Prospective study for an ISPRS database

D. Burette, GDTA, France
Ch-H. Latarche, GDTA, France
D. Pouyllau, CEGET-CNRS, France
J.H. Ten Haken, ITC, the Netherlands

Commission VI, Invited paper, 16th ISPRS congress, Kyoto, Japan

0 - INTRODUCTION

The establishment of a photogrammetry and remote sensing database was the subject of a resolution taken by Commission VI at the Hamburg Congress in 1980. Slowly but surely, the idea has germinated. The different stages can be traced, from the Mainz symposium in 1982 with a significant paper presented by J.H Ten Haken titled "An investigation of available on-line databases in the field of photogrammetry and remote sensing", to the Rio congress in 1984 with the paper by Prof. J. Hothmer "The ISPRS-IRS information retrieval system for literature and factual data" (WG VI-4), of fundamental importance to the implementation of the project.

In parallel, C-H. Latarche presented an updated inventory of existing databases at the Rio congress in 1984, concluding that the bases were spread across a large number of international hosts. In 1986, the Badagry symposium confirmed the previous resolutions.

Today, our intentions cannot fail to be reinforced by the increasing development of remote sensing, particularly with its adoption of the new image processing technologies, and by the recent advances made in photogrammetry and the contribution to Geographic Information Systems (GIS).

In addition to the scattered locations of existing databases and the consequently increasing number of access procedures and languages, we are also witnessing a proliferation of different hardware, standards and formats due to the explosion of microcomputing and the appearance of a large number of new subject-specific databases.

In view of this background, we consider that it is now urgent for us to firmly assert our wish for joint action and coordination, with a view to defining a single set of characteristics for a unified and standardized database.

The ISPRS congress is the decision-maker, the entity responsible for the overall guidance of the project and the taking of decisions. The aim of the ISPRS-IRS database feasibility study is to provide the Kyoto congress with the political, technical and economic background by which the project itself can be justified, advice given regarding the choice between the various configurations (or approaches), and funding obtained. The working group, having given consideration to the timeliness of implementing the ISPRS-IRS database, has conducted a study into the existing resources based on previous work (mentioned above) and conducted a survey in February-March 1988 among the organizations potentially interested in the project. The results of the survey and of a questionnaire sent to 600 organizations were used to study users' information requirements and to attempt to estimate the potential user community. The working group had thus produced specific information on users' needs, and the broad outline of a proposed ISPRS-IRS database was beginning to take shape.

The feasibility study puts forward several possible configurations, each leading to project implementation based on a different network of producers. At the end of the complete study it should be possible to select a strategy. This would be followed by an ISPRS-IRS database design study, producing an executive guideline and a set of technical specifications. To help define this later stage, a simulation was done on 50 bibliographic records covering papers read at the *"SPOT-1, Utilisation des Images, Bilan, Résultats"* conference (Paris, November 1987). The implementation of the executive guideline should lead to the choice of hardware and human resources, the initialization of the management software, and the launching of the application, to be followed by tests and an assessment of the operational result. This stage will also include input of data to the chosen host.

1 - STUDY OF EXISTING RESOURCES

The work indicated above, slightly updated, revealed two categories of partners likely to be interested in the project: centers that produce computerized databases, and centers with non-computerized documentary resources. A survey concerning the existing resources and their potential was sent to 30 organizations who were invited to indicate their degree of interest in the project and to describe their resources. This frequently led to requests for further information, either by telephone or by visits. The questionnaire results have been analyzed, providing information on the documentary resources held, and on actual access to source documents, the existence of computerized databases and projects, the documentary tools used, and/or being designed, the services provided and the interest in participation in a future producers' network.

1.1 - Computerized databases

1.1.1 - Computerized databases in general

This category covers major producers, the majority of whom generate very large, multidisciplinary information systems. The RESORS database in Canada is the only file dealing exclusively with remote sensing and photogrammetry. It contains 60,000 items {articles (47%), papers read at symposia (26%), theses (15%) and manuals (11%)}, and 6,500 slides. Some 94% of the records are in English.

In view of the richness of RESORS, and the fact that the database is presently available on the main Canadian host only, close collaboration with the Canadian center to achieve greater dissemination seems essential.

