Remote Sens. **2015**, *7*, 4499-4526; doi:10.3390/rs70404499

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Assessing Field Spectroscopy Metadata Quality

Barbara A. Rasaiah [1,*], Simon. D. Jones [1], Chris Bellman [1], Tim J. Malthus [2] and Andreas Hueni [3]

[1] Remote Sensing and Photogrammetry Research Centre, RMIT University,
Melbourne, VIC 3001, Australia; E-Mails: simon.jones@rmit.edu.au (S.D.J.);
chris.bellman@rmit.edu.au (C.B.)

[2] CSIRO Land and Water, Canberra, ACT 2601, Australia; E-Mail: tim.malthus@csiro.au

[3] Remote Sensing Laboratories, University of Zurich, CH-8057 Zurich, Switzerland;
E-Mail: andreas.hueni@geo.uzh.ch

* Author to whom correspondence should be addressed; E-Mail: Barbara.rasaiah@gmail.com;
Tel.: +61-3-9925-2419.

Academic Editors: Eyal Ben-Dor and Prasad S. Thenkabail

Received: 31 January 2015 / Accepted: 3 April 2015 / Published: 15 April 2015

Abstract: This paper presents the proposed criteria for measuring the quality and completeness of field spectroscopy metadata in a spectral archive. Definitions for metadata quality and completeness for field spectroscopy datasets are introduced. Unique methods for measuring quality and completeness of metadata to meet the requirements of field spectroscopy datasets are presented. Field spectroscopy metadata quality can be defined in terms of (but is 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 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 case studies reveal that the metadata in publicly available spectral datasets are underperforming on the quality and completeness measures. This paper is part two in a series examining the issues central to a metadata standard for field spectroscopy datasets.

Keywords: metadata; databases; data quality; field spectroscopy

1. Introduction

1.1. The Importance of Field Spectroscopy Metadata

The completeness and quality of metadata are central to designing a common platform for the exchange and sharing of field spectroscopy datasets within the remote sensing community on a global basis. Informing users and stakeholders of field spectroscopy datasets of the impact of high-quality metadata in the context of Earth observing data systems is a challenge facing the remote sensing community [1]. Quality assurance of field spectroscopy datasets necessitates oversight and standardization, both at local, national, and international levels and is a way of ensuring reliable field spectroscopy datasets are identified and available for research and other applications.

There is no standardized methodology for documentation of field spectroscopy data or metadata [1–5]. The need for a standardized methodology for collecting—and assuring the quality of—field spectroscopy metadata has increased with the emergence of data sharing initiatives such as NASA's EOSDIS (Earth Science Data and Information System), the LTER (Long Term Ecological Research) network, the Australian Terrestrial Ecosystem Research Network (TERN), SpecNet [6] and some of the smaller *ad hoc* spectral libraries and databases created by remote sensing communities internationally. The absence of a formal standard prohibits efficient and viable intercomparison and fusibility of datasets generated from quantitative field observations [7]. This applies to data and metadata generated for discipline-agnostic information sharing systems and for discipline-specific databases [8]. Consequently, a data user is not sufficiently empowered to assess the quality of field spectroscopy metadata retrieved from field spectroscopy libraries, databases, or other information sharing systems.

A core set of metadata parameters critical to all field spectroscopy campaigns have been recently introduced in [1], the first in this series of papers. They include viewing geometry, location, general target and sampling properties, illumination, instrument properties, reference standards, calibration, hyperspectral signal properties, atmospheric conditions, and general project details. These were identified by an international expert panel of 90 scientists that comprised a diverse group experienced in gathering spectroscopy data across a wide range of disciplines. It is this core set of metadata parameters that was used as a benchmark to assess the quality and completeness of metadata in the case studies presented in this paper.

1.2. Defining Metadata Quality

It is important 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 they are typically documented within the metadata referencing the dataset, whether the dataset is a raster image, coverage, or recorded spectrum [9,10]. 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 paper 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.

1.3. Quality and Completeness within Existing Metadata Standards

Geospatial metadata quality has not been formally defined either in any standard or by any advisory body [9,11–14] 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" [11].

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 [9,10].

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" [15] (p. 106), important to information seeking activities [16], an expression of fitness for purpose [17] and supportive of interoperability and long-term curatorship and preservation [18].

Methods for assessing information quality have been applied to studies of metadata quality. These methods most commonly include 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 [19–21]. 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) [20].

