Problems in Merging Earth Sensing Satellite Data Sets

Paul H. Smith
Michael J. Goldberg
Problems in Merging Earth Sensing Satellite Data Sets

Paul H. Smith
Michael J. Goldberg

National Space Science Data Center
Goddard Space Flight Center
Copyright © 1998 IEEE. All rights reserved.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>2. SITUATION</td>
<td>1</td>
</tr>
<tr>
<td>3. PROBLEMS, APPROACH AND STATUS</td>
<td>1</td>
</tr>
<tr>
<td>3.1 Data Management</td>
<td>3</td>
</tr>
<tr>
<td>3.2 Data Dissemination</td>
<td>3</td>
</tr>
<tr>
<td>3.3 Multisource Registration</td>
<td>5</td>
</tr>
<tr>
<td>3.4 Data Merging</td>
<td>5</td>
</tr>
<tr>
<td>4. THE NEXT GENERATION</td>
<td>5</td>
</tr>
<tr>
<td>4.1 Recommendations for Future Systems</td>
<td>5</td>
</tr>
<tr>
<td>4.2 Pilot Development</td>
<td>7</td>
</tr>
<tr>
<td>4.3 Concerns</td>
<td>7</td>
</tr>
<tr>
<td>5. FUTURE/SOLUTIONS</td>
<td>7</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>9</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>10</td>
</tr>
</tbody>
</table>

PRECEDING PAGE BLANK NOT FILMED
1. INTRODUCTION

Satellite remote sensing systems provide a tremendous source of data flow to the Earth science community. These systems can provide scientists with data of types and on scales previously unattainable. Because of these opportunities, there is a growing emphasis in Earth science research to ask questions which are multidisciplinary in nature and global in scale. Many of these research projects emphasize the physical mechanisms of interactions between the land surface, the atmosphere, and the oceans. Conducting these research endeavors requires large and complex data sets and teams of multidisciplinary scientists, often working at widely dispersed locations. Unless adequate data and information systems are developed and employed, the full realization of the potential of satellite remote sensing, such as that expected through the Earth Observing System, Space Station and others, will be severely handicapped. In this paper, an overview is given of the problems expected to persist, at least into the near future, in merging these large volumes of Earth sensing satellite data into spatially referenced and manageable data sets.

2. SITUATION

The data and information situation in the remote sensing scientific community centers around three fundamental aspects that can be fairly simply stated. The situation is that there are (1) huge volumes of data, (2) many diverse data sources, and (3) a globally distributed and heterogeneous user community.

The first of the three aspects to consider is the volume of the data that must be handled. Present remote sensing systems such as NOAA's Advanced Very High Resolution Radiometer (AVHRR), the NASA-developed Thematic Mapper (TM), The French SPOT and others have already generated roughly a hundred terabits (i.e., $100 \times 10^{12}$ bits) of sensor data. The sensors of the 1990's will be even larger suppliers of data with the Aircraft Imaging Spectrometer (AIS), the Synthetic Aperture Radar (SAR), the Moderate-Resolution Imaging Spectrometer (MODIS), and other sensor systems in the Space Station era promising to provide five times the total cumulative amount of remote sensed data (i.e., $500 \times 10^{12}$ bits) every year!

Secondly, there are many satellite sensors or sensor systems which are and will be employed. These use different imaging technologies, have different spectral bands of various resolutions, and often present calibration difficulties and differences. The production of usable data products from these diverse sensors is receiving increasing attention and care. However, the data groups responsible for the products are becoming globally distributed and often are functioning with specific and unique purposes. The multitude of data formats continues to be a vexing problem.

Thirdly, not only are the sources of these remotely sensed data globally distributed with the user now needing to merge, in some appropriate fashion, data from more than one of the sources, but the users themselves are distributed around the world. The user communities are different. Four such user-group types are government, university, big industry, and small commercial. Even within a given type of user community, multi-purpose uses can be found, such as monitoring, mapping, specific scientific analysis or global change research studies. In addition, a given user or user institution will have widely varying capabilities for using the data, e.g. computers, software, expertise, time, money.

As is illustrated by the puzzled user in Figure 1, questions such as those posed are often not satisfactorily answered.

3. PROBLEMS, APPROACH AND STATUS

The situation described in Section 2 describes broadly-based and technically-deep problems that are beyond the scope of any single paper. However, it can be generally stated that Earth scientists do not want to overlook an
WHO HAS THE DATA?
WHERE CAN I GET IT?
CAN I USE IT?
HOW CAN I READ IT?
IS THERE ANY OTHER?

