

AUTOMATED DATA DISCOVERY AND USAGE



July 21, 2003

Christopher Lynnes
NASA Goddard Space Flight Center
Code 902, Greenbelt, MD 20771

Abstract	2
Introduction	2
Challenges in Achieving Automated Data Discovery and Usage	4
Data Detection and Location	4
Data Evaluation	5
Data Access	5
Data Usage	5
Pervasive Standardization	6
Universal Translators	6
Data Mapping	7
Future Vision	7
Conclusion	8

ABSTRACT

Adding new datasets into research and applications is a mostly manual function today. In order to attain the knowledge-building systems of the future, however, intelligent archives must be able to support the automated detection/location, evaluation, access and usage of the data. Detection is primarily a matter of catalogs with sufficiently rich metadata, particularly where data quality is concerned. Data evaluation is more difficult to automate, though sample data and reference datasets may make this problem tractable. Data access is straightforward for online data, but is more difficult for offline or nearline data. Asynchronous interfaces to tape libraries have proven resistant to standardization, but pseudo-synchronous gateways or extensions to Grid technologies may provide a path to a solution. Finally, automatically incorporating, i.e., using, novel datasets presents a major challenge. We present three approaches: pervasive standardization, or application of standards to all aspects of a data product; universal translators (gateways that translate from native format to a client-friendly one); and data mapping, in which machine-usable descriptions are created for the data and metadata structures to sufficient detail to serve as a “map” of the data product. The eventual vision of an intelligent archive is one that can automatically detect the presence of new, useful data sources, evaluate their usefulness, determine how to access them, and provide them in an understandable format to the archive’s users.

INTRODUCTION

The key to science is access to the right sources of data and (especially in the case of near-real-time applications) at the right time. Discovering such sources of data and incorporating them into research and applications is currently a lengthy, laborious, almost completely manual enterprise. For remote sensing data, the process appears at first glance to be at least straightforward. There are, after all, a limited number of well-known remote sensing satellites, normally with tightly controlled data streams and a well-defined set of data products. However, the reality is more complex. For example, a number of satellite instruments (e.g., AVHRR, MODIS, AIRS) broadcast data directly to antennas on the ground (known as direct broadcast or direct readout stations). This fosters the proliferation of regional remote sensing datasets, whose limited range is compensated by decreased latency and sometimes increased spatial resolution (as in the case of SeaWiFS High-Resolution Picture Transmission). In turn, the products derived from direct broadcast data streams often diverge from “standard” products from the same satellite

due to different algorithms. For example, calibrated radiance data from MODIS Direct Broadcast may be in one of (at least) three formats: Hierarchical Data Format-Earth Observing System (HDF-EOS), International MODIS/AIRS Processing Package (IMAPP¹) HDF, or a flat binary. Furthermore, remote sensing data must typically be accompanied by in situ data to provide ground truth, correlative and ancillary data. By their very nature, in situ data tend to be widely distributed and diverse in format, structure and metadata.

The Intelligent Archives project seeks to create a next generation conceptual archive architecture that is able (among other things) to increase overall data utilization and automate the transformation of data to information and knowledge². Intelligent archives play a key role in an overall knowledge building system by providing intelligent data management, persistence and understanding services, allowing users to focus on research and applications rather than data and data system manipulation (Fig. 1). An important challenge to intelligent data management and understanding is to mitigate or eliminate the manual process of discovering novel useful data sources, and just as important, incorporating those datasets into analyses, models and applications. Intelligent archives of the future should be able to detect the presence of newly available data anywhere in the world, determine the usefulness of the data, learn how to access them and ultimately provide the data to users or applications in a form in which they can use it.