The CDST-CNRS, the French producer of PASCAL, scans the majority of the main journals dealing with remote sensing. At the end of 1987, the file contained over 16,000 remote sensing records, mainly oriented toward earth sciences.

Mention must also be made of GEOBASE, the online GeoAbstracts file with over 8000 records in part G (Remote Sensing, Photogrammetry and Cartography) alone.

COMPENDEX, dealing largely with engineering, has 7800 remote sensing records.

The multidisciplinary NASA database, basically aerospace, contained over 7000 remote sensing records in 1987.

NTIS, also multidisciplinary, covering aeronautics and the earth sciences, contains over 7000 records.

INSPEC is oriented toward electronics and computers, but now has over 8000 remote sensing records.

1.1.2 - Specific databases

Specific databases each with several thousand remote sensing records include GEOREF (dedicated to geology and earth sciences), AGRIS (FAO) in agriculture, TULSA (US) in energy, ASFA in water resources and marine issues, and Water Resources Abstract (water and water resources). Another group of databases each contain under 3000 records; these include GEOLINE (geology, FRG), CAB (agriculture, GB), AGRICOLA, and IBISCUS (dedicated to developing countries).

1.2 - Non-computerized databases

(NB: the scope of the discussion is restricted to the centers that answered our questionnaire)

The ITC has significant dedicated resources, and in particular some 800 theses or reports. It scans 37 journals of the 93 we surveyed.

The GDTA also has resources comprising papers read at the different symposia it organizes, documents and grey literature (some 120 DESS students' study reports), and 400 technical reports. The GDTA also receives the major remote sensing journals.

The IGN *(Institut National Géographique français)* already possesses its own computerized database of over 8000 records, including photogrammetry data, forming part of a wider system comprising 60,000 works and approximately 2000 theses and 500 technical reports.

The GSTS *(Groupement Scientifique de Télédétection de Strasbourg)* has a semi-computerized stock of over 200 publications on remote sensing, 300 sets of congress proceedings, and 50 theses or study reports (1800 records in all).

The ORSTOM *Laboratoire d'Informatique Appliquée* is already co-producing databases with 200 remote sensing records so far.

The CRTO *(Centre Régional de Télédétection de Ouagadougou)* has a large number of unpublished regional monographs.

ESRIN in Frascati has significant documentary resources, mainly oriented toward space technology.

1.3 - Potential of existing resources

The importance of close ties between the RESORS database and the ISPRS-IRS project was stated above. An analysis of the 93 major journals dealing with remote sensing (see paper presented by H. Ten Haken in Mainz, 1982, updated 1988) indicates that RESORS scans 20 remote sensing journals, the majority being international or North American publications. Few European journals are scanned. Among the multidisciplinary databases, PASCAL (which scans 36 of the 93 journals) and GEOBASE (54 out of 93) give considerable coverage to remote sensing. These are less "technical" than the American databases such as NASA, NTIS and COMPENDEX and the British INSPEC. They should therefore be considered for inclusion in the project. However, collaboration with these databases can only be envisioned in terms of an agreement encompassing computer reformatting, additional indexing and the payment of royalties on interrogations in accordance with standard practice. Such collaboration is covered in greater detail below in § 4 under configurations B, C and D.

Specific databases deal in particular with earth sciences, geography, agriculture, hydrology and water resources. These can be seen as complementary to the ISPRS-IRS project, permitting subject specialists to do searches in their particular areas. We do not consider it reasonable to further develop collaboration in this area, except for agreements for the scanning of dedicated journals.

The non-computerized centers covered are mainly located in Europe. Broadening the scope of the survey was impossible without risking too great a degree of dispersion and making the feasibility study unrealistic. Extending the survey is, however, possible and perhaps desirable. The centers concerned plan to computerize in the fairly short term. The chief benefit of establishing a network would be the initiating of ongoing harmonization of such projects. The network would probably help in terms of practical implementation by proposing a basic form for description and scanning purposes, an indexing macro-vocabulary, and possibly joint management of the files. Technology and staff training in computerized processing at such centers, which are often small-scale, is highly variable. A joint proposal for coordinated computerization is urgently needed; this would permit access to invaluable documentary resources comprising papers produced at universities and research centers, technical reports, congress proceedings and grey literature.

