Towards a metadata standard for field spectroscopy datasets

Barbara Rasaiah
BSc, Computer and Mathematical Science
MAppSci, Geospatial Information

School of Mathematical and Geospatial Sciences
College of Science, Engineering and Health
RMIT University
Melbourne, Australia

This thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy.

June 2014

Author’s Declaration

I hereby declare that except where due acknowledgement has been made, the work is my own, the work has not been submitted previously, in whole or in part, to qualify for any other academic award and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Signature: …………………………………………………………………………….

Name: Barbara Rasaiah

Date: …………………………………………………………………………….

Abstract

The volume of information derived from *in situ* field spectroradiometers, across a broad variety of, often costly, applications and instrumentation, grows each year. There is a recognized need within the international remote sensing community to document, store, and share field spectroscopy data and metadata in consistent formats within dedicated data sharing and other intelligent archiving systems (Committee on Earth Observing Satellites, 2013; Group on Earth Observations, 2014). Establishing and maintaining optimal integrity of the data is a key priority to ensure effective re-use of the data, and to enable more efficient and higher impact research.

Metadata is an important component in the cataloguing and analysis of field spectroscopy datasets because of their central role in identifying and quantifying the quality and reliability of spectral data and the products derived from them. There is currently no international standard methodology for collecting field spectroscopy metadata. This makes rich and flexible metadata capabilities a critical factor in the interoperability and quality assurance of datasets.

This thesis identifies the core components for a field spectroscopy metadata standard to facilitate discoverability, interoperability, reliability, quality assurance and extended life cycles for datasets being exchanged in a variety of data sharing platforms. The research is divided into five parts: 1) an overview of the importance of field spectroscopy, metadata paradigms and standards, metadata quality and geospatial data archiving systems; 2) definition of a core metadataset critical for all field spectroscopy applications; 3) definition of an extended metadataset for specific applications; 4) methods and metrics for assessing metadata quality and completeness in spectral data archives; 5) recommendations for implementing a field spectroscopy metadata standard in data warehouses and 'big data' environments.

Part 1 of the thesis is a review of the importance of field spectroscopy in remote sensing; metadata paradigms and standards; field spectroscopy metadata practices, metadata quality; and geospatial data archiving systems. The impact of field spectroscopy as a foundation to scientific operations and research is examined. Definitions of metadata from across disciplines are presented, and the usefulness of metadata as a tool for making datasets discoverable, shareable, and interoperable is explored. The unique metadata requirements for field spectroscopy are discussed. Conventional definitions and metrics for measuring metadata quality are presented. Finally, geospatial data archiving systems for data warehousing and intelligent information exchange are explained.

Part 2 of the thesis presents a core metadataset for all field spectroscopy applications. The core metadataset is derived from the results of an international expert panel survey. The survey respondents helped to identify a metadataset critical to all field spectroscopy campaigns, as well as recommend additional metadata to increase the versatility of a metadataset, both for application-specific metadata and general campaign metadata. These results form the foundation of a field spectroscopy metadata standard that is practical, flexible enough to suit the

purpose for which the data is being collected, and/or has sufficient legacy potential for long-term sharing and interoperability with other datasets.

Part 3 presents an extended metadataset for specific application areas within field spectroscopy. The key metadata is presented for three applications: tree crown, soil, and underwater coral reflectance measurements. The performance of existing metadata standards in complying with the field spectroscopy metadataset was measured. Results show they consistently fail to accommodate the needs of both field spectroscopy scientists in general as well as the three application areas. A hybrid standard that serves as a 'best of breed' incorporating useful modules and parameters within the standards is proposed.

Part 4 presents criteria for measuring the quality and completeness of field spectroscopy metadata in a spectral archive. Existing methods for measuring quality and completeness of metadata were scrutinized against the special requirements of field spectroscopy datasets. Field spectroscopy metadata quality can be defined in terms of (but not limited to) logical consistency, lineage, semantic and syntactic error rates, compliance with a quality standard, quality assurance by a recognized authority, and reputational authority of the data owners/data creators. Two spectral libraries were examined as case studies of operationalized metadata, and the degree to which they comply with the needs of field spectroscopy scientists. The case studies revealed that publicly available datasets are underperforming on the quality and completeness measures.

Part 5 presents recommendations for adoption and implementation of a field spectroscopy standard, both within the field spectroscopy community and within the wider scope of IT infrastructure for storing and sharing field spectroscopy metadata within data warehouses and big data environments. The recommendations are divided into two main sections: community adoption of the standard, and integration of standardized metadatasets into data warehouses and big data platforms.

In conclusion, this thesis has identified the core components of a metadata standard for field spectroscopy. The metadata standard serves overall to increase the discoverability, reliability, quality, and life cycle of field spectroscopy metadatasets for wide-scale data exchange. It also presents recommendations for a formal adoption of the standard by the field spectroscopy community and the steps forward for its integration into data warehouses and big data platforms.

Table of Contents

Preface

The work presented herein was completed as a body of work for this thesis and is substantially my own work. Publications and contributions from others are detailed below:

The content in Chapter 3 was published in part as:

Rasaiah, B.; Jones, S.D.; Bellman, C.; Malthus, T. (2014). Critical Metadata for Spectroscopy Field Campaigns, *Remote Sensing,* 6, 3662-3680. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2012). Critical Metadata Protocols in Hyperspectral Field Campaigns for Building Robust Hyperspectral Datasets*, Proceedings of the XXII ISPRS Congress,* August 26 – September 1 in Melbourne, Australia 2012. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2011). Designing a Robust Hyperspectral Dataset: The Fundamental Role of Metadata Protocols in Hyperspectral Field Campaigns*. Proceedings of Geospatial Science Research Symposium*, November 28-30 in Melbourne, Australia. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2011). Building Better Hyperspectral Datasets: The Fundamental Role of Metadata Protocols in Hyperspectral Field Campaigns. *Proceedings of the Surveying & Spatial Sciences Conference,* November 21-25 in Wellington, New Zealand 2011. (peer-reviewed)

The content in Chapter 4 was published in part as:

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2012). A Novel Metadata Schema for *in situ* Marine Spectroscopy, *Proceedings of Geospatial Science Research Symposium 2*, December 10-12 in Melbourne, Australia, 2012 (peer-reviewed).

The content in Chapter 6 was published in part as:

Rasaiah, B.; Jones, S.D.; Chisholm, L.; Hueni, A.; Bellman, C.; Malthus, T.J.; Chisholm, L.; Gamon, J.; Huete, A.; Ong, C.; Phinn, S.; Roelfsema, C.; Suarez, L.; Townsend, P.; Trevithick, R.; Wyatt, M. (2013). Approaches to Establishing a Metadata Standard for Field Spectroscopy Datasets. *Proceedings of IGARSS 2013*, July 21-26 in Melbourne, Australia. (peer-reviewed)

Acknowledgements

Thank you to my supervisors (Simon Jones, Professor, RMIT University; Chris Bellman, Associate Professor, RMIT University; Tim Malthus, Program Leader and Research Scientist, Land and Water, CSIRO) for your support, guidance, and expertise over the course of my research. You were available to answer my questions, review my ideas, and spend your evenings, lunch hours, weekends and holidays reading my thesis drafts -- I will never cease to admire the stamina it takes to supervise a PhD! Thank you for leading me through your proven and rigorous scientific approach to research.

Simon – as my senior supervisor, thank you for supporting me in a research topic of my choice; for giving me so much freedom to explore the ideas that interested me; for the opportunity to travel and work with scientists from around the world; for the opportunity to participate in and lead workshops and the conferences that broadened my view in the most enjoyable and memorable way. This aside from the many ways that you educated me on how to carry out doctoral research!

Chris – thank for your ever insightful viewpoints and meticulousness in examining my work, and for taking the time to write comments in the margins of my chapter drafts with a broken wrist. You are one of the most exacting reviewers and thesis editors I have met. I hope you continue to put your talents as a skilled negotiator to use!

Tim – thank you especially for your instrumental role in proposing an initial metadataset that directed the course of analysis for my thesis; for organizing the workshops (both in Australia and abroad) that introduced me to the international experts with whom I collaborated; for all your contributions to my papers, thesis, and guidance on my presentations.

You are, each of you, thesis wizards with different hats. It has been an honour working with you.

Thank you also, to –

The 90 scientists who participated in the online survey – you gave freely of your time and expertise. You gave me the first stepping stone for a metadata standard.

The scientists and experts who attended the spectral library workshops – it was a pleasure to spend time with you and learn how you create, store, and share your data. Specifically: Brian Curtiss, John Gamon, Phil Townsend, Chris MacLellan, Robert Hewson, Andy Hueni, Alfredo Huete, Laurie Chisholm, Stuart Phinn, Cindy Ong, Chris Roelfsema, Mariela Soto-Berelov, Lola Suarez, Rebecca Trevithick, Davina White, Matthew Wyatt, Carlos Aya.

Andy Hueni of Remote Sensing Laboratories, Switzerland – for your expert insights on metadata, especially in field spectroscopy.

Uta Heiden of DLR, Germany – for your all your efforts and assistance in my examination of the Spectral Archive data.

Anthony Bedford of SMGS Consultancy, RMIT University – for your advice on the best statistical analyses for my datasets.

Eliza Cook of SMGS Administration, RMIT University – for your guidance in all the administrative tasks involved in completing a PhD.

John Hearne and the school of SMGS, RMIT University – for enabling me to complete my research with the support of a tuition scholarship.

Finally, thank you to everyone, both acknowledged and unacknowledged, who in their unique way contributed to making my research a thoroughly enjoyable and educational experience.

List of Figures

List of Tables

List of Equations

Chapter 1 Introduction

1.1 Introduction

Field spectroscopy metadata is a central component in the quality and reliability of spectral data and the products derived from it. The impact of the quantity and quality of metadata created at this fundamental stage of spectral research is amplified as spectral data exchange becomes more important and widespread in the international remote sensing community. Cataloguing, data mining, and interoperability of these datasets rely upon the robustness of metadata protocols for field spectroscopy. Currently, no standard methodology for collecting field spectroscopy metadata exists. There is an immediate need within the international remote sensing community to establish a metadata standard for field spectroscopy that ensures high quality, interoperable metadatasets that can be archived and shared efficiently within Earth observation data sharing systems.

Field spectroscopy metadata consists of those data elements that explicitly document the spectroscopy dataset and field protocols, sampling strategies, instrument properties and environmental and logistical variables. Field spectroscopy datasets are dependent upon their associated metadata for ensuring their quality, reliability, and longevity. A superior quality metadataset can describe a broad range of observed field data, across a range of applications. Such metadata is vital since it can influence 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 a tool for potentially mitigating these errors. The logistics of collecting sufficiently reliable and complete metadata, as well as the requisite volume of metadata, is a central consideration for creating a standardized methodology for defining and documenting metadata. A practical metadata standard must be closely aligned to field spectroscopy data collection practices adopted by remote sensing research communities around the world.

There is urgency in acquiring continuous high quality spectroscopy data to solve problems in Earth systems science (Milton *et al.*, 2009). Informing users and stakeholders of field spectroscopy datasets of the impact of high-quality data and metadata in the context of Earth observing data systems is an additional challenge facing the remote sensing community. Quality assurance of field spectroscopy datasets necessitates oversight and standardization, both at local, national, and international scales and is a way of ensuring robust metadata protocols for field spectroscopy. The need for a standardized methodology for collecting field spectroscopy metadata has increased with the emergence of data sharing initiatives such as NASA's EOSDIS (Earth Science Data and Information System) LTER (Long Term Ecological Research) network, Australian Terrestrial Ecosystem Research Network (TERN), SpecNet, and some of the smaller *ad hoc* spectral libraries and databases created by remote sensing communities internationally.

Careful examination of all stages of metadata collection and analysis can inform a robust metadata standard that is widely applicable to field campaigns. This thesis presents the core components of a metadata standard that serve to increase the discoverability, reliability, quality, and life cycle of field spectroscopy datasets for wide-scale data exchange.

1.2 Research objectives

The main objectives of this research are: i) to identify core components of a metadata standard for field spectroscopy to enhance discoverability, reliability, quality, and longevity of datasets; and ii) to derive methods and metrics for evaluating metadata completeness and quality in field spectroscopy datasets.

1.3 Research questions

The following research questions are answered in this thesis:

1. What are the key elements of a core metadataset for all field spectroscopy applications?

2. Is additional metadata required for specific field spectroscopy applications and to support interoperability with other metadata standards?

3. What are the criteria for measuring the quality and completeness of field spectroscopy metadata in a spectral archive?

4. What are the issues related to adoption of the proposed field spectroscopy metadata standard?

1.4 Thesis outline

The thesis consists of seven chapters, four of which are research chapters, and a final chapter of conclusions.

Chapter 2 presents a literature review addressing: the importance of field spectroscopy, metadata paradigms across a range of disciplines, field spectroscopy metadata practices, conventional definitions and metrics for measuring metadata quality, and geospatial data archiving systems. It seeks to present gaps in knowledge within the remote sensing community in the context of addressing needs of field spectroscopy scientists for quality-assured metadata documentation and spectral data exchange.

Chapter 3 identifies elements of a core metadataset for all field spectroscopy applications. Results from an international expert panel survey produced a core metadataset that serves as the central component of a field spectroscopy metadata standard. The core metadataset is necessary for ensuring that a field spectroscopy metadata standard is practical, flexible enough to suit the purpose for which the data is being collected, and/or has sufficient legacy potential for long-term sharing and interoperability with other datasets.

Chapter 4 identifies additional metadata for specific applications in field spectroscopy. The key metadata is presented for three applications: tree crown, soil, and underwater coral reflectance measurement. The suitability of existing metadata standards in supporting the core field spectroscopy metadataset presented in

Chapter 3 is measured, and a hybrid standard that serves as a 'best of breed' incorporating useful modules and parameters within the standards is proposed.

Chapter 5 investigates criteria for measuring the quality and completeness of field spectroscopy metadata in a spectral archive. Existing methods for measuring quality and completeness of metadata are presented. These are scrutinized against the special requirements of field spectroscopy datasets. Two spectral libraries are examined as case studies of operationalized metadata, and the degree to which they align with the needs of field spectroscopy scientists is assessed.

Chapter 6 investigates issues to adoption of the proposed field spectroscopy metadata standard. It presents a set of recommendations for community adoption of the standard. A proposed way forward for the integration of standardized metadatasets into data warehouses and big data platforms is also discussed.

Chapter 7 presents the conclusions of the research by providing an overview of outcomes from the four research questions.

Chapter 2 Literature review

2.1 Introduction

This chapter presents a review of literature on the importance of field spectroscopy, metadata paradigms and standards, field spectroscopy metadata documentation practices, metadata quality, and geospatial data sharing systems.

2.2 The importance of field spectroscopy

2.2.1 What is field spectroscopy?

Field spectroscopy falls within the science of remote sensing. Remote sensing can be defined as using instruments to gather data about an object or an area from a distance (ESA, 2013; NOAA, 2013*d*) (Figure 2.1).

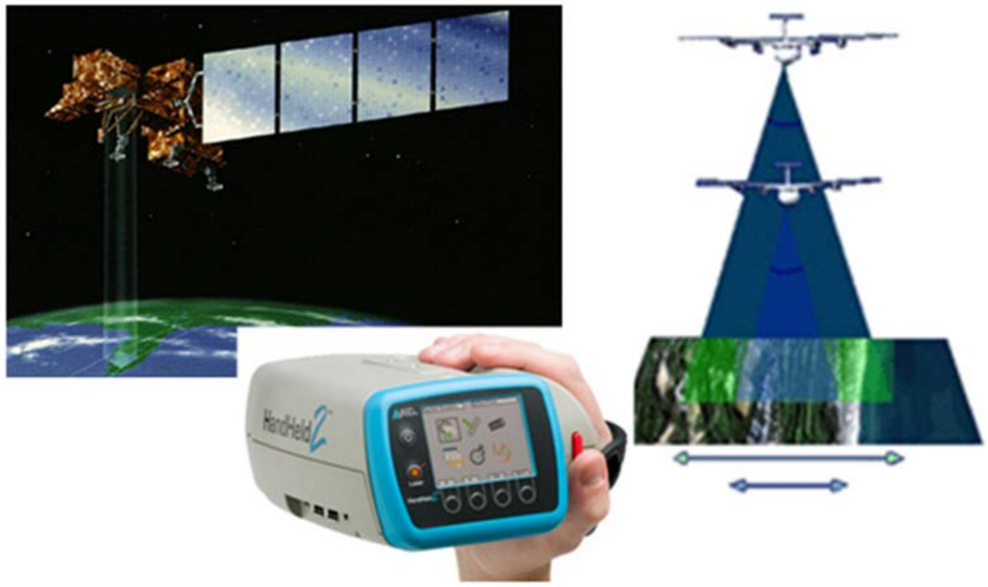


Figure 2.1 Examples of satellite-based, airborne, and in situ remote sensing instruments (left) NASA's Landast 7 satellite scanning along its orbital path (center) handheld field spectroradiometer (right) airborne multispectral imaging
Source: (NASA, n.d.; ASD, 2012; Channel Systems, 2010)

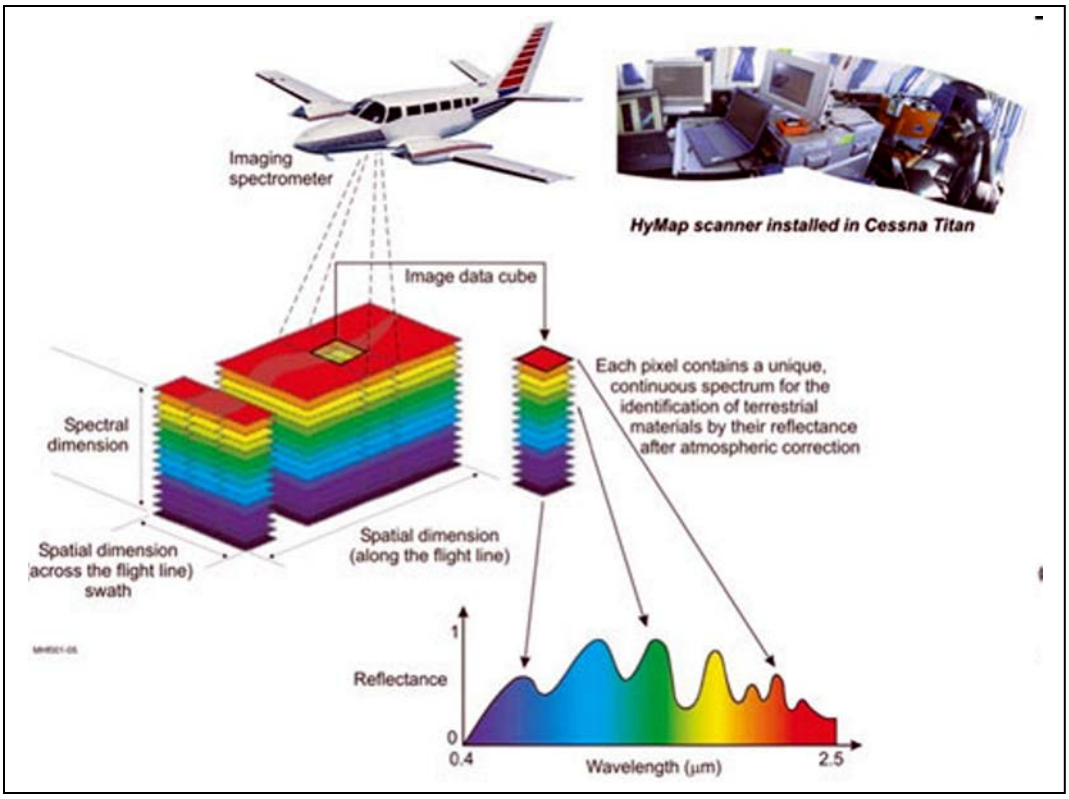

Imaging spectrometer
HyMap scanner installed in Cessna Titan
Image data cube
Each pixel contains a unique, continuous spectrum for the identification of terrestrial materials by their reflectance after atmospheric correction
Spectral dimension
Spatial dimension (across the flight line) swath
Spatial dimension (along the flight line)
Reflectance
1
0
0.4
Wavelength (µm)
2.5

Figure 2.2 Reflectance signature generated by an airborne HyMap sensor
Source: (HyVista Corporation, 2012)

Figure 2.2 illustrates how a passive hyperspectral sensor (HyMap) on board a plane generates an electromagnetic (em) reflectance signature for the surface of the Earth along the flight path of the plane.

Field spectroscopy takes place in what can be considered the natural environment, where em reflectance, radiance, irradiance and transmission of features in natural settings (vegetation, seagrass, rocks, soils, snow, and rooftops) are measured (Mac Arthur, 2011) (Figure 2.3). Principles underlying field spectroscopy, as with remote sensing in general, are well defined and a variety of measurement and data analysis techniques are used, depending on the specific application domain (Li and Strahler, 1992; Lewis and Barnsley, 1994; Sandmeier and Itten, 1999; Martonchik *et al*., 2000;

Nolen and Dozier, 2000; Sandmeier, 2000; Dangel *et al*., 2005; Peltoniemi *et al*., 2005*a*, 2005*b*; Schaepman-Strub *et al*., 2006, 2009; Schaepman, 2007; Schopfer *et al*., 2008; Jacquemoud *et al*., 2009; Kokaly *et al*., 2009; Milton *et al*., 2009; Dekker *et al*., 2010; Dumont *et al*., 2010). A field spectroradiometer, or alternately, a 'spectrometer' or 'field radiometer' is the device used to measure the em signals, and is a passive sensing instrument.

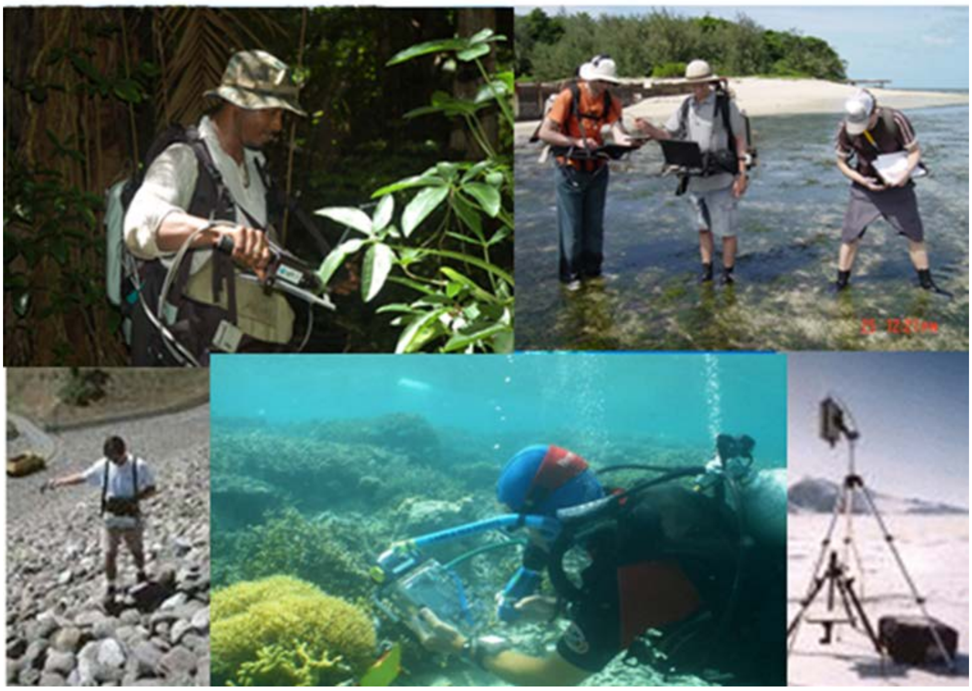


Figure 2.3 Common types of field spectroscopy campaigns Clockwise from left: vegetation, estuarine, snow, underwater coral, and geological. *Source: (ASD, 2013c; CSER, 2012; NERC FSF, n.d.; Biophysical Remote Sensing Group, 2011; USGS, 2002)*

Prototype multispectral field spectroradiometers emerged in the 1960s for use within the research community and the first commercially available research grade portable spectroradiometer for the field was made available by ASD (Analytical Spectral Devices, now PANalytical Boulder, Inc.) in the early 1990s (Milton *et al*., 2009). Other manufacturers today include Spectra Vista Corporation, Ocean Optics, Skye Instruments, GmbH, and TriOS.

A spectroradiometer can be characterized by technical specifications including its spectral range (the em wavelength span its sensor bank can respond to), spectral resolution (the smallest distance in wavelengths that can be discriminated), number of spectral channels or bands, radiometric resolution (sensitivity to the magnitude of the em signal, expressed as a bit value), and manufacturer. For example, the ASD Field Spec Hi-Res Spectroradiometer has a spectral range of 350-2500 nm, with a spectral resolution ranging from 3 nm to 8nm across 2151 bands and radiometric resolution of 16 bits (ASD, 2013*a*). The SVC (Spectra Vista Corporation) GER 1500 has a spectral range of 350-1050 nm, with a spectral resolution of 3.2nm across 512 bands and a radiometric resolution of 16 bits (SVC, n.d.). Other distinguishing characteristics include field of view, sensor head size, signal-to-noise ratio, integration time, photo diode composition, full-width-half-maximum measure, and foreoptics available. This review focuses on passive hyperspectral in *situ* spectroradiometers with sensitivity in the 0.35 -2.5 nm range. Hyperspectral can be defined as sensitivity in hundreds or thousands of bands versus multispectral instruments, which are sensitive in multiple bands, such as the MODIS (36 12-bit bands), Landsat 8 (11 12-bit bands) and MERIS (15 16-bit bands) satellite sensors.

2.2.2 Field spectroscopy's contribution to remote sensing

Field spectroscopy serves as the fundamental stage for primary research and operational applications (Herold *et al*., 2005; McCoy, 2005; Mazel, 2006; Milton *et al*., 2009; Viscarra Rossel *et al*., 2009; Asner *et al*., 2011) and provides critical input for calibration, validation, and other data modelling activities within the research community. It can be used to obtain information about the spectral characteristics of elements (leaf, rock, snow, asphalt) within a natural or urban scene (forest, riverbed, glacier, city); as ground-truth for calibration of airborne and spaceborne sensors; provide input for models (biophysical, radiative transfer); and provide reference signatures for spectral libraries and databases (Hueni *et al*., 2009; Milton *et al*., 2009, Mac Arthur, 2011). This data is ideally objective and replicable, and can be used in a diverse range of applications including chlorophyll estimation in water, vegetation biomass, spatial variation in atmospheric constituents, and geomorphic mapping (Mac Arthur, 2011).

There is a continual stream of studies and projects around the world that make direct use of field spectroscopy data to enhance scientific understanding both within the realm of remote sensing and in interdisciplinary contexts. A sample of studies for a one-year period ending in 2012 was retrieved from the top three remote sensing journals (by impact factor over a five-year period representing the average number of citations received per paper published in that journal during the preceding five years) (Senf, 2013). Table 2.1 classifies the studies (n=20) retrieved from Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing over a one-year period by the purpose for which the field spectroscopy data was used and the research area (e.g. vegetation, snow, etc.).

Research area*	Application model / algorithm validation	classification / mapping
Generic	2	1
Atmosphere	1	0
Marine/Estuarine	2	0
Vegetation	7	6
Snow	0	1
Soils	1	0
Wetlands	1	0

Table 2.1 Summary of application and research area for field spectroscopy data for remote sensing journal articles over 2012 (n=20)

**where an individual study applied to more than one research area (e.g. snow and forest mapping), each research area was counted individually for that study*

**Source: (Cheng et al., 2012 ; Ciganda et al., 2012; Gonzalez et al., 2012; Hernandez-Clemente et al., 2012 ; Hesketh et al., 2012 ; Inoue et al., 2012 ; Jiao et al., 2012 ; Knaeps et al., 2012 ; Knox et al., 2012 ; Mazzoni et al., 2012 ; Mishra et al., 2012 ; Niemi et al., 2012 ; Pascucci et al., 2012 ; Pisek et al., 2012 ; Rodger et al., 2012 ; Sayer et al., 2012 ; Serrano et al., 2012 ; Thorp et al., 2012 ; Tits et al., 2012 ; Zhang et al., 2012)*

The analysis of the studies shows that they are biased towards vegetation research (65%), and are more often used in model and algorithm validation (65%) than for target classification or mapping. None of the referenced studies explicitly stated that the data derived from field spectroscopy was used for other applications including sensor calibration (Pegrum *et al*., 2006; Green, 2010; Pacheco-Labrador *et al*., 2014) or population of spectral libraries (Becvar *et al*., 2006; Pfitzner *et al*., 2006, 2010; Hueni *et al*. 2009, 2010, 2012; USGS, 2006; Haselwimmer and Fretwell, 2009; Zomer *et al*. 2009; Iordache *et al*., 2010) that have been documented in other studies. The breadth of applications for field spectroscopy grows with the increased reliance on remote sensing to answer scientific questions, the development and availability of new and more advanced spectroradiometers, the refinement of protocols for specific applications, the volume of field data being generated, and continually evolving data sharing capabilities among researchers.

2.2.3 Field spectroscopy protocols

Campaigns, or the operations and activities involved in the field spectroscopy data collection for a given application, can be defined and differentiated by their logistics (including equipment and ease-of-access to the target location), instrumentation, operators involved and purposes for which the data is being collected. This diversity stems from large potential variations in instrument setup and calibration, viewing geometry, reference standards, target sampling strategies, and environmental variables. It is widely acknowledged that these factors influence the spectral measurements and should be documented to allow mitigation and intercomparison (Duggin, 1985; Kerekes, 1998; McCoy, 2005; Stuckens *et al*., 2009).

There are laboratories, research agencies and organizations that provide documentation for good practice in the field (Table 2.2). The degree of their prescriptiveness and assumptions about the instrument operator and principal investigator varies. There are guides that are comprehensive, especially for specific applications, and some that assume that the principal investigator has an advanced understanding of the principles of sampling (viewing geometry strategies, bi-directional distribution functions) with little background information about field spectroscopy science provided.

Name of document	Topics addressed					
	Application specific	**Em theory**	**Instrument optimization**	**Recommended viewing geometry**	**Sampling strategy**	**Field data documentation protocol**
NERC FSF instrument guides (ASD Field Spec Pro, GER1500, GER3700) (Mac Arthur, 2006, 2007*a*, 2007*b*)			X			
Australian Government Department of Sustainability, Environment, Water, Population and Communities: Standards for reflectance spectral measurement of temporal vegetation plots (Pfitzner *et al*., 2011)	X	X	X	X	X	X
University of Queensland Field Spectrometer and Radiometer Guide (Phinn *et al*., 2007)	X	X	X	X	X	X
Spectranomics Protocol: Leaf Spectroscopy (350-2500nm) (Carnegie Spectranomics, 2010)	X				X	
ASD instrument guides and FAQ (ASD 2012, 2013*b*)		X	X	X		

Table 2.2 Comparison of field spectroscopy good practice guides

The amount of advice given and its explicitness varies across the good practice guides and illustrates the spectrum of opinions about what constitutes good sampling strategy. The comparison shows that the application-specific guides (Australian Government Department of Sustainability, Environment, Water, Population and Communities: Standards for reflectance spectral measurement of temporal vegetation plots and the University of Queensland Field Spectrometer and Radiometer Guide) discuss the broadest range of topics for field spectroscopy and are more explicit in their instructions for field protocol and how to document it (with the exception of the Spectranomics Protocol: Leaf Spectroscopy (350-2500nm) guide). The other guides leave it to the researchers to decide what viewing geometries and sampling strategies are ideal, and omit references to field data documentation.

NERC FSF (National Environmental Research Council Field Spectroscopy Facility) states in its online instrument (ASD Field Spec Pro, GER1500, GER5700) guides that it is unable to recommend sampling strategies due to varying requirements across projects, and that this responsibility ultimately lies with the principal investigator (Mac Arthur, 2006, 2007*a*, 2007*b*), but it does advise on sampling strategies in its training courses (NERC FSF, 2014). It does provide general guidance about warming up the spectroradiometer prior to measuring samples, the importance of calculating the field of view, secure mounting of the instrument, and taking white reference measurements for the ASD Field Spec Pro (Fogwill 2005; Mac Arthur 2006, 2007*a*). PANalytical Boulder (formerly ASD Inc.), the leading world manufacturer of field spectroradiometers, maintains an online document repository on the physics of field spectroscopy, as well as general guidance for instrument optimization, and viewing geometry in its instrument guides (ASD 2012, 2013*b*).

Others provide more explicit guidance on field protocol. The Australian government Department of Sustainability, Environment, Water, Population and Communities provides a detailed protocol for spectral measurement of temporal vegetation plots (Pfitzner *et al*., 2011). It includes a background on em theory for field spectroscopy, and recommends the number of average signals per sample, optimal viewing geometries, stabilizing equipment setup, methods for cleaning the white reference panel, and a protocol for measuring an instrument’s conformance to manufacturer specifications (including warm-up time for illumination laps, average spectrums and white reference measurements taken). The Carnegie Spectranomics lab provides detailed protocol for leaf collection and spectroscopy, but omits any discussion on em theory (Carnegie Institute for Science, 2010).

The University of Queensland provides a detailed protocol for marine campaigns that includes advice about the instruments (ASD, Ocean Optics, TriOS Ramses) best suited to the type of signal being recorded (*in situ* marine spectral reflectance, down- or up-welling irradiance for depth profiles) (Phinn *et al*., 2007). It also presents optimal sampling strategies and ways of minimizing influencing environment effects on the signal, including: specific references to CSIRO-recommended viewing geometries, proper communication with divers operating the instrument, ways to avoid splashing water on the instrument, minimizing reflecting effects of wet samples and surrounding environments, and measuring the water surface and column before each white reference measurement to counteract their influence (Phinn *et al*., 2007).

Research groups around the world, each taking samples according to their own ‘good practice’ protocols is not sufficient to guarantee consistent measurement and output, even when the target is a controlled variable. For example, Jung *et al*. (2010) reported on a simple scenario with a single non-variant object, in which fifteen spectroscopy laboratories used the same instrument, targets, and a consistent instrument calibration protocol to record the spectral reflectance of the targets. A marked variation in output reflectance was noticeable, suggesting potential consequences for the inter-comparability of the spectra.

A more complex scenario has the potential for increased variation in measured output. In cases where the sample remains the same but the instrument and white reference panel changes, systematic differences are introduced by the device and reference panels (Jung *et al*., 2010). As an additional example of the impact of the instrument, PANalytical Boulder Inc. supplies a device known as a 'Scrambler' for its FieldSpec models to compensate for spectral discontinuities due to non-uniformity of field-of-view across the sensor bank fibreoptics (Mac Arthur *et al*., 2012). Good practice would assume that a FieldSpec user has accounted for this in their field protocol, especially in the cases where they are intercomparing datasets for the same samples generated from other instruments.

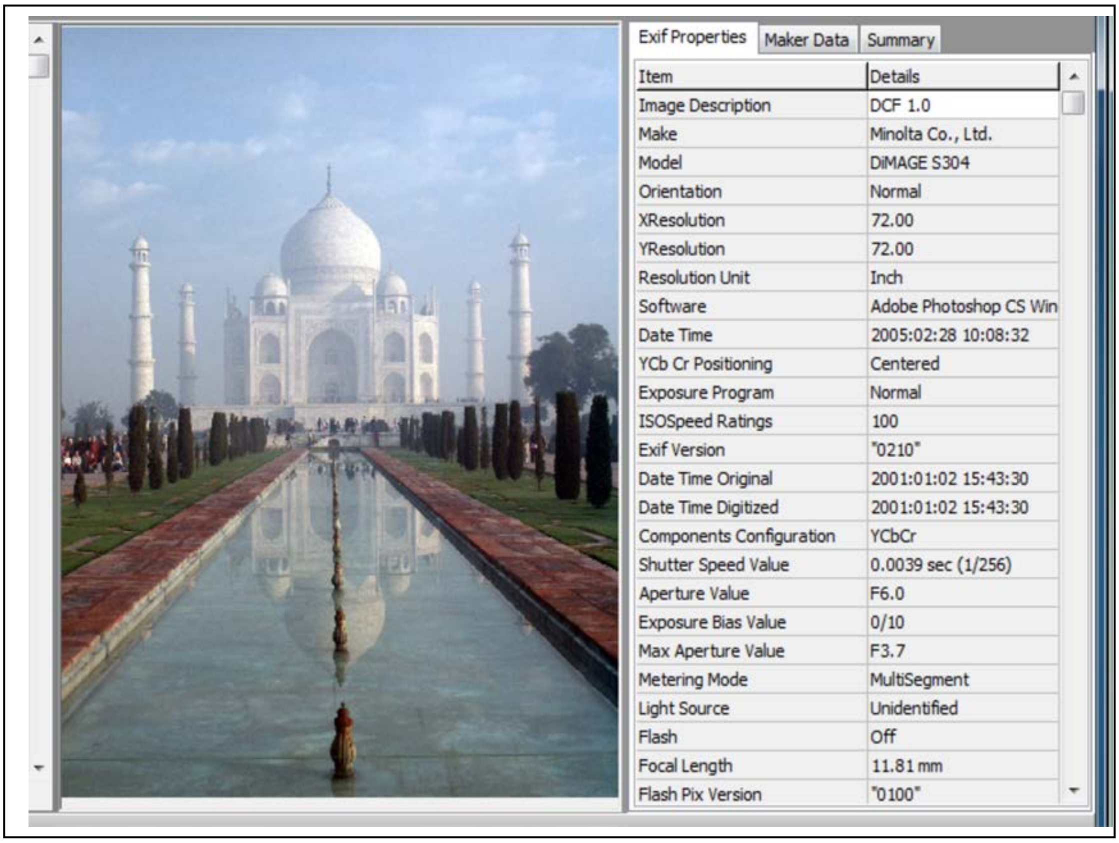
In general, activities undertaken to produce reflectance products -- including data preprocessing, the choice of atmospheric correction algorithms, illumination and viewing angles, and radiometric calibration -- can each contribute to inconsistent measurements for the same sample (Schaepman-Strub *et al*., 2006). Documenting the source of these differences in the derived reflectance products, in a standardized terminology for the benefit of data users, remains a challenge within the remote sensing community (Schaepman-Strub *et al*., 2006). For this reason, "whether the methodology is designed for a one-off sample for correlation with airborne or satellite multispectral or hyperspectral image data, or temporal measurements, spectral data must be collected in a well-designed and consistent manner" (Pfitzner *et al.*, 2006, pp. 89-90).

The review presented in this section has helped to demonstrate that field spectroscopy is an essential activity in remote sensing, in applications including modelling activities, classification and mapping, population of spectral libraries, and sensor calibration. Internationally, research groups adhere to different protocols for carrying out campaigns, with different opinions on what constitutes 'good practice' for viewing geometry, sampling strategy, and documentation of field data, among others. However, a lack of standardized protocols, and no community consensus on how to document them (i.e. what metadata to provide), ultimately may serve as a hindrance to intercomparison of field spectroscopy datasets and quality assurance.

2.3 Metadata paradigms and standards

2.3.1 Introduction

Metadata, in its broadest definition, is 'data about data' or 'information about information'. It originates in the discipline of computer science, from a 1968 book written on computer languages used for electronic database searching (Bagley, 2013). This definition has been expanded and refined within different disciplines. It generally refers to information that functions to make datasets discoverable, interoperable regardless of source, software and hardware platforms that manage and maintain the data, and for the archiving and preservation of information (MIT, n.d.; NISO, 2004, 2007; Higgins, 2007; ANDS, 2011). Figure 2.4 is an example of metadata generated automatically by a digital camera.



Item	Details
Image Description	DCF 1.0
Make	Minolta Co., Ltd.
Model	DiMAGE S304
Orientation	Normal
XResolution	72.00
YResolution	72.00
Resolution Unit	Inch
Software	Adobe Photoshop CS Win
Date Time	2005:02:28 10:08:32
YCb Cr Positioning	Centered
Exposure Program	Normal
ISOSpeed Ratings	100
Exif Version	"0210"
Date Time Original	2001:01:02 15:43:30
Date Time Digitized	2001:01:02 15:43:30
Components Configuration	YCbCr
Shutter Speed Value	0.0039 sec (1/256)
Aperture Value	F6.0
Exposure Bias Value	0/10
Max Aperture Value	F3.7
Metering Mode	MultiSegment
Light Source	Unidentified
Flash	Off
Focal Length	11.81 mm
Flash Pix Version	"0100"

Figure 2.4 Image metadata automatically generated as an EXIF (Exchangeable Image File Format) file by a digital camera. It includes information about the date the image was taken, camera properties and settings, and image dimensions and resolution. *Source: (Williams, 2012)*

Information science, a discipline with a history of and vested interest in metadata research, defines metadata as information that serves to manage, preserve and distribute information resources. It can be divided into the following categories: i) descriptive (information about the subject matter, creators, type of data) for the purposes of discoverability; ii) operations metadata (technical information about the data management of the organization and distribution of digital objects); iii) preservation metadata (archiving information); iv) rights metadata (security access, legal restrictions, copyrights for publishing and viewing); v) administrative (describes organizing workflow for preparing and publishing electronic resources) (MIT, n.d.; NISO, 2004, 2007; Higgins, 2007; ANDS, 2011).

This definition has also been extended to accommodate a unique category of metadata referred to as scientific metadata, which is “all the information that is very specific to the study, and is needed to use and interpret the data collected” (ANDS, 2011) and “auxiliary information, ranging perhaps from the experimenter and the time and place that the experiment was conducted to arcane calibration details” (Davenhall, n.d.)

Metadata can be used to reference any digital or physical object. Metadata can be created for such entities as museum artifacts, biological specimens, chemicals, data tables, music, films and web documents, among others. Within computer science, metadata commonly refers to the identification and handling of data elements. For example, in the management of large datasets within databases and data warehouses, metadata can be categorized as business metadata (descriptive information about such entities as tables, calculations used for derived attributes), technical metadata (load and performance statistics, data quality problems) and process execution metadata (transfer duration, logging information) (Green, 2009; Gamji, 2011). Audio files also have their own metadata. The ID3 tag is a metadata container with information including the title of the song, artist, album and track number, and is encoded within audio files (commonly .mp3 files) and readable within software (Windows Media Player, iTunes) and hardware players (Creative Zen, iPod) (O'Neill, 2013). Museums worldwide adhere to specific protocols for creating and documenting metadata for their biological specimens, archaeological specimens, and works of art (Getty Research Institute, 2006). Across the disciplines and applications, metadata can be generated automatically or created manually, depending on the type of object or entity being described (digital or physical), user community preferences, and whether software exists to support the documentation and archiving of metadata.

Metadata can have multiple uses for the reason that it makes datasets documentable and discoverable. It can have significant implications in instances of intellectual property and legal matters, as examples. The validity and admissibility of evidence from electronic systems are dependent upon the existence of metadata (author, data of creation, location) (Gezler, 2008). At the other end of this spectrum, there are instances where it is preferable to minimize discoverability of metadata. In the protection of intellectual rights and medical privacy for example, for some datasets descriptive metadata published publicly should be kept to a minimum (Slamanig and Stingl, 2008). These examples illustrate that metadata can be a powerful tool in the discovery and sharing of datasets, as wells as establishing their provenance.

2.3.2 Metadata standards and structure

A metadata standard can be defined as the set of elements used to perform specific functions (description, preservation, access and operations information, administration) within a metadataset. Standards will vary, therefore, on the number of metadata elements or metadata fields, and the application for which they were designed. Any individual or organization can develop a standard, but not all are officially recognized, adopted, and/or implemented. Figure 2.5 illustrates some of the metadata standards developed for different disciplines.

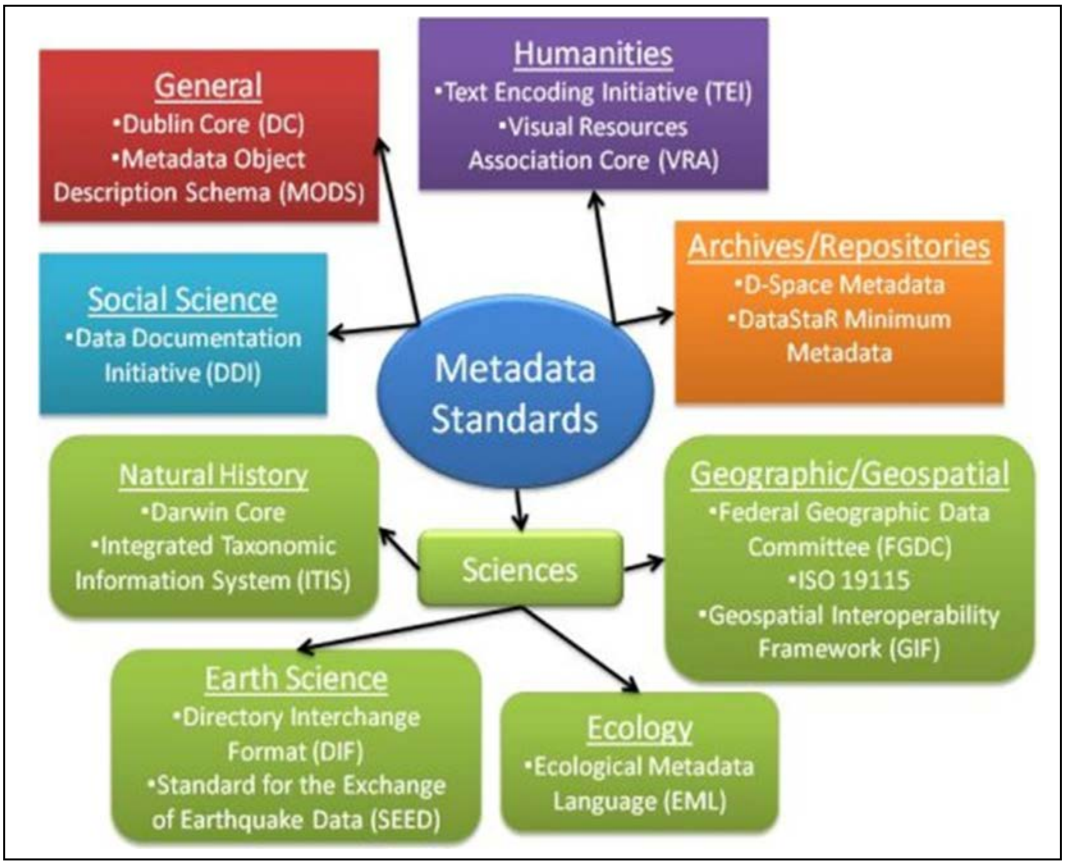

General
•Dublin Core (DC)
•Metadata Object Description Schema (MODS)
Humanities
•Text Encoding Initiative (TEI)
•Visual Resources Association Core (VRA)
Archives/Repositories
•D-Space Metadata
•DataStaR Minimum Metadata
Social Science
•Data Documentation Initiative (DDI)
Metadata Standards
Natural History
•Darwin Core
•Integrated Taxonomic Information System (ITIS)
Sciences
Geographic/Geospatial
•Federal Geographic Data Committee (FGDC)
•ISO 19115
•Geospatial Interoperability Framework (GIF)
Earth Science
•Directory Interchange Format (DIF)
•Standard for the Exchange of Earthquake Data (SEED)
Ecology
•Ecological Metadata Language (EML)

Figure 2.5 Metadata standards across the disciplines *Source: (Tarbet, 2012)*

Metadata standards can be categorized into generic standards, applicable to all datasets for the purposes of archiving and discoverability (Darwin Core, Dublin Core, D-Space Metadata) and more specialized standards of use for a given user community (Access to Biological Collections Data Schema 2.06 for ecology, ANZLIC Metadata Profile 1.1 (Geographic dataset core) for geospatial datasets). Each standard is designed with different objectives for the use of the metadata, for different user groups, with unique vocabularies, taxonomies (discipline-specific classifications based on ontologies among metadata elements), and granularity (the specificity or level of detail at which each metadata field is expressed). The variety of standards illustrates that there is no ‘one size fits all’, and the utility of a standard is directly linked to the preferences and needs of data users, and the purposes for which the metadata will be utilized.

Metadata standards can be structured according to a specific schema, with unique taxonomies, syntax, and granularity. The term ‘standard’ has often been used interchangeably with the term ‘schema’, but there are differences between the two. Metadata schema are the specifications for representing metadata elements in digital format (Higgins, 2007). The schema can include document format (HTML, XML, SGML), syntax (controlled vocabularies), taxonomies, and granularity. Figure 2.6 illustrates the relationship between metadata standards and schema for the ISO 23081 Records Management standard.

Standards and their schema play the greatest role in interoperability between metadatasets. Examining the complexity of schemas helps illustrate this. Schemas can be categorized into three levels of complexity:

1) simple (highest degree of interoperability with other metadata schemas, generally multidisciplinary and non granular, with 15-25 metadata fields);

2) simple/moderate (interoperability is inversely correlated with the specific needs of an application or discipline, granular with more metadata fields);

3) complex (interoperability requires expertise, hierarchical, granular, and extensive, with more than 100 metadata fields) (Greenberg, 2012).

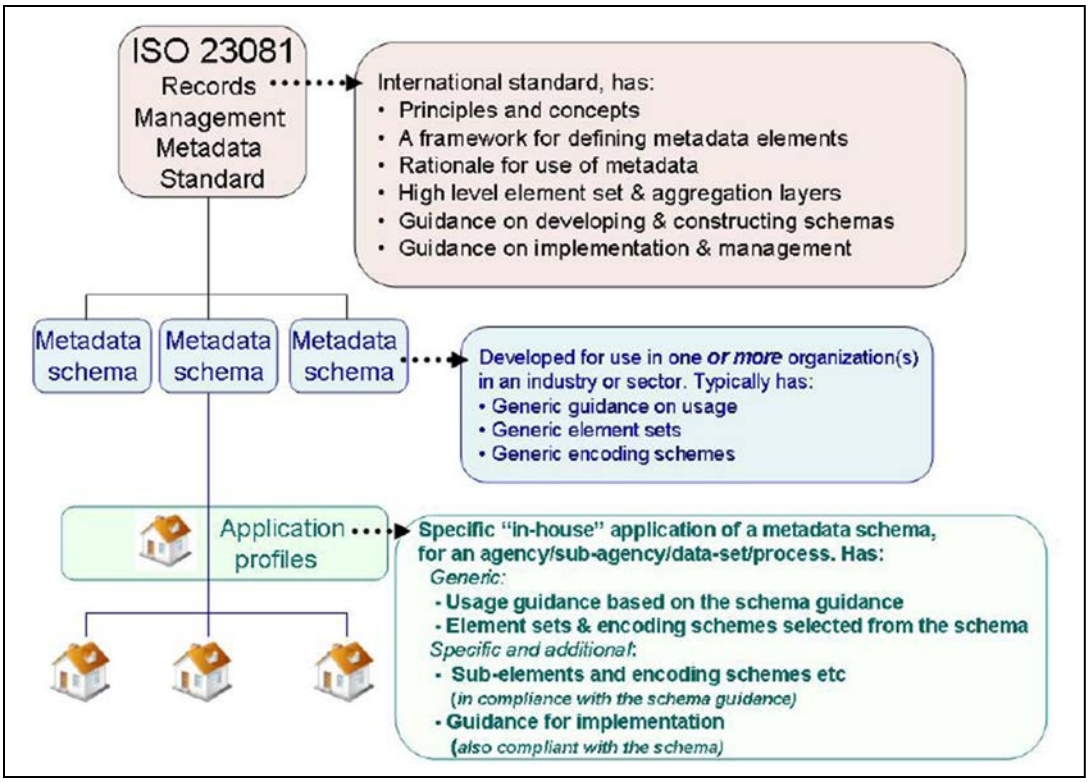
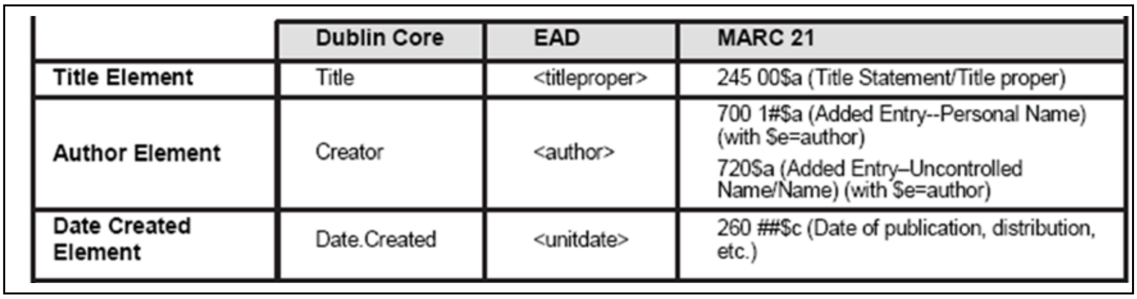

ISO 23081
Records
Management
Metadata
Standard
International standard, has:
• Principles and concepts
• A framework for defining metadata elements
• Rationale for use of metadata
• High level element set & aggregation layers
• Guidance on developing & constructing schemas
• Guidance on implementation & management
Metadata schema
Metadata schema
Metadata schema
Developed for use in one or more organization(s) in an industry or sector. Typically has:
• Generic guidance on usage
• Generic element sets
• Generic encoding schemes
Application profiles
Specific "in-house" application of a metadata schema, for an agency/sub-agency/data-set/process. Has:
Generic:
- Usage guidance based on the schema guidance
- Element sets & encoding schemes selected from the schema
Specific and additional:
- Sub-elements and encoding schemes etc
(in compliance with the schema guidance)
- Guidance for implementation
(also compliant with the schema)

Figure 2.6 ISO 23081 Records Management standard, illustrates the relationships between metadata standards, schema, and application profiles *Source: (ISO/TC, 2008)*

For example, Dublin Core 1.1 has fifteen elements at a single level of granularity, whereas ABCD 2.06 has 1004 elements defined within hierarchies. Mapping and intercomparison of metadata elements between these two standards would be no simple exercise and implies that much consideration must be given to adopting the most suitable metadata standard for a given dataset. Therefore, the complexity of a schema must accommodate the user's needs and the purposes for which the metadata will be used (discoverability, archiving, other).

The ability to find a dataset in a digital repository and assess its usefulness for a given application is dependent in part on its underlying schema. Consider the simple scenario of a database user conducting a search for geospatial datasets by entering keywords for the criteria – this criteria might include geographic extent, description of the datasets, nature of the scientific study for which the dataset was generated, and the instrumentation and sampling protocols used. It is possible that two similar datasets in the database meet the criteria for a user's needs, but their respective metadata is structured according to different taxonomies, vocabularies, and levels of granularity-- perhaps one of these metadatasets adheres to a schema that is unrecognized by the database or insufficiently designed to describe the underlying dataset in a way that this is most useful to the data user. As a result, automated data mining algorithms may filter out the unrecognized metadataset or the data user is presented with search results that do not provide them with enough information to ascertain that both datasets in fact meet their criteria for usability.

There are ongoing efforts to translate metadata from one standard or schema to another to avoid such problems – this is also known as 'crosswalk mapping' (NISO, 2004). Figure 2.7 shows a simple mapping, at a uniform level of granularity, for Dublin Core, EAD (Encoded Archival Description) and MARC 21 (MAchine-Readable Cataloging) standards used in library applications. Schemas have also been extended or adapted for specific applications. OGG (Open Geospatial Consortium) adopted GML (Geographic Markup Language) and KML (Keyhole Markup Language) as schemas based on the XML-format for geographic datasets and 3D map software, respectively.



	Dublin Core	EAD	MARC 21
Title Element	Title	<titleproper>	245 00$a (Title Statement/Title proper)
Author Element	Creator	<author>	700 1#$a (Added Entry--Personal Name) (with $e=author) 720$a (Added Entry–Uncontrolled Name/Name) (with $e=author)
Date Created Element	Date.Created	<unitdate>	260 ##$c (Date of publication, distribution, etc.)

Figure 2.7 Crosswalk mapping for Dublin Core, EAD, MARC21 *Source: (NISO, 2004)*

2.3.3 The role of metadata in data discoverability, sharing and distribution

Metadata is a central factor in data mining, sharing and distribution of datasets. Metadata is about “controlling the quality of data entering the data stream” (Mailvaganam, 2007). It is important for data producers, owners, and managers to ensure that a metadataset is as complete and high quality as possible before it is uploaded to databases, datawarehouses, cloud platforms, or otherwise made available for distribution (Orr, 1998; Bruce and Hillman, 2004; Loshin, 2010; da Cruz *et al*., 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.

Any effective metadata system should provide good correspondence between the description of the resource by the cataloguer and the strategies of the searcher (Wason and Wiley, 2000). This can be considered a ‘gold standard’ for any metadata policy that would necessitate addressing specific requirements (taxonomy, granularity, hierarchical structure, extensiveness of a metadata schema) for making a metadataset as discoverable as possible. For example, metadata spaces is one conceptualization of ideally structured metadata for maximum discoverability in data repositories. The 'best' metadata spaces in a data repository are defined by metadata fields that are orthogonal (independent) to each other where resolution (ability to differentiate between two separate datasets), precision (measure of metadata detail) and repeatability (ability to describe a dataset the same way on two or more occasions) are balanced with the querying approach and objectives of the data user (Wason and Wiley, 2000). The Learning Object Metadata standard has incorporated the concept of metadata spaces for digital learning environments (Wason and Wiley, 2000; IEEE, 2002). However, for the time being, topics such as maximum discoverability and metadata spaces must remain as high-level concepts that are worth noting, but difficult to examine further on a practical level given that there is no community understanding within remote sensing of the fundamental requirements for field spectroscopy metadata.

The simplest way of making datasets accessible through metadata is with metadata registries or clearinghouses, which are databases of metadata. They contain descriptive, access, and preservation data for information sources (other databases, text documents, pdfs, music, videos, museum artifacts, etc.). Any digital or physical object that has associated metadata can be referenced in a metadata registry. The user can execute searches through an online interface. Examples of metadata registries are the Distributed Archive Centre for Biogeochemical Dynamics, for field campaigns, regional and global data, land validation products, environmental numeric data models (Oak Ridge National Laboratory, 2013); the Knowledge

Network for Biocomplexity for biodiversity ecosystem data biodiversity and ecosystem data across multiple habitats (KNB, 2013); the United States Health Information Knowledgebase for healthcare data and standards (Agency for Healthcare Research and Quality, 2013), and NASA's Global Change Master Directory, which holds a catalog of all of NASA's Earth science data set and service descriptions, one of the largest public metadata inventories in the world (NASA, 2013*c*).

2.4 Field spectroscopy metadata documentation practices

2.4.1 What is field spectroscopy metadata?

Metadata in the context of field spectroscopy can be defined as those data elements that explicitly document the primary spectroscopy dataset and field protocols that capture sampling strategies, instrument properties and environmental and logistical variables, all of which are integral to assessing fitness-for-purpose of the spectral measurements (Milton *et al*., 2009; Dekker *et al*., 2010; Malthus *et al.*, 2010, Pfitzner *et al.*, 2011)**.** In a broader context, this definition is also aligned with the purpose and scope of geospatial metadata standards such as the FGDC Content Standard for Digital Geospatial Metadata: Shoreline Metadata Profile, used "to capture critical processes and conditions that revolve around creating and collecting shoreline data, and to help define and qualify shoreline data for use" (FGDC, 2001, p.1); Ecological Metadata Language 2.1.1, used to describe the dataset in fine detail as well as the methodology, including field and sampling methods, applied to obtain the dataset (KNB, 2013), and the ANDS definition of scientific metadata as "all the information that is very specific to the study, and is needed to use and interpret the data collected" (ANDS, 2011). The provision within metadata standards for the documentation of protocols is accepted as good practice as scientists across disciplines acknowledge that such activities, considered to be in the background of research problems, are not commonly presented with the data outputs generated from field data collections, but are an important part of the research process and therefore must be captured and available for sharing (Wynholds *et al*., 2012). Figure 2.8 is a conceptualization of the interrelationships among metadata and their effect on the analysis of spectral measurements.

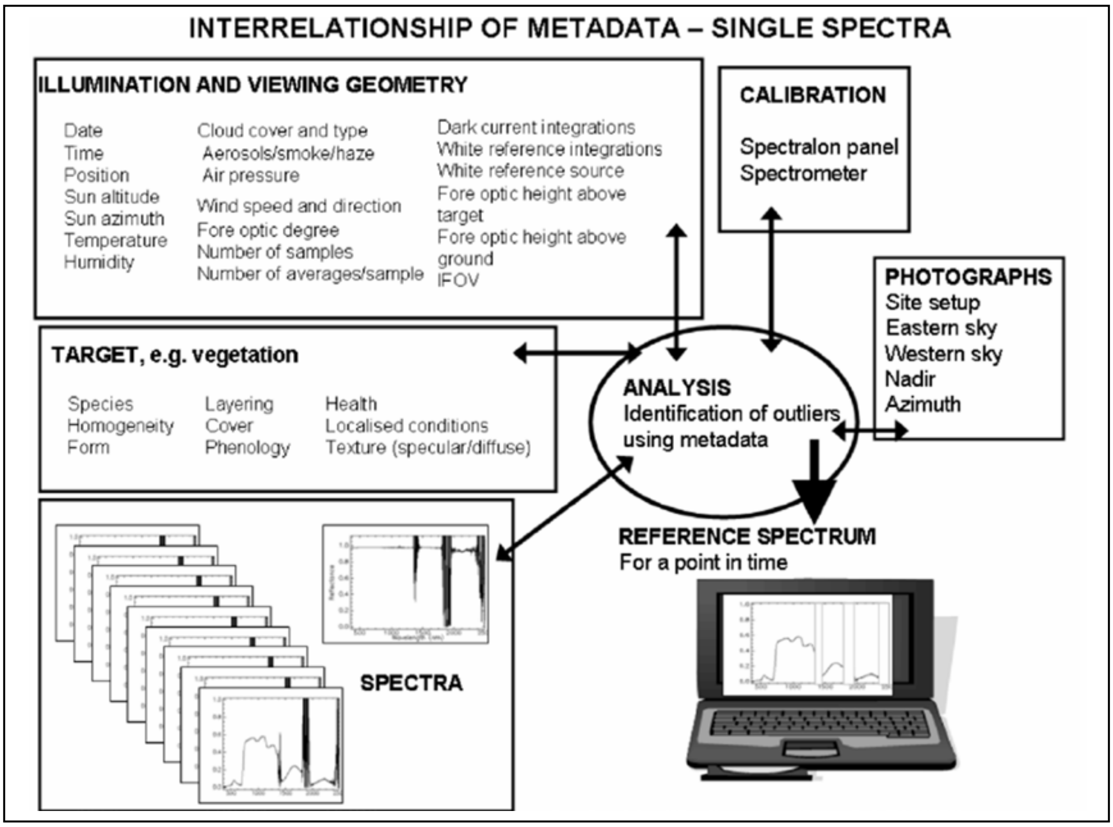


Figure 2.8 A conceptualization of the interrelationships among metadata and their effect of the analysis of spectral measurements *Source: (Pfitzner et al., 2006)*

2.4.2 Field spectroscopy metadata standards and conventions

There are no national or international standards for the documentation of field spectroscopy metadata, the minimum set required, or any quality assurance process for metadata. Metadata modelling techniques and standards have been proposed by numerous bodies overseeing and advising the geospatial sciences but fail on several fronts to address the relevant aspects of field spectroscopy datasets. Many are based on the ISO 191__ standard family relating to storage, encoding, and quality evaluation of geographic data. 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 (INSPIRE, 2009; OGC 2012). While this helps to solve many problems in the intercomparison of geospatial metadatasets in general, field spectroscopy datasets are not represented in these protocols.

Although providing general guidelines, geospatial metadata standards do not explicitly address the metadata requirements of field spectroscopy collection techniques, or the ontologies and data dependencies required to model the complex interrelationships among the observed phenomena as data and metadata entities. For example, a logical semantic model would express dependencies between metadata entities such as user-controlled viewing conditions including sensor orientation, height above the target, and area of target in the field of view, all three of which have a relationship with the spectral measurements. Weaknesses in field spectroscopy data collection and their implication for the need for a metadata standard have been identified by both users and providers of field spectroscopy data, particularly in the European remote sensing community; these include a lack of quality assurance and calibration information for sensors; no real capability to define accuracy or validation for data processing; and a lack of agreed standards in data processing (Reusen *et al*., 2007). Steps forward in confronting these challenges must begin with investigating how the field spectroscopy community currently documents metadata.

2.4.3 Documenting field spectroscopy metadata

Methods of documenting and storing metadata vary across research groups and as with field protocols, are done on an *ad hoc* basis. Worldwide practice for recording metadata relating to the instrument properties, illumination and viewing angles, reference standards and general project information is done according to a group's own definition of what constitutes a suitable metadataset (Dekker *et al*., 2010; Mac Arthur, 2007*a*; Pfitzner *et al.,* 2011). When inconsistent sampling and measurement protocols remain undocumented through metadata, any valid intercomparison of datasets is compromised. The impact of these variables across datasets has not yet been fully identified within the remote sensing community nor can it be properly quantified in many instances, further necessitating the recording of adequate metadata. The time invested in metadata collection is surpassed by its benefits in reducing system bias and variability (Pfitzner *et al*., 2006). While most users recognize this, what is required are standards and techniques to facilitate easy recording of this data. Capturing such data is therefore central to ensuring reliability, legacy, re-use, and interoperability of field spectroscopy datasets. Application-agnostic metadata has been recommended within the field spectroscopy community (Bojinski *et al.*, 2003; Milton *et al*., 2009) but there is no consensus in the literature on conventions to use, or how inclusive it should be.

Metadata can be documented manually, and concurrently with the spectral measurements, or generated automatically and obtained post-campaign. The choice to document concurrently or retrospectively can be a result of prioritizing metadata due to constraints of time and conditions under which the measurements are being taken (Fogwill 2005; CSER, 2006; Mac Arthur 2006, 2007*a*, 2007*b*; Phinn *et al*., 2007) (Figure 2.9). For example, information relating to viewing geometry, which includes the height and angle of the sensor above the target, and height of the sensor above ground, the field of view, and foreoptics used – is best documented in the same window of time as the em signatures being recorded, since this data is difficult to obtain post-fact and prone to error if done from memory alone.

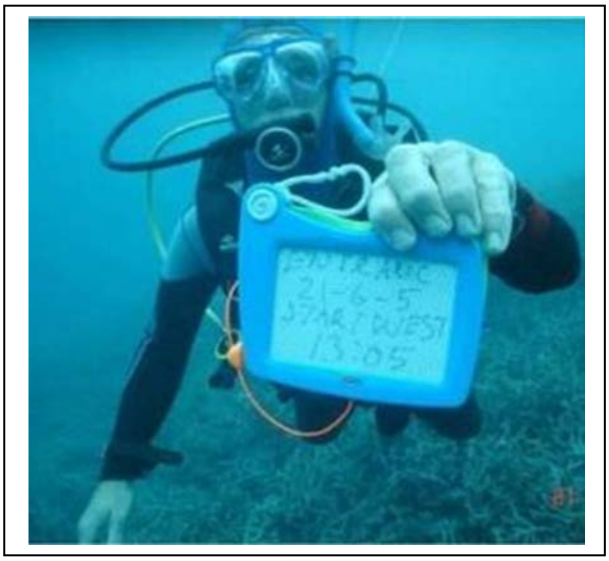


Figure 2.9 Documenting metadata underwater *Source: (CSER, 2006)*

Retrospectively documented metadata is most often information that is unlikely to change over the duration of the campaign, or information that has been documented elsewhere by a third party. Meteorological agencies, oceanographic institutes (NOAA, University of Hawaii Sea Level Center), and weather stations have online data relating to local weather information, solar angles, and tides.

Automatically generated metadata include those from the field spectroradiometers themselves that encode instrument and signal properties information within their native files that can then be exported as metadata to a local or central database or other data repository. If a single instrument is used for multiple campaigns, the information can be documented once, and then referenced through the metadata by instrument serial number or other identifying key (Hueni, 2011).

2.4.4 Storing and sharing field spectroscopy metadata

Popular avenues for the storage of field spectroscopy metadata include log sheets, text documents, and excel files. NERC FSF provides metadata log sheets for generic use in the field, but the number and type of metadata fields vary by instrument (Fogwill 2005; Mac Arthur 2006, 2007). The University of Queensland Centre for Spatial Environment Research supplies field metadata sheets for vegetation campaigns, but the metadata fields do not adhere to officially recognized norms other than the parameters pertaining to sky conditions for describing cloud amount (oktas) and type (based on recommendations from the UK Meteorological Office) and viewing geometry parameters (based on CSIRO recommendations) (Phinn *et al*., 2007). It also supplies excel files for loading spectral measurements with their associated metadata for marine campaigns (CSER, 2006). CSIRO provide datasheets for spectral measurements in shallow benthic habitats and for some metadata fields, proper documentation is restricted to pre-defined keywords to describe water colour, water type, substrate density of cover, epiphytic growth, cloud cover as a percentage (CSIRO, n.d.). These metadata proforma in general are inconsistent in the number of metadata fields they recommend, the information they represent, and how these metadata fields should be expressed (specificity, naming conventions, keywords).

There are online repositories of field spectroscopy data and metadata for scientists and general members of the public to access. The USGS Spectral Library (http://speclab.cr.usgs.gov/spectral-lib.html) is available online for download. The library was developed to support imaging spectroscopy studies of the Earth and other planets (USGS, 2006). Functionally it is an html-based directory of spectra with associated metadata. There are 820 spectra, categorized into mineral, vegetation, man-made, mixture, volatile, microorganism, and plant samples. Each spectrum is stored as an image plot and metadata including sample name, description, chemical formula, sample donor, location, xrd analysis, with up to 24 metadata fields stored in pre-defined templates for each category of target. It is a static library in the sense that the data is read-only, and members of the public cannot upload new spectra or perform updates. It is one of three spectral libraries (both field and lab-based spectroscopy) within the ASTER Spectral Library (http://speclib.jpl.nasa.gov/).