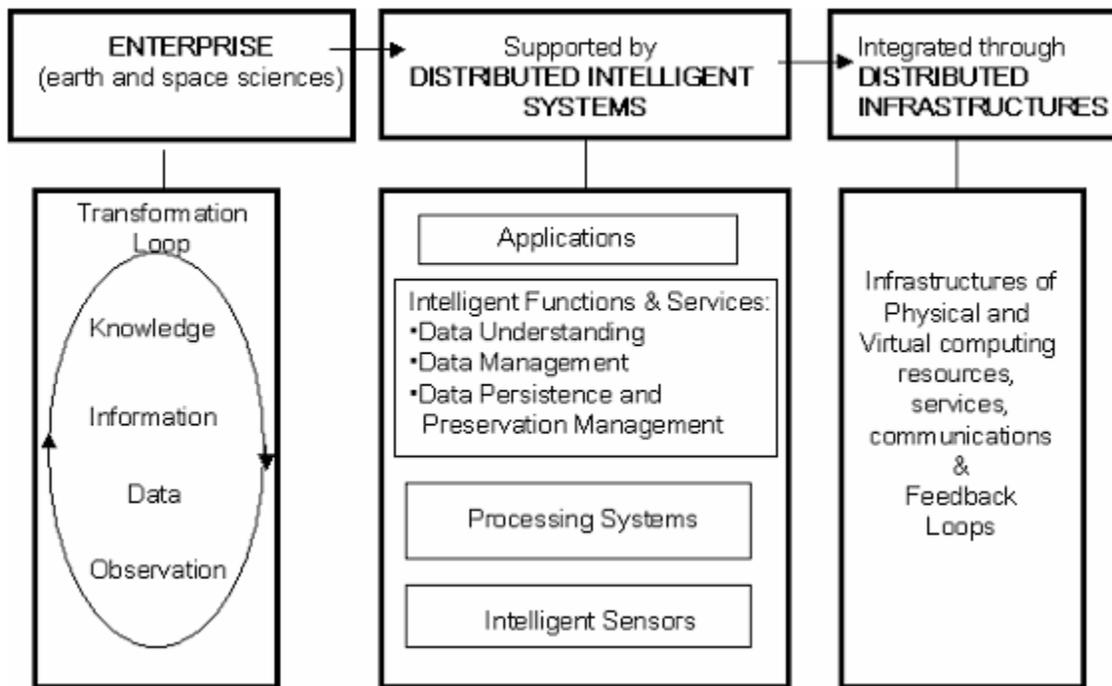


Figure 1. Context of Intelligent Archives, which provide Data Understanding, Management and Persistence in support of the transformation of observations to knowledge.

In order to accomplish this objective, we must first identify the obstacles to automated data discovery and usage today. The most fundamental problem, of course, is simply detecting the existence of potentially useful data; a closely related problem is identifying the network location of such data. The MODIS Direct Broadcast case provides a vivid example. Because these stations are sold by commercial entities, or even locally built in some cases, the total number of stations world wide is unknown, though it is believed to be several dozen; furthermore, only a small subset of these stations make their data available to others, with a varying set of restrictions.

Once data have been identified and located, assessing their quality represents another hurdle. The quality is a function of not only the sensor which collected the data, but also the science algorithm used to

process it, the version of calibration tables that go into it, the production rules, and the general operational environment. For example, data produced for low-latency applications must typically do without useful ancillary data (e.g. global model output), producing a lower quality product than research-driven products that wait for and incorporate those ancillary products.

The next hurdle to overcome lies in the mechanisms by which the data are obtained. The simplest mechanism is public online data available through standard synchronous protocols such as anonymous FTP or HTTP. However, some data resides in near-line or offline archives, forcing an asynchronous order-retrieve sequence; other data may have restricted distribution or require payment.

Finally, perhaps the most challenging hurdle is acquiring a detailed enough understanding of the data format and internal structure to use the data without resorting to additional software development for each dataset. This requires an understanding of the overall format of the data (e.g. Hierarchical Data Format); the location of particular geophysical parameters within that format, as well as the metadata describing them; the location of required ancillary data (such as geolocation); the type of interleaving used for multi-dimensional parameters; and finally how to transform the actual stored numbers into a parameter with known units. For example, many discrete geophysical parameters (e.g., MODIS cloudmask³) lend themselves to storage as complicated bit masks, requiring a detailed understanding of how the individual bytes themselves are structured.

While the problems outlined above are thorny, none of them seem insoluble. In the next section, we address the key challenges of Detection, Location, Evaluation, Access and Usage in more detail, together with potential technological solutions to address these problems in a limited fashion. The following section will illustrate a future vision if these technologies can solve these key problems.