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 [22]. Metrics for metadata quality are mostly generated through automated processes and take various forms including the following:

- an ordinal scale "good/moderate/poor/unusable" describing the overall quality of the metadataset [23]

- quantification of the problems themselves (ambiguity, inaccuracy, inconsistency, redundancy) as percentage of occurrence within a recordset [20]
- accuracy as a measure of semantic distance between a metadata instance and the textual information it references [21]
- 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 [20]

Metrics are limited only by the inventiveness of the metadata analysts and how informative these measures are 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, or search ability and discoverability by users, among others. Simply put, metadata fields for a dataset, however numerous, are not relevant for all resources [21]. 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" [17] (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" [24] (p.182), enables the user to "locate entities by the attributes the user intends to use" [16] (p. 116), and "makes [a dataset] more trustworthy" [25]. 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.

In summary, 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 data ware houses. 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.

Therefore, metadata quality and completeness must be defined in a way with the greatest utility and relevance to users of field spectroscopy datasets and encompass a set of criteria that relates to a baseline set of parameters derived from existing standards and those unique to field spectroscopy metadata.

1.4. 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. 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 [19–21], while at the same time conforming in some respect to concepts of data quality proposed for geospatial datasets by geospatial science advisory bodies [9–14].

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. These are therefore a suitable set of parameters by which to measure metadata completeness and quality in field spectroscopy datasets.

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 [1] 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.

Applying the proposed quality and completeness measures presented in this section to existing spectral libraries illustrates 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.

2. Methods

2.1. 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. 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 1 provides a general overview of the two selected data libraries.

Table 1. Overview of USGS and SPECCHIO spectral data libraries.

Spectral Library	Agency	Purpose	Year Created	Format	# of Campaigns	# of 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

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 [26]. 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 SPECTRO radio meters in a central repository [27]. 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. SPECCHIO 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). Table 1 presents a summary profile of the USGS Spectral Library and the SPECCHIO database used in the metadata quality and completeness analysis.

Both the SPECCHIO and USGS datasets had to be prepared for analysis. A database backup copy of SPECCHIO was provided by the RSL data centre 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 violating 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 SPECCHIO 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 the 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 [28,29].

2.2. Data Analysis

Assessing the quality and completeness in the data libraries was a combination of qualitative and quantitative methods based on the parameters proposed for measuring metadata quality presented in Section 1.4.

2.2.1. Metadata Completeness Analysis Methods

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.

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.

Additionally, SPECCHIO spectrum-level metadata was analysed to determine if there were patterns of variance for completeness levels. 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.

Investigating those categories where the database users were inconsistent in populating metadata categories—or patterns of variance—can 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 method of analysis chosen for identifying variance in SPECCHIO spectrum-level completeness was dimensional scaling of the data, to better understand the variance and co-variance relationships among the SPECCHIO metadata elements. 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 [30–32]. All zero-variance metadata elements were excluded, and these were "Is Reference", "Reference Serial Number" and "Reference Brand Name" (all referring to the reference standard used while taking measurements) and the "Required Quality Level" and "Quality Level" 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 [33].

2.2.2. Metadata Quality Analysis Methods

The quality measure was an assessment based on the five proposed metadata quality parameters presented in Section 1.4 (logical consistency, error rate, quality assurance, lineage, reputational authority). 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 (1,110,233 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.

It was not possible to assess metadata quality to the same rigorous extent for each of the five parameters, for reasons explained in more detail in the results section. However, two of the quality parameters—reputational authority and logical consistency—were assessed using unique methods. 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 more 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. 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 a given 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.

Logical consistency was the second metadata quality parameter that required unique and rigorous analysis of the data. Logical consistency for a metadata instance can be defined as "the degree to which it matches the metadata standard definition" [21] (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 [34] 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 [35] (Equation (1))

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

$$H(X) = -p(x)\log p(x) \tag{1}$$

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 [34]. Entropy and cumulative entropy are useful for metadata quality analysis because they can be used to identify changes in data entry characteristics [34] and as a measure of the diversity of information being stored [16].

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.

3. Results and Discussion

3.1. Mappings for Metadata Elements between the Core Metadataset and SPECCHIO and USGS Spectral Library

The successful mappings for metadata elements between the core metadataset [1] and SPECCHIO and USGS Spectral Library are presented in Tables 2 and 3. The mappings were used to assess metadata completeness and quality in Section 3.2. For brevity, metadata parameters in the core metadataset that could not be mapped to either SPECCHIO or the USGS Spectral Library are not shown. A full listing of metadata parameters in the core metadataset are presented in [1].

