THE ROLE OF HYPERSPECTRAL METADATA IN HYPERSPECTRAL DATA EXCHANGE AND WAREHOUSING

RASAIAH Barbara[a], MALTHUS Tim J. [b], JONES Simon D.[c], BELLMAN Chris[c]

[a]PhD Candidate, Centre for Remote Sensing, RMIT University, Melbourne, Australia – barbara.rasaiah@rmit.edu.au
[b]CSIRO Land and Water, Canberra, Australia – Tim.Malthus@csiro.au
[c] RMIT University, Melbourne, Australia – (simon.jones, chris.bellman)@rmit.edu.au

Abstract –Hyperspectral databases and data exchange across wide area networks are becoming increasingly prolific in the remote sensing community as a way of sharing, cataloguing, and mining data from a variety of sensors. A data warehousing model, encompassing distributed spectral libraries and databases has become necessary for integrating a broad range of hyperspectral data formats and sources. Metadata is an important component in terms of maintaining both the quality and reliability of a hyperspectral dataset and the products derived from it. Metadata affects all tiers of data processing and user roles in such a system, from the recording of *in situ* measurements, to instrument calibration and data validation. This paper is an introduction to doctoral research investigating the necessity for a coordinated evolution of hyperspectral metadata protocols, field spectroscopy methods and data exchange standards for a new phase of collaboration in the hyperspectral remote sensing community.

Keywords – *Datawarehouse, Databases, Digital, Hyperspectral, Metadata, Field Spectroscopy, Interoperability*

1. INTRODUCTION

Hyperspectral metadata plays a significant role in any hyperspectral dataset, whether it refers to *in situ* observations or those derived from airborne and spaceborne sensors. It can describe the properties of the target being viewed, illumination and viewing geometry, sensor calibration, and environmental conditions -- all of which are influencing factors that affect standardized measurements (Pfitzner et al, 2006). In a conceptual sense, field spectroscopy acts as the fundamental stage for primary research and operational applications (Milton, 1987). The radiometric data in itself is not sufficient in providing the potential for accurate, consistent environmental modelling and long-term data legacy. Both the imaging conditions and instrument itself can introduce systematic and random errors on recorded radiance, target discriminability and contrast, prompting the need for ancillary information including conditions of observation and field techniques (Duggin, 1985). Metadata can be an effective tool in describing and quantifying these errors, and potentially mitigating them. The time invested in metadata collection is outweighed by its benefits in reducing system bias and variability (Pfitzner et al, 2006). Therefore, metadata must be a central consideration when creating reliable hyperspectral datasets with legacy potential. Currently no standardized methodology for collecting *in situ* spectroscopy data or metadata protocols exist.

Creating a standardized platform for sharing valid and reliable hyperspectral datasets requires synchronization with the development of the crucial *in situ* spectroscopy protocols and hyperspectral metadata standards. Datawarehousing can offer the solution to these interrelated requirements. Datwarehousing is a specialized datastore model that provides a single-point interface for data mining. It can be defined as a "complete intelligent data storage and information delivery or distribution solution enabling users to customize the flow of information through their organization" (Ouyang and Wang, 2008). It aggregates data from multiple databases and in varying formats to a single point of access for a large population of users. In the context of hyperspectral data, a standardized datawarehousing model would serve the remote sensing community by providing a central interface for hyperspectral data from a pool of databases and spectral libraries. Independent from hardware or operating system platforms, datawarehousing software can run on multiple servers for superior performance (Ponniah, 2001). By definition datawarehousing encourages collaboration between communities of users.

Hyperspectral datawarehousing can be modelled as a cascade of information, with its origin in hyperspectral datasets obtained both *in situ* and via airborne and satellite sensors. This data flow continues to the spectral libraries (for example the USGS Digital Spectral Library), databases (SPECCHIO, Hyperspectral.info), and any other online data repositories and access points either on private or public networks, where users upload and retrieve data. Beyond these access points the potential for hyperspectral datasets becomes limitless, as users can generate campaign-specific end-products, such moisture content analysis in limnology studies or fuse them with other remote sensing datasets such as LiDAR point clouds for vegetation canopy modelling and DEMs. Other common uses for the hyperspectral datasets include calibration and validation activities and retrospective analysis (Malthus and Shirinola, 2009). Figure 1 illustrates one possible datawarehousing model as applied to hyperspectral data, including the role of metadata during each state of data processing.