CHALLENGES IN ACHIEVING AUTOMATED DATA DISCOVERY AND USAGE

DATA DETECTION AND LOCATION

The detection and location of useful data must be approached on two different levels: the dataset level and the individual data item level. In other words, an automated client must first determine that a collection of data provides one or more useful geophysical parameters. Once collections are discovered, the application must zero in on the particular sub-units of the collection that cover the spatio-temporal area required by the application. While there are many thousands of potentially useful remote-sensing-related data collections throughout the world, the number of individual data items (most commonly files, but potentially database records or file subsets) is on the order of 10^8 and growing rapidly.

There exist today a number of data directories providing information at the dataset level. NASA's Global Change Master Directory, for instance, has information about 11,000+ datasets⁴, primarily (but not exclusively) related to remote sensing and global change research. Similar directories include the Federal Geospatial Data Committee Clearinghouse⁵ and the Master Environmental Library⁶. These provide a wealth of structured and unstructured information about data collections.

Information at the data item (granule) level is more difficult to come by. The Earth Observing System Data and Information System (EOSDIS) provides a set of search and order protocols to EOSDIS and partner entities, allowing item-level information to be queried and obtained from individual data centers. More recently, the EOS Clearinghouse (ECHO) has been developing a capability to accept item-level metadata from data centers, providing a single point of contact for developing tailored clients for querying.

DATA EVALUATION

Simply detecting and locating data is not enough for automated data discovery. The suitability of that data for a particular application must be evaluated, especially when multiple similar sources for a geophysical parameter are available. Typically, evaluation criteria break down into quality and timeliness, with the former usually trading off against the latter. Quality criteria include sensor accuracy, algorithmic accuracy, spatial resolution, temporal resolution and quality assessment results. Timeliness is simpler to define, in terms of the time differences between data collection, data product generation and data availability. However, it is a function of a number of complex factors (most of them relating to quality), such as how long the producer waits for ancillary data, the speed of the algorithm, the resolution (lower resolution data takes less time to transfer), the level of quality assurance applied, and the form in which data are available (online vs. offline).

In order to automatically select data based on quality and timeliness, the above criteria must be available in a machine-parseable form. This can be challenging due to the imprecise nature of terms related to quality and accuracy. Clearly some standardization of accuracy and other quality terms would be useful. Another avenue would be the establishment of standard benchmark data sets against which new data sources could be compared. By checking a novel data set where it coincides with enough known benchmark data, one might have confidence that the data are also usable elsewhere. Such a benchmarking effort would need to combine ground validation sites with satellite measurements whose quality has been rigorously controlled.

DATA ACCESS

Once suitable data have been located, automated data access becomes the next critical step. For data that are available online, this is relatively straightforward, typically involving Internet Universal Resource Locators and the Hypertext Transfer Protocol or File Transfer Protocol. However, offline (or nearline) data introduces a complication: this interaction follows an asynchronous sequence beginning with a data order, followed by staging the data, notification to the user that the data are ready for retrieval, and finally retrieval of the data. Some attempts are being made to standardize ordering protocols, such as the PRODUCT_REQUEST protocol of EOSDIS⁷ and the Web Services-based Order Entry Service of the EOS Clearinghouse (ECHO)⁸. However, the notification messages from each data provider tend to have a unique format. Consequently, anyone developing automated applications parsing these messages is forced into new coding to accommodate new providers. Standardization of distribution notifications would ameliorate this situation.

An alternative is to provide gateways that emulate synchronous interfaces: such an interface would serve as a proxy by submitting the order, waiting for the notification, and retrieving the data on behalf of the client. Ideally, this emulation would be an extension of accepted synchronous service standards. For example, the Storage Resource Broker (SRB)⁹ is an emerging standard for remote mass storage data access in Grid computing, but it currently supports only synchronous access to mass storage with limited queuing features. This makes it unsuitable for highly read-active robotic tape libraries like those of the EOSDIS Distributed Active Archive Centers, which rely on batching and intelligent queue management to minimize both tape mounts and seeks. However, an SRB-based synchronous-asynchronous gateway (or intelligent queuing) could provide safe and efficient Grid access to active libraries.