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 (Table 2).

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 chooses 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. Metadata fields defined within SPECCHIO that do not exist within the core metadataset and were therefore not mapped can be found in Appendix A.

The number of successful mappings for metadata elements between the USGS Spectral Library and the core metadataset (Table 3) were fewer than for SPECCHIO. Instrument, Reference Standard, Calibration, Hyperspectral Signal Properties, Illumination Information, Viewing Geometry, Atmospheric Conditions categories in the Core metadataset could not be 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. Metadata fields defined within the USGS Spectral Library that do not exist within the core metadataset and were therefore not mapped can be found in Appendix B.

Table 2. Mappings between core metadataset parameters and SPECCHIO metadata.

Core Metadataset Category	Core Metadataset Parameter	SPECCHIO Metadata Parameter
Instrument	Make and model	description
	Dark signal correction	time since last dc
	Integration time	integration time
	Mode (cos-conical, bi–conical)	Bidirectional/Directional-conical/Directional-hemispherical/ Conical-directional/Biconical/Conical-hemispherical/ Hemispherical-directional/Hemispherical-conical/Bihemispherical
	Manufacturer	manufacturer_name
	Serial number	serial_number
	Gain settings (Automatic/Manual)	instr_setting_type_id
	Signal averaging (instrumental)	InternalAverageCount
Reference Standard	No reference standard used	IsReference
	Reference material	name
	Serial number	serial_number
Hyperspectral Signal Properties	Data type (Reflectance, Radiance…)	Reflectance/Radiance/Absorbance/Transmittance/DN/ Wavelength/Mueller10/Mueller20/Irrance
	Wavelength interval	avg_wavelength
	Wavelength data	avg_wavelength
Illumination Information	Source of illumination (e.g., sun, lamp)	IlluminationSourceID
Viewing Geometry	Distance from target	sensor_distance
	Illumination zenith angle	illumination_zenith
	Illumination azimuth angle	illumination_azimuth
	Sensor zenith angle	sensor_zenith
	Sensor azimuth angle	sensor_azimuth
Atmospheric Conditions	Cloud cover (%)	cover_in_octas
	Humidity	relative_humidity
	Wind speed	wind_speed

Table 2. *Cont.*

Core Metadataset Category	Core Metadataset Parameter	SPECCHIO Metadata Parameter
General Project Information	Relevant websites	Specchio_User WWW, InstituteWWW
	Project participants	Specchio FirstName, SpecchioUser_LastName, Institute_Name
	Name of experiment/Project	Campaign Name
Location Information	Location Description	location_name; landcover cover_desc
	Longitude	longitude
	Latitude	latititude
	Altitude	altitude

Table 3. Mappings between metadata elements in the Core metadataset and the USGS Spectral Library v. splib06a metadata template profiles.

Core Metadataset Category	Core Metadataset Parameter	USGS Metadata Template Profile					
		Man Made	**Volatile**	**Minerals**	**Micro-Organisms**	**Mixtures**	**Plants**
		Metadata Parameter	**Metadata Parameter**	**Metadata Parameter**	**Metadata Parameter**	**Metadata Parameter**	**Metadata Parameter**
General Project Information	Project participants	original donor					
Location Information	Location Description	collection locality					
	Referencing Datum	datum					
	Longitude	collection longitude					
	Latitude	collection latitude					
General Target Sampling Information	Description of target/sample	title, material, formula, sample description		title, mineral, formula, sample description	title, micro-organism, latin name, sample description	title, mixture, mixture formula , sample description	title, plant, plant latin name, sample description
	Target type (vegetation, mineral, aquatic, *etc.*)	material type		mineral type	micro-organism type	mixture type	plant type
	Target ID	sample id					
	Target photograph	image of sample					

3.2. Metadata Completeness Analysis Results

The results of the mapping analysis are presented in Table 4. 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 would not be aware of the core metadataset, and therefore, have not implemented it in their metadata policy.

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.

Table 4. Metadata completeness report for SPECCHIO and USGS Spectra Library.

