, magnetic tape), including size of tape, recording mode, density, block (physical record) structure, interblock gaps, and file marks; (2) overall structure of the file, including character sets and header and trailer labels; (3) format (structure) of individual (logical) records, including the mechanisms(s) for identifiying particular units within each record and specifying their lengths and positions; (4) the types of information to be included with a record (i.e., the data fields or data elements); and (5) the exact definition of the type of information to be included in a data field, its form (or style), and authority files associated with it. Standards for levels 1 and 2 are reasonably well established within the computing community; the following covers levels 3 to 5.

Standards for record formats (or structures) have the advantage of reducing significantly the effort needed to "crack" a new data file and, in addition, can save time and effort when designing a file. Such a standard, established by consensus of suppliers and users of exchanged data, ensures completeness of the information required to decode a tape. The standard structure should permit the recording, without distortion, of all types of data whose inclusion can be anticipated. To the extent possible, such a structure should be hardware and software independent.

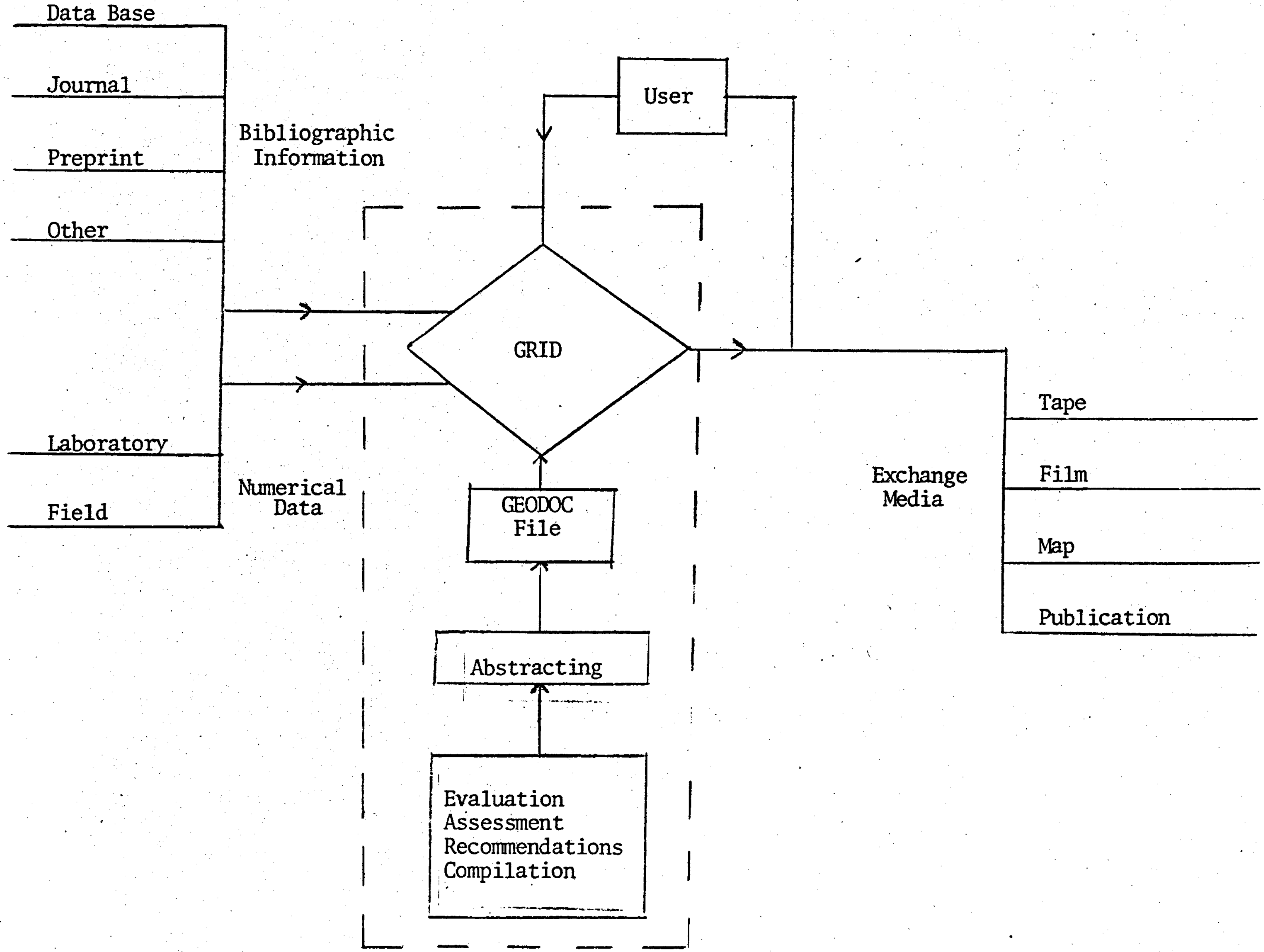

Figure 1. National Geothermal Information Resource (GRID)