DATA USAGE

Of all the challenges, the most difficult is to automatically incorporate new data products without adding new code. NASA's science community has been attempting to define data format standards for more than a decade to make this kind of universal access possible, from the Common Data Format¹⁰ (CDF) to Hierarchical Data Format¹¹ and the HDF-EOS¹² standard of the EOS project. Other communities have been engaged in similar fashions, leading to such formats as netCDF¹³. However, many of these efforts have run into difficulties reconciling strict adherence to standards with the flexibility necessary to accommodate new data products.

Specifying a format is not always enough: a number of other elements can be critical to understanding a dataset. One example is the semantic meaning of discrete variables, such as the bit field values in a cloudmask product. Similar challenges arise when interpreting the metadata that describe the data. Another complicating factor is the location of necessary ancillary data. For example, different MODIS products vary widely in the approach to geolocation information. Some products, such as the 0.5-km calibrated radiance, contain Latitude and Longitude for each field-of-view within the calibrated radiance file. Others, such as the 1-km calibrated radiance, include only a subsampled version of the geolocation radiance. Yet others, like the Level 2 ocean products, contain no internal geolocation information, relying instead on external geolocation products. Thus, in order to automatically incorporate such data, information on where or how to find the geolocation data is vital. Similar problems arise when with ancillary data such as quality control parameters. In this case, the problem is compounded by a lack of standardization in terminology.

There are essentially three types of approach to solving the data usage interoperability problem: pervasive standardization, universal translators, and data mapping.

PERVASIVE STANDARDIZATION

This approach moves beyond specification of data formats to apply standards to all aspects of a data product. These aspects begin with the overall format, data structures and metadata model, but also include:

- Scientific units
- Metadata terminology
- Discrete variable semantics
- Geolocation structures and location
- Ancillary data structures and location
- Quality control semantics

This approach has the advantage of being straightforward. However, it is not always clear beforehand how exhaustive and rigorous a standards specification must be to enable truly automated data usage. This can be addressed to some degree by constructing additional layers of standards. The original specification of HDF went a long way toward a uniform format for the EOS project, but it has so much flexibility in the storage of datasets that it is not possible to write generalized programs to operate usefully on every HDF dataset. This was followed by HDF-EOS, a set of data structure specifications within HDF, such as Swath and Grid, which enable some generalized subsetting. Even that retains enough flexibility to render the HDF-EOS subsetting capabilities of limited use for certain data products (such as the MODIS L2 oceans, with no internal geolocation information). One possible mitigation of this uncertainty is to develop a reference application to read and apply some basic interpretation to data products using the standard specification. As it is applied to more and more products, it should become clear where the standard is insufficiently rigorous.

A second difficulty with the pervasive standardization approach is the difficulty of achieving community consensus on standards, particularly in areas related to semantics. This gives rise to a high overhead in accommodating novel datasets, as the often-lengthy standardization process must be followed to conclusion before beginning to generate the new products.

Perhaps the greatest problem with pervasive standardization is how to handle existing, or heritage, data. Generally, such data do not follow the standard, and so must be reformatted to be usable in an automated sense. This reformatting process is typically difficult and expensive.

UNIVERSAL TRANSLATORS

In this approach, gateways provide translation from datasets in their native format to one of a small set of standard formats, often upon transfer from the data supplier to the user. In this respect, it incorporates exhaustive standardization, but applies it on the fly. The OpenGIS approach, for instance, supports a few

basic formats (e.g., GeoTIFF, HDF-EOS)¹⁴. A similar approach is the Open Data Access Protocol (OpenDAP), which has both a standard data transport protocol and an application program interface available for client programs. The approach puts the burden of supplying the translation on the server side, and thus the data supplier. The data supplier must either furnish a server to support the data transfer protocol, or some translation information to an off-the-shelf server (as is the case with OpenDAP). There are potential performance penalties to translating the data for transfer, and indeed both OpenGIS and OpenDAP protocols include provisions for subsetting the data at the server before transfer, thus limiting the penalty somewhat. Also, it is important for the data transfer standard to be exhaustive enough to support automated data interpretation; as we have seen, specifying HDF-EOS, for instance, is insufficient for many datasets. As with exhaustive standardization, the existence of a reference application could be useful to ensuring the standard is sufficiently rigorous. However, a key advantage of the universal translator approach is its ability to handle heritage data, by supplying the necessary server or translation information.

