International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XXXIX-B2, 2012
XXII ISPRS Congress, 25 August – 01 September 2012, Melbourne, Australia

Critical Metadata Protocols in Hyperspectral Field Campaigns for Building Robust Hyperspectral Datasets

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

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

[c]CSIRO Land and Water, Black Mountain, ACT 2601, Australia – tim.malthus@csiro.au

Commission II, Working Group II /IV

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

ABSTRACT

Field spectroscopic metadata is a central component in the quality and reliability of hyperspectral data and the products derived from it. The impact of the quantity and format of metadata created at this fundamental stage of hyperspectral research is amplified as hyperspectral data exchange becomes prolific in the international remote sensing community. Cataloguing, mining, and interoperability of these datasets rely upon the robustness of metadata protocols for field spectroscopy. Currently no standardized methodology for collecting *in situ* spectroscopy data or metadata protocols exist. This paper presents initial results of an international expert panel survey investigating metadata protocols in field spectroscopy. Field measurement methods, data representativeness and their expression as metadata entities are examined across a range of campaigns. Consensus between expert groups and variance in agreement on criticality are also investigated. The survey is part of a doctoral research project to investigate approaches to a coordinated evolution of hyperspectral metadata protocols, field spectroscopic methods and data exchange standards within the hyperspectral remote sensing community.

1 INTRODUCTION

Hyperspectral datasets are dependent upon their associated metadata for ensuring their quality, reliability, and longevity. To varying degrees, *in situ* hyperspectral datasets are uniformly sensitive to the integrity of their metadata. A superior quality metadataset can describe a broad range of observed field data, including environmental conditions, properties of the target being viewed, sensor specifications and calibration activities, and illumination and viewing geometry, among others. Such metadata are vital because they are all influencing factors that affect standardized measurements (Pfitzner *et al.*, 2006). Metadata can serve numerous other functions including describing and quantifying errors introduced into the spectra, and as tool for potentially mitigating these errors. The logistics of collecting sufficiently reliable metadata, as well as the requisite volume of metadata, is a central consideration for creating a standardized methodology for defining and storing metadata that are also closely aligned to *in situ* data collection practices adopted by remote sensing research communities around the world. There is an urgency in acquiring continuous high quality spectroscopy data to solve problems in the Earth sciences and to inform users and stakeholders of the value of such data (Schaepman *et al*., 2009). Weaknesses in hyperspectral data collection and sharing have been identified by users in the European remote sensing community and include 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). The need for a standardized methodology for collecting *in situ* hyperspectral metadata has increased with the emergence of data sharing initiatives such as US LTER (Long Term Ecological Research) network, Australian Terrestrial Ecosystem Research Network (TERN), SpecNet, and some of the smaller *ad hoc* spectral libraries frequently created by remote sensing communities internationally. Currently no such methodology for collecting *in situ* spectroscopy data or metadata protocols exist. A fundamental step in designing a protocol for *in situ* metadata collection requires that the remote sensing community identify and define user needs for quality assurance, and a standardized protocol for hyperspectral metadata storage and data exchange.

Quality assurance of field datasets necessitates oversight and standardization, both at local, national, and international scales for creating robust metadata protocols for field spectroscopy. This allows for the most efficient and successful cataloguing, mining, and interoperability of these datasets. Addressing user needs for quality assurance requires some measure of consensus from the remote sensing community on defining critical metadata thresholds for creating high quality and long-term hyperspectral dataset.

These concerns prompt the need for creating a mechanism to determine the metadata fields that are critical for valid and reliable field spectroscopic datasets. Such a mechanism must establish the required metadata with enough integrity to generate datasets for long-term cataloguing and data warehousing across a range of campaigns. Ideally it should also be easily accessible to an international audience with a broad spectrum of expertise.

2 CONSULTING THE EXPERTS

An online and duplicate hardcopy survey was launched in early 2011 in the form of a user-needs analysis for field spectroscopy metadata. The purpose of the survey was to determine, based on the input of experts in the field, the metadata fields that are critical for creating valid and reliable field spectroscopic datasets, with enough integrity to generate datasets for long-

term cataloguing and data warehousing across a range of campaigns. The metadata fields (approximately 200 in the survey) are closely associated with common *ad hoc* field spectroscopy protocols practiced by remote sensing communities around the world. The audience was an international panel of scientists with expertise in *in situ* hyperspectral remote sensing, who were asked to respond on an anonymous basis. Each participant assessed the criticality of several categories of metadata fields, and could propose additional metadata fields that they believed could enhance the quality of a hyperspectral dataset generated in the field. Open-ended comments were available throughout the survey for further input in each metadata category.

