ENHANCED DATA DISCOVERABILITY FOR *IN SITU* HYPERSPECTRAL DATASETS

B. Rasaiah[a,*], C. Bellman[b], R.D. Hewson[c], S. D. Jones[d], T. J. Malthus[e]

[a]University Corporation for Atmospheric Research, Tuscaloosa, AL 35401, USA – rasaiah@ucar.edu

[b,c,d] Centre for Remote Sensing, RMIT University Melbourne, VIC 3001, Australia – chris.bellman@rmit.edu.au, simon.jones@rmit.edu.au

[e]CSIRO Land and Water, Canberra, ACT 2601, Australia – tim.malthus@csiro.au

Commission IV, Working Group IV/4

KEYWORDS: Databases, Data mining, Hyperspectral, Metadata, Calibration, Data Quality, Interoperability, Standards

ABSTRACT

Field spectroscopic metadata is a central component in the quality assurance, reliability, and discoverability of hyperspectral data and the products derived from it. Cataloguing, mining, and interoperability of these datasets rely upon the robustness of metadata protocols for field spectroscopy, and on the software architecture to support the exchange of these datasets. Currently no standard for *in situ* spectroscopy data or metadata protocols exist. This inhibits the effective sharing of growing volumes of *in situ* spectroscopy datasets, to exploit the benefits of integrating with the evolving range of data sharing platforms. A core metadataset for field spectroscopy was introduced by Rasaiah *et al*., (2011-2015) with extended support for specific applications. This paper presents a prototype model for an OGC and ISO compliant platform-independent metadata discovery service aligned to the specific requirements of field spectroscopy. In this study, a proof-of-concept metadata catalogue has been described and deployed in a cloud-based architecture as a demonstration of an operationalized field spectroscopy metadata standard and web-based discovery service.

1. INTRODUCTION

1.1 Background

Hyperspectral datasets are dependent upon their associated metadata for ensuring their quality, reliability, and discoverability. To varying degrees, *in situ* hyperspectral datasets are thus uniformly sensitive to the integrity of their metadata. There remains no standardized methodology for documentation of field spectroscopy data or metadata (Rasaiah *et al*., 2011-2015) or for the exchange and discoverability of such datasets. The need for a standardized methodology for collecting, storing, sharing – and assuring the quality of field spectroscopy metadata has increased with the emergence of data sharing initiatives such as NASA's EOSDIS (Earth Science Data and Information System), the LTER (Long Term Ecological Research) network, the Australian Terrestrial Ecosystem Research Network (TERN), SpecNet (Gamon, 2006) and several smaller *ad hoc* spectral libraries and databases created by remote sensing communities internationally.

The absence of a formal standard prohibits efficient and viable intercomparison and fusibility of datasets generated from quantitative field observations (Jung *et al*., 2012). Additionally, the absence of a data exchange and metadata standard inhibits discoverability of field spectroscopy datasets. This applies to data and metadata generated for discipline-agnostic information sharing systems and for discipline-specific databases (Ben-Dor *et al*., 2015). In the context of a supporting hardware and software architecture, effective dissemination and exchange of *in situ* hyperspectral datasets across data sharing platforms is achievable when 1) metadata is comprehensive and high quality and 2) a metadata discovery service exists to expose datasets to users.

1.2 *In Situ* Hyperspectral Metadata

The generation of a core metadataset for field spectroscopy as a foundation for a metadata standard, was introduced and developed by Rasaiah *et al.* (2011-2015). A superior quality metadataset can describe a broad range of observed field data, including environmental conditions, properties of the target being viewed, sensor specifications / calibration activities, and illumination conditions / viewing geometries. Such metadata are vital because they are all influencing factors that affect standardized measurements (Pfitzner *et al.*, 2006). Metadata can also serve to describe and quantify errors introduced into the spectra, and be a tool for potentially mitigating these errors. Metadata quality parameters for field spectroscopy were presented by Rasaiah *et al*. (2015) as a set of qualitative and quantitative measures that provide the data user with information that allows them to decide on the suitability of the metadata and associated dataset for a particular purpose. Metadata in general, can serve numerous other functions (e.g., identification, discovery, administration, version control) built on a framework of specific categories of defined metadata elements (Higgins, 2007). A comprehensive suite of metadata parameters aligned to ISO 19115 (ISO, 2014) and OGC standards (OGC, 2015) for geospatial metadata, was introduced in Rasaiah *et al.* (2015*b*), encompassing the critical field spectroscopy metadataset in addition to quality parameters and dataset parameters.