Recent developments in relational spectral databases (some of which are online and publicly accessible) have allowed a more structured storage for spectral measurements and their associated metadata (Pfitzner *et al.*, 2006; Hueni *et al.*, 2007). The SPECCHIO (http://www.specchio.ch/) database is available online for members of the public and can also be downloaded as a local instance. SPECCHIO was created by Remote Sensing Laboratories at the University of Zurich to store reference spectra and campaign data obtained by spectroradiometers in a central repository (Hueni *et al.,* 2009). It is accessible through a Java application, and all data is stored in a MySQL database. The public can upload spectra and metadata and make edits to their own datasets. It contains 111,023 spectra across 55 campaigns that are available for viewing and download. Metadata is stored at both the spectrum and campaign level, some of which is auto-generated. Users have the option of additional metadata they wish to populate, either at the spectrum level (including viewing geometry, target homogeneity, environment information) or campaign level (including description, associated institute).

The DLR (German Aerospace Center) Spectral Archive (http://cocoon.caf.dlr.de/intro_en.html) is a publicly available database created in 2006 to serve as a tool for archiving, managing and using spectral signals collected from a variety of campaigns in the field and/or in the laboratory (Becvar *et al*., 2006). It has 152 campaigns with 1609 spectra, with metadata defied at the campaign and spectrum level. The data is published as html files accessible through any web browser. To upload data, users must comply with the metadata formatting provided, but there is no minimum required metadataset specified.

None of these spectral databases, however, have a full suite of standardized metadata definitions, and nor do they provide quality assessment or assurance for the metadataset (SPECCHIO allows users to assign a quality designation to their metadataset but this option has not been used for any of the stored datasets).

There are other databases that are in the process of being developed or being enhanced. The SSD (Supervising Scientist Division, Australia) Spectral Vegetation Database was developed in 2006 for the objective of providing a database of reference spectral signatures in the 200-2500 nm range for vegetative ground covers (Pfitzner *et al*., 2006). It is not available to the public currently. The Carnegie Spectranomics Lab has published a database online for spectroscopy data and metadata for tropical forest canopy trees, lianas, vines, hemi-epiphytes, and other lifeforms that are normally inaccessible to scientific researchers; spectroscopy data is not yet available (Carnegie Spectranomics, 2013). SPECCHIO is undergoing an enhancement to become a national spectral database for Australia to ensure the long-term storage of data and support scientists in data analysis activities (Hueni et *al*., 2012). The purposes for which these spectral libraries and databases have been created are not consistent; while some have been created as a repository of reference data (USGS Spectral Library), others are a tool simply for sharing data (DLR Spectra Library, SPECCHIO). This in turn creates inconsistency in quality control policy for input datasets and their associated metadata, in those cases where quality control policy exists and/or is enforced. In order for any current or future data metadata exchange mechanisms and platforms to be useful to data users as tools for making informed choices about a dataset's usability, issues of metadata quality must be addressed.

2.5 Geospatial data sharing systems

There are international initiatives to share geospatial data among researchers and the public. 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.

EOSDIS (Earth Observing System Data Information System) is a network of data centres, metadata repositories, middleware providers and directory services for NASA's Earth science data (Kuo, 2010). It provides datasets and products derived from satellites, aircraft, field measurements and other sources. As of September 2012, it offered 6,886 unique data products, with an average daily archive growth of 5.4 TB, and total archive volume of 7.4 PB (NASA, 2013*a*). EOSDIS incorporates the Global Change Master Directory, a metadata clearinghouse for NASA Earth observation datasets, documents, and services. NASA has also collaborated with OCC (Open Cloud Consortium) on Project Matsu to create cloud-serviced Earth Observing satellite image processing for global flood and fire monitoring (Grossman *et al.,* 2012). Figure 2.10 is a high-level overview of NASA's Earth science data operations, showing the data flow from source data acquisition, to processing, and distribution through EOSDIS for access to researchers and members of the public.

TERN (Terrestrial Ecosystem Research Network) is an Australian initiative to coordinate a national data network with quality assured observational data from the terrestrial domain (TERN, 2013). It was built for Australia's ecosystem science community to share and manage data from a network of research facilities including plot-based monitoring systems and remotely sensed data time-series products (AusCover Facility), coastal ecosystem datasets (The Australian Coast Ecosystems Facility) and multidisciplinary ecosystem observatories (The Australian Supersite Network), among others (TERN, 2013).

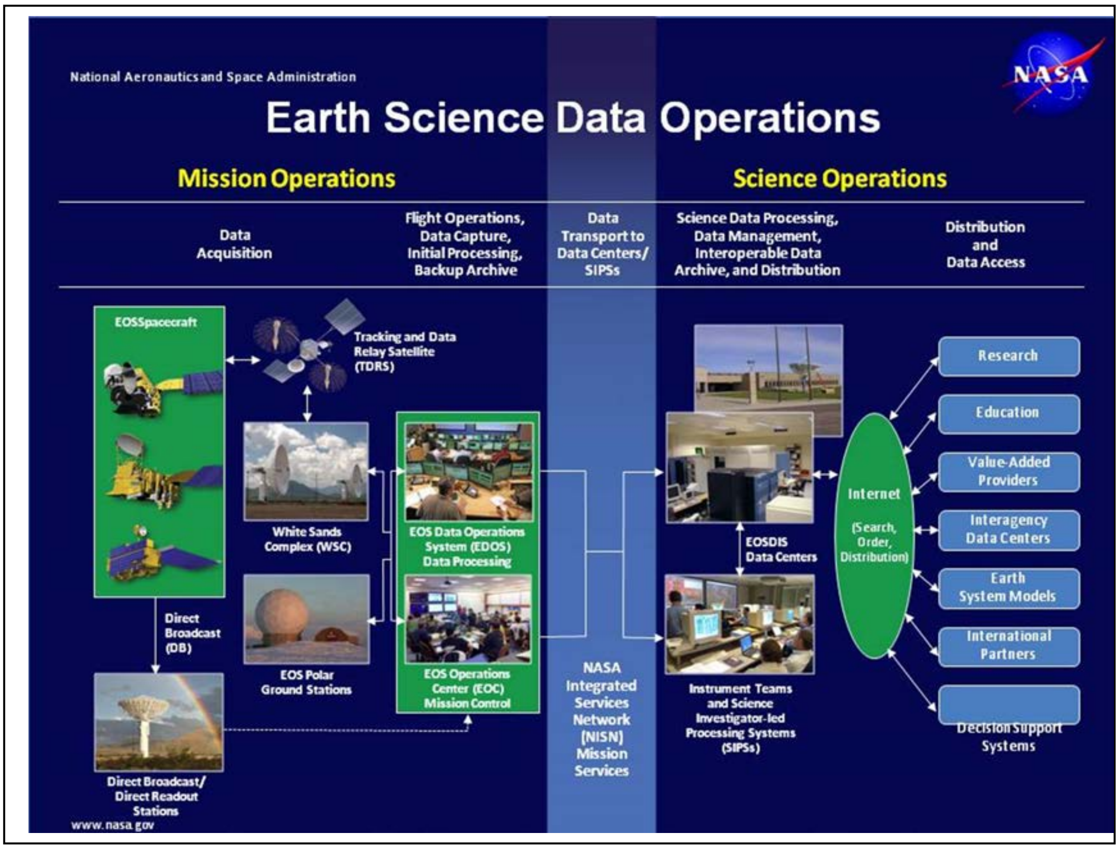

National Aeronautics and Space Administration
NASA
Earth Science Data Operations
Mission Operations
Science Operations
Data Acquisition
Flight Operations, Data Capture, Initial Processing, Backup Archive
Data Transport to Data Centers/ SIPSs
Science Data Processing, Data Management, Interoperable Data Archive, and Distribution
Distribution and Data Access
EOS Spacecraft
Tracking and Data Relay Satellite (TDRS)
White Sands Complex (WSC)
EOS Data Operations System (EDOS) Data Processing
EOS Polar Ground Stations
EOS Operations Center (EOC) Mission Control
Direct Broadcast (DB)
Direct Broadcast/ Direct Readout Stations
NASA Integrated Services Network (NISN) Mission Services
EOSDIS Data Centers
Instrument Teams and Science Investigator-led Processing Systems (SIPSs)
Internet (Search, Order, Distribution)
Research
Education
Value-Added Providers
Interagency Data Centers
Earth System Models
International Partners
Decision Support Systems
www.nasa.gov

Figure 2.10 NASA's Earth Data Science Operations, incorporating EOSDIS, and distribution through clearinghouses including GCMD
Source: (NASA, 2011)

The TERN Data Discovery Portal, a metadata clearinghouse for datasets sourced from each facility, is the access point for these datasets. In 2012, TERN teamed with Google to leverage their cloud computing services to create an online vegetation monitoring tool through the Google Earth Engine for land managers (States News Service, 2012). Figure 2.11 provides a data flow overview for facilities participating in TERN.

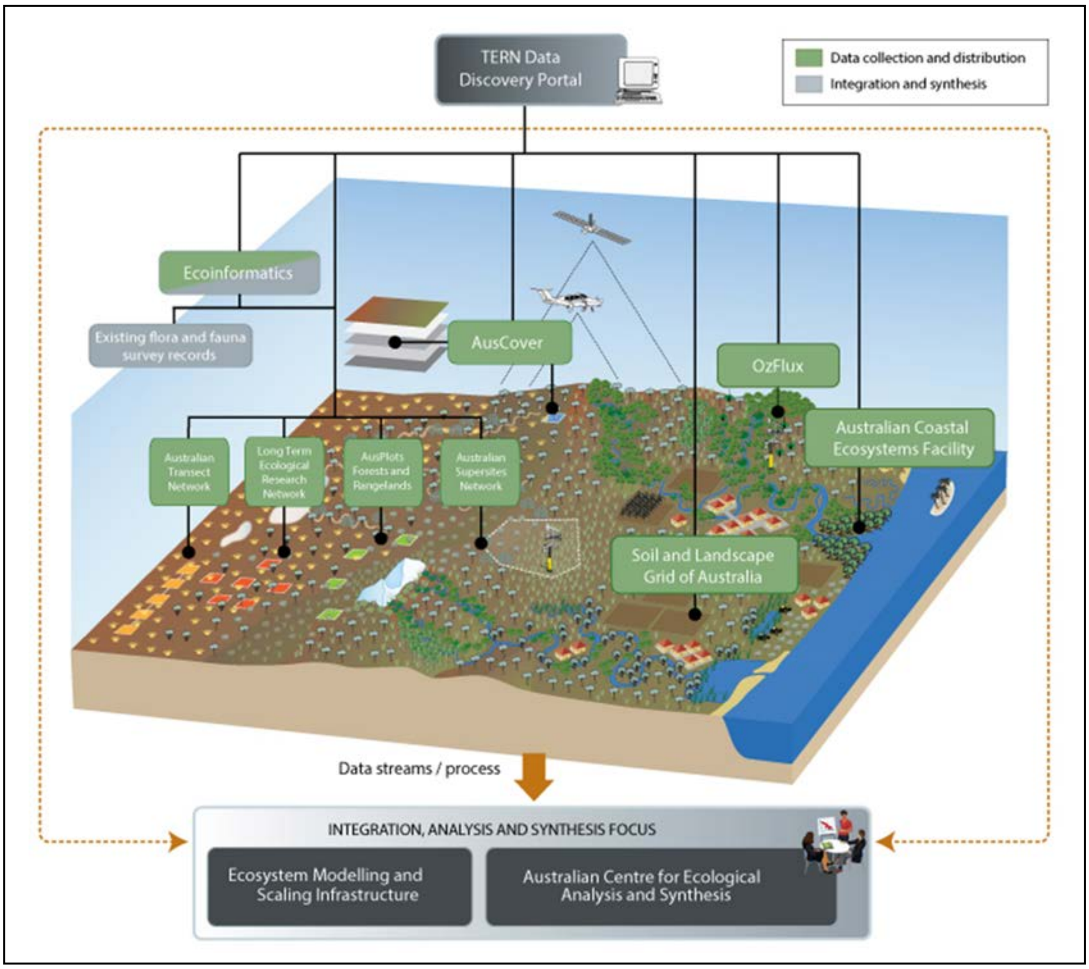

TERN Data Discovery Portal
Data collection and distribution
Integration and synthesis
Ecoinformatics
Existing flora and fauna survey records
AusCover
OzFlux
Australian Coastal Ecosystems Facility
Australian Transect Network
Long Term Ecological Research Network
AusPlots Forests and Rangelands
Australian Supersites Network
Soil and Landscape Grid of Australia
Data streams / process
INTEGRATION, ANALYSIS AND SYNTHESIS FOCUS
Ecosystem Modelling and Scaling Infrastructure
Australian Centre for Ecological Analysis and Synthesis

Figure 2.11 A hierarchical overview of the facilities contributing to the TERN dataset accessible through the TERN Data Discovery Portal *Source: (TERN, 2013)*

NOAA’s National Climatic Data Centre is a digital archive of global weather and climate data. It originated in 1951 as a weather records bureau and has evolved to become the world’s largest provider of land-based, marine, model, radar, weather balloon, satellite, and paleoclimatic datasets (NOAA, 2013*a*). NCDC’s digital archive has increased from 1 petabyte to 6 petabytes in the past 10 years and is expected to exceed 15 petabytes by 2020 with the continual introduction of new remote sensing technologies (NOAA, 2013*a*).

Though not comprehensive, the list above exemplifies the largest operational geospatial data sharing systems, with the greatest variety of source data. These systems present the possibilities that exist for quality assured sharing of geospatial datasets, metadata, and derived products for the research community and the public on national and international scales. They are examples of on-demand access that were developed to meet the needs of scientific communities that require a central archiving and distribution platform for their data and metadatasets to assist in answering research questions both individually and as a group. The speed at which the data repositories continue to grow within these systems also emphasizes the need to implement standards and policies at the level of the data producers, data owners, data managers, and within the IT infrastructure to maintain complete, high-quality, and up-to-date metadatasets.

2.6 Metadata quality

2.6.1 Introduction

It is important here to differentiate between concepts of *data* quality and *metadata* quality. Data quality refers to the characteristics of the dataset referenced by the metadata. Within geospatial applications, this can include parameters such as positional accuracy, precision, and timeliness, and are typically documented within the metadata referencing the dataset, whether the dataset is a raster image, coverage, or recorded spectrum (FGDC 2002; ISO, 2003). However, metadata quality refers to the characteristics of the metadataset itself, recognizing it as a distinct body of data that can be analyzed separately. Metadata quality makes no direct reference to the underlying spectra or field data collection protocols, as the case may be, for field spectroscopy applications. Therefore data quality will not be addressed further in any substantive manner as it is not within the scope of this thesis and instead the focus will be on metadata alone.

The concepts of metadata quality and completeness arise within the framework of metadata standards and it is on this foundation that they must be defined and developed as useful measures with meaning for data users. There is no established definition of quality and completeness for field spectroscopy metadata. Evaluation of existing standards can serve as a starting point to creating logical, rational, and useful quality and completeness criteria for such datasets.

2.6.2 Quality and completeness within existing metadata standards

Geospatial metadata quality has not been formally defined either in any standard or by any advisory body (FGDC, 2002; ISO, 2002, 2011; ANZLIC, 2007; INSPIRE 2009) responsible for issuing these standards. Rather, metadata fields assigned to the 'quality' modules or classes within existing standards refer to the quality of the dataset (such as a coverage or raster image), not the metadata itself. For example, the ISO 19113:2002 standard for quality principles for geographic data

> *''is applicable to data producers providing quality information to describe and assess how well a dataset meets its mapping of the universe of discourse as specified in the product specification, formal or implied, and to data users attempting to determine whether or not specific geographic data is of sufficient quality for their particular application"* (ISO, 2002).

This definition of quality is often expressed in quantitative and qualitative terms describing the positional accuracy, temporal accuracy, thematic accuracy, logical consistency, and completeness of the original dataset (FGDC 2002; ISO, 2003).

The concept of metadata quality is more commonly referenced in literature relating to general information science and research on the design and utility of metadata for digital data repositories. Even here however, the definition of metadata quality is an oblique one and has been characterized variously as "a true representation of the resource" (Margaritopoulos, 2008, p. 106), important to information seeking activities (Stvilia *et al*., 2004), an expression of fitness for purpose (Park, 2009) and supportive of interoperability and long-term curatorship and preservation (NISO, 2007).

Methods for assessing information quality have been applied to studies of metadata quality. These methods most commonly includes measures of dimensions referring to accuracy, conformance to expectations, logical consistency and coherence, accessibility, and timeliness, with up to thirty two individual items proposed within these categories (Bruce and Hillman, 2004; Stvilia *et al*., 2007; Ochoa and Duval, 2009). These quality dimensions can be further grouped into classes representing the causes underlying quality variance on each dimension, specifically those causes that are intrinsic (referring to a standard within a data user's conventions, norms, and language), relational (relationships between objects and their context) and reputational (the merit and reputation of the metadataset and its creators) (Stvilia *et al.*, 2007).

Quantification of metadata quality can provide information, whether directly or implicitly, about the metadataset, its suitability for a given purpose, the data repository in which it is stored, and the creators and/or owners of the data. Quantifying is useful to highlight challenging-to-acquire components of specification (Liolios *et al*., 2012). Metrics for metadata quality are mostly generated through automated processes and take various forms including:

- an ordinal scale 'good/moderate/poor/unusable' describing the overall quality of the metadataset (Currier *et al*., 2004)
- quantifying the problems themselves (ambiguity, inaccuracy, inconsistency, redundancy) as percentage of occurrence within a recordset (Stvilia *et al.*, 2007)
- accuracy as a measure of semantic distance between a metadata instance and the textual information it references (Ochoa and Duval, 2009)

- reputation of a metadataset as a linear combination of weighted sub-parameters including number of unique editors, edits, connectivity, reverts, registered user edits, anonymous user edits (Stvilia *et al.*, 2007)

Metrics are limited only by the inventiveness of the metadata analysts and the degree of informativeness these measures provide to data users.

Agreement on what constitutes metadata completeness is even more difficult to achieve than that for metadata quality. The reasons for this arise mostly out of the numerous and varied applications that metadata is created and used for, as well as the diverse standards inherently related to these applications, whether the applications are bibliographic, machine readability, searchability and discoverability by users, among others. Simply put, metadata fields for a dataset, however numerous, are not relevant for all resources (Ochoa and Duval, 2009). What defines completeness is “conditioned by characteristics of the resource type ... specifically by local metadata guidelines and best practices ... and modulated by characteristics of local communities” (Park, 2009, p. 220). Metadata completeness is described more consistently in terms of the advantages of creating a complete set in conforming to a given standard. A complete metadataset “should describe the resource as fully as possible” (Goovaerts and Leinders, 2012, p. 182), enables the user to “locate entities by the attributes the user intends to use” (Stvilia *et al*., 2004, p. 116), and “makes [a dataset] more trustworthy” (ANDS, 2011). Completeness metrics are almost exclusively derived through automated data mining processes and have most often been expressed as individual or combinations of weighted percentages of compliance statistics with a requisite set of metadata fields.

It can be summarized that quality and completeness parameters ultimately serve to give a data user the necessary information to make decisions about the utility of the metadata for a given purpose. These two attributes can be viewed as complimentary but individual measures that, in combination, provide a data user with a more comprehensive and less ambiguous assessment of a metadataset than either measure would on its own. For example, a metadataset assessed within the confines of a single metadata standard for a given application may be evaluated as high quality due to its logical consistency and ontological compliance, but can be incomplete according to the requirements of a data user. Likewise, a metadataset may be complete, but corrupted by syntactic and semantic errors. Therefore, both measures are necessary to enable the user to make intelligent and informed choices.

Metadata quality and completeness are factors that determine whether a metadataset is available for discovery in a metadata clearinghouse, or whether it passes through the data filtering systems of datawarehouses. In the context of sharing and distributing metadatasets for research and public access, it is incumbent upon the designers and managers of IT infrastructure software policies to ensure that they provide the data users with as rich and complete metadatasets as possible to permit them to make informed choices about whether a dataset is usable for a given purpose.

2.7 Conclusions

This review confirms the need for creating a metadata standard for field spectroscopy. A lack of formally identified metadata, the insufficiency of existing metadata standards in meeting the requirements of field spectroscopy standards, and limited or no implementation of quality control of field spectroscopy metadata in spectral databases and data sharing platforms results in scientists being unable to make informed decisions about whether datasets are suitable for a given purpose. It also reduces the potential for datasets to be discovered, shared, and re-used for multiple purposes.

A review of field spectroscopy metadata practices revealed that campaigns differ according to the purpose for which the data is collected, the geographic and environmental variables, and the target being sampled. The impact of inconsistent sampling and measurement protocols and the fact that these protocols largely remain undocumented through metadata together compromise valid intercomparison of datasets. The variety, but limited number of metadata documentation practices by field spectroscopy scientists and opinions on what constitutes application-specific and application-agnostic metadata means that there is no formally identified metadataset for field spectroscopy.

Existing geospatial metadata standards do not meet the requirements of the field spectroscopy community. They do not explicitly address field spectroscopy collection techniques, or the ontologies and data dependencies required to model the complex interrelationships among the observed phenomena as data and metadata entities.

There is no consensus on definitions of metadata quality and completeness, but what is clear is that they can be defined to suit the requirements of field spectroscopy metadata users. Issues of metadata completeness and quality are given little attention in spectral archives, and in cases where the user has the option for some degree of quality assessment, they are not enforced by the system.

As data sharing becomes more prolific and data users expect on-demand access, it is vital that data sharing exchange mechanisms and platforms incorporate metadata quality metrics and quality control for field spectroscopy datasets. This allows data users to make the best choices when searching for and selecting a dataset for a given application.

This review has identified the gaps in knowledge within the remote sensing community about what constitutes field spectroscopy metadata, how to document it, and how to meet data users' requirements for interoperability and quality assurance. From this emerges a framework for specific areas of enquiry to respond to these problems, where to look for guidance on building a standard, and the unique components of a field spectroscopy metadata standard that require focus.

Global-scale metadata exchange, intelligent and automated archiving processes for datasets, and quality controlled distribution of field spectroscopy data all require a concerted effort from the field spectroscopy community to first of all, identify their needs for robust, complete, and high quality metadatasets. Metadata standards, polices and quality control processes can then be developed and implemented on this foundation. Only when these conditions are met can field spectroscopy datasets be released into data warehouses and other data sharing platforms with maximum potential for discoverability and re-use.

Chapter 3 Critical metadata for spectroscopy field campaigns

Published in part as:

Rasaiah, B.; Jones, S.D.; Bellman, C.; Malthus, T. (2014). Critical Metadata for Spectroscopy Field Campaigns. *Remote Sensing Open Access,* 6, 3662-3680. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D., Bellman, C. (2012). Critical Metadata Protocols in Hyperspectral Field Campaigns for Building Robust Hyperspectral Datasets. *Proceedings of the XXII ISPRS Congress*, August 26 – September 1 in Melbourne, Australia. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2011). Designing a Robust Hyperspectral Dataset: The Fundamental Role of Metadata Protocols in Hyperspectral Field Campaigns. *Proceedings of the GSR 1*, November 28 – 30 in Melbourne, Australia. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2011). Building Better Hyperspectral Datasets: The Fundamental Role of Metadata Protocols in Hyperspectral Field Campaigns. *Proceedings of the Surveying & Spatial Sciences Conference,* November 21-25 in Wellington, New Zealand. (peer-reviewed)

3.1 Introduction

This chapter addresses research question #1, ‘What are the key elements of a core metadataset for all field spectroscopy applications?’ The results and analysis of an international expert panel survey (n=90) are presented on the use and utility of metadata for field spectroscopy sampling. Next, a core set of metadata parameters for all spectroscopy campaigns is proposed based on the survey analysis.

A review of field spectroscopy protocols, including their diversity and commonalities, as well as the rationale for a metadata standard is presented in Chapter 2. A standardized methodology for defining and storing metadata must be closely aligned to *in situ* data collection practices, but currently, no such methodology for documenting *in situ* spectroscopy metadata exist. To address the requirements for efficient and viable intercomparison and fusibility of datasets generated from quantitative field observations, it is necessary to identify which metadata parameters are common to all campaigns, which are unique to specific applications, and which among these are critical to all campaigns. The aim of this chapter is to present a way of prioritizing metadata that can be applied to any *in situ* field spectroscopy metadata standard that is practical, flexible enough to suit the purpose for which the data is being collected, and/or has sufficient legacy potential for long-term sharing and interoperability with other datasets.

A field spectroscopy metadata standard must handle both generic and application-specific information. Figure 3.1 illustrates a conceptualized prototype field spectroscopy metadata standard that can be informed by the research in this chapter and subsequent chapters. It consists of fundamental building blocks including core metadata that is common to and requisite for all applications, application-specific metadata, and additional metadata modules from existing standards and paradigms to enhance quality, discoverability, and interoperability.

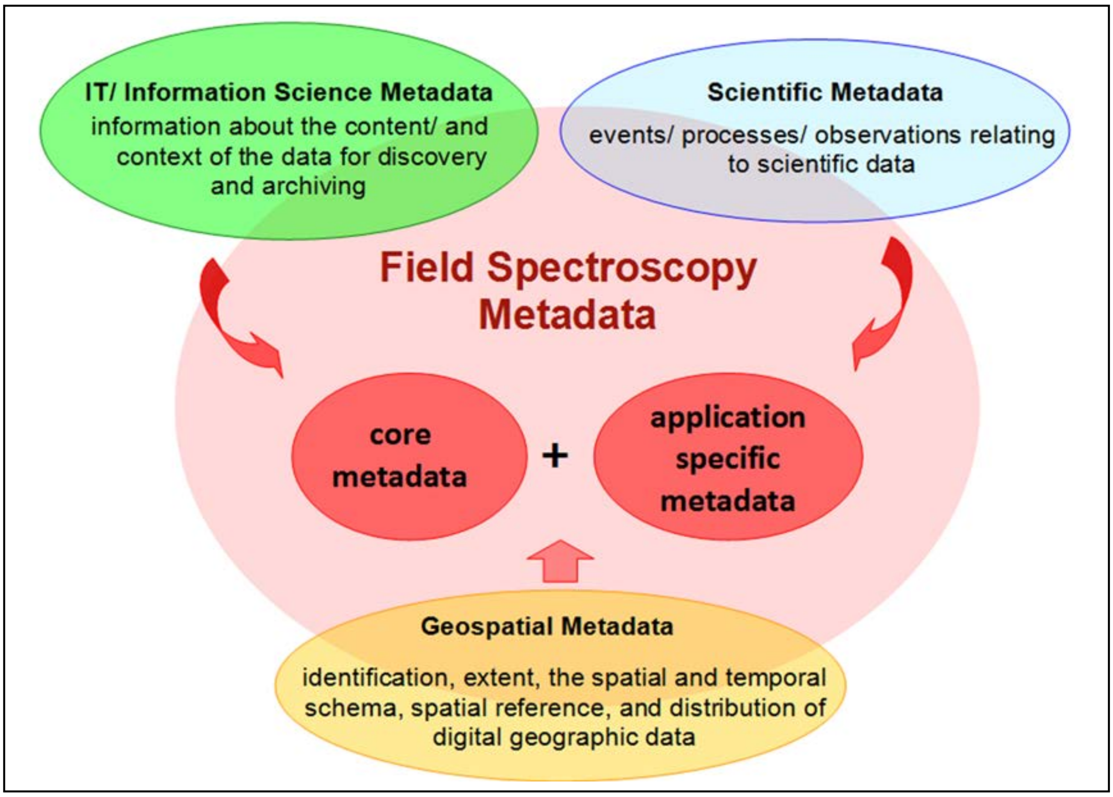

IT/ Information Science Metadata
information about the content/ and context of the data for discovery and archiving
Scientific Metadata
events/ processes/ observations relating to scientific data
Field Spectroscopy Metadata
core metadata
+
application specific metadata
Geospatial Metadata
identification, extent, the spatial and temporal schema, spatial reference, and distribution of digital geographic data

Figure 3.1 A conceptualized prototype field spectroscopy metadata standard

In order to create a full suite of metadata definitions, first the unique conditions under which field spectroscopy campaigns operate must be identified and described. Secondly, these definitions must be robust and sufficiently versatile to accommodate the breadth of campaigns commonly conducted. Thirdly, the metadata standard must overcome the obstacles to interoperability and quality assurance expressed by data users within the remote sensing community (Reusen *et al.*, 2007). To solve these issues, a survey of spectroscopy experts was conducted.

3.2 Consulting the experts

To define a common set of metadata standards, the opinion of the spectroscopic science community was canvassed. An expert panel was convened for guidance through the process. To establish membership in this group, one, or more, of the following criteria was met by each participant: 1) be an established investor in the quality of the spectroscopic metadata; 2) have experience in, and possess understanding of theory and methods of spectroscopic data capture; and, 3) express an interest in developing techniques for increased sharing and intercomparison of their datasets with other remote sensing research groups. The group was representative and comprised a broad spectrum of expertise, but was not comprehensive.

A pilot survey was introduced to a group of remote sensing scientists at the 7th EARSeL (European Association of Remote Sensing Laboratories) workshop in Edinburgh, Scotland, in 2011. Refinements to the survey were made based on the response from the test group and an improved and expanded online survey was launched later in 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 exchange across a range of campaigns. Approximately 200 metadata fields were presented to the survey participants. A large proportion of the fields were obtained from Malthus and Shirinola (2009) and appended with application-specific metadata proposed by select field spectroscopy experts through personal interviews. Appendix A lists all the metadata elements in each category. Table 3.1 is a listing of the generic and application-specific metadata categories included in the survey.

Generic campaign metadata	Application-specific metadata
instrument	vegetation
reference standards	woodland and forest
calibration	agriculture
spectral signal properties	soil
illumination information	mineral exploration
viewing geometry	snow
environment information	urban environments
atmospheric conditions	marine and estuarine
general project information	underwater substratum targets
location information	
general target and sampling information	

Table 3.1 Categories of metadata fields in the survey

The audience was an international panel of scientists with expertise in *in situ* field spectroscopy, who were asked to respond on an anonymous basis. The survey was completed by 90 participants from organizations and institutes with a history of research on the relevant topics and included the NERC FSF (National Environment Research Council Field Spectroscopy Facility, UK), DLR (German Aerospace Center, Germany), CSIRO (Commonwealth Scientific and Industrial Research Organisation, Australia), RSL (Remote Sensing Laboratories, Switzerland), EPA (Environmental Protection Agency, USA), numerous other North American and European university research labs and participants from the commercial sector. 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 possible throughout the survey for further input in each metadata category. A copy of the survey is supplied in Appendix A.

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 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, each defined accordingly:

- '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 dataset);
- 'not useful now but has legacy potential' (not directly relevant to the associated field spectroscopy data but potentially has use for a related hyperspectral product)
- 'not applicable' (this metadata is not relevant)

These four rankings were chosen to inform a prioritization model for criticality for a metadata standard.

3.3 Results of the survey

This section presents an analysis of the survey participants' responses. Sections 3.3.1 (metadata categories) and 3.3.2 (metadata elements) present the quantitative results and Section 3.3.3 is a synopsis of the free-form comments from the participants. The quantitative analysis enabled identifying a core metadataset for all applications, which includes 'Viewing Geometry', 'Location Information', 'General Target and Sampling Information', 'Illumination Information', 'Instrument', Reference Standards', 'Calibration', 'Hyperspectral Signal Properties', 'Atmospheric Conditions', and 'General Project Information'. Quantitative analysis also identified the varying degrees of consensus, both inter- and intra- category, among the generic and application-specific categories. The following section provides a more detailed discussion on the derivation of the core metadataset and examples of responses to specific categories.

3.3.1 Quantitative results – metadata categories

Figure 3.2 identifies the areas of expertise of the participants. Each respondent was asked to designate themselves as experts in one or more fields. Areas of spectroscopy research beyond this scope, as stated by the respondents, included atmospheric studies, calibration and validation activities for airborne sensors, and wetlands and peatlands research (all grouped within the 'other' category). The largest group of experts were from the agriculture (40), forest/woodland (39), and soils (27). The smallest sample was from the snow research area (2). The range of group sizes sampled required both parametric and non-parametric statistical methods to analyze the results.

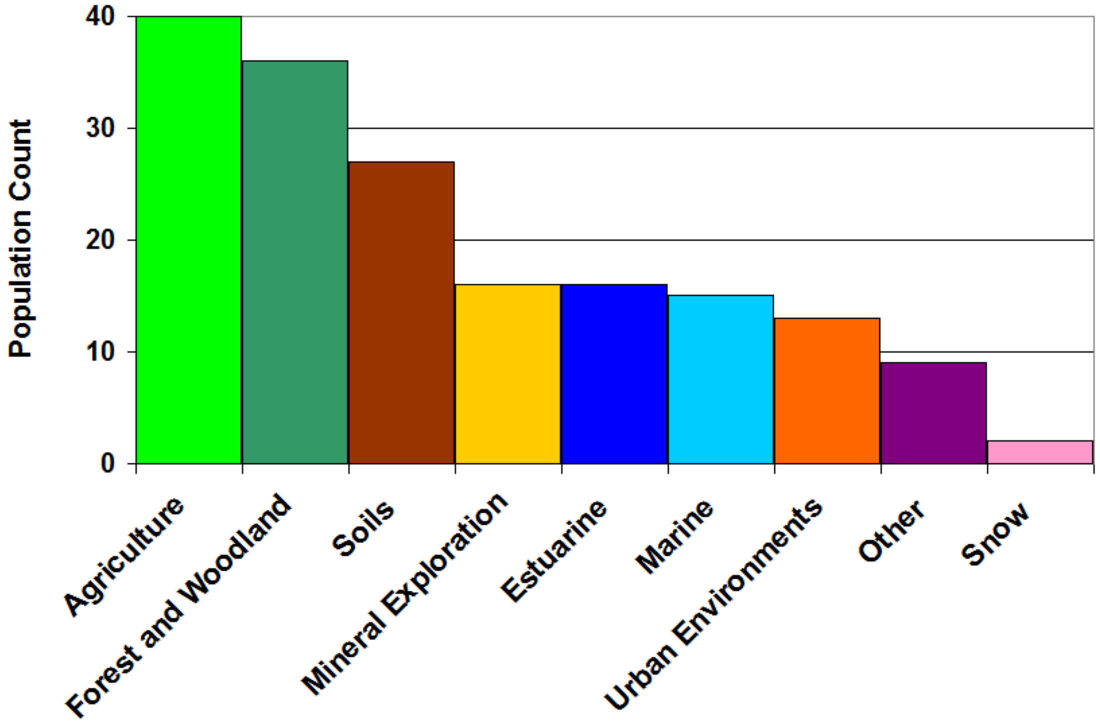


Figure 3.2 Areas of expertise self-nominated by survey respondents (*n=90)*

The survey was designed to gather information on two metadata categories – generic and application-specific (Table 1). Generic campaign metadata refers to subsets of metadata common to all campaigns, regardless of the application or purpose for collection and includes, for example: instrument (Figure 3.3), calibration (Figure 3.4), reference standards, and viewing geometry information. Application-specific (or target) metadata is associated with the purpose of the campaign and the type of target being measured; this category is separated into subsets including vegetation, snow, soil, mineral exploration and marine targets. For each subset of metadata, whether in the general or target specific categories, a criticality index of four measures ('critical'/'useful'/'not useful now but has legacy potential'/'not applicable') was used by each respondent. The ordinal criticality (or degree of importance) rankings were standardized to numerical values (ranging from 0 for 'N/A' to 3 for 'critical' to permit statistical analysis of variance. The criticality rankings for all metadata categories are presented in Appendix B.

Variation in ranking of criticality varied for each metadata category. As examples, Figures 3.3 and 3.4 depict the frequency of ranking for two subsets of general campaign metadata fields in the 'instrument' (2) and 'calibration' (3) metadata categories, responded to by the scientists.

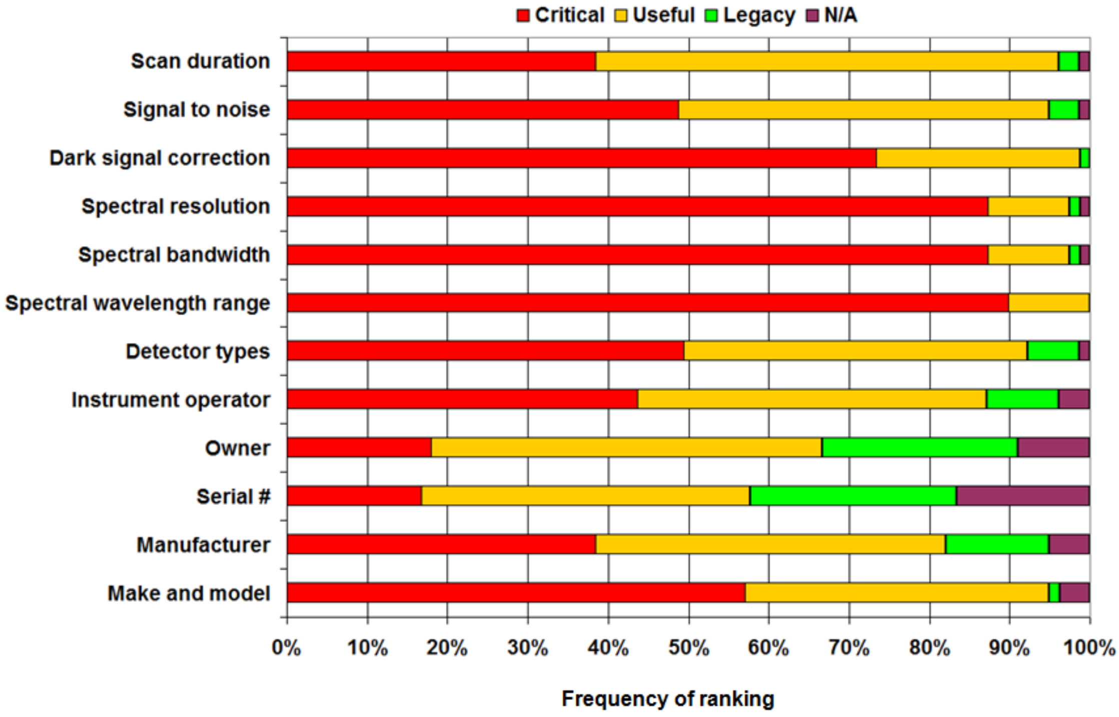


Figure 3.3 Frequency of criticality ranking for 'instrument' metadata (*n=79*)

In the 'instrument' category, assignment of 'critical' to a given metadata field ranges from 90% for 'spectral wavelength range' to less than 20% for [instrument] 'serial number'. The former field is highlighted as the only one with no 'N/A' or 'legacy potential' ranking, suggesting that it is regarded as a fundamentally crucial metadata field and warrants inclusion in all field spectroscopy metadata protocols. The latter field, ‘serial number’ implies that it is perceived as less important to respondents, despite its crucial role in databases and other information systems in tracking the history of use and calibration of an instrument. Respondents’ familiarity with and knowledge of metadata storage within information systems may have an impact on the frequency of ‘serial number’ being ranked ‘critical’. The implication is that most respondents do not have direct use for ‘serial number’ but in many cases it is vital to maintaining an accurate history of the use and calibration of instrument, and could be used within an information system to automatically populate detailed metadata for a given instrument.

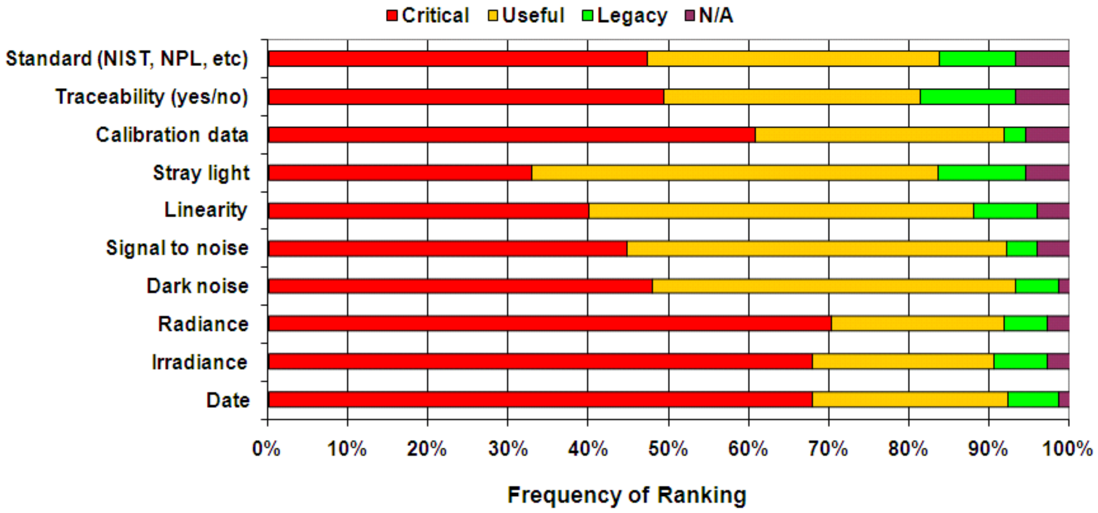


Figure 3.4 Frequency of criticality ranking for ‘calibration’ metadata (*n=68*)

In the ‘calibration’ category, there is a lower disparity in assignment of ‘critical’ ranking across the fields with a range between 70% (‘radiance’) to 32% (‘stray light’). This implies a greater degree of consensus opinion on the influencing factors of calibration activities on both the hyperspectral data and the end-products for which the data and metadata will be utilized e.g. end-member retrieval, land cover classification, satellite sensor validation and BRDF modelling. It must be noted that different interpretations of a given metadata element as well as the number of metadata elements provided in a category may have influenced the ranking given by some respondents. For example, comments provided by the respondents for the 'calibration' category indicated that for one respondent, there was ambiguity as to whether the 'calibration data' field referred to the spectral measurements against the calibration standard, and another suggested that additional fields should be provided indicating whether the calibration was relative (to a spectral plate as is) or absolute to a specific NIST-traceable spectralon panel.

Some of the variation in the 'instrument' category 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. Figure 3.5 shows preferred instruments by expert group.

The unique technical aspects of each instrument may have a bearing on the particular metadata fields that an operator chooses to include in their metadataset; these may include instrument housing (for extreme weather conditions or non-terrestrial campaigns), the degree to which an instrument has been customized for a particular application and whether it is a prototype, and sensor behaviour affected by manufacturer design. As an example, PANalytical Boulder, Inc. (formerly ASD) supplies a device known as a 'Scrambler' for its FieldSpec models to compensate for spectral discontinuities due to non-uniformity of field-of-view across the sensor bank

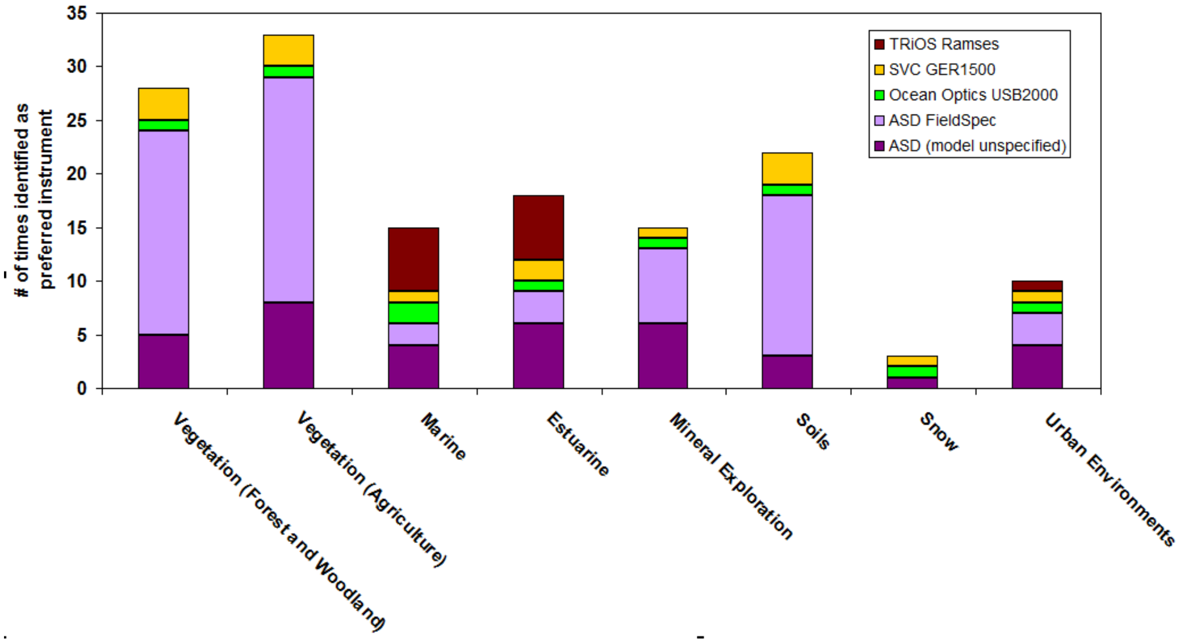


Figure 3.5 Top 5 preferred instruments by expert group

fibreoptics. It may be worthwhile investigating the proportion of FieldSpec users who incorporate biased field-of-view calculations into their spectral data modelling and how this impacts their calibration, viewing geometry, and reference standards protocols and the subsequent designation of metadata that are critical to account for this.

Levels of agreement between respondents across all categories were measured using the intraclass correlation coefficient (ICC) (Tabachnick and Fidell, 2007). This method was most amenable to the ordinal rankings and adjusted for the scale of measurement, which varied across the categories. Figure 3.6 shows a measure of consensus among the respondents, from highest to lowest, across the metadata categories. The trend for consensus is determined mostly by the population size and composition of the respondents for each group. Generally, the smaller and more specialized the expert group and the more specialized the metadata category, the higher the degree of consensus within it. The four metadata category groups with almost perfect consensus were ‘Underwater Substratum Target’ (ICC=0.922), ‘Marine and Estuarine’ (ICC=0.847), ‘Snow Campaign’ (ICC=0.824), and ‘Agriculture Campaign’ (ICC=0.802). The top eight metadata categories for consensus ranking were all application-specific. The ‘Vegetation Campaign’ metadata category is the only application-specific category that exhibits ‘Fair’ consensus (ICC=0.381).

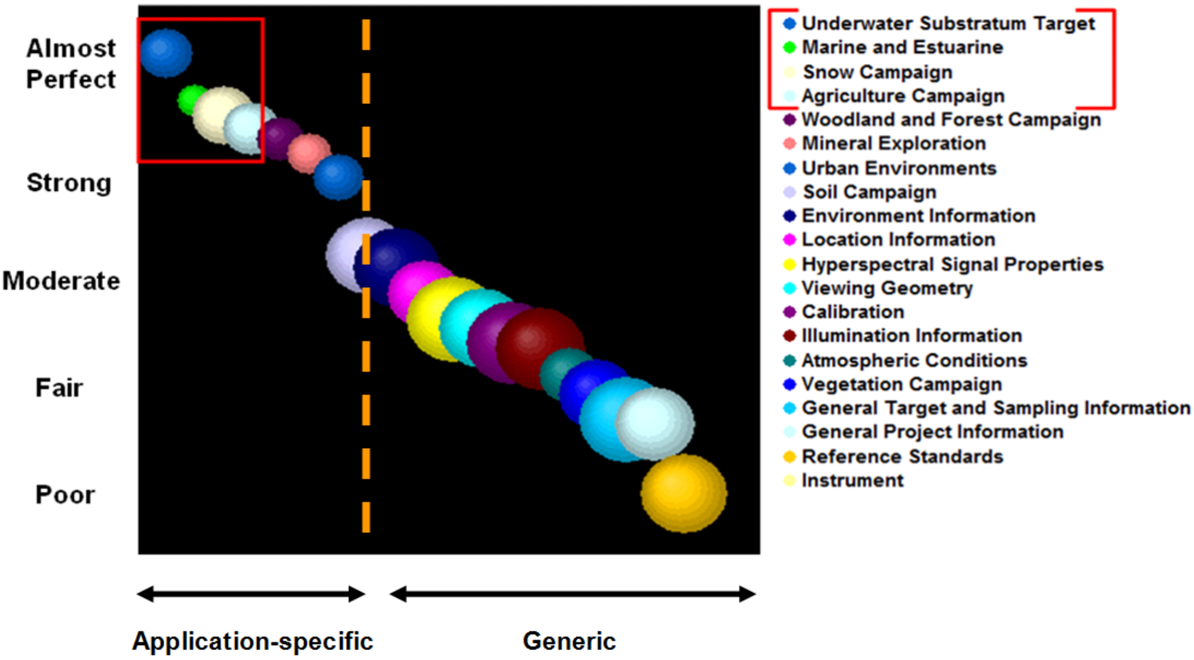


Figure 3.6 Group consensus measure for metadata field criticality among the respondents, from highest consensus (‘almost perfect’) to lowest (‘poor’), across the metadata categories. Size of sphere denotes numbers of respondents in each category.

Note: The estimator for each intraclass coefficient of variance measure is the same, whether the interaction effect is present or not. ICC can be interpreted as follows: 0-0.2 indicates poor agreement: 0.3-0.4 indicates fair agreement; 0.5-0.6 indicates moderate agreement; 0.7-0.8 indicates strong agreement; and >0.8 indicates almost perfect agreement.

As an example of responses to application-specific metadata, Figure 3.7 depicts the frequency of ranking for marine 'substratum target' metadata which was responded to by a smaller population of scientists (40). The substratum target metadata category most commonly refers to submerged biological marine targets such as seagrass and corals, but can include any target on a submerged surface. For all fields in this category, there was a greater consensus between the four available rankings than in the non-specialized metadata categories, 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 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. The ranking results for all metadata categories are found in Appendix A.

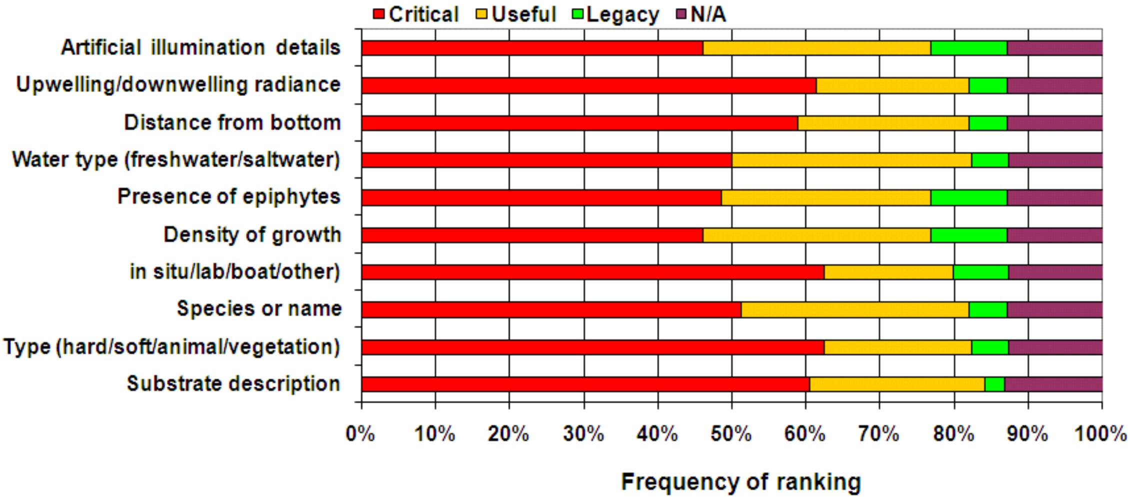


Figure 3.7 Frequency of criticality ranking for 'substratum target' metadata (*n=40*)

The criticality rankings 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 3.8, group means ranged between 'useful' and 'critical' for both the marine and non-marine scientists.

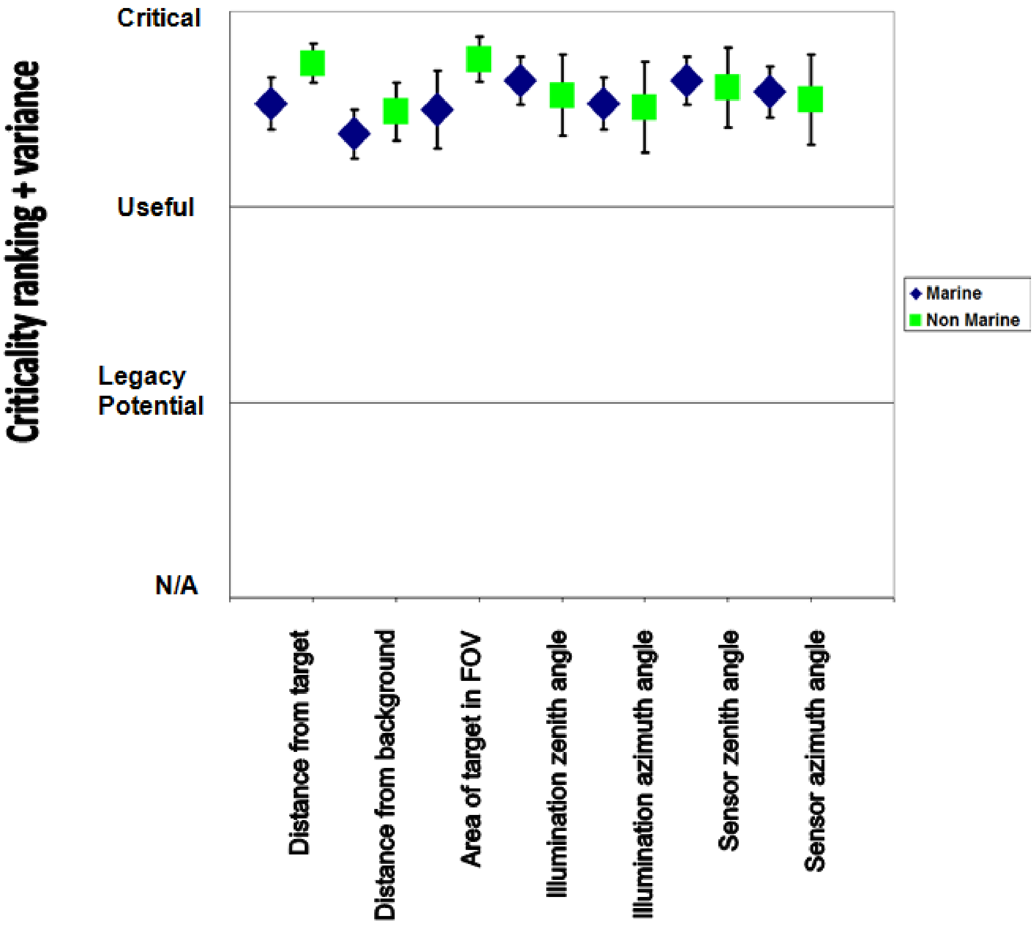


Figure 3.8 Group means and variance in 'viewing geometry' metadata (Marine and Estuarine *n_1=18,* Non Marine *n_2=49)*

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. There was 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 reflectance anisotropy, either in the atmosphere, water column, or due to the target surface properties, is of equal importance to all campaigns. Variance in criticality ranking for viewing geometry metadata was consistently higher among non-marine scientists, implying that there exists greater consensus among field spectroscopy scientists from the same expert group.

Figure 3.9 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 were uniformly higher for all metadata fields in this category, and variance was uniformly lower than for rankings assigned by non-marine scientists. Underwater campaigns can vary in terms of the application of the data being investigated and the protocol necessary to capture the required data. Targets can include seagrass, macro-algae, corals and sponges and spectral measurements may be taken above surface or below surface; opinions differ on how inclusive a metadataset must be to document environmental and target properties (Bhatti *et al*., 2009 and Dekker *et al*., 2010). The unique complexities of measuring targets and controlling influencing variables in a marine environment can be understood best by the scientists with in-field expertise (details of these complexities are discussed in

Chapter 2). These considerations and the results of the survey strengthen the implication that consensus and agreement are dependent upon the respondents' area of expertise.

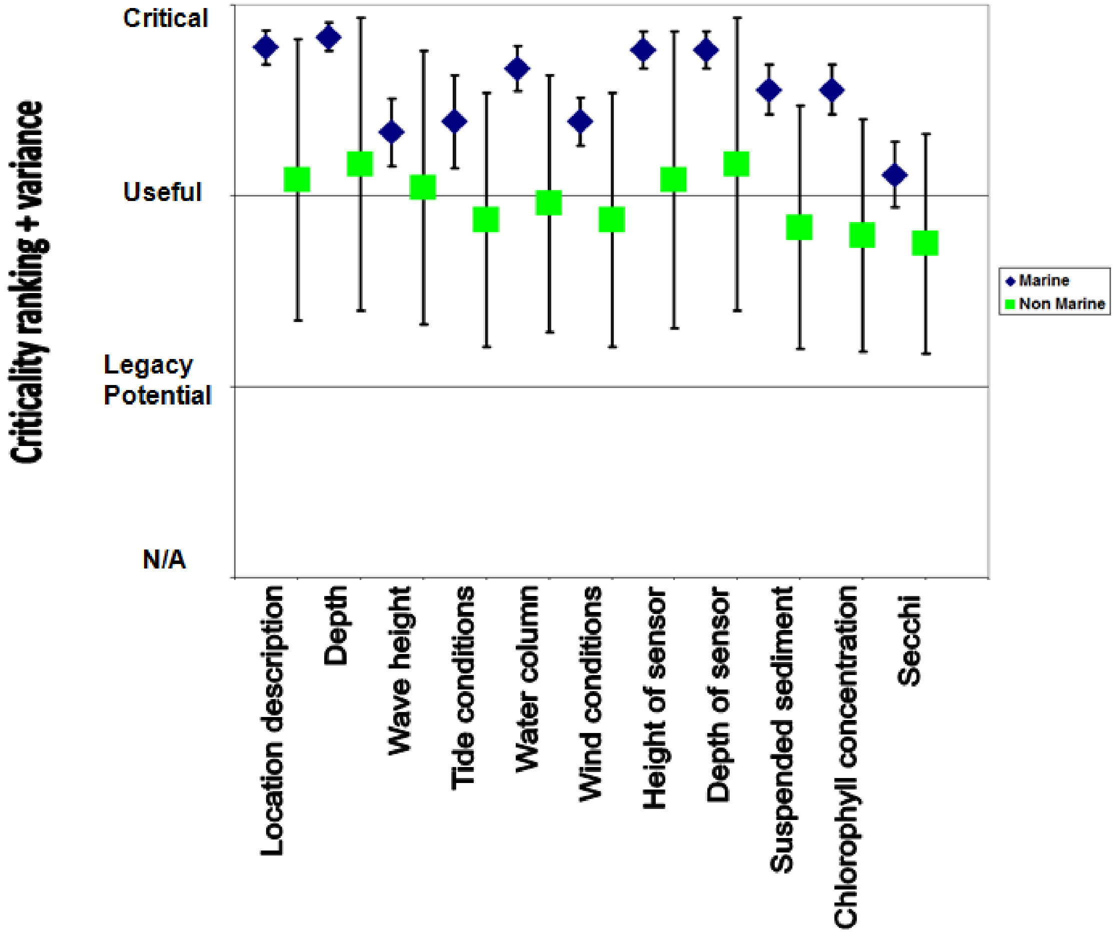


Figure 3.9 Group means and variance in 'marine and estuarine environmental conditions' metadata (Marine and Estuarine n_1=18, Non Marine n_2=49)

3.3.2 Quantitative results – metadata elements

An optimal standard that would meet basic requirements for practical implementation, flexibility, and longevity of a dataset, would be constructed using the most essential ('critical') fields that are common to all campaigns. Such a standard would need to accommodate variation in response by expert groups across the metadata categories, as well as the logistics, aims, and goals inherent to each campaign. To explore this, thresholds for inclusion of a metadata field in a protocol were determined based on its criticality. The ordinal and non-parametric nature of the data necessitated a suitable suite of tests that could adjust for the scale of measurements in each metadata category and permit repeatability and intercomparison for all categories. The first phase of analysis was conducted using a stringent test for calculating the likelihood of a dichotomous outcome -- either a field must be included in a protocol (the 'critical' fields extracted from the responses) or it is excluded – and this was achieved via binomial analysis. The second phase identified additional metadata fields that demonstrated ranking aberrant to the other fields in the category; this was accomplished via scale statistics for describing internal consistency and interrelation between items in a given category; in other terms, the usefulness of every metadata item being in that particular category.

Table 3.2 shows the binomial test results and scale statistics analysis on calibration metadata where the frequency of critical rankings were compared to the non-critical ('useful'/'legacy potential'/'NA'). The binomial tests were conducted such that the proportion of 'critical' ratings were compared with a baseline proportion of 0.5. Metadata fields that have been designated as critical more than 50% of the time have been highlighted in bold. They comprise half of the metadata fields within the category and include 'Date', 'Dark Noise', 'Signal to Noise', 'Stray Light' and 'Calibration Data'.

While binomial analysis is useful for generating a dichotomous outcome, it fails to calculate the proportion of individual 'useful'/'legacy potential'/'NA' measures and

Metadata category		Observed Prop.	p-value	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Date	Critical	0.68	0.002	**26.443**	**0.596**	**0.888**
	Non-critical	0.32				
Irradiance	Critical	0.32	0.002	25.786	0.649	0.885
	Non-critical	0.68				
Radiance	Critical	0.30	0.001	25.909	0.660	0.885
	Non-critical	0.70				
Darknoise	Critical	0.52	0.818	**26.210**	**0.686**	**0.884**
	Non-critical	0.48				
Signal to Noise	Critical	0.55	0.422	**26.638**	**0.579**	**0.890**
	Non-critical	0.45				
Linearity	Critical	0.40	0.105	25.623	0.636	0.886
	Non-critical	0.60				
Stray Light	Critical	0.67	0.005	**25.063**	**0.660**	**0.884**
	Non-critical	0.33				
Calibration Data	Critical	0.61	0.081	**25.416**	**0.620**	**0.887**
	Non-critical	0.39				
Traceability (yes/no)	Critical	0.49	1	24.547	0.631	0.887
	Non-critical	0.51				
Standard (NIST/NPL, etc.)	Critical	0.47	0.728	**24.097**	**0.723**	**0.880**
	Non-critical	0.53				

Table 3.2 Binomial test results and scale statistics analysis for 'calibration' metadata (n=78)

Metadata fields that have been designated as critical more than 50% of the time through binomial testing have been highlighted in bold. Scale statistics examined those metadata fields that generated relatively extreme values for Corrected Item-Total Correlation and Cronbach's Alpha, or a strong effect on the scale mean and variance if they were deleted; these have also been highlighted in bold.

therefore an additional method of analysis using scale statistics is necessary for this purpose. Scale statistics examined those items that generated relatively extreme values for corrected item-total correlation and Cronbach's Alpha, or a strong effect on the scale variance if they were deleted; these also have been highlighted in Table 2. Cronbach's Alpha, α, is a reliability coefficient that is a useful measure of internal consistency of inter-rater agreement (Bland 1997, Santos 1999, Gliem and Gliem, 2003) on the metadata fields in each category, and ranges from 0 ≤ α < 6 for unacceptable and/or poor internal consistency and any value α ≥0.9 is considered excellent. The corrected item-total correlation is the correlation of the metadata field with the summated score for all other metadata fields in the category. Identifying those metadata fields that have the strongest effect on the inter-rater agreement levels by their effect on Cronbach's Alpha if they are removed warrants investigation as to why they exhibit a trend of rating different from other metadata fields in that category, and invites consideration for their exclusion from the category or potentially being assigned a status more important than the other items (Howard and Forehand, 1962; Henrysson, 1963).

In the 'Calibration' category, 'Cronbach's Alpha if Item Deleted' value is relatively high for all metadata fields in the category, representing good internal consistency within the category. However, 'Standard' is a field not previously identified through binomial testing but indicates the need for further study into this metadata field as to the causes for its impact on the degree of agreement among respondents. It is closely correlated with the results for 'Traceability'. This implies that for most of the survey participants, documenting frequency and results of calibration is important, but the details of the reference standard less so. The survey data permits speculation only at this point but the reason for choosing not to document the reference standard may arise from the level of significance a scientist assigns to the instrument itself being a factor in the recorded spectra, and the extent to which they are willing to collect and analyze ancillary calibration data to mitigate any spectral discrepancies resulting from the instrument. Similar results from all metadata categories illustrate the ambiguity presented by fields that lie below or near the threshold, therefore not being representative of the majority of respondents but having been identified as ‘critical’ frequently enough to support inclusion in a customized metadata standard for given campaigns.

A foundation for a standard can be established by including those metadata categories with high overall rankings of criticality and internal consistency. Table 3.3 illustrates ranking, from highest to lowest, of metadata categories in terms of field ranking means, variances, and analysis of variance between fields.

			ANOVA between item means for respondents (per metadata category)		
Metadata category	**Item means**	**Item variances (mean)**[a]	**df**	**Friedman Test**	**p-value**
Environment Information	1.822	0.597	4	17.704	0.001
Mineral Exploration	1.822	0.841	15	189.355	<0.001
Snow Campaign	1.890	1.183	10	58.645	<0.001
Soil Campaign	2.057	0.657	20	192.433	<0.001
Woodland and Forest Campaign	2.068	0.705	8	40.058	<0.001
General Project Information	**2.103**	0.469	5	146.004	<0.001
Atmospheric Conditions	**2.153**	0.425	6	117.125	<0.001
Urban Environments	2.189	0.866	10	116.875	<0.001
Marine and Estuarine	2.199	1.014	10	69.282	<0.001
Underwater Substratum Target	2.216	1.139	9	28.481	0.001
Vegetation Campaign	2.231	0.462	15	159.044	<0.001
Agriculture Campaign	2.242	0.690	8	84.810	<0.001
Hyperspectral Signal Properties	**2.352**	0.561	18	294.364	<0.001
Calibration	**2.379**	0.605	9	54.067	<0.001
Reference Standards	**2.388**	0.628	6	79.651	<0.001
Instrument	**2.393**	0.484	18	310.47	<0.001
Illumination Information	**2.420**	0.474	5	86.771	<0.001
General Target and Sampling Information	**2.477**	0.446	13	105.327	<0.001
Location Information	**2.489**	0.464	7	53.578	<0.001
Viewing Geometry	**2.571**	0.358	6	12.624	0.049

Table 3.3 Ranking of metadata categories by frequency of critical rankings and between-field variances

All metadata categories that surpassed the threshold mean (2.0) for inclusion in the model metadata standard (ten categories were identified) have been highlighted in bold.

A Friedman Test was run against the non parametric data to measure variability in ordinal criticality rankings in each category (Tabachnick and Fidell, 2007). The Friedman Test measures the difference between the observed rankings per respondent for each metadata category against a baseline of uniform rankings between respondents with α = 0.05 and the results show that for each category the differences between respondents is statistically significant for values p<0.05.

It can be assumed that for any given campaign, an ideal or model metadata standard would include both the generic campaign metadata (up to eleven categories) and at least one application-specific category, creating a total of 12 metadata categories. The item mean for a given category, as shown in Table 3.3, incorporates the compound measure of the frequency of 'critical', 'useful', 'legacy potential', and 'N/A' rankings. The more often that given fields in the metadata category were ranked 'critical' or 'useful', the higher the item means values for that category. This accounts for metadata categories with low inter-item consensus between respondents, such as 'Reference Standards' (ICC=0.224) and 'Instrument' (ICC=0.185) but high overall rankings for the metadata fields in that category. In Table 3.3 all generic (non-target-specific) metadata categories that surpassed the threshold mean (2.0) for inclusion in the model metadata standard (ten categories were identified) have been highlighted in bold**.** A mean greater than 2.0 for a given category means that on average, the fields in that category have a minimum overall ranking exceeding 'useful' (2.0) and a maximum overall ranking of 'critical' (3.0). All general campaign metadata categories surpassed this threshold, except for 'Environment Information', with a mean criticality ranking of 1.822. Therefore, those metadata categories necessary for inclusion in the model metadata standard are (in order of mean criticality ranking): ‘Viewing Geometry’, ‘Location Information’, ‘General Target and Sampling Information’, Illumination Information’, ‘Instrument’, ‘Reference Standards’, ‘Calibration’, ‘Hyperspectral Signal Properties’, ‘Atmospheric Conditions’, and ‘General Project Information’.

3.3.3 Additional qualitative feedback from the survey participants

Designing a standard benefits from both the quantitative data and the recommendations provided by the respondents. The comments section in the survey was an alternative, non-systematic method of canvassing the opinion of the field spectroscopy community on metadata, what purpose they believe it serves for their research activities, and what they believe are best practices for metadata documentation. The spectrum of free-form commentary ranged from general remarks (the utility and benefits of the survey, the necessity to create a standardized way of documenting field spectroscopy activities and variables, and additional considerations for creating a metadata standard) to suggestions on additional metadata to include for specific applications. Some of the suggestions and comments from the participants included:

“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”

“there is no end to metadata!”

(Rasaiah, 2011)

More than fifty additional metadata fields across many categories were suggested by the respondents. They provide a strong ancillary set of data to the quantitative results for informing design of a robust standard capable of accommodating a broad selection of campaigns in field spectroscopy. The application-specific metadata recommended by participants was incorporated into the metadatasets presented in Chapter 4.

3.4 The core metadataset

A core set of metadata is proposed based on those categories and fields identified as critical most often by the respondents. The results indicate that the categories meeting this criteria are 'Viewing Geometry', 'Location Information', 'General Target and Sampling Information', 'Illumination Information', 'Instrument', 'Reference Standards', 'Calibration', 'Hyperspectral Signal Properties', 'Atmospheric Conditions', and 'General Project Information'. Consequently, the core metadataset must include the minimum ten generic metadata categories and at least one application-specific category, for a total of eleven (Figure 3.10).