Software includes operating systems, and, to an increasing degree, data management systems. Few data management systems are hospitable to all types of data; nearly every one has some restrictions, around which its users must work. Examples of restrictions imposed by data management systems are those on the number of data fields for a record, on the length of data fields, and on the number of occurrences of a particular data field or group of fields. It should be kept in mind that such a structure is for interchange only and says nothing about structures for internal use. The format in use in bibliographic data interchange may be briefly described as one in which a fixed format directory (specifying each type of data present, its length and position within the record) precedes a variable-length area (containing the actual data). Although the use of this structure for hierarchical or otherwise linked information is possible, the usual implementation is linear (i.e., with little or no linking of data elements). The principal barrier to implementing such a standard is the investment in local formats for already established data bases.

Establishing standards for the types of information to be carried in a record is even more difficult. Probably the most difficult aspect is anticipating the degree with which information should be delimited. As an example some bibliographic systems tend to include in one undelimited field all of the journal reference information while in others the journal name, volume, issue, page, and date are in separate fields. The higher the degree of separation of discrete units, the greater the flexibility in manipulation (including format checking) possible. In order to produce data that can be used by other groups, GRID has followed the INIS list of types of data. The two major stumbling blocks to adoption of a standard list of data elements are, again, the investments in existing systems, and the costs of processing a detailed structure, whose value may not be apparent initially.

Finally, perhaps the greatest difficulty lies in standardizing the contents of the data fields. These specify, for instance, the order of authors' names; abbreviations for journal names; forms of institutional names; and codes for recurring information, such as journal CODEN, corporate author identifiers, and country codes, which are a final check on the consistency of the entries. The major advantage of standardization in this area is ease of understanding and searching the data, which should be unambiguous. Probably the most difficult standardization problem in the bibliographic area is that of the means of describing the subjects of a document; this problem includes both the style of the subject description and the particular terms or categories used.

Summary

Standardization of data for interchange means that a product generated for one set of users can be used, without needless effort, by a wider community. Aspects subject to standardization include physical characteristics of the medium of exchange; overall structure of the file; structure of the individual records; types of information to be included within a record; and style of the information and authority files associated with it. Standards for the exchange of bibliographic data are well established and in active use throughout the world. GRID utilizes these standards for bibliographic work, thus enhancing the value of its product and promoting the active interchange of data with other groups. Similar geoscience data standards would take the bibliographic experience into account and would result in an even wider level of compatibility for the interchange of data.

Conclusion

After reviewing current standards for the exchange of machine-readable data, some general observations are appropriate. Present-day standards, although adequate in many respects, still have a number of shortcomings. The "ideal" system for standard data exchange should include:

1. Data elements developed by a consensus among users, cataloguers, and data centers. This would facilitate the collection, exchange and utilization of geothermal information.
2. Uniformity in magnetic tape formats. Much time and expense are devoted to cracking a tape from one data center so that the tape information can be utilized by another data center.
3. Delegation of responsibility to each participating data center for one or more aspects of the exchange program. For example, it is desirable to have one data center put all of one type of data into one acceptable format for dissemination to the other participants.
4. Continually updating of authority lists so that they are compatible with current usage. For example, many nations have changed their methods for mail addresses (e.g., ZIP codes), and these changes should be reflected early by a change in the authority list.

5. One set of commonly accepted units for reporting physical measurements. Conversion of many units to a common unit is a time-consuming procedure which also impedes evaluation of the data.

6. Development of tags or flags to indicate the level of quality of numerical data. Often, erroneous or unevaluated data are worse than no data at all.

7. A listing of data descriptors in a thesaurus of terms which will impart to the user an idea of the magnitude of the numerical data, the experimental procedure used to obtain the data, and the materials (e.g, chemicals) used in the experiment.

References

1. International Standards Organization (ISO), "Documentation - Format for Bibliographic Information Interchange on Magnetic Tape," ISO-2709-1973 (E).

2. UNISIST/ICSU-AB Working Group on Bibliographic Descriptions, "UNISIST Reference Manual for Machine-Readable Bibliographic Descriptions," SC.74WS/20, UNESCO, Paris, 1974.

3. International Atomic Energy Agency, "INIS: Magnetic Tape Specifications and Record Format," IAEA-INIS-9 (Rev. 0), Jan. 1970.

4. International Atomic Energy Agency, "INIS: Descriptive Cataloguing Rules," IAEA-INIS-1 (Rev. 2), Jun 1972.

5. J. G. Mulvihill, "GeoRef Developments," presented at the Tenth Annual Meeting of the Geoscience Information Society, Salt Lake City, 21 Oct 1975.