Enabling metadata to fulfill its potential to the broad range of data users requires exposure through a metadata discovery service (IBM & BEA, 2004; Hauch *et al*., 2005; Vaughn, 2011; Oak Ridge National Laboratory, 2015; Patroumpas *et al*., 2015). Additionally, the metadata must be sufficiently comprehensive and complete to empower data users to make informed decisions about the most suitable dataset for them.

* Corresponding author

1.3 Application-specific metadata

In addition to a core metadataset that is critical to all campaigns (Rasaiah *et al*., 2014) an extended metadataset is required to support specific applications of the data. As there is no metadata standard for field spectroscopy, there is none for specific applications, such as for agriculture, soil studies, or geological surveys.

Discoverability, in the context of metadata, can be defined to be those metadata elements and supporting discovery services that enable exposure of a dataset (and its metadata) to users searching for it (Fegraus *et al*., 2005; Mann, 2006; Vaughn, 2011). The richer and larger the metadataset, the greater its potential for discovery, establishing ontological relationships with other metadatasets, and the more empowered data users are to determine whether the underlying dataset is suitable for a given purpose. Discoverability of metadata is possible when the metadata is consistent with the taxonomies, syntax, and metadata granularity (the specificity or level of detail at which each metadata field is expressed) unique to a given application. For example, a benthic-specific metadataset for field spectroscopy (Rasaiah *et al.*, 2015b) must document environmental factors and additional logistics both above the water surface and below, including, tide conditions, wave attenuation, turbidity, and a modified and attenuated light field that are not generally a consideration for terrestrial campaigns.

Establishing standardized guidelines and formats for soil and geological field spectroscopy metadata is particularly relevant for *in-situ* hyperspectral surveys. It is potentially wasteful if the large number field spectral measurements of soil and rock outcrops, undertaken by multiple government agencies and private industry, cannot be effectively compared and accumulated, because of the lack of meaningful metadata.

As with several environmental applications, geoscience field spectroscopy metadata includes the standard information regarding the spectrometer instrument, calibration methodology, measurement, geometry, and illumination conditions. However in addition, specific geoscience metadata requires such information on the nature and extent of lichen / moss within the measurement field of view, the weathered or broken nature of its surface, and presence of obvious mineralogical and texture features. *In situ* soil measurement metadata should also include whether it includes a disturbed or undisturbed surface/crust, the soil horizon measured, its soil classification, and ideally its moisture content. The incorporation of subsequent laboratory analysis on collected geoscience samples as metadata, such as their geochemistry or mineralogy, is also highly relevant information from hyperspectral geoscience surveys.

As key stakeholders of the data, field spectroscopy scientists have a vested interest in the development of a metadata standard and metadata discovery services most suitable to their needs as both metadata data creators and users of this data.

1.4 Data formats

An operationalized metadata standard and its complimentary data discovery service must accommodate a variety of data sources and formats. Digital-format metadata can include automatically generated metadata from field spectroradiometer measurements. In the approach presented here, encoded instrument and signal properties information can be incorporated within their native files and later exported as metadata to a local or central database or other data repository. For example, source files originating from an instrument can consist of metadata stored in the header of binary files, which then can be extracted and loaded as individual records in a database or, alternatively, encoded in an XML file for storage on a file server. The strengths and weaknesses of different data encoding formats becomes a valuable debate when operationalizing an *in situ* hyperspectral metadata schema. The digital format of metadata format is a factor in its potential for large-scale archiving, mining, and sharing across platforms. For example, datawarehousing models can function in support of a given metadataset when external metadatasets are aligned to its metadata schemas (Rasaiah *et al.*, 2011).