2 - USER REQUIREMENTS AND ATTITUDE OF POTENTIAL NETWORK MEMBERS

The basic demand for information on remote sensing can be seen through the results of the enquiry into the centers and the requirements of the potential user community as revealed by a questionnaire sent to 600 organizations.

2.1 - Information requirements

Although potential users consider databases useful, they do not routinely interrogate them. Remote sensing and photogrammetry information is dispersed among various multidisciplinary and specific databases, causing a certain amount of inertia since several databases on different hosts have to be interrogated. Typical criticisms concern the different command languages and thesauri, and the difficulty of accessing source documents. The existing databases are considered to be suitable as far as journal articles are concerned, but little coverage is given to low-circulation and non-mainstream literature. A definite need exists and should be satisfied, providing that a suitable operational system is also set up for access to source documents.

It is not always easy to distinguish between subject needs, satisfied to a greater or lesser extent by interrogating subject databases, and those relating to a particular remote sensing or photogrammetry file.

This is the long-standing debate as to whether databases should be split on a subject basis or transversely, for example by technology, methodology, culture, etc. While the scattering across several hosts is observed to have a restraining influence on the use of databases, the choice between languages (English/French) and possible difficulties in the interrogation procedure are not considered to be major problems, probably because of the highly technical content.

2.2 - Attitude of potential members of producer network

2.2.1 - Interest in project

There are two types of "members": those who have actually declared an interest and those who may be interested. Organizations who intend, or are starting, to computerize their documentary resources generally express considerable interest. This applies to several European centers. Examples from those who responded to the March 1988 questionnaire include the ITC, GDTA, IGN, GSTS (Strasbourg) and ORSTOM. These centers clearly indicate their interest in harmonization, either in terms of task-sharing to achieve greater synergy, or for the definition of description and indexing procedures using the same documentary tools.

The Canadian RESORS remote sensing database producer is interested in greater dissemination in Europe, and is open to proposals that would help meet that objective. The size of the database led us to submit configuration C below which meets such a requirement.

Potential members are to be determined by analyzing the existing resources of multi-disciplinary producers that give significant coverage to our subject. The CDST-CNRS, the French producer of PASCAL (available on the Questel, ESA-IRS and DIALOG hosts) is open to all forms of collaboration, either by downloading or by input from PASCAL. The respective configurations are considered below. GeoAbstracts Ltd. of Norwich (GB), the producer of GEOBASE (accessible on DIALOG), also seems a potentially useful member.

2.2.2 - Possible involvement of organizations

Involvement is naturally proportional to motivation. Several organizations have stated a willingness to take on technical responsibilities in the operation of the future network. Once our congress has selected a configuration, the national decision-makers responsible for these organizations must confirm the commitment at their level, in particular by allocating the human and technical resources needed to ensure that the chosen network operates properly. Given the present climate, namely computerization already in hand or planned at many centers (and resources allocated), and a desire for coordination and more widespread dissemination, this should be feasible.

3 - PROPOSAL FOR AN ISPRS-IRS REMOTE SENSING AND PHOTOGRAMMETRY DATABASE

The desirable technical specifications for such a database were described by J. Hothmer at the Rio congress in 1984. On the basis of that work, and taking account of the new environment (existing resources, requirements, state of information systems), certain issues can now be further defined.

3.1 - General characteristics of database; creation and input

3.1.1 - Types and volume of information

The study of existing resources has shown that remote sensing articles in journals were in general covered (by RESORS, GEOBASE and PASCAL). However, greater coverage should be given in the only dedicated database, RESORS, to articles from Europe and developing countries. An initial survey of the journals is providing an opportunity to quantify the degree of inadequacy.

The organizations who have declared an interest in the future network each have significant documentary resources, as described in § 1.2 above. The most interesting documents, excluding major collections of journals (which are easier to locate and scan) are, of course, low-circulation documents which all the database producers are interested in but few provide. This applies to the documentary resources mentioned above belonging to the ITC, the GDTA, the IGN, etc, i.e. organizations with large stocks of universities papers, technical reports, congress proceedings and grey literature in general. An

initial effort should produce descriptions of 6000 to 10,000 documents of this type, with annual increments of 1000 to 2000.