DATA MAPPING

The Data Mapping approach consists of machine-parseable descriptions of data. This would include the layout of data fields within files or databases, location of geolocation and ancillary data, descriptions of the metadata model, and transformational equations to convert data variables into standard units. These descriptions would be supplemented by thesauri to handle semantic mappings for discrete variables, metadata terms and quality control information. The result would be a virtual map of the data product that could be used by applications to locate and interpret any data variable within it.

To some extent, the Data Mapping approach is similar to the Semantic Web¹⁵, which seeks to build a language for describing information and services in a machine-usable form. This “language” is the Resource Description Framework (RDF), which provides a mechanism for describing classes, properties and domains. A semantic web then requires domain-specific descriptors represented within RDF, based on the ontology of that domain. The Earth Science Information Partners Federation (ESIP-Fed) is currently developing a Semantic Web for Earth and Environmental Terminology for the earth science domain¹⁶. In order to enable usage, a machine-usable description of the underlying data formats is also needed. This is the intent of the Earth Science Markup Language¹⁷, which uses XML to describe data variables within a data file. Limited transformational equations are also supported to convert into standard units; however, transforms for discrete variables are not yet addressed.

Although only beginning to take root, the data mapping approach circumvents many of the disadvantages of Exhaustive Standardization and Universal Translators. For instance, although community consensus on terminology is useful for these purposes, it is not essential. Indeed, more than one data map may exist for the same data product depending on the needs or expectations of the end-user application. Also, heritage data need not be reformatted, merely described in machine-usable terms. While data mapping may be used to help implement universal translators, they may also be used by the end application, thus avoiding the performance penalty of the translation.

FUTURE VISION

While it is difficult to predict the future of earth science data, we can speculate on potential evolution from developments today. Many of these developments point toward more data collection and distribution by small operators. For example, NASA has made a concerted effort to foster smaller, more distributed data providers through the ESIP Federation and the REASoN CAN. Another example comes from the first few years of the EOS Terra and Aqua missions, which has spawned a large number of Direct Broadcast stations around the world (upwards of 40 as of 2002), largely driven by the MODIS instrument. These stations enable station operators to acquire a rich data set with low latency, who collect the data for a variety of reasons, from applications uses to regional redistribution. Unfortunately, it is difficult to identify all the stations, let alone determine which stations provide data and in what format.

In the future, it is even possible that data collection could move into the realm of small business and individuals, particularly farmers and ranchers. We can envision a data collection grid, whereby individuals collect in situ weather data such as temperature, rainfall, humidity and wind, making the data available to the science and applications communities. Indeed, a similar scheme has been operating since 1853 on the high seas, namely the Volunteer Observing Ships scheme that recruits merchant vessels to collect meteorological data¹⁸.

Individual data collection has the potential to be useful, but its non-institutional nature also presents several challenges to data usage. Standards can be more difficult to enforce, for example. Also, individual data collection points are perforce more ephemeral in nature. The ephemeral nature of both the data collection systems and the datasets themselves represents a key challenge to automated data discovery and usage. To be of maximum use, the appearance of new datasets must be detected quickly. For applications usage, the data must also be followed shortly thereafter by incorporation.

Thus in our future vision of automated data discovery and usage, we may see the sudden appearance of new and ephemeral data sources and data sets, such as a new direct broadcast station (or new products at an existing one), a ship borne observation platform during a voyage, or a new in situ observation site. Such a source would advertise its existence, along with its data or services, sample data and a data map (see previous section). The IA would detect the existence of these new data and services within minutes of their initial appearance, most likely by subscribing to a directory such as GCMD. The new data products and services would be evaluated by metadata inspection plus comparison of sample data to reference benchmark datasets wherever possible. The IA would then determine the availability of standard access methods (e.g., FTP, HTTP, SRB) for acquiring the data or invoking the service, and would then poll users (human and machine) for interest in the data. For research and some application users, the IA would then serve as a data translator, both providing the data to the user in a client-friendly format and storing the data for later reanalysis or reprocessing. This persistence management is essential for data sources where the continued existence of the data is in doubt. Low-latency applications users, on the other hand, may handle data access directly to avoid the latency increase from translation.