The datawarehousing model is more advanced and efficient than a database or network of data repositories, because it interfaces multiple computer servers on different platforms, irrespective of the geographic origin of the data or encoding format. Information access is rapid and secure, and can be latent or updated real-time. Datawarehousing also provides software and protocols for aggregating distributed digital data repositories, data transformation services, and front-end access for flexible reporting. It supports multi-dimensionality, facilitating fusion of different data schemas. Volume handling is superior to databases because modern data warehouses can facilitate up to 100 terabytes of more of data (Hadrian and Greenidge, 2009). Oracle, IBM, and Teradata are commercial vendors of datawarehousing software that has been used since the 1990s, primarily in the business intelligence community.

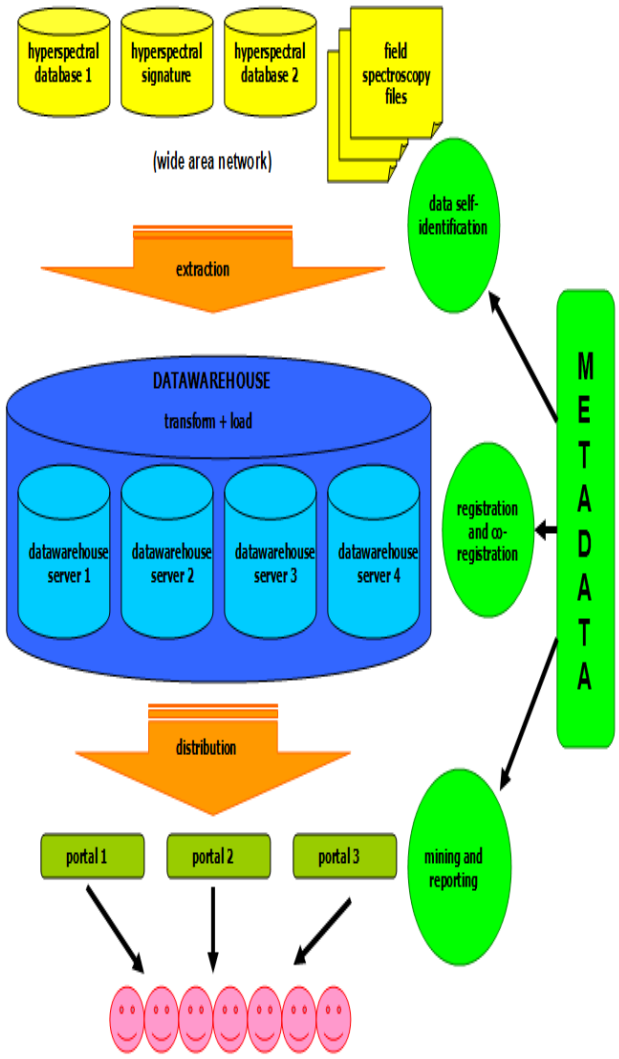

Figure 1: Data warehouse model for hyperspectral datasets

Common to all digital information architectures with multiple tiers of data storage and processing, the potential for error propagation and magnification increases with complexity (Król and Kukla, 2009). For this reason, the format and quality of data input to a hyperspectral datawarehouse has significant impact on the downstream end-products.

2. METADATA

2.1 Standards and protocols

Viewed within a holistic context, the importance of standardized data definitions, *in situ* data collection protocols, metadata ontologies, and interoperability of formats for hyperspectral data becomes apparent. Identifying user needs for a range of campaigns, -- among them vegetation, limnology, aquatic, geological, and atmospheric - is necessary before protocols can be standardized. HYRESSA (HYperspectral REmote sensing in Europe - specific Support Actions) has identified highly individual preferences among hyperspectral users in Europe, with each application group (vegetation, atmosphere, land, water) queried assigning varying importance to spectral, geometric, radiometric, and temporal parameters, and placing unique emphasis on spectral calibration quality and preferred observation time (Nieke et al, 2007).

Obstacles to harmonization of data definitions and protocols must also be addressed. Users in the European remote sensing community have identified weaknesses such as a lack of quality assurance and calibration information for sensors; no real capability to define accuracy or validation for data processing; a lack of agreed standards in data processing, and the need for more transparency on calibration processes (Reusen et al, 2007). These findings highlight the need to develop metadata protocols for ensuring data validity and reliability critical to all campaigns, as well as protocols that address the individual phenomenological, observation conditions, sensor specifications and requirements of individual campaigns.