According to the configuration chosen for the implementation of the network discussed below, we can envision a database essentially generated by downloading and reformatting journal articles, and by "local" input for other types of documents and for articles from Europe and developing countries. It is difficult to estimate the number of reformatted articles, as this would depend on future agreements, but it will be recalled that RESORS has 60,000 records, Pascal 16,000 and Geobase 8000. It can be noted that the cost of reformatting the CEGET database, merged with IBISCUS in 1987-8, is estimated at FF 17.00 per record, excluding the cost of the magnetic media and on the basis of 25,000 records. This information can be used to guide later agreements.

3.1.2 - Analytical or non-analytical?

This fundamental issue deserves serious consideration. The respective arguments are well-known; most databases, for example RESORS, are non-analytical (i.e. do not contain abstracts), while PASCAL and GEOBASE are generally analytical. It is thus clear that reasonably detailed indexing is necessary. We suggest that a basic indexing thesaurus be implemented fairly rapidly so that the first 500 records entered in the database can be indexed. Such basic indexing with controlled terms will be accompanied, if necessary, by additional indexing using candidate descriptors. After a careful evaluation of the frequency of usage, these will enable a more comprehensive and sophisticated thesaurus to be produced. This is now the conventional method by which thesauri evolve dynamically.

The feasibility of producing a basic indexing thesaurus shows that the job is partly facilitated by previous works, including the RESORS dictionary, the GDTA classification scheme, the ITC remote sensing CDU and the PASCAL-GEODE glossary (specific thesauri), and the IBISCUS-CEGET Thesaurus for generic terms. We propose that a preliminary digest be made of these glossaries, to be translated into the three ISPRS-IRS languages (English, French and German).

The final specifications will specify the potential and the cost of machine translation of the keywords. The advantage is that indexing can be done in one language with the keywords being translated at a later stage. This will partly depend on the type of management office and on the host to be adopted.

The physical description of the documents (cataloging) will take account of the currently-recognized international standards in force. A cataloging and indexing/scanning manual and a form will be produced by the working group under the framework of the executive guideline (this work is also linked to the choice of alternatives put forward in § 4).

The indexing of a record, the preparation of a form and the data entry itself represent on average 30-35 minutes of work. The creation of a record with an abstract requires 50-60 minutes. This information should permit an evaluation of the human resources that the members of the network will need to provide. These facilities have an impact on the choice between analytical and non-analytical.

3.1.3 - Official language

Alternative indexing language(s) were mentioned above, together with the possibility of machine translation using suitable software and multilingual glossaries, as done by PASCAL. Decisions must nevertheless be taken as to (a) whether records are to be presented both with the original title and with the title translated into a to-be-determined language, (b) whether the bibliographic details are to be recorded in English (or in French or German), and (c) in which language a possible abstract is to be drafted, this depending on whether the database is analytical or non-analytical (the use of two languages could also be considered for the drafting of abstracts). These points require further attention.

3.2 - Access to database, dissemination

3.2.1 - Types of products

The products most frequently produced by computerized systems are well-known, as they tend to be offered by the majority of producers: interactive interrogation, answer-providing services and Selective Dissemination of Information (SDI), and printed bulletins with several indexes obtained by sorting on one or more areas. Decisions on such matters are closely related to the type of management office to be set up, since an SDI answer-providing service, though very useful, in particular for facilitating access from

developing countries or users wishing to receive regular updates of their bibliographies, requires the allocation of at least 1.5 or 2 documentary assistants. Print products could be envisaged, but we feel this is best left to the initiative of the centers for their internal needs, i.e. such requirements will be better determined as a function of demand.

3.2.2 - Interactive Interrogation

This leads to the question of supply by the hosts. Contact already made, and the interest shown by ESA in our project, logically points to the ESA-IRS host for the dissemination of the database. A study is in hand to define the conditions for uploading. In particular, the cost of interrogation by network members needs to be specified. One possible model is the IBISCUS network system. Here, members play either a financial role or a production role; charges are modulated (or weighted) by the awarding of points according to the contribution made by members in their respective roles. The idea submitted by J. Hothmer, by which members playing a financial role should be found, is still relevant. Contact must also be sought with international and national partners such as the United Nations organizations, ACCT, ministries, etc. so that access can be optimized, particularly for our partners in the developing countries.