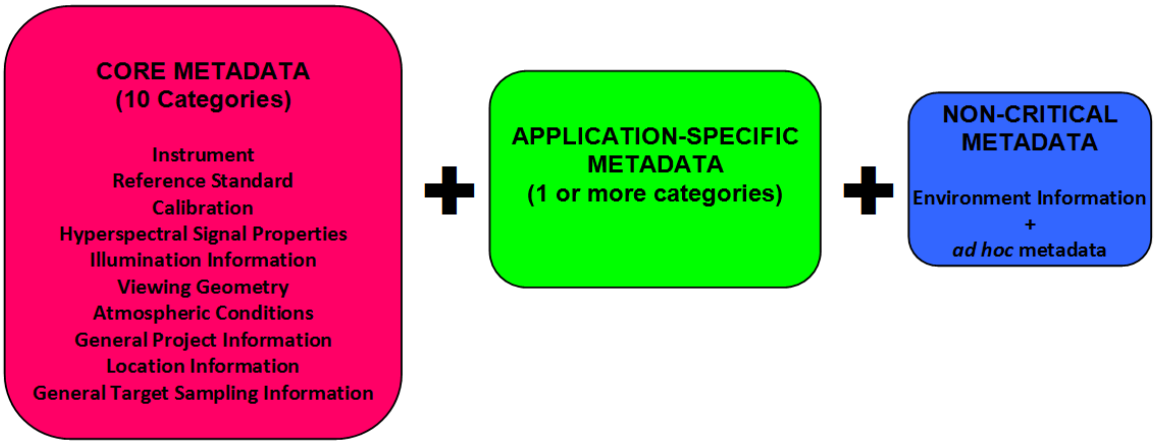


Figure 3.10 A metadataset for a given field spectroscopy campaign, including the core set common to all campaigns, application-specific metadata, and non-critical metadata

Other categories, within both the generic and application-specific metadata divisions, may be included to enhance the usefulness and legacy potential of the field spectroscopy metadataset. Appendix A presents the core metadataset of generic campaign metadata with critical and optional fields within each category, generic campaign metadata outside the core metadataset that may be included to enhance the robustness of a metadataset, and the critical and optional metadata elements for each application-specific campaign. Among the generic campaign metadata categories, only one category met the threshold for exclusion from the core set, ‘Environment Information’, and none of the metadata elements within the category were designated overall to be critical.

Application-specific metadata are presented in Appendix A and show a mix of critical and non-critical fields as designated by respondents for each target. ‘Underwater Substratum Target’ and ‘Agriculture’ have the highest ratios of critical to optional fields, in contrast to ‘Woodland and Forest’ target metadata where no fields were designated overall as critical.

For both generic and application-specific categories, there are subsets of critical metadata fields, identified by both binomial analysis and scale statistics, and ambiguously ranked metadata fields that warrant further investigation as to their inclusion or exclusion. Establishing 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 be especially useful for those fields that have been designated as both ‘critical’ and ‘N/A’ in almost equal proportion.

The versatility of a metadataset can be increased by including both the critical fields and those difficult to identify as critical. Group membership is an influencing factor on criticality rankings within a metadata category. Consensus is highest among expert groups for those categories directly related to their area of specialization, as exemplified by the high consensus and low variance for the marine scientists in the 'marine conditions' category. This indicates that a metadata standard designed for specific applications is best informed by the expert group most closely associated with research involving those applications. Overall the results provide an informed and detailed summary of what is required across many campaigns, with the fields identified as critical most often by respondents being the core metadata set that must be including in all standards.

3.5 Conclusions

The survey results provide the key elements of a metadataset that can be applied to any field spectroscopy metadata standard that is practical, flexible enough to suit the purpose for which the data is being collected, and/or has sufficient legacy potential for long-term sharing and interoperability with other datasets. The survey respondents helped to identify the key elements of a core metadataset critical to all field spectroscopy campaigns, as well as recommend additional metadata to increase the versatility of a metadataset, both for application-specific metadata and generic metadata.

A core metadataset must include 'Viewing Geometry', 'Location Information', 'General Target and Sampling Information', 'Illumination Information', 'Instrument', 'Reference Standards', 'Calibration', 'Hyperspectral Signal Properties', 'Atmospheric Conditions', and 'General Project Information' and at least one application-specific metadata category, depending on the type of target being sampled. The inclusion of additional categories, relating to both generic and application-specific metadata, serve to enhance the robustness of the dataset. The composition of each category is a factor of those metadata fields that were easily identified as critical (through binomial analysis in the 'Calibration' category, for example) and those that are difficult to designate. Overall, the results from the binomial and scale measurement testing prompt two important questions: i) whose opinion among the experts can be used as a basis for designating a metadata field as critical, and supported by what rationale?; ii) Is fitness-for-purpose an additional dynamic that must be accounted for when designing a metadata standard?

Consensus is highest among experts within the same field, and within categories most closely related to their area of knowledge. This was illustrated by marine scientists who showed lower variance in response and higher overall criticality rankings in the 'Marine and Estuarine Environmental Conditions' metadata category than did their non-marine counterparts in the same category. The trend for consensus amongst all categories, measured using the intraclass correlation coefficient, demonstrates that application-specific metadata with smaller but more specialized groups of experts have the highest level of agreement between respondents on the criticality rankings for each field.

The survey results and subsequent analysis provide answers to the problem of identifying critical field spectroscopy metadata with the following information:

- metadata categories that have the highest overall criticality rankings
- metadata fields that can be easily identified as critical to all campaigns

- metadata fields that are identified ‘critical’/’useful’/’legacy potential’/’NA’ most frequently
- the impact of group membership on determination of what is critical in a given metadata category
- consensus trends among groups in both generic and application specific metadata categories

Adapting the core metadataset as a standard for facilitation in data exchange is the best way forward for ensuring interoperability, intercomparison, and wide-scale sharing of high quality field spectroscopy metadata. This is the ideal solution to the problem of absent or ill-defined geospatial metadata standards currently in place that do not address the specific needs of field spectroscopy scientists.

Chapter 4 Identifying additional metadata for specific field spectroscopy applications and supporting interoperability with other metadata standards

Published in part as:

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2012). A Novel Metadata Schema for *in situ* Marine Spectroscopy. *Proceedings of Geospatial Science Research Symposium 2*, December 10-12 in Melbourne, Australia. (peer-reviewed)

4.1 Introduction

This chapter addressed research question #2, 'Is additional metadata required for specific field spectroscopy applications and to support interoperability with other metadata standards?' The key metadata is presented for three applications: tree crown, soil, and underwater coral reflectance. The performance of existing metadata standards in supporting the proposed core field spectroscopy metadataset is assessed, and a hybrid standard that serves as a 'best of breed' incorporating useful modules and parameters within the standards is proposed.

In Chapter 2, metadata in the context of field spectroscopy was defined as those data elements that explicitly document the primary spectroscopy dataset and field protocols that capture sampling strategies, instrument properties and environmental and logistical variables, all of which are integral to the assessment of fitness-for-purpose of the spectral measurements**.** This definition is aligned with the purpose and scope of metadata standards such as the FGDC Content Standard for Digital Geospatial Metadata: Shoreline Metadata Profile, used "to capture critical processes and conditions that revolve around creating and collecting shoreline data, and to help define and qualify shoreline data for use" (FGDC, 2001, p.1); Ecological Metadata Language 2.1.1, used to describe ecological datasets in fine detail as well as the method of data collection, including field and sampling methods (KNB, 2013); and the ANDS definition of scientific metadata as "all the information that is very specific to the study, and is needed to use and interpret the data collected" (ANDS, 2011). Please refer to Chapter 2 for a more comprehensive discussion on metadata.

The analysis of the results of the metadata survey presented in Chapter 3 demonstrate that there is a core metadataset that is critical to all campaigns, and scientists with expertise relating to specific applications are best informed about what belongs in a metadata standard relating to those applications of interest. As illustrated previously in consensus and variance analysis across metadata categories, a marine scientist, for example, has the requisite knowledge, expertise, and experience to provide a credible opinion on a metadata standard for substratum targets such as seagrass and coral. Feedback from the survey participants included arguments for further refining the metadata presented because in certain cases, the recording of metadata is dependent on the purposes for which the data is being collected. This information can be used as a basis for adapting and expanding the metadataset originating in the survey results to make it useful for specific user communities. As stakeholders of the data, field spectroscopy scientists have a vested interest in adopting a standard most suitable to their needs as both metadata data creators and users of these data.

There are several core principles that must be adhered to when designing a 'good' metadata standard. These include:

- identification of the needs of users who will access and use the data
- identification of an application profile
- direct involvement of interested stakeholders
- extension or refinement of existing standards that may not entirely meet the requirements of users
- enabling modularity for logical and consistent organization of the data

- facilitation of data discovery, retrieval, and re-use
- elimination of redundancy in data documentation so that data is collected only once

(Duval *et al.*, 2002; ANZLIC, 2007; ISO/TC 2008; INSPIRE, 2009 ; ANDS 2011; ISO 2011).

Therefore, the best approach for building a user-centric metadata standard is to begin by identifying the needs of the scientists who are being asked, potentially, to implement and use it. This has been accomplished for the core metadataset presented in Chapter 3, and what follows in this section concerns identifying specialist needs for field spectroscopy metadata.

4.2 Identifying key metadata required for soil, tree crown, and underwater coral reflectance

Defining the key metadata for specific applications requires firstly identifying the user community, and secondly, consulting them directly on what they judge to be critical metadata within the applications they would use this metadata for.

4.2.1 Identifying a user community

In 2012, an expert panel of field spectroscopy data stakeholders from the Australian and international community was convened at the TERN ACEAS 'Bio-optical data: Best practice and legacy datasets' workshop in Brisbane, Australia held on June 18-

22 in 2012 (please see Appendix B.1 for a list of attendees). The purpose of the workshop was to "drive best practice in field measurement and to lay the foundations of an international standard for the exchange of spectral datasets" (Malthus, 2012, p. 1). The workshop participants included scientists with expertise in vegetation, marine, estuarine, mineralogical, and soil. Based on the collective expertise in the group, panel discussions were structured to identify key metadata for soil, tree crown, and underwater coral applications.

4.2.2 Method

Three teams were formed to provide their input for a field spectroscopy metadata standard for three application domains: 7 vegetation scientists (tree crown), 2 marine scientists (substratum coral) and 3 soil scientists (soil). Individually assigned to the vegetation and soil teams were two IT consultants (including a data governance expert) who were stakeholders in field spectroscopy data management.

Each team was presented with a baseline metadataset derived from the survey results from the previous chapter, field data collection protocols unique to each application, and proposed metadata obtain through personal interviews with field spectroscopy scientists prior to the workshop. The objective of the activity was to derive the elements of a standard for each application that would incorporate the core metadataset, application-specific metadata, and optional metadata as proposed by each team for enhancing exchange and usability (Figure 4.1).

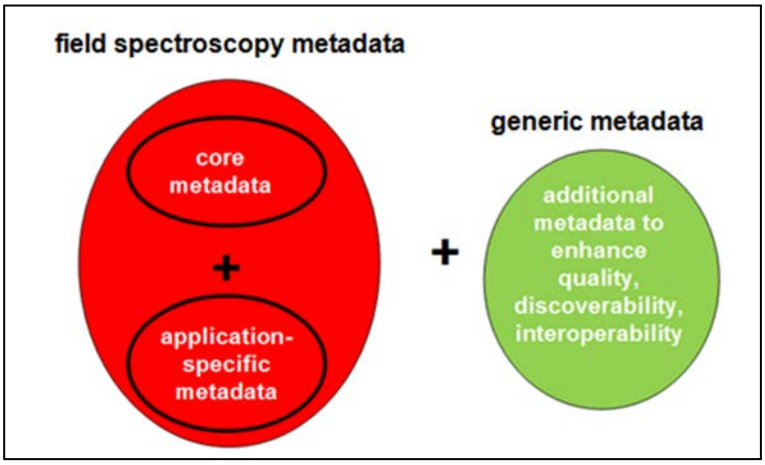

field spectroscopy metadata
core metadata
+
application-specific metadata
+
generic metadata
additional metadata to enhance quality, discoverability, interoperability

Figure 4.1 Profile of an application-specific field spectroscopy metadataset

Once presented with the baseline metadataset, the participants were asked two questions: 1) ‘If you were to create the highest quality metadataset possible, for use in either calibration or validation activities, which fields would be critical, and which would be optional? 2)’Do you recommend any new fields?’ For the first question, ‘highest quality’ was defined to be a dataset that was:

1) *comprehensive*: accurately documents the protocol executed to obtain the data;

2) *complete*: inclusive of all metadata critical to that metadataset;

3)*interoperable (digitally and semantically)*: comprises metadata elements expressed in a manner conforming to commonly accepted terminologies and ontologies to accommodate fusion with other datasets and exchange across data platforms;

4) *explicit*: captures the requisite metadata to a granularity that minimizes potential for recording ambiguous metadata (granularity in this context is the smallest unit of metadata defined for capturing a given unit of information)

Prior to panel discussions on application-specific metadata, the above parameters had been defined and discussed with the participants during a presentation given on methods and criteria for a ‘best fit’ metadata standard for field spectroscopy datasets. Calibration and field validation activities were used as a point of reference, as they are widely acknowledged within the field spectroscopy community to require the most stringent adherence to best practices in data collection (Schaepman-Strub *et al*., 2006; Milton *et al*., 2009). Field protocol, or the sampling and methodology used to generate the field spectroscopy datasets, was selected for inclusion in the metadatasets because it is an integral component in the collection of *in situ* spectroscopy data. Section 2.4 discusses the rationale and importance of including field protocol in field spectroscopy metadatasets.

For each metadata field presented within the baseline set, the scientists were asked to provide a reason for inclusion or comments, categorize the fields as critical or optional, provide an example, and to specify the data type for each field (Boolean/text/numeric/other). Providing an example and specifying a data type allowed the scientists to customize the metadataset in accordance with the taxonomies and vocabularies of their discipline.

The human perspective on metadata was a central consideration in the design of the panel discussions. Rather than being an exercise simply in documenting information related to a given application, it was important that the scientists provide direct input into the semantic structure of the metadata. Best practice for creating an application profile requires identifying specific requirements of the community that is going to use the application profile. Using scenarios and case studies, and defining the obligation of data elements, with the emphasis on human-generated metadata developed by skilled classifiers ensuring more precise and high-quality metadata (Syn and Spring, 2007; Malta and Baptista, 2012), is the advised approach. It has been previously demonstrated that in the interpretation and application of a metadata standard, people can easily confuse a concept with the designation used to represent it (Davies *et al*., 2008). For this reason, the metadata elements were expressed at a single level of atomization, with no subclasses or formally defined ontological interdependencies among metadata elements. The structure of the discussions served to minimize any potential confusion about what information is being documented and to enable the scientists to define the metadata elements in a way that is least ambiguous and most meaningful to them. Additionally, each team was also invited to volunteer any new fields that may be suitable.

4.2.3 Results: Key metadata for soil, tree crown, and underwater coral applications presented

The application-specific metadatasets, as amended and expanded by each team, are presented in Sections 4.2.3.1-3. These metadata elements present the critical elements (for brevity) as designated by the scientists. A more comprehensive presentation of the metadatasets, with both critical and optional elements, is provided in Appendix B.

4.2.3.1 Underwater coral reflectance

The underwater coral reflectance metadata list is the most comprehensive of the three applications, because it includes metadata elements relating to location and environment conditions in addition to application-specific parameters. This more voluminous metadataset is a result of additional metadata elements recommended by 1) the marine and estuarine scientists in the survey presented in the previous chapter and 2) scientists participating in the workshop. The critical elements for underwater coral reflectance are presented in Table 4.1

There are parameters in the location category directly relating to the unique and complex conditions under which marine spectroscopy operates and the environmental factors influencing the spectral measurements that are absent from terrestrial campaigns (these include tide conditions, above- and sub-surface conditions, and water column profile data). There is an almost even distribution of critical and optional designations, and two parameters (wind speed and direction) have been ranked as critical in the special case of severe conditions.

There are fifteen fields relating to coral properties, nearly half of which have been designated as critical. Two fields refer to a photo for additional data 'Homogeneity/heterogeneity' and 'Presence of epiphytes' (presented in the comprehensive list in Appendix B). This is illustrative, in part, of the difficulty of recording metadata *in situ* for marine campaigns and the use of alternate methods (such as analysis of a photo taken onsite) to add metadata retrospectively.

METADATA FIELD	REASON FOR INCLUSION / COMMENTS	EXAMPLE	DATA TYPE
GPS coordinates	Permits referencing to aerial/satellite/other campaigns; Difficult to do in situ; done on the dive site; Coordinates, datum + projection can be determined from Google Earth	x,y,z	numeric
Location description (in situ/on boat/in lab)	Critical to quantifying environmental factors to spectral measurement	Lab/boat/in situ	text
Reference to photo of local relevant environment + target	Provides additional visual data where recording additional metadata of target and environment is not possible or feasible	photo # or filename	text
Depth	From lowest astronomical tide	18 m	numeric
Tide conditions H or L	Input for determining true depth relative to datum and wave lensing effects	6:36 PM	time
Wave height and period (for reflectance measures)	Input for determining true depth relative to datum and wave lensing effects	0.25 m	numeric
Wind speed	Critical in severe conditions	5 kn	numeric
Wind direction	Critical in severe conditions	Ssw	text
Distance from bottom/substrate	Critical if 3D structure present (seagrass, branching coral)	20 m	numeric
Substratum height	Input parameter for determining upwelling radiance/ background reflectance affecting spectral measurements	4 m	numeric
Height of sensor from surface	Critical for water column profiles	1.75 m	numeric
Depth of sensor from surface	Critical for water column profiles	7 m	numeric
Distance of operator from sensor	Only applies if there is presence of shading from operator's body	0.25 m	numeric
CDOM spectral slope	Coloured dissolved organic matter; critical for water column profiles	-S value	numeric
CDOM concentration	Coloured dissolved organic matter; critical for water column profiles	A 440 nm	numeric
Detritus concentration	Critical for water column profiles	1200 μg C•l -1	numeric
Phytoplankton species/classes	Critical for water column profiles	Gymnodinium spp.	text
Target ID	Code identifier/tag for sample	Name code	text
Type	Qualitative descriptor of target type	Coral algae etc.	text
Species or name	Coral species	Diploria strigosa	text
Density of growth	Quantitative measure of density of target	2.94 g cm^{-3}	text

Table 4.1 Underwater coral reflectance metadataset (critical metadata elements)

The illumination category, a component of the core metadataset presented in the previous chapter, has also been expanded for the underwater coral reflectance; these include non-critical metadata fields presented in Appendix B including ‘natural canopy shading’ and ‘artificial canopy effect’. There are four parameters relating to viewing geometry that are normally not required for terrestrial campaigns – ‘distance from bottom/substrate’, ‘distance of operator from sensor’, ‘height of sensor from surface’, and ‘depth of sensor from surface’. The latter three are critical only in cases of shading by the operator's body or where data is required for profiling the water column.

4.2.3.2 Tree crown reflectance

This metadataset was originally presented to the seven scientists on the vegetation team as a 'tree crown reflectance' standard but was changed by the respondents to 'vegetation reflectance'. This group spent the most amount of time (approximately 2 hours) debating the inclusion of the proposed metadata elements. The critical fields are presented in Table 4.2. There are five metadata fields that have been designated as critical and these include 'Collected within 1 week of campaign', 'Position in canopy', 'Illuminated leaves', 'Target or scale' that denote sampling protocol steps that must be completed in accordance with good practice, or a recognized protocol.

METADATA FIELD	REASON FOR INCLUSION / COMMENTS	EXAMPLE	DATA TYPE
Collected within 1 week of aerial campaign	Minimizes any detectable changes in leaf phenology (this can be referenced via a protocol citation)	Yes; ABCD Organization Tree Crown Reflectance Protocol, 2012	text
Position in canopy	Corresponds to visible canopy in an aerial hyperspectral campaign (this can be referenced via a protocol citation)	Emergent leaves on top third of canopy; ABCD Organization Tree Crown Reflectance Protocol, 2012	text
Illuminated leaves	(this can be referenced via a protocol citation)	Yes; ABCD Organization Tree Crown Reflectance Protocol, 2012	text
Target or scale (single leaf, branches, mature leaves, etc.)	Ensures consistent phenological state for all samples and sufficient leaf size for integrating sphere measurement (this can be referenced via a protocol citation)	Yes; ABCD Organization Tree Crown Reflectance Protocol, 2012	boolean
Tree species		Eucalyptus aquatica	text

Table 4.2 Tree crown reflectance metadataset (critical metadata elements)

Of the three teams, the vegetation team had the least consensus among members, with the lowest proportion of metadata elements agreed upon for inclusion in the standard. In a group discussion following the team activity, the vegetation scientists stated that it was difficult to arrive at a conclusive optionality designation without knowing “what the purpose of the campaign is”. Consequently, no exemplar metadataset for tree crown reflectance could being developed within the timeframe of the workshop. As a solution, one of the team members with expertise in tree crown reflectance was consulted after the workshop to specify optionality for the metadata elements. For this reason the metadataset retained its original designation, ‘tree crown reflectance’ for utility as an exemplar metadataset.

4.2.3.3 Soil reflectance

The soil team designated the largest proportion of metadata fields as critical (over 75%) (Table 4.3). All the fields in the soil reflectance metadataset refer almost exclusively to the properties of the soil and not environmental conditions in the field or sampling protocol. Unlike the metadata elements in the previous applications, most of the parameters in the soil metadataset can only be obtained retrospectively (chemical constituents, alkalinity, etc.) but were designated by the team to be critical for creating a ‘highest quality’ metadataset. Comments or reasons for inclusion in the metadataset were not provided for every field.

4.2.4 Discussion

Overall, the results re-assert the findings from the previous chapter that defining a dataset for specific applications requires the input of those who know best about the

METADATA FIELD	REASON FOR INCLUSION / COMMENTS	EXAMPLE	DATA TYPE
Description		ferri-soil	text
Sample #		1	text
Name	can be extracted from a taxonomic list / soil series name	calcic orthid	text
Weight	can be used to describe wet or dry weight	dry weight to moisture	numeric
Volume	derived from soil cans	134.5 cm^3	numeric
Mineral bulk density	also can be designated 'soil bulk density'	msd/vd	numeric
Particle density		265g/cm^3	numeric
Order		Aridisol	text
Type		loam	text
Horizon		A'	text
Grain size		3 parts	numeric
Texture	sand/silt/clay	sieving	text
Surface roughness	necessary for BRDF/erosion calculations	0.025	numeric
Colour	MUNSELL units/ colour chips can be used	10 YR 6/4	alphanumeric
Level surface/rough/inclined	aspect should be included	10˚ or 10`0	numeric
Moisture content	gravimetric or volumetric	57%	numeric
Humus content		3.40%	numeric
Nitrogen content		20 ppm	numeric
Clay content		20%	numeric
Sand content		5%	numeric
Silt content		5%	numeric
pH in H20		7.0pH	numeric
Water retention (field capacity)			numeric
Wilting point		0.44 cm^3/cm^3	numeric
Total alkalinity		10 mg L^{-1}	numeric
Conductivity		8 dS/m	numeric
Porosity		0.45	numeric
Contamination (none/mining/agriculture/etc)		mining	text

Table 4.3 Soil reflectance metadataset (critical metadata elements)

field practices and properties of the feature being sampled. The metadatasets also demonstrate that the proportion of metadata that are obtained retrospectively varies from a minimum, as demonstrated by the tree crown reflectance metadataset, to a maximum, as demonstrated by the soil metadataset.

The relative difficulty in creating a definitive standard for tree crown reflectance suggests that 1) consensus can be difficult to achieve and may not always be a prerequisite in building a good standard for a given application or objective and 2) there is a threshold at which a prescriptive standard becomes restrictive.

Despite the low consensus among scientists as to which fields should be designated as optional or critical, the tree crown reflectance metadataset remains a valid baseline of comprehensive metadata relating to both quantitative and qualitative field data that are commonly recorded in field spectroscopy campaigns in accordance with good practice. For the purpose of meaningful analysis in this study, the tree crown reflectance metadataset retained its original name, despite it being changed by the scientists to ‘vegetation campaign’. The reference to vegetation is too broad a term (as it can include applications such as those in agriculture or estuarine environments) and it was necessary to restrict the metadataset to a single and specific application, especially for use in further examination in the next section.

The degree of prescriptiveness of a standard is worth considering as a measure of its value to a given community. Prescriptiveness has the potential to guide good practice for metadata documentation in the field. However, it is possible that requiring a user to record protocol steps, or target properties in multiple metadata fields at too fine a granularity may in fact be prohibitive and result in an inflexible and onerous standard. This can arise first from draining resources of time in the field by forcing the user to comply with the proposed standard. Secondly, aligning the metadataset to a field data collection protocol that prevents an expert user from making their own informed choices about what is good practice is counter to principles of sound and innovative research.

4.3 Do current geospatial metadata standards accommodate the needs of specific applications (soil/tree crown/underwater coral reflectance)?

4.3.1 Geospatial metadata standards as a test case

Assessing the usefulness of existing standards to accommodate the needs of specific user communities requires a comparison of the core metadataset and the three field spectroscopy metadatasets (underwater coral, tree crown, soil) presented in the previous section with standards commonly implemented within geospatial science.

Choosing a standard for analysis was based on an informal survey of those endorsed by agencies involved in research in geospatial science or geospatial data standards and include FGDC (Federal Geographic Data Committee), NASA's Earth Science Division, CSIRO (Commonwealth Scientific and Industrial Research Organisation), and INSPIRE (Infrastructure for Spatial Information in the European Community), among others. Many of these standards are implemented in popular commercial GIS packages, signifying the implication of an inadequately performing standard when applied to widely distributed geospatial datasets Table 4.4 provides an overview of the seven standards chosen for analysis.

Standard	Date created (initial version)	Creator(s)	Purpose	External standards incorporated	# of elements
Dublin Core 1.1	1995	Dublin Core Metadata Initiative	for use in resource description for a wide range of resources (DCMI, 2013)		15
Access to Biological Collections Data Schema 2.06	2006	ABCD Task Group	support the exchange and integration of detailed primary collection and observation data (ABCD Task Group, 2013)	• Dublin Core • FGDC Content standards for digital spatial metadata • FGDC-STD-005 Vegetation Classification and Information Standards • FGDC-STD-001-1998 • FGDC Content Standard for Digital Geospatial Metadata: Biological Data Profile • SPECTRUM • Abstract Syntax Notation One	1004
Ecological Metadata Language 2.1.1	2000	National Center for Ecological Analysis and Synthesis	provide the ecological community with an extensible, flexible, metadata standard for use in data analysis and archiving that will allow automated machine processing, searching and retrieval (KNB, 2013)	• Dublin Core • CSDGM • CSDGM Biological Profile • ISO 19115 • ISO 8601 Date and Time Standard • GML • STMML • XSIL	562
Darwin Core	1998	Darwin Core Task Group	• provide a stable standard reference for sharing information on biological diversity • provide stable semantic definitions with the goal of being maximally reusable in a variety of contexts (Darwin Core Task Group, 2013)	Dublin Core	45
Content Standard for Digital GeoSpatial Metadata: Remote Sensing Extension	1998	Federal Geographic Data Committee	provide a common terminology and set of definitions for documenting geospatial data obtained by remote sensing (FGDC, 2002)	ISO 19115	360
Content Standard for Digital GeoSpatial Metadata: Shoreline Metadata Profile	2001	• Federal Geographic Data Committee • Marine and Coastal Spatial Data Subcommittee	capture the critical processes and conditions the revolve around creating and collecting shoreline data, and to help define and qualify shoreline data for use (FGDC, 2001)	FGDC-STD-001-1998	33
ANZLIC Metadata Profile 1.1 (Geographic dataset core)	2007	ANZLIC	create metadata records that provide information about the identification, spatial and temporal extent, quality application schema, spatial reference system, and distribution of digital geographic data (ANZLIC, 2007)	• AS/NZS ISO 19115:2005 • ISO 19115:2003/Cor.1:2006 • ISO/TS 19139:2007	45

Table 4.4 Geospatial standards selected for analysis

Section 2.3 presents a discussion on metadata standards in general. Generic geographic metadata standards such as ISO 19115 were already incorporated in part or in whole in several of the standards selected (EML 2.1.1, CDGSM, ANZLIC Metadata Profile) and therefore were not directly examined to avoid redundant analysis.

4.3.2 Method

Measuring the capacity of an existing geospatial metadata standard to document the requisite metadata for a given campaign type (tree crown/underwater coral target/soil) was done by answering a single question: for campaign-level data, how many metadata fields (metadata elements) in each existing standard could be used to capture the information specified in the metadatasets presented in Section 4.2?

The purpose of the analysis was to accomplish more than simply examining whether the field spectroscopy metadata elements could be operationalized as a metadataset conforming to an existing standard. Operationalizing a field spectroscopy metadataset as an existing standard could entail storing the metadata elements wherever available within the standard, including generic free-text parameters (such the value-eml-text field in EML 2.1.1 standard). This could be possible in cases where more explicit metadata elements relating specifically to field spectroscopy were unavailable. However, this manner of over-simplistic analysis would fail to yield any meaningful results as it would not be an accurate measure of instances where an existing standard succeeds or fails to correspond to the field spectroscopy dataset.

Rather, the analysis was used to determine how well an existing standard can be mapped to, unidirectionally, on a metadata element-by-metadata element basis, to the field spectroscopy metadatasets. Figure 4.2 shows a successful mapping for metadata elements in two existing standards to metadata elements in the proposed field spectroscopy metadataset.

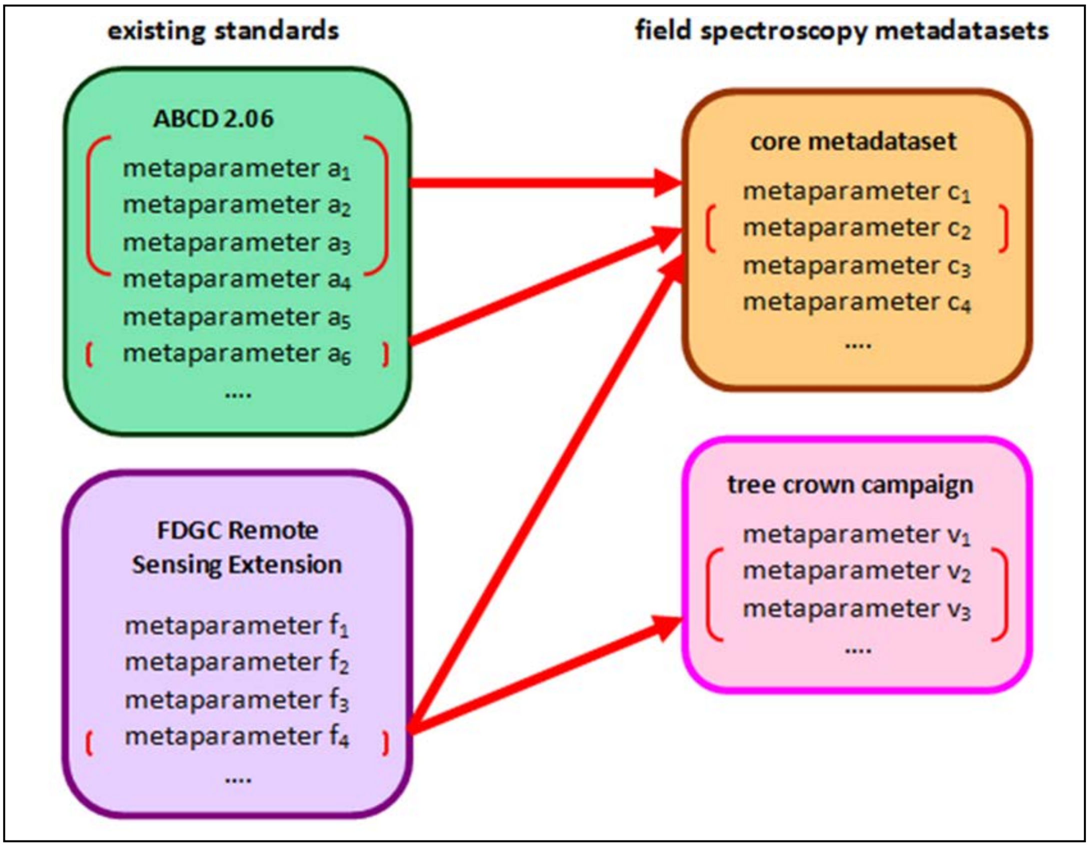


Figure 4.2 A conceptual example of a successful mapping from two existing geospatial metadata standards to the proposed field spectroscopy metadata standard

Criteria were applied to define successful mapping. These are explained in detail in Table 4.5. Metadata elements specified at the smallest level of granularity or atomization in the standard were chosen. This was done to allow a uniform comparison among the proposed and existing standards. For example, the 'Date' field in the field spectroscopy metadataset is expressed as a single unit of metadata,

ACCEPT	REJECT
Explicit reference *Example: The 'Wind speed' metadata element in the FDGC Marine Extension standard was successfully mapped to 'Wind Speed' in the coral target metadataset.*	**Undefined or ambiguous metadata element** *Example: Where the parameter description was absent or too vague to determine its purpose, it was not counted as a suitable metadata element. For example, in ABCD standard user guidelines, the 'Method' field within the '/DataSets/DataSet/Units/Unit/Sequences/Sequence/' class has no definition.*
Implicit reference *Example: Instrument category metadata elements ('Make', 'Model', 'Serial Number') could be recorded in the EML 2.1.1 'Instrumentation' metadata field in both the 'Protocol' and 'Methods' module.*	**Incorrect parent or container class** *Example: The 'Viewing Geometry' category in the proposed core metadataset is comprised of critical elements relating to sensor viewing angles. A mapping was not successful if counterparts in an existing standard were in the wrong parent or container classes. Sensor azimuth and zenith angle parameters exist within the FGDC Remote Sensing Extension but are defined within the 'Satellite' container class and therefore could not be mapped to sensor geometry metadata in the core dataset.*
	Manually-defined classes or fields *Example: Instances of the EML 2.1.1 'attribute' parameter that could defined by the user to record any campaign metadata.*
	Generic metadata element *Example: Any metadata elements within an existing standard that referred to data that could be extracted from a generic data table, such as those referenced by the EML 2.1.1 'dataset' module; the 'measurementValue', 'Attribute', 'dynamicProperties' metadata fields in Darwin Core 1.1 that could be applied to any numeric or text metadata parameter.*

Table 4.5 Criteria for accepting or rejecting a metadata element in an existing standard for mapping

whereas the ABCD standard for Date data (in the /DataSets/DataSet/Units/Unit/Identifications/Identification/Date container class) has nine subfields (DateText, TimeZone, ISODateTimeBegin, DayNumberBegin, TimeOfDayBegin, ISODateTimeEnd, DayNumberEnd, TimeOfDayEnd, PeriodExplicit) used to capture this information.

Using the finest granularity was true for all cases where the documentation for the standard defined parameters to this level of granularity. This was the baseline against which all standards were measured. All other standards needed to be reduced to the same level of granularity for analysis, taking into account both explicit and implicit references to a given metadata element. The definition of each element was used as the determining factor for mapping. For example, EML 2.1.1 specifies that the ‘instrumentation’ metadata element in the ‘Methods’ module can include information about the quality control and quality assurance for the instrument, therefore it could be mapped to the instrument calibration metadata category in the proposed core metadataset.

Unique and non-unique mappings were counted. A unique mapping occurs when a metadata element (e_1) in an existing standard has been mapped to one and only one metadata element (p_1) in the field spectroscopy dataset (core/ tree crown/ underwater coral/ soil). An example of this is the ‘Wind direction’ field for above-surface marine conditions in the FGDC Marine Shoreline Data Extension that was mapped to the ‘Wind direction’ field in the underwater coral metadataset, with no other mappings to other fields in the underwater coral metadataset. If metadata element e_1 can be mapped to multiple metadata elements in a proposed dataset, this is considered a non-unique mapping. An example of this is two metadata fields in ABCD 2.06 that can be mapped to both [target]‘Species or name’ and ‘Phytoplankton species/classes’ in the underwater coral set. Counting unique and non-unique mappings is useful for determining the requisite explicitness for an existing standard to successfully capture information in a field spectroscopy metadataset.

4.3.3 Results

The results of the mappings are summarized in Figure 4.3.

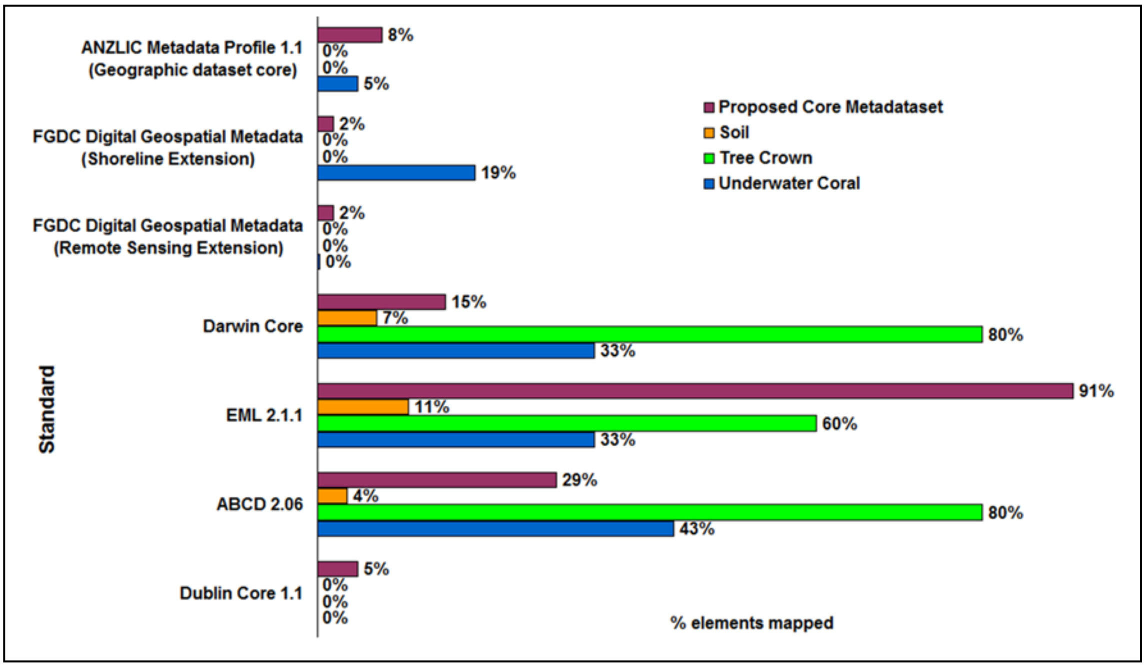


Figure 4.3 Successful mappings from existing standards to the field spectroscopy metadatasets as a percentage of the total number of elements mapped in the proposed core and application-specific metadatasets

4.3.3.1 Dublin Core 1.1

Fifteen metadata fields within Dublin Core were examined. The number of successful mappings ranged from 0-5 % of the target metadatasets. The consistency in high failure rates across the four field spectroscopy metadatasets could be accounted for by Dublin Core's primary purpose to identify a dataset at the collection-level with parameters whose scope are limited to content (i.e. subject, description), intellectual property (i.e. publisher, rights), and instantiation (i.e. format, identifier). The mapping had some success (5%) with the core metadataset, specifically within a subset of the core metadataset relating to project information, of which four metadata elements could be mapped to (given that the owner of the dataset would choose to use the project/experiment details as identifiers for the dataset as well).

4.3.3.2 Access to Biological Collections Data Schema 2.06

One thousand and four metadata elements were examined in ABCD 2.06. Success ranged from a minimum of 4% of critical elements with the soil metadataset to 80% for tree crown with the mean value of elements mapped being 39% with σ=32%. It mapped to 29% of the core metadataset and 43% of the critical elements in the coral metadataset. Dublin Core has been wholly incorporated into ABCD 2.06 so a minimum of successful mappings to the core metadataset is guaranteed. The mandate for ABCD is to facilitate “access and exchange” of “primary biodiversity data” (ABCD Task Group, 2013), of which the underwater coral reflectance metadataset has the highest proportion in terms of biological sample parameters (including species, specimen id) compared to the core, soil, and tree crown sets. The tree crown has a higher proportion of sampling protocol parameters which can be captured in several of the /DataSets/DataSet/Units/Unit/Gathering/ modules.

4.3.3.3 Ecological Metadata Language 2.1.1

Four hundred eighty four elements in EML 2.1.1 were examined. It had the highest overall success with all four metadatasets: 91% for core, 60% for critical elements in the tree crown metadataset, 11% soil, and 33% underwater coral the mean value of elements mapped being 49% with σ=35%. As with ABCD 2.06, it is biased towards biological data collection. Mappings to soil and underwater coral can increase (up to 100% for soil) if the 'table dataset value' element, referring to an associated table with target characteristics, is selected to store parameters such as clay content (soil) and chlorophyll concentration (coral). This element was ignored for successful mappings as it was classed as too generic, according to the criteria in Table 4.5.

Its success with the core metadataset can be accounted for in part the by the fact that it has a larger amount of dataset-level metadata elements that can be mapped to the 'project information' subset, and instrumentation metadata that can be populated in the 'methods' module 'instrumentation' metadata element, which accommodates description of any instruments used in the data collection. The sampling protocol metadata elements in the tree crown metadataset ('illuminated leaves', 'position in canopy') can also be captured either in the 'methods' or 'protocols' modules. According to the EML documentation, either parameter is suitable, based on how the protocols are described: "'methods' is descriptive (often written in the declarative style: "I took five subsamples...") whereas 'protocol' is prescriptive (often written in the imperative mood: "Take five subsamples...")" (KNB, 2013).

4.3.3.4 Darwin Core

Forty five elements in Darwin Core were examined. It had the highest success with tree crown (80% of the critical elements), 33% for coral, 15% for core and 7% for soil with the mean value of elements mapped being 34% with σ=33%. Those parameters referring to sample properties have been semantically structured for biodiversity data, hence its relative success with coral data. There were no explicit or implicit references to instrument properties (within the core metadataset), and the 'method' parameter was considered insufficient in scope by the author to be suitable for sampling protocol or viewing geometry.

4.3.3.5 FGDC Content Standard for Digital Geospatial Metadata (Remote Sensing Extension)

Three hundred and sixty elements in FGDC Content Standard for Digital Geospatial Metadata (Remote Sensing Extension) were examined. The Remote Sensing Extension could be mapped only to 2% of the core metadataset with no mappings to the three specific applications. Mappings to the core were for dataset-level metadata, given that the experiment information (name, date) could be used to identify the metadataset at this level. However, this hypothetical dataset would be empty as no target properties could be documented within the standard. The Remote Sensing Extension is designed for digital geospatial data (obtained from satellite and airborne sensors primarily), and has no suitable parameters to capture sampling techniques, viewing geometry, or instrument information for *in situ* sensors.

4.3.3.6 FGDC Content Standard for Digital Geospatial Metadata: Shoreline Metadata Profile

Thirty three elements in the Shoreline Metadata Profile were examined. It had the highest success with critical elements in the underwater coral reflectance (19%), and core (2%), but no elements were mapped to either the tree crown or soil metadatasets. Even though this standard applies to digital geospatial metadata, when examined on its own, it is useful for recording location and environment parameters (wind speed, tide, above surface conditions) for the underwater coral campaign. It is noteworthy that this standard has no 'depth' parameter. The metadata elements mapped to the core metadataset related to a subset of location and environment parameters.

4.3.3.7 ANZLIC Metadata Profile 1.1 (Geographic dataset core)

Forty five elements in the ANZLIC Metadata Profile 1.1 (Geographic dataset core) were examined. Successful mappings were restricted to the core dataset (8%) and 5% of the critical elements in the coral reflectance metadataset. This is due to the fact that ANZLIC standards are primarily for cataloguing services, and in the context of the geographic dataset core standard, document information about the "identification, spatial and temporal extent, quality, application schema, spatial reference system, and distribution of digital geographic data" (ANZLIC, 2007). The few core metadataset parameters that were mapped to relate to project and experiment profile information, or the special case of GPS coordinates categorized as spatial reference information for underwater coral reflectance.

4.3.4 Measuring flexibility

An additional measure was included in the analysis to determine whether an existing standard's flexibility had an effect on how much information it could capture in the field spectroscopy metadatasets (core/tree crown/soil/underwater coral). In this context, flexibility is defined as the potential for a metadata element in an existing standard to be re-used (or re-mapped) to multiple metadata fields in a field spectroscopy metadataset. For example, according to the user guidelines for EML 2.06 (KNB, 2013), in the 'Sampling' module, the metadata element 'instrumentation' can be mapped to all parameters for instrument metadata defined in the core metadataset. This is considered a non-unique mapping. On the other hand, the 'Wind speed' metadata element in the FGDC Shoreline Metadata Profile standard can be successfully mapped to one and only one metadata element ('Wind Speed') in the coral reflectance metadataset. This is considered a unique mapping. The more explicit a metadata element in the existing standard is, the greater the likelihood of a unique mapping for that field. An average of unique (UM/me) and non-unique (NUM/me) mappings per total number of mapped elements for each dataset was calculated. These averages were then correlated to a standard's success in capturing information in the field spectroscopy metadatasets (core/tree crown/soil/underwater coral) (Figure 4.4).

The datasets shows that the correlation between the amount of data captured by an existing standard (% elements mapped in the dataset) and average mappings per element is stronger for non-unique mappings (r=0.365 n=28, p=0.001) than for

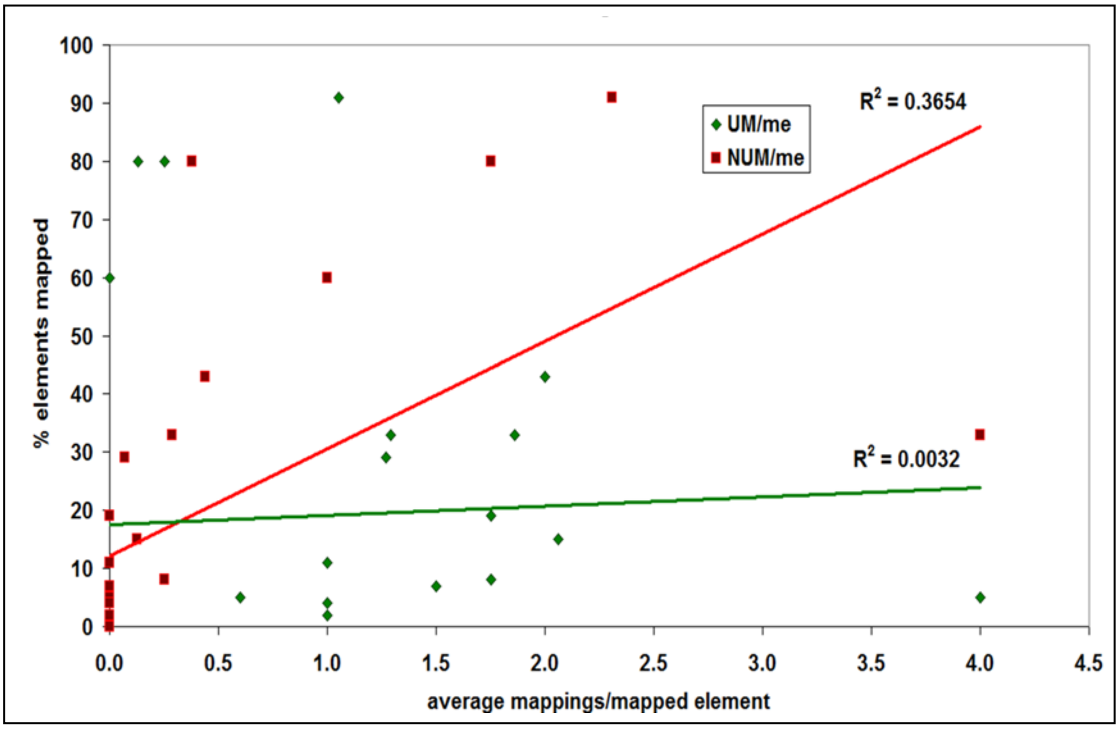


Figure 4.4 Correlation of mappings per element (both unique and non-unique) to the percentage of total elements mapped in the dataset

unique mappings (r=0.003 n=28, p=0.001). This suggests that in the context of the standards studied, generally, the less prescriptive or explicit an existing standard is, the more likely it is to capture a larger amount of information in the field spectroscopy metadataset. These results are significant to the formal adoption and implementation of a field spectroscopy metadata standard. First of all, a balance must exist between the generality of metadata parameters (for capturing the maximum amount of information necessary for a dataset) and the granularity of metadata parameters (so that datasets can be described in sufficient detail). Secondly, the interoperability between a field spectroscopy metadata standard and other metadata standards is dependent in part on the prescriptiveness of the field spectroscopy metadata standard. These two considerations must be addressed to enable data users to share and intercompare datasets.

4.3.5 Discussion

The mapping results, from the seven existing standards, demonstrate that they are almost uniformly lacking in meeting the needs of field spectroscopy scientists in the context of the four field spectroscopy metadatasets. The overall compliance levels, in decreasing order, are tree crown (μ=31%, σ=40%), proposed core metadataset (μ=22%, σ=32%), coral (μ=19%, σ=18%) and soil (μ=3%, σ=4%) applications. In no instances were the critical metadata elements for any of the datasets captured in their entirety. Field spectroscopy metadata has a large proportion of protocol and sampling information that is commonly documented in biological data metadata standards (hence the relative success with EML 2.1.1) but these are absent from dataset-level specific standards such as Dublin Core 1.1 and the ANZLIC Metadata Profile 1.1 (Geographic dataset core). There was a consistent lack of explicit references to critical field metadata such as instrument properties, viewing geometry, and reference standards. The metadata model in the FGDC Content Standard for Digital Geospatial Data (Remote Sensing Extension) for satellite and airborne sensors was the most closely aligned with requirements for field spectroradiometers. Despite the deficiencies in the existing standards, many of them have dataset-level modules and parameters (literature citations, quality assessment reports) that may be useful in enhancing a field spectroscopy metadataset's potential for discoverability and re-use.

The correlation tests for unique and non-unique mappings show that flexibility has a positive effect on a standard's success in capturing more information. These results have the greatest implication for metadata that documents field or sampling protocol, as these are most likely to be non-standard and dependent upon the purpose for which the data being is being collected. This was exemplified in part in Section 4.2 by the low consensus among vegetation scientists as to which protocol steps and other metadata to include in the tree crown dataset. Therefore an instrument operator or other campaign participant recording the metadata must be able to document field protocol unambiguously (with sufficient explicitness) without the restrictions imposed by a metadata standard that cannot be adhered to because it is not aligned with the operator's field methods.

4.4 A hybrid standard

The work presented in this chapter so far has met four of the criteria for building a good standard presented in the introduction. The application profiles (tree crown, soil, underwater coral application) and needs of the data users (metadatasets presented in Section 4.2 and the core metadataset in the previous chapter) have been identified; they were asked for assistance in building the standard, and as stakeholders, they were directly involved. The remaining criteria -- extend or refine existing standards that may not entirely meet the requirements of users; enable modularity for logical and consistent organization of the data; facilitate data discovery, retrieval, and re-use; eliminate redundancy in data documentation so that data is collected only once – are satisfied in this section through the creation of a hybrid field spectroscopy standard that integrates the metadatasets presented in Section 4.2 and metadata elements from the standards scrutinized in Section 4.3.

Examination of existing geospatial metadata standards demonstrates that although they are deficient in meeting the needs of field spectroscopy scientists, they are comprised of modules and parameters that are useful for enabling and enhancing the robustness, discoverability, quality assurance, and interoperability of the field spectroscopy datasets. These include metadata relating to dataset-level information (title, abstract, keywords, contacts, maintenance history, purpose), data quality (logical consistency, completeness, lineage), access rights (copyrights, levels of access for user groups), revision history, literature citations, and physical format data, among others.

Digital provenance information is especially significant for long-term preservation of datasets, and research scientists have demonstrated a preference for long-term storage capabilities (i.e. over five years) over short-term storage (i.e. less than twelve months) and commonly share datasets from 1-3 months to 2-5 years after findings have been published (Guenther, 2010; Chao, 2012). Documenting this metadata has benefit within and outside the field spectroscopy community. It enables logging of the use of the dataset, promotes greater understanding of research inquiries, provides those responsible for its governance with information for forecasting the use of the dataset, who in turn endorse services to support data access (Chao, 2012).

Figure 4.5 shows a proposed hybrid metadata standard that fuses the metadatasets identified as requisite by the vegetation, soil, and marine scientists, and additional metadatasets imported from the standards examined in Section 4.3 that can serve as a ‘best of breed’ standard.

Figure 4.5 A proposed hybrid standard fusing the four field spectroscopy metadatasets (core and application-specific) with elements from the standards examined

The new modules that have been imported and customized from existing standards are:

dataset module: broad-scope information that describes the entire dataset and includes title of the dataset, metadata standard name and version, revision history, keywords, purpose, and other general descriptors, for the main purpose of cataloguing and discoverability. Imported from the ANZLIC Metadata Profile (Geographic dataset core) metadata element, ABCD 2.06 metadata module, EML 2.1.1.dataset module.

resource module: information about the creators/owners/distributors of the data, lineage information, and contact information for the data resources. Imported from ANZLIC Metadata Profile (Geographic dataset core) metadata element; ABCD 2.06 metadata module; Dublin Core 1.1 publisher metadata element, EML 2.1.1.dataset module.

access module: specifies access rights to groups or particular users. Includes information about copyrights, trademarks, licenses, sequestered/classified datasets. Imported from Dublin Core 1.1 rights metadata element, EML 2.1.1 access module.

project module: information about the research context and purpose, experiment design, funding and sponsorship. Imported from the EML 2.1.1 project module.

applications module: databases/datawarehouses/online repositories where the data can be accessed, and software recommended for viewing or analyzing the associated dataset. These can be references to EOSDIS Reverb|ECHO, Carnegie Spectranomics, TERN Data Discovery Portal, DLR Spectral Archive (for data access), ViewSpec Pro, SPECCHIO, MATLAB (for data analysis). Imported from the EML 2.1.1 software module.

data quality module: reports, indices, and assurances on the completeness, quality, and logical consistency of the data. Imported from the FGDC Content Standard for Digital Geospatial Metadata (Remote Sensing Extension).

citations module: relevant literature, publications, reports, journal articles, etc. cited in the metadataset or specifications about how the dataset itself should be cited externally. Imported from the EML 2.1.1 literature module.

protocol module: documentation of (or references to) the sampling and field protocols used in the collection of the field data, such as those for hyperspectral ground calibration, leaf sampling, underwater coral sampling. Can also include taxonomies, nomenclatures, and classification systems used in the protocol such as the AASHTO/FAO/USDA/Canadian/Australian soil classification systems for soil applications. Imported from the EML 2.1.1 literature module.

The protocol module is especially relevant to field spectroscopy. Section 4.2 demonstrated that in many cases, sampling techniques for a single target are dependent on the purposes for which the data is being collected, and Section 4.3 established the value of flexibility in a standard in capturing the requisite metadata for a given campaign. Including a protocol module in a field spectroscopy standard allows the user to choose the protocol (with associated metadata elements) they want to apply to their metadataset, and in cases where they are creating one *ad hoc*, the baseline metadataset for the application is available and can be customized accordingly to the campaign.

4.5 Conclusions

Three user communities within field spectroscopy were identified and interviewed to help design a metadataset for three applications – tree crown, soil, and underwater coral. Three metadatasets were created, with descriptions and rationale for each metadata element, optionality rankings, and preferred data formats. Consensus within the tree crown group was lowest on which metadata should be included in their metadataset, based on the argument that knowledge of what the dataset will be used for determines the metadata elements that are required. It was established that some parameters are difficult to obtain *in situ* and can only be populated retrospectively, as illustrated with the underwater coral application, which is carried out under conditions and in environments unique to marine campaigns.

Seven metadata standards, selected as being representative of standards within geospatial science and its applications were examined for their ability to support proposed field spectroscopy metadatasets. These were: Dublin Core 1.1, Access to Biological Collections Data Schema 2.06, Ecological Metadata Language 2.1.1, Darwin Core, Content Standard for Digital GeoSpatial Metadata (Remote Sensing Extension), Content Standard for Digital GeoSpatial Metadata (Shoreline Metadata Profile) and ANZLIC Metadata Profile 1.1 (Geographic dataset core). The results show they consistently fail to accommodate the needs of both field spectroscopy scientists in general as well as the three user communities (vegetation, soil, marine). Mappings from each standard to the field spectroscopy metadatasets were, on average, 22% of the proposed core metadataset, 31% tree crown, 3% soil, and 19% of the coral metadatasets. Flexibility analysis revealed that the less prescriptive or explicit an existing standard is, the more likely it is to capture a larger amount of information in the field spectroscopy metadatasets.

By building upon the knowledge of scientists in ecology, marine science, the physical sciences and data governance experts who helped to develop existing geospatial standards, a hybrid standard can be created. Elements describing and documenting the dataset, resources, access, applications, data quality, citations, and protocols can enrich a field spectroscopy standard and make it adaptable to multiple data infrastructures. This entirely new field spectroscopy metadata standard addresses the specific needs of field spectroscopy data stakeholders with sufficient robustness to facilitate documentation, quality assurance, discoverability and data exchange within large-scale data sharing platforms.

Chapter 5 Field spectroscopy metadata quality

5.1 Introduction

This chapter addresses research question #3, 'What are the criteria for measuring the quality and completeness of field spectroscopy metadata in a spectral archive?' Unique methods for measuring quality and completeness of metadata to meet the requirements of field spectroscopy datasets are presented. Two spectral libraries are examined as case studies of operationalized metadata policies, and the degree to which they are aligned with the needs of field spectroscopy scientists.

The previous chapter suggested that a hybrid model is best for a metadata standard that accommodates the uniqueness of field spectroscopy data sets and permits documentation, quality assurance, and maximizes potential for discoverability and data exchange within large-scale data sharing platforms. A hybrid model incorporates the core metadataset and application-specific metadata as defined by remote sensing scientists in surveys and interviews presented in Chapters 3 and 4. It also adopts metadata quality modules from existing geospatial metadata standards that include reports, indices, and assurances on the completeness, quality, and logical consistency of the data.

It was also established in the previous chapter that field spectroscopy metadata cannot be discretized in the same manner as defined by generic metadata standards (Dublin Core 1.1, Darwin Core) or those within geospatial science (Content Standard for Digital GeoSpatial Metadata: Remote Sensing Extension, ANZLIC Metadata Profile

1.1: Geographic dataset core). Likewise, the principle of fusing the best elements of existing standards with the requisite core metadata presented in earlier chapters extends to assessing the quality and completeness of a field spectroscopy metadataset. Therefore, metadata quality and completeness must be defined in a way of greatest utility and relevance to users of field spectroscopy datasets and encompass a set of criteria that relates to a baseline set of parameters from existing standards and those unique to field spectroscopy metadata.

5.2 A quality and completeness definition for field spectroscopy metadata

In the context of field spectroscopy stored within digital libraries and databases, metadata can be described in both its completeness and quality. Please refer to Section 2.6 for a discussion on general concepts of metadata completeness and quality. In the absence of a formal definition of quality and completeness for field spectroscopy metadata, a definition is required that is a) useful, informative, and understandable to users of this metadata, b) can quantify the success of a given metadataset or data repository in meeting users' needs, and c) provides information about the reputability of the repository or the data creators as a source of complete and high quality metadata.

Field spectroscopy metadata quality can therefore be defined 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. Ideally this set includes parameters that have been identified as most important in information science studies on metadata (Bruce and Hillman, 2004; Stvilia *et al*., 2007; Ochoa and Duval, 2009), while at the same time conforming in some respect to concepts of data quality proposed for geospatial datasets by geospatial science advisory bodies (FGDC 2002; ISO, 2002, 2003, 2011; ANZLIC, 2007; NISO, 2007; INSPIRE 2009). At the intersection between geospatial and information science metadata, there exists a set of parameters that are most commonly identified as essential: logical consistency (metadata elements are expressed using ontologies, taxonomies, data types and relationships conforming to an informed consensus rationale); lineage (the source of the metadataset, responsible parties, citations and metadata revision history); error rate (documents semantic and syntactic errors in the metadata); compliance with a metadata quality standard; quality assurance by a recognized authority; and reputational authority of the data owners/data creators. While this is not a comprehensive list of all the possible metadata quality parameters, it serves as a suitable compromise between the two disciplines, and satisfies the criteria for a field spectroscopy metadata quality definition presented earlier in this section.

Field spectroscopy metadata completeness can be defined as a two-fold measure consisting of a) conformance with the core metadataset and application-specific metadata presented in Chapters 3 and 4 and b) compliance with the standards of the data infrastructure in which they are stored. The former sets a consistent benchmark for all field spectroscopy metadatasets. The latter is a fluctuating target dependent upon the benchmarks defined by the database/data repository designers; it provides implicit reputational information about the database/data repository because it measures how well (or if it all) a data repository complies with its own completeness rules.

5.3 How do current spectral libraries perform in terms of the proposed quality and completeness measures?

Applying the proposed quality and completeness measures in Section 5.2 to existing spectral libraries gives an illustration of how well existing datasets meet the needs of the field spectroscopy community. The results of the analysis also reveal areas of potential change to metadata policies for future implementation of spectral data repositories.

5.3.1 Datasets

An investigation into publicly available field spectroscopy libraries that hold a range of spectra with associated metadata revealed that few exist that can be considered suitable for analysis. These include the ASTER Spectral Library v. 2.0, DLR Spectral Archive, USGS Spectral Library v. splib06a, and SPECCHIO v. 2.2 (a more detailed discussion on spectral libraries can found in Section 2.4.4). Of these, tests cases were chosen based on their diversity of spectra, volume of data, and availability of the metadataset for download and analysis. The two chosen were the USGS Spectral Library and the SPECCHIO database. The DLR Spectral Archive could not be analysed concurrently with SPECCHIO and the USGS Spectral Library because data could not be obtained from the DLR data center in a suitable format in time for analysis. USGS Spectral Library, as a subset of the ASTER Spectral Library, was chosen as a suitable, more appropriate proxy than the entire ASTER Spectral Library itself, given that USGS Spectral Library has a larger proportion of field spectroscopy data. Table 5.1 provides a general overview of the two selected data libraries.

The USGS Spectral Library (http://speclab.cr.usgs.gov/spectral-lib.html) is available online for any member of the public to download. The library was developed to support imaging spectroscopy studies of the Earth and other planets (USGS, 2006). Functionally, it is an html-based directory of spectra with associated metadata. There are 820 spectra, categorized into mineral, vegetation, man-made, mixture, volatile, microorganism, and plant samples. Each spectrum is stored as an image plot and metadata including sample name, description, chemical formula, sample donor, location, xrd analysis, with up to 24 metadata elements stored in pre-defined templates for each category of target. It is a static library in the sense that the data is read-only, and members of the public cannot upload new spectra or perform updates.

The SPECCHIO (http://www.specchio.ch/) database is available online for members of the public and can also be downloaded as a local instance. SPECCHIO was created by Remote Sensing Laboratories at the University of Zurich to store reference spectra and campaign data obtained by spectroradiometers in a central repository (Hueni *et al.*, 2009). It is accessible through a Java application, and all data is stored in a MySQL database. The public can upload spectra and metadata and make edits to their own datasets. It contains 111,023 spectra across 71 campaigns. Metadata is stored at both the spectrum and campaign level, some of which is auto-generated.

Spectral Library	Agency	Purpose	Year Created	Format	Campaigns	Spectra	Explicit quality assurance	Mandatory metadataset
USGS	USGS	• used as reference for material identification in remote sensing images • cataloguing of field and laboratory observations	2003	Static archive (online)	Data not defined at campaign level	820	No	Yes; pre-formatted templates
SPECCHIO	RSL	• designed to hold reference spectra and spectral campaign data obtained by spectroradiometers • rich metadataset in the data model for ensuring the longevity of spectral data and enables the sharing of spectral data between research groups •cataloguing of field observations	2007	Open access database (online and as a single or multi-user instance)	71	111,023	No	Yes; at campaign and spectrum level

Table 5.1 Overview of USGS and SPECCHIO spectral data libraries

Users have the option of additional metadata they wish to populate, either at the spectrum level (including viewing geometry, target homogeneity, environment information) or campaign level (including description, associated institute).

Both the SPECCHIO and USGS datasets had to be prepared for analysis. A database backup copy of SPECCHIO was provided by the RSL data center at the University of Zurich. The entire SPECCHIO database was restored as a local MySQL instance. The database schema required some redesign caused by data in the SPECCHIO database that violated the original schema as specified by the database designer. These schema violations were due to the fact that since becoming publicly available, data had been loaded into SPECCHIO by members of the public with no oversight as to whether it conformed to the original schema. Once the schema underwent a small amount of redesign, all the data that currently resides in SPECCHO could be loaded in to the local instance. A total of 111,023 spectra categorized into 55 campaigns (the SPECCHIO website advertises 71 campaigns but only 55 are available for analysis), with metadata stored across 61 interrelated tables were loaded into the local instance.

The USGS metadata was downloaded from the USGS Spectroscopy Lab website as a set of html files. It was then extracted from the html files (one file per spectrum) and transferred to a custom designed MySQL database. This was a time-intensive process as each file of spectrum data had to be extracted individually and then loaded into a database schema conforming to the dataset. As such, 90 spectra were chosen, comprising a random selection of 10% of the total datasets from each sample category (mineral, mixture, vegetation, micro-organism, man-made, volatile). Random numbers were generated using SPSS v. 21 software random number generator module to select the sample set. In the only category that had two spectra (volatile), both spectra were used to permit statistical analysis for that category. Using a range of samples permitted a more equable comparison with the SPECCHIO datasets, which are also varied in sample type. The number of samples was chosen based on statistically acceptable thresholds for sampling sizes in data mining (SAS 2010; Khandar and Dani, 2011).

5.3.2 Method

Assessing the quality and completeness in the data libraries was based on the parameters proposed for measuring metadata quality presented in Section 5.2. Completeness measures were entirely quantitative. The evaluation could be implemented as an automated process to individually evaluate the completeness of the metadata for 111,023 spectra in the SPECCHIO database and the 90 spectra in the USGS dataset using data querying utilities within MySQL. A mapping of metadata between SPECCHIO, USGS Spectral Library and the core metadataset is found in Appendix C.

The quality measure was an assessment based on the five proposed parameters (logical consistency, error rate, quality assurance, lineage, reputational authority) presented in Section 5.2. The choice to use a qualitative assessment for both test cases was based on the manner in which measures for logical consistency and error rates are typically derived. Both comprise counts of instances where a metadataset contains contradictory information or inconsistent formatting for the same unit of information. Both require a pre-defined vocabulary and a baseline set of reference metadata against which to verify semantic and syntactic errors and consistency. A reference metadataset in this case would be defined based on knowledge of the correct formatting and spelling of metadata elements such as names of data owners, campaign locations, and dates, none of which were specified in either the SPECCHIO or USGS database design or user guidelines; nor could they reasonably be expected to be provided by the database owners based upon the volume of data and the diversity of sources from which they originate (the SPECCHIO database, for example, has a single database administrator responsible for managing all data). These factors prohibited a practical implementation of an automated process to check the metadataset for each spectrum (111,0233 in SPECCHIO, 90 in USGS Spectral Library) for presence of errors or measures of logical consistency. Rather, analysis was applied to derive results that *implied* logical consistency or degrees of reputational authority, and included analysis such as cumulative entropy calculations for populated metadata parameters and completeness measures per database user and institute.

Both datasets underwent filtering and cleaning prior to analysis. In preparation for completeness analysis, metadata had to be searched for every instance of fields that would qualify as non-populated. These included fields with entries of null values, ‘None’, ‘none available’, ‘unknown’, ‘Not done yet’ and other similar variations. This was relevant mostly to the USGS data. The bulk of the data loaded into the library had been acquired via metadata templates that had undergone several iterative changes over the lifetime of the library, with each subsequent iteration being an expanded version of those before, therefore unpopulated fields in earlier datasets had default null values. In newer iterations of the metadata templates, users had the option of manually populating most metadata fields, and where there was no data for the user to enter, the user either left it blank, or explicitly stated that there was no data. In SPECCHIO, in most instances, if metadata is not entered by the user, it is automatically stored as a null value in the database.

5.3.3 Results

The completeness and quality reports for SPECCHIO and the USGS Spectral Library are presented in this section. The mappings for metadata elements between the core metadataset and SPECCHIO and USGS Spectral Library are found in Appendix C.

5.3.3.1 Metadata Completeness Analysis

A summarized completeness report for both SPECCHIO and the USGS Spectral Library is provided in Table 5.2. SPECCHIO and the USGS Spectral Library both show higher compliance with their internal metadata requirements (SPECCHIO at 59.3 % at campaign level and 52% at spectrum level metadata; USGS at an average 72% compliance for all samples) than with the proposed core metadataset (SPECCHIO at 18% and USGS at 7.7%). This is expected, as the SPECCHIO and USGS Spectral Library data managers do not have knowledge of what the core metadataset is, therefore have not implemented it in their metadata policy.

SPECCHIO metadata is defined at the campaign and spectrum level (16 and 35 metadata elements respectively). Almost every metadata element in the SPECCHIO set can be mapped to elements in the core metadataset – the majority of these pertained to viewing geometry, instrument information, location information, atmospheric information, illumination information, and general project information.

Spectral Library/Database		Completeness Statistics									
			Campaign Completeness (internal metadata policy)					Core Metadataset Compliance (campaign + spectrum)			
		# of campaigns examined	# of parameters	min	max	avg %	stdev	min	max	avg %	stdev
SPECCHIO		**55**	15	6	15	59.3%	12.7%	11	21	18.4%	1.3%
			Spectrum Completeness (internal metadata policy)								
		# of spectra examined	# of parameters	min	max	avg %	stdev				
		111 023	35	10	20	51.7%	4.0%				
			Spectrum Completeness (internal metadata policy)					Core Metadataset Compliance (spectrum)			
		# of spectra examined	# of parameters	min	max	avg %	stdev	min	max	avg %	stdev
USGS	Man-made*	**11**	24	15	18	68.8%	5.4%	7	8	7.2%	0.5%
	Microorganism*	**2**	19	11	14	65.8%	11.1%	7	10	8.2%	2.0%
	Minerals*	**44**	25	16	20	74.8%	4.8%	7	8	8.2%	0.5%
	Mixture*	**13**	25	16	23	82.0%	12.0%	6	11	8.0%	0.3%
	Plant*	**18**	19	11	16	62.6%	7.4%	7	10	7.1%	0.8%
	Volatile*	**2**	25	17	22	78.0%	14.0%	8	8	7.7%	0.0%
	Average					72.0%	9.1%			7.7%	0.7%

Table 5.2 Metadata Completeness report for SPECCHIO and USGS Spectra Library

In SPECCHIO, there is no calibration information metadata. In cases where a spectrum completeness measure for SPECCHIO exceeded the core metadataset completeness measure, this was due to additional metadata elements in SPECCHIO that do not exist within the proposed core metadataset. These include but are not limited to: three additional metadata elements pertaining to metadata quality – the ‘required quality level’ and ‘quality level’ flags at the spectrum level, and ‘quality comply’ flag at the campaign level; air pressure/ambient temperature/wind direction metadata in the environmental conditions category; ‘illumination distance’ in the sampling geometry category and database user, institute, and instrument manufacturer information (postal address, email, etc.) all of which are not explicitly referenced in the proposed core metadataset.