OGC (Open Geospatial Consortium) and INSPIRE (Infrastructure for Spatial Information in the European Community) have both adopted architecture and data interoperability protocols for geospatial metadata based on EN ISO 19115 and EN ISO 19119. Although providing general guidelines, neither of these explicitly address the metadata requirements of hyperspectral field collection techniques, or the ontologies and data dependences required to model the complex interrelationships among the observed phenomena as data and metadata entities. Dependencies include the influence of environmental phenomena such as wind speed and cloud cover on the recorded spectrometer signal, or user-controlled viewing conditions such as sensor orientation and height above the target. More specific metadata schema for vegetation observations have been proposed (Pfitzner *et al* 2006 and Hüni *et al* 2007) but mostly on an *ad hoc* basis and do not yet encompass the full spectrum of common field specstroscopy campaigns.

2.2 Metadata sharing and interoperability

Metadata definitions must also provide the flexibility for users to create reliable datasets on an *ad hoc* basis, and to export their data to enterprise-scale databases and data warehouses for other users. Coupled with this is the need to preserve the integrity and quality of the original dataset and safeguard it from data corruption and information loss as it cascades through the extraction, transformation and loading processes in a datawarehouse. Lack of standardization in this area is symptomatic across the remote sensing community, with ongoing efforts to provide a solution. NASA's SPG (Earth Science Data Systems Standards Process Group) is investigating candidate standards for data access protocols and interoperability between OGC (Open Geospatial Consortium) Catalog Services for the Web and Web Coverage Services protocols and is working on a development of a NASA Earth Science Missions Data System reference architecture (Ullman and Enloe, 2010). The IEEE GRSS (Geoscience and Remote Sensing Society) Data Archiving And Distribution Technical Committee is currently exploring data archiving developing data availability methods and online discovery functionality for distributed datasets, as well as standards for information and systems (Rochon et al, 2010). OGC has also adopted the Sensor Observation Service standard and Sensor Model Language for modelling the interface between *in situ* sensors, dynamic remote sensors, and sensor networks, and for enabling data retrieval either in real-time or through data archives.

With the emergence of online data archives and the proliferation of web-interfacing applications, it is necessary to use a metadata encoding format that is sufficiently robust to capture the required data elements and provide hierarchical information, while being popular, compatible with spectroradiometric software (ViewSpecPro, SpecWin, SVC), with the capability for efficient transport across wide-area networks for integration into

external archives. An XML (Extensible Markup Language)-based exchange format for spectroradiometric metadata has been proposed because it facilitates searching and selection, it is human and machine readable, platform independent, convertible to other formats and allows quick assessment of suitability for other research products (Malthus and Shironola, 2009). The XML format is easily accommodated in a datawarehouse. XML is self-descriptive with extensibility features (Mahboubi and Darmont, 2010), facilitating integration with multi-dimensional remote sensing data sets. One of its greatest strengths is platform independence, and a framework for XML-based data interchange is espoused in the Common Warehouse Metamodel, which includes XMI (XML Metadata Interchange) standards for datawarehouses (Mangisengi *et al*, 2001 and Torlone, 2009).

3. SPECTRAL LIBRARIES, DATABASES, DATAWAREHOUSES, AND CLOUD COMPUTING

Identifying the best direction to take for hyperspectral data and metadata protocols for archiving and sharing requires a general overview of the facilities currently in place within the remote sensing community and IT architectures that can offer potential solutions.

3.1 Spectral Libraries

Publicly available spectral libraries such as NASA's ASTER Spectral Library and the USGS Digital Spectral Library offer downloadable data for a broad range of hyperspectral signatures in the form of image files of plots and descriptive text for each signature. Although comprehensive and easily navigable, these libraries are the most static of the data archiving models and do not support the hierarchical dependencies of metadata components to the hyperspectral data.

3.2 Databases

Hyperspectral databases created in the last few years include SPECCHIO and Hyperspectral.info. SPECCHIO offers more sophisticated capabilities for storing, retrieving, and analyzing hyperspectral data than a spectral library. SPECCHIO is a MySQL database with a Java client application for automated metadata retrieval, metadata editing and instrumentation administration, as well as reports, with support for multiple spectradiometer file formats (Hüni and Kneubühler, 2010). SPECCHIO provides efficient storing and reporting mechanisms for hyperspectral data and metadata input by its users.