However, it is not possible to implement or mandate the adoption of a single data encoding format for field spectroscopy metadata. Such standardization is precluded by the wide variety of data sources and file format preferences of data creators and users within the remote sensing community. A practical approach is to support discovery of metadata through software services that are flexible and sufficiently robust to expose the maximum volume of field spectroscopy datasets to data users while accommodating the diversity of formats and data platforms from which the data originates.

1.4 Existing hyperspectral data repositories

There are international initiatives to share geospatial data online. These include TERN AusCover Data Discovery Portal (http://portal.tern.org.au), NASA's EOSDIS WorldView and Data Portal (http://earthdata.nasa.gov/labs/worldview/), and GEO GEOSS (Group on Earth Observations Global Earth Observation System of Systems) Portal (http://www.geoportal.org). Their architecture is a mixture of metadata registries, databases, datawarehouses, and cloud platforms. These systems were built with the objective of providing reference datasets and products for researchers and the public, enabling sharing of datasets in a quality controlled manner, and facilitating the distribution of datasets and their metadata through a single point of access. Among those systems that do catalogue *in situ* hyperspectral datasets, the metadata is not sufficiently comprehensive to align with identified community needs (Rasaiah *et al*., 2014). Likewise, there are online metadata catalogues customized for geospatial datasets (Oak Ridge National Laboratory, 2015; Patroumpas *et al*., 2015) but these do not accommodate the metadata requirements for field spectroscopy datasets.

2. A SOFTWARE ARCHITECTURE FOR FIELD SPECTROSCOPY METADATA DISCOVERY

Figure 1 presents the cloud-based software architecture for the field spectroscopy metadata discovery service implemented as a proof-of-concept. This software system exists within the Amazon Web Services cloud with databases deployed as a Relational Database Service (RDS) and application servers deployed on an Elastic Compute (EC2) platform.

Its core components are:

- Distributed data sources in different data formats (text files, web feature service, database) (multiple RDS instances)
- Field spectroscopy metadata database (Postgres 9.4)
- GeoServer, a Java-based geospatial data server (Windows Server 2012)

- Metadata catalogue, a Java-based application for metadata discovery (Windows Server 2012)
- Web server and web services for publishing the catalogue online (Windows Server 2012)

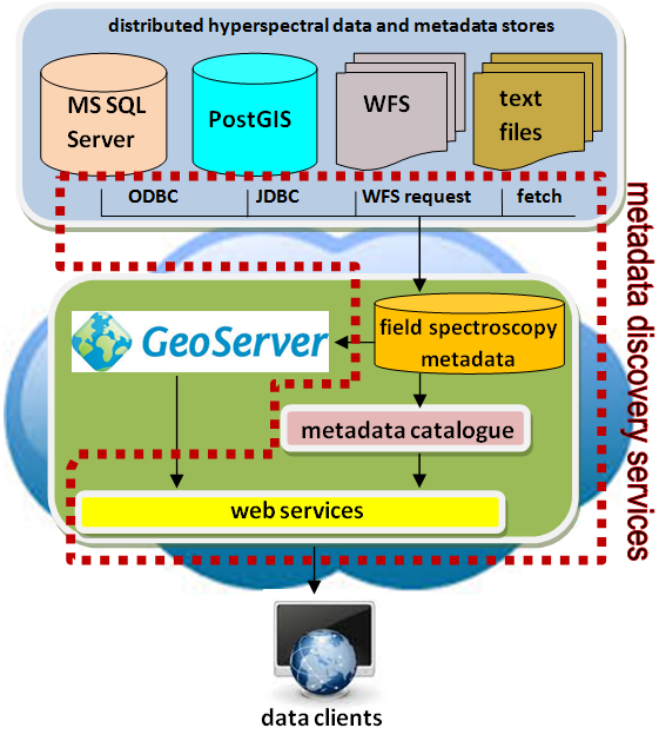

Figure 1 A software architecture describing a metadata discovery service for field spectroscopy datasets

Data access is managed by GeoServer through ODBC, JDBC, and WFS requests to data servers within the AWS cloud. Metadata is stored in the field spectroscopy metadata database (Postgres 9.4), and accessed by an online metadata catalogue for discovery. The catalogue is aligned to the ISO 11179 standard for metadata registries (ISO, 2009, OGC, 2007).