Figure 1.
important data set, wait a long time for data delivery, or spend a large proportion of their time and money converting data into useable form. To minimize these problems, current systems for the management, dissemination, and registration of multisource satellite and non-satellite data must be improved. This section will focus on the three sets of problems, (data management, data dissemination, and multisource registration), shown schematically in Figure 2, highlight key problems in each set, describe some approaches being taken to address the key problems, and give a brief status of those approaches.

3.1 Data Management

Data management systems for remote sensing basically organize satellite data, (and information about those data), in such a manner that scientists can rapidly determine what data are available for their particular research needs. Most present remote sensing data management systems do not support a wide enough variety of sensors or a flexible enough range of queries to satisfy the needs of interdisciplinary scientists. Historically, the earliest of these systems were set up only to manage data archives for particular satellites or sets of sensors. Each system gave access to a limited subset of remotely sensed data and provided no clues to the contents or even the existence of other related databases. Users had to have an intimate knowledge of specific sensor data characteristics in order to query the database and understand the system responses. More recent systems contain information about large groups of sensors related to a particular area of study, such as the atmosphere, land surface, or oceans. These systems provide on-line information about sensor data characteristics and provide a reasonably user-friendly and uniform method for querying the database about specific data holdings, regardless of the sensor. But the data holdings of these systems still do not support scientific investigations of land/atmosphere, ocean/atmosphere, and other interactions which cross discipline boundaries. In addition, several types of database queries which correspond to questions that scientists frequently ask about data are not generally available. Two of the most common questions relate to the availability of data for a specified portion of the Earth's surface and the quality (cloud cover, etc.) of such data. Even these queries are not supported for all sensor data types. The basic approach that is being taken toward the development of interdisciplinary Earth science data management systems is one of enhancement and integration of existing systems. The major enhancement that is being proposed is the development of a greater range of database queries. In particular, queries by latitude and longitude coordinates which describe rectangles or other closed boundaries are being developed so that multisource data can be identified for a given study site. The basic integration methodology that is being proposed is the development of automated directories that point to other database systems. This will at least provide the users of a given system with information on the existence, availability, and operation of other data management systems that may be of interest.

3.2 Data Dissemination

Earth science data dissemination systems are responsible for providing remote users with scientific data sets or with information about these data sets. Thus, these dissemination systems provide communications with data management systems as well as transmission or delivery of copies of selected data sets from their storage facilities to the scientific users. Current dissemination systems for geophysical data do not provide scientists with fast enough access to data management systems or to each others’ data. While scientific data networks, telecommunications links, and mail facilities have increasing capabilities for transmitting data quickly over long distances, logistical problems in providing remote data management service and in establishing data interchange standards remain. Data management service is difficult to set up over distributed locations because of questions about payment for service, (particularly in the loosely structured scientific community), responsibility for database administration, and responsibility for day-to-day maintenance and operations. Meanwhile, data interchange is hampered by wide differences in hardware/software configurations, communications protocols, and user applications at various network nodes.
Figure 2.
Slow but steady progress is being made toward the resolution of data dissemination issues. In response to stated Earth science goals, institutions responsible for the management of very large geophysical databases, (e.g. NASA, NOAA, USGS, CCRS, etc.), are beginning to set up policies and procedures for interdisciplinary data exchange. Spatial data interchange standards adapted from the Landsat family of data types have been proposed. These standards combine the "common format family" and the "data definition language" approaches to data representation.

3.3 Multisource Registration

Multisource registration is the process of converting spatial (e.g. image and map) data sets into a common structure, geographic orientation, areal coverage, and resolution to enable easy intercomparison. The technology for the registration of multisource, interdisciplinary data is more advanced than that for data management or data dissemination, but scientists still spend too large a percentage of their time and money getting their data sets into usable form. As is the case with data dissemination, the technical advances, (in this case high-speed, high-volume image processing systems and sophisticated digitizers), are offset by logistical problems. Two major cost-producers are the needless duplication of effort incurred in writing and running image reformatting software and the labor involved in correcting minor inconsistencies in digitized maps. Because remotely sensed images are archived in such a wide variety of formats, scientists usually bear the cost of writing and running software for converting them into a form that is compatible with the local image processing system. Meanwhile, the correction of digitized maps is still highly subject to human error.

Progress is being made on both of the above mentioned problems. The cost of image data reformatting can be reduced in two ways. First, development of a generic image display format could be developed in conjunction with data interchange standards. Second, archive-to-image reformatting could be performed as a service by major data suppliers. A promising approach to the automatic editing and digitization of maps involves the use of knowledge engineering using expert systems technologies to duplicate the procedures used by experts to detect and correct errors.