Mappings to the core metadataset incorporated SPECCHIO metadata elements at both the spectrum and campaign level, since much of the campaign level metadata can be mapped to the ‘General Project Information’ category in the core metadataset (including campaign description, relevant websites, and project participants). The database user who loaded the campaign into the database was designated as a project participant when mapped to the core metadataset. However, there was no metadata in SPECCHIO indicating who the field operators were. The core metadataset distinguishes between project participants, affiliates, and field instrument operators.

SPECCHIO spectra could not be categorized into individual sample types because there is no field describing the sample type (vegetation/mineral/aquatic/other) and the sample name in most cases is not informative. There is no information about the sample itself other than the ‘target name’ metadata field. The campaign description, in some cases, provides minimal information about the types of samples and purpose of the campaign.

Information about the hyperspectral signal properties is limited to type (reflectance/radiance/absorbance/transmittance/DN/wavelength/mueller10/muelle-r20/irradiance), wavelength interval, and wavelength data, that are assigned mostly to the ‘measurement type’ and ‘sensor’ metadata categories (SPECCHIO distinguishes between sensor and instrument information). The SPECCHIO user interface, via a Java application, does provide access to additional instrument and signal properties encoded within the instrument-native files (ASD binary, GER signature files, SVC HR-1024 files, among others), but these are not enforced by the internal SPECCHIO metadata policies. Rather, it is assumed that the user can load these retrospectively if they have a local installation of the database and they had customized it to allow additional metadata fields for instrument, sensor, and signal properties. Therefore SPECCHIO makes assumptions that users may not wish to populate all metadata at once, or do not need to view all metadata available while searching for the dataset of their choice.

None of the quality flags for SPECCHIO metadata were populated. These flags reference the level of completeness of the metadata only. At the spectrum level, both the ‘required quality level’ and ‘quality level’ can be populated. There are two rankings for both the ‘required quality level’ and ‘quality level’ parameters -- Level A (not defined or implemented in the current version of SPECCHIO) and Level B, which is defined to be a metadataset that “should make spectral data useable by third persons who were not directly involved in the capturing process and are thus not familiar with the sampling circumstances” (Hueni, 2011, p. 15). According to the SPECCHIO metadata policies, Level B metadata comprise campaign investigator, sensor, instrument, foreoptic, landcover, target homogeneity, measurement unit, sampling environment, measurement type, latitude, longitude, altitude, cloud cover, sensor/illumination azimuth and zenith, and target type. At the campaign level, the 'quality comply' flag is not defined. There is no SPECCHIO metadata policy that requires a minimum metadataset, and the metadata, once loaded, is not reviewed by the database administrator.

USGS Spectral Library metadata is populated according to templates categorized by sample type: man-made (rooftop shingles, asphalt, concrete, etc.), microorganism (lichen, bacteria, etc.), minerals (zinc, calcite, etc.), mixture (andradite, siderite. etc.), plant (trees, flowers, grasses, etc.), volatile (water, melting snow, etc.), each with varying degrees of maximum allowable metadata elements. The majority of the metadata describe the sample itself (sample ID, mineral type, Latin name, formula, etc.) including image metadata.

Remaining metadata refer to the location where the spectra were recorded (if outdoors), former and current sample location, original donor, and results of xrd and chemical analysis, where applicable. The original donor field was considered a project participant when mapped to the 'General Project Information' category in the core metadataset. Metadata referring to instrument, hyperspectral signal properties, calibration, viewing geometry, or illumination information do not exist within the metadata templates; such information is only available if the user chooses to include these in the 'Sample Description' metadata field. The metadata does not specify that the data itself is a reflectance measure, but this is stated on the USGS

Spectral Library website information pages. Instrument information, including wavelengths used in the measurement and spectral resolution, can be obtained from the SpectraProc files that are available separately from the USGS website. As with SPECCHIO, there is no specified minimum completeness level for metadata, nor is there any explicit evidence that the data is reviewed once loaded. The library is acknowledged not to have "…all samples completely characterized. The characterization of samples will continue as our resources allow, and results will be added in future releases of the database" (USGS, 2006). There are no completeness or quality flags in the metadata.

Investigating those categories where the database users were inconsistent in populating metadata categories can serve to inform the future design of metadata policies within databases, especially in those parameters relating to core metadata. When users are consistent in the way they populate the same set of metadata fields (they either populate them or not with little variance), it can be assumed that the users have a consensus opinion on whether these metadata are critical or not. Otherwise the cause can be attributed in part to system design. In the case where users are consistently populating the same fields, the database interface encourages or at the very least makes it easier for the user to populate those fields and conversely, inhibits users where consistently unpopulated fields are concerned. It is necessary here to assume that for metadata elements that are being inconsistently populated, users who are not populating these fields are technically literate and/or capable enough not to be inhibited by poor database user interface, and are not populating them of their own volition. Investigating users' motives is beyond the scope of this discussion, but it remains worthwhile, to highlight any patterns of variance, specifically, why certain users consider a given set of metadata fields important, while others do not.

The SPECCHIO spectrum-level metadataset is the most useful for this kind of analysis, because of the large number of spectra (111,023) and the uniform number of metadata elements (35) associated with each sample (the USGS Spectral Library is not uniform in its metadata policy for sample types). While SPECCHIO metadata is not sufficiently discretized to allow the segregation of users into specific groups, it is possible to identify those fields that contribute to the greatest variance.

SPECCHIO spectrum-level metadata was analysed to determine if there were patterns of variance for completeness levels. The method of analysis chosen was dimensional scaling of the data, to better understand the variance and co-variance relationships among the SPECCHIO metadata elements for spectrum-level completeness. This was accomplished with categorical principal component analysis (with ordinal measurement) to determine those metadata parameters that cluster together, by their proportionate variance, for completeness measure. Principal components analysis generates linear combinations (dimensions) of the original variables (metadata elements) expressed as proportions of variance. Categorical principal components is a method specialized for categorical data ('populated' or 'not populated') and does not require normal distributions for input (Linting, *et al.*, 2007; Meulman and Heiser, 1989, 2012; Starkweather, 2012). All zero-variance metadata elements were excluded, and these were 'IsReference', ‘ReferenceSerialNumber’ and ‘ReferenceBrandName’ (all referring to the reference standard used while taking measurements) and the ‘RequiredQualityLevel’ and ‘QualityLevel’ fields (these were not populated for any spectra).

The analysis yielded seven dimensions for the spectrum-level metadata. The choice of seven dimensions was based on prior factor analysis testing that showed seven factors was the threshold at which 85-90% of the cumulative variance could be accounted for. Only factors with eigenvalues greater than 1 were extracted (Kaiser, 1960). Table 5.3 shows the (metadata element) loadings for each dimension.

The highest loading for each metadata element has been highlighted in bold. The results show that dimension 1 is principally viewing geometry (‘SensorAzimuth’, ‘SensorDistance’, ‘SensorZenith’, ‘IlluminationZenith’, ‘IlluminationZenith’), hyperspectral signal properties (‘MeasurementType’, ‘InternalAverageCount’), and location information (‘SamplingEnvironmentName’, ‘LocationName’). Dimension 2 is almost exclusively environmental conditions ('AirPressure', 'AmbientTemperature', 'RelativeHumidity', 'WindDirection', 'WindSpeed'). Dimension 3 is exclusively instrument information (‘ManufacturerName’,’ ManufacturerShortName’, ‘SensorDescription’, ‘SensorName’, ‘SensorNoOfChannels’). Dimension 4 is primarily location information, (‘LandcoverDescription’, ‘Altitude’, 'Latitude', 'Longitude') with two elements of instrument information ('InstrumentName', 'InstrumentSerial_number') and one from sample properties (‘TargetHomogeneity’). Dimension 5 has one metadata element from viewing geometry ('IlluminationDistance') and one from instrument information

(‘ManufacturerWWW’). Dimension 6 has its highest loading for one parameter from environmental conditions (‘CloudCoverInOctas’), and dimension 7 is primarily hyperspectral signal properties (‘MeasurementUnit’).

Metadata element	**Dimension**						
	1	2	3	4	5	6	7
AirPressure	.034	**.797**	-.555	-.228	-.010	-.047	-.003
Altitude	-.299	-.415	.034	**-.670**	-.111	-.338	.207
AmbientTemperature	.034	**.797**	-.555	-.228	-.010	-.047	-.003
CloudCoverInOctas	.066	-.015	-.011	-.114	-.090	**.600**	-.001
IlluminationAzimuth	**.553**	-.304	-.107	-.516	-.432	.090	-.199
IlluminationDistance	-.073	.003	.037	-.125	**.439**	-.219	-.420
IlluminationZenith	**.872**	-.219	-.133	-.177	.036	-.099	.188
InstrumentName	.457	.052	-.075	**.618**	-.463	-.199	-.027
InstrumentSerialNumber	.457	.052	-.075	**.618**	-.463	-.199	-.027
InternalAverageCount	**-.892**	-.060	.158	-.201	.053	-.065	-.126
LandcoverDescription	-.257	-.042	.065	**-.485**	.116	.441	.271
Latitude	.543	-.311	-.101	**-.570**	-.422	.121	-.167
LocationName	**.903**	-.197	-.127	-.085	.157	-.191	.042
Longitude	.543	-.311	-.101	**-.570**	-.422	.121	-.167
ManufacturerName	.226	.511	**.810**	-.126	-.104	-.042	.011
ManufacturerShortName	.226	.511	**.810**	-.126	-.104	-.042	.011
ManufacturerWWW	.227	.274	.383	.087	**.452**	.315	-.211
MeasurementType	**.918**	.086	-.128	.243	.118	.177	.054
MeasurementUnit	.125	-.003	.037	-.076	.038	-.052	**.720**
RelativeHumidity	.034	**.797**	-.555	-.228	-.010	-.047	-.003
SamplingEnvironmentName	**.904**	-.124	.000	-.150	.309	-.049	.026
SensorAzimuth	**.919**	-.089	-.028	-.034	.251	-.045	-.025
SensorDistance	**.921**	-.005	-.071	.084	.239	.052	-.001
SensorZenith	**.929**	-.083	-.027	-.020	.271	-.049	-.010
SensorDescription	.226	.511	**.810**	-.126	-.104	-.042	.011
SensorName	.226	.511	**.810**	-.126	-.104	-.042	.011
SensorNoOfChannels	.226	.511	**.810**	-.126	-.104	-.042	.011
TargetHomogeneity	-.186	-.337	.219	**-.644**	.273	-.379	-.054
WindDirection	.034	**.797**	-.555	-.228	-.010	-.047	-.003
WindSpeed	.034	**.797**	-.555	-.228	-.010	-.047	-.003

Table 5.3 Dimension loadings for SPECCHIO spectrum level metadata completeness using categorical principal components analysis with variable principal normalization. The highest loading for each metadata element has been highlighted in bold.

The first three dimensions account for 63% of the total variance (with progressively diminishing variance loading on the remaining four dimensions). The first three dimensions relate strongly to viewing geometry, instrument information, hyperspectral signal properties and environmental conditions, all of which are elements of the core metadataset. These findings invite future investigation as to why database users are not consistent in populating metadata in these three categories that have been identified by their peers (in Chapters 3 and 4) as critical to all field spectroscopy metadatasets. Unpopulated metadata in these categories is fundamentally compromising the overall quality, interoperability, and intercomparison of these datasets. These findings also invite data managers and stakeholders to educate data creators about the importance and implications of metadata completeness, and to implement metadata policies within data sharing platforms that force data creators to comply with given levels of completeness.

5.3.3.2 Metadata Quality Analysis

In the absence of metadata quality flags in both SPECCHIO and the USGS Spectral Library, a metadata quality analysis was completed on parameters including logical consistency, error rate, lineage, quality assurance, and reputational authority. A comprehensive analysis was not possible for all parameters, and this is discussed in more detail in the sections below.

Quality Assurance

Neither SPECCHIO nor USGS Spectral Library have any metadata quality measures (aside from those discussed in Section 5.3.3.1 that are inapplicable) or quality assurance parameters.

Lineage

Neither SPECCHIO nor USGS Spectral Library have lineage metadata for records.

Reputational Authority

SPECCHIO and the USGS Spectral Library do not have explicit reputational authority metadata for data creators. Reputational authority can be established if metadata about the data creator or owner includes information about their professional affiliations, publications, projects on which they have worked, and other similar data that allows user to make value judgements about whether the data creator has sufficient gravitas within the research community to produce reliable datasets. However, when this metadata is absent, there are ways of establishing reputational authority implicitly or indirectly. This is the case for the SPECCHIO database, in which each spectrum is associated with both a database user and the institute under which they are registered (multiple users can belong to one institute). Measuring data owner or data creator compliance to metadata policies supplies the data user with some information on which to form an opinion about the reliability of the data creator. The premise for this argument being, if a data creator is being diligent in complying with metadata policies, then they are likely to be diligent in producing reliable and higher quality datasets than their counterparts.

Analysis of variance was one method of determining the effect of user and institute on completeness measures. This was computed on normalized completeness measures using a one-way between subjects ANOVA. Completeness measures used were those for SPECCHIO campaign and spectrum-level metadata, and for the proposed core metadataset. There was a significant effect of user on completeness measures at the $p<0.001$ level for the 26 users examined $F(25,110997) = 280.45$ for SPECCCHIO spectrum, $F(25,46174) = 1488.79$ for SPECCHIO campaign, and $F(25,110997) = 337.75$ for the proposed core metadataset. There was a significant effect of institute on completeness measures at the $p<0.001$ level for the 15 institutes examined $F(14,111008) = 289.81$ for spectrum, $F(14,46185) = 1325.23$ for campaign, and $F(14,111008) = 348.79$ for core.

Z-scores were calculated from the raw completeness measures to determine whether they differed across users and institutes. A Z-score quantifies the original completeness values in terms of the number of standard deviations that that value is from the mean of the distribution. It is useful for identifying any users or institutes which have values below or above the mean. Z-scores above zero indicate that a given user or institute populates metadata to a higher level of completeness than their peers; the reverse is true for Z-scores below zero.

Figures 5.1 and 5.2 show the mean Z-score for spectrum, campaign, and core dataset completeness for each user and institute, respectively, with the mean for all scores at $y=0$. The Z-score is a calculation of the distance of each user or institute from the mean completeness score for all users or institutes.

The Z-scores for database users (Figure 5.1) show overall poor completeness levels for spectrum, campaign, and core metadataset completeness. They indicate that spectrum-level and core metadataset compliance exhibit similar scores, mostly due to the fact that a large proportion of the spectrum-level metadata is a subset of the core. The mean Z-score ranges were 12.6 for the proposed core metadataset completeness, 10.3 for SPECCCHIO campaign-level completeness and 13.9 for SPECCHIO spectrum-level completeness.

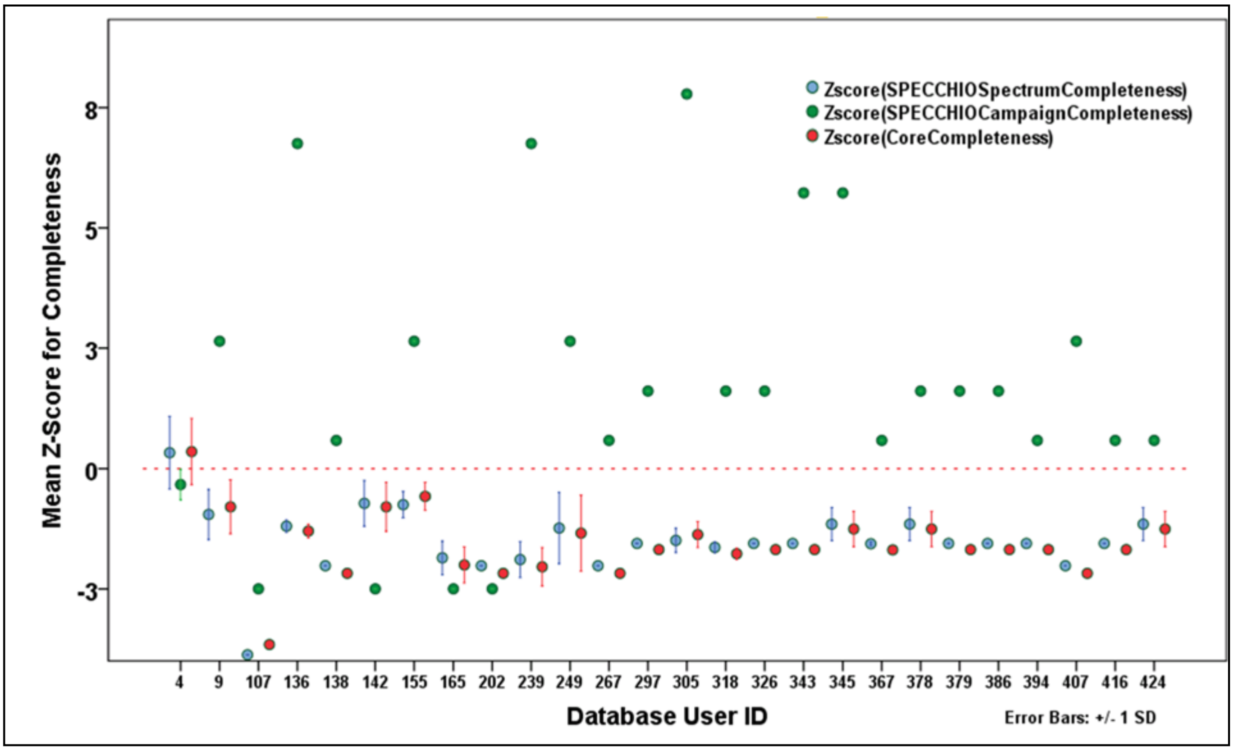


Figure 5.1 Mean Z-scores for completeness by database user

The highest mean Z-scores for spectrum-level and core metadataset completeness belong to user 4 (accounting for 82% of the spectra), user 142 (< 1% of the spectra) and user 155 (<1 % of the spectra). The lowest mean Z-scores for spectrum-level and core metadataset completeness belong to user 107 (<1% of the spectra), user 267 (<1% of the spectra) and user 407 (1% of the spectra). The highest campaign completeness scores belong to user 136 (<1% of the spectra), user 239 (<1% of the spectra) and user 305 (<1% of the spectra). The results show that a high spectrum-level completeness does not imply the same degree of campaign completeness for a given user, therefore the must be considered separately when assessing reputational authority.

The Z-scores for the institute associated with each spectrum (Figure 5.2) indicate the same degree of similarity between spectrum-level and core metadataset completeness as with the Z-scores for database users, but again, overall poor performance for completeness.

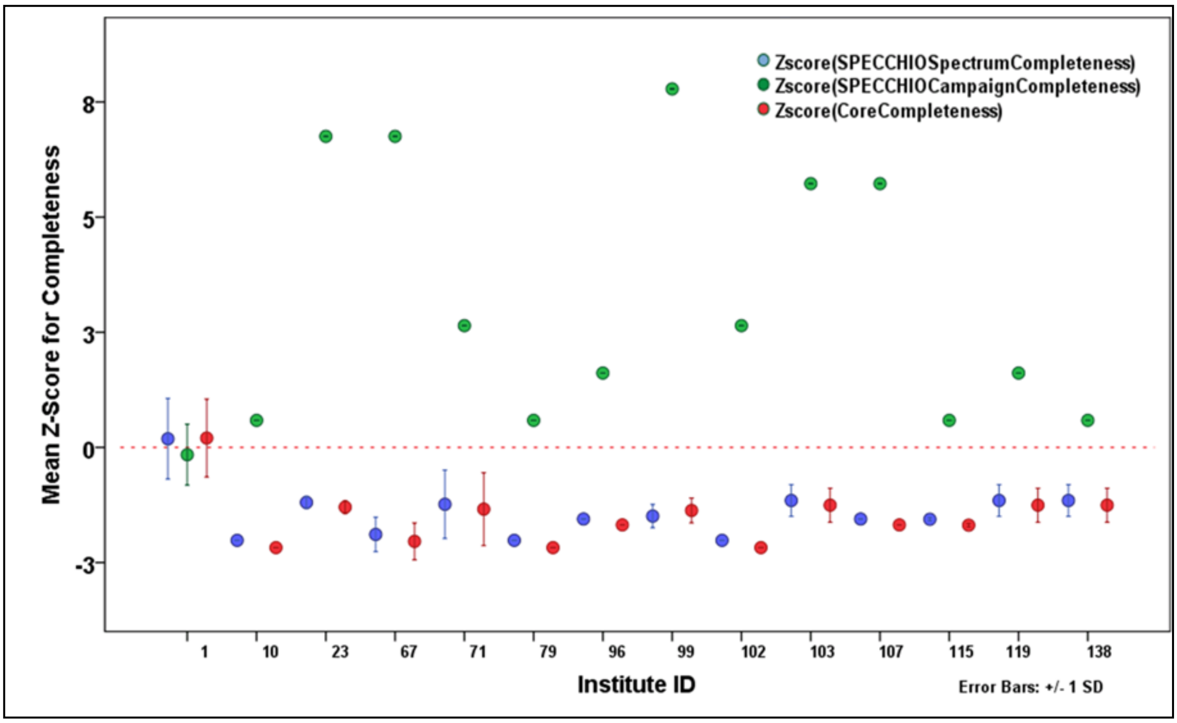


Figure 5.2 Mean Z-scores for completeness by institute

The highest mean Z-scores for spectrum-level and core metadataset completeness belong to institute 1 (89% of the spectra), institute 103 (<1 % of the spectra), institutes 119 and 138 (<1 % of the spectra). The lowest mean Z-scores for spectrum-level and core metadataset completeness belong to institutes 10, 79, and 102, each accounting for 1% or less of the spectra). The highest campaign completeness scores belong to institutes 23, 67, 99, each accounting for less than 1% of the spectra. The highest-performing institute for spectrum-level and core metadataset completeness, Institute 1, is also associated with the top-scoring users for spectrum and core metadataset completeness (users 4, 155) and is not associated with any of the lowest-scoring users.

The results show that in the absence of explicit information relating to the reputational authority of the metadata creators, it is still possible for a data user to form an opinion about the reliability of the data creator. For example, in the SPECCHIO database, the highest-ranking database users and institutes for metadata completeness could be identified. Since they were demonstrably diligent in complying with metadata policies, it can be assumed that they are likely to be diligent in producing reliable and higher quality datasets than their counterparts. These results suggest that in order to aid the data user in making informed choices about the suitability of a dataset, the conventional definition of reputational authority can be expanded to include implicit measures.

Error rates

A systematic assessment of syntactic and semantic error rates was not possible due to the absence of a reference dataset for either SPECCHIO or the USGS Spectral Library, as discussed in more detail in Section 5.3.2. Instances of metadata that were presumed to be erroneous are noted here for illustration purposes only. This was mostly relevant to the USGS Spectral Library, due to its numerous free-form text metadata elements. Examples of presumed semantic errors include: 'image sample' metadata field left null when image is attached (BR93-33Arecord); 'XRD analysis' metadata not clear about whether data does not exist or the analysis did not yield results: 'See/'Unknown' (multiple records). Examples of presumed syntactic errors include: variations of the spellings in 'original donor' field presumably representing the same entity: 'Greg Swayze'/'Gregg Swayze' (multiple records).

Logical Consistency

Logical consistency for a metadata instance can be defined as "the degree to which it matches the metadata standard definition" (Ochoa and Duval, 2009, p. 9). It can be measured in part by the type and amount of information that users are entering into the metadata fields. Inconsistencies can be caused by incompetent data entry, or fundamental systematic problems in the metadata policy. Ruling out incompetent data entry, the effects of systematic problems can be manifest if one group of users is recording metadata in a markedly differently way than other users, whether by populating a given field with too little or too much information, or with information not within the standard definition of what that metadata field is designed to represent. This can suggest that the metadata policy is not consistent with their needs as a user group.

The USGS Spectral Library, based on its numerous free-form text fields, and metadata templates specialized by sample type, permits this kind of examination.

The metadata instance chosen for analysis was ‘sample description’. This metadata element is used by the USGS Spectral Library users to provide details about a given spectroscopic sample, including a physical description, core compounds, trace elements, main spectral features etc. The two groups chosen for comparison were the vegetation community (designated as all data users who populated the vegetation metadatasets) and the non-vegetation community (all other users). Inspection of USGS records revealed that vegetation spectroscopy on live samples documented in the USGS Spectral Library was more likely to be done in the field (rather than many of the mineral samples that were examined in the lab). Therefore, the vegetation sample metadata would be a more accurate reflection of metadata arising from a field campaign.

The method of analysis was a comparison of cumulative entropy measurement (Simon, 2010) for the ‘sample description’ text length between vegetation and non-vegetation groups. Text length was used as a measure of how much data users are entering into the sample description. The reasonable assumption was that a larger text length denoted more data. Since there was no pre-defined vocabulary or a baseline set of reference metadata within the USGS Spectral Library against which to verify the kind of information that users should input for ‘sample description’, text length was the most the suitable measure given the data available.

Entropy is a concept derived from thermodynamics used to describe the possible microstates of a system. It has been extended to information theory and computer science to be defined as the amount of information required on average to describe a random variable (Cover and Thomas, 1991) (Equation 5.1)

The entropy H(X) of a discrete random variable X is defined by

$$H(X) = - \Sigma\, p(x) \log p\,(x) \qquad \textbf{(Equation 5.1)}$$

Entropy is calculated in log base 2 as a quantity in bits for computer science applications. When applied to a discrete variable representing categories of information (in this instance, the ‘sample description’ field), entropy is large when each category has roughly the same proportion, but small when the probability is concentrated in a few specific categories (Simon, 2010). Entropy and cumulative entropy are useful for metadata quality analysis because they can be used to identify changes in data entry characteristics (Simon, 2010) and as a measure of the diversity of information being stored (Stvilia *et al*., 2004).

Prior to entropy analysis, the probability of each ‘sample description’ text length had to be calculated. All null values were changed to 0 (indicating that user had entered no data, therefore having a text length of 0). The non-vegetation group was separated from the vegetation samples, and a subset was randomly selected as a training set. Within the training set, 20 bins for text length were created, based on percentiles, at a width of 5%. The 20th percentile included those values higher than 1231 characters (the maximum text length in the training set). A probability was then assigned to each bin based on the number of occurrences of values within a given bin. Based on the training dataset, the largest value expected was 1300 characters in length. For the purpose of analysis, both the lowest cut-off (0) and highest cut-off (1300) were considered to have no bounds and extend into either negative or positive infinity.

The probabilities derived from the training dataset were then assigned to the vegetation and non-vegetation groups. Cumulative entropy was calculated on two sets of data: a non-vegetation-only group, and a mixed vegetation and non-vegetation group. The cumulative entropy graph is shown in Figure 5.3.

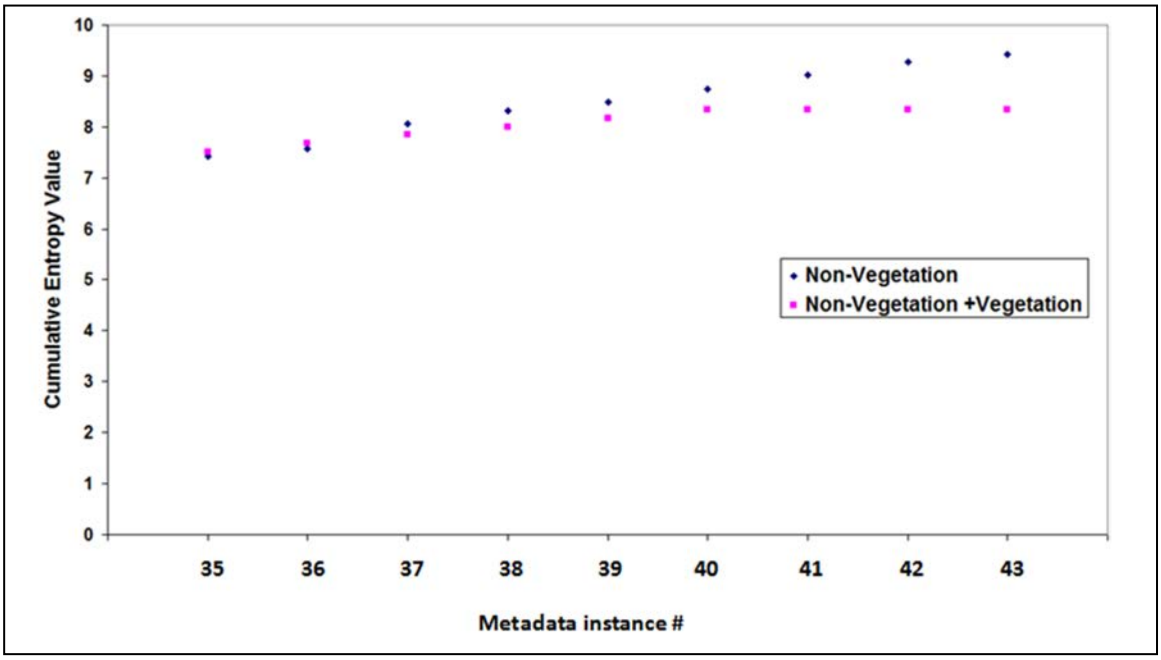


Figure 5.3 Cumulative entropy for non-vegetation and mixed groups

The metadata instance represents an individual 'sample description' field. A bifurcation is visible at approximately the 35th metadata instance, after the vegetation group is introduced to the mixed group. With each vegetation instance added, the cumulative entropy remains nearly constant at a value of 8, whereas cumulative entropy for the non-vegetation group continues to rise. This is explained by the fact that the text length for the vegetation ‘sample description’ instances had an overall lower probability of occurring, because they were beyond the normal expected length (1300 characters) derived from the training dataset. Two vegetation instances in particular had the highest ‘sample description’ text length in the entire metadataset, at 1742 and 6082 characters.

Closer examination as to what was causing such large values for text length revealed that the vegetation group is using this metadata element to store detailed and explicit information about field data collection protocol including viewing geometry, sensor information, illumination information, target homogeneity and atmospheric conditions. This suggests that the vegetation metadata template in the USGS Spectral Library is insufficiently structured and lacks the richness required to permit users to store the information for vegetation field spectroscopy in a logically and semantically consistent way.

5.3.4 Discussion

The results show that in the completeness and quality measures, SPECCHIO and the USGS Spectral Library are not aligning well with the needs of field spectroscopy scientists as identified in Chapters 3 and 4. Overall, the low scores on completeness and generally poor metadata quality in both cases are a hindrance to discoverability of the data, interoperability with other datasets, and make it difficult for data user to assess whether a given metadataset is suitable for their purpose.

SPECCHIO has an average completeness measure of 51.7% at the spectrum level, 59.3 % at the campaign level, and 18.4% compliance with the core metadataset. The USGS Spectral Library has an average completeness level of 72% across the metadata templates and 7.7% compliance with the core metadataset. The two datasets fail to comply completely with their internal metadata policies, with 59.3% compliance with the campaign-level and 51.7% with spectrum-level completeness for SPECCHIO, and average of 72% compliance for samples in the USGS Spectral Library. There are no metadata quality parameters in either dataset, aside from two spectrum-level quality parameters in SPECCHIO that describe metadata completeness; the third quality parameter, at the campaign level, is undefined. None of the quality parameters in SPECCHIO have been populated for any dataset in the database.

The five metadata quality parameters selected to assess SPECCHIO and the USGS Spectral Library were 1) logical consistency, 2) lineage, 3) semantic and syntactic error rates, 4) quality assurance by a recognized authority, and 5) reputational authority of the data owners/data creators. However, only two (logical consistency and reputational authority) could be evaluated based on the datasets available. In both SPECCHIO and the USGS Spectral Library, there is a lack of metadata quality assurance or lineage information. Presumed semantic and syntactic errors could be identified within the USGS Spectral Library given the numerous free-form text fields used within its metadata templates, but for both the USGS Spectral Library and SPECCHIO, it was not possible to automate this process given the lack of a reference dataset or metadata dictionary to use for comparison.

A preliminary estimate of reputational authority was established within SPECCHIO by identifying the highest and lowest completeness measures for spectrum-level and campaign-level metadata by user and by institute. Logical inconsistency within the USGS Spectral Library metadata was identified by entropy analysis which showed that vegetation spectroscopy metadata is being populated by users in a very different manner from non-vegetation metadata.

The methods and algorithms used in these test cases for quality and completeness assessment could be used on any field spectroscopy metadataset, given that in the special case of semantic and syntactic errors, a reference metadata dictionary is available for identifying such errors. This kind of analysis would serve database designers, standards organizations, and the field spectroscopy community in identifying areas where users are not educated on which metadata are critical, and in identifying systematic problems with metadata policies.

5.5 Conclusions

Metadata quality and completeness measures for field spectroscopy can be defined using numerous criteria. In order to be useful for data mining, they must be informative for users who will make decisions on the usefulness of the data for their application/purpose. Field spectroscopy metadata completeness can be defined as a two-fold measure consisting of: a) compliance with the core metadataset and application-specific metadata (presented in Chapters 3 and 4); and, b) compliance with the standards of the data infrastructure in which they are stored. Metadata quality can be defined in terms of (but not limited to) logical consistency, lineage, semantic and syntactic error rates, compliance with a quality standard, quality assurance by a recognized authority, reputational authority of the data owners/data creators.

Publicly available datasets are underperforming on these quality and completeness measures. The two test cases examined, SPECCHIO and the USGS Spectral Library, neither have quality assurance metadata, nor do they align to any considerable degree with the proposed core metadataset (SPECCHIO at 18% and USGS at 7.7%). Lineage metadata was consistently negligible or absent for both datasets, and an examination of the USGS Spectral Library revealed logical inconsistencies in the metadata being populated by the users, as well as semantic and syntactic errors. Reputational authority associated with SPECCHIO could be established using completeness measures by user and institute.

The metadata quality and completeness measures presented here can be easily implemented for wide-scale assessment of metadatasets. They were developed with a focus on the users' needs in discovering metadata and assessing it as suitable for their purposed, an underlying principle currently lacking in existing metadata standards (Goodchild, 2007). Adopting these metadata quality and completeness measures as a standard can be of great service to the field spectroscopy community. They are built on a foundation of a metadataset established as critical by the field spectroscopy community and have incorporated additional elements of metadata quality parameters that serve to enhance the discoverability and interoperability of datasets.

Given that the spectral libraries examined in this chapter are state-of-the-art for publicly available field spectroscopy datasets, their shortcomings identified here highlight the urgency with which metadata policies, database design and user education need to be addressed in the context of quality assured metadata for discovery, interoperability, and sharing.

Chapter 6 Issues to adoption of a field spectroscopy metadata standard

Published in part as:

Rasaiah, B.; Jones, S.D.; Chisholm, L.; Hueni, A.; Bellman, C.; Malthus, T.J.; Chisholm, L.; Gamon, J.; Huete, A.; Ong, C.; Phinn, S.; Roelfsema, C.; Suarez, L.; Townsend, P.; Trevithick, R.; Wyatt, M. (2013). Approaches to Establishing a Metadata Standard for Field Spectroscopy Datasets. *Proceedings of IGARSS 2013*, Melbourne, Australia, 2013. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2012). Critical Metadata Protocols in Hyperspectral Field Campaigns for Building Robust Hyperspectral Datasets, *Proceedings of the XXII ISPRS Congress*, Melbourne, Australia, 2012. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2011). Designing a Robust Hyperspectral Dataset: The Fundamental Role of Metadata Protocols in Hyperspectral Field Campaigns*.* *Proceedings of Geospatial Science Research Symposium*, Melbourne, Australia, 2011. (peer-reviewed)

Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. (2011). Building Better Hyperspectral Datasets: The Fundamental Role of Metadata Protocols in Hyperspectral Field Campaigns. *Proceedings of the Surveying & Spatial Sciences Conference,* Wellington, New Zealand, 2011. (peer-reviewed)

6.1 Introduction

This chapter addresses research question #4, 'What are the issues related to adoption of the proposed field spectroscopy metadata standard?' It integrates the outcomes from the three preceding research questions, and the lessons learned from other disciplines in their respective metadata practices, into a set of recommendations for adoption and implementation of a field spectroscopy metadata standard. These recommendations apply to both the field spectroscopy community and in the wider scope of IT infrastructure for storing and sharing field spectroscopy metadata in data warehouses and big data environments. The recommendations are divided into two main sections: approaches to community adoption of the standard, and integration of standardized metadatasets into data sharing platforms. The recommendations are summarized in Table 6.4 at the conclusion of Section 6.3.

6.2 Recognizing obstacles to community adoption

6.2.1 Identifying obstacles that must be addressed

Implementation of a standard consisting of the core metadataset (proposed in Chapter 3), the application-specific metadata (proposed in Chapter 4), and extended metadata modules (proposed in Chapter 4) requires, first of all, acceptance and adoption by the field spectroscopy community. The community in this context encompasses the field operators, instrument providers, scientists and associated data creators who create the initial metadata set, the data owners, managers and stakeholders, and the organizations and advisory bodies responsible for formally recognizing and recommending the standard.

The adoption of, and compliance to, metadata standards has encountered difficulties in both the geospatial and non-geospatial disciplines (Wayne, 2001; Barton *et al*., 2003; Brownfield and Oliver, 2003; Green *et al*., 2005; Palmer *et al*., 2007; Devillers *et al*., 2010; Qin *et al*., 2012; Hendler, 2013). These difficulties can arise from causes that include a lack of knowledge within the community about the importance of metadata, logistical obstacles to recording metadata, no clear objective or purpose for metadata collection, and a lack of IT infrastructure (software and hardware) for supporting and enforcing the standard. Obstacles to creating metadata can exist in the field while collecting the spectral measurements, and in the period after the field spectroscopy campaign is complete when the metadata is uploaded to a local or central data repository.

6.2.2 Logistical obstacles in field spectroscopy

Respondents to the metadata survey presented in Chapter 3 and scientists participating in the workshops described in Chapter 4, identified logistical obstacles to documenting metadata in the field. These obstacles have implications for creating a metadataset that is as complete as possible, and therefore, impact a dataset's relevance, longevity and re-use. For example, the marine environment presents unique challenges for measuring objects and documenting variables that influence the measurements. These variables can include water turbidity, wave lensing, and the inherent difficulty of having the resources on hand, such as paper or electronic devices (laptop, tablet, mobile phone) to record the metadata in water or near water environments. Marine scientists have been creative in using data from photos to document metadata retrospectively, such as sample properties (size and colour of samples, underwater conditions).

Metadata relevant for all applications, such as those pertaining to instrument information in the core metadataset, can also include metadata that are not feasible to document concurrently with the spectral measurements in their entirety due to constraints of time during the campaign. These metadata elements may include instrument serial number, manufacturer, instrument housing (for extreme weather conditions or non-terrestrial campaigns), the degree to which an instrument has been customized for a particular application and whether it is a prototype, and sensor behaviour affected by manufacturer design.

As an example, PANalytical Boulder (formerly ASD) supplies a device known as a ‘Scrambler’ for its FieldSpec models to compensate for spectral discontinuities due to non-uniformity of field-of-view across the sensor bank fibreoptics. Documenting information about the Scrambler will help users of the data to know 1) that the FieldSpec models do not have uniform field-of-view and 2) whether the field operators compensated for this appropriately when taking spectral measurements. It will ultimately help a data user to determine whether they wish to use a dataset generated under these conditions. While it is impossible to account for every combination of instrument, environment, and sampling strategy variables, the core metadataset presented in Chapter 3, the application-specific metadatasets presented in Chapter 4, and the protocol module for specifying sampling strategies in the hybrid model also presented in Chapter 4 are as comprehensive as possible in addressing the range of variables inherent to field practice. The protocol module allows field operators to document details about field methods that are not explicitly specified in the core or application-specific metadatasets.

6.2.3 Lessons learned from others

Compliance tests for data in the USGS Spectral Library and SPECCHIO database against the internal metadata policies for these databases show inconsistent and occasionally low compliance by data producers. Reasons for this can only be speculated upon, but may include a lack of metadata on the part of the data producer, data producers choosing not to populate the metadata fields, and/or a lack of understanding among data producers about the importance of documenting metadata. Additional reasons for low compliance can arise from limitations of the user interface for populating metadata fields, or the metadata policy itself being ill suited to accommodate the spectrum of metadata that could be provided by the data producer. Overwhelming standards, requirements of excessive time and resources, and few perceived benefits and incentives have been identified as obstacles to adherence to geospatial metadata policies (Wayne, 2001).

Finding the right balance between enough metadata to assist users or applications to find relevant data and not overdefining the specification to make it brittle and unwieldy is a challenge metadata developers are just beginning to take on (Hendler, 2013). Creation of metadata within geospatial science in general is a time-consuming task, with the most successful applications of metadata being related to discoverability of datasets based on keywords (Devillers *et al*., 2010). As an example of an adopted standard with geospatial community, CSDGM (Content Standard for Digital Geospatial Metadata) is generally used in its truncated format, rather than implemented as a full and extensive schema (Qin *et al*., 2012). Extrapolating this to the field spectroscopy standard, compliance can be reasonably predicted to be less than complete (i.e. less than 100%).

Implementation of metadata standards in other disciplines has revealed that complete compliance is neither possible nor is it necessarily a sensible goal, given the time and resources required for complete compliance. For example, in online learning applications, the generation of adequate metadata for resource discovery appears to be universally problematic. In a trial project to create a metadata repository for digital resources using a 15-element metadata standard (Brownfield and Oliver, 2003) uncompliant datasets resulted from: a lack of skill among the metadata creators to describe datasets; a limited set of keywords to describe contact; and a lack of controlled vocabulary for metadata creators to use. In an analysis of the adoption of metadata standards within the learning object repositories and open e-Print archives, lack of compliance was attributed to factors including metadata creators not understanding the purpose or value of metadata, nor the context of their datasets (Barton *et al*., 2003). Within the ecological community, 'good' metadata is not the cultural norm as research data is not typically published or shared (Green *et al.*, 2005). In the library science community, an institution's adoption of, and level of compliance to, a metadata standard is affected by the choices of its peer institutions, and whether the standard is suitable for their information archiving systems; local metadata practices were found to conflict with standards designed to increase interoperability (Palmer *et al.*, 2007).

6.2.4 A way forward for overcoming obstacles

A general solution to overcoming potential compliance problems for a field spectroscopy metadata standard is to encourage good practice, and not impose onerous and time-consuming requirements that could discourage the community from adhering to the proposed standard. Good practice can be supported through specific activities and community behaviour supportive of a metadata standard.

These include:

- metadata prioritization (§6.3.1)
- allocating the responsibility of metadata creation to more than one party (§6.3.2)
- identifying the role of metadata in a user's decision making process (§6.3.3)
- building software tools and information systems supportive of standard compliant metadata (§6.3.4)
- education initiatives for the community (§6.3.5)
- certification of datasets (§6.3.6)
- additional approaches (§6.3.7)

The following sections elaborate on these practices.

6.3 Approaches to community adoption of a field spectroscopy metadata standard

6.3.1 Metadata prioritization

Metadata prioritization, or preference of certain categories over others, is a common behaviour among metadata creators (Stvilia *et al*., 2004; Heath *et al*., 2005; Liolios *et al*., 2012). The reasons for this include logistical constraints, personal preference, and complacency. The core metadataset, within the proposed field spectroscopy metadata standard, has addressed the need for a truncated format of the standard by identifying the priority metadata, i.e., the minimum and most essential metadata parameters required for a given field spectroscopy dataset. Likewise, the critical metadata elements of the application-specific metadatasets serve the same purpose for the coral, tree crown, and soil application areas.

Additionally, metadata creators in field spectroscopy applications can prioritize metadata that should be documented in the field over that which can be populated retrospectively. This would balance the need for complete metadatasets with the constraints of time and conditions under which the measurements are being taken. For example, numerous instruments encode instrument and signal properties information within their native files that can then be exported as metadata to a local or central database or other data repository. Other metadata parameters, including local weather information, or environmental conditions such as tide information, which are components of the core metadataset and application-specific metadatasets respectively, can also be documented post-campaign. A metadata creator could specify that these types of metadata are available upon request, or available through a third party when they make a dataset available for sharing.

6.3.2 Allocation of responsibility for metadata creation

Allocating responsibility for creating metadata to multiple parties reduces the individual workload for metadata documentation and ensures that distinct categories of metadata (calibration, field protocol, security and access rights, project information) are the responsibility of those best qualified to document those categories. Coordinated metadata stewardship is one of many recommended best practices for data governance within organizations (DAMA, 2010; Digital Curation Centre, 2010; ISO/IEC, 2012; USGS, 2013*b*; ANDS, 2014; MIT Libraries, 2014).

Past implementations of metadata management systems have yielded similar recommendations. Assigning panels of participating scientists to facilitate exchange of metadata between scientists in the field and data managers at data centers, as proposed for the US JGOFS (Joint Global Ocean Flux Study), permits collection of missing metadata and maintains quality control (Glover *et al*., 2006). Sharing metadata expertise and identifying meta-tagging experts also ensures consistency across datasets (Brownfield and Oliver, 2003). To illustrate this point, information specialists have demonstrated a better understanding of the purpose of metadata and generate more complete metadatasets than non-information specialists but have difficulty with contextual aspects of the metadata (Barton *et al*., 2003).

Extrapolating this to field spectroscopy, as an example of a dataset generated for vegetation campaign, those with the knowledge of instrument calibration activities would be responsible for calibration metadata, vegetation specialists would be responsible for populating vegetation sample metadata in the field, and non-domain specific metadata relating to provenance and security and access rights would be assigned to IT specialists and legal administrators respectively. Table 6.1 is one possible example of the allocation of responsibility for different modules within the proposed metadata standard for a tree crown reflectance metadataset.

	Metadata module	Examples of metadata included	Parties assigned responsibility
Core metadataset	Instrument	model, manufacturer name, spectral bandwidth, etc.	Instrument manufacturer
	All other core metadata	viewing geometry, illumination information, hyperspectral signal properties, location information, etc.	Field operators
Application-specific metadata	Target properties	tree species, tree crown size, DBH, etc.	Field operators
Ancillary	Project	information about the research context and purpose, experiment design, funding and sponsorship	Principal investigator
	Protocol	documentation of (or references to) the sampling and field protocols used in the collection of the field data, such as those for hyperspectral ground calibration, leaf sampling, underwater coral sampling	Principal investigator/Field operators
	Citations	relevant literature, publications, reports, journal articles, etc. cited in the metadataset or specifications about how the dataset itself should be cited externally	Principal investigator
	Dataset	broad-scope information that describes the entire dataset and includes title of the dataset, metadata standard name and version, revision history, keywords, purpose, and other general descriptors, for the main purpose of cataloguing and discoverability	Data managers
	Resource	information about the creators/owners/distributors of the data, lineage information, and contact information for the data resources	Data managers
	Access	access rights to groups or particular users, information about copyrights, trademarks, licenses, sequestered/classified datasets	Data owners/Legal administrators
	Applications	databases/datawarehouses/online repositories where the data can be accessed, and software recommended for viewing or analyzing the associated dataset	Data owners/Data managers
	Quality	reports, indices, and assurances on the completeness, quality, and logical consistency of the metadata	Metadata certification authorities

Table 6.1 One example of allocation of responsibility for metadata documentation for a tree crown reflectance metadataset

Instrument manufacturers could be required to supply a minimum set of instrument related metadata as specified by the proposed core metadataset, and accordingly, the output data files could be certified as compliant with the proposed field spectroscopy metadata standard. Digital camera manufacturers, for example, comply with EXIF standard for photo metadata (Williams, 2012), discussed in more detail in Section 2.3. Data managers (individuals in the role of custodian of data ultimately responsible for its stewardship and quality management within an information sharing system) would be allocated responsibility for more high-level metadata pertaining to the dataset itself. Metadata quality parameters could include input from other experts in the field of research (e.g., ecologists, soil scientists) to provide measures of external consistency with other datasets with metrics for how well concepts or classes in a given dataset relate to another (Comber *et al*., 2007). It would be the responsibility of the data stakeholders (principal investigator, data owners, data managers, etc.) to coordinate the stewardship of the metadata modules. The modules include those presented in the hybrid model discussed in Chapter 4. The example given allocates responsibility to roles and not to individuals (i.e. it is possible in some cases that one individual may be both the principal investigator and the field operator).

6.3.3 Identifying the role of metadata in data users' decision making processes

6.3.3.1 Metadata and fitness-for-use

Understanding how a data user evaluates metadata, and the purpose for which data was collected and used allows data creators within the field spectroscopy community to decide for themselves the extent to which they reasonably need to apply the proposed metadata standard to their datasets. This is a vital step in prioritizing metadata elements that are necessary for a metadataset to be considered compliant, by both data creators and data users for a given application. Consider the multiple cases for which a metadataset may be created and used. In the simplest case, a spectral signature may be collected for use as a reference signature in a data library, with a minimal set of associated metadata documenting the sample name, and wavelengths used, such as the signatures and metadata in the USGS spectral library. A more extensive metadataset would be required for signatures used in end-member retrieval, and validation and calibration activities. Each method of analysis requires different levels of metadata completeness, and accepts different error thresholds. A more formal investigation of the effect of fitness-for-use on expectations for completeness was investigated with two expert panels.

6.3.3.2 Field spectroscopy scientists and metadata quality

A prototype metadata quality analysis tool was created for user feedback from the field spectroscopy scientific panel at the TERN ACEAS ‘Bio-optical data: Best practice and legacy datasets’ workshop in Brisbane in 2012 and at the Geospatial Science Research Symposium Spectral Libraries workshop in Melbourne in 2012. The prototype was proof-of-concept software to demonstrate that metadata quality analytics could be implemented as a user-friendly application to assess metadatasets imported from databases and data libraries. The tool generated a metadata completeness and quality report for individual spectrum metadatasets obtained from SPECCHIO, the DLR Spectral Archive, and the USGS Spectral Library. Some of the data was simulated (error counts, quality schema) in cases where this data was unavailable from the source datasets, or to obtain variability in quality statistics for illustrative purposes. Quality and completeness indices were also generated to give a snapshot assessment of a given metadataset. Core functions of the metadata quality analysis tool are outlined in Table 6.2.

Function	Details
completeness report	provides completeness measures (in percentage compliance) on the core metadataset and database-native metadata
completeness index	derived from the completeness report as a linear combination of populated metadata with larger weighting given to the core metadataset
quality report	provides a list of semantic and syntactic errors (as percentage occurrence in the metadataset) and presence of quality assurance flags
quality measure	derived from the completeness report as a linear combination of weighted quality parameters
categorical ranking (optimal or suboptimal)	* Optimal = presence of quality assurance flag and quality measure ≥95 and completeness index ≥ 75 * Suboptimal = no quality assurance flag or quality measure<95 or completeness index <75
quality schema import	user-adjustable specification where pre-defined quality and completeness thresholds can be imported from other metadata standards
metadata completeness threshold	user-adjustable specification where the completeness threshold can be one of the following: * complete in all parameters * complete only in core metadataset * complete in compliance with a specific standard
quality assurance specification	user-adjustable specification where quality index can include or exclude requirement for a quality assurance flag

Table 6.2 Details of each core function in the prototype metadata quality analysis tool

The tool was demonstrated to both panels, with the intention of determining if the core functions would be useful to the scientists. The second panel (Geospatial

Research Science Symposium Spectral Libraries Workshop in Melbourne, 2012) was asked an additional question: 'if no quality assurance were available for a metadataset, under what conditions (number of syntax and semantic errors, for example) would you accept or reject a metadataset?' A discussion followed. Each participant in the panel was encouraged to provide an opinion on which functions they found useful and why. The group feedback is outlined in Table 6.3.

Function	Panel feedback (is this useful?)
completeness report	yes; statistics referring to specific parameters are useful
completeness index	yes; given that an explanation of how it is derived and the associated completeness report is available
quality report	yes; statistics referring to specific parameters are useful
quality measure	no; users should make this decision for themselves
categorical ranking (optimal or suboptimal)	no; users should make this decision for themselves
quality schema import	yes
metadata completeness threshold	yes
quality assurance specification	yes

Table 6.3 Panel feedback obtained for each of the main functions of the prototype metadata quality analysis tool

Both panels found the majority of the functions useful, particularly the completeness and quality reports which provided a listing of statistics on compliance with specific metadata categories, presence of errors, and presence of quality assurance flags in the metadata. The consensus in the second panel was that data users want to make decisions for themselves in determining whether a given metadataset is useful for their purpose and are therefore reluctant to rely on indices such as a categorical ranking or quality measure for metadata quality. This is particularly relevant in cases where the user may assign a different weighting to the parameters that comprise a quality or completeness index. Panel respondents stated that they would choose to use suboptimal metadata (high number of semantic errors, or high degree of incompleteness for example) if the data is rare or unique and/or necessary for their application – for example, if it is the only dataset in existence meeting their criteria. Respondents emphasized that the reputation of the data creators also played a role in whether they would choose to use data where the metadata may be incomplete. Users may also choose a dataset whose metadataset is incomplete if they were aware that requisite metadata may be obtained retrospectively (such as instrument properties, which a data creator might not load into a spectral database but can provide to other scientists if requested).

6.3.3.3 Metadata and metadata quality as an aid to decision making

Any data discovery tool or quality or completeness assessment must empower a user with enough information about the metadata to make a judgement about whether the dataset is useful to *them* (Section 2.6.2). It is important to consider that the metadata, and any quantitative or qualitative assessments of it must serve to aid, not hinder, the decision making process. It has been previously documented that completeness of datasets can be considered a prerequisite to appropriateness (Sicilia *et al.*, 2005), and this concept can be extended to metadatasets, as evaluated by data users. User-oriented metadata relating to quality measures should provide suitable information to enable users to understand the limitations of analysis for a given dataset, and potentially, linking uncertainty with the quality assessments (Comber *et al*., 2007). Methods of quality assessment of metadatasets are discussed in more detail in Section 2.6.

The use and interpretation of quality information have been shown (Watts *et al*., 2009) to be affected by the type of user, the context of the inquiry, and the amount of information supplied to the data users, among other factors. Supplying a data user with quality metadata along with its associated dataset results in different decisions being made by the data user than when using the underlying information alone.

A study of information systems professionals and their use of data quality information in decision making revealed that the level of expertise, and domain of expertise, had an effect on the degree to which they relied on data quality indicators to make decisions about the information in a dataset (Ballou *et al.*, 2003). For example, experts used data quality indicators more often than non-experts; there was less consensus among expert data users in the choices made based on data quality indicators; those without domain-specific experience made greater use of the data quality indicators, with the suggestion that domain-specific experience may inhibit use of data quality indicators when making decisions; and information systems managers, who decide which datasets are suitable for databases and datawarehouses, benefit most from data quality indicators (Ballou *et al.*, 2003).

The results of these studies suggest that several considerations must be taken into account when creating and provisioning metadata with the flexibility for multiple purposes – these include addressing the needs and expertise of a broad range of data users, and supplying metadata (both the raw metadataset and quality and completeness assessments) in a suitable format for interpretation and usability. Section 6.3.6 discusses quality assurance certificates that can be issued for field spectroscopy metadatasets, in tandem with the raw metadatasets, that include a range of quality indicators designed to accommodate the needs of and non-expert data users across domains.

6.3.4 Building standard-enabling software tools and information systems

6.3.4.1 The need for standard-enabling software and information systems

Building software tools and data sharing systems that support and endorse standard-compliant metadata for field spectroscopy, while meeting users' needs, is an additional effort towards good practice. Studies of scientific data collection management have shown that preferred systems are initiated by scientists respected among their community where domain knowledge is a prerequisite for proper management and documentation of datasets by scientists and researchers (Anderson, 2004). Therefore, building standard-compliant or standard-enabling information systems with direct input from the field spectroscopy community is an enticement for field spectroscopy researchers to produce and publish good metadatasets. This approach also helps to prevent implementation of data sharing systems that are not optimal for metadata creators (Palmer *et al.*, 2007). It is essential too that the metadata be formatted in a manner that is useful and informative for data users.

6.3.4.2 Software tools

The provision of software tools to guide metadata creators through the process of creating high quality metadatasets is a small and relatively simple step towards standard-compliant metadata. For example, USGS has an online metadata validation service that accepts text, XML, and SGML metadata files and returns a report identifying discrepancies with the FGDC metadata standard (USGS, 2013*a*). Metadata editors such as Morpho for EML and Metavist for the Biological Data Profile metadata standard are also available to the public for download (Huettmann, 2009). Similar small-scale applications could be made available online for field spectroscopy research centers and institutes to validate their metadatasets before releasing them for distribution.

In a larger context, distribution portals, data discovery tools and metadata clearinghouses must be sufficiently robust in design to accommodate the needs of multiple users. Experts and non-experts alike must be provisioned with sufficient metadata and information for domain and non-domain-specific applications. Expert field spectroscopy data users (e.g. a scientist with 15 years experience in research), who is searching for data for a specific application, (e.g. spectral signatures for sensor calibration activities) can be predicted to rely to some degree on a quality and completeness report but also bypass it to access the raw metadataset and underlying dataset decide for themselves if they wish to use a dataset. A non-expert user (e.g. an undergraduate student searching for an exemplar leaf reflectance signature for a university assignment) can be predicted in most cases to ignore altogether any quality completeness information and rely only on the most basic metadata parameters to choose a dataset, such as the species of leaf and the em wavelengths. Scientists analysing large numbers of spectra or geospatial information data managers need to be supplied with quality and completeness reports for quality control of distributed datasets, as this may be the only information they use (or have the time to use) on which to base decisions for accepting or rejecting datasets in bulk.

6.3.4.3 Useful and informative metadata formats

A secondary consideration for distributing and sharing metadata is that it is delivered in a suitable format for interpretability and usefulness. Balance must be also be maintained between providing informative summaries and overwhelming the data user with metadata. Volume does not necessarily imply information, and in a poorly designed data discoverability tool, a user may be subject to a tidal wave of 'metacrap' (Doctorow, 2001, p. 1) with limited ability to discern its value or applicability.

The completeness and quality measures provide a summary description of the metadata; in the case where a data user wishes to bypass the completeness measure in their decision making process, they can view the raw metadataset to decide whether it is complete enough or of suitable quality for their purpose. This hybrid approach of supplying the data user with metadata summaries and description (e.g. completeness and quality measures through an automated process) along with the metadataset itself to provide manual assessment, is perhaps the most empowering for the data user because it gives them both a macro and micro view of the metadata. This enables efficient use of the data user's time in the search process by enabling quick decisions to be made with a macro-level metadata assessment, and providing the user with further information to engage in a personalized assessment of completeness at the micro-level if they choose. It has been proposed that visualization methods to exhibit patterns of incompleteness in metadatasets can be a useful tool for a data user to determine whether a metadataset is sufficiently complete in the categories relevant to their purpose for use (Daas *et al*., 2010).

Automated assessments of completeness and quality of metadatasets are likely to be more commonplace with time. This is the case especially in future implementations of larger and more complex information processing and sharing systems (more advanced than the spectral libraries and databases currently in operation) where field spectroscopy datasets can be used in system-automated processes to create 'synthetically-derived' data products (i.e. created with no human intervention) for large-scale distribution.

Metadata that includes quality assessments must also be designed and presented in a manner that is useful to a data user. Within geospatial applications especially, data quality parameters must make it possible for the user to link the data quality statement to the quality of the results potentially derived from a dataset (Frank, 1998). By extension, the same principle of utility can be applied to metadata quality parameters. Quality scores and indices are perhaps the most basic way of presenting quality information. However, scores alone meet limitations and challenges, especially in the case of intercomparing datasets and determining normative quality scores so that quality information can be made understandable for a user (Daas *et al*., 2010).

There are alternative ways of presenting quality information. The field spectroscopy metadata quality assessment tool presented to the expert panels in Section 6.3.3.2 metadata quality offers a template for supplying metadata quality information, but is not necessarily comprehensive in terms of the kind of information that data users may useful. Quality assessment ideally integrates contextual and objective quality assessment processes (Watts *et al*., 2009).

Examples of the types of contextual and non-contextual quality assessments implemented in other disciplines include the LTER Network Information System, which assigns five categories (0-4) to datasets based on how well they meet the needs of specific data products (Michener and Jones, 2012); crowd-sourced record-per-record basis data quality assessment of chemical compounds within ChemSpider (Williams *et al*., 2012); manual, automated, and global quality assessments (using domain-specific expertise) of datasets within OpenTox, an online platform for the discovery and exchange of toxicity data (Fu *et al*., 2011); and assessments of the scientific impact of datasets archived by NASA's ORNL (Oak Ridge National Laboratory) DAAC (Distributed Active Archive Center for Biogeochemical Dynamics) (NASA, 2014). Although not all of these are directly related to metadata quality for field spectroscopy, they are examples of the many ways in which quality information can be expressed and presented to users, as an alternative to relying on indices or qualitative descriptions alone. Designers of future data sharing systems for field spectroscopy datasets and metadata benefit from exploring the range of possibilities for producing useful quality assessments to data users.

6.3.4.4 Standard-supportive information systems

For field spectroscopy scientists to participate in the mainstream IT environment of large-scale data distribution, complete and high-quality metadata (enforced in part by the proposed field spectroscopy metadata standard) is essential. Integrating the field spectroscopy metadata standard within large-scale information systems requires, first of all, addressing data management concepts around which a recommendation can be formulated.

In order for a metadata standard to facilitate the exchange, discoverability, and promote the extended life cycle of a dataset, the IT infrastructure must exist to support it. The IT infrastructure can consist of the data stores, data access services, and the organizations responsible for maintaining them. An IT infrastructure also serves to ensure some quality control over the creation, ownership, and management of metadata. Data stores (including spectral libraries, databases, datawarehouses, and other data repositories within big data platforms) facilitate the discovery and distribution of metadata and its associated datasets. Their potential for making datasets visible to users through search engines and other discovery tools is maximized when metadata-rich, standards-compliant datasets exist for discoverability. IT infrastructure can be defined at the institutional, government, and discipline-specific level (ANDS, 2013). They vary primarily in size of the network, number of support personnel assigned to it, and the degree to which they promote collaboration across institutions (more commonly found within the government and discipline-specific infrastructures).

In order to determine the ways forward for integrating standards-compliant metadatasets within large-scale IT infrastructure, first it is important to examine the data archiving structures currently being used to store field spectroscopy data metadata, and the larger and more complex structures within which they can be stored in the future. Figure 6.1 illustrates the evolution of data archiving structures for field spectroscopy datasets by their data volume and infrastructure complexity.

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

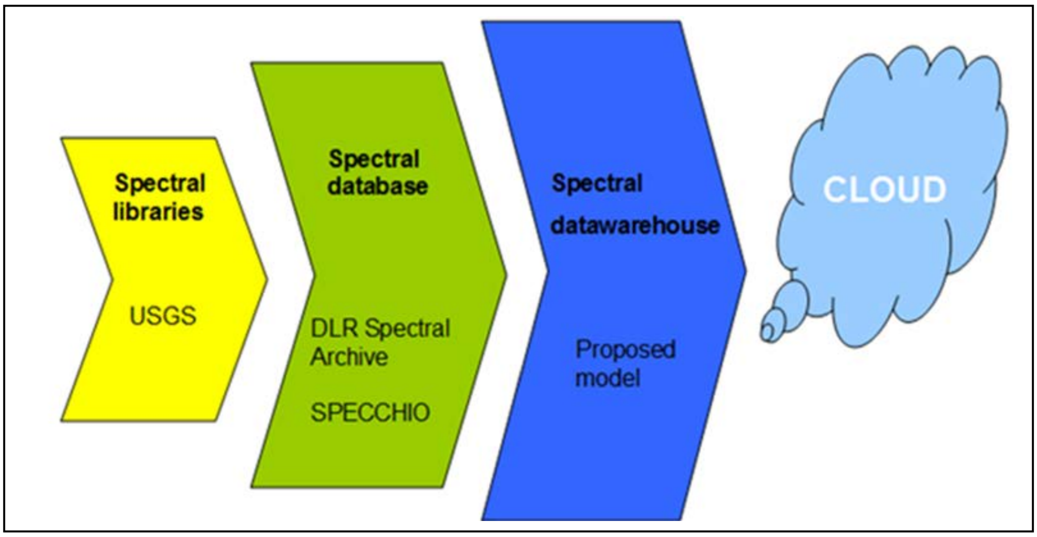


Figure 6.1 Archiving structures (existing and potential) for field spectroscopy data

most static of the data archiving models and do not support the hierarchical dependencies of metadata components for field spectroscopy metadata.

Spectral Databases

In a database data is structured, updated, and edited by database management software and can store almost anything digital including textual information, videos, and images. An example is a geodatabase that stores information such as satellite images and digital elevation data for users to access, analyze, and model through a front-end application (such as a GIS). It can be used by a single user or installed as an enterprise application for many users. Hyperspectral databases created in the last few years include SPECCHIO, the DLR Spectral Archive 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 spectroradiometer file formats (Hueni and Kneubühler, 2010). SPECCHIO provides efficient storing and reporting mechanisms for hyperspectral and field spectroscopy metadata input by its users.

In practical application, large-scale data sharing platforms do not use standalone databases as a direct and single source of data (Ponniah, 2007; Harrington, 2009). This is because large-volume transactions are restricted by the operating system and speed and bandwidth of network connections between the databases. Also, scalability is dependent upon the computer where a given database instance is installed, as well as the operating system and infrastructure resources. Therefore, it is more useful to examine the implications of metadata within more complex data archiving architectures.

Datawarehouses

A datawarehouse 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. Downstream data transactions are affected by the metadata at the data sources. Figure 6.2 illustrates the data flow through a data warehouse.

The datawarehouse is presented here as a proposed model for efficient and quality controlled distribution of large volumes of field spectroscopy data and metadata. A datawarehousing model does not yet exist for field spectroscopy datasets. In the context of field spectroscopy metadata, a datawarehousing model would serve the remote sensing community by providing a central interface for field spectroscopy data and metadata 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, in itself a strong incentive for adoption of common standards for data and metadata exchange.

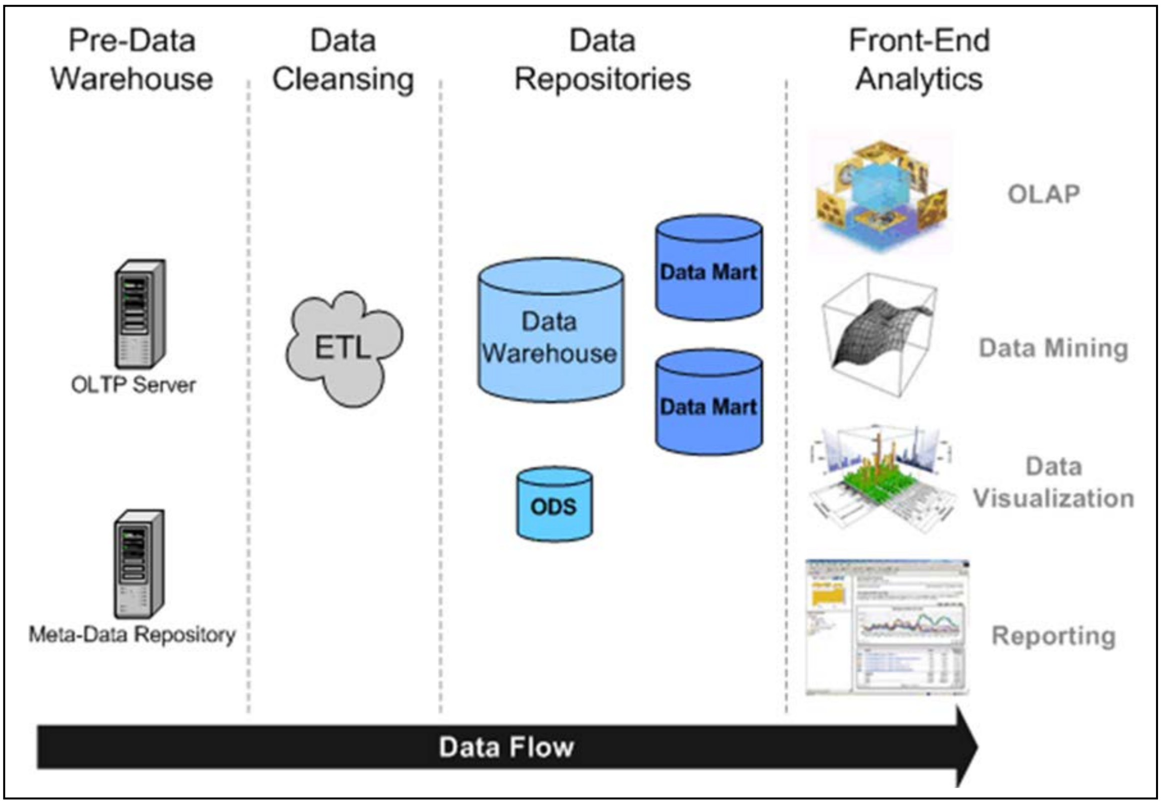


Figure 6.2 Overview of data flow within a datawarehousing infrastructure
Source: (Mailvaganam, 2007)

The cloud

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). Challenges including security and trustworthiness have been indentified for geospatial data users using cloud services, since these are magnified within the cloud environment (Yang *et al*., 2011). Because of limited standardization and consequently, no mechanism for quality assurance for field spectroscopy datasets, cloud computing at this time is not a suitable candidate as a one-size-fits all data sharing solution for field spectroscopy data users. However, the potential for cloud computing for future distribution of field spectroscopy datasets is discussed in more detail in Section 6.4.