3.3 - Access to source documents

This key problem must be solved, the usefulness of a database generally being related to the operation of such a service. According to the type of management office chosen and established, the following types of solutions are possible:

3.3.1 - Management office coordinates service.

3.3.2 - Management office coordinates and provides service.

3.3.3 - A network member provides service, management office coordinates (includes configuration 3.3.1).

3.3.4 - Members jointly provide service, management office coordinates (includes configuration 3.3.1).

3.3.5 - The service providing access to documents is sub-contracted to a specialist organization. Now that large national systems are developing skills in this area, this idea seems feasible. Coordination with the network members is provided by the management office (includes configuration 3.3.1).

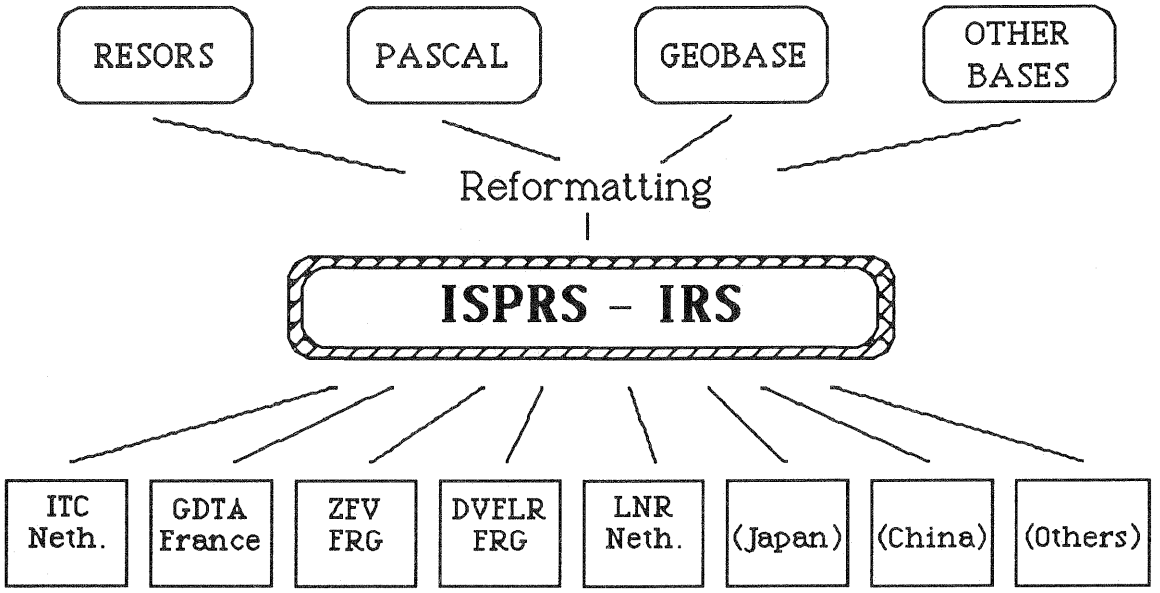
The staff requirement at the management office level for configurations 3.3.1, 3.3.3, 3.3.4 and 3.3.5 is around 0.5 persons. In view of the facilities required, configuration 3.3.2 does not seem feasible. The cost of copying documents will be established in the specifications according to the market supply conditions.

4 - POSSIBLE CONFIGURATIONS FOR IMPLEMENTING AN ISPRS-IRS DATABASE

A careful study of several strategies for implementing our project produced four proposals which constitute the conclusion to our work. Configuration A, a complete international network, is the ideal. Configurations B and C are feasible and should be studied in detail. Our commission should probably choose between these two. Configuration D is the minimum, whereby the ISPRS network would not actually create a database but would provide input to an existing base. We shall conclude our discussion of these configurations by outlining how the future network and its management office would operate.