Respondents had the option of participating in the categories of their choice, and were also asked to nominate themselves as experts in one or more areas of field spectroscopy application. This self-nomination of area of expertise did not in any way limit the categories available to each participant, and primarily served the purpose of informing correlational analysis between a participant's area of expertise and their assessment of metadata criticality. Metadata fields presented in the survey could be given one and only one ranking:

- 'critical' (required metadata field for a field spectroscopy campaign; without this data the validity and integrity of the associated spectroscopy data is fundamentally compromised)
- 'useful' (not required, but enhances the overall value of the campaign)
- 'not useful now but has legacy potential' (not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product
- 'not applicable' (this metadata is not relevant to this campaign)

3 SURVEY RESULTS & DISCUSSION

The survey was completed by 90 participants, with many comments provided in each category of expertise and opinions on metadata and *in situ* protocols.

Figure 3.1 illustrates the areas of expertise identified by the survey respondents. Areas of spectroscopy research beyond this scope, as stated by the respondents, include atmospheric study, calibration and validation activities for airborne sensors, and wetlands and peatlands research.

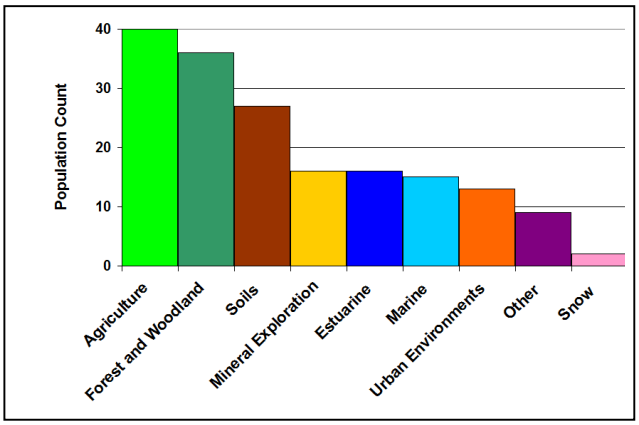

Figure 3.1 Areas of expertise self-nominated by survey respondents *(n=90)*

Variance in ranking of criticality varied for each metadata category. The ordinal criticality rankings (critical/useful/legacy potential/NA) were standardized to numerical values (ranging from 0 for 'N/A' to 3 for 'critical' to permit statistical analysis of variance. Figure 3.2 depicts the frequency of ranking for a subset of metadata fields in the 'instrument' metadata category, responded to by 79 of the scientists. Assignment of 'critical' to a given metadata field ranges from less than 20% for 'detector types' to 90% for 'spectral bandwidth'. The latter field is highlighted as the only one with no 'N/A' or 'legacy potential' ranking, suggesting that is a fundamentally crucial metadata field and warrants inclusion in all *in situ* metadata protocols.

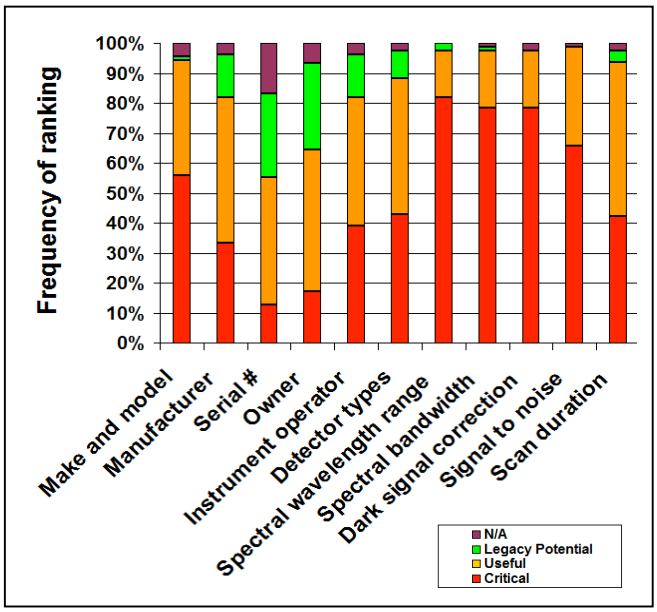

Figure 3.2 Frequency of criticality ranking for 'instrument' metadata (*n=79*)

Some of the variation may be accounted for by the choice of instrument listed by the participants of the survey; more than twenty different instruments were identified as being commonly used for *in situ* campaigns, with the top four being ASD models, Ocean Optics USB2000, SVC GER1500, and TRiOS Ramses, in addition to others designed in-house. The unique technical aspects of each instrument may have an influence on the particular metadata fields that an operator chooses to include in their metadataset.