Existing scientific data sharing platforms

There are several IT infrastructure models that have been adopted for the sharing and distribution of scientific research in general and for geospatial data specifically. Since members across the field spectroscopy community are increasingly engaging each other on an international platform, government-level and discipline-specific IT infrastructure models are the preferred models to examine here.

For example, NASA's Global Change Master Directory is a public metadata inventory of a broad spectrum of Earth science data and more specifically, authoring tools, data discovery, and metadata transformation and conversion tools in accordance with ISO, FGD, ESRI, Dublin Core, ANZLIC standards for geospatial metadata (NASA, 2013*b*). iVEC (Interactive Virtual Environments Centre) is a joint venture among partners including CSIRO and Australian universities to provide supercomputing and data storage services to researchers across Australia and enable data discoverability through rich metadata cataloguing (iVEC, 2013). NOAA's National Coastal Development Center hosts MerMAid (Metadata Enterprise Resource Management Aid), a platform-independent application that provides an online service to build databases to generate, manage, and publish metadata in accordance within FGDC standards, EML, and MARC (MAchine-Readable Cataloging) (NOAA, 2013*b*). LTER

(Long-Term Ecological Research) is a network of researchers and agencies including USDA Forest Service and Agricultural Research Services, NASA, USGS, the US Environmental Protection Agency, providing public-access ecological data compliant to EML metadata standards (LTER, 2013).

These IT infrastructure models share the following characteristics: a network of distributed data centers incorporating research from multiple disciplines; IT support and administrative personnel with knowledge of common data management protocols; a history of engagement with the public, industry, and research agencies; and necessary funding for the hardware, software, and personnel resources required for long-term data storage and distribution. These models are currently suited to accommodate field spectroscopy metadatasets adhering to the proposed metadata standard. Integrating field spectroscopy metadatasets need not be a challenging task given that the data stakeholders have an understanding of the value of storing and sharing their data on such a platform, and that they have the desire to make their datasets available.

Future integrations of spectral information systems such as the proposed GEOSS (Global Earth Observation System of Systems) (Group on Earth Observations, 2013) would greatly benefit from adherence to international metadata standards, and with this is mind, according standardizations should be pursued with alacrity; existing systems should be upgraded accordingly, once such standards have been defined and/or recognized. With the support of the proposed field spectroscopy metadata standard, these systems could be adapted to incorporate existing spectroscopy databases (SPECCHIO, DLR Spectral Archive) and spectral libraries (USGS Digital Spectral Library). Figure 6.3 illustrates how field spectroscopy metadata would flow through the GEOSS.

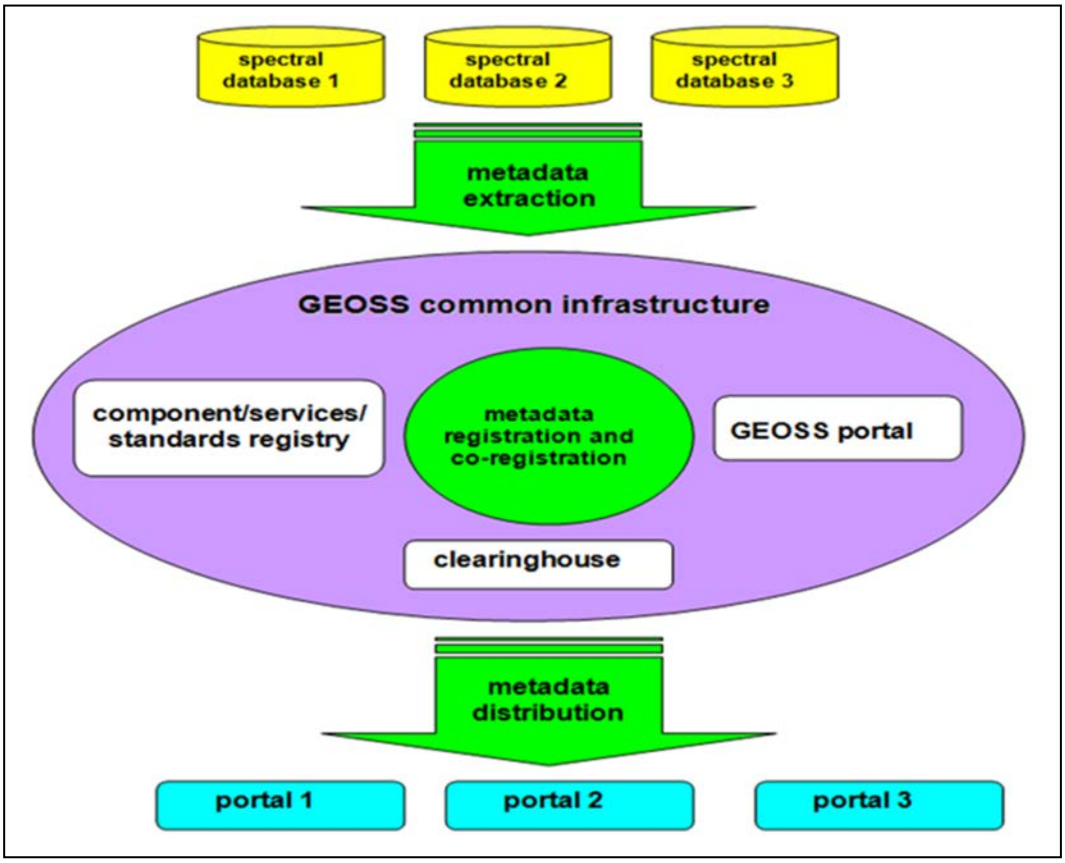


Figure 6.3 An adapted GEOSS data model illustrating the flow of field spectroscopy metadata through the GEOSS data infrastructure integrated with current archives

6.3.5 Educating the community about the importance of metadata

Having established the needs of field spectroscopy scientists for metadata, the importance of metadata needs to be framed within the broader context of large-volume data storage and exchange to assist in community support for a metadata standard and its subsequent adoption.

Promoting support for the standard can be accomplished through educational efforts and promotion at national and international events, similar to those utilized by agencies and organizations within the geospatial community. NOAA's NCDDC (National Coastal Data Development Center) provides outreach programs and online training for individual researchers and organizations to understand how to create and utilize metadata for their datasets (NOAA, 2013*c*). DataONE (Data Observation Network for Earth), an international consortium of geospatial data providers, is engaged in projects for data management training for online data dissemination (DataONE, 2013). iVEC regularly hosts 'Data Clinics', and data management workshops to educate researchers about procedures, tools and practices for sharing data and making it discoverable within the IVEC infrastructure (iVEC, 2013). EDiNA (University of Edinburgh national data centre) facilitates workshops for using their GoGeo geospatial data portal and creating metadata within the ISO 19115, INSPIRE, UK GEMINI 2.1, and UK AGMAP 2.1 guidelines for spatial data (EDiNA, 2013). Similar workshops, in the theme of the previous events discussed in Chapter 4, can be developed for the field spectroscopy community with the involvement of data stakeholders and data producers.

Research in ecoinformatics has revealed that adoption of metadata standards and principles of good informatics practice requires that scientists be made aware of informatics tools, how to use them, and that funding agencies that are stakeholders need to demonstrate a long-term commitment for data repositories through encouragement of data sharing and stewardship (Michener and Jones, 2012). In general, highlighting organizational and individual benefits, providing training, publishing organizational efforts, and building administrative support are ways of changing community culture (Wayne, 2005). Introducing a metadata standard to the field spectroscopy community that encompasses a larger metadataset than researchers are currently using can reasonably be expected to entail some challenges. Following the models and tools of agencies responsible for issuing metadata standards may aid in successful wide-ranging implementation of the proposed metadata standard presented in this research.

6.3.6 Certification of metadatasets

Certification of metadatasets is important in their quality assurance as they are exchanged and disseminated widely throughout datawarehouses and other archiving and data sharing systems. Certification confirms to the data user that a given metadataset adheres (either partially or completely) to the field spectroscopy metadata standard. It also limits duplication of data generation efforts and eliminates risk of legal liability (Joshi and Joshi, 2013). Overall it provides data users with the confidence that a metadataset is reliable and suitable for a certain purpose. Certification can occur on two levels – quality assurance of the metadataset itself, and of the repository managing the metadata. The field spectroscopy metadata standard can exist as a hybrid model (Chapter 4) that incorporates modules (dataset, resource, access information) from existing metadata standards to include information about the source of the metadata (data producers, data owners, data repository), revision history (number and type of edits, who made them, and when), and access information (legal, copyright, security, or privacy rights and restrictions).

Establishing reputational authority of data producers and data repositories is available through automated metrics as discussed in Chapter 5.

There are few formal certifications for geospatial metadata itself, and none for field spectroscopy metadata. There are, however, methods for certifying or providing reports on compliance to standards for geospatial data and software within the geospatial community. OGC provides compliance testing with the mandate to increase system interoperability for database, server, client software and encoding schemas with an annual license fee for certification (OGC, 2012). They provide a listing of all OGC-certified companies and software as well as those that have implemented OGC specifications (but have not been certified) on their website. ESRI's ArcGIS suite of software provides a metadata validator for compliance to XML schemas for standards specified by the user (ESRI, 2013). USGS also has an online metadata validation tool (USGS, 2013*a*) referred to earlier.

Certification for field spectroscopy metadatasets would need to be carried out either by a geospatial advisory body (FGDC, ISO, OGC) that has endorsed the field spectroscopy standard or an agency or a team of recognized experts within the field spectroscopy community with the reputational authority to do so. This could take several forms, whether as a simple quality assurance flag present in the attached metadata, or as a full certificate with information about the certifying body, date of certification, and all standards it complies to. Table 6.4 is a hypothetical example of a quality assurance certificate provided for a metadataset obtained from a spectral database.

Certification Details	Metadataset X
Certificates	Standard: ISO 19XXX Compliance level: fully compliant Certificate expiry: 01-01-201x Certifier: Abcd Efgh Email: Abcdefgh@agency1.com
	Standard: Field Spectroscopy Metadata Standard Compliance level: fully compliant Certificate expiry: 05-05-201x Certifier: Jklm Nopq Email : jklmnopq@agency2.com
Quality Assurance	Assurer: Dr. Spectral Investigator Date of assurance : 01-01-201x Level of internal quality compliance: 2 Institute: Spectral Investigations, Inc. Address: 45 Investigation Suite, Spectral City, Rainbow 8888 Email : spectral.investigator@spectralinvestigations.org
Lineage	Database user: DB User Institute: Spectral Databases, Inc. Address: 25 Spectral Databases Suite 52, Database City, Binary 1001 Email: db.user@spectraldatabases.com
	Metadataset producer: Field Spectroscopy Scientist Institute: Field Spectroscopy University Address: 88 Field Spectroscopy Drive, Spectral City, Rainbow 8888 Email: fieldspectroscopy.scientest@fsu.edu
	Metadataset owner: Spectral Databases, Inc. Address: 44 Spectral Database Drive, Spectralville, Democratic Republic of Signatures, w8w 8w8 Email: equireies@sd.com
	Metadataset creation date: 01-01-20xx
	Revision history: 01-01-20xx Viewing geometry updated 01-01-20xx Database user information updated 01-01-20xx Project information details deleted
Quality Parameters Investigated	Logical consistency, semantic/syntactic errors, reputational authority, completeness
Quality Report	Available for download

Table 6.4 Hypothetical example of a quality assurance certificate for a metadataset downloaded from a spectral database

While not comprehensive in all the quality and completeness parameters that are provided in the field spectroscopy standard (this could be provided in separate reports for the data users), Table 6.4 is illustrative of the most relevant information in a quality assurance certificate. It includes information about organizations that have certified the metadataset, levels of compliance to given standards, quality assurance, lineage information, and quality parameters that were assessed. Such a certificate provides information for the data user to help them decide whether the dataset meets their criteria for a given application, and data managers to decide whether such datasets are of suitable quality for their data archiving and sharing systems. It is vital that in the life cycle of a metadataset, its certification occurs before distribution through clearinghouses and other online portals (i.e. no later than within the domain of datawarehouses) to maintain quality control.

Alternately, following the OGC model, field spectroscopy datawarehouses and archiving systems that choose to implement the proposed standard could themselves be certified, with the implication that all metadatasets they store, generate, and distribute are automatically compliant with the standard. The field spectroscopy metadata standard includes provisions for quality and completeness reporting where automated metrics, implemented by way of algorithms including database crawls, can be used to establish reputational authority of both the data producers and the spectral library itself (discussed in more detail in Chapter 5).

6.3.7 Additional approaches

Additional approaches have been suggested with the aim of encouraging metadata creators both within geospatial science and in other disciplines to adopt metadata standards. These include emphasizing the cost and benefit of metadata creation, and the consequences of using unsuitable metadatasets (Barton *et al*., 2003) and relying on experts (information systems specialists, statisticians, risk analysts) other than geospatial researchers themselves to determine formulas and complex scenarios for strengths, weaknesses, and opportunities in the metadata creation process (Devillers *et al*., 2010).

Implementing a standards-supportive publication process for scholarly research, where documentation of datasets in the literature is coupled with research standards has been proposed for the genomics community (Garrity *et al*., 2008). A standards-supportive publication process should become a more prominent objective with time as increasingly more scientific data originates in digital form, so opportunities must be exploited to leverage the common digital properties of scientific data and information (Anderson, 2004). However, focus should be concentrated initially on those activities that require the most direct engagement with metadata creators and the metadata creation process for promoting a standards-supportive culture. Investing in low-quality metadata that fails to accurately and comprehensively describe its associated dataset has proven fatal to 'mega-science' initiatives due to large maintenance costs, lost potential associated with poor metadata, and the global use of such datasets (Huettmann, 2009), so it is incumbent upon researchers and data distributors to adopt a culture of valuing and creating the best possible metadata.

Information management projects and similar initiatives within geospatial science, and other scientific applications have revealed that good practice is difficult to achieve, but identifying obstacles to good practice is an important first step. Only then can an adoption strategy be implemented that addresses the needs of field spectroscopy community scientists and provides the tools, resources, and IT infrastructure possible for the whole community to participate. To enable good metadata practice, the field spectroscopy community should focus initially on the recommendations summarized in Table 6.5.

	RECOMMENDATION	RATIONALE/EXAMPLES	PRIORITY
1	Metadata prioritization p. 168	Prioritizing metadata that is a) essential (core metadataset, critical elements of the application-specific metadatasets) b) and/or can be documented concurrently with field data collection (e.g. viewing geometry) versus metadata that can be documented retrospectively for achieving the most complete metadatasets possible	Primary
2	Collaborative stewardship of metadata pp. 169-171	Assigning of responsibility of creating and maintaining metadata to multiple individuals and stakeholder (researchers, IT specialists, data managers) according to their domain of expertise	Primary
3	Identifying a purpose for metadata collection and use pp. 171-177	Allows metadata creators the flexibility to set thresholds for quality and completeness within domain and purpose-specific contexts	Primary
4	Standards-compliant software tools and information systems pp. 177-179	Data sharing systems and metadata editors that enable and enforce creation and distribution of metadatasets compliant with the field spectroscopy metadata standard	Primary
5	Metadata completeness and quality assessments pp. 179-182	Metadata completeness and quality reports provisioned with datasets to aid decision making for data users; a minimum of completeness metrics for the field spectroscopy core metadataset is required	Primary
6	IT infrastructure and management pp. 182-189	Data distribution system that provision quality-controlled discoverability and distribution of field spectroscopy metadatasets	Primary
7	Education initiatives pp. 189-191	Workshops and training programs for researchers and field spectroscopy data stakeholders	Primary
8	Metadata certification pp. 191-194	Assigns a level of quality assurance to metadatasets with a range of quality indicators to accommodate a range of data users	Primary
9	Cost-benefit analyses p. 194	Demonstrates to the field spectroscopy community the impact, benefits, and losses associated with variable quality metadata	Secondary
10	Standards-supportive publication protocols p. 195	Couples research publications with metadata-compliant datasets; promotes improved metadata capture and completeness of metadatasets, traceability of data-derived results, and evaluation of scientific impact of datasets	Secondary

Table 6.5 Strategy for adopting, implementing, and integrating a metadata standard for the field spectroscopy community

Priority designations provide guidance on those areas that require preliminary focus, and have been detailed in the preceding sections. They are a synthesis of inputs derived from the field spectroscopy expert panel workshops and from studies in other disciplines elaborated upon earlier.

6.4 Current and future opportunities for field spectroscopy datasets

6.4.1 Big data and the cloud

All IT infrastructures will at some point need to prepare for the challenges of operating as a 'big data' environment. A relatively new data conceptualization, big data is characterized by its volume (e.g. banking transactions for national financial institutions, traffic flow sensor data), velocity (generated quickly over short windows of time or continuously, such as GPS tracks), and variability (e.g. text, images, raw feeds from satellite-based sensors) (Dumbill, 2012). Big data is expected to continue to increase in all three of these dimensions. In 2005, it was estimated that the global digital data inventory was 130 exabytes; in 2010, 1,277 exabytes, and in 2015, it is predicted to be 7,910 exabytes ('No end in sight', 2011). Recognizing and addressing these trends and challenges is an opportunity for field spectroscopy scientists to take an active role in future-proofing their datasets and ensuring that their data will be distributed in a quality-controlled manner.

Traditional data infrastructures are ill suited to handle the storage and processing of big data. Cloud computing has been proposed as a suitable architecture because it does not rely on a single party or organization to fund and maintain the infrastructure (software and hardware). The option of sharing resources is an attractive one considering that in 2020, the number of servers (virtual and physical) around the world is predicted to increase 10-fold, the amount of data management by data centers will increase 50-fold, and the number of files they will have to process will increase at least 75-fold, and almost 20% of information will be 'touched' by the cloud by 2015 somewhere in a byte's journey from the data creator to its destination ('No end in sight', 2011).

Implications for maintaining integrity of metadata are magnified in the big data environment. For example, it is possible to increase the value of searches and rapid data retrieval for scientific data discovery by bundling original datasets and their associated publications in the search results; however, intuitive, user-centric interfaces must be developed to resolve semantic ambiguities between disciplines to facilitate this kind of discovery (Tolle *et al*., 2011). This breed of intelligent searching is only possible with metadata-rich datasets, standardized metadata that is interoperable on a broad scale, and with platforms and search engines that facilitate visibility to source data repositories. Provenance metadata plays a crucial role in tracing the evolution of a datasets in big data environments (Buneman, 2013), and is of special significance in research applications where it is necessary to know the source of a dataset, who created it, and any changes that have been made to it. The impetus for new and evolving metadata standards to meet these challenges grows stronger with the proliferation of datasets in the public sphere and the demands by data users to access them.

6.4.2 Towards an integrated model

There exist data exchange networks among the geospatial community that are evolving towards an integrated datawarehousing, cloud-based, big data model. Among these are:

- EOSDIS (Earth Observing System Data Information System), a network of data centers, metadata repositories, 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
- EUFAR (European Facility for Airborne Research), a transnational initiative to create databases and streamlined data exchange standards for airborne hyperspectral research (EUFAR, 2009)

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, including field spectroscopy scientists. Leveraging the capabilities of these systems enables field spectroscopy scientists to share their datasets with a wide audience of users in a quality controlled environment. However, it is incumbent upon the field spectroscopy community to actively participate in the design and implementation of such systems, which includes supporting the proposed field spectroscopy metadata standard for maximizing the discoverability and quality assurance of their datasets.

6.5 Conclusions

Field spectroscopy scientists, and lessons learned from other disciplines, have provided valuable insight into how to proceed with the adoption, implementation, and integration of a field spectroscopy metadata standard. Community adoption of new standards in other disciplines has proven difficult. A simple approach to good practice (not perfect practice) is best, beginning with the recognition that obstacles exist, and will persist. However, many of these obstacles can be overcome by adopting the strategy for community adoption presented in this chapter.

Prioritizing metadata that can be documented concurrently with field data collection balances the need for complete metadatasets with the constraints of time and conditions under which the measurements are being taken, and is an important step towards achieving the most complete metadatasets possible. Allocating stewardship of metadata to multiple parties reduces individual workload for metadata documentation and ensures that distinct categories of metadata (calibration, field protocol, security and access rights, project information) are the responsibility of those with the relevant expertise.

Establishing a purpose for metadata collection and use allows metadata creators the flexibility to set thresholds for quality and completeness within domain and purpose-specific contexts. Metadata completeness and quality reports provisioned with datasets aids decision making for data users across domains and with varying levels of expertise. A minimum requirement of completeness metrics for the field spectroscopy core metadataset provisioned with each dataset is recommended.

Educational efforts (workshops, training programs) and promotion of the field spectroscopy metadata standard at national and international events helps research and data stakeholders to understand the value and specific activities of good metadata practice. Certification confirms to the data user that a given metadataset adheres (either partially or completely) to the proposed field spectroscopy metadata standard. Certification of metadatasets is important in their quality assurance as they are exchanged and disseminated throughout datawarehouses and other archiving and data sharing systems.

In order to facilitate the exchange, discoverability, and life cycle of dataset and their associated metadata, the IT infrastructure must exist to support it. The IT infrastructure can consist of the data stores, data access services, and the organizations responsible for maintaining them, and serves to ensure some quality control over the creation, ownership, and management of metadata.

Chapter 7 Conclusions

7.1 Introduction

This thesis has proposed the core components of a metadata standard for field spectroscopy. The metadata standard was built through engagement with subject matter experts and aims to increase the discoverability, reliability, quality, and life cycle of field spectroscopy datasets for wide-scale data exchange. The main components of the metadata standard are a core metadataset for all applications, an extended metadataset for specific applications, and additional modules imported from existing standards to enhance robustness and interoperability. Weaknesses in existing metadata standards both within geospatial science and related disciplines were examined, and metrics tailored for analysing field spectroscopy metadata quality and completeness parameters were presented, both at the level of individual records and at the level of a spectral library as a whole. Recommendations focused on overcoming obstacles to a formal adoption of the standard by the field spectroscopy community and steps forward for its integration into data warehouses and big data platforms.

This chapter presents an overview of the results of research questions in the preceding chapters, and the recommendations for adoption and implementation of a field spectroscopy metadata standard. As a result of conducting the research, the following outcomes were produced: i) a proposed core metadataset for all field spectroscopy applications (Chapter 3); ii) an extended metadataset for three specific applications (Chapter 4); iii) a hybrid metadata standard incorporating modules from existing modules for increased robustness (Chapter 4); iv) methods and metrics for evaluating metadata completeness and quality in spectral libraries (Chapter 5); and, v) recommendations for adoption and integration of the proposed metadata standard (Chapter 6).

7.1.1 Research question 1: What are the key elements of a core metadataset for all field spectroscopy applications?

An international panel of experts was surveyed for their opinion on the metadata that must be documented for field campaigns to ensure that all the information for maximizing the integrity of the dataset and ensuring legacy potential for long-term sharing and interoperability with other datasets is captured. The survey respondents helped to identify a core metadataset critical to all field spectroscopy applications, as well as recommend additional metadata to increase the versatility of a metadataset, both for application-specific metadata and generic campaign metadata.

The survey established that a core metadataset must include 'Viewing Geometry', 'Location Information', 'General Target and Sampling Information', 'Illumination Information', 'Instrument', 'Reference Standards', 'Calibration', 'Hyperspectral Signal Properties', 'Atmospheric Conditions', and 'General Project Information' and at least one application-specific metadata category, depending on the type of target being sampled. The inclusion of additional categories, relating to both generic and application-specific metadata, serve to enhance the robustness of the dataset. The composition of each category is a factor of those metadata fields that were clearly identified as critical (through binomial analysis in the 'Calibration' category, for example) and those that are more difficult to designate. Overall, the results from the binomial and scale measurement testing prompt two important questions: i) whose opinion among the experts can be used as a basis for designating a metadata field as critical, and supported by what rationale?; ii) Is fitness-for-purpose an additional dynamic that must be accounted for when designing a metadata standard? This was discussed in more detail in Section 6.3.3, which addressed the importance of identifying a clear purpose for metadata collection.

Consensus was highest among experts within the same field, and within categories most closely related to their area of knowledge. This was illustrated by expert groups such as marine scientists who showed lower variance in response and higher overall criticality rankings in the ‘Marine and Estuarine Environmental Conditions’ metadata category than did their non-marine counterparts in the same category. The trend for consensus amongst all categories, measured using the intraclass correlation coefficient, demonstrated that application-specific metadata with smaller but more specialized groups of experts have the highest level of agreement between respondents on the criticality rankings for each field.

The survey results and subsequent analysis provided answers to the problem of identifying critical field spectroscopy metadata with the following data: i) metadata categories that have the highest overall criticality rankings; ii) metadata fields that can be easily identified as critical to all campaigns; iii) metadata fields that are identified ‘critical’/’useful’/’legacy potential’/’NA’ most frequently; iv) the impact of group membership on determination of what is critical in a given metadata category;

v) consensus trends among groups in both generic and application-specific metadata categories.

7.1.2 Research question 2: Is additional metadata required for specific field spectroscopy applications and to support interoperability with other metadata standards?

Three user communities within field spectroscopy were identified and interviewed to help identify key metadata for three target applications – tree crown, soil, and underwater coral. Three metadatasets were created, with descriptions and rationale for each metadata element, optionality rankings, and preferred data formats. Consensus within the tree crown group was lowest on which metadata should be included in their metadataset, based on the argument that knowledge of what the dataset will be used for determines the metadata elements that are required. It was established that some parameters are difficult to obtain *in situ* and can only be populated retrospectively. It was also established that campaigns for each target application have unique logistics and considerations for carrying out spectral measurement, as illustrated best with the underwater coral targets, which are carried out under conditions and in environments exceptional to marine campaigns. Metadata requirements were presented for three application domains: tree crown, soil, and underwater coral reflectance studies.

Seven metadata standards, selected as being representative of standards within geospatial science and information science were examined for their suitability in accommodating the proposed field spectroscopy metadatasets. These were the Dublin Core 1.1, Access to Biological Collections Data Schema 2.06, Ecological Metadata Language 2.1.1, Darwin Core, Content Standard for Digital GeoSpatial Metadata (Remote Sensing Extension), Content Standard for Digital GeoSpatial Metadata (Shoreline Metadata Profile) and ANZLIC Metadata Profile 1.1 (Geographic dataset core) standards. The results show they consistently fail to accommodate the needs of both field spectroscopy scientists in general as well as the three user communities (tree crown, soil, marine). Mappings of metadata fields from each standard to the field spectroscopy metadatasets were, on average, 22% of the core metadataset, 31% tree crown, 3% soil, and 19% of the coral target metadatasets. Flexibility analysis revealed that the less prescriptive or explicit an existing standard is, the more likely it is to capture a larger amount of information in the field spectroscopy metadatasets. Additional modules and parameters from these standards were proposed for inclusion in a field spectroscopy metadata standard for increased robustness.

By building upon the knowledge of scientists in ecology, marine science, the physical sciences and data governance experts who helped to develop existing geospatial standards, a hybrid standard was proposed. Elements describing and documenting the dataset, resources, access, applications, data quality, citations, and protocols can enrich a field spectroscopy standard and make it adaptable to multiple data infrastructures.

7.1.3 Research question 3: What are the criteria for measuring the quality and completeness of field spectroscopy metadata in a spectral archive?

Conventional methods for measuring quality and completeness of metadata were scrutinized against the special requirements of field spectroscopy datasets. Two spectral libraries and their metadata policies were evaluated as test cases for their compliance with the needs of field spectroscopy scientists.

Metadata quality and completeness measures for field spectroscopy can be defined by multiple parameters and using a range of metrics. In order to be useful for data mining, they must be informative for users who will make decisions on the fitness of the data for their purpose. Field spectroscopy metadata completeness can be defined as a two-fold measure consisting of a) compliance with the core metadataset and application-specific metadata presented in Chapters 3 and 4; and b) compliance with the standards of the data infrastructure in which they are stored. Metadata quality for field spectroscopy metadata can be defined in terms of (but not limited to) logical consistency, lineage, semantic and syntactic error rates, compliance with a quality standard, quality assurance by a recognized authority, and reputational authority of the data owners/data creators.

Publicly available datasets are underperforming on these quality and completeness measures. The two test cases examined, SPECCHIO and USGS Spectral Library, have neither quality assurance metadata, nor do they comply to any considerable degree with the core metadataset (SPECCHIO at 18% and USGS at 7.7%). Lineage metadata is consistently minimal or absent for both libraries, and an examination of the USGS

Spectral Library revealed logical inconsistencies in the metadata being populated by the users, as well as semantic and syntactic errors. Reputational authority can be established in SPECCHIO using completeness measures by user and institute.

7.1.4 Research question 4: What are the issues related to adoption of the proposed field spectroscopy metadata standard?

A synthesis of results from this research, field spectroscopy scientists and lessons learned from other disciplines have provided valuable input on how to proceed with the adoption, implementation, and integration of a field spectroscopy metadata standard. Recommendations are divided into two main sections: community adoption of the standard and integration of standardized metadatasets into data sharing platforms. Primary steps forward for promoting good metadata practice among field spectroscopy scientists include approaches to prioritization of metadata, collaborative stewardship of metadata, quality assurance, identifying a purpose for metadata collection and use, metadata completeness and quality assessments, education initiatives, and building IT infrastructure to enable distribution of standard-supportive datasets.

7.2 Final words

Much potential exists for adapting and improving current geospatial data exchange environments for the unique requirements of the field spectroscopy community. Before widespread adoption can proceed, user needs for quality assurance must be formally recognized, and a standard adopted by the field spectroscopy community and geospatial data advisory bodies and data management agencies. A collaborative and innovative spirit can bring great benefits to international efforts for providing the data sharing capabilities and quality control for the field spectroscopy community. The importance of creating a metadata standard can be summarized by participant feedback from the field spectroscopy metadata survey, “Congratulations for your effort in this work ... It is of great interest to find out about commonalities and create a minimum standard set of metadata for all occasions” (Rasaiah, 2011).

References

ABCD Task Group (2013). *An Introduction to the ABCD Schema v2.0*. Retrieved September 03, 2013 from http://wiki.tdwg.org/twiki/bin/view/ABCD/AbcdIntroduction

Agency for Healthcare Research and Quality (2013). *USHIK: United States Health Information Knowledgebase.* Retrieved January 08, 2014 from http://ushik.ahrq.gov/mdr/portals

Anderson, W. L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal*, 3, 191-202.

ANDS (n.d.). *Metadata Guide*. Retrieved August 27, 2013 from http://www.ands.org.au/guides/metadata-working.html

ANDS (2011). *Metadata*. Retrieved January 07, 2014 from http://ands.org.au/guides/metadata-awareness.html

ANDS (2014). Data Management Planning. Retrieved May 26, 2014 from http://ands.org.au/resource/data-management-planning.html

ANZLIC (2007). *ANZLIC Metadata Profile Version 1.1*. Retrieved August 27, 2013 from http://spatial.gov.au/sites/default/files/legacy/osdm.gov.au/Metadata/ANZLIC%2BMetadata%2BProfile/default.html

Arafat, S.; Aboelghar M.; Ahmed, E. (2013). Crop Discrimination Using Field Hyper Spectral Remotely Sensed Data. *Advances in Remote Sensing*, 2, 63-70.

ASD (n.d.). *Handheld 2* [Image]. Retrieved January 03, 2014 from http://www.asdi.com/products/fieldspec-spectroradiometers/handheld-2-portable-spectroradiometer

ASD (2012). *ASD Support Central*. Retrieved January 05, 2013 from http://support.asdi.com/Document/Documents.aspx

ASD (2013). *Field Spec Hi-Res Spectroradiometer.* Retrieved January 05, 2014 from http://www.asdi.com/products/fieldspec-spectroradiometers/fieldspec-4-hi-res

ASD (2013). *FAQ*. Retrieved January 03, 2014 at http://www.asdi.com/resource-center/faqs

ASD (2013). *Prospere Using the Contact Probe with Leaf Clip* [Image]. Retrieved February 20, 2014 from http://discover.asdi.com/bid/93530/Establishing-Spectral-Libraries-of-Wetland-Vegetation-in-Jamaica-A-Goetz-Recipient-s-Efforts

Asner, G. P.; Martin, R. E.; Knapp, D. E.; Tupayachi, R.; Anderson, C.; Carranza, L.; Weiss, P. (2011). Spectroscopy of canopy chemicals in humid tropical forests. *Remote Sensing of Environment*, 115, 3587-3598.

Bagley, S. (2013). *Philip R. Bagley (1927-2011).* Retrieved January 07, 2014 from http://stevenbagley.net/blog/philip-r-bagley.html.

Barton, J.; Currier, S.; Hey, J. (2003). Building quality assurance into metadata creation: an analysis based on the learning objects and e-prints communities of practice. *Proceedings of the 2003 Dublin Core Conference*, September 28 October 2, 2003 in Seattle, USA.

Becvar, M.; Hirner, A.; Heiden, U. (2006) *DLR Spectral Archive* [Database]. Retrieved January 1, 2012 from http://cocoon.caf.dlr.de/intro_en.html

Bhatti, A. M.; Rudnquist, D.; Schalles J., Ramirez, L.; Nasu, S. (2009). A comparison between above-water surface and subsurface spectral reflectances collected over inland waters. *Geocarto International,* 24, 133-14.

Biophysical Remote Sensing Group (2011). *Development of Underwater Spectrometer 2* [Image]. Retrieved February 21, 2014 from http://www.gpem.uq.edu.au/brg-research-roelfsemaetal6.

Bland, J.M.; Altman, D.G. (1997). Statistics Notes: Cronbach's alpha. *BMJ*, 314:572.

Bojinski, S.; Schaepman, M.; Schlaepfer, D. (2003). SPECCHIO: a spectrum database for remote sensing applications. *Computers & Geosciences*, 29, 27-38.

Brownfield, G.; Oliver, R. (2003). Factors influencing the discovery and reusability of digital resources for teaching and learning. *Proceedings of the Annual Conferences of the Australasian Society for Computers in Learning in Tertiary Education*, December 7-10, 2003 in Adelaide, Australia.

Bruce, T.R.; Hillmann, D.I. (2004). The Continuum of Metadata Quality: Defining, Expressing, Exploiting. In Hillmann D. & Westbrooks, E. (Eds.) *Metadata in Practice*. Retrieved from eCommons@Cornell.

Buneman, P. (2013). The Providence of Provenance. In G. Gottlob et al.,(Eds.), *Big Data: Lecture Notes in Computer Science* (pp 7-12). Retrieved from SpringerLink.

Carnegie Spectranomics (2013). *Spectranomics Database* [Database]. Retrieved July 31, 2013 from http://spectranomics.stanford.edu/Spectranomics_Database

Carnegie Institute for Science (2010). *Spectranomics Protocol: Leaf Spectroscopy (350-2500nm)*. Retrieved January 06, 2014 from http://spectranomics.stanford.edu/technical_information.attachment/103

Centre for Spatial Environmental Research (CSER) (2012). *Centre for Spatial Environmental Research*. Retrieved January 17, 2014 from http://www.gpem.uq.edu.au/cser

Centre for Spatial Environmental Research (CSER) (2006). *Underwater Spectrometer System 2006 (UWSS04).* Retrieved January 06, 2014 from http://ww2.gpem.uq.edu.au/CRSSIS/publications/UW%20Spec%20Manual%2029August06.pdf

Committee on Earth Observing Satellites (CEOS) (2013). CEOS Strategic Guidance Version: November 2013. Retrieved June 07, 2014 from http://www.ceos.org/images/CSS/CEOS_Strategic_Guidance_Nov_2013.pdf

Channel Systems (2010). *AISA Airborne Hyperspectral Systems* [Image]. Retrieved February 19, 2014 from http://www.channelsystems.ca/SpectralImaging-AISA.cfm

Chao, T.C. (2012) Exploring the Rhythms of Scientific Data Use. *Proceedings of the 2012 iConference*, February 7-10, 2012 in Toronto, Canada.

Cheng, Y. B.; Middleton, E. M.; Zhang, Q.; Corp, L. A.; Dandois, J.; Kustas, W. P. (2012). The photochemical reflectance index from directional cornfield reflectances: Observations and simulations. *Remote Sensing of Environment*, 124, 444-453.

Ciganda, V. S.; Gitelson, A. A.; Schepers, J. (2012). How deep does a remote sensor sense? Expression of chlorophyll content in a maize canopy. *Remote Sensing of Environment*, 126, 240-247.

Comber, A. J.; Fisher, P. F.; Wadsworth, R. A. (2007). User-focused metadata for spatial data, geographical information and data quality assessments. *Proceedings* of the *10th AGILE International Conference on Geographic Information Science*, May 8-11, 2007 in Aalborg, Denmark.

Cover, T.M.; Thomas, J.A. (1991). *Elements of Information Theory* (pp 112-13.). Retrieved from Wiley Online Library Online Books.

CSIRO Land and Water (n.d.). *A Field Guide for Spectral Measurements of Australian Shallow Benthic Habitats*. Retrieved January 03, 2014 from http://www.ozcoasts.gov.au/nrm_rpt/pdf/Guide.pdf

Currier, S. ; Barton, J. ; O'Beirne, R.; Ryan, B. (2004). Quality assurance for digital learning object repositories: issues for the metadata creation process. *Research in Learning Technology*, 12, 5-20.

Daas, P. J.; Ossen, S. J.; Tennekes, M. (2010). Determination of administrative data quality: Recent results and new developments. *Proceedings of the European Conference on Quality in Official Statistics*, May 4-6, 2010 in Helsinki, Finland.

da Cruz, S. M. S.; Paulino, C. E.; de Oliveira, D.; Campos, M. L. M.; Mattoso, M. (2011). Capturing distributed provenance metadata from cloud-based scientific workflows. *Journal of Information and Data Management*, 2, 43-50.

DAMA (2010). *Vision, Mission and Goals*. Retrieved May 26, 2014 from http://www.dama.org/i4a/pages/index.cfm?pageid=3369

Dangel, S.; Verstraete, M. M.; Schopfer, J.; Kneubuhler, M.; Schaepman, M.; Itten, K. I. (2005). Toward a direct comparison of field and laboratory goniometer measurements. *Geoscience and Remote Sensing, IEEE Transactions on*, 43, 2666-2675.

Darwin Core Task Group (2013). *Darwin Core*. Retrieved September 01, 2013 from http://rs.tdwg.org/dwc/

DataONE (n.d.). *Community Education and Engagement*. Retrieved December 30, 2013 from http://www.dataone.org/working_groups/community-education-and-engagement

Davenhall, C. (n.d.). *Scientific Metadata*. Retrieved January 07, 2014 from http://www.dcc.ac.uk/resources/curation-reference-manual/chapters-production/scientific-metadata

Davies, J., Harris, S., Crichton, C., Shukla, A., & Gibbons, J. (2008). Metadata standards for semantic interoperability in electronic government. *Proceedings of the 2nd international conference on Theory and practice of electronic governance*, December 1-4, 2008 in Cairo, Egypt.

DCMI (2013). *DCMI Home: Dublin Core Metadata Initiative*. Retrieved September 1, 2013 from http://dublincore.org/

Dekker, A.; Brando, V.; Anstee, J.; Botha, H.; Park, Y.J.; Daniel, P.; Malthus, T.; Phinn, S.; Roelfsema, C.; Leiper, I.; Fyfe, S. (2010). A comparison of spectral measurement methods for substratum and benthic features in seagrass and coral reef environments. *Proceedings of the Art, Science and Applications of Reflectance Spectroscopy Symposium*, February 23-25, 2010 in Boulder, USA.

Devillers, R.; Stein, A.; Bédard, Y.; Chrisman, N.; Fisher, P.; Shi, W. (2010). Thirty years of research on spatial data quality: achievements, failures, and opportunities. *Transactions in GIS*, 14, 387-400.

Digital Curation Centre (2010). Template for a Data Management Plan v 1.2. Retrieved May 26, 2014 from http://www.dcc.ac.uk/sites/default/files/documents/tools/dmpOnline/DMP_template_v1.2_100106.doc

Doctorow, C. (2001). *Metacrap: Putting the torch to the seven straw-men of the meta-utopia*. Retrieved March 5, 2014 from http://www.well.com/~doctorow/metacrap.htm

Domenico, B.; Caron, J.; Davis, E.; Nativi, S.; Bigagli, L. (2006). GALEON: Standards-based Web Services for Interoperability among Earth Sciences Data Systems. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, July 31- August 04, 2006 in Denver, USA.

Duggin, M. J. (1985). Factors limiting the discrimination and quantification of terrestrial features using remotely sensed radiance. *International Journal of Remote Sensing*, 6, 3-27.

Dumbill, E. (2012). *Planning for Big Data.* Retrieved from Safari Books Online.

Dumont, M.; Brissaud, O.; Picard, G.; Schmitt, B.; Gallet, J. C.; Arnaud, Y. (2010). High-accuracy measurements of snow Bidirectional Reflectance Distribution Function at visible and NIR wavelengths–comparison with modelling results. *Atmospheric Chemistry and Physics*, 10, 2507-2520.

Duval, E.; Hodgins, W.; Sutton, S.; Weibel, S.L. (2002). Metadata Principles and Practicalities. *D-Lib Magazine*, 8, 16.

Edina (2013). *GoGeo Metadata Workshops*. Retrieved December 30, 2013 from http://www.gogeo.ac.uk/gogeo/metadata/wshop.htm

ESA (2013). *What is remote sensing?* Retrieved January 03, 2014 from http://www.esa.int/SPECIALS/Eduspace_EN/SEMF9R3Z2OF_0.html

ESRI (2013). *ArcGIS Help 10.2 --Support for ISO metadata standards in ArcGIS for Desktop*. Retrieved December 31, 2013 from http://resources.arcgis.com/en/help/main/10.2/index.html#//003t00000039000000

FGDC (2001). *Shoreline Metadata Profile of the Content Standards for Digital Geospatial Metadata.* Retrieved August 24, 2013 from http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/shoreline-metadata/sp_endorsed.pdf

FGDC (2002). *Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata*. Retrieved August 24, 2013 from http://www.fgdc.gov/standards/projects/FGDC-standards-projects/csdgm_rs_ex/MetadataRemoteSensingExtens.pdf

Fisher, C. W.; Chengalur-Smith, I.; Ballou, D. P. (2003). The impact of experience and time on the use of data quality information in decision making. *Information Systems Research*, 14, 170-188.

Fogwill, F. (2005). *Field Guide for the GER3700, Version 3.0*. Retrieved January 03, 2014 from http://fsf.nerc.ac.uk/resources/guides/pdf_guides/3700_guide_v3_rocky.pdf

Frank, A. U. (1998). Metamodels for data quality description. In Goodchild, M. & Jeansoulin, M. (Eds.) *Data quality in Geographic Information: From error to uncertainty*, (pp 15-30). Paris: Hermes.

Fu, X.; Wojak, A.; Neagu, D.; Ridley, M.; Travis, K. (2011). Data governance in predictive toxicology: A review. *Journal of cheminformatics*, 3, 1-16.

Gamji, V. (2011). *The Good, The Bad & The Ugly of Metadata Management*. Retrieved January 07, 2014 from http://sqlscape.wordpress.com/2011/03/15/the-good-the-bad-the-ugly-of-metadata-management/

Garrity, G. M.; Field, D., Kyrpides, N.; Hirschman, L.; Sansone, S. A.; Angiuoli, S.; White, O. (2008). Toward a standards-compliant genomic and metagenomic publication record. *OMICS A Journal of Integrative Biology*, 12, 157-160.

Getty Research Institute (2006). *The CDWA and Other Metadata Standards*. Retrieved January 07, 2014 from http://www.getty.edu/research/publications/electronic_publications/cdwa/moreinfo.html

Gezler, R.D. (2008). Metadata, Law, and the Real World: Slowly, the Three Are Merging. *Journal of AHIMA* 79, 56-57, 64.

Gliem, J. A.; Gliem, R. R. (2003). Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. *Proceedings of the Midwest Research-to-Practice Conference in Adult, Continuing, and Community Education,* October 8-10 in Columbus, USA.

Glover, D. M.; Chandler, C. L.; Doney, S. C.; Buesseler, K. O.; Heimerdinger, G.; Bishop, J. K. B.; Flierl, G. R. (2006). The US JGOFS data management experience. *Deep Sea Research Part II: Topical Studies in Oceanography*, 53, 793-802.

González, C.; Mozos, D.; Resano, J.; Plaza, A. (2012). FPGA implementation of the N-FINDR algorithm for remotely sensed hyperspectral image analysis. *Geoscience and Remote Sensing, IEEE Transactions on*, 50, 374-388.

Goodchild, M. F. (2007). Beyond metadata: Towards user-centric description of data quality. *Proceedings, Spatial Data Quality 2007 International Symposium on Spatial Data Quality,* June 13-15 in Enschede,The Netherlands.

Goovaerts, M.; Leinders, D. (2012). Metadata quality evaluation of a repository based on a sample technique. In *Metadata and Semantics Research (pp 181-89)*. Berlin and Heidelberg: Springer.

Green, J. L.; Hastings, A.; Arzberger, P.; Ayala, F. J.; Cottingham, K. L.; Cuddington, K.; Neubert, M. (2005). Complexity in ecology and conservation: mathematical, statistical, and computational challenges. *BioScience*, 55, 501-510.

Green, M. (2009). *Metadata implementation with Ab Initio EME*. Retrieved January 07, 2014 from http://developer.teradata.com/tools/articles/metadata-implementation-with-ab-initio-eme

Green, R. O. (2010). The Use of a Portable Spectrometer in Support of the Calibration of AVIRIS, the Moon Mineralogy Mapper and other High Uniformity Imaging Spectrometers. *Proceedings of the Art, Science and Applications of Reflectance Spectroscopy Symposium*, February 23-25, 2010 in Boulder, USA.

Greenberg, J. (2012). *NISO/DCMI Webinar* [Presentation slides]. Retrieved January 07, 2014 from http://www.slideshare.net/BaltimoreNISO/metadata-for-managing-scientific-research-data

Grossman, R. L.; Greenway, M.; Heath, A. P.; Powell, R.; Suarez, R. D.; Wells, W.; Mambretti, J. J. (2012). The Design of a Community Science Cloud: The Open Science Data Cloud Perspective. In *Proceedings of the High Performance Computing, Networking, Storage and Analysis (SCC*), November 10-16, 2012 in Salt Lake City, USA.

Group on Earth Observations (GEO) (2013). *What is GEOSS?: The Global Earth Observation System of Systems*. Retrieved January 10, 2013 from https://www.earthobservations.org/geoss.shtml

Group on Earth Observations (GEO) (2014). *Report on Progress 2011-2013*. Retrieved June 07, 2014 from https://www.earthobservations.org/docs_pub.shtml

Guenther, R. (2007). *Metadata standards at the library of congress*. Retrieved September 04, 2013 from http://www.oscars.org/science-technology/council/projects/metadata-symposium/media/dmpms_07_guenther.pdf

Halevy, A.; Franklin, M.; Maier, D. (2006). Principles of dataspace systems.

Proceedings of the Twenty-Fifth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, June 26-28, 2006 in Chicago, USA.

Haselwimmer, C.; Fretwell, P. (2009). Field reflectance spectroscopy of sparse vegetation cover on the Antarctic peninsula. *Proceedings of the IEEE First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing,* August 26-28, in Grenoble, France.

Hartig, K. (2009). What is Cloud Computing. *Cloud Computing Journal*. Retrieved April 29, 2011 from http://cloudcomputing.sys-con.com/node/579826

Heath, B. P.; McArthur, D. J.; McClelland, M. K.; Vetter, R. J. (2005). Metadata lessons from the iLumina digital library. *Communications of the ACM*, 48, 68-74.

Hendler, J. (2013). Peta Vs. Meta. *Big Data*, *1*, 82-84.

Henrysson, S. (1963). Correction of item-total correlations in item analysis. Psychometrika, 28, 211-218.

Hernández-Clemente, R.; Navarro-Cerrillo, R. M.; Zarco-Tejada, P. J. (2012). Carotenoid content estimation in a heterogeneous conifer forest using narrow-band indices and PROSPECT+ DART simulations. *Remote Sensing of Environment*, 127, 298-315.

Hesketh, M.; Sánchez-Azofeifa, G. A. (2012). The effect of seasonal spectral variation on species classification in the Panamanian tropical forest. *Remote Sensing of Environment*, 118, 73-82.

Herold, M.; Roberts, D. (2005). Spectral characteristics of asphalt road aging and deterioration: implications for remote-sensing applications. *Applied Optics*, 44, 4327-4334.

Higgins, S. (2007). *What are Metadata Standards*. Retrieved January 07, 2014 from http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards.

Howard, K. I.; Forehand, G. A. (1962). A method for correcting item-total correlations for the effect of relevant item inclusion. Educational and Psychological Measurement, 22, 731-735.

Hueni, A.; Nieke, J.; Schopfer, J. Kneubuehler, J.; Itten, K.I. (2009). The spectral Database SPECCHIO for Improved Long-Term Usability and Data Sharing. *Computers & Geosciences,* 35, 557-565.

Hueni, A.; Kneubühler, M. (2010). The Spectral Database SPECCHIO In Support of Cal/Val Activities. *Proceedings of the ESA WORKSHOP*, March 17-19, 2010 in Frascati, Italy.

Hueni, A. (2011). *SPECCHIO User Guide V. 2.1.2*. Retrieved March 23, 2012 from http://specchio.ch/user_guides.php

Hueni, A.; Chisholm, L.; Suarez, L.; Ong, C.; Wyatt, M. (2012). Spectral information system development for Australia. *Proceedings of the Geospatial Science Research Symposium*, December 10 – 12, 2012 in Melbourne, Australia.

Huettmann, F. (2009). The Global Need for, and Appreciation of, High-Quality Metadata in Biodiversity Database Work. In Spehn, E.M. & Korner, C. (Eds.) *Data Mining for Global Trends in Mountain Biodiversity* (pp 25-28). Retrieved from COMPUTERSCIENCEnetBASE.

HyVista Corporation (2012). *Hyperspectral Data Collection* [Image]. Retrieved January 03, 2014 from http://www.hyvista.com/wp_11/wp-content/uploads/2008/08/hdc.png

IEEE (2002). *Draft Standard for Learning Object Metadata.* Retrieved January 08, 2014 from http://ltsc.ieee.org/wg12/files/LOM_1484_12_1_v1_Final_Draft.pdf

Inoue, Y.; Sakaiya, E.; Zhu, Y.; Takahashi, W. (2012). Diagnostic mapping of canopy nitrogen content in rice based on hyperspectral measurements. *Remote Sensing of Environment*, 126, 210-221.

INSPIRE (2009). *INSPIRE Metadata Implementing Rules:*
Technical Guidelines based on ENISO 19915 and ENISO 19119. Retrieved May 17, 2010 from http://inspire.jrc.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf

Iordache, M. D.; Plaza, A.; Bioucas-Dias, J. (2010). On the use of spectral libraries to perform sparse unmixing of hyperspectral data. *Proceedings of the IEEE Second Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing,* June 14-16, in Reykjavik, Iceland.

ISO (2002). *ISO 19113:2002: Geographic information -- Quality principles*. Retrieved October 13, 2013 from http://www.iso.org/iso/catalogue_detail.htm?csnumber=26018

ISO (2003). *Project Information -- Fact Sheet 19113: 19113 Geographic information -- Quality principle*. Retrieved October 13, 2013 from http://www.isotc211.org/Outreach/Overview/Factsheet_19113.pdf

ISO (2011). ISO/WD 19159 Geographic Information – Calibration and validation of remote sensing imagery sensors and data. Lysaker: ISO/TC 211 Secreteriat.

ISO/IEC (2012). *ISO/IEC 38500:2008 - Corporate governance of information technology*. Retrieved May 26, 2014 from http://www.iso.org/iso/catalogue_detail?csnumber=51639

ISO/TC (2008). *Where to start – advice on creating a metadata schema*. Retrieved August 27 2013 from http://www.niso.org/apps/group_public/download.php/5271/N800R1_Where_to_start_advice_on_creating_a_metadata_schema.pdf

iVEC (n.d.). *iVEC History*. Retrieved December 30, 2013 from http://www.ivec.org/about/history/

Jacquemoud, S.; Verhoef, W.; Baret, F.; Bacour, C.; Zarco-Tejada, P. J.; Asner, G. P.; Ustin, S. L. (2009). PROSPECT+ SAIL models: A review of use for vegetation characterization. *Remote Sensing of Environment*, 113, S56-S66.

Jiao, H.; Zhong, Y.; Zhang, L. (2012) Artificial DNA Computing-Based Spectral Encoding and Matching Algorithm for Hyperspectral Remote Sensing Data. *Geoscience and Remote Sensing, IEEE Transactions on*, 50, 4085-4104.

Joshi, P.K. ; Joshi, A. (2013). Certification of Geospatial Data. *Current Science*, 105, 759-760.

Jung, A.; Goetze, C.; Glasser, C. (2010). White-reference based post-correction method for multi-source spectral libraries. *Photogrammetrie - Fernerkundung – Geoinformation*, 7, 363-369.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, 20, 141-151.

Kerekes, J. P. (1998). Error Analysis of Spectral Reflectance Data From Imaging Spectrometer Data. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, July 6-10, in Seattle, USA.

Khandar, P.V.; Dani, S.V. (2011). Knowledge Discovery and Sampling Techniques with Data Mining for Identifying Trends in Data Sets*. International Journal on Computer Science & Engineering*, Special Issue February 2011, 7-11.

Kokaly, R. F.; Asner, G. P.; Ollinger, S. V.; Martin, M. E.; Wessman, C. A. (2009). Characterizing canopy biochemistry from imaging spectroscopy and its application to ecosystem studies. *Remote Sensing of Environment*, 113, 78-91.

Knaeps, E.; Dogliotti, A. I.; Raymaekers, D.; Ruddick, K.; Sterckx, S. (2012). *In situ* evidence of non-zero reflectance in the OLCI 1020nm band for a turbid estuary. *Remote Sensing of Environment*, 120, 133-144.

KNB (2013). *The KNowledge Network for BioComplexity*. Retrieved January 08, 2014 from https://knb.ecoinformatics.org/index.jsp

Knox, N. M., Skidmore, A. K.; Prins, H. H.; Heitkönig, I.; Slotow, R.; van der Waal, C.; de Boer, W. F. (2012). Remote sensing of forage nutrients: Combining ecological and spectral absorption feature data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 72, 27-35.

Krol, D. ; Kukla, G. (2009). Quantitative Analysis of the Error Propagation Phenomenon in Distributed Information Systems. *Proceedings of the First Asian Conference on Intelligent Information and Database Systems*, April 1-3, 2009 in Dong hoi, Vietnam.

Kuo, K. S. (2010). Experiences with NASA Earth Science Data Information Systems and Suggestions for Improvements from a Scientist User Perspective. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, July 25-30, 2010 in Honolulu, USA.

Jiao, H.; Zhong, Y.; Zhang, L. (2012). Artificial DNA computing-based spectral encoding and matching algorithm for hyperspectral remote sensing data. *Geoscience and Remote Sensing, IEEE Transactions on*, 50, 4085-4104.

Jung, A.; Götze, C.; C. Gläßer, C. (2010). White reference tour 2009. A round-robin test for better spectral libraries. *DGPF Tagungsband,* 19, 433-439.

Lewis, P.; Barnsley, M. J. (1994). Influence of the sky radiance distribution on various formulations of the earth surface albedo. *Proceedings of the 6th International Symposium on Physical Measurements and Signatures in Remote Sensing, ISPRS,* January 17-21 in Val d'Isère France.

Li, X.; Strahler, A. H. (1992). Geometric-optical bidirectional reflectance modeling of the discrete crown vegetation canopy: Effect of crown shape and mutual shadowing. *Geoscience and Remote Sensing, IEEE Transactions on*, 30, 276-292.

Linting, M.; Meulman. J.J.; Groenen, P.J.; van der Koojj, A.J. (2007). Nonlinear principal component analysis: introduction and application. *Psychological Methods* 12, 36-58.

Liolios, K.; Schriml, L.; Hirschman, L.; Pagani, I.; Nosrat, B.; Sterk, P.; Field, D. (2012). The Metadata Coverage Index (MCI): A standardized metric for quantifying database metadata richness. *Standards in Genomic Sciences*, 6, 438-447.

Lloveria, R.M.; Perez-Cabello, F.; Garcia-Martin, A.; de la Riva F., J. (2008). Combined Methodology Based on Field Spectrometry and Digital Photography for Estimating Fire Severity. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 1, 266-274.

Loshin, D. (2010). *Effecting data quality improvement through data virtualization*. Accessed June 16, 2014 from http://dataqualitybook.com/kii-content/DataQualityDataVirtualization.pdf

LTER (2013). *LTER: Network Overview.* Retrieved December 30, 2013 from http://www.lternet.edu/network/

Lubin, D. (2010). Shortwave Spectroscopy for Climate and Radiation Budget Studies in the High Arctic. *Proceedings of the Art, Science and Applications of Reflectance Spectroscopy Symposium*, February 23-25, 2010 in Boulder, USA.

Mac Arthur, A. (2006). *Field Guide for the GER1500 - Single Beam Mode Radiance/irradiance measurements, Version 1.0.* Retrieved January 03, 2014 from http://fsf.nerc.ac.uk/resources/guides/pdf_guides/1500_sb-Rad-Irrad-guide_v1_toughbook.pdf

Mac Arthur, A. (2007*a*). *Field Guide for the ASD Field Spec Pro -- White Reference Mode*. Retrieved January 03, 2014 from http://fsf.nerc.ac.uk/resources/guides/pdf_guides/asd_guide_v2_wr.pdf

Mac Arthur, A. (2007*b*). *Field Guide for the GER1500 with underwater housing, Version 2.* Retrieved January 03, 2014 from http://fsf.nerc.ac.uk/resources/guides/pdf_guides/1500_uw-guide_v2_toughbook.pdf

Mac Arthur, A. (2011). *Introduction to Field Spectroscopy.* Retrieved January 03, 2014 from http://cost-es0903.fem-environment.eu/uploads/MacArthur_Introduction%20to%20Field%20Spectroscopy.pdf

MacArthur, A.; MacLellan, C. J.; Malthus, T. (2012). The fields of view and directional response functions of two field spectroradiometers. Geoscience and Remote Sensing, IEEE Transactions on, 50, 3892-3907.

Mailvaganam, M. (2007). *Introduction to Metadata*. Retrieved January 07, 2014 from http://www.dwreview.com/Articles/Metadata.html

Malta, M.C.; Baptista, A.A. (2012). State of the art on methodologies for the development of a metadata application profile. *International Journal of Metadata, Semantics and Ontologies*, 8, 332-341.

Malthus, T.; Shirinola, A. (2009). An XML-based format of exchange of spectroradiometry data*. Proceedings of the EARSeL Imaging Spectroscopy SIG*, March 16-19, 2009 in Tel Aviv, Israel.

Malthus, T.; Brando, V.; Jones, S; Held, A.; Dekker, A. (2010). Australian activities in calibration and validation for hyperspectral sensors. *Proceedings of the Hyperspectral 2010 Workshop*, March 17-19, 2010 in Frascati, Italy.

Malthus, T. (2012). *Agenda for the 'Bio-optical data: best practice and legacy datasets' workshop*. June 18-12, 2012 in Brisbane, Australia.

Margaritopoulos, T., Margaritopoulos, M., Mavridis, I., & Manitsaris, A. (2008). A Conceptual Framework for Metadata Quality Assessment. *Proceedings of the International Conference On Dublin Core And Metadata Applications*, September 22-26, 2008 in Berlin, Germany.

Martonchik, J. V.; Bruegge, C. J.; Strahler, A. H. (2000). A review of reflectance nomenclature used in remote sensing. *Remote Sensing Reviews*, 19, 9-20.

Mazel, C.H. (2006). In situ measurement of reflectance and fluorescence spectra to support hyperspectral remote sensing and marine biology research. *Oceans 2006*, 1-4.

Mazzoni, M.; Meroni, M.; Fortunato, C.; Colombo, R.; Verhoef, W. (2012). Retrieval of maize canopy fluorescence and reflectance by spectral fitting in the O_2-A absorption band. *Remote Sensing of Environment*, 124, 72-82.

McCoy, R.M. (2005). *Field Methods in Remote Sensing*. New York: The Guilford Press.

Meulman, J.J.l Heiser, W.J. (1989, 2012). *IBM SPSS Categories 21*. Retrieved January 14, 2014 from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/en/client/Manuals/IBM_SPSS_Categories.pdf

Michener, W. K., & Jones, M. B. (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in ecology & evolution*, 27, 85-93.

Millerance, F.; Bowker, G.C. (2009). *Trajectories and Enactment in the Life of an Ontology*. Retrieved September 04, 2013 from http://interoperability.ucsd.edu/docs/09millerandBowkerStds.pdf

Milton, E. J. (2009). Field Spectroscopy. In Atkinson, P. M. (Ed.), *GeoInformatics, Volume 1*. Retrieved January 03, 2014 from http://www.eolss.net/sample-chapters/c01/e6-64-02-02.pdf

Milton, E. J.; Schaepman, M.E.; Anderson, K, Kneubühler, M.; Fox, N. (2009). Progress in Field Spectroscopy. *Remote Sensing of Environment*, 113, 92-109.

Mishra, D. R.; Cho, H. J.; Ghosh, S.; Fox, A.; Downs, C.; Merani, P. B.; Mishra, S. (2012). Post-spill state of the marsh: Remote estimation of the ecological impact of the Gulf of Mexico oil spill on Louisiana Salt Marshes. *Remote Sensing of Environment*, 118, 176-185.

MIT Libraries (n.d.). *About Metadata.* Retrieved January 07, 2014 from http://libraries.mit.edu/metadata/role.html

MIT Libraries (2014). *Data Management and Publishing*. Retrieved May 26, 2014 from http://libraries.mit.edu/guides/subjects/data-management/plans.html

NASA (n.d.). *Landsat 7* [Image]. Retrieved January 03, 2014 from http://science.hq.nasa.gov/kids/imagers/teachersite/satelite.html

NASA (2011). *Earth Science Data Operations* [Image]. Retrieved December 20, 2013 from https://earthdata.nasa.gov/about-eosdis

NASA (2013*a*). *EOSDIS FY2012 Annual Metrics Report*. Retrieved January 09, 2014 from https://earthdata.nasa.gov/about-eosdis/performance/eosdis-annual-metrics-reports

NASA (2013*b*). Global Change Master Directory: Metadata Protocols and Standards Retrieved December 20, 2013 from http://gcmd.nasa.gov/add/standards/index.html

NASA (2013*c*). *Learn About GMCD*. Retrieved December 20, 2013 from http://gcmd.gsfc.nasa.gov/learn/index.html

NASA (2014). *ORNL DACC News (Winter 2014).*
Retrieved March 20, 2014 from
http://daac.ornl.gov/news/DAAC_newsletter_Winter14.pdf

NERC Field Spectroscopy Facility (n.d.). *GER1500* [Image]. Retrieved January 03, 2014 from http://fsf.nerc.ac.uk/

NERC Field Spectroscopy Facility (n.d.). *Logsheets*. Retrieved January 03, 2014 from http://fsf.nerc.ac.uk/resources/logsheets/

NERC Field Spectroscopy Facility (2014). *Introduction to Field Spectroscopy*. Retrieved June 12, 2014 from
http://fsf.nerc.ac.uk/resources/general/IntroToFS_CourseFlyer2014.pdf

Niemi, K.; Metsämäki, S.; Pulliainen, J.; Suokanerva, H.; Böttcher, K.; Leppäranta, M.; Pellikka, P. (2012). The behaviour of mast-borne spectra in a snow-covered boreal forest. *Remote Sensing of Environment*, 124, 551-563.

NISO (2004). *Understanding Metadata*. Retrieved January 07, 2014 from http://www.niso.org/publications/press/UnderstandingMetadata.pdf

NISO (2007). *A Framework of Guidance for Building Good Digital Collections*. Retrieved January 07, 2014 from http://www.niso.org/publications/rp/framework3.pdf

No end in sight to the digital data deluge (2011). *Information Management Journal,* 45, 20.

NOAA (2013*a*). *Data Access - National Climatic Data Center (NCDC).* Retrieved January 09, 2014 from http://www.ncdc.noaa.gov/data-access

NOAA (2013*b*). *Metadata Enterprise Resource Management Aid*. Retrieved December 30, 2013 from http://www.ncddc.noaa.gov/metadata-standards/mermaid/

NOAA (2013*c*). *NCDDC Metadata Training*. Retrieved December 30, 2013 from http://www.ncddc.noaa.gov/metadata-standards/metadata-training/

NOAA (2013*d*). *What is remote sensing?* Retrieved January 03, 2014 from http://oceanservice.noaa.gov/facts/remotesensing.html

Nolin, A. W.; Dozier, J. (2000). A hyperspectral method for remotely sensing the grain size of snow. *Remote Sensing of Environment*, 74, 207-216.

Oak Ridge National Laboratory (2013). *About ORNL DAAC Data Products and Services*. Retrieved January 08, 2014 from http://daac.ornl.gov/cgi-bin/search/asearch.pl

Ochoa, X.; Duval, E. (2009). Automatic evaluation of metadata quality in digital repositories. *International Journal on Digital Libraries*, 10, 67-91.

OCSD (n.d.). *OCC Project Matsu*. Retrieved January 09, 2014 from http://matsu.opensciencedatacloud.org/

OGC (2012). *Process to get OGC Certification*. Retrieved December 31, 2013 from http://cite.opengeospatial.org/getCertified

O'Neill, D. (2013). *ID3.org - The MP3 Tag Standard*. Retrieved January 07, 2014 from http://id3.org/

Orr, K. (1998, February). Data quality and systems theory. Communications of the ACM, 41(2), 66-71.

Ouyang, H.; Wang, J. (2008). Data Warehouse Software. In Tomei, L.A. (Ed.) *Encyclopedia of Information Technology Curriculum Integration* (pp 179-184). Retrieved from IGI Global InfoSci-Books.

Pacheco-Labrador, J.; Martín, M. P. (2014). Nonlinear Response in a Field Portable Spectroradiometer: Characterization and Effects on Output Reflectance. *Geoscience and Remote Sensing, IEEE Transactions on*, 52, 920-928.

Painter, T.H. (2010). Complexity of the Spectral Reflectance of Snow Cover from Field and Imaging Spectroscopy. *Proceedings of the Art, Science and Applications of Reflectance Spectroscopy Symposium*, February 23-25, 2010 in Boulder, USA.

Palmer, C. L.; Zavalina, O. L.; Mustafoff, M. (2007). Trends in metadata practices: a longitudinal study of collection federation. *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, June 18-23, 2007 in Vancouver, Canada.

Park, J.R. (2009). Metadata Quality in Digital Repositories: A Survey of the Current State of the Art. *Cataloging & Classification Quarterly*, 47, 213-228.

Pascucci, S.; Belviso, C.; Cavalli, R. M.; Palombo, A.; Pignatti, S.; Santini, F. (2012). Using imaging spectroscopy to map red mud dust waste: The Podgorica Aluminum Complex case study. *Remote Sensing of Environment*, 123, 139-154.

Peter, H.; Greenidge, C. (2009). Aligning the Warehouse and the Web. In Wang, J. (Ed.) *Encyclopedia of Data Warehousing and Mining*, *Second Edition* (pp 18-24). Retrieved from IGI Global InfoSci-Books.

Pegrum, H.; Fox, N.; Chapman, M.; Milton, E. (2006). Design and testing a new instrument to measure the angular reflectance of terrestrial surfaces. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, July 31- August 04, 2006 in Denver, USA.

Peltoniemi, J. I.; Kaasalainen, S.; Näränen, J.; Matikainen, L.; Piironen, J. (2005). Measurement of directional and spectral signatures of light reflectance by snow. *Geoscience and Remote Sensing, IEEE Transactions on*, 43, 2294-2304.

Peltoniemi, J. I.; Kaasalainen, S.; Näränen, J.; Rautiainen, M.; Stenberg, P.; Smolander, H.; Voipio, P. (2005*b*). BRDF measurement of understory vegetation in pine forests: dwarf shrubs, lichen, and moss. *Remote Sensing of Environment*, 94, 343-354.

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.

Pfitzner, K.; Bollhofer, A.; Esparon, A.; Bartolo, R.; Staben, G., (2010). Standardised spectra (400-2500 nm) and associated metadata: An example from northern tropical Australia. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, July 25-30 2010 in Honolulu, USA.

Pfitzner, K.; Bartolo, R.; Carr, G.; Esparon, A.; Bollhoefer, A. (2011). *Standards for reflectance spectral measurement of temporal vegetation plots.* Retrieved January 03, 2014 from http://www.environment.gov.au/system/files/resources/bf8002d0-2582-48a1-820f-8e79d056faed/files/ssr195.pdf

Phinn, S.; Scarth P.; Gill, T.; Roelfsema, C.; Stanford, M. (2007). *Field Spectrometer & Radiometer Guide, Version 7*. Retrieved January 03, 2014 from http://ww2.gpem.uq.edu.au/CRSSIS/publications/CRSSIS_Field_Spectrometry_Guide_010508.pdf

Pisek, J.; Rautiainen, M.; Heiskanen, J.; Mõttus, M. (2012). Retrieval of seasonal dynamics of forest understory reflectance in a Northern European boreal forest from MODIS BRDF data. *Remote Sensing of Environment*, 117, 464-468.

Ponniah, P. (2001). *Data Warehousing Fundamentals: A Comprehensive Guide for IT Professionals*, John Wily & Sons, 2001.

Qin, J.; Ball, A.; Greenberg, J. (2012). Functional and architectural requirements for metadata: supporting discovery and management of scientific data. *Proceedings of the Twelfth International Conference on Dublin Core and Metadata Applications*, September 3-7, 2012 in Kuching, Malaysia.

Rank, R,; McCormick, S.; Cremidis, C. (2010). NOAA Enterprise Archive Access Tool (NEAAT): Accelerated Application Development (XAD). *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, July 25-30, 2010 in Honolulu, USA.

Rasaiah, B. (2011). *Field Spectroscopy Metadata Survey*. Retrieved October 9, 2011 from https://www.surveymonkey.com/

Redkar, T. (2009). *Windows Azure Platform* (pp 1-51). Retrieved from SpringerLink eBooks.