4.1 - Configuration A: international ISPRS-IRS database

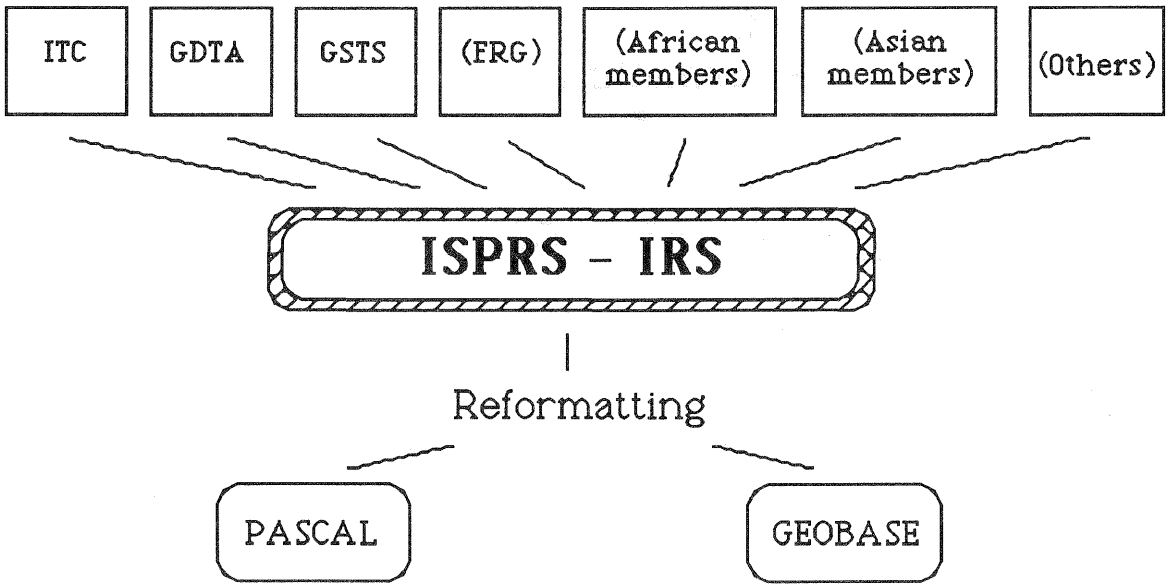
This is the ideal, as presented by J. Hothmer. The international database would materialize by reformatting data from other files and contributions from a very large network of organizations. An example of configuration A is shown below:



Such a configuration is attractive but is probably difficult to implement, in particular due to the extensive reformatting software required. It would call for a large management office, probably around 10-20 persons, to centralize information and provide the network with the necessary impetus.

4.2 - Configuration B: ISPRS-IRS database operated by European, African and Asian centers

This configuration takes explicit account of RESORS, which has 95% of its literature in English and a considerable stock of 60,000 records. Input to the ISPRS-IRS database would thus be directly from centers that are not presently, or only to a small extent, computerized, and possibly by the reformatting of such databases as PASCAL and GEOBASE. Additional indexing could be envisioned to provide reformatting with added value (around 10 minutes of work per record). An example of configuration B is shown below:



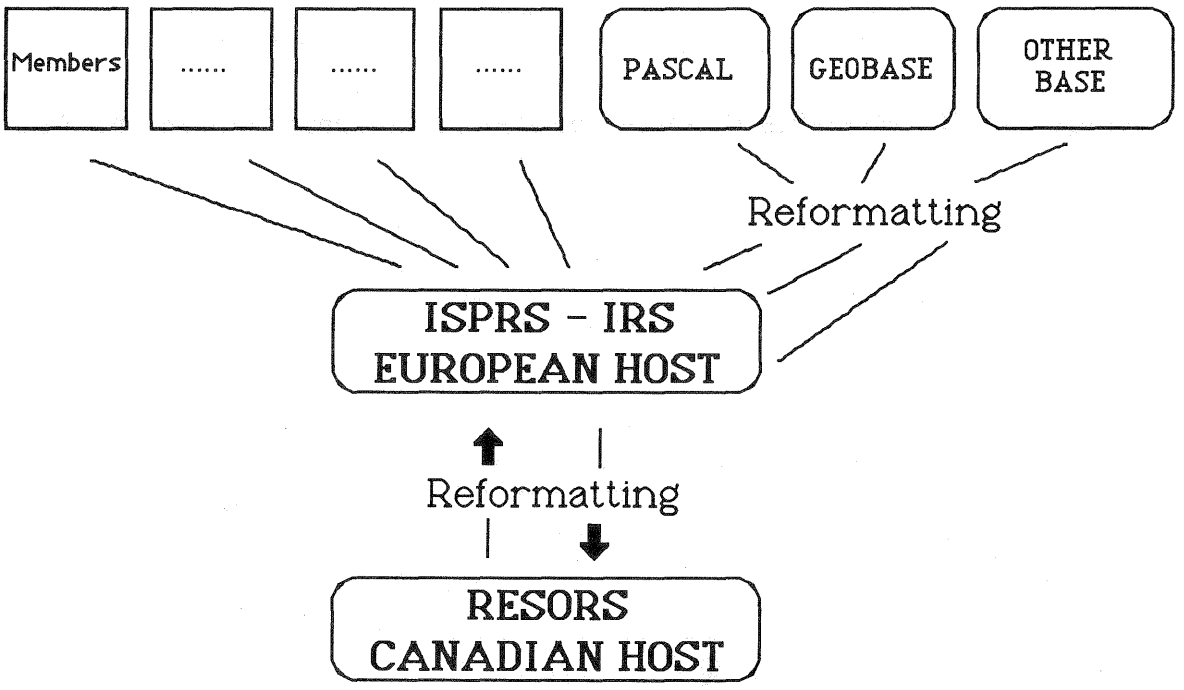
This configuration reduces the complexity of operating the network, just one or two reformats being required on a routine basis. The form and the screen format used could be based on those used at