The main advantage of this architecture is its focus on discovery of field spectroscopy data, rather than centralization of data and software resources. This enables deployment of the catalogue on any platform, with no reliance on the location of data sources. Additionally, source data does not need to be replicated to a central data store in order for it to be mined and catalogued.

Figure 2 is an example screen image of the developed prototype online metadata catalogue within the GeoServer architecture.

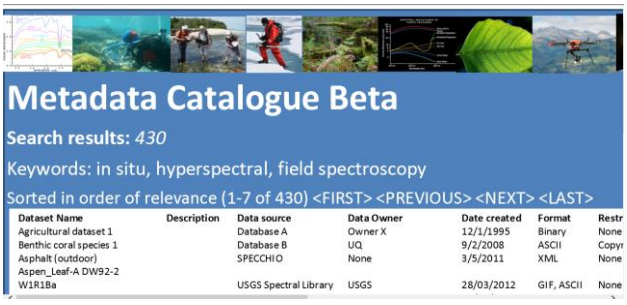

Figure 2 A screen image of the prototype metadata catalogue with search results

Its main features include a search function for metadata in distributed data stores, and comprehensive metadata associated with each spectral measurement dataset. Search results will include a metadata snapshot with dataset name, data source, owner, date of creation, format, access rights, and provides the data user with options for viewing a comprehensive metadataset for a give field-spectroscopy dataset (e.g. instrument properties, viewing geometry, illumination information, environmental conditions). Additionally, a metadata standard compliance report is available to data users, reporting compliance with the core metadataset, its application-specific extensions, and relevant geospatial metadata standards (Figure 3). The uniqueness of this catalogue is that it is the only one in existence that accommodates identified needs and preferences of field spectroscopy data users.

This model will aid field spectroscopy data users in searching for and identifying datasets distributing globally, rather than divesting effort and time to search the data sources individually. It will aid the general community in enriching their data searches with field spectroscopy datasets not otherwise accessible through a single discovery service. This metadata discovery service presents an evolution for field spectroscopy data sharing in that it complements the earlier research by the authors in identifying metadata requirements for data creators with a newly developed model for data users' accessibility to these datasets.

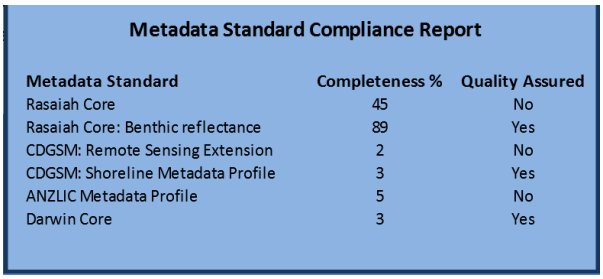

Figure 3 A screen image of a metadata standard compliance report for a selected field spectroscopy metadataset

3. CONCLUSIONS

As the volume of hyperspectral datasets grows together with the diversity of data sharing platforms, it is vital that data and metadata discovery services are adopted and aligned with user requirements. A platform-agnostic metadata discovery service for *in situ* hyperspectral datasets is ideal for interoperability with non-uniform system architectures. The metadata discovery service presented here is unique in its utility for field spectroscopy data users because it aligns with their identified metadata requirements, for assessing datasets appropriate for a given application. Moving forward, coupling this metadata discovery service with operationalized field spectroscopy metadata schemas will ensure increased access to and quality assurance of *in situ* hyperspectral datasets.

4. REFERENCES

Ben-Dor, E.; Ong, C.; Lau, I. C, 2015. Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma*, 245, 112-124.