Reusen I.; HYRESSA Team (2007). Towards an improved access to hyperspectral data across Europe. *Proceedings of the ISIS Meeting*, November 16-17, 2007, in Hilo, USA.

Rodger, A.; Laukamp, C.; Haest, M.; Cudahy, T. (2012). A simple quadratic method of absorption feature wavelength estimation in continuum removed spectra. *Remote Sensing of Environment*, 118, 273-283.

Salminen, M.; Pulliainen, J.; Metsämäki, S.; Kontu, A.; Suokanerva, H. (2009). The behaviour of snow and snow-free surface reflectance in boreal forests: Implications to the performance of snow covered area monitoring. *Remote Sensing of Environment*, 113, 907-918.

Sandmeier, S. R.; Itten, K. I. (1999). A field goniometer system (FIGOS) for acquisition of hyperspectral BRDF data. *Geoscience and Remote Sensing, IEEE Transactions on*, 37, 978-986.

Sandmeier, S. R. (2000). Acquisition of bidirectional reflectance factor data with field goniometers. *Remote Sensing of Environment*, 73, 257-269.

Santos, J. R. A. (1999). Cronbach's Alpha: A Tool for Assessing the Reliability of Scales. *Journal of Extension*, 37, 1-5.

SAS Institute (2010). *Data Mining and the Case for Sampling*. Retrieved October 15, 2013 from http://www.inst-informatica.pt/servicos/informacao-e-documentacao/dossiers-tematicos/dossier-tematico-no-8-business-intelligence-abril-2010/estudos-de-caso/data-mining-and-the-case-for-sampling

Sayer, A. M.; Thomas, G. E.; Grainger, R. G.; Carboni, E.; Poulsen, C.; Siddans, R. (2012). Use of MODIS-derived surface reflectance data in the ORAC-AATSR aerosol retrieval algorithm: Impact of differences between sensor spectral response functions. *Remote Sensing of Environment*, 116, 177-188.

Schaepman, M. E. (2007). Spectrodirectional remote sensing: From pixels to processes. *International Journal of Applied Earth Observation and Geoinformation*, 9, 204-223.

Schaepman, M. E.; Kneubühler, M.; Bartholomeus,H.; Malenovský, Z.; Damm, A.; Schaepman-Strub, G. (2010). Scaling Spectroscopic Approaches – From Leaf Albedo to Ecosystems Mapping. *Proceedings* of the *Art, Science and Applications of Reflectance Spectroscopy Symposium*, February 23-25, 2010 in Boulder, USA.

Schaepman-Strub, G.; Schaepman, M. E.; Painter, T. H.; Dangel, S.; Martonchik, J. V. (2006). Reflectance quantities in optical remote sensing—Definitions and case studies. *Remote sensing of environment*, 103, 27-42.

Schaepman-Strub, G.; Schaepman, M.; Martonchik, J.; Painter, T.; Dangel, S. (2009). Radiometry and reflectance: From terminology concepts to measured quantities. In T. Warner, M. Nellis, & G. Foody (Eds.), *The SAGE handbook of remote sensing.* (pp. 215-229). London: SAGE Publications Ltd.

Schopfer, J.; Dangel, S.; Kneubühler; M.; Itten, K. I. (2008). The improved dual-view field goniometer system FIGOS. *Sensors*, 8, 5120-5140.

Senf, C. (2013). *5-Year impact factor updates (2012) on remote sensing journals*. Retrieved January 02, 2014 from http://corneliussenf.wordpress.com/2013/08/05/5-year-impact-factor-updates-2012-on-remote-sensing-journals/

Serrano, L.; González-Flor, C.; Gorchs, G. (2012). Assessment of grape yield and composition using the reflectance based Water Index in Mediterranean rainfed vineyards. *Remote Sensing of Environment*, 118, 249-258.

Shafique, N.A.; Fulk, F.A.; Cormier, S. M. ; Autrey B. C. (2002). Coupling Hyperspectral Remote Sensing With Field Spectrometry to Monitor Inland Water Quality Parameters. *Proceedings of the AVIRIS Earth Science and Applications Workshop*, March 5-8, 2002 in Pasadena, USA.

Simon, S. (2010). *P.Mean: Checks for data quality using metadata*. Retrieved October 26, 2013 from http://www.pmean.com/08/CheckMetadata.html

Slamanig, D.l.; Stingl, C. (2008). Privacy Aspects of eHealth. *Proceedings of the Third International Conference on Availability, Reliability and Security*, March 4-7 2008 in Barcelona, Spain.

Spectra Vista Corporation (n.d.). *SVC GER 1500*. Retrieved January 05, 2014 from http://www.precisionphotometrics.com/pdf-files/SVC_GER1500(2).pdf

Starkweather, J. (2012). *RSS SPSS Short Course Module 9 Categorical PCA*. Retrieved August 30, 2013 from http://www.unt.edu/rss/class/Jon/SPSS_SC/Module9/M9_CATPCA/SPSS_M9_CATPCA.htm

States News Service (2012). *Scientists Team Up with Google in an Australian First*. Retrieved January 20, 2014 from Academic One File.

Stuckens, J.; Somers, B.; Verstraeten, W.W.; Swennen, R.; Coppin, P. (2009). Normalization of Illumination Conditions For Ground Based Hyperspectral Measurements Using Dual Field of View Spectroradiometers and BRDF Corrections in *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, July 12-17, in Cape Town, South Africa.

Stvilia, B.; Gasser, L.; Twidale, M. B.; Shreeves, S. L. Cole, T. W. (2004). Metadata Quality For Federated Collections. *Proceedings of the Ninth International Conference on Information Quality*, November 5-7, 2004 in Cambridge, USA.

Stvilia, B.; Gasser, L.; Twidale, M. B.; Smith, L. C. (2007). A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58, 1720-1733.

Syn, S.Y.; Spring, M.B. (2008). Can a system make novice users experts? Analysis of metadata created by novices and experts with varying levels of assistance. *International Journal of Metadata, Semantics and Ontologies*, 3, 122-131.

Tabachnick, B.G.; Fidell, L.S. (2007). *Using Multivariate Statistics*, 5th Ed. New York: Pearson Education.

Tarbet, A. (2012). *Metadata Concept Map* [Image]. Retrieved January 07, 2014 from http://hackescilibship.wordpress.com/category/metadata-schema/

TERN (2013). *TERN*. Retrieved January 09, 2014 from http://www.tern.org.au/rs/7/sites/998/user_uploads/File/TERN%20Corporate%20brochure.pdf

The University of Queensland Library (2011). *Metadata*. Retrieved January 07, 2014 from http://www.library.uq.edu.au/research-support/metadata.

Thorp, K. R.; Wang, G.; West, A. L.; Moran, M. S.; Bronson, K. F.; White, J. W.; Mon, J. (2012). Estimating crop biophysical properties from remote sensing data by inverting linked radiative transfer and ecophysiological models. *Remote Sensing of Environment*, 124, 224-233.

Tits, L.; De Keersmaecker, W.; Somers, B.; Asner, G. P.; Farifteh, J.; Coppin, P. (2012). Hyperspectral shape-based unmixing to improve intra-and interclass variability for forest and agro-ecosystem monitoring. *ISPRS Journal of Photogrammetry and Remote Sensing*, 74, 163-174.

Tolle, K.M.; Tansley, D.; Hey, A.J.G. (2011). The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. *Proceedings of the IEEE*, *99*, 1334-1337.

USGS (2002). *Photo showing dam face surface and spectral measurement technique* [Image]. Retrieved January 7, 2014 from http://pubs.usgs.gov/of/2002/ofr-02-0199/PCcalib_NEW4-02_PUBLISH.html

USGS (2006). *USGS Digital Spectral Library splib06a*. Retrieved October 26, 2013 from
http://speclab.cr.usgs.gov/spectral.lib06/ds231/index.html

USGS (2013*a*). *Geospatial Metadata Validation Service*. Retrieved December 25, 2013 from http://geo-nsdi.er.usgs.gov/validation/

USGS (2013*b*). *USGS Data Management - Data Stewardship: Roles and Responsibilities*. Retrieved May 26, 2014 from
http://www.usgs.gov/datamanagement/plan/stewardship.php

Viscarra Rossel, R. A.; Cattle, S. R.; Ortega, A.; Fouad, Y. (2009). In situ measurements of soil colour, mineral composition and clay content by vis–NIR spectroscopy. *Geoderma*, 150, 253-266.

Wason, T.D.; Wiley, D. (2000). Structured Metadata spaces. *Journal of Internet Cataloging*, 3, 263-277.

Watts, S.; Shankaranarayanan, G.; Even, A. (2009). Data quality assessment in context: A cognitive perspective. *Decision Support Systems*, 48, 202-211.

Wayne, Lynda. *Institutionalize metadata before it institutionalizes you*. Retrieved March 10, 2014 from http://www.fgdc.gov/metadata/documents/InstitutionalizeMeta_Nov2005.doc

Williams, A. J.; Ekins, S.; Tkachenko, V. (2012). Towards a gold standard: regarding quality in public domain chemistry databases and approaches to improving the situation. *Drug discovery today*, 17, 685-701.

Williams, M. (2012). *Browse your images and their EXIF metadata with Photo Data Explorer* [Image]. Retrieved January 07, 2014 from http://www.softwarecrew.com/2012/06/browse-your-images-and-their-exif-metadata-with-photo-data-explorer/

Wynholds, L.A.; Wallis, J.C.; Borgman, C.L.; Sands, A.; Traweek, S. (2012). Data, Data Use, and Scientific Inquiry: Two Case Studies of Data Practices. *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, June 10-14 in Washington, USA.

Yang, C.; Goodchild, M.; Huang, Q.; Nebert, D.; Raskin, R.; Xu, Y.; Fay, D. (2011). Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing?. *International Journal of Digital Earth*, 4*,* 305-329.

Zhang, Y.; Slaughter, D. C.; Staab, E. S. (2012). Robust hyperspectral vision-based classification for multi-season weed mapping. *ISPRS Journal of Photogrammetry and Remote Sensing*, 69, 65-73.

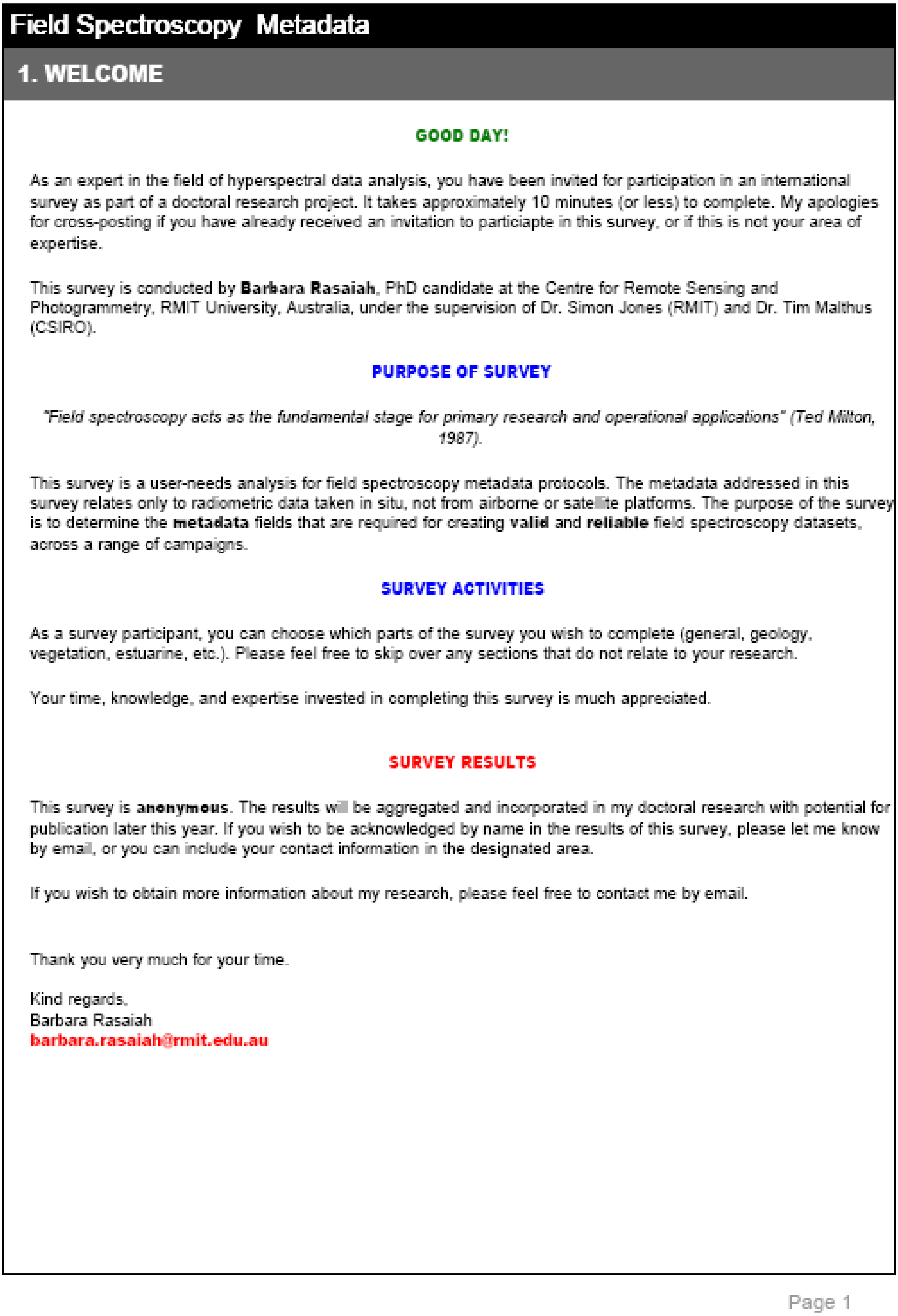
Zheng, Y.; Zhao, K.; Stylianou, A. (2013). The impacts of information quality and system quality on users' continuance intention in information-exchange virtual communities: An empirical investigation. *Decision Support Systems*, 56, 513-524.

Zomer, R. J., Trabucco, A., Ustin, S. L. (2009). Building spectral libraries for wetlands land cover classification and hyperspectral remote sensing. *Journal of Environmental Management*, 90, 2170-2177.

Appendix A Field Spectroscopy Metadata Survey (Questions and Results)

Appendix A.1 Survey Questions

This section contains all questions submitted to the expert panel participating on the online field spectroscopy metadata survey (a total of 25 pages, including the introduction to the survey and an explanation of the criticality rankings).



Field Spectroscopy Metadata

1. WELCOME

GOOD DAY!

As an expert in the field of hyperspectral data analysis, you have been invited for participation in an international survey as part of a doctoral research project. It takes approximately 10 minutes (or less) to complete. My apologies for cross-posting if you have already received an invitation to particiapte in this survey, or if this is not your area of expertise.

This survey is conducted by **Barbara Rasaiah**, PhD candidate at the Centre for Remote Sensing and Photogrammetry, RMIT University, Australia, under the supervision of Dr. Simon Jones (RMIT) and Dr. Tim Malthus (CSIRO).

PURPOSE OF SURVEY

"Field spectroscopy acts as the fundamental stage for primary research and operational applications" (Ted Milton, 1987).

This survey is a user-needs analysis for field spectroscopy metadata protocols. The metadata addressed in this survey relates only to radiometric data taken in situ, not from airborne or satellite platforms. The purpose of the survey is to determine the **metadata** fields that are required for creating **valid** and **reliable** field spectroscopy datasets, across a range of campaigns.

SURVEY ACTIVITIES

As a survey participant, you can choose which parts of the survey you wish to complete (general, geology, vegetation, estuarine, etc.). Please feel free to skip over any sections that do not relate to your research.

Your time, knowledge, and expertise invested in completing this survey is much appreciated.

SURVEY RESULTS

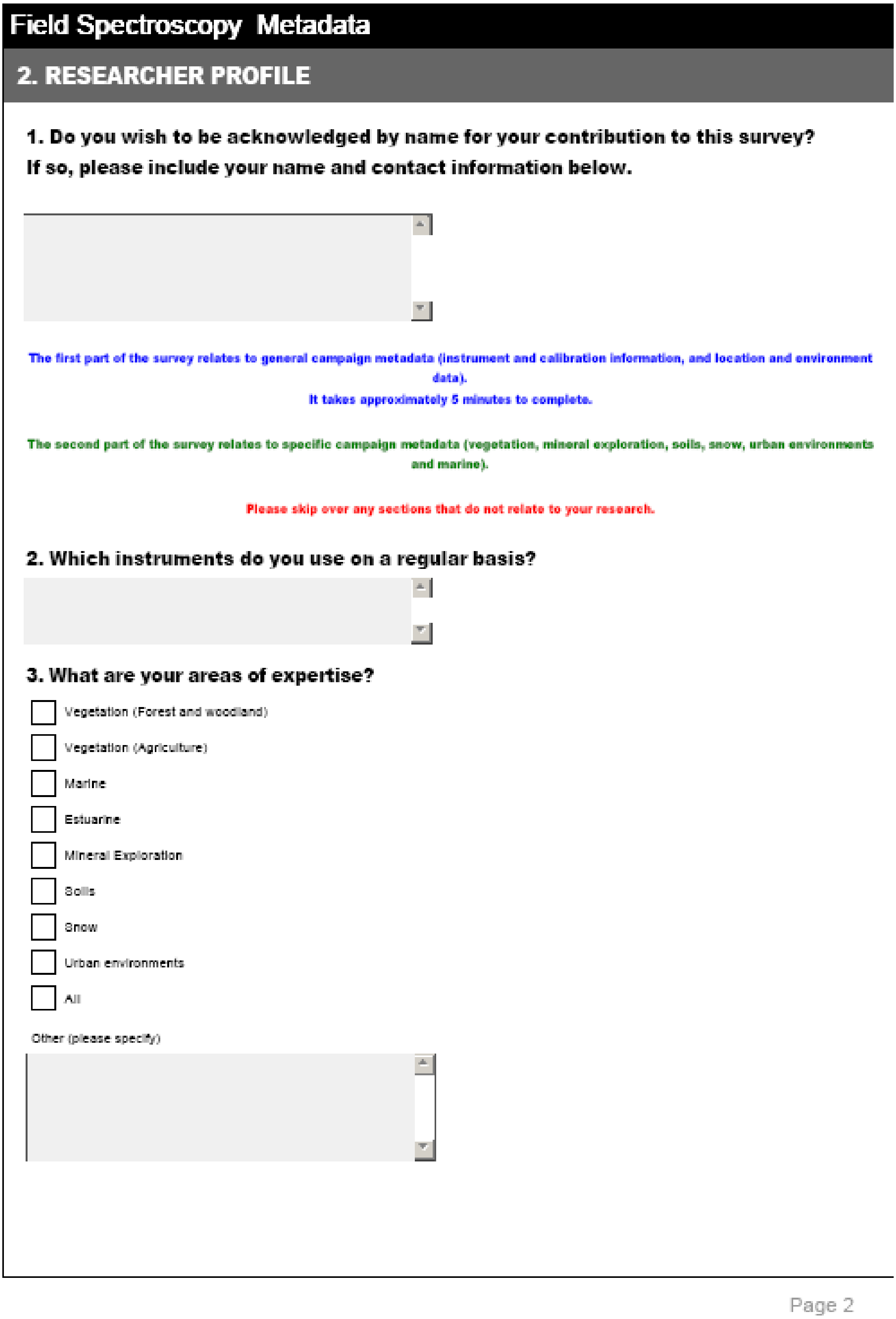
This survey is **anonymous**. The results will be aggregated and incorporated in my doctoral research with potential for publication later this year. If you wish to be acknowledged by name in the results of this survey, please let me know by email, or you can include your contact information in the designated area.

If you wish to obtain more information about my research, please feel free to contact me by email.

Thank you very much for your time.

Kind regards,
Barbara Rasaiah
barbara.rasaiah@rmit.edu.au

Page 1



Field Spectroscopy Metadata

2. RESEARCHER PROFILE

1. Do you wish to be acknowledged by name for your contribution to this survey? If so, please include your name and contact information below.

The first part of the survey relates to general campaign metadata (instrument and calibration information, and location and environment data).
It takes approximately 5 minutes to complete.

The second part of the survey relates to specific campaign metadata (vegetation, mineral exploration, soils, snow, urban environments and marine).

Please skip over any sections that do not relate to your research.

2. Which instruments do you use on a regular basis?

3. What are your areas of expertise?

□ Vegetation (Forest and woodland)

□ Vegetation (Agriculture)

□ Marine

□ Estuarine

□ Mineral Exploration

□ Soils

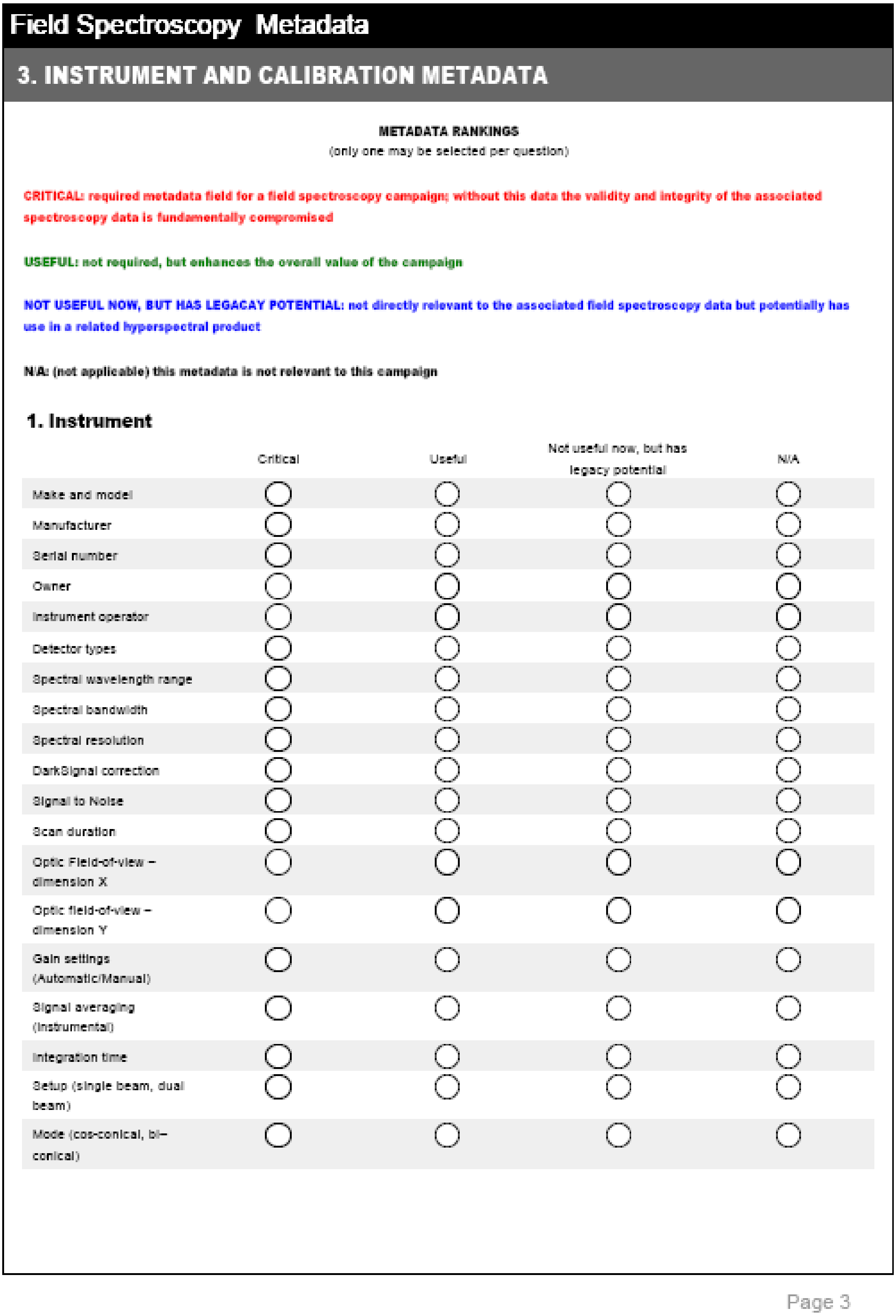
□ Snow

□ Urban environments

□ All

Other (please specify)

Page 2



Field Spectroscopy Metadata

3. INSTRUMENT AND CALIBRATION METADATA

METADATA RANKINGS
(only one may be selected per question)

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

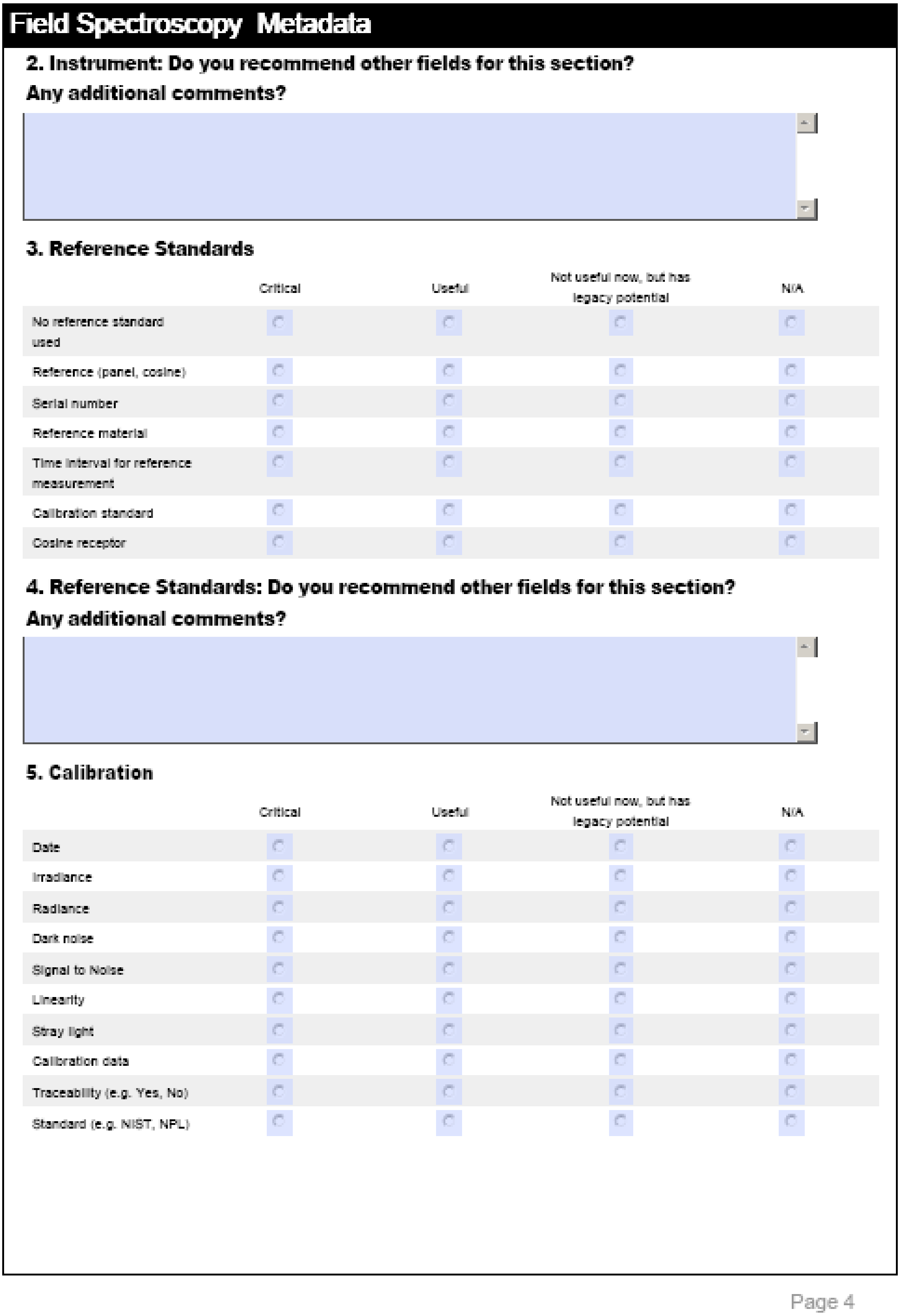
NOT USEFUL NOW, BUT HAS LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

N/A: (not applicable) this metadata is not relevant to this campaign

1. Instrument

	Critical	Useful	Not useful now, but has legacy potential	N/A
Make and model	○	○	○	○
Manufacturer	○	○	○	○
Serial number	○	○	○	○
Owner	○	○	○	○
Instrument operator	○	○	○	○
Detector types	○	○	○	○
Spectral wavelength range	○	○	○	○
Spectral bandwidth	○	○	○	○
Spectral resolution	○	○	○	○
DarkSignal correction	○	○	○	○
Signal to Noise	○	○	○	○
Scan duration	○	○	○	○
Optic Field-of-view – dimension X	○	○	○	○
Optic field-of-view – dimension Y	○	○	○	○
Gain settings (Automatic/Manual)	○	○	○	○
Signal averaging (instrumental)	○	○	○	○
Integration time	○	○	○	○
Setup (single beam, dual beam)	○	○	○	○
Mode (cos-conical, bi-conical)	○	○	○	○

Page 3



Field Spectroscopy Metadata

2. Instrument: Do you recommend other fields for this section?

Any additional comments?

3. Reference Standards

	Critical	Useful	Not useful now, but has legacy potential	N/A
No reference standard used	○	○	○	○
Reference (panel, cosine)	○	○	○	○
Serial number	○	○	○	○
Reference material	○	○	○	○
Time interval for reference measurement	○	○	○	○
Calibration standard	○	○	○	○
Cosine receptor	○	○	○	○

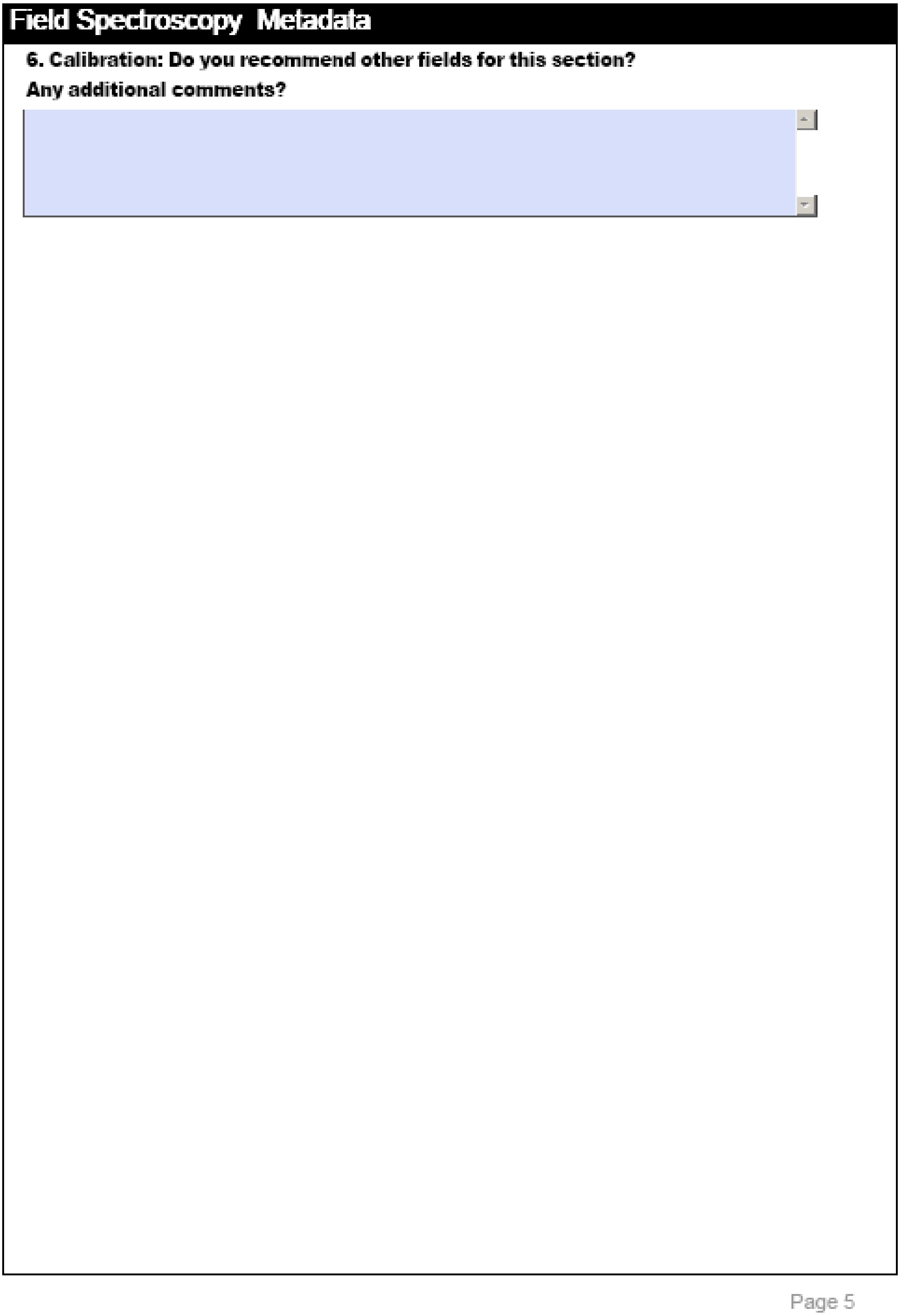
4. Reference Standards: Do you recommend other fields for this section?

Any additional comments?

5. Calibration

	Critical	Useful	Not useful now, but has legacy potential	N/A
Date	○	○	○	○
Irradiance	○	○	○	○
Radiance	○	○	○	○
Dark noise	○	○	○	○
Signal to Noise	○	○	○	○
Linearity	○	○	○	○
Stray light	○	○	○	○
Calibration data	○	○	○	○
Traceability (e.g. Yes, No)	○	○	○	○
Standard (e.g. NIST, NPL)	○	○	○	○

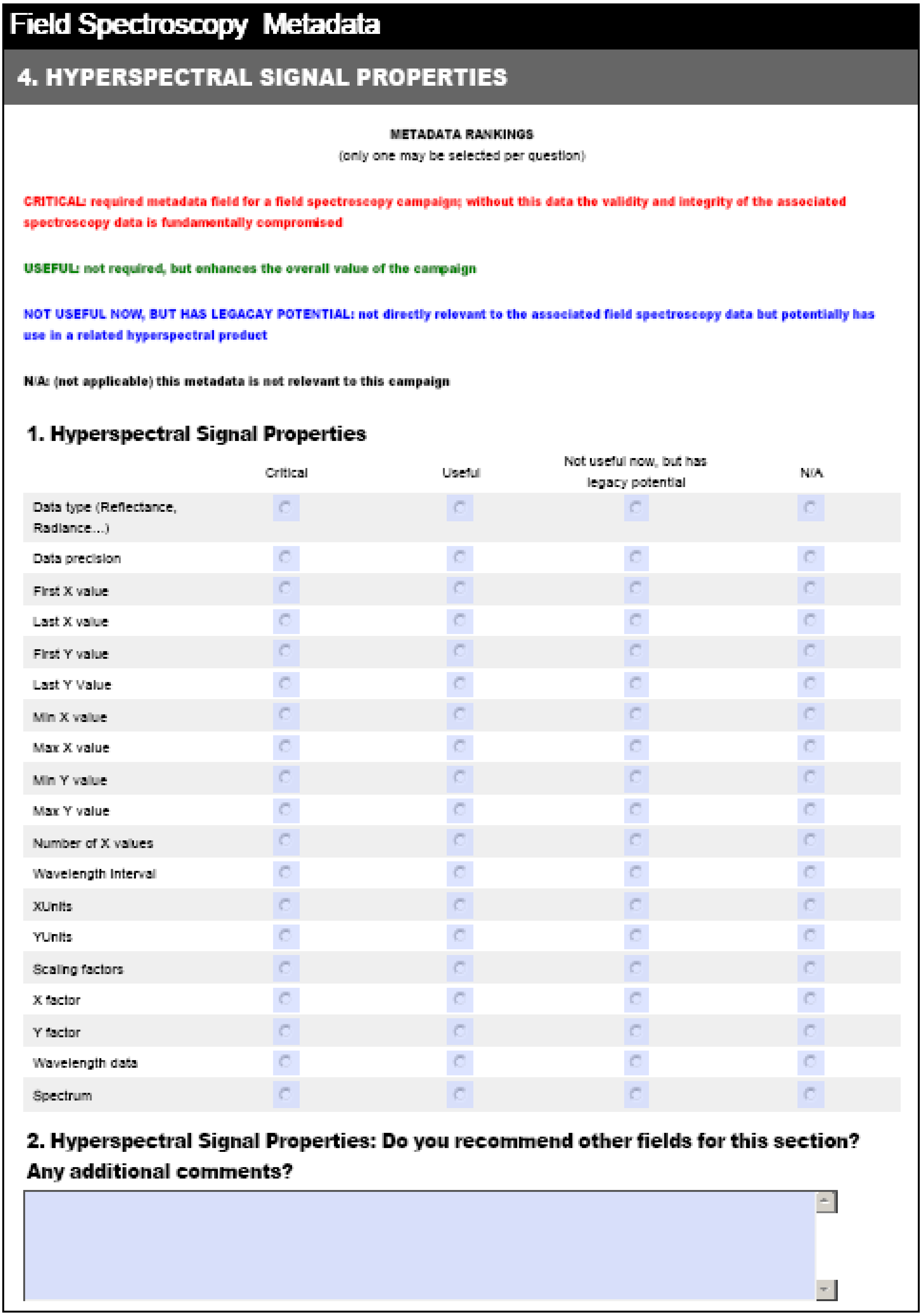
Page 4



Field Spectroscopy Metadata

6. Calibration: Do you recommend other fields for this section?

Any additional comments?



Field Spectroscopy Metadata

4. HYPERSPECTRAL SIGNAL PROPERTIES

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

N/A: (not applicable) this metadata is not relevant to this campaign

1. Hyperspectral Signal Properties

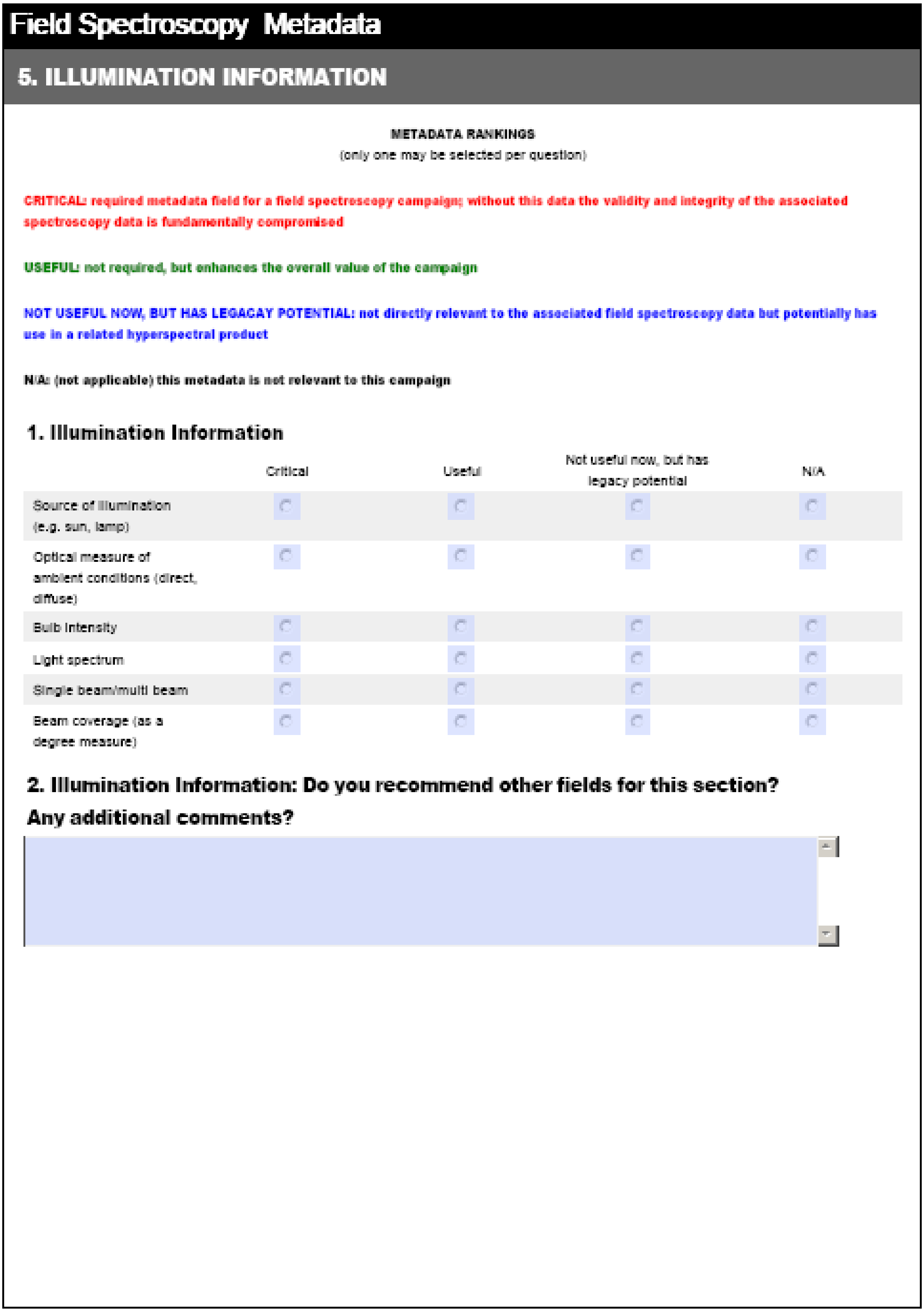
	Critical	Useful	Not useful now, but has legacy potential	N/A
Data type (Reflectance, Radiance...)	○	○	○	○
Data precision	○	○	○	○
First X value	○	○	○	○
Last X value	○	○	○	○
First Y value	○	○	○	○
Last Y Value	○	○	○	○
Min X value	○	○	○	○
Max X value	○	○	○	○
Min Y value	○	○	○	○
Max Y value	○	○	○	○
Number of X values	○	○	○	○
Wavelength Interval	○	○	○	○
XUnits	○	○	○	○
YUnits	○	○	○	○
Scaling factors	○	○	○	○
X factor	○	○	○	○
Y factor	○	○	○	○
Wavelength data	○	○	○	○
Spectrum	○	○	○	○

2. Hyperspectral Signal Properties: Do you recommend other fields for this section? Any additional comments?

Page 6

Field Spectroscopy Metadata

Page 7



Field Spectroscopy Metadata

5. ILLUMINATION INFORMATION

METADATA RANKINGS
(only one may be selected per question)

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

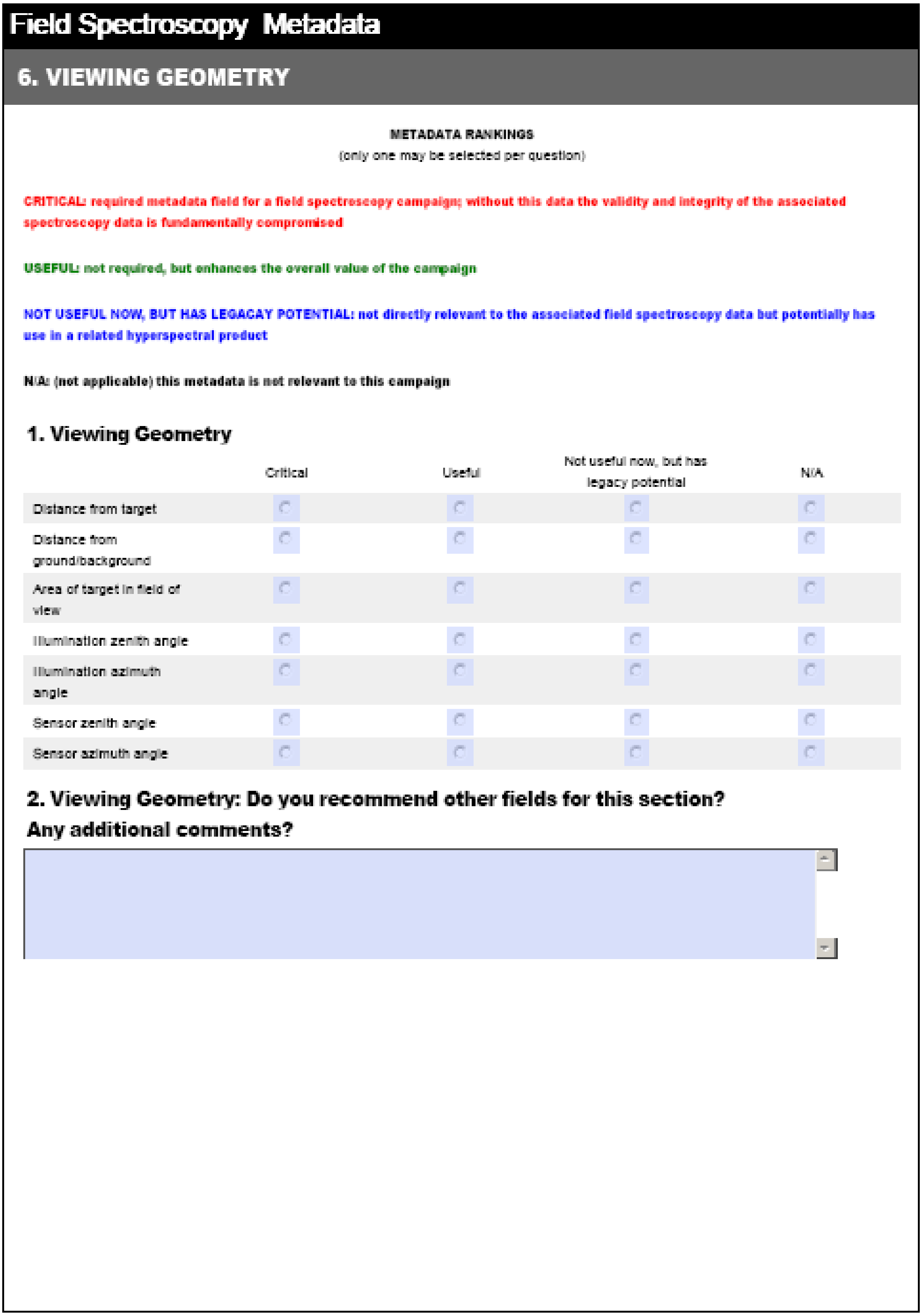
NOT USEFUL NOW, BUT HAS LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

N/A: (not applicable) this metadata is not relevant to this campaign

1. Illumination Information

	Critical	Useful	Not useful now, but has legacy potential	N/A
Source of illumination (e.g. sun, lamp)	○	○	○	○
Optical measure of ambient conditions (direct, diffuse)	○	○	○	○
Bulb intensity	○	○	○	○
Light spectrum	○	○	○	○
Single beam/multi beam	○	○	○	○
Beam coverage (as a degree measure)	○	○	○	○

2. Illumination Information: Do you recommend other fields for this section? Any additional comments?



Field Spectroscopy Metadata

6. VIEWING GEOMETRY

METADATA RANKINGS

(only one may be selected per question)

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 LEGAGAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

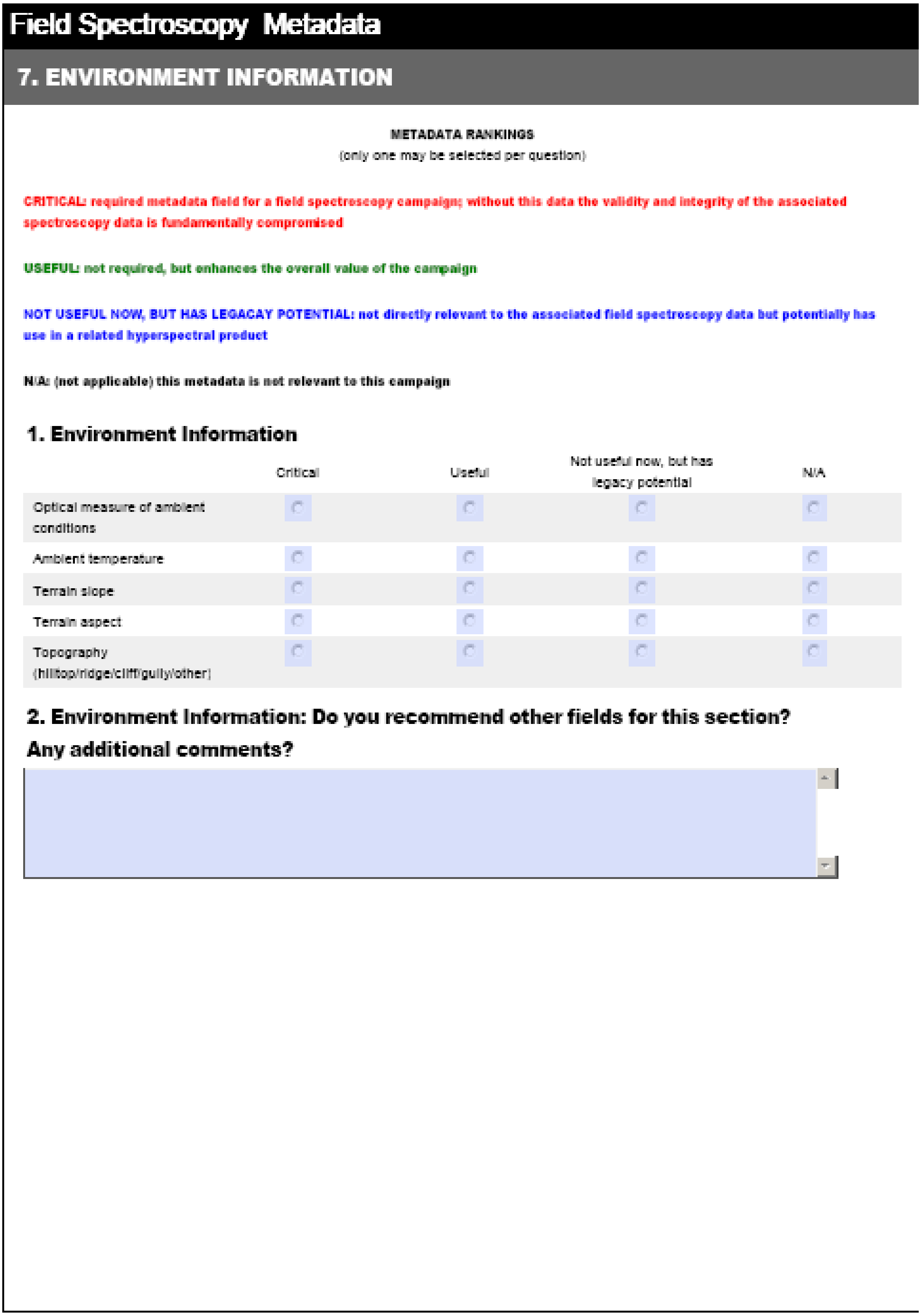
N/A: (not applicable) this metadata is not relevant to this campaign

1. Viewing Geometry

	Critical	Useful	Not useful now, but has legacy potential	N/A
Distance from target	○	○	○	○
Distance from ground/background	○	○	○	○
Area of target in field of view	○	○	○	○
Illumination zenith angle	○	○	○	○
Illumination azimuth angle	○	○	○	○
Sensor zenith angle	○	○	○	○
Sensor azimuth angle	○	○	○	○

2. Viewing Geometry: Do you recommend other fields for this section? Any additional comments?

Page 9



Field Spectroscopy Metadata

7. ENVIRONMENT INFORMATION

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

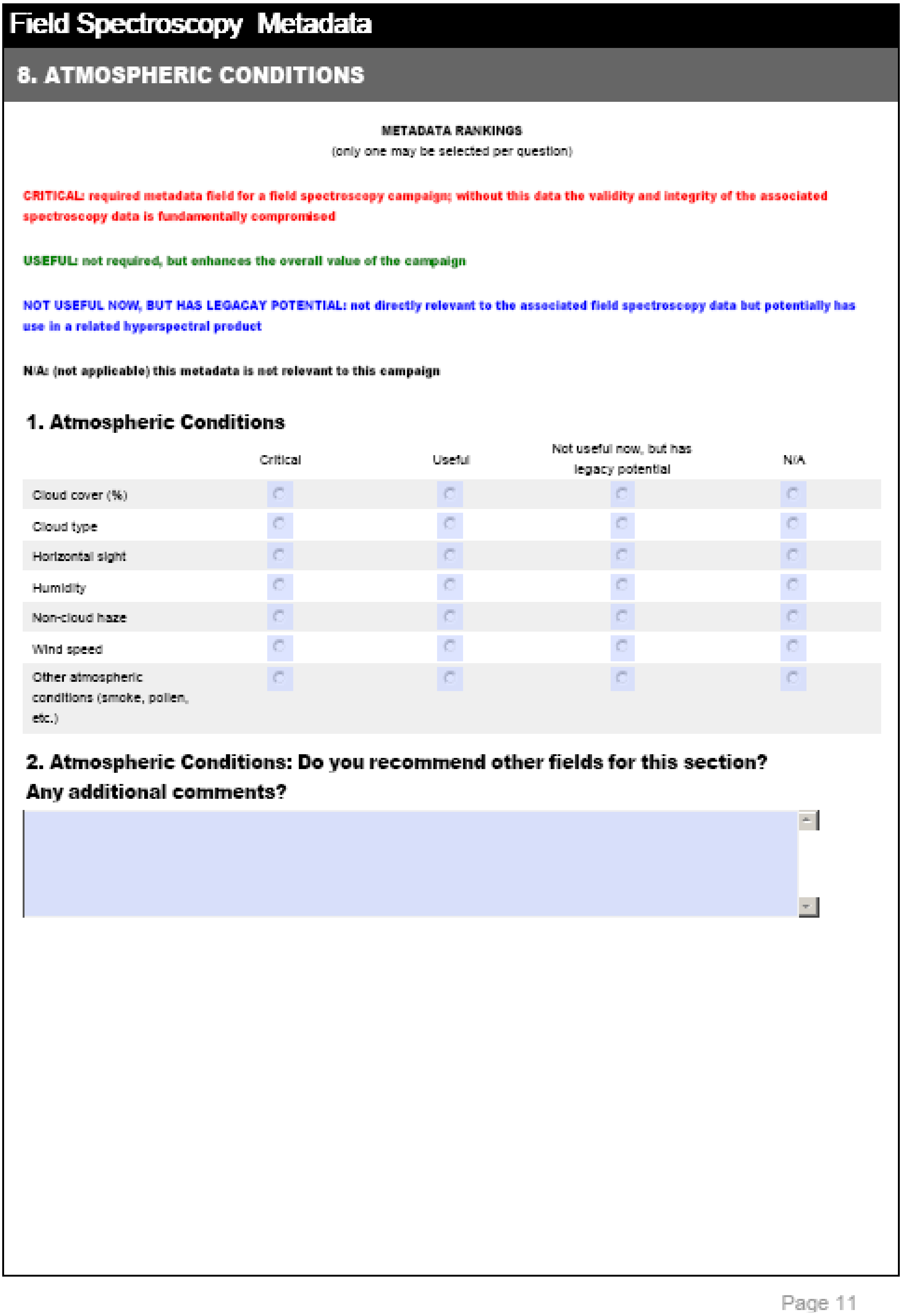
N/A: (not applicable) this metadata is not relevant to this campaign

1. Environment Information

	Critical	Useful	Not useful now, but has legacy potential	N/A
Optical measure of ambient conditions	○	○	○	○
Ambient temperature	○	○	○	○
Terrain slope	○	○	○	○
Terrain aspect	○	○	○	○
Topography (hilltop/ridge/cliff/gully/other)	○	○	○	○

2. Environment Information: Do you recommend other fields for this section? Any additional comments?

Page 10



Field Spectroscopy Metadata

8. ATMOSPHERIC CONDITIONS

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

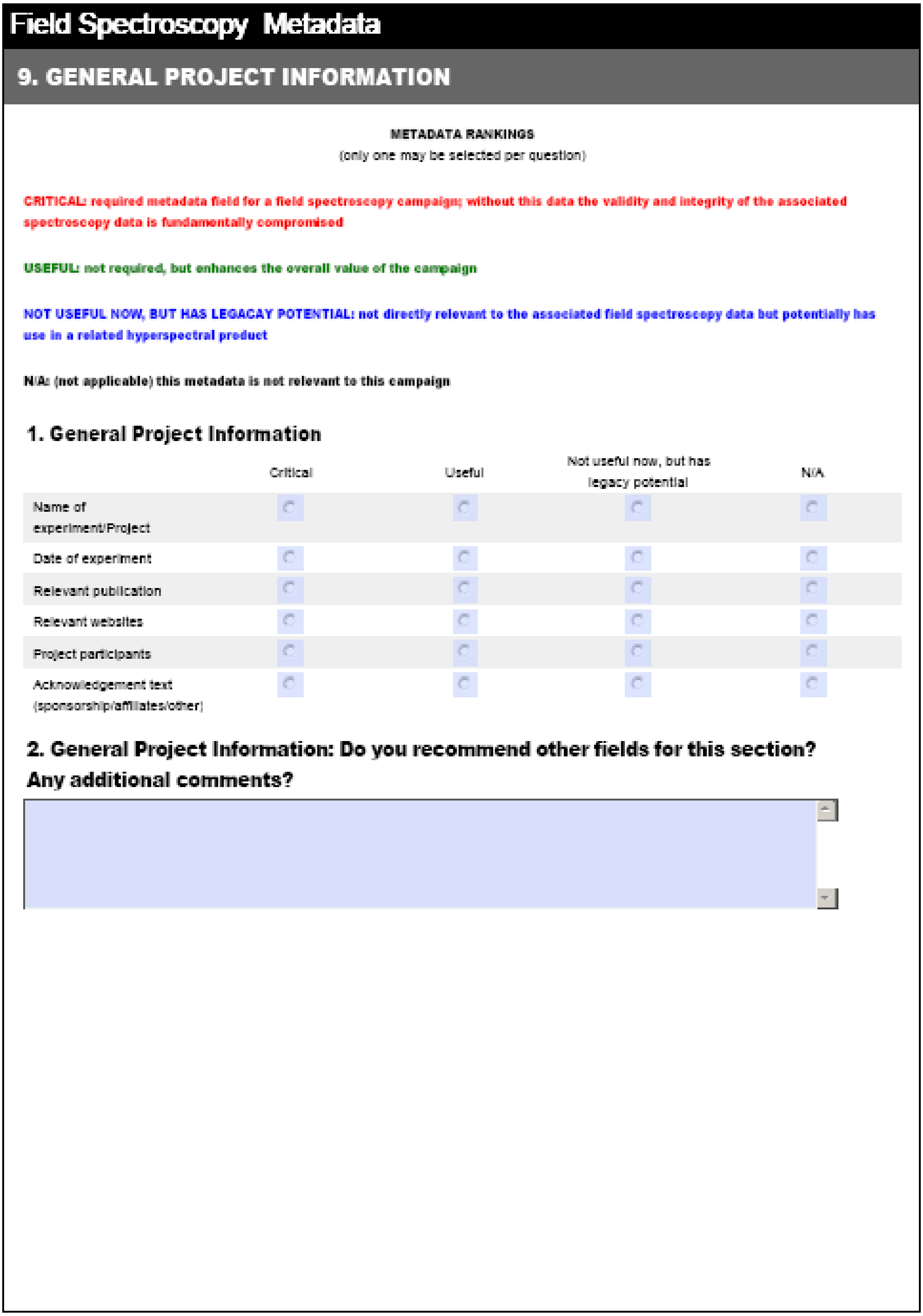
N/A: (not applicable) this metadata is not relevant to this campaign

1. Atmospheric Conditions

	Critical	Useful	Not useful now, but has legacy potential	N/A
Cloud cover (%)	○	○	○	○
Cloud type	○	○	○	○
Horizontal sight	○	○	○	○
Humidity	○	○	○	○
Non-cloud haze	○	○	○	○
Wind speed	○	○	○	○
Other atmospheric conditions (smoke, pollen, etc.)	○	○	○	○

2. Atmospheric Conditions: Do you recommend other fields for this section? Any additional comments?

Page 11



Field Spectroscopy Metadata

9. GENERAL PROJECT INFORMATION

METADATA RANKINGS
(only one may be selected per question)

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

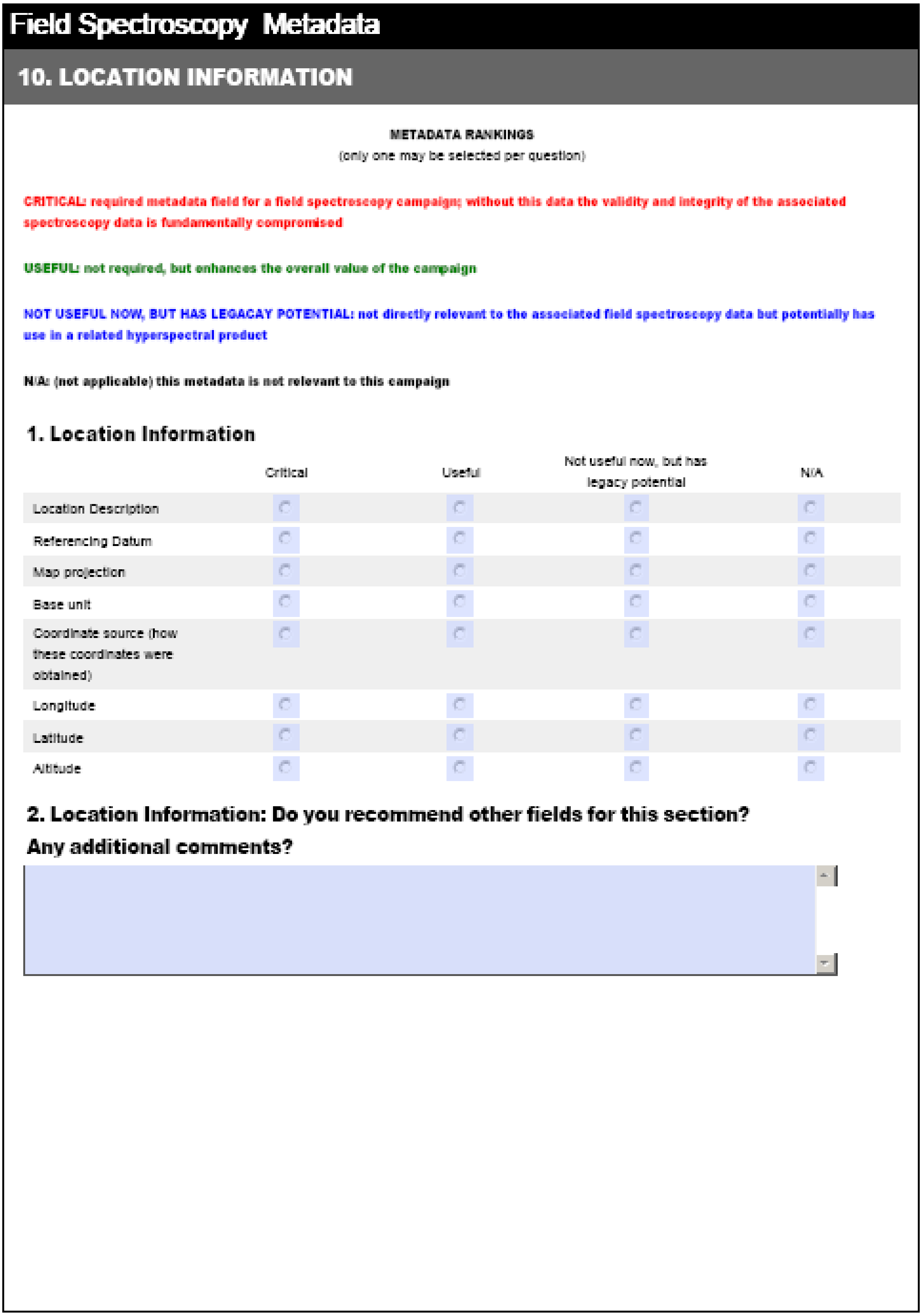
NOT USEFUL NOW, BUT HAS LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

N/A: (not applicable) this metadata is not relevant to this campaign

1. General Project Information

	Critical	Useful	Not useful now, but has legacy potential	N/A
Name of experiment/Project	○	○	○	○
Date of experiment	○	○	○	○
Relevant publication	○	○	○	○
Relevant websites	○	○	○	○
Project participants	○	○	○	○
Acknowledgement text (sponsorship/affiliates/other)	○	○	○	○

2. General Project Information: Do you recommend other fields for this section? Any additional comments?



Field Spectroscopy Metadata

10. LOCATION INFORMATION

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

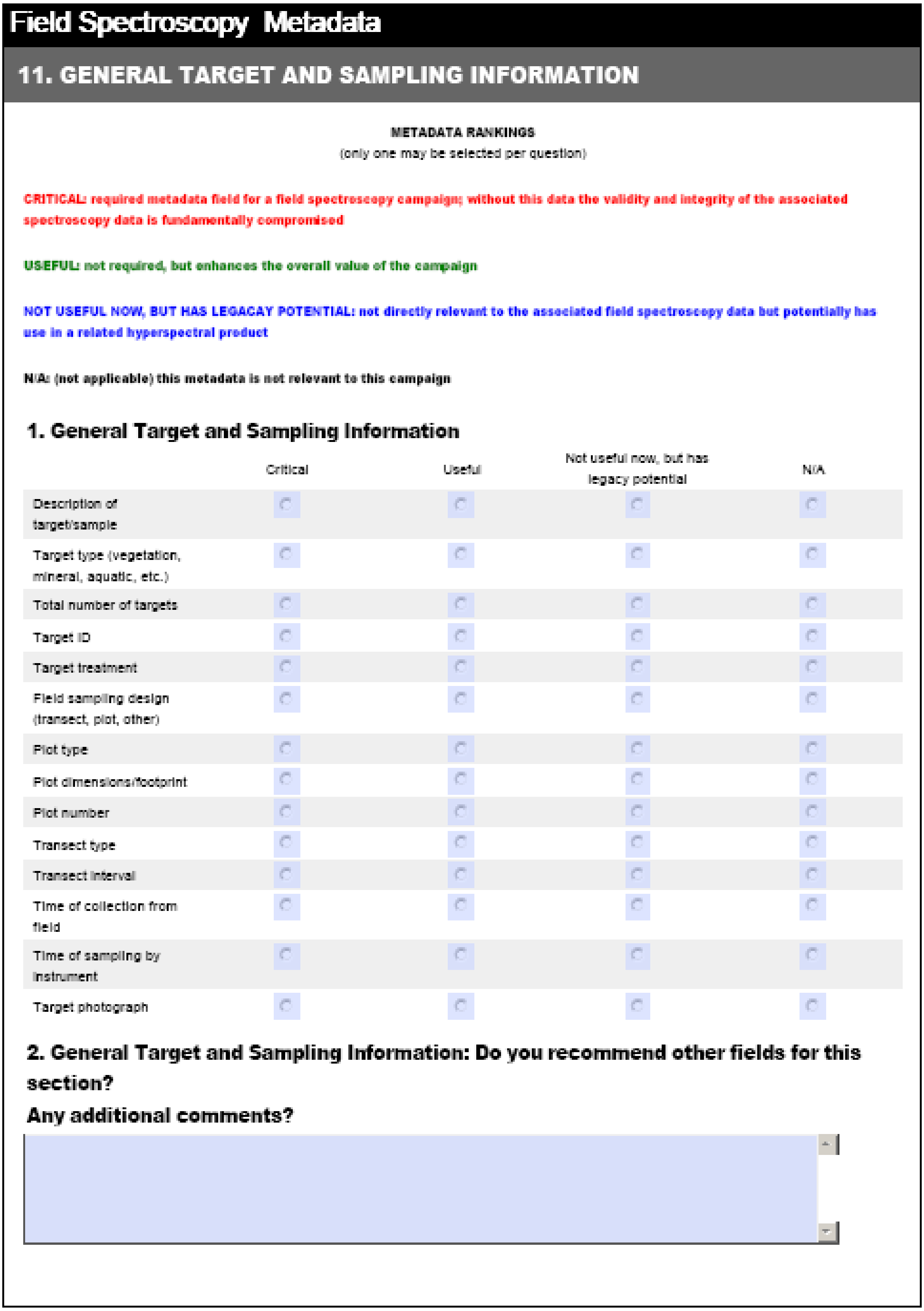
N/A: (not applicable) this metadata is not relevant to this campaign

1. Location Information

	Critical	Useful	Not useful now, but has legacy potential	N/A
Location Description	○	○	○	○
Referencing Datum	○	○	○	○
Map projection	○	○	○	○
Base unit	○	○	○	○
Coordinate source (how these coordinates were obtained)	○	○	○	○
Longitude	○	○	○	○
Latitude	○	○	○	○
Altitude	○	○	○	○

2. Location Information: Do you recommend other fields for this section? Any additional comments?

Page 13



Field Spectroscopy Metadata

11. GENERAL TARGET AND SAMPLING INFORMATION

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

N/A: (not applicable) this metadata is not relevant to this campaign

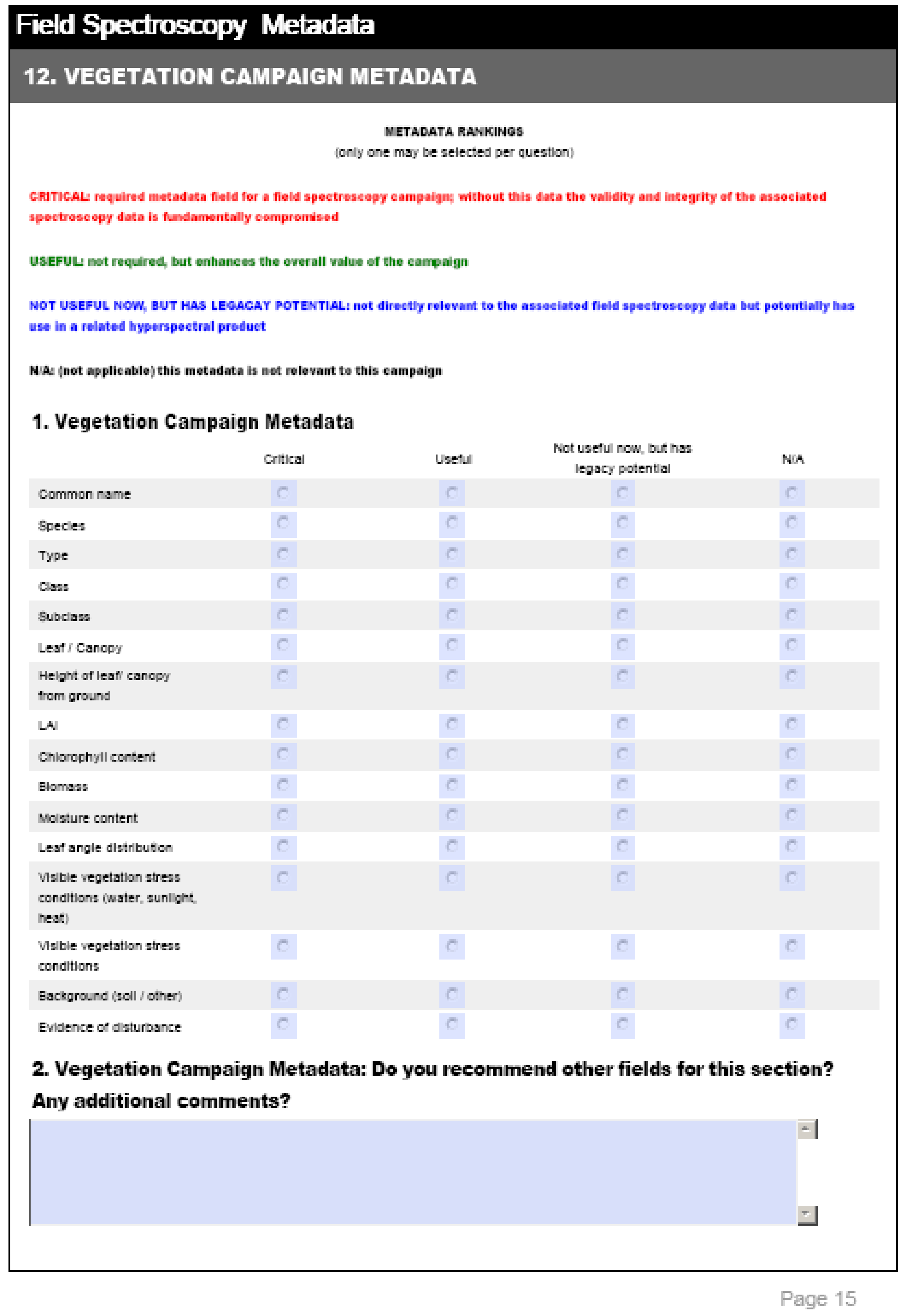
1. General Target and Sampling Information

	Critical	Useful	Not useful now, but has legacy potential	N/A
Description of target/sample	○	○	○	○
Target type (vegetation, mineral, aquatic, etc.)	○	○	○	○
Total number of targets	○	○	○	○
Target ID	○	○	○	○
Target treatment	○	○	○	○
Field sampling design (transect, plot, other)	○	○	○	○
Plot type	○	○	○	○
Plot dimensions/footprint	○	○	○	○
Plot number	○	○	○	○
Transect type	○	○	○	○
Transect interval	○	○	○	○
Time of collection from field	○	○	○	○
Time of sampling by instrument	○	○	○	○
Target photograph	○	○	○	○

2. General Target and Sampling Information: Do you recommend other fields for this section?

Any additional comments?