Figure 3.3 depicts the frequency of ranking for marine 'substratum target' metadata, which was responded to by a smaller population of scientists (40), mostly those with a background in marine campaigns. For all fields in this category, there is a more consistent proportion between the four available rankings, and further investigation revealed that most of the 'N/A' rankings were assigned by respondents whose primary expertise lay outside of the marine sciences.

Among all the metadata fields presented throughout the survey, from generic campaign to specialized campaign categories, every field was designated as 'critical' by at least a small subset of respondents, regardless of their area of expertise.

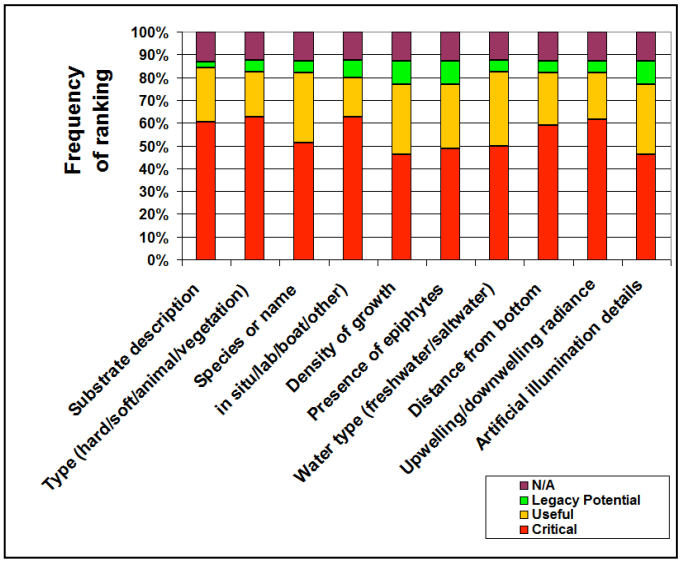

Figure 3.3 Frequency of criticality ranking for 'substratum target ' metadata (*n=40*)

The results also indicate that group membership has an impact on the degree of variance in response. An example among the marine and estuarine scientists demonstrates the variability in their responses from the other expert groups, with group differences between the two being amplified in the marine-specific metadata categories.

In the viewing geometry metadata category, shown in Figure 3.4, group means range between 'useful' and 'critical' for both the marine and none-marine scientists. The non-marine scientists rate the first three metadata fields 'distance from target', 'distance from ground' and 'area of target in FOV' as 'critical' more often than the marine group.

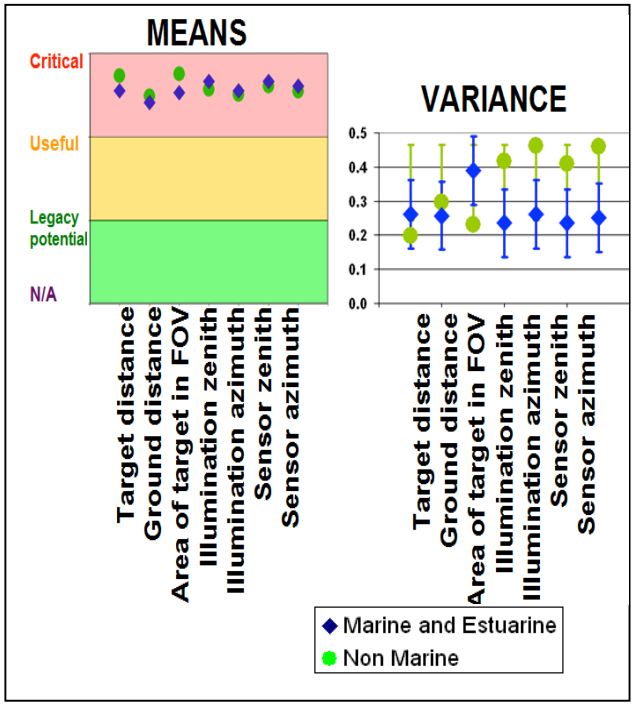

Figure 3.4 Group means and variances in 'viewing geometry' metadata category
(Marine and Estuarine n_1=*18,* Non Marine n_2=*49)*