existing databases. The management office is still essential, but can operate with 4-7 persons. Significant computing capacity is needed, particularly for reformatting.

4.3 - Configuration C: ISPRS-IRS / RESORS combination

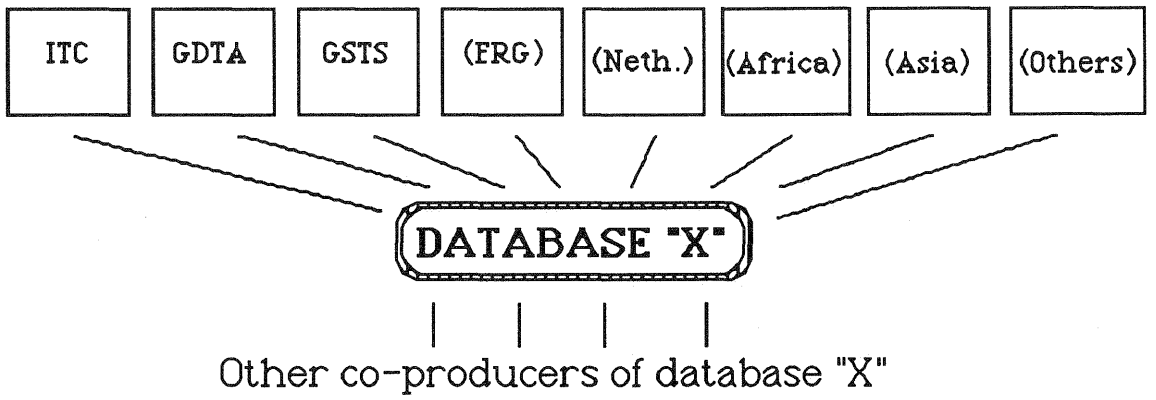
In view of the interest shown by the Canadian Remote Sensing Center for better dissemination of its database, collaboration should be considered. Configuration B would be selected (entirely or totally) for the generation of the ISPRS-IRS database, and a merged file would be input to the ESA-IRS database after initial reformatting (total or partial) of the RESORS data. To facilitate reformatting, the ISPRS-IRS format would have to be similar to that used by RESORS. A second reformat would permit the ISPRS-IRS part to be sent to RESORS for input in Canada. The management office would have an important role under such a configuration, and temporary support would be needed to reformat the substantial existing stock of 60,000 records. The indexing thesaurus would also have to be similar to that of the RESORS dictionary.