Page 14



Field Spectroscopy Metadata

12. VEGETATION CAMPAIGN METADATA

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

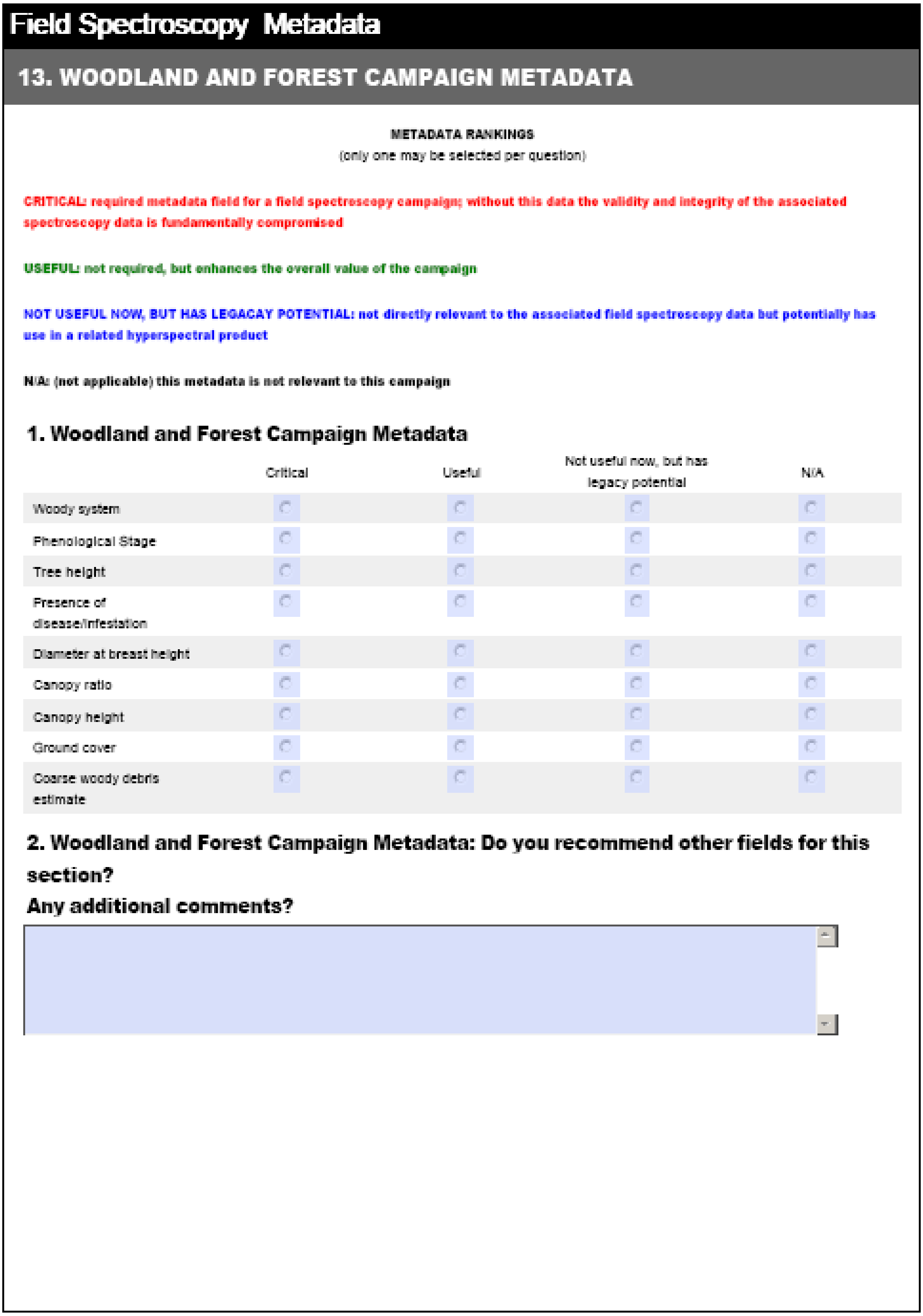
N/A: (not applicable) this metadata is not relevant to this campaign

1. Vegetation Campaign Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Common name	○	○	○	○
Species	○	○	○	○
Type	○	○	○	○
Class	○	○	○	○
Subclass	○	○	○	○
Leaf / Canopy	○	○	○	○
Height of leaf/ canopy from ground	○	○	○	○
LAI	○	○	○	○
Chlorophyll content	○	○	○	○
Biomass	○	○	○	○
Moisture content	○	○	○	○
Leaf angle distribution	○	○	○	○
Visible vegetation stress conditions (water, sunlight, heat)	○	○	○	○
Visible vegetation stress conditions	○	○	○	○
Background (soil / other)	○	○	○	○
Evidence of disturbance	○	○	○	○

2. Vegetation Campaign Metadata: Do you recommend other fields for this section? Any additional comments?

Page 15



Field Spectroscopy Metadata

13. WOODLAND AND FOREST CAMPAIGN METADATA

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

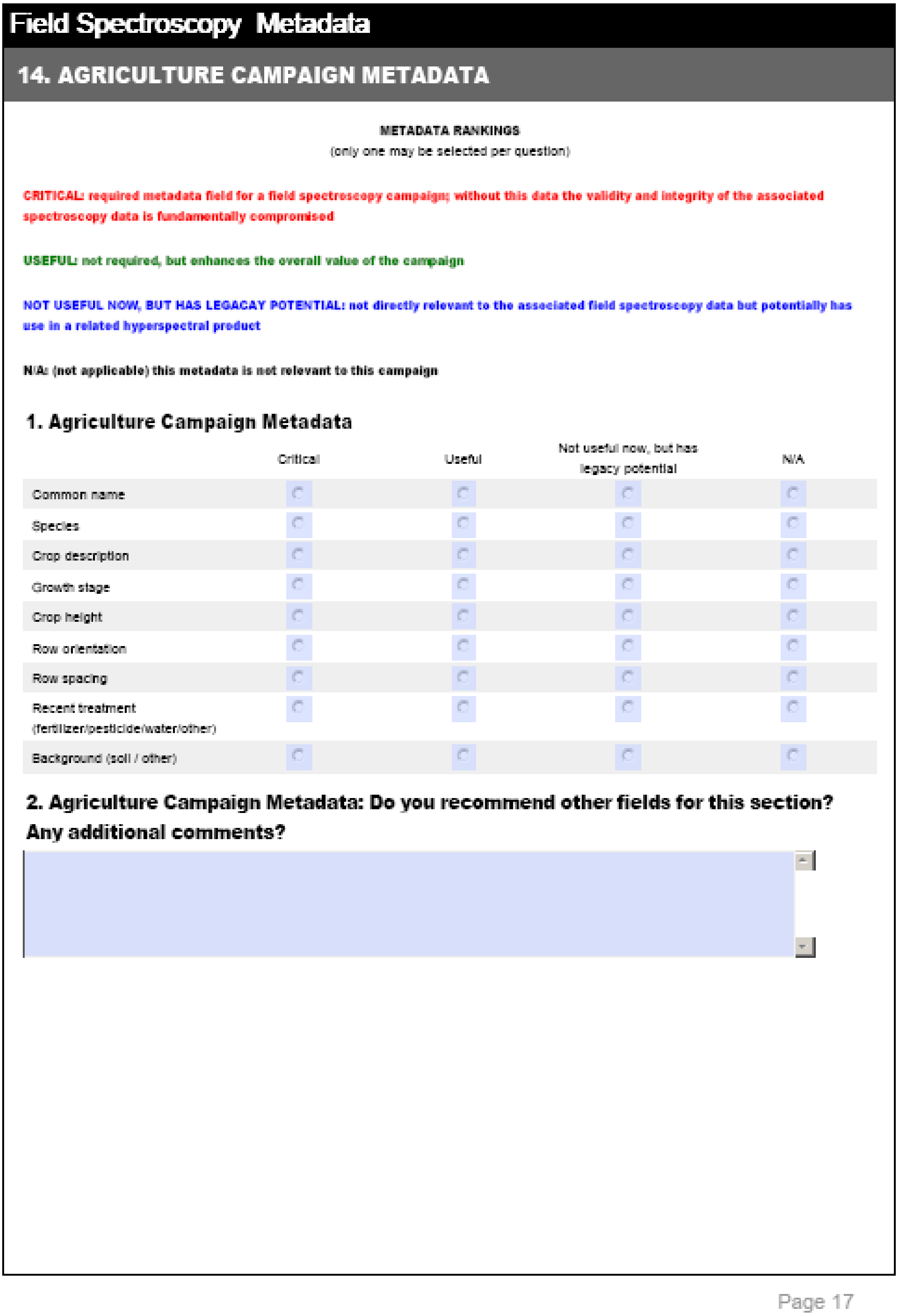
N/A: (not applicable) this metadata is not relevant to this campaign

1. Woodland and Forest Campaign Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Woody system	○	○	○	○
Phenological Stage	○	○	○	○
Tree height	○	○	○	○
Presence of disease/infestation	○	○	○	○
Diameter at breast height	○	○	○	○
Canopy ratio	○	○	○	○
Canopy height	○	○	○	○
Ground cover	○	○	○	○
Coarse woody debris estimate	○	○	○	○

2. Woodland and Forest Campaign Metadata: Do you recommend other fields for this section?
Any additional comments?

Page 16



Field Spectroscopy Metadata

14. AGRICULTURE CAMPAIGN METADATA

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

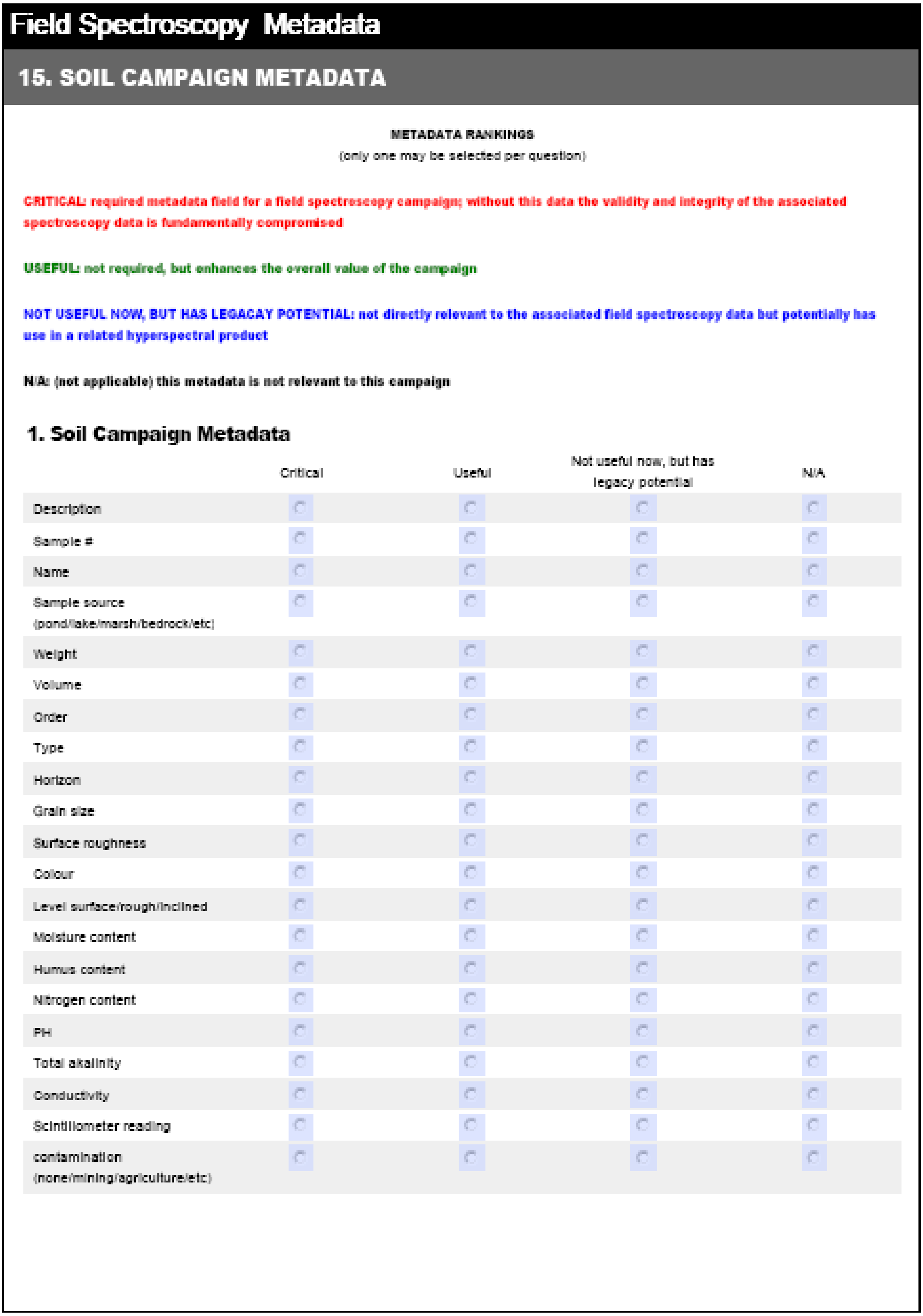
N/A: (not applicable) this metadata is not relevant to this campaign

1. Agriculture Campaign Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Common name	○	○	○	○
Species	○	○	○	○
Crop description	○	○	○	○
Growth stage	○	○	○	○
Crop height	○	○	○	○
Row orientation	○	○	○	○
Row spacing	○	○	○	○
Recent treatment (fertilizer/pesticide/water/other)	○	○	○	○
Background (soil / other)	○	○	○	○

2. Agriculture Campaign Metadata: Do you recommend other fields for this section? Any additional comments?

Page 17



Field Spectroscopy Metadata

15. SOIL CAMPAIGN METADATA

METADATA RANKINGS
(only one may be selected per question)

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

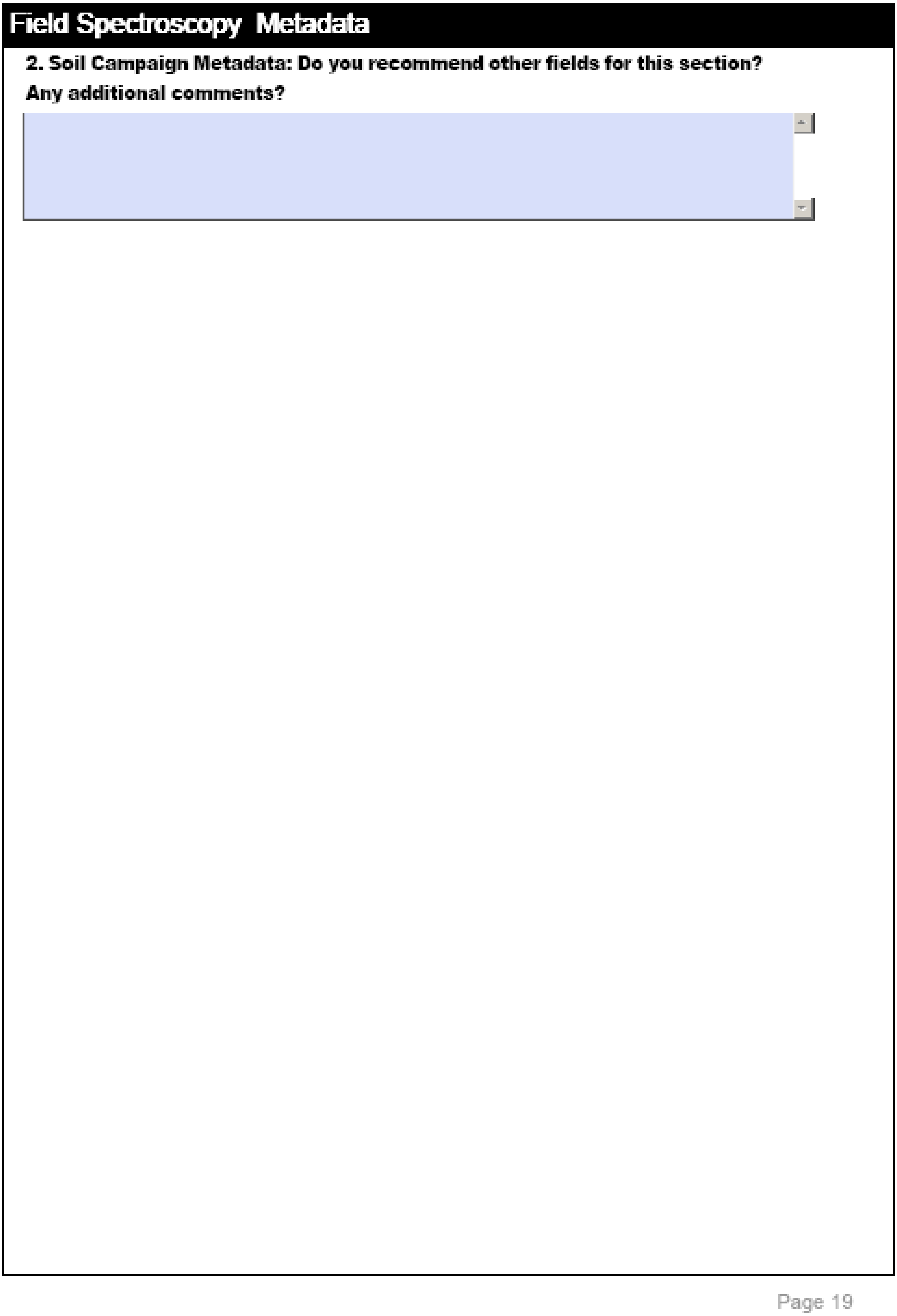
NOT USEFUL NOW, BUT HAS LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

N/A: (not applicable) this metadata is not relevant to this campaign

1. Soil Campaign Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Description	○	○	○	○
Sample #	○	○	○	○
Name	○	○	○	○
Sample source (pond/lake/marsh/bedrock/etc)	○	○	○	○
Weight	○	○	○	○
Volume	○	○	○	○
Order	○	○	○	○
Type	○	○	○	○
Horizon	○	○	○	○
Grain size	○	○	○	○
Surface roughness	○	○	○	○
Colour	○	○	○	○
Level surface/rough/inclined	○	○	○	○
Moisture content	○	○	○	○
Humus content	○	○	○	○
Nitrogen content	○	○	○	○
PH	○	○	○	○
Total akalinity	○	○	○	○
Conductivity	○	○	○	○
Scintillometer reading	○	○	○	○
contamination (none/mining/agriculture/etc)	○	○	○	○

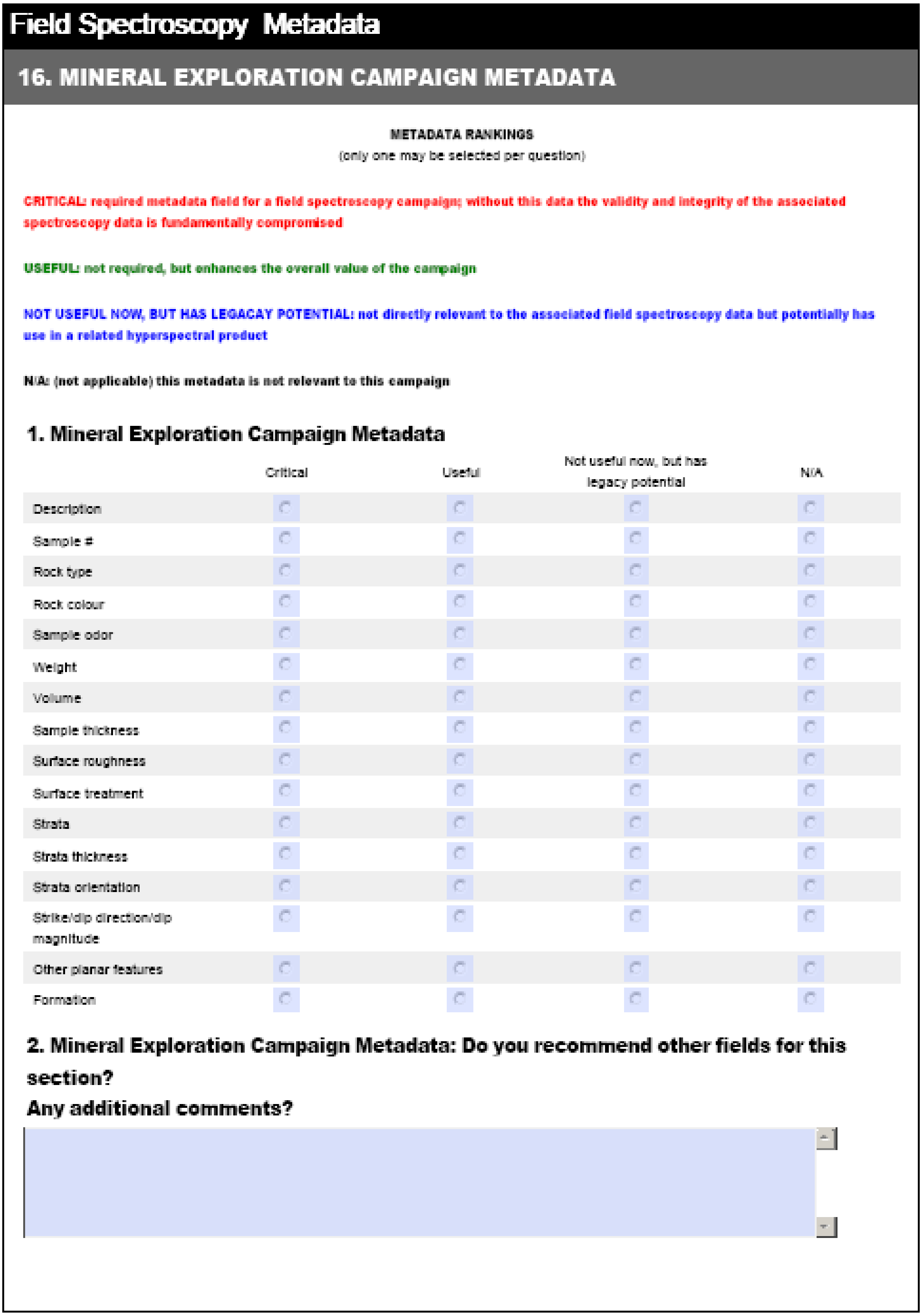
Page 18



Field Spectroscopy Metadata

2. Soil Campaign Metadata: Do you recommend other fields for this section? Any additional comments?

Page 19



Field Spectroscopy Metadata

16. MINERAL EXPLORATION CAMPAIGN METADATA

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

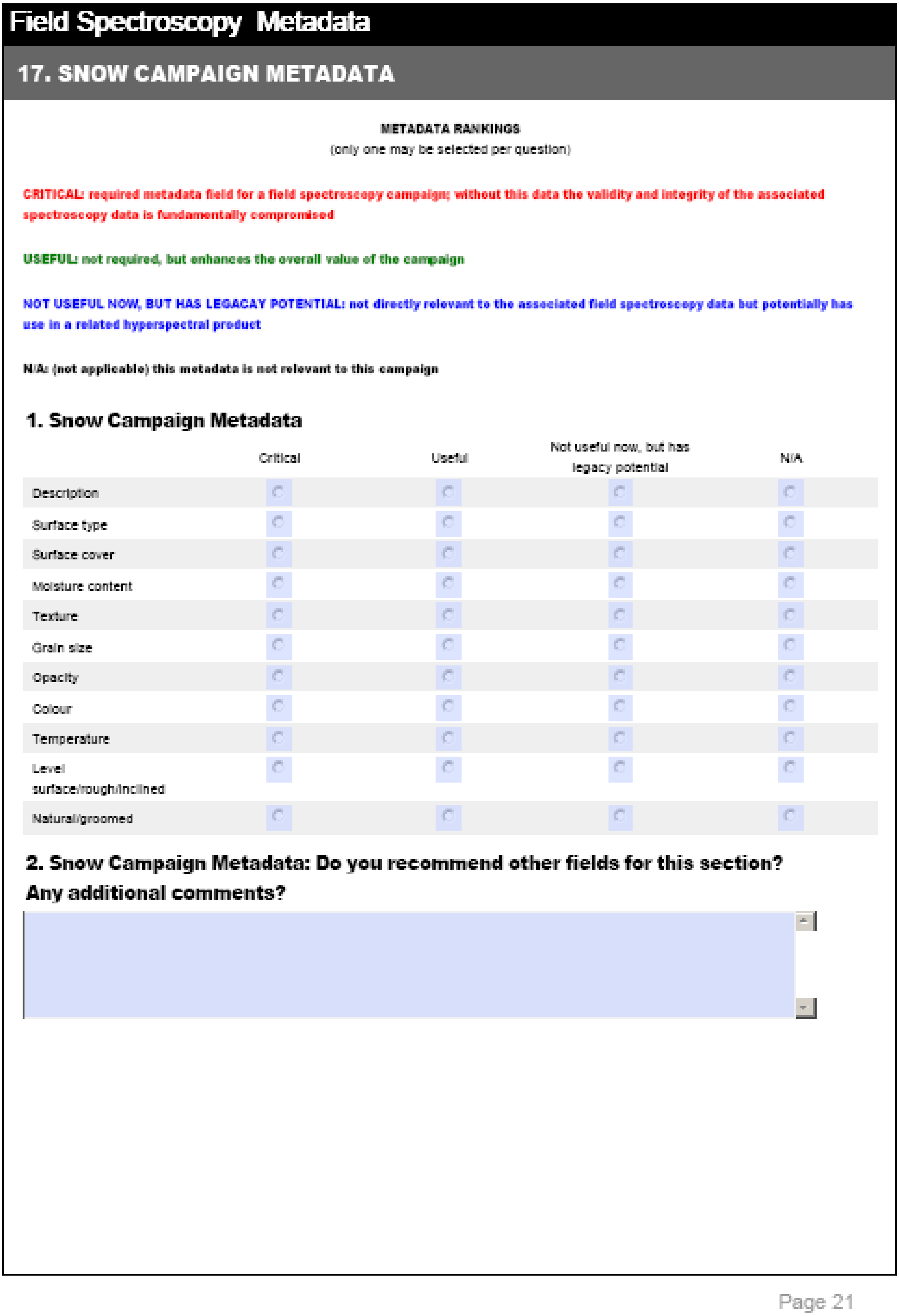
N/A: (not applicable) this metadata is not relevant to this campaign

1. Mineral Exploration Campaign Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Description	○	○	○	○
Sample #	○	○	○	○
Rock type	○	○	○	○
Rock colour	○	○	○	○
Sample odor	○	○	○	○
Weight	○	○	○	○
Volume	○	○	○	○
Sample thickness	○	○	○	○
Surface roughness	○	○	○	○
Surface treatment	○	○	○	○
Strata	○	○	○	○
Strata thickness	○	○	○	○
Strata orientation	○	○	○	○
Strike/dip direction/dip magnitude	○	○	○	○
Other planar features	○	○	○	○
Formation	○	○	○	○

2. Mineral Exploration Campaign Metadata: Do you recommend other fields for this section?
Any additional comments?

Page 20



Field Spectroscopy Metadata

17. SNOW CAMPAIGN METADATA

METADATA RANKINGS
(only one may be selected per question)

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 LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

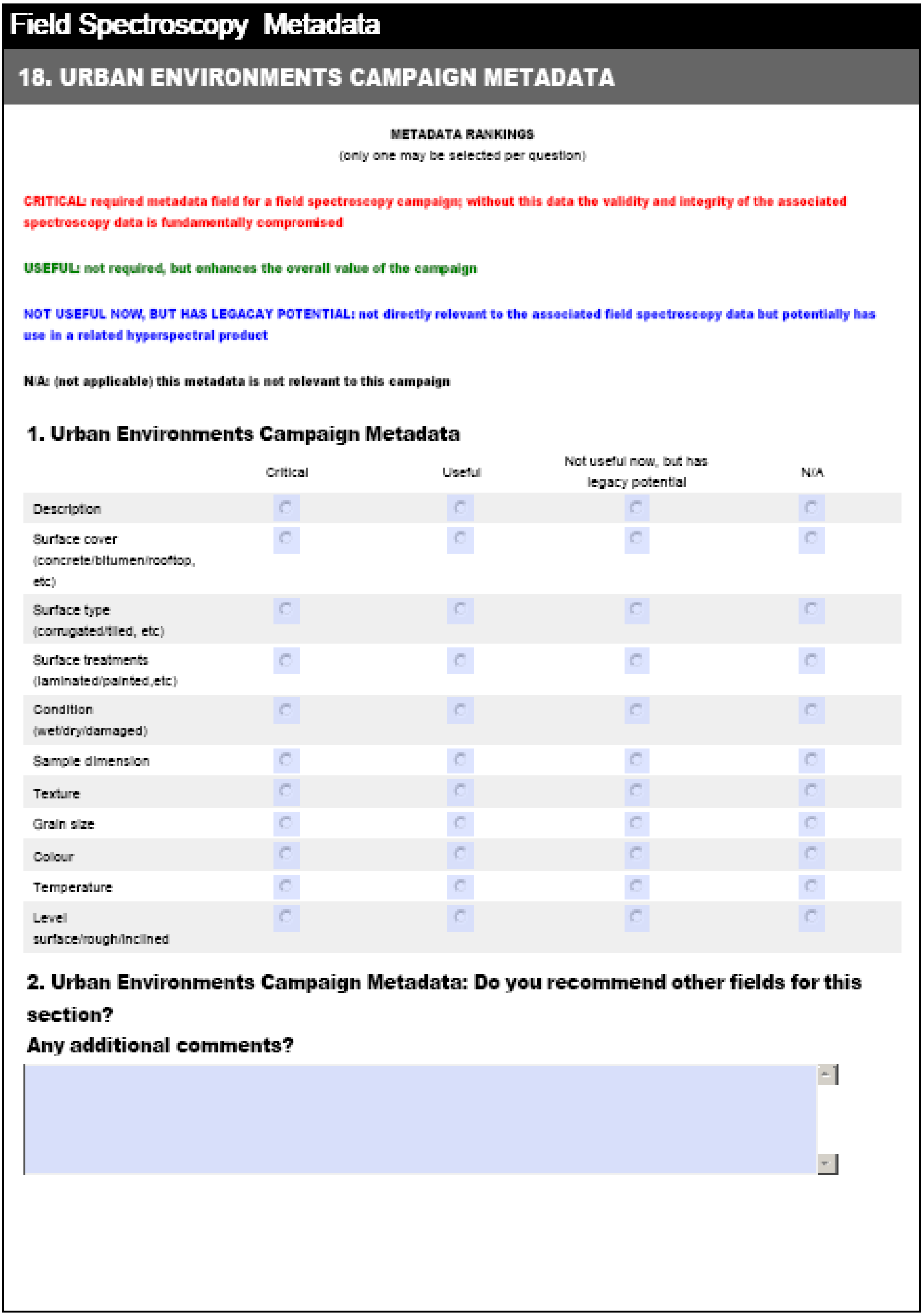
N/A: (not applicable) this metadata is not relevant to this campaign

1. Snow Campaign Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Description	○	○	○	○
Surface type	○	○	○	○
Surface cover	○	○	○	○
Moisture content	○	○	○	○
Texture	○	○	○	○
Grain size	○	○	○	○
Opacity	○	○	○	○
Colour	○	○	○	○
Temperature	○	○	○	○
Level surface/rough/inclined	○	○	○	○
Natural/groomed	○	○	○	○

2. Snow Campaign Metadata: Do you recommend other fields for this section? Any additional comments?

Page 21



Field Spectroscopy Metadata

18. URBAN ENVIRONMENTS CAMPAIGN METADATA

METADATA RANKINGS
(only one may be selected per question)

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

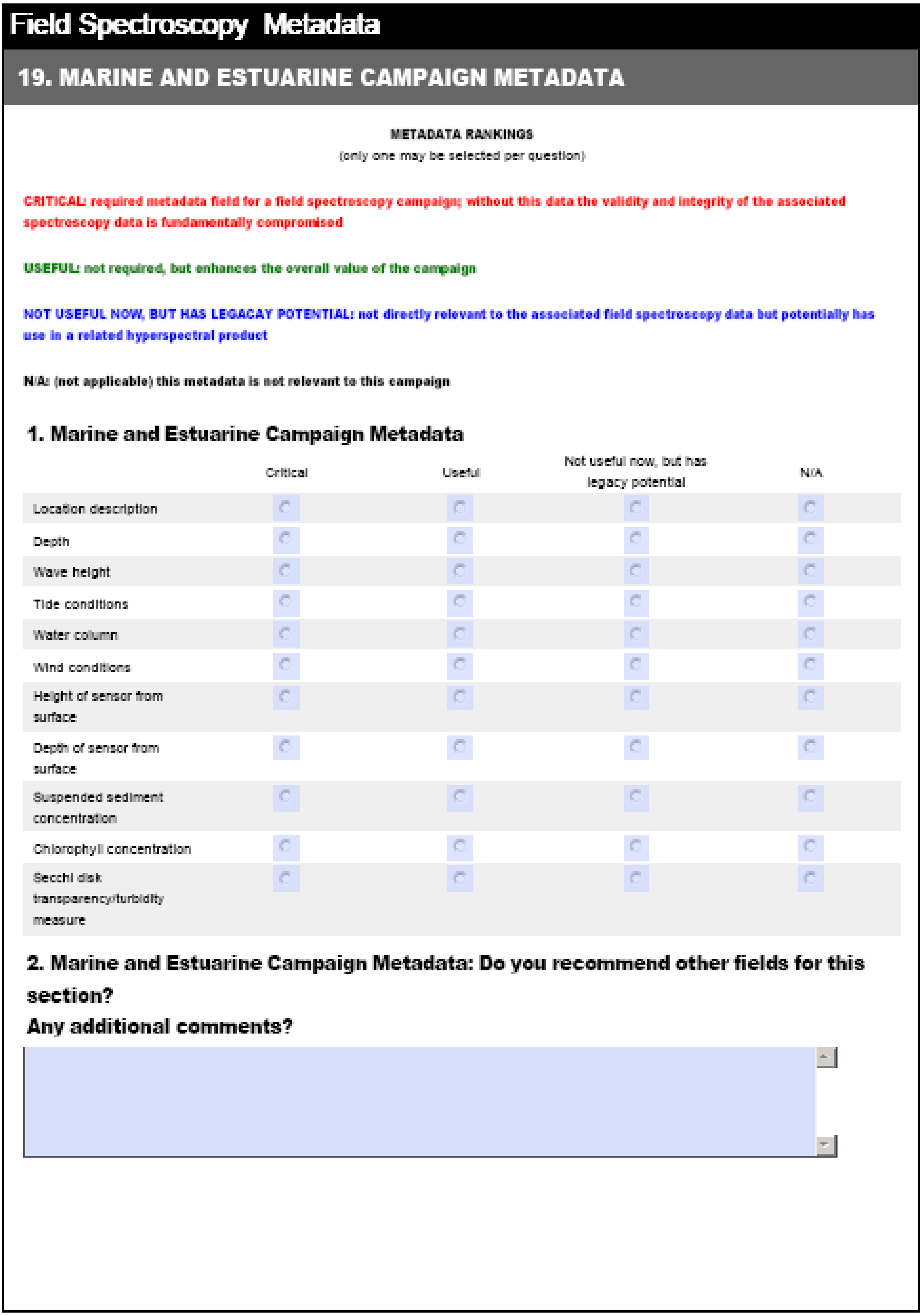
NOT USEFUL NOW, BUT HAS LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

N/A: (not applicable) this metadata is not relevant to this campaign

1. Urban Environments Campaign Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Description	○	○	○	○
Surface cover (concrete/bitumen/rooftop, etc)	○	○	○	○
Surface type (corrugated/tiled, etc)	○	○	○	○
Surface treatments (laminated/painted,etc)	○	○	○	○
Condition (wet/dry/damaged)	○	○	○	○
Sample dimension	○	○	○	○
Texture	○	○	○	○
Grain size	○	○	○	○
Colour	○	○	○	○
Temperature	○	○	○	○
Level surface/rough/inclined	○	○	○	○

2. Urban Environments Campaign Metadata: Do you recommend other fields for this section?
Any additional comments?



Field Spectroscopy Metadata

19. MARINE AND ESTUARINE CAMPAIGN METADATA

METADATA RANKINGS
(only one may be selected per question)

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

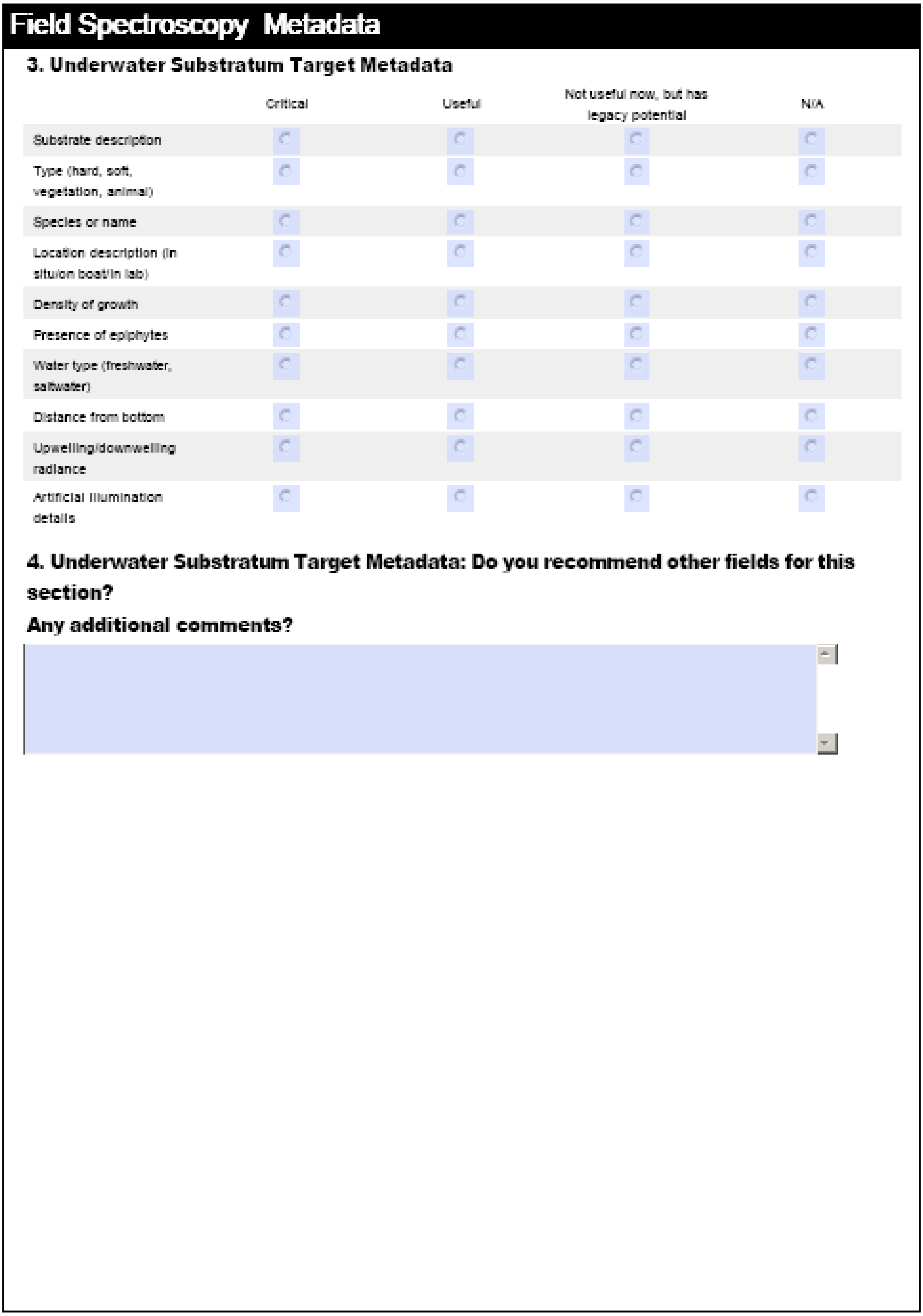
NOT USEFUL NOW, BUT HAS LEGACAY POTENTIAL: not directly relevant to the associated field spectroscopy data but potentially has use in a related hyperspectral product

N/A: (not applicable) this metadata is not relevant to this campaign

1. Marine and Estuarine Campaign Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Location description	○	○	○	○
Depth	○	○	○	○
Wave height	○	○	○	○
Tide conditions	○	○	○	○
Water column	○	○	○	○
Wind conditions	○	○	○	○
Height of sensor from surface	○	○	○	○
Depth of sensor from surface	○	○	○	○
Suspended sediment concentration	○	○	○	○
Chlorophyll concentration	○	○	○	○
Secchi disk transparency/turbidity measure	○	○	○	○

2. Marine and Estuarine Campaign Metadata: Do you recommend other fields for this section?
Any additional comments?



Field Spectroscopy Metadata

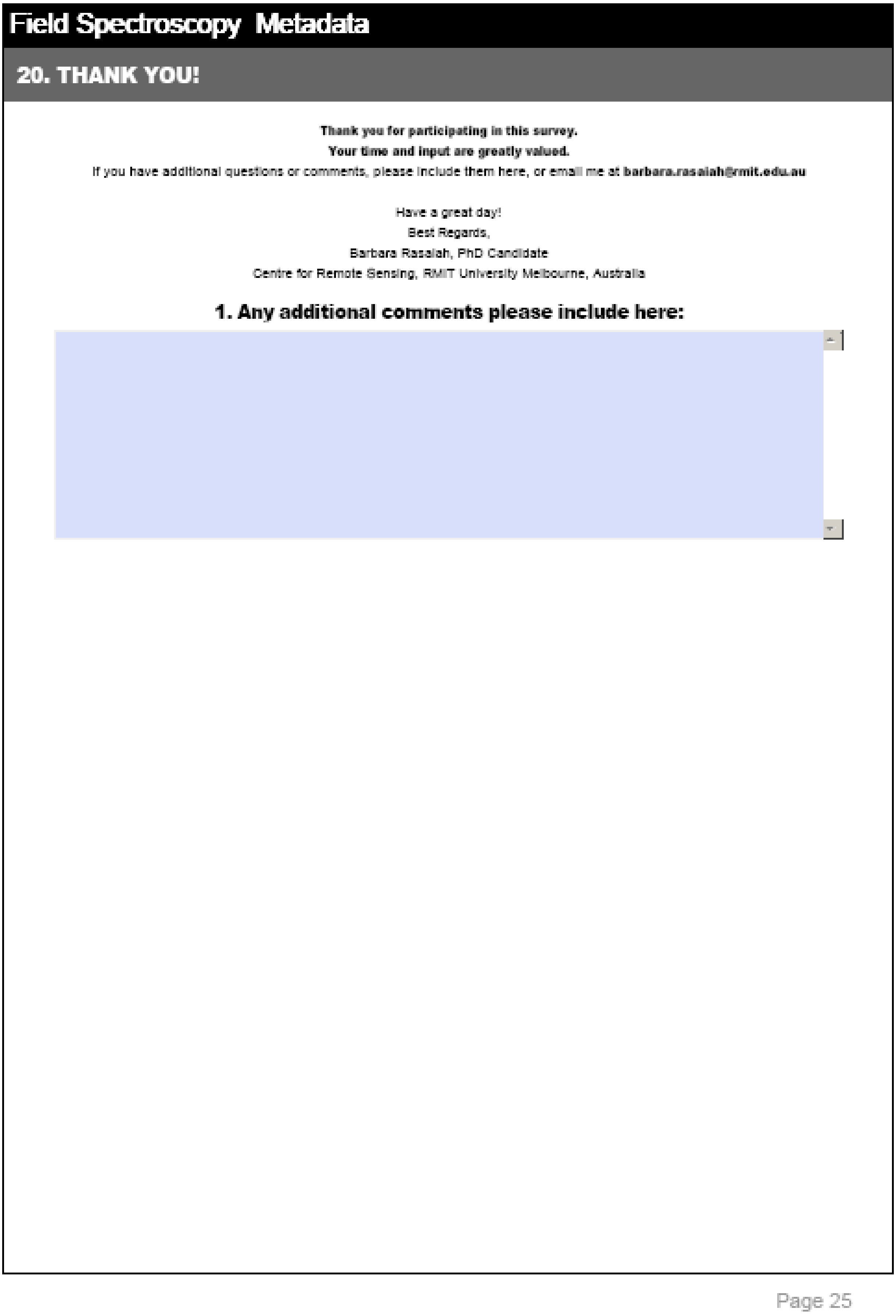
3. Underwater Substratum Target Metadata

	Critical	Useful	Not useful now, but has legacy potential	N/A
Substrate description	○	○	○	○
Type (hard, soft, vegetation, animal)	○	○	○	○
Species or name	○	○	○	○
Location description (in situ/on boat/in lab)	○	○	○	○
Density of growth	○	○	○	○
Presence of epiphytes	○	○	○	○
Water type (freshwater, saltwater)	○	○	○	○
Distance from bottom	○	○	○	○
Upwelling/downwelling radiance	○	○	○	○
Artificial illumination details	○	○	○	○

4. Underwater Substratum Target Metadata: Do you recommend other fields for this section?

Any additional comments?

Page 24



Field Spectroscopy Metadata

20. THANK YOU!

Thank you for participating in this survey.
Your time and input are greatly valued.
If you have additional questions or comments, please include them here, or email me at **barbara.rasaiah@rmit.edu.au**

Have a great day!
Best Regards,
Barbara Rasaiah, PhD Candidate
Centre for Remote Sensing, RMIT University Melbourne, Australia

1. Any additional comments please include here:

Page 25

Appendix A.2 Survey Results

This section presents criticality ranking results for all metadata categories in the online field spectroscopy metadata survey.

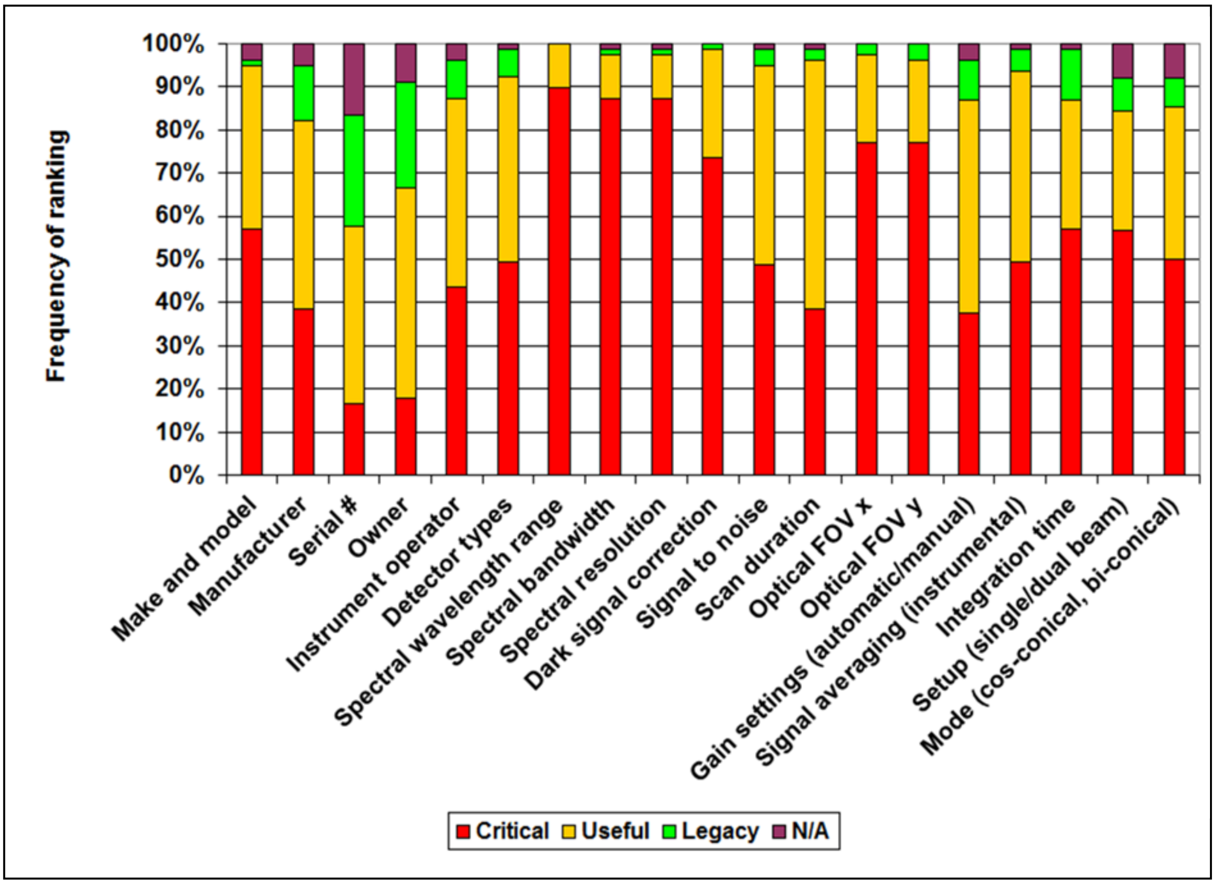


Figure A.1 Instrument (n=79)

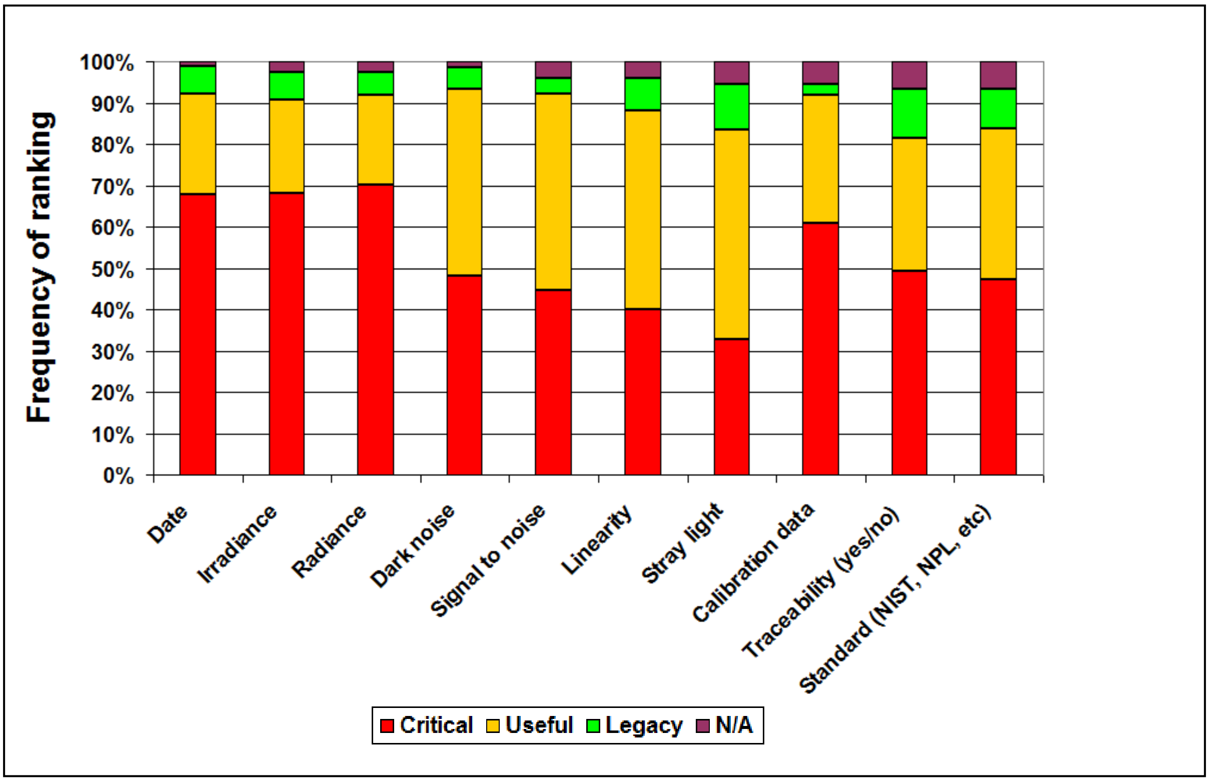


Figure A.2 Calibration (n=68)

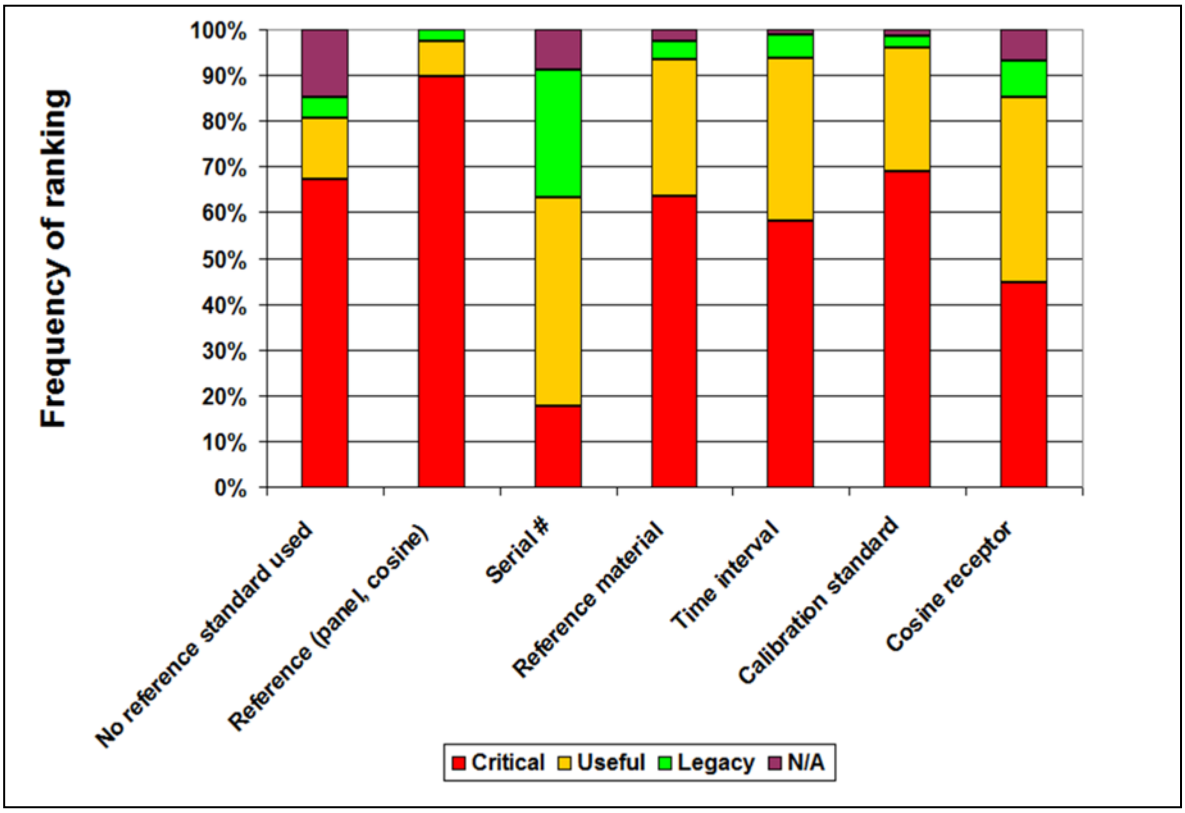


Figure A.3 Reference standards (n=79)

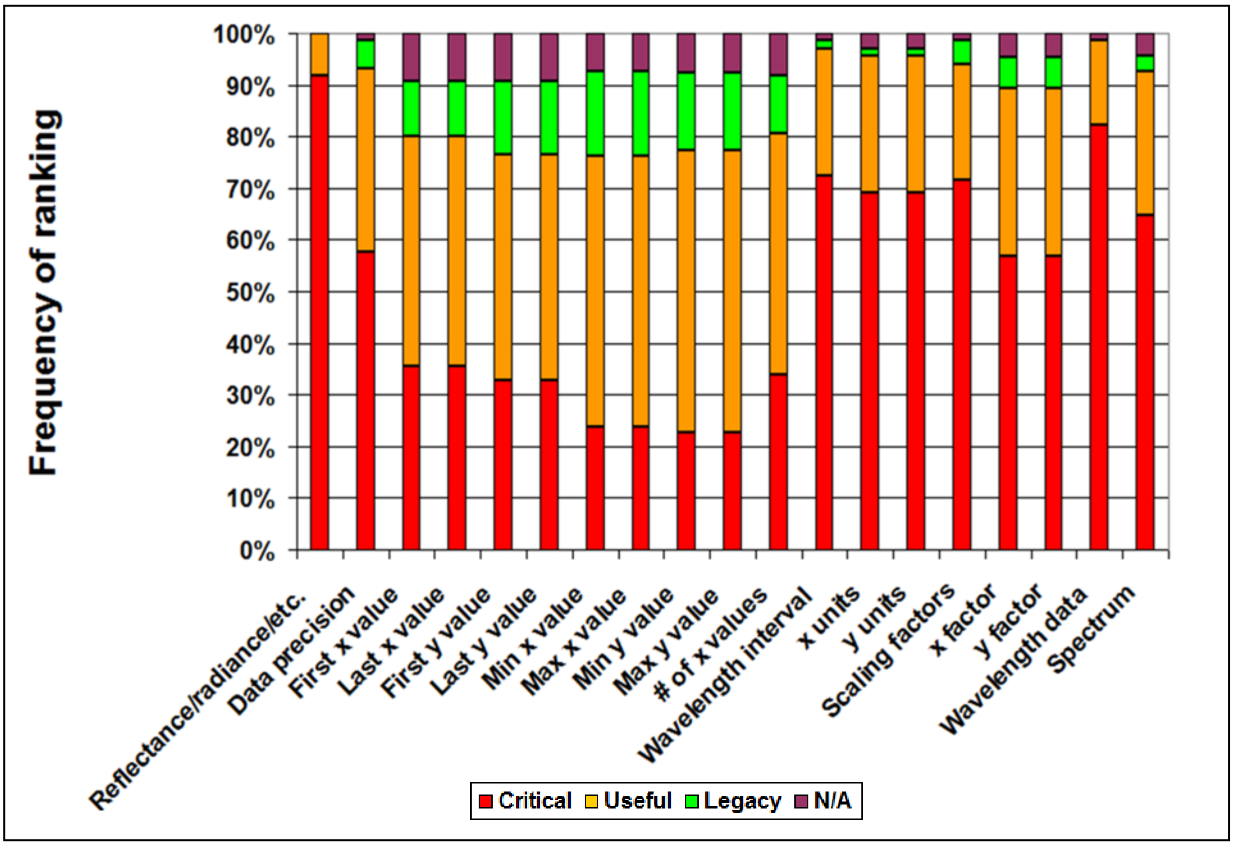


Figure A.4 Hyperspectral signal properties (n=73)

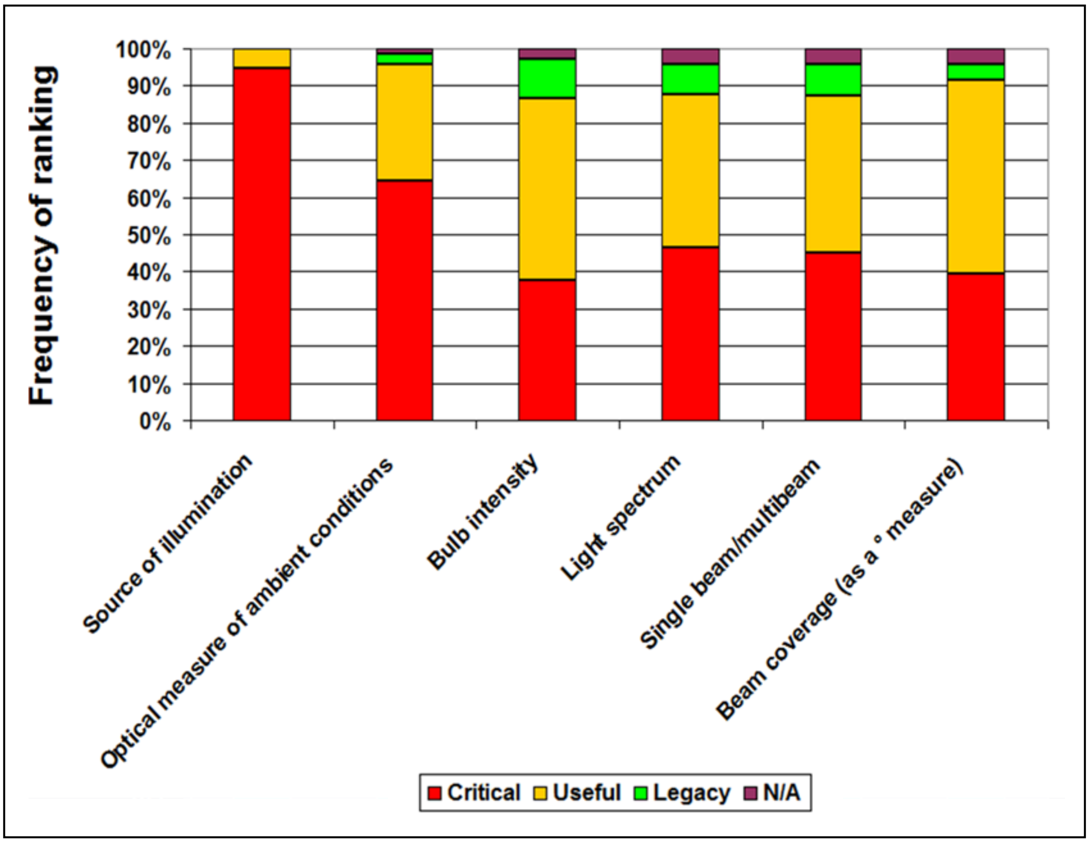


Figure A.5 Illumination information (n=75)

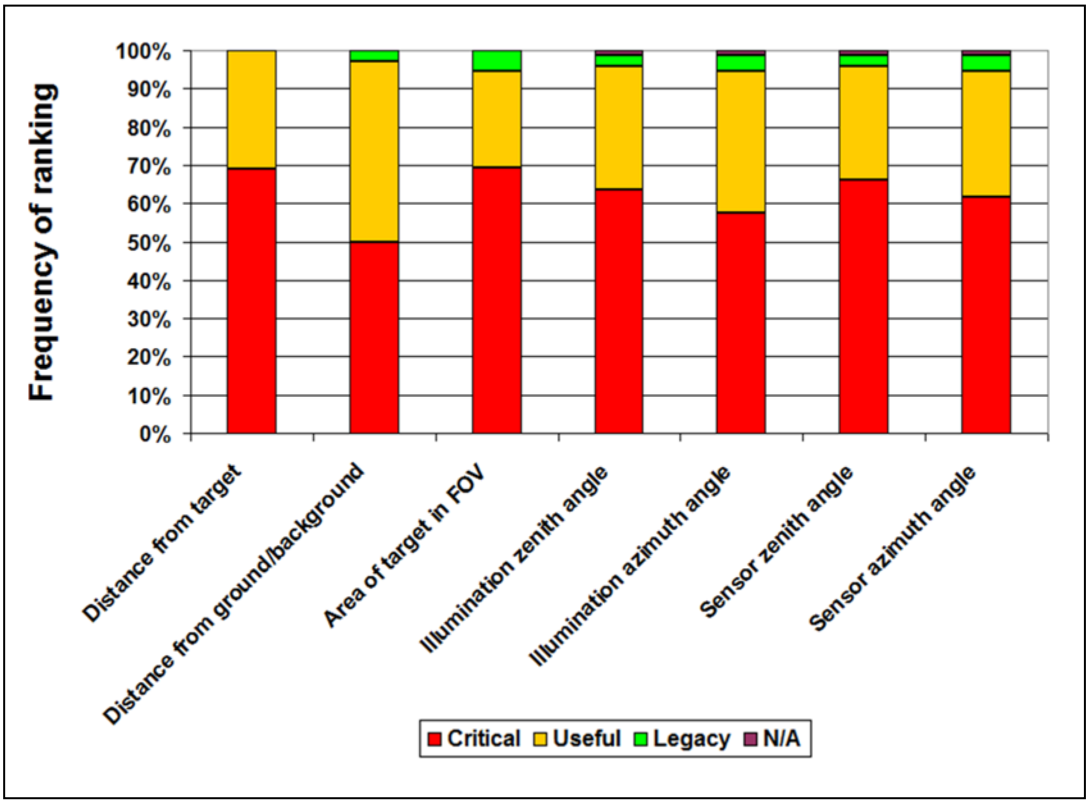


Figure A.6 Viewing geometry (n=74)

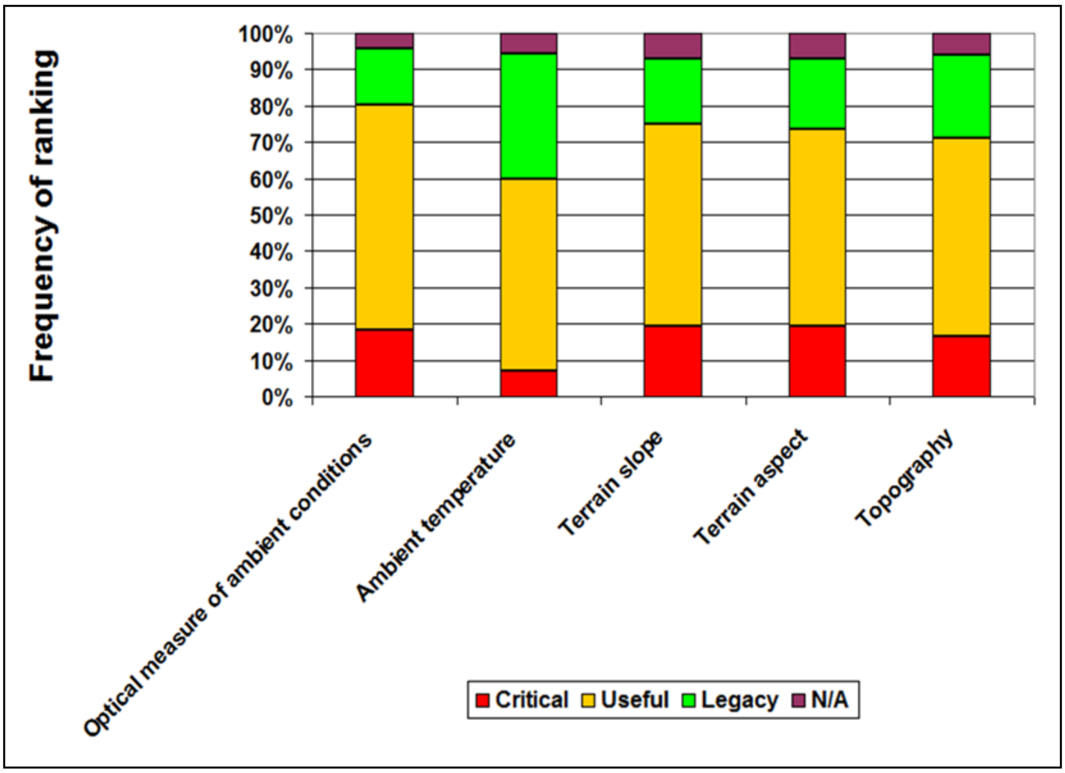


Figure A.7 Environment information (n=72)