There are limitations and restrictions imposed by databases as a platform for data storage and sharing in the broader remote sensing community. These include: the requirement for users to install an instance of the database in order to query, import, or export data; metadata relationships, data encoding, and import/export utilities are heavily reliant on the architecture and embedded functions of the database; large-volume transactions are restricted by the operating system and speed and bandwidth of network connections between the databases; scalability is depended upon the computer where a given database instance is installed, as well as the operating system and infrastructure resources. It is a point-to-point method of sharing data between instances of the database, rather than being a single-point access for multiple users, with no mechanism for tracking transactions and updates across multiple instances, thereby compromising quality assurance and disaster recovery efforts. The database model does not address one of the central issues to sharing of hyperspectral data remote sensing community, which is the need for a common data exchange protocol across networks and platforms with quality assurance.

3.3 Datawarehousing

Earlier discussion detailed the suitability of the datawarehousing model for the large-scale archiving sharing, and mining of hyperspectral data and metadata from a pool of sources. It enables *ad hoc* querying for end-users, aggregation and merging of large volumes of data from a variety of platforms and file formats, utilities for inconsistency reconciliations and intensive use of metadata (Peter and Greenidge, 2009). Its other strengths include: all transactions are traceable, thereby providing a basis for quality control; responsibility for physical data ownership does not rest with a single user or group once the data has been warehoused, enabling disaster recovery should a dataset be corrupted or lost; datawarehouses rely on metadata mining as a basic function of data aggregation, and therefore could be the impetus for standardizing quality control protocols for hyperspectral metadata. Processing loads are reduced by separating online transactions from analytical processing. The datawarehousing model has been successfully implemented in many business intelligence communities and is a sufficiently mature technology for adoption by hyperspectral data users.

3.4 Cloud Computing

Cloud computing is a large networked environment of shared software, databases, and other computing resources from a variety of architectures. The focus is on providing services to users who are not required to have a vested interest in the implementation or the management of the data (Hartig, 2009). IBM and Google are examples of this scalable data centre infrastructure, where the user is aware only of the services provided by the cloud, and not the back end servers, databases or data exchange mechanisms (Redkar, 2009). The OCC (Open Cloud Consortium) exists to develop standards and best practices, but as the large-scale implantation of the cloud computing concept is not yet fully realized, protocols for data exchange are unclear. Since many of the hyperspectral data users in the remote sensing community are also the creators and owners of the data, they may find difficulty in assessing the validity and reliability of a hyperspectral dataset and its metadata as it moves through the cloud. Because of limited standardization and no mechanism for quality assurance, cloud computing at this time is not a suitable candidate for data sharing by hyperspectral data users in the remote sensing community

4. TOWARDS DATA WAREHOUSING

There exist data exchange networks among the remote sensing community that are evolving towards the datawarehousing model. Among these are:

- EOSDIS (Earth Observing System Data Information System), a network of data centers, metadata repository, middleware providers and directory services for NASA's Earth science data (Kuo, 2010)
- GALEON (Geo-interface for Atmosphere, Land, Earth, and Ocean netCDF) Interoperability Experiment, an OGC initiative to specify standard interfaces for interoperability between data sets used by GIS communities and those used by Earth scientists (Domenico *et al*, 2006)
- TERN (Terrestrial Ecosystem Research Network) an Australian initiative to coordinate a national data network with quality assured observational data from the terrestrial domain
- NOAA Enterprise Archive Access Tool (NEAAT), which allows archive managers to supply data to users without modifying archiving systems or data presentation (Rank *et al*, 2010)

These data exchange initiatives demonstrate both the necessity and feasibility of defining and streamlining protocols and IT infrastructure for creating a new generation of advanced data repositories with a centralized interface for a broad range of users.

5. CONCLUSIONS

Much potential exists for adapting and improving current geospatial metadata standards for the unique requirements of the hyperspectral remote sensing community. Datawarehousing remains the best option for the management of hyperspectral data and metadata. Before this can occur on a large scale, user needs for quality assurance must be formally identified, as well as standardized protocol for hyperspectral metadata storage and data exchange. Steps towards greater collaboration in the hyperspectral remote sensing community include:

- determining the metadata requirements for field spectroscopy for the full range of hyperspectral campaigns
- standardizing field spectroscopy protocols for accuracy and consistency
- establishing file exchange protocols allowing flexibility to capture hyperspectral metadata and enabling fusion with other remote sensing products
- aligning data sharing and mining within the remote sensing community with datawarehousing practices

A collaborative and innovative spirit can bring great benefits to international efforts for providing the data sharing capabilities and quality control tracking for the hyperspectral remote sensing community.