There is more agreement in the remaining metadata fields relating to solar and sensor angles, suggesting that regardless of a respondent's area of expertise, metadata relating directly to radiative transfer modelling is of equal importance to all campaigns. Variance in criticality ranking for viewing geometry metadata is generally consistently higher among non-marine scientists, implying that there exists greater consensus among field spectroscopy scientists from the same expert group.

Figure 3.5 illustrates group means and group variances for criticality rankings in the 'marine and estuarine environmental conditions' metadata category. This is a more specialized campaign category, where it can be justifiably assumed that the marine scientists have a better informed opinion as to the metadata that most impacts the validity and reliability of *in situ* marine datasets. The group mean rankings for marine scientists are uniformly higher for all metadata fields in this category, and variance is uniformly lower than for rankings assigned by non-marine scientists. These results strengthen the implication that consensus and agreement are dependent upon the respondents' area of expertise. Within-group variance may be explained by investigating correlations between an *in situ* dataset's fitness-for-purpose and the metadata fields critical to fulfilling that purpose.

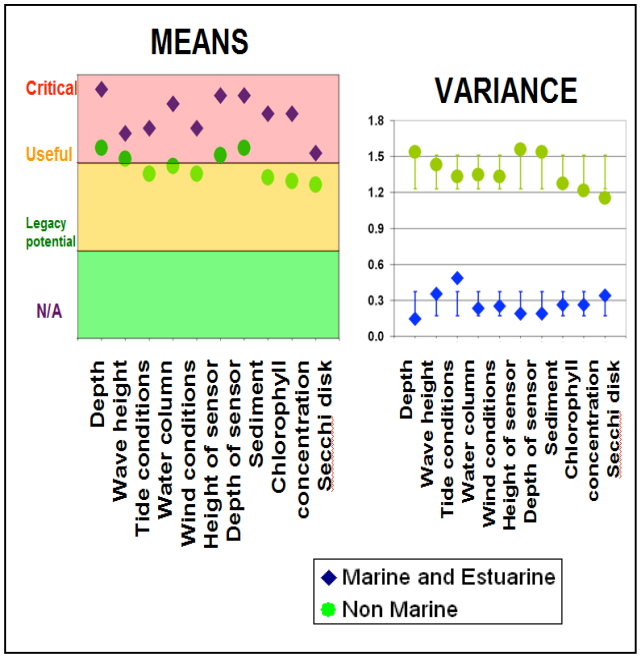

Figure 3.5 Group means and in 'marine and estuarine environmental conditions' category
(Marine and Estuarine n_1=*18,* Non Marine n_2=*49)*

Determining thresholds for including a metadata field in a protocol based on its criticality may require an involved process. This is apparent via a binomial test executed on responses for calibration metadata in Table 3.1, where all critical rankings were compared to non-critical ('useful'/'legacy potential'/'NA'). The null hypothesis was set at p=0.5. Metadata fields that have been designated as critical less than 50% of the time have been highlighted in red. Similar results from all categories prompt two important questions: whose opinion can be used as a basis for designating a metadata field as critical, and under what circumstances?

		Observed Prop.	p-value
Date	Critical	.68	.002
	Non-critical	.32	
Irradiance	Critical	.32	.002
	Non-critical	.68	
Radiance	Critical	.30	.001
	Non-critical	.70	
Darknoise	Critical	.52	.818
	Non-critical	.48	
Signal to Noise	Critical	.55	.422
	Non-critical	.45	
Linearity	Critical	.40	.105
	Non-critical	.60	
Stray Light	Critical	.67	.005
	Non-critical	.33	
Calibration Data	Critical	.61	.081
	Non-critical	.39	
Traceability (yes/no)	Critical	.49	1.000
	Non-critical	.51	
Standard (NIST/NPL, etc.)	Critical	.47	.728
	Non-critical	.53	