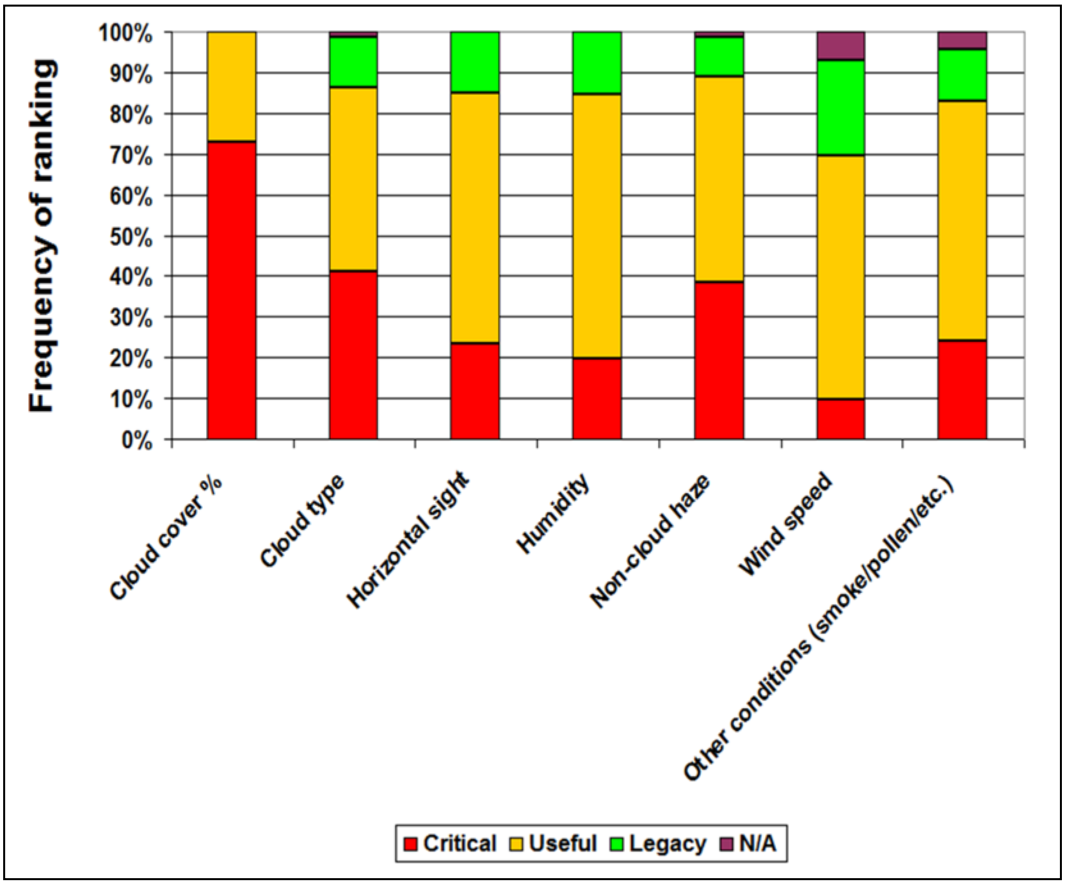


Figure A.8 Atmospheric conditions (n=74)

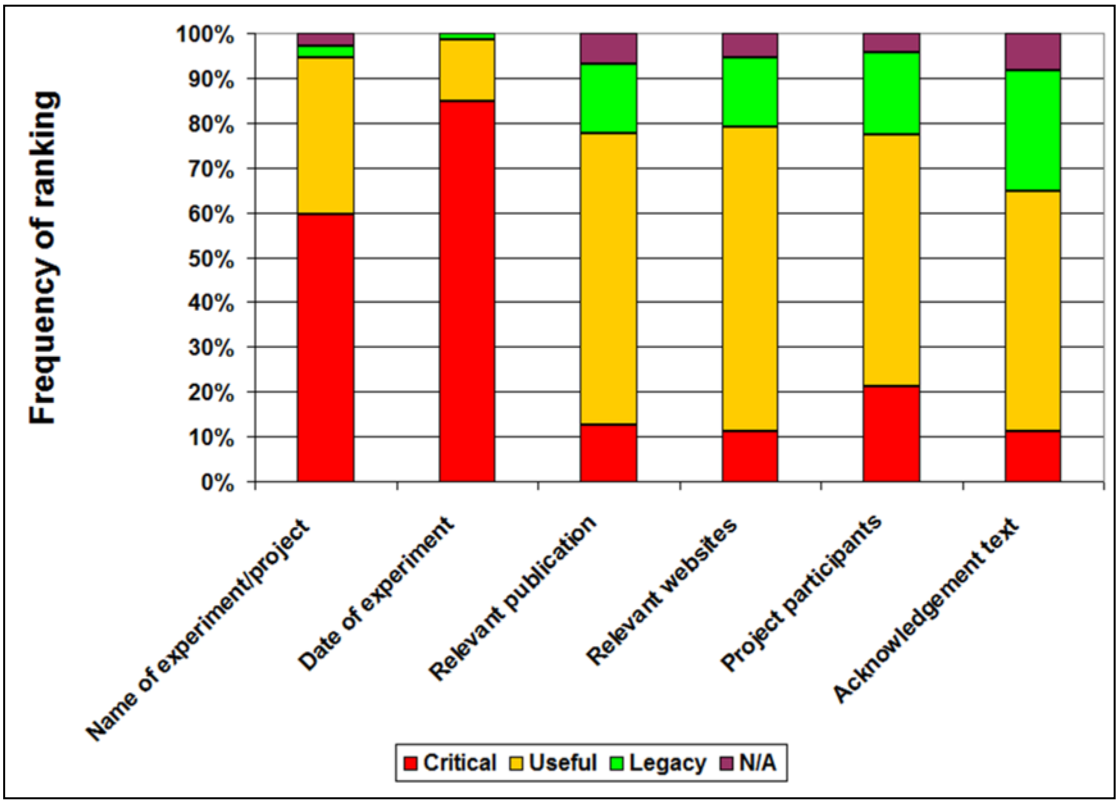


Figure A.9 General project information (n=73)

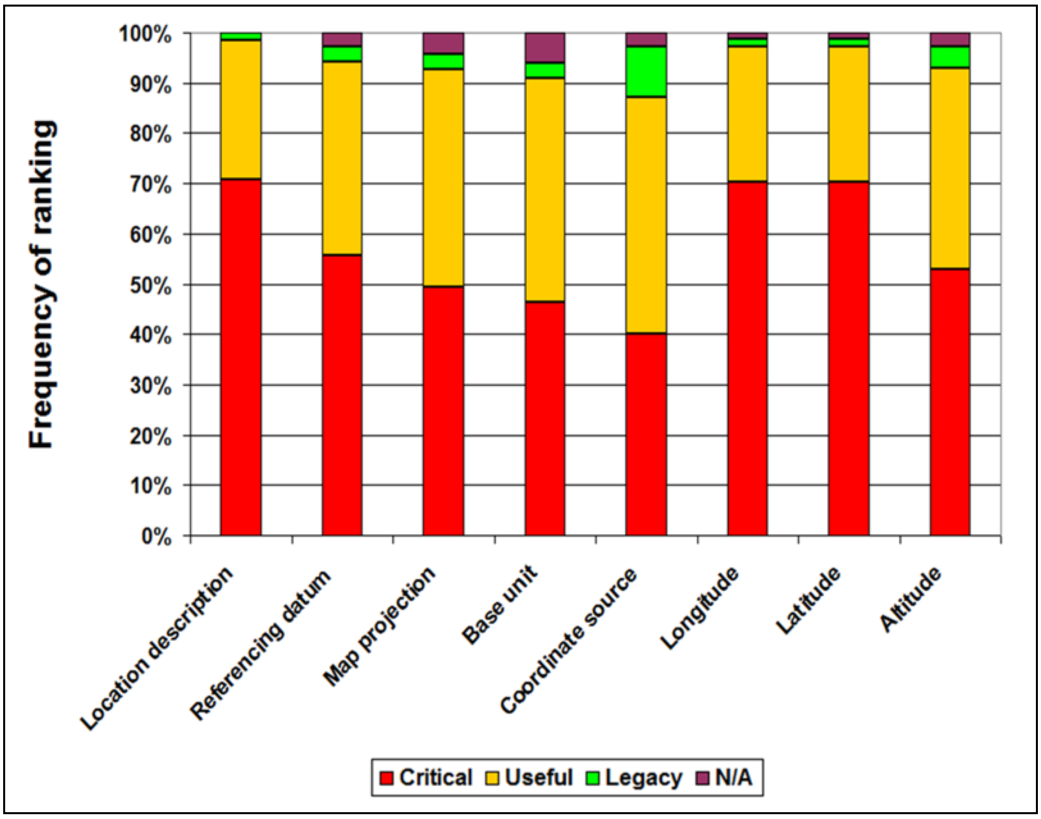


Figure A.10 Location information (n=74)

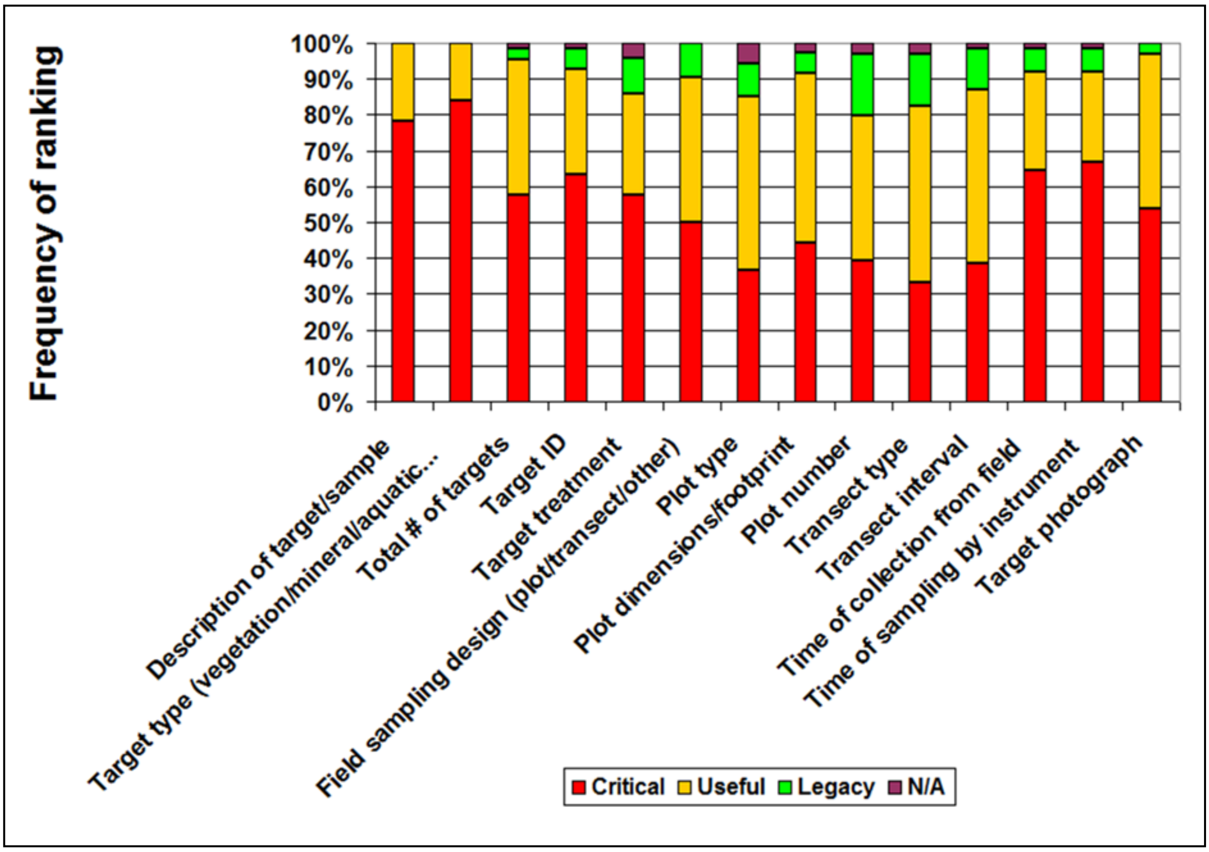


Figure A.11 General target and sampling (n=74)

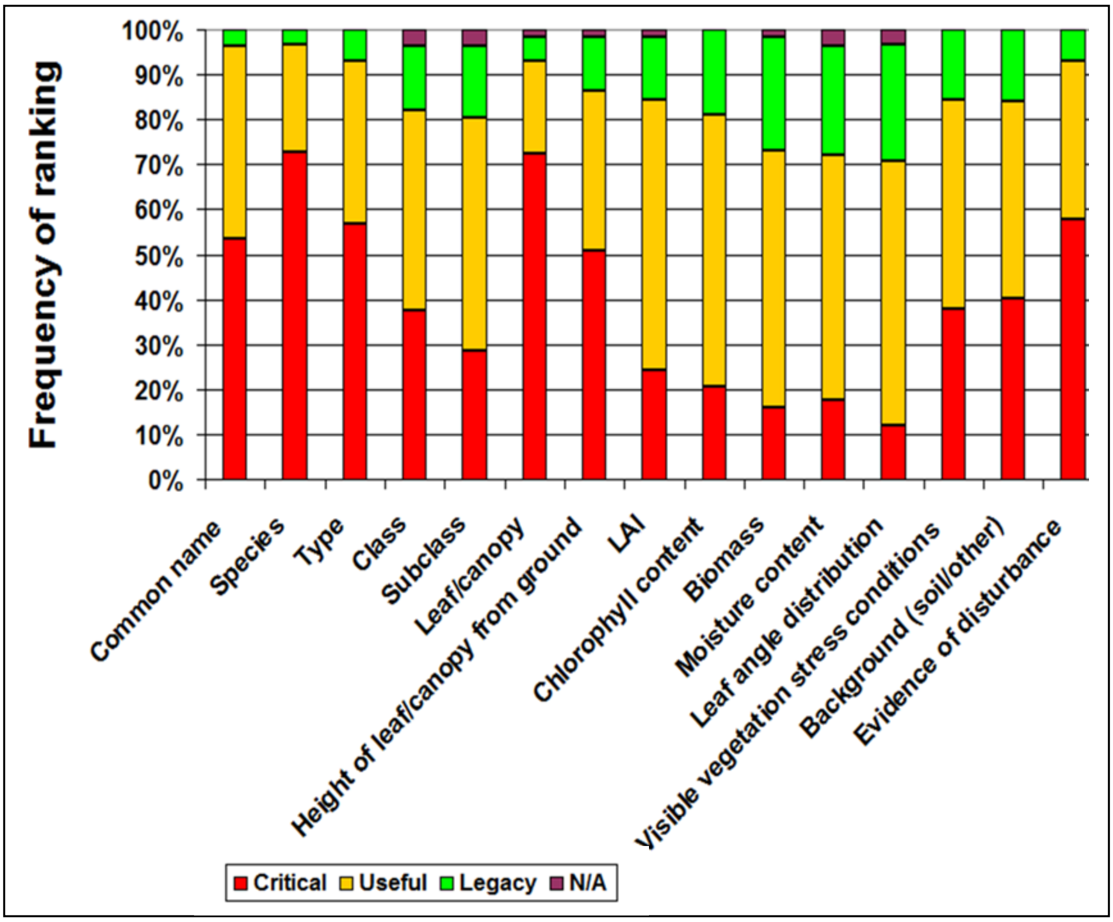


Figure A.12 Vegetation campaign (n=59)

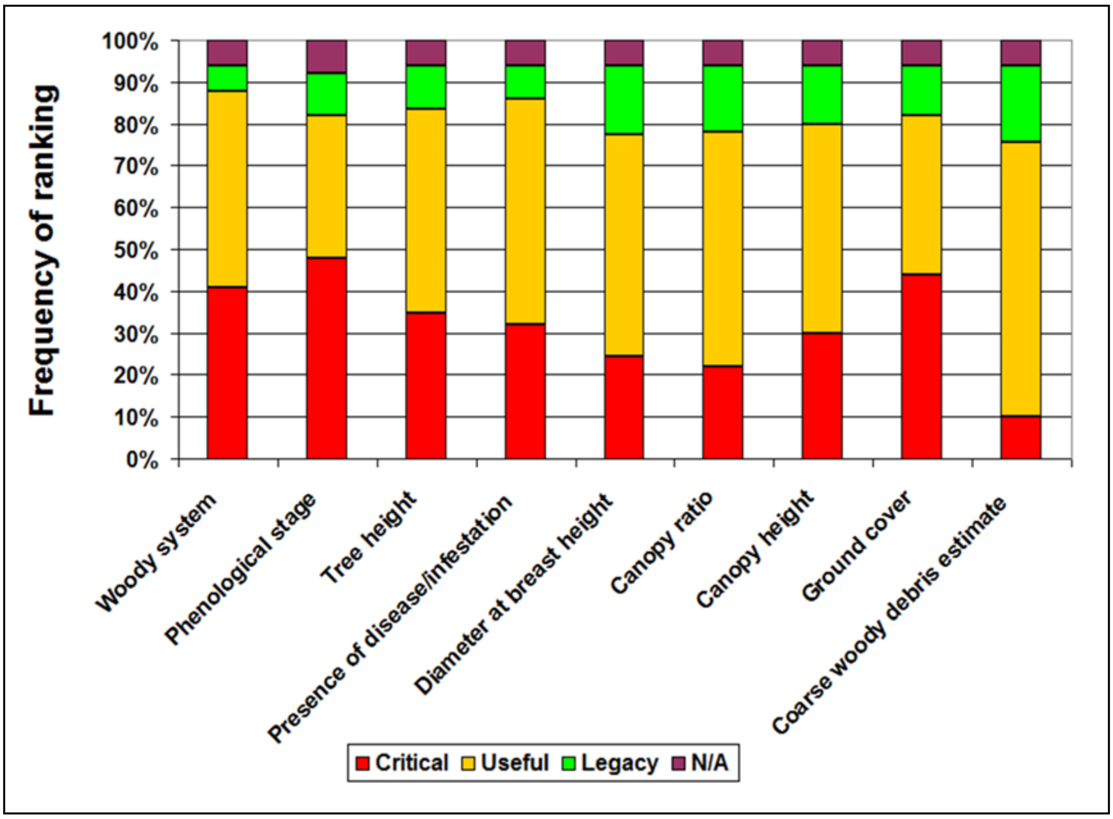


Figure A.13 Woodland and forest (n=50)

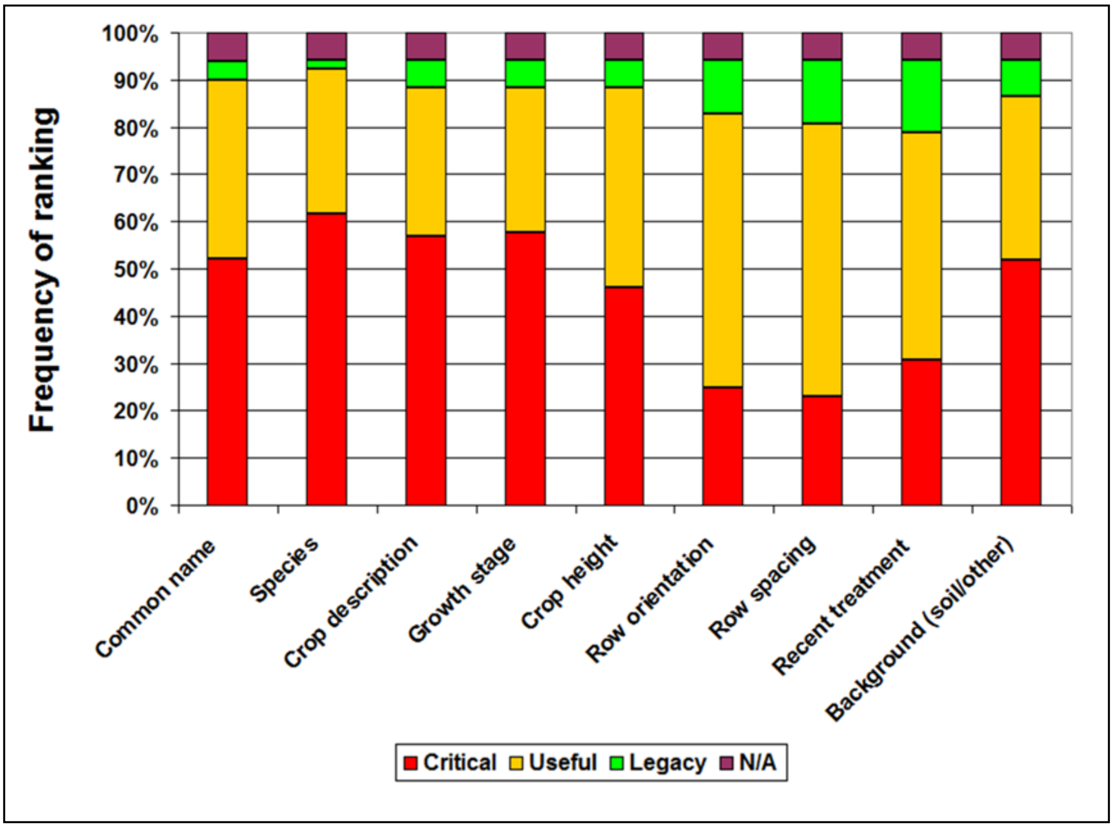


Figure A.14 Agriculture (n=52)

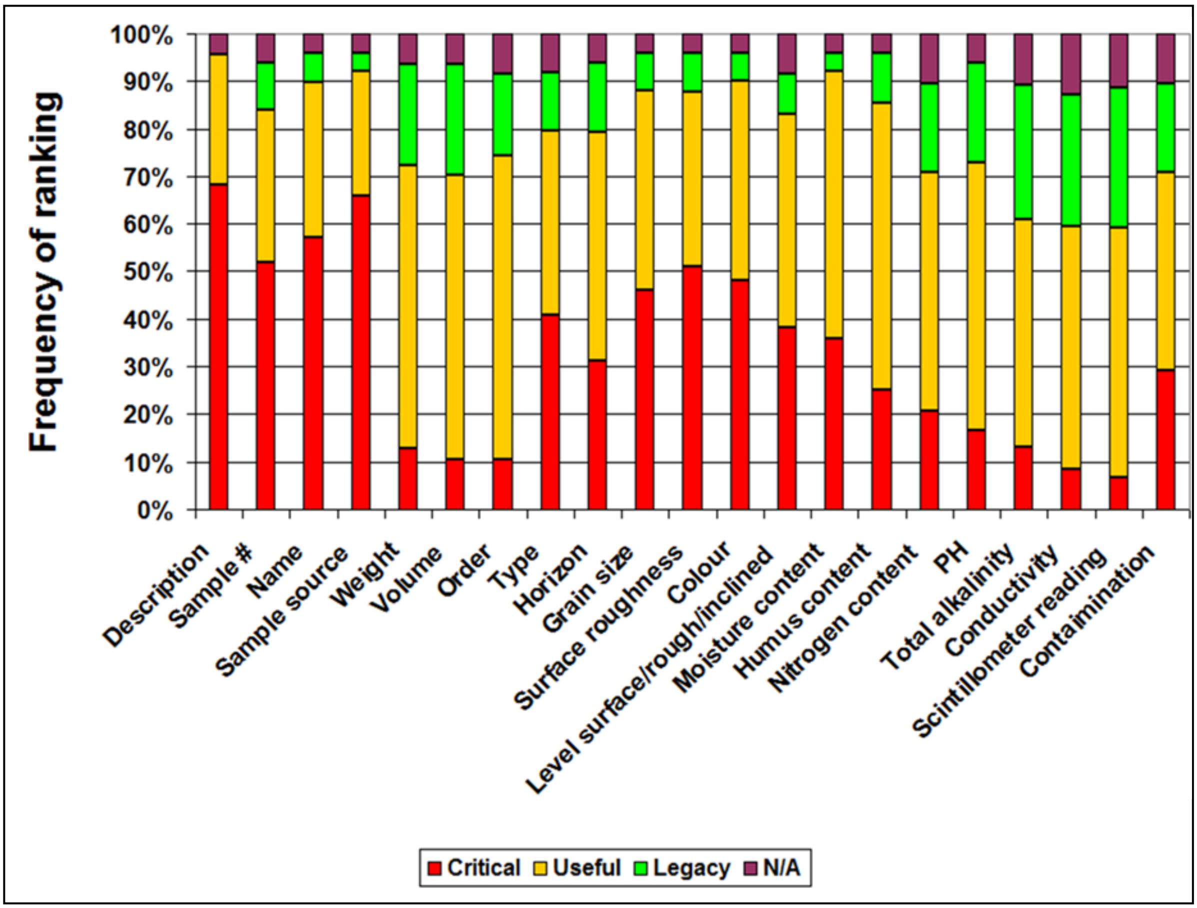


Figure A.15 Soil (n=50)

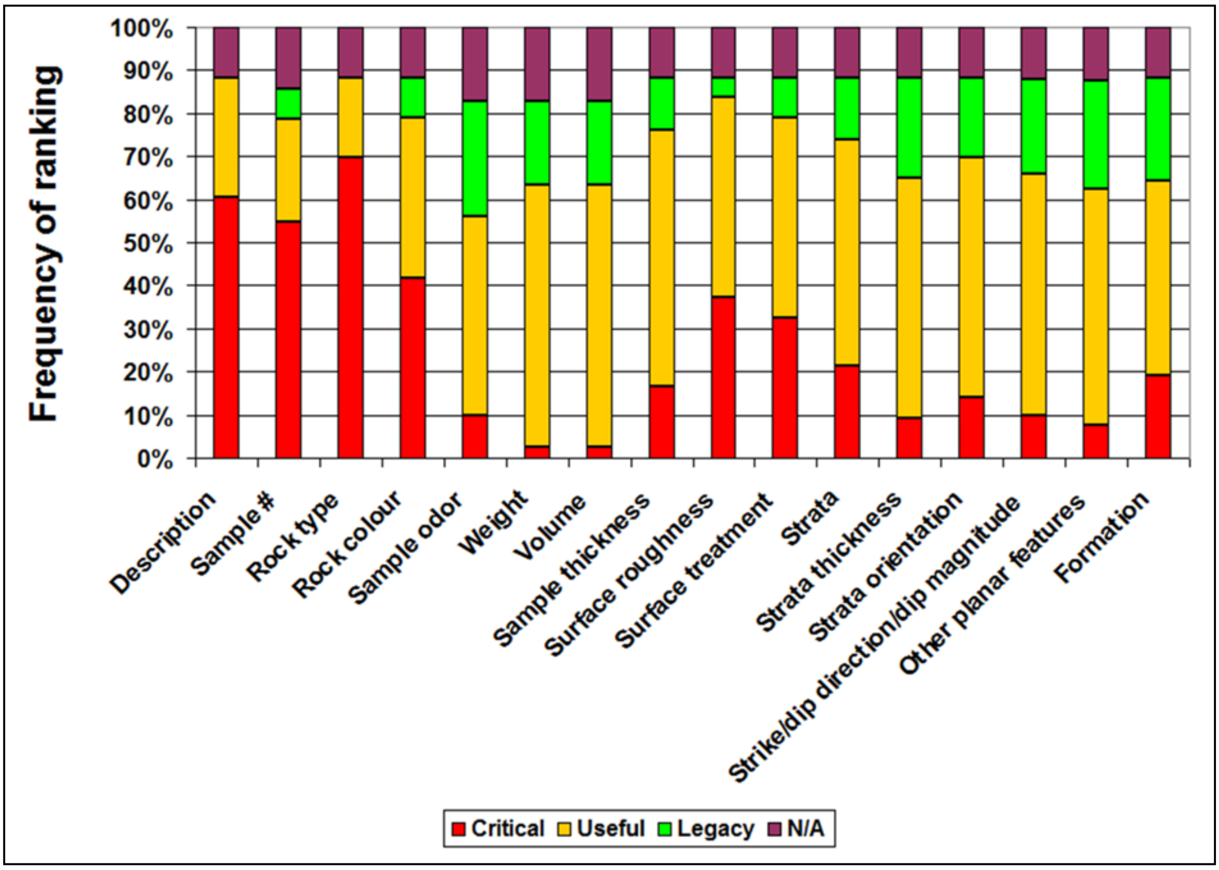


Figure A.16 Mineral exploration (n=43)

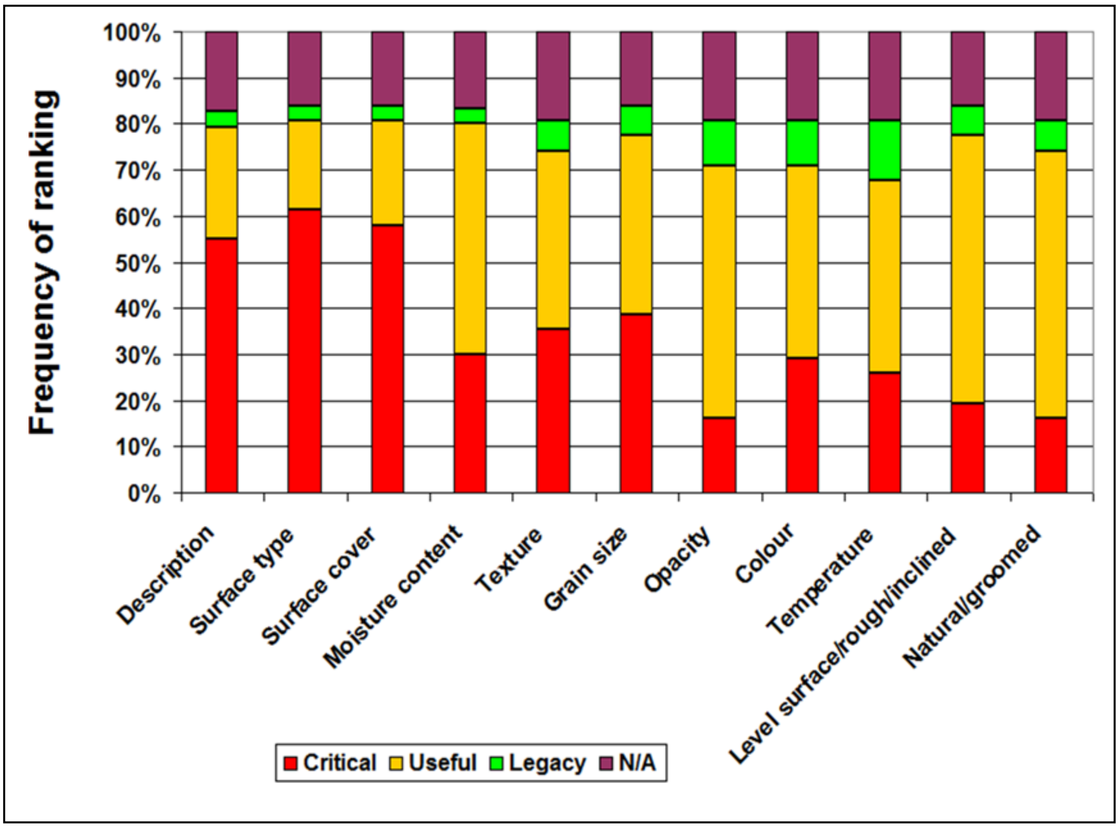


Figure A.17 Snow (n=31)

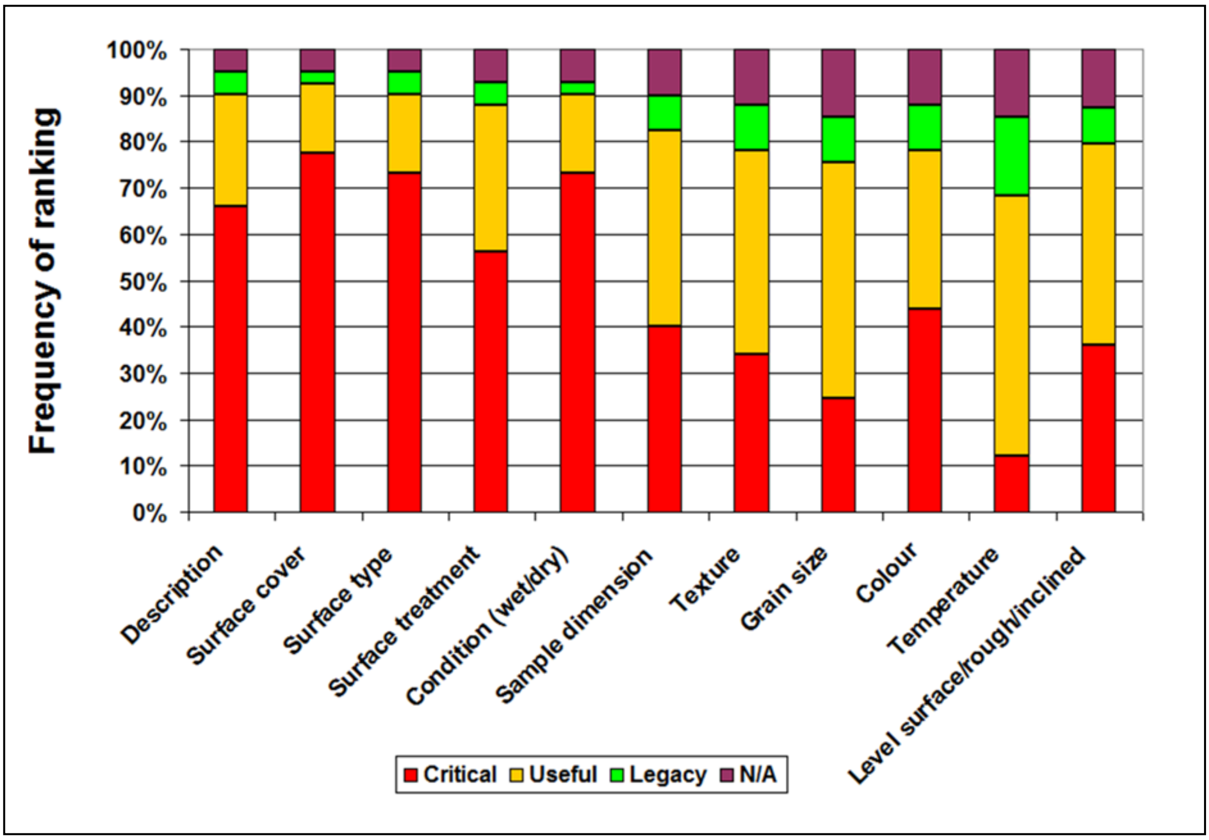


Figure A.18 Urban environments (n=41)

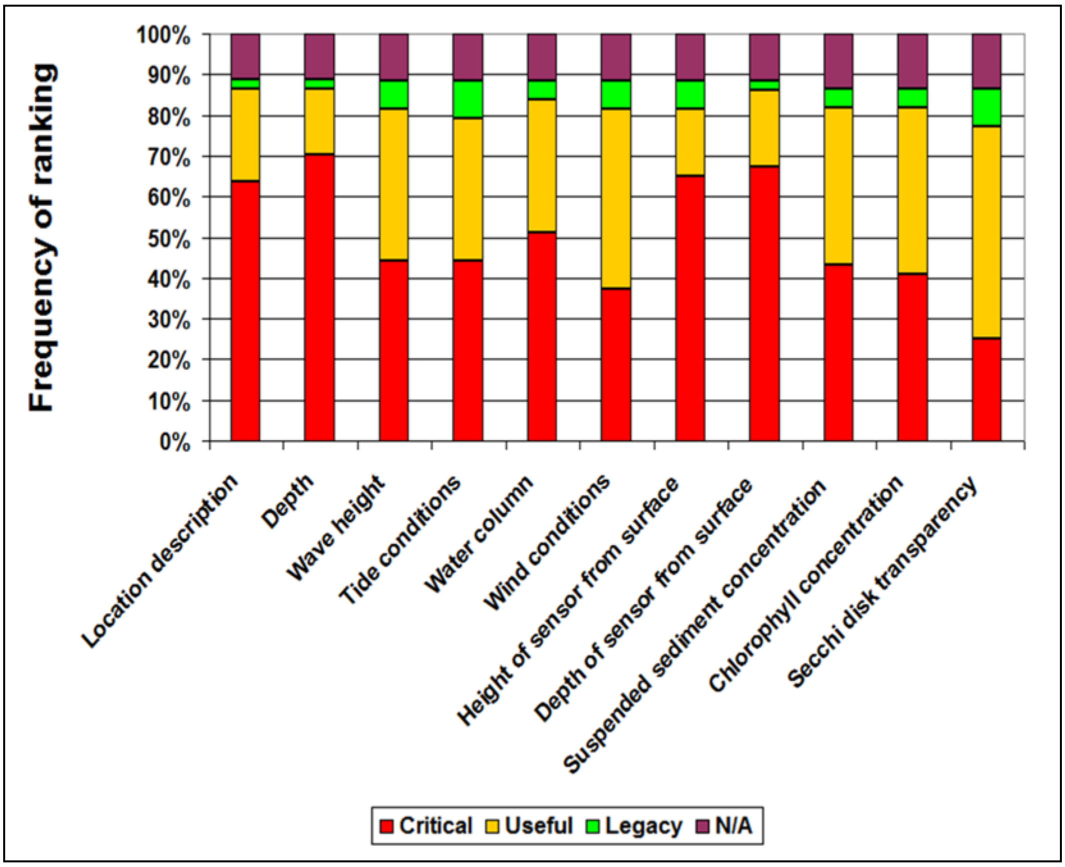


Figure A.19 Marine and estuarine (n=44)

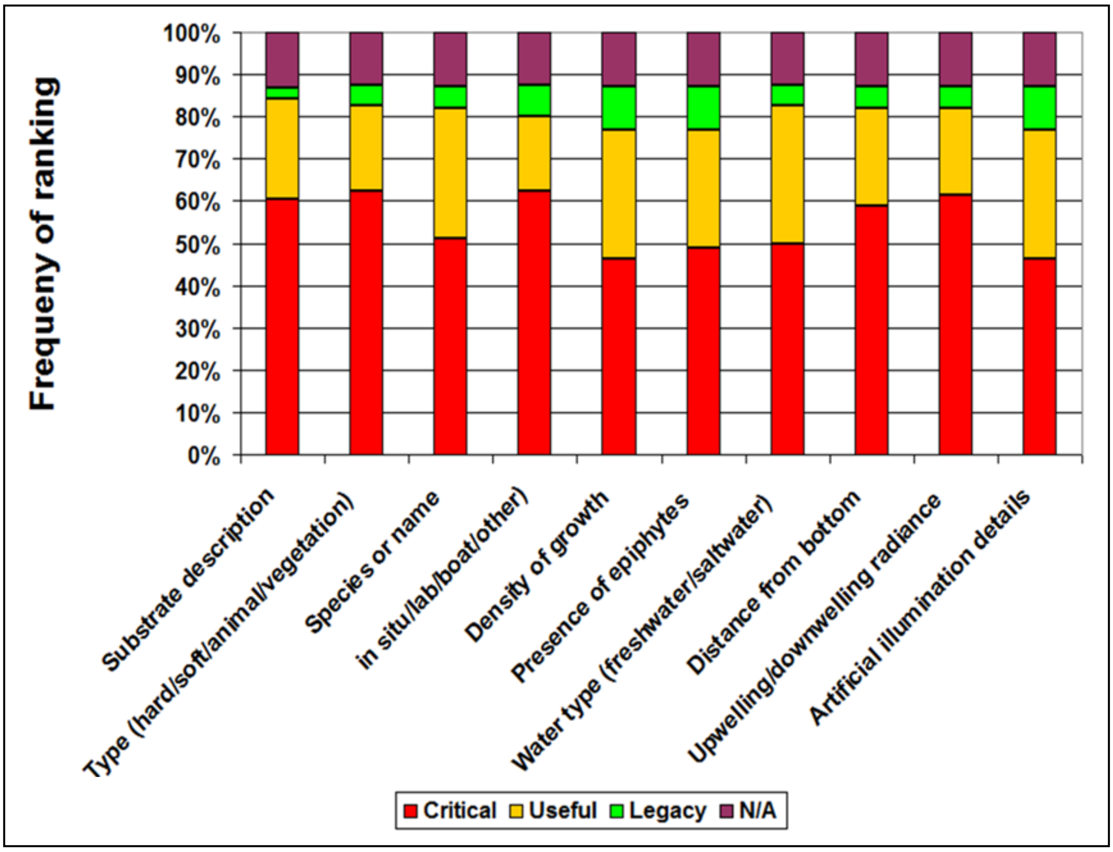


Figure A.20 Underwater substratum target (n=40)

Appendix B Spectral Library Workshop Outcomes

This section presents the attendees of the spectral library workshops that served to inform the research in Chapters 4, 5, 6, the application-specific metadatasets derived from the workshops, and the mappings from the seven examined metadata standards to the Core metadataset and the application-specific metadataset.

Appendix B.1

This section presents the attendees of the spectral libraries workshops held in Australia in 2012.

Name	Institution	Role / Expertise
Tim Malthus	CSIRO Division of Land and Water, Canberra	PI, field spectroscopy, calibration and validation
John Gamon	University of Alberta, Canada	Convenor of SpecNet community
Phil Townsend	University of Wisconsin, USA	Vegetation spectroscopy
Chris MacLellan	NERC Field Spectroscopy Facility, University of Edinburgh, UK	Calibration and validation
Andy Hueni	RSL, University of Zurich, Switzerland	Writer of SPECCHIO software
Alfredo Huete	University of Technology Sydney	Spectroscopy for phenological studies
Laurie Chisholm	University of Wollongong	Field spectroscopy
Simon Jones	Royal Melbourne Institute of Technology	Vegetation spectroscopy
Stuart Phinn	University of Queensland	Terrestrial and aquatic spectroscopy
Cindy Ong	CSIRO Earth Science and Resource Engineering, Perth	Geological and mineral spectroscopy
Barbara Rasaiah	Royal Melbourne Institute of Technology	Metadata and informatics (PhD student)
Chris Roelfsema	University of Queensland	Aquatic spectroscopy
Lola Suarez	Royal Melbourne Institute of Technology	Remote sensing of vegetation
Rebecca Trevithick	Department of Science, Information Technology, Innovation and the Arts, Queensland	Informatics and data archiving
Matthew Wyatt	IVEC, Western Australia	Metadata and informatics
Carlos Aya	Intersect, NSW	Senior IT developer

Table B.1 Attendees of the TERN ACEAS 'Bio-optical data: Best practice and legacy datasets' workshop, held June 18-22 2012 in Brisbane, Australia, led by Dr. Tim Malthus

Name	Institution	Role / Expertise
Chris Bellman	RMIT University	Photogrammetry
Laurie Chisolm	University of Wollongong	Field spectroscopy
Robert Hewson	RMIT University	Remote sensing of vegetation
Andy Hueni	RSL, University of Zurich, Switzerland	Writer of SPECCHIO software
Simon Jones	RMIT University	Vegetation spectroscopy
Barbara Rasaiah	RMIT University	Metadata and informatics (PhD student)
Mariela Soto-Berelov	RMIT University	Remote sensing of land use change
Lola Suarez	RMIT University	Remote sensing of vegetation
Rebecca Trevithick	Department of Science, Information Technology, Innovation and the Arts, Queensland	Informatics and data archiving
Phil Wilkes	RMIT University	Remote sensing of vegetation (PhD student)
Will Woodgate	RMIT University	Remote sensing of vegetation (PhD student)

Table B.2 Attendees of the Spectral Libraries Workshop, Held December 10 2012 in Melbourne, Australia, hosted by RMIT University and led by Barbara Rasaiah

Appendix B.2

This section presents the application-specific metadatasets discussed in Chapter 4.

METADATA FIELD	REASON FOR INCLUSION / COMMENTS	OPTIONALITY	EXAMPLE	DATA TYPE
Collected within 1 week of aerial campaign	Minimizes any detectable changes in leaf phenology (this can be reference via a protocol citation)	Critical	yes	text
Position in canopy	Corresponds to visible canopy in an aerial hyperspectral campaign (this can be reference via a protocol citation)	Critical	Emergent leaves on top third of canopy	text
Illuminated leaves	(this can be reference via a protocol citation)	Critical	Yes	text
Target or scale (single leaf, branches, mature leaves, etc.)	Ensures consistent phenological state for all samples and sufficient leaf size for integrating sphere measurement (this can be reference via a protocol citation)	Critical	yes	Boolean
Tree species		Critical	Eucalyptus aquatica	text
Healthy leaves (absent of fungal or pest infection)	Permits most accurate leaf chemical analysis and spectral measurement	Optional	yes	Boolean
Tree ID	Used for correspondence to sample bags and spectra files	Optional	5885	text
Tag trees with marker	Permits correspondence to aerial/satellite imagery	Optional	yes	Boolean
Tree DBH	Trunk diameter at chest height (cm) / Provides additional information about tree properties and health	Optional	80cm	numeric
Tree height	Height of tree (m) / Provides additional information about tree properties and health	Optional	55m	numeric
Approx crown Ø	Approximation of tree crown diameter (m)	Optional	8m	numeric
E/C/I	crown position in the field with respect to the surrounding tree crowns	Optional	E	text
% cover	estimated percentage of the leaf fractional cover in the crown	Optional	25%	numeric

Table B.3 Comprehensive list of the metadata elements (critical and optional) in the tree crown reflectance metadataset

Store first set of 50g samples in air-sealed bag with moisturized tissue	For spectral analysis; prevents moisture loss	Optional	yes	Boolean
Store second set of 50g samples in air-sealed bag and store in dry ice (N2O)	For additional chemical analysis	Optional	yes	Boolean
Wet weight	weight measure the same day the leaves are collected from the tree (g)	Optional	5g	numeric
Dry weight	weight of the same leaves measured after drying them in the oven (g)	Optional	3.8g	numeric
Leaf area	Area corresponding to the same leaves computed from the scanned image (cm2)	Optional	8.5cm2	numeric
SLA	Specific leaf area, calculated as (Wet weight/Leaf area) in g/cm2	Optional	0.6g/cm2	numeric
Photo of samples, bough, and canopy	Visual record of samples	Optional	photo # or name	text
Obtain a total of X samples per tree		Optional	5 samples per tree	text

Table B.3 (continued) Comprehensive list of the metadata elements (critical and optional) in the tree crown reflectance metadataset

METADATA FIELD	REASON FOR INCLUSION / COMMENTS	OPTIONALITY	EXAMPLE	DATA TYPE
Description		Critical	ferri-soil	text
Sample #		Critical	1	text
Name	Can be extracted from a taxonomic list / soil series name	Critical	calcic orthid	text
Weight	can be used to describe wet or dry weight	Critical	dry weight to moisture	numeric
Volume	derived from soil cans	Critical	134.5 cm^3	numeric
Mineral bulk density	also can be designated 'soil bulk density'	Critical	msd/vd	numeric
Particle density		Critical	265g/cm^3	numeric
Order		Critical	Aridisol	text
Type		Critical	loam	text
Horizon		Critical	A'	text
Grain size		Critical	3 parts	numeric
Texture	sand/silt/clay	Critical	sieving	text
Surface roughness	necessary for BRDF/erosion calculations	Critical	0.025	numeric
Colour	MUNSELL units/ colour chips can be used	Critical	10 YR 6/4	alphanumeric
Level surface/rough/inclined	aspect should be included	Critical	10˚ or 10`0	numeric
Moisture content	gravimetric or volumetric	Critical	57%	numeric
Humus content		Critical	3.40%	numeric
Nitrogen content		Critical	20 ppm	numeric
Clay content		Critical	20%	numeric
Sand content		Critical	5%	numeric
Silt content		Critical	5%	numeric
pH in H20		Critical	7.0pH	numeric
Water retention (field capacity)		Critical		numeric
Wilting point		Critical	0.44 cm^3/cm^3	numeric
Total alkalinity		Critical	10 mg L^{-1}	numeric
Conductivity		Critical	8 dS/m	numeric
Porosity		Critical	0.45	numeric
Contamination (none/mining/agriculture/etc)		Critical	mining	text
Sample source (pond/lake/marsh/bedrock/etc)		Optional	pond	text
Lower plastic limit		Optional	10	numeric
Upper plastic limit		Optional	12	numeric
pH in CaCl2		Optional	6.3 pH	numeric
pH buffering capacity		Optional	1250 LBC	numeric
Scintillometer reading		Optional	75 c/s	numeric
Loss on ignition (carbon content)	this is a redundant field	Optional	30%	numeric

Table B.4 Comprehensive list of the metadata elements (critical and optional) in the soil reflectance metadataset

METADATA FIELD	REASON FOR INCLUSION / COMMENTS	OPTIONALITY	EXAMPLE	DATA TYPE
GPS coordinates	Permits referencing to aerial/satellite/other campaigns; Difficult to do in situ; done on the dive site; Coordinates, datum + projection can be determined from Google Earth	Critical	x,y,z	numeric
Location description (in situ/on boat/in lab)	Critical to quantifying environmental factors to spectral measurement	Critical	Lab/boat/in situ	text
Reference to photo of local relevant environment + target	Provides additional visual data where recording additional metadata of target and environment is not possible or feasible	Critical	photo # or filename	text
Depth	From lowest astronomical tide	Critical	18 m	numeric
Tide conditions H or L	Input for determining true depth relative to datum and wave lensing effects	Critical	6:36 PM	time
Wave height and period (for reflectance measures)	Input for determining true depth relative to datum and wave lensing effects	Critical	0.25 m	numeric
Wind speed	Critical in severe conditions	Critical	5 kn	numeric
Wind direction	Critical in severe conditions	Critical	Ssw	text
Distance from bottom/substrate	Critical if 3D structure present (seagrass, branching coral)	Critical	20 m	numeric
Substratum height	Input parameter for determining upwelling radiance/ background reflectance affecting spectral measurements	Critical	4 m	numeric
Height of sensor from surface	Critical for water column profiles	Critical	1.75 m	numeric
Depth of sensor from surface	Critical for water column profiles	Critical	7 m	numeric
Distance of operator from sensor	Only applies if there is presence of shading from operator's body	Critical	0.25 m	numeric
CDOM spectral slope	Coloured dissolved organic matter; critical for water column profiles	Critical	-S value	numeric
CDOM concentration	Coloured dissolved organic matter; critical for water column profiles	Critical	A 440 nm	numeric
Detritus concentration	Critical for water column profiles	Critical	1200 µg C•l -1	numeric
Phytoplankton species/classes	Critical for water column profiles	Critical	Gymnodinium spp.	text
Target ID	Code identifier/tag for sample	Critical	Name code	text
Type	Qualitative descriptor of target type	Critical	Coral algae etc.	text
Species or name	Coral species	Critical	Diploria strigosa	text
Density of growth	Quantitative measure of density of target	Critical	2.94 g cm^{-3}	text
Bulb intensity	Input parameter for downwelling radiance calculation	Optional	100 W	numeric
Light spectrum	Range of irradiance spectrum	Optional	VIS/NIR	text
Wave lensing	Can’t be measured in situ; Will know this from wave height data	Optional	yes/no	boolean
Natural canopy shading	Only in seagrass, branching corals	Optional	seagrass shadowing	text
Artificial canopy effect	Shadowing with diver’s body to eliminate influences (e.g. Wave lensing) If measurement is from a boat, then boat may shade	Optional	shadowing of target from diver	text

Table B.5 Comprehensive list of the metadata elements (critical and optional) in the underwater coral reflectance metadataset

Size (diameter)	Size of target	Optional	30 cm	numeric
Homogeneity/heterogeneity	Qualitative description of degree of homogeneity of target being sampled	Optional	homogeneous	text
Homogeneity/heterogeneity (photo)	Attached photo can be used as a reference	Optional	photo # or filename	text
Presence of epiphytes	Useful for endmember analysis of spectral measurements	Optional	Numerous epiphytes	text
Presence of epiphytes(photo)	Attached photo can be used as a reference	Optional	photo # or filename	text
Benthic microalgae (absence/presence)	Useful for endmember analysis of spectral measurements	Optional	Chla sampling	text
Slope	Input parameter for determining upwelling radiance/ background reflectance affecting spectral measurements	Optional	5%	numeric
Strike	Input parameter for determining upwelling radiance/ background reflectance affecting spectral measurements	Optional	25˚	numeric

Table B.5 (continued) Comprehensive list of the metadata elements (critical and optional) in the underwater coral reflectance metadataset

Appendix B.3

This section provides tables of mappings from the seven standards to the proposed Core metadataset, and the critical elements of the tree crown, soil, and underwater coral reflectance metadatasets.

Core Metadataset		**ABCD v2**
Instrument	Instrument operator	/DataSets/DataSet/Units/Unit/Gathering/Agents/GatheringAgent/Person/FullName
Viewing Geometry	Distance from target Distance from ground/background Area of target in field of view Illumination zenith angle Illumination azimuth angle Sensor zenith angle Sensor azimuth angle	/DataSets/DataSet/Units/Unit/Gathering/Method
Project Information	Relevant publication	/DataSets/DataSet/Metadata/IPRStatements/Citations/Citation/Text
	Relevant websites	/DataSets/DataSet/Units/Unit/Gathering/Project/Contact/URIs/URL
	Project participants	/DataSets/DataSet/Units/Unit/Gathering/Project/Contact/Organisation/Name/Representation/Text /DataSets/DataSet/Units/Unit/Gathering/Agents/GatheringAgent/Person/FullName
	Acknowledgement text (sponsorship/affiliates/other)	/DataSets/DataSet/Metadata/IPRStatements/Acknowledgements/Acknowledgement/Text
	Name of experiment/Project	/DataSets/DataSet/Units/Unit/Gathering/Project/ProjectTitle
	Date of experiment	/DataSets/DataSet/Units/Unit/Gathering/DateTime/DateText /DataSets/DataSet/Units/Unit/Gathering/DateTime/DayNumberBegin /DataSets/DataSet/Units/Unit/Gathering/DateTime/DayNumberEnd
Location Information	Location Description	/DataSets/DataSet/Units/Unit/Gathering/NamedAreas/NamedArea /DataSets/DataSet/Units/Unit/Gathering/AreaDetail /DataSets/DataSet/Units/Unit/Gathering/LocalityText
	Referencing Datum	/DataSets/DataSet/Units/Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLong/SpatialDatum
	Longitude	/DataSets/DataSet/Units/Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLong/LongitudeDecimal
	Latitude	/DataSets/DataSet/Units/Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLong/LatitudeDecimal
	Altitude	/DataSets/DataSet/Units/Unit/Gathering/Altitude/MeasurementOrFactAtomised/Parameter /DataSets/DataSet/Units/Unit/Gathering/Altitude/MeasurementOrFactAtomised/LowerValue /DataSets/DataSet/Units/Unit/Gathering/Altitude/MeasurementOrFactAtomised/UpperValue /DataSets/DataSet/Units/Unit/Gathering/Altitude/MeasurementOrFactAtomised/UnitOfMeasurement
	Coordinate source	/DataSets/DataSet/Units/Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinateMethod

Table B.6 Mappings from Access to Biological Collections Data Schema 2.06 to the Core metadataset

General Target and Sampling Information	Target ID	/DataSets/DataSet/Units/Unit/ObservationUnit/ObservationUnitIdentifiers/ObservationUnitIdentifier
	Target treatment	/DataSets/DataSet/Units/Unit/SpecimenUnit/Preparations/Preparation/PreparationType /DataSets/DataSet/Units/Unit/SpecimenUnit/Preparations/Preparation/PreparationProcess /DataSets/DataSet/Units/Unit/SpecimenUnit/Preparations/Preparation/PreparationMaterials
	Field sampling design (transect, plot, other) Plot type Plot dimensions/footprint Transect type Transect interval	/DataSets/DataSet/Units/Unit/Gathering/Method
	Time of sampling by instrument	/DataSets/DataSet/Units/Unit/Gathering/DateTime/ISODateTimeBegin /DataSets/DataSet/Units/Unit/Gathering/DateTime/TimeOfDayBegin /DataSets/DataSet/Units/Unit/Gathering/DateTime/ISODateTimeEnd /DataSets/DataSet/Units/Unit/Gathering/DateTime/TimeOfDayEnd
	Time of collection from field	/DataSets/DataSet/Units/Unit/Gathering/SiteMeasurementsOrFacts/SiteMeasurementOrFact/MeasurementOrFactAtomised/MeasurementDateTime /DataSets/DataSet/Units/Unit/Gathering/DateTime/ISODateTimeBegin /DataSets/DataSet/Units/Unit/Gathering/DateTime/ISODateTimeEnd /DataSets/DataSet/Units/Unit/Gathering/DateTime/TimeOfDayBegin /DataSets/DataSet/Units/Unit/Gathering/DateTime/TimeOfDayEnd
	Target photograph	/DataSets/DataSet/Units/Unit/MultiMediaObjects/MultiMediaObject/ID /DataSets/DataSet/Units/Unit/MultiMediaObjects/MultiMediaObject/FileURI /DataSets/DataSet/Units/Unit/MultiMediaObjects/MultiMediaObject/Format /DataSets/DataSet/Units/Unit/MultiMediaObjects/MultiMediaObject/Comment

Table B.6 (continued) Mappings from Access to Biological Collections Data Schema 2.06 to the Core metadataset

Core Metadataset		ANZLIC Metadata Profile 1.1
General Project Information		
	Relevant websites	On-line resource
	Project participants	Dataset responsible party, Metadata contact individual name , Metadata contact organisation, Metadata contact position , Metadata contact role
	Name of experiment/Project	Dataset title
	Date of experiment	Dataset reference date
Location Information	Location Description	Geographic location of the resource (by description)
	Longitude	West longitude , East longitude, Geographic location of the dataset (by four coordinates or by description)
	Latitude	South latitude , North latitude, Geographic location of the dataset (by four coordinates or by description)
	Altitude	Vertical extent information for the dataset

Table B.7 Mappings from ANZLIC Metadata Profile 1.1 (Geographic dataset core) to the Core metadataset

Core Metadataset		Darwin Core
General Project Information	Relevant publication	bibliographicCitation, references, associatedReferences
	Project participants	institutionID, institutionCode, ownerInstitutionCode, recordedBy
	Name of experiment/Project	datasetName
	Date of experiment	eventDate, startDayOfYear, endDayOfYear, year, month, day, verbatimEventDate
Location Information	Location Description	habitat, locationRemarks, locality
	Referencing Datum	verbatimSRS, geodeticDatum
	Longitude	verbatimLongitude,decimalLongitude
	Latitude	verbatimLatitude, decimalLatitude,
	Altitude	verbatimElevation, minimumElevationInMeters, maximumElevationInMeters
General Target and Sampling Information	Description of target/sample	occurrenceRemarks
	Target ID	individualID, materialSampleID
	Target treatment	preparations
	Time of sampling by instrument	eventTime
	Total number of targets	individualCount
	Time of collection from field	eventTime
	Target photograph	associatedMedia

Table B.8 Mappings from Darwin Core to the Core metadataset

Core Metadataset		Dublin Core
General Project Information	Project participants	Contributor
	Acknowledgement text (sponsorship/affiliates/other)	Contributor
	Name of experiment/Project	Title
	Date of experiment	Date
Location Information	Location Description	Coverage

Table B.9 Mappings from Dublin Core 1.1 to the Core metadataset

Table B.10 Mappings from Ecological Metadata Language 2.1.1 to the Core metadataset

Table B.10 (continued) Mappings from Ecological Metadata Language 2.1.1 to the Core metadataset

Table B.10 (continued) Mappings from Ecological Metadata Language 2.1.1 to the Core metadataset

Core Metadataset		FGDC Remote Sensing Extension
General Project Information		
	Relevant publication	Science_Paper (Description_Documentation module)
	Date of experiment	Time_Period_of_Content (Identification_Information module)

Table B.11 Mappings from FGDC Content Standard for Digital Geospatial Metadata (Remote Sensing Extension) to the Core metadataset

Core Metadataset		FGDC Marine Shoreline Extension
Atmospheric Conditions	Wind speed	Wind Speed
Location Information	Location Description	Description of Geographic Extent

Table B.12 Mappings from FGDC Content Standard for Digital Geospatial Metadata (Shoreline Metadata Profile) to the Core metadataset

Mappings to the soil reflectance metadataset*

** No mappings were possible from ANZLIC Metadata Profile 1.1 (Geographic dataset core), Dublin Core 1.1, FGDC Content Standard for Digital Geospatial Metadata (Remote Sensing Extension) or FGDC Content Standard for Digital Geospatial Metadata (Shoreline Metadata Profile)*

Soil Reflectance Metadataset	ABCD v2
Sample #	DataSets/DataSet/Units/Unit/ObservationUnit/ObservationUnitIdentifiers/ObservationUnitIdentifier

Table B.13 Mappings from Access to Biological Collections Data Schema 2.06 to the soil metadataset

Soil Reflectance Metadataset	Darwin Core
Sample #	individualID, materialSampleID
Weight	ObservedWeight

Table B.14 Mappings from Darwin Core to the soil reflectance metadataset

Soil Reflectance Metadataset	EML 2.1.1.
Description	Specimen (coverage module)
Sample #	referencedEntityId (methods module)
Name	commonName (coverage module)

Table B.15 Mappings from Ecological Metadata Language 2.1.1 to the soil reflectance metadataset

Mappings to the tree crown reflectance metadataset*

** No mappings were possible from ANZLIC Metadata Profile 1.1 (Geographic dataset core), Dublin Core 1.1, FGDC Content Standard for Digital Geospatial Metadata (Remote Sensing Extension) or FGDC Content Standard for Digital Geospatial Metadata (Shoreline Metadata Profile)*

Tree Crown Reflectance Metadataset	ABCD v2
Collected within 1 week of aerial campaign (reference to protocol)	/DataSets/DataSet/Metadata/IPRStatements/Citations/Citation/Text
Position in canopy (reference to protocol)	/DataSets/DataSet/Metadata/IPRStatements/Citations/Citation/Text
Illuminated leaves (reference to protocol)	/DataSets/DataSet/Metadata/IPRStatements/Citations/Citation/Text
Tree species	/DataSets/DataSet/Units/Unit/SpecimenUnit/NomenclaturalTypeDesignations/NomenclaturalTypeDesignation/TypifiedName/FullScientificNameString

Table B.16 Mappings from Access to Biological Collections Data Schema 2.06 to the tree crown reflectance metadataset

Tree Crown Reflectance Metadataset	Darwin Core
Collected within 1 week of aerial campaign (reference to protocol)	samplingProtocol, measurementMethod
Position in canopy (reference to protocol)	samplingProtocol, measurementMethod
Illuminated leaves (reference to protocol)	samplingProtocol, measurementMethod
Mature, dark green leaves have been collected (reference to protocol)	CollectingMethod, measurementMethod

Table B.17 Mappings from Darwin Core to the tree crown reflectance metadataset

Tree Crown Reflectance Metadataset	EML 2.1.1
Collected within 1 week of aerial campaign (reference to protocol)	citation (methods module)
Position in canopy (reference to protocol)	citation (methods module)
Illuminated leaves (reference to protocol)	citation (methods module)

Table B.18 Mappings from Ecological Metadata Language 2.1.1 to the tree crown reflectance metadataset

Mappings to the underwater coral reflectance metadataset*

** No mappings were possible from Dublin Core or FGDC Content Standard for Digital Geospatial Metadata (Remote Sensing Extension)*

Table B.19 Mappings from Access to Biological Collections Data Schema 2.06 to the underwater coral reflectance metadataset

Coral Reflectance Metadataset	ANZLIC Metadata Profile 1.1 (Geographic dataset core)
GPS coordinates	West longitude, East longitude, South latitude, North latitude

Table B.20 Mappings from ANZLIC Metadata Profile 1.1 (Geographic dataset core)to the underwater coral reflectance metadataset

Coral Reflectance Metadataset	Darwin Core
Location description (in situ/on boat/in lab)	locationRemarks
GPS coordinates	verbatimLatitude, verbatimLongitude, decimalLatitude, decimalLongitude
Reference to photo of local relevant environment + target	associatedMedia
Depth	verbatimDepth, minimumDepthInMeters, maximumDepthInMeters, minimumDistanceAboveSurfaceInMeters, maximumDistanceAboveSurfaceInMeters
Phytoplankton species/classes	specificEpithet
Target ID	individualID, materialSampleID
Species or name	specificEpithet

Table B.21 Mappings from Darwin Core to the underwater coral reflectance metadataset

Coral Reflectance Metadataset	EML 2.1.1
GPS coordinates	longitude(spatialReference module), name (angleUnits)(spatialReference module), value(spatialReference module), name (lengthUnits)(spatialReference module)
Depth	depthDatumName (spatialReference module) depthResolution (spatialReference module) depthDistanceUnits (spatialReference module) depthEncodingMethod (spatialReference module)
Height of sensor from surface (if characterizing water column properties)	methodStep, substep, sampling, qualityControl, description (methods module), , proceduralStep (protocol module), protocol (protocol module)
Depth of sensor from surface (if profiling water column)	methodStep, substep, sampling, qualityControl, description (methods module), , proceduralStep (protocol module), protocol (protocol module)
Distance from bottom/substrate	methodStep, substep, sampling, qualityControl, description (methods module), , proceduralStep (protocol module), protocol (protocol module)
Distance of operator from sensor	methodStep, substep, sampling, qualityControl, description (methods module), proceduralStep (protocol module), protocol (protocol module)
Target ID	referencedEntityId(methods module)

Table B.22 Mappings from Ecological Metadata Language 2.1.1 to the underwater coral reflectance metadataset

Coral Reflectance Metadataset	**FGDC Marine Shoreline Extension**
Wave height and period (for reflectance measures)	Wave Height
Tide conditions H or L	Time of Low Tide, Time of High Tide, Tidal Datum, Range of Tide
Wind speed	Wind Speed
Wind direction	Wind Direction

Table B.23 Mappings from FGDC Content Standard for Digital Geospatial Metadata (Shoreline Metadata Profile) to the underwater coral reflectance metadataset

Appendix C Metadata Mappings for SPECCHIO and USGS Spectral Library

Mappings of metadata elements between the Core metadataset and the SPECCHIO and USGS Spectral Library metadatasets are shown here.

Table C.1 Mappings between metadata elements in the Core metadataset and default SPECCHIO v. 2.2 metadata definitions

Table C.1 (continued) Mappings between metadata elements in the Core metadataset and default SPECCHIO v. 2.2 metadata definitions

Note:

Most of the hyperspectral signal properties data within the Core metadataset can be populated retrospectively within SPECCHIO via import of native instrument files, if the user choses to create new metadata fields to store this data. As these metadata fields were not defined in the default metadataset supplied by SPECCHIO, they were not mapped.

There are metadata fields defined within SPECCHIO that do not exist within the core metadataset and therefore were not mapped, and these include: campaign_id*, CampaignDescription, CampaignQualityComply, EnvironmentalConditionID*, ForeopticID*, IlluminationSourceID*, institute_id*, InstituteCity, InstituteCountry, InstituteDepartment, InstituteName, InstitutePOCode, InstituteStreetNo, InstituteStreet, InstrumentID*, LandCoverID*, MeasurementTypeID*, MeasurementUnitID*, NumberOfSpectra, PositionID*, QualityLevelID*, ReferenceID*, RequiredQualityLevelID*, SamplingEnvironmentID*, SamplingGeometryID*, SensorID*, SpecchioUserEmail, SpecchioUserInsitituteID*, SpecchioUserTitle, spectrum_id*, TargetHomogeneity, user_id*.

*These metadata fields are internal database key identifiers for dependent fields (e.g.: SamplingGeometryID is the key identifier for all the viewing geometry metadata parameters dependent on it). In cases where the dependent fields could be mapped to the core metadataset, the key identifier was considered redundant and non-informative, and therefore not mapped.

Table C.2 Mappings between metadata elements in the Core metadataset and the USGS Spectral Library v. splib06a metadata template profiles

Note:

Instrument, Reference Standard, Calibration, Hyperspectral Signal Properties, Illumination Information, Viewing Geometry, Atmospheric Conditions categories in the Core metadataset could not me mapped to the USGS Spectral Library metadata template profiles Only those elements in the remaining categories (General Project Information, Location Information, General Target Sampling Information) that could be mapped to are shown.

There are metadata fields defined within the USGS Spectral Library that do not exist within the core metadataset. These relate mostly to results of spectroscopic and chemical analysis of the samples and include:

COMPOSITION (New Total)
COMPOSITION Al2O3 (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION BaO (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION CaO (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION Cellulose
COMPOSITION Chlorophyll_A
COMPOSITION Chlorophyll_B
COMPOSITION Cl (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION CO2 (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION Cr2O3 (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION F (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION Fe2O3 (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION FeO (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION H2O (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION H2O- (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION H2O+ (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION K2O (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION Li2O (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION Lignin
COMPOSITION LOI (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION MgO (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION MnO (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION Na2O (Oxide ASCII, Amount, wt%, Oxide html)
COMPOSITION NiO (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION Nitrogen

COMPOSITION NNO2

COMPOSITION O=Cl,F,S (Oxide ASCII, Amount, wt%, #correction for Cl, F)

COMPOSITION P2O5 (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION S (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION SiO2 (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION SO3 (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION SrO (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION TiO2 (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION Total

COMPOSITION Total_Chlorophyll

COMPOSITION V2O3 (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION volatile

COMPOSITION Water

COMPOSITION YYO2

COMPOSITION ZnO (Oxide ASCII, Amount, wt%, Oxide html)

COMPOSITION_DISCUSSION

COMPOSITION_TRACE

COMPOSITIONAL_ANALYSIS_TYPE

CURRENT_SAMPLE_LOCATION

FORMULA_HTML

LIB_SPECTRA

LIB_SPECTRA_HED

MICROSCOPIC_EXAMINATION

SPECTRAL_PURITY (1_2_3_4_ # 1= 0.2-3, 2= 1.5-6, 3= 6-25, 4= 20-150)

SPECTROSCOPIC_DISCUSSION

TRACE_ELEMENT_ANALYSIS

TRACE_ELEMENT_DISCUSSION

ULTIMATE_SAMPLE_LOCATION

XRD_ANALYSIS