Fegraus, E. H., Andelman, S., Jones, M. B.; Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: an introduction to Ecological Metadata Language (EML) and principles for metadata creation. *Bulletin of the Ecological Society of America*, 86(3), 158-168.

Gamon, J. A.; Rahman, A. F.; Dungan, J. L.; Schildhauer, M.; Huemmrich, K. F., 2006. Spectral Network (SpecNet)—What is it and why do we need it? *Remote Sensing of Environment*, 103, 227-235.

Hauch, R.; Miller, A.; Cardwell, R., 2005. Information intelligence: metadata for information discovery, access, and integration. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (pp. 793-798).

Higgins, S. What are Metadata Standards. 2007. Available online:http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards (accessed on 7 January 2014).

IBM & BEA, 2004. Enterprise Metadata Discovery. Available online: http://public.dhe.ibm.com/software/dw/library/j-emd/EnterpriseMetadataDiscovery_v0.12.pdf (accessed 10 December 2015).

ISO, 2009. ISO/IEC 11179-1:2004. Available online: http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35343 (accessed 10 December 2015).

ISO, 2014. ISO 19115-1:2014. Available online: http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798 (accessed 10 December 2015).

Jung, A.; Götze, C. Glässer, C., 2012. Overview of Experimental Setups in Spectroscopic Laboratory Measurements–the SpecTour Project. *Photogrammetrie-Fernerkundung-Geoinformation* 4, 433-442.

Mann, T., 2006. The Changing Nature of the Catalog and Its Integration with Other Discovery Tools, Final Report, March 17, 2006, Prepared for the Library of Congress by Karen Calhoun: A Critical Review. AFSCME, Local 2910.

Oak Ridge National Laboratory, 2015. Mercury: Distributed Metadata Management, Data Discovery and Access System. Available online: http://mercury.ornl.gov/?q=overview (accessed 10 December 2015).

OGC, 2007. Catalogue Service. Available online: http://www.opengeospatial.org/standards/cat (accessed 10 December 2015).

OGC, 2015. OGC Standards. Available online: http://www.opengeospatial.org/docs/is (accessed 10 December 2015).

Patroumpas, K.; Georgomanolis, N.; Stratiotis, T.; Alexakis, M.; Athanasiou, S., 2015. Exposing INSPIRE on the Semantic Web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35, 53-62.

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

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C., 2011*a*. Building better hyperspectral datasets: The fundamental role of metadata protocols in hyperspectral field campaigns. In *Proceedings of the Surveying & Spatial Sciences Conference*, Wellington, New Zealand, 21–25 November 2011.

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C., 2011*b*. Designing a robust hyperspectral dataset: The fundamental role of metadata protocols in hyperspectral field campaigns. *In Proceedings of the GSR 1 Research Symposium*, Melbourne, Australia, 28–30 November 2011.

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C., 2012. Critical metadata protocols in hyperspectral field campaigns for building robust hyperspectral datasets. In *Proceedings of the XXII ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences Congress*, Melbourne, Australia, 26 August–1 September 2012.

Rasaiah, B.A.; Jones, S.D.; Bellman, C.; Malthus, T.J., 2014. Critical Metadata for Spectroscopy Field Campaign. *Remote Sens.* **2014**, *6*, 3662–3680.

Rasaiah, B.A.; Jones, S.D.; Bellman, C.; Malthus, T.J.; Hueni, A., 2015*a*. Assessing Field Spectroscopy Metadata Quality. *Remote Sens.* **2015**, *7*, 4499–4526.

Rasaiah, B.A.; Bellman, C.; Jones, S.D.; Malthus, T.J.; Roelfsema, C., 2015*b*. Towards an Interoperable Field Spectroscopy Metadata Standard with Extended Support for Marine Specific Applications. *Remote Sens.* **2015**, *7*, 15668-15701.

Vaughan, J., 2011. EBSCO discovery services. Library Technology Reports, 47(1), 30-38.

Vaughan, J. (2011). Web scale discovery what and why?. Library Technology Reports, 47(1), 5-11.