Table 3.1 Binomial test results for 'calibration' metadata (*n=78*)

Protocol design cannot be informed solely by the quantitative data and the comments and suggests can provide additional recommendations. Some of the suggestions and comments from the participants include:

- "the context of inquiry must be specific enough to address the variety of type of radiometric data (reflectance, radiance, irradiance, transmission, etc.) and the purpose of the measurements (field survey, algorithm development)"

- "regardless [of] the applications of the field spectroscopy, metadata should contain sufficient information for users 1) to repeat the sampling (or in the least to imagine the measurements and its surrounding condition), 2) to cite and pinpoint the dataset for the reference, and 3) to explore the data as much flexible as possible, even beyond its original purpose"

- "depending on the campaign and available budget and instrumentation different [metadata] points become critical and other[s] useful or negligible"

- "there's a need for an integrated 'quality flag' so that people can rapidly assess whether to utilise the data or not"

CONCLUSIONS

Criticality ranking alone may not be sufficient for building a protocol. This is demonstrated by high variability in response for metadata categories common to all campaigns (instrument, viewing geometry, illumination conditions, etc.). Investigation into what the data is being collected for (activities such as population of a spectral library, calibration and validation) may help determine whether protocols must be streamlined for fitness-for-use within each campaign. This may help navigation through the more ambiguous fields that have been designated as both 'critical' and 'N/A' in almost equal proportion. While a robust dataset with long-term reliability would include at least every metadata field presented in this survey, there are limitations to the investment of time and resources required to record these metadata *in situ*. It may be prudent to scale down a universal metadata set to one that is more ideally suited to the purpose for which it will be used. This would require the informed judgement of both the creator and the users of the campaign data.

TOWARDS A STANDARDIZED *IN SITU* METADATA PROTOCOL

Utilising the expert panel input in the field spectroscopy metadata survey will provide building blocks for *in situ* metadata protocols that are robust enough to meet the requirements of a range of hyperspectral campaigns, while ensuring accuracy and consistency. The avenues for solutions for a universal protocol are multifold, and as demonstrated by some of the survey results presented here, highly dependedent on a given dataset's intended purpose. Creating a dataset that is amenable to installation in a large-scale data library for cataoluging and mining can benefit from an extensible Markup Language (XML)-based exchange format for spectroradiometric metadata because XML 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 can be easily accommodated in a variety of data archiving schema and software, including spectral libraries, databases, and datawarehouses.

A possible solution for the need for quality flagging and interoperability may be addressed by such schema as Water ML 2.0, now an OGC (Open Geospatial Consortium) specification, an example of an XML schema that could be modified and adopted for hyperspectral *in situ* data; it is used for encoding hydrological observation and measurement data, and accommodates quality flagging by incorporating metadata (contaminated sample, holding time exceeded, etc.) that affects the data's fitness for use in future applications (Terhorst, 2009).

The strengths and weaknesses of data encoding format becomes a valuable debate when designing an *in situ* hyperspectral metadata protocol, because both the representation of the metadata in a digital format and its subsequent compatibility with schema such as the datawarehousing model for large-scale archiving, sharing, and mining of hyperspectral metadata need to be considered (Rasaiah *et al.,* 2011). Preliminary results from the survey indicate that there although there exist consensus on the criticality of some metadata, building a universal schema for data storage, exchange and interoperability requires a more refined analysis on the fitness-for-purpose for each dataset.

ACKNOWLEDGEMENTS

- Anonymous survey participants who generously devoted their time and expertise
- Adrian Schembri, *RMIT University, Australia*

REFERENCES

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

B. Rasaiah, T.J. Malthus, S.D. Jones, C. Bellman, "The Role of Hyperspectral Metadata In Hypspectral Data Exchange and Warehousing", *Proceedings of the 7th EARSeL Workshop*, Edinburgh, Scotland, 2011 (*in press*) .

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

M. E. Schaepman, S.L. Ustin, A.J. Plaza, T. H. Painter, J. Verrelst, S. Liang, (2009). Earth system science related imaging spectroscopy—An assessment. Remote Sensing of Environment, 113:S123–S137.

A. Terhorst. "WaterML 2.0.", GEOSS Sensor Web Workshop Tsukuba, Japan, 21-22 May 2009. https://portal.opengeospatial.org/files/?artifact_id=34678 (accessed August 02, 2011)