3.4 Data Merging

While data merging is not an easy task, high quality products of merged data can be found. One such example is the Data Mosaic of the Western United States from AVHRR sensors during May, 1984 shown in Figure 3. This product is an Albers Equal Area Projection in color and is courtesy of R. J. Thompson, United States Geological Survey, EROS Data Center in Sioux Falls, South Dakota.

4. THE NEXT GENERATION

4.1 Recommendations for Future Systems

In the past few years, a number of studies, reports, workshops, and pilot systems have focused on the issue of improving handling and management of scientific satellite data. Most recently these activities are addressing the enormous difficulties referred to in Section 2.

One such activity is the recently published EOS Data Panel Report which has been recently published. Mike Ward, EOS Ground Systems Study Manager, NASA/GSFC, has kindly provided a summary view of the recommendations that have resulted from the EOS Data Panel’s studies. The EOS Recommendations are for:

- Geographically distributed system
- Sites of varying capabilities and responsibilities
- Flexible, modular, layered architecture
  - expandable, evolvable system
- Centralized coordination and information management
- Distributed data archives
- Telescience - remote and interactive electronic access
- Research, operational, and commercial needs
- Transparent access to other NASA and non-NASA data sets
- Timely access to the science data
  - real-time data
  - archival data
- Data standards which allow data autonomy

The reader is referred to the full report of the EOS Data Panel for a complete discussion of the Panel’s recommendations.

4.2 Pilot Development

NASA has just begun to build the Pilot Land Data System (PLDS). This is expected to be one of the most advanced data and information systems in the world for providing integrated data management, networking and communications, system access capability, land analysis software and special processes.

PLDS will include a central on-line directory of remote databases and data archives, distributed data catalog system, full information query capability about the data sets, state-of-the-art spatial data management functions, high speed communication lines servicing over thirty scientific investigator nodes by 1987, uniform access to a variety of NASA computer systems, use of both super computers and powerful workstations, library of software tools, comprehensive image analysis software, and advanced processing techniques and storage devices.

The functional concept of the distributed PLDS is shown in Figure 4. This illustrates distributed nodes at a few of the organizations expected to be participating with PLDS, with the user at any one of the nodes being able to access information and data at different places in the network to use various software tools for processing and analysis, and to perform the computations on the appropriate computer systems whether it be workstations or super computers.

4.3 Concerns

While there are many technical and scientific obstacles to overcome and great expectation of succeeding, some concerns remain which organizations such as CODATA and other international scientific groups might wish to consider. These are not easily answered or simply solved. Some of the concerns that come to mind are these: institutional problems that result in system incompatibilities; conflicting purposes and/or needs of remotely sensed data such as rapid use in forecasting, long-term studies of climatic change, detailed scientific investigations of physical mechanisms and studies of global change processes; data quality assessments that are at times non-existent or, at best, non-uniform; data archived at inappropriate levels of processing such that subsequent merging is difficult if not impossible, and cost effectiveness. Are the benefits in being able to merge remotely sensed data worth the cost required to solve the various difficulties?

5. FUTURE/SOLUTIONS

In summary, the problems of merging large volume data into spatially referenced and manageable data sets are formidable, but the future holds solutions to these difficulties and tremendous benefits can be derived. These
Figure 4. PLDS Distributed Functional Concept
difficulties can be overcome through the definition of common expectations, establishment of multidisciplinary teams, adoption of integrated system design and standard, development of distributed systems and the application of advanced technology.

ACKNOWLEDGMENTS

We gratefully acknowledge the contributions toward the preparation of this review-style paper made by the following people:

David M. McKeown, Jr.
Carnegie-Mellon University
Donna J. Peuquet
Pennsylvania State University
Robert H. Rogers
Environmental Research Institute of Michigan
Philip H. Swain
Purdue University

A special note of appreciation goes to R. J. Thompson, USGS EROS Data Center, for providing significant materials for this paper. The contributions of Michael Ward and the authors of the papers cited in the Bibliography are also acknowledged.
BIBLIOGRAPHY


Rogers, R. H., personal communication, ERIM, 1986.


Thompson, R. J., personal communication, USGS EDC, 1986.
Satellite remote sensing systems provide a tremendous source of data flow to the Earth science community. These systems provide scientists with data of types and on a scale previously unattainable. Looking forward to the capabilities of Space Station and the Earth Observing System (EOS), the full realization of the potential of satellite remote sensing will be handicapped by inadequate information systems. There is a growing emphasis in Earth science research to ask questions which are multidisciplinary in nature and global in scale. Many of these research projects emphasize the interactions of the land surface, the atmosphere, and the oceans through various physical mechanisms. Conducting this research requires large and complex data sets and teams of multidisciplinary scientists, often working at remote locations. A review of the problems of merging these large volumes of data into spatially referenced and manageable data sets is presented in this paper.