Here, we approach the limits of what is feasible in terms of reformatting, but this could be achieved by a competent and highly methodical group. If reformatted PASCAL and GEOBASE data were input to the ISPRS-IRS database itself, it would be difficult to achieve a coordinated indexing thesaurus for three different databases. The configuration is very tempting since it comes close to the ideal (configuration A), yet is probably feasible. However, we should be warned by the difficulties encountered by IFREMER-DOCOCEAN in this type of work. An example of configuration C is shown below:



4.4 - Configuration D: ISPRS-IRS network provides input to existing database

This minimum solution is probably the simplest and thus the easiest to implement. After agreement on who provides what, the network members would input documentary forms, for records in their possession, to a database (to be determined). Entry of the forms could be done by the database producer or by the network members using a software format compatible with the database receiving the input. Centers could thus manage their resources independently on microcomputers. The remote sensing and photogrammetry content of the growing database would thus increase. An example of configuration D is shown below:



The management office would be very small-scale, but basic coordination (managing the allocation of scanning tasks, maintenance of indexing thesaurus) would still be necessary.

4.5 - Operation of network

Operating rules are vital if our network is to have an ongoing existence. A suitable legal status must be found, or agreements signed, so that we act in complete harmony. Some form of Memorandum, containing the statutes and house rules, must lay down:

- rules for producing the database;
- rules for using the database;
- system for supplying copies of source documents;
- internal and external invoicing system;
- rules for reaching agreement and decisions.

According to the configuration chosen, the management office in charge of managing the network and operations will take one of the following forms:

(i) significant size, centralized: required for configurations A, B and C. Only possible if sufficient financial support obtained.

(ii) significant size, centralized, with persons seconded from the database producers (each center supplies staff and continues to provide salaries). One of the centers could provide premises, as in case (i). Suitable for configurations A, B and C.

(iii) significant or average size, centralized, implemented by one of the centers. The "host" organization is the main producer, members' responsibility being confined to drafting, and possibly entry, of forms.

(iv) average or small size, distributed system. Each center responsible for one of the following functions: entry and maintenance of forms, computer processing for database management, maintenance of indexing thesaurus, SDI service, access to source documents. Has the advantage of sharing costs, but a number of previous cases indicate problems in the full coordination of such a system.

The configurations outlined above and the different types of management offices obviously impact resource allocations. This, in turn, directly affects the choice. In particular, the hardware and software will only be defined after the choice is made, due to the considerable differences between the various configurations.

Conclusion

This paper arises naturally out of the deliberations at Mainz and Rio. It should permit our commission to take on its role of decision-maker and motivator, and indicate its choice of proposals in the closing resolution of the congress.

Such decisions on general guidelines are essential if the representatives of the national societies and organizations are to decide on their level of involvement in the project. We remind the Commission of the major choices to make and the areas to consider:

- type of information to enter in ISPRS-IRS database;
- ISPRS-IRS database proper (configurations B and C), combination with existing database(s) (configuration C), or input to existing system (configuration D);
- analytical or non-analytical database, indexing thesaurus, indexing language, language for abstract, database language;
- statute for management office, system for reaching agreement and decision;
- dissemination, products, host;
- access to source documents;
- brand (commercial) name for ISPRS-IRS database.

These decisions, mentioned in the feasibility study, are essential as they affect the future development of our project. An overall design study and set of specifications should be formulated so that the various operations can be coordinated and dispersed activity avoided. Irrespective of commercial aspects of products on the market, our scientific community must be seriously concerned about such coordination, this being one of the basic objectives of the ISPRS-IRS.

Toulouse, March 1988

* * *

Bibliography

France (1988) *Centre National d'Etudes Spatiales* (CNES), Paris. *"SPOT-1, Utilisation des Images, Bilan, Résultats"* , Paris, November 1987. Cepadues Editions, Toulouse. 1552 pp. ISBN 2.85428.188.8.

J. Hothmer (1984). The ISPRS-IRS information retrieval system for literature and factual data" (Report of Working Group VI-4). 15th ISPRS Congress, Rio de Janeiro.

C-H. Latarche, A. Buffeteau (1984). *"Contribution du GDTA à l'évaluation des bases de donnnées en ligne pour la télédétection"*. 15th ISPRS Congress, Rio de Janeiro.

J.H. Ten Haken (1982). "An investigation of available on-line data bases in the field of photogrammetry and remote sensing". Mainz symposium, ISPRS Commission VI.

J.H. Ten Haken (1980) "Automated storage and retrieval systems for photogrammetry and remote sensing". 14th ISPRS congress, Hamburg.