CONCLUSION

The growth in distributed, heterogeneous data providers is exposing a need for a way to automatically discover and use these new sources of data. Automation is particularly critical for low-latency applications and for those using a wide variety of data, such as certain kinds of data assimilation and mining applications. All of the key steps in the process, from detection/location and evaluation through access and usage, require substantial development to achieve true automation. However, all of them seem tractable in the long run.

The role of the intelligent archive in this future data collection realm is somewhat ambiguous. The rise of small organizations and individuals as data providers seems at first glance to diminish the role of archive centers. However, the ephemeral nature of many of these smaller data collectors casts them more in the role of interim archives, and may actually push the intelligent archive center into a role of active data acquisition, so that it becomes the key collection point on behalf of data users in order to provide a *stable persistent* repository. Thus intelligent archive centers would be themselves users of the technologies supporting data detection/location, evaluation and access. The archive centers would also likely provide the Universal Translators or Data Mapping information to enable automated usage by their various users.

¹ Strabala, K., I., L. E. Gumley, T. Rink, H.-L. Huang, R. Dengel, 2002. MODIS direct broadcast products and applications, SPIE Remote Sensing of the Atmosphere, October 2002.

-
- ² Ramapriyan, H. K., G. McConaughy, C. Lynnes, S. Kempler, K. McDonald, R. Harberts, L. Roelofs, and P. Baker, 2002. "Conceptual Study of Intelligent Archives of the Future", Report prepared for the Intelligent Data Understanding program, 39 p., http://daac.gsfc.nasa.gov/IDA/IA_report_8-27-02_baseline.pdf.
- ³ MODIS Atmospheres: MOD35_L2: Content & Format, http://modis-atmos.gsfc.nasa.gov/MOD35_L2/format.html.
- ⁴ FAQ: What is the Global Change Master Directory (GCMD) and how can it help me? http://gcmd.gsfc.nasa.gov/Aboutus/gcmd_faq/about.html.
- ⁵ FGDC Geospatial Data Clearinghouse Activity, <http://www.fgdc.gov/clearinghouse/clearinghouse.html>.
- ⁶ Master Environmental Library, <http://mel.dmsi.mil/>.
- ⁷ Schessler, J., 2001. Interface Control Document for ECS Interfaces That Support External Subsetters Located at DAACs, 209-CD-036-001, 74 p.
- ⁸ ECHO – EOS Clearinghouse, <http://www.echo.eos.nasa.gov/>.
- ⁹ Wan, M., A. Rajasekar, R. Moore, P. Andrew, 2003. A Simple Mass Storage System for the SRB Data Grid. 20th IEEE/ 11th NASA Goddard Conference on Mass Storage Systems & Technologies (MSST2003) San Diego, California, April 7-10, 2003, <http://www.npaci.edu/DICE/Pubs/tapeMS2003.pdf>.
- ¹⁰ NSSDC's CDF Homepage, http://nssdc.gsfc.nasa.gov/cdf/cdf_home.html.
- ¹¹ NCSA HDF Home Page, <http://hdf.ncsa.uiuc.edu/>.
- ¹² HDF-EOS Tools and Information Center, <http://hdfeos.gsfc.nasa.gov/>.
- ¹³ Unidata NetCDF, <http://www.unidata.ucar.edu/packages/netcdf/>
- ¹⁴ OpenGIS Consortium, Inc., 2002. OpenGIS ® Web Coverage Service (WCS) Implementation Specification, <http://www.opengis.org/techno/02-024r1.pdf>.
- ¹⁵ Berners-Lee, T., J. Hendler and O. Lassila, 2001. The semantic web, *Scientific American*, 285(5), p. 35-43.
- ¹⁶ Raskin, R., Semantic Web for Earth and Environmental Technology, <http://sweet.jpl.nasa.gov/>.
- ¹⁷ Ramachandran, R., M. Alshayeb, B. Beaumont, H. Conover, S. J. Graves, X. Li, S. Movva, A. McDowell, M. Smith, 2001. Earth Science Markup Language: A Solution for Generic Access to Heterogeneous Data Sets, NASA Earth Science Technology Conference 2001, August 28, 2001.
- ¹⁸ The WMO Voluntary Observing Ships (VOS) Scheme, <http://www.vos.noaa.gov/wmo.html>.