REFERENCES

B. Domenico, J. Caron, E. Davis, S. Nativi, L. Bigagli, "GALEON: Standards-based Web Services for Interoperability among Earth Sciences Data Systems" *IEEE International Conference on Geoscience and Remote Sensing Symposium,* August 2006.

M.J. Duggin, "Factors limiting the discrimnation and quantification of terrestrial features using remotely sensed radiance", *, International Journal of Remote Sensing*, 6: 1, 3-27, 1985.

D. Krol and G. Kukla, "Quantitiave Analysis of the Error Propagation Phenomenon in Distributed Information Systems", *First Asian Conference on Intelligent Information and Database Systems,* pp.202-207, April 2009.

K. S. Kuo, "Experiences with NASA Earth Science Data Information Systems and Suggestions for Improvements from a Scientist User Perspective", *IEEE International Conference on Geoscience and Remote Sensing Symposium,* July 2010.

K. Hartig, "What is Cloud Computing", *Cloud Computing Journal,* December 13 2009, http://cloudcomputing.sys-con.com/node/579826

A. Hüni, M. Kneubühler, "The Spectral Database SPECCHIO In Support of Cal/Val Activities", ESA WORKSHOP, 2010.

A. Hüni , J. Nieke, J. Schopfer, M. Kneubühler and K. I. Itten, "Metadata of Spectral Data Collections" *Proceedings of the 5th EARSeL Workshop on Imaging Spectroscopy*. Belgium, April 2007.

H. Mahboubi and J. Darmont, "Optimization in XML Data Warehouses", pp 232-253 *E-Strategies for Resource Management Systems: Planning and Implementation*, E. Alkhalifa (Ed.), University of Bahrain, 2010.

T. Malthus and A. Shirinola, "An XML-based format of exchange of spectroradiometry data", EARSeL Imaging Spectroscopy SIG, Tel Aviv, March 2009.

O. Mangisengi, J. Huber, C. Hawel, W. Essmayr, "A Framework for Supporting Interoperability of Data warehouse Islands Using XML", Lecture Notes in Computer Science, 2001 *Data Warehousing and Knowledge Discovery*, Volume 2114, 328-338, 2001.

E.J. Milton, "Principles of Field Spectroscopy", *International Journal of Remote Sensing*, 8:12, pp1807-1827, 1987.

J. Nieke et al., "User-driven requirements of the European Hyperspectral Remote Sensing Community", *HYRESSA Workshop*, Davos, Swtizerland, March 2007.

H. Ouyang, and John Wang, "Data Warehouse Software", *Encyclopedia of Information Technology Curriculum Integration,,* Lawrence A. Tomei (ed.), Robert Morris University, 2008.

H. Peter and C. Greenidge, "Aligning the Warehouse and the Web" pp 18-24, *Encyclopedia of Data Warehousing and Mining, Second Edition.* J. Wang (Ed.), Montclair State University, 2009.

K. Pfitzner, A Bollhöfer, and G. Carr, "A Standard Design for Collecting Vegetation Reference Spectra: Implementation and Implications for Data Sharing" *Spatial Science*, 52:2, 79-92, December 2006.

P. Ponniah, "Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals", John Wily & Sons, Inc, 2001.

R. Rank, S. McCormick, C. Cremidis, "NOAA Enterprise Archive Access Tool (NEAAT): Accelerated Application Development (XAD)", *IEEE International Conference on Geoscience and Remote Sensing Symposium, 2330-2332, July 2010.*

T. Redkar, Windows Azure Platform, pp 1-51, Springer-Verlag, 2009 .

I. Reusen et al., "Towards an improved access to hyperspectral data across Europe", *ISIS meeting,* Hilo, 2007.

G. L. Rochon, H. Ramapriyan, E. F. Stocker, R. Duerr, R. Rank, S. Nativi, "Advances in Spatial Data Infrastructure, Acquisition, Analysis, Archiving & Dissemination", *IEEE International Conference on Geoscience and Remote Sensing Symposium,* 1980-1983, July 2010.

R. Torlone, "Encyclopedia of Database Systems", Part 9, p. 1560-1564, Springer Science + Business Media, LLC, 2009.

R. Ullman and Y. Enloe, "NASA's Standards Process for Earth Science Data Systems", *IEEE International Conference on Geoscience and Remote Sensing Symposium,* 4232 - 4235, July 2010.