Spectral Library/Database	**Completeness Statistics**										
SPECCHIO				**Campaign Completeness (Internal Metadata Policy)**				**Core Metadataset Compliance (Campaign + Spectrum)**			
		# of campaigns examined	# of parameters	min	max	avg %	stdev	min	max	avg %	stdev
		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%				
USGS				**Spectrum Completeness (Internal Metadata Policy)**				**Core Metadataset Compliance (Spectrum)**			
		# of spectra examined	# of parameters	min	max	avg %	stdev	min	max	avg %	stdev
	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%

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" [36] (p. 15). SPECCHIO also provides a quality flag that relates to the quality of the reference measurements. The "IsReference" field *was* intended as a flag that can be set by the database administrator to denote "prime data". This metadata field was mapped to the "no reference standard used" metadata field in the core metadataset because of the ontological correspondence between these two fields; the quality of the spectrum can be derived implicitly by data users based on the value stored in the "IsReference" and "no reference standard used" fields, but does not relate to the quality of the metadata itself. 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 SPECPR 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 metadata 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" [26]. There are no completeness or quality flags in the metadata.

Measuring variance in SPECHHIO spectrum-level completeness was accomplished with categorical principal component analysis (with ordinal measurement) to determine those metadata parameters that cluster together, by their proportionate variance, for the completeness measure (please see the Methods section for a description of this method). Table 5 shows the (metadata element) loadings for each dimension.

Table 5. 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.

Metadata Element	**Dimension**						
	1	2	3	4	5	6	7
AirPressure	0.034	**0.797**	−0.555	−0.228	−0.010	−0.047	−0.003
Altitude	−0.299	−0.415	0.034	**−0.670**	−0.111	−0.338	0.207
AmbientTemperature	0.034	**0.797**	−0.555	−0.228	−0.010	−0.047	−0.003
CloudCoverInOctas	0.066	−0.015	−0.011	−0.114	−0.090	**0.600**	−0.001
IlluminationAzimuth	**0.553**	−0.304	−0.107	−0.516	−0.432	0.090	−0.199
IlluminationDistance	−0.073	0.003	0.037	−0.125	**0.439**	−0.219	−0.420
IlluminationZenith	**0.872**	−0.219	−0.133	−0.177	0.036	−0.099	0.188
InstrumentName	0.457	0.052	−0.075	**0.618**	−0.463	−0.199	−0.027
InstrumentSerialNumber	0.457	0.052	−0.075	**0.618**	−0.463	−0.199	−0.027
InternalAverageCount	**−0.892**	−0.060	0.158	−0.201	0.053	−0.065	−0.126
LandcoverDescription	−0.257	−0.042	0.065	**−0.485**	0.116	0.441	0.271
Latitude	0.543	−0.311	−0.101	**−0.570**	−0.422	0.121	−0.167
LocationName	**0.903**	−0.197	−0.127	−0.085	0.157	−0.191	0.042
Longitude	0.543	−0.311	−0.101	**−0.570**	−0.422	0.121	−0.167
ManufacturerName	0.226	0.511	**0.810**	−0.126	−0.104	−0.042	0.011
ManufacturerShortName	0.226	0.511	**0.810**	−0.126	−0.104	−0.042	0.011
ManufacturerWWW	0.227	0.274	0.383	0.087	**0.452**	0.315	−0.211
MeasurementType	**0.918**	0.086	−0.128	0.243	0.118	0.177	0.054
MeasurementUnit	0.125	−0.003	0.037	−0.076	0.038	−0.052	**0.720**
RelativeHumidity	0.034	**0.797**	−0.555	−0.228	−0.010	−0.047	−0.003
SamplingEnvironmentName	**0.904**	−0.124	0.000	−0.150	0.309	−0.049	0.026
SensorAzimuth	**0.919**	−0.089	−0.028	−0.034	0.251	−0.045	−0.025
SensorDistance	**0.921**	−0.005	−0.071	0.084	0.239	0.052	−0.001
SensorZenith	**0.929**	−0.083	−0.027	−0.020	0.271	−0.049	−0.010
SensorDescription	0.226	0.511	**0.810**	−0.126	−0.104	−0.042	0.011
SensorName	0.226	0.511	**0.810**	−0.126	−0.104	−0.042	0.011
SensorNoOfChannels	0.226	0.511	**0.810**	−0.126	−0.104	−0.042	0.011
TargetHomogeneity	−0.186	−0.337	0.219	**−0.644**	0.273	−0.379	−0.054
WindDirection	0.034	**0.797**	−0.555	−0.228	−0.010	−0.047	−0.003
WindSpeed	0.034	**0.797**	−0.555	−0.228	−0.010	−0.047	−0.003

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").

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 encourage future investigation as to why database users are not consistent in populating metadata in these three categories that have been identified by their peers as critical to all field spectroscopy metadatasets [1] Unpopulated metadata in these categories is fundamentally compromising the overall quality, interoperability, and inter comparison of these datasets. These findings also encourage 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.

3.3. Metadata Quality Analysis Results

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.

3.3.1. Quality Assurance

Neither SPECCHIO nor USGS Spectral Library have any metadata quality measures (aside from those discussed in Section 3.2 that are inapplicable) or quality assurance parameters.

3.3.2. Lineage

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

3.3.3. Reputational Authority

Analysis of variance 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 the proposed core metadataset.

Z-scores were calculated from the raw completeness measures to determine whether they differed across users and institutes. 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 1 and 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 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.

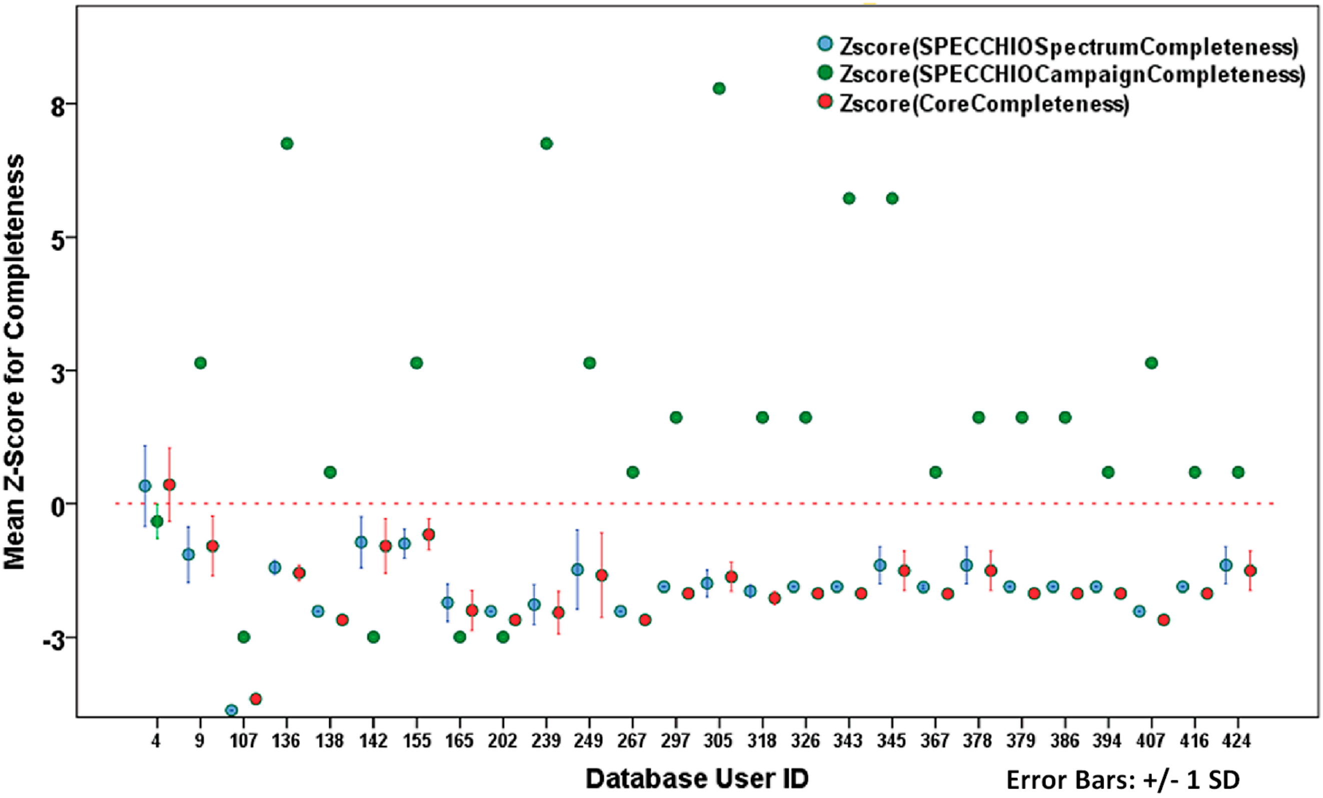

Figure 1. Mean Z-scores for completeness by database user.

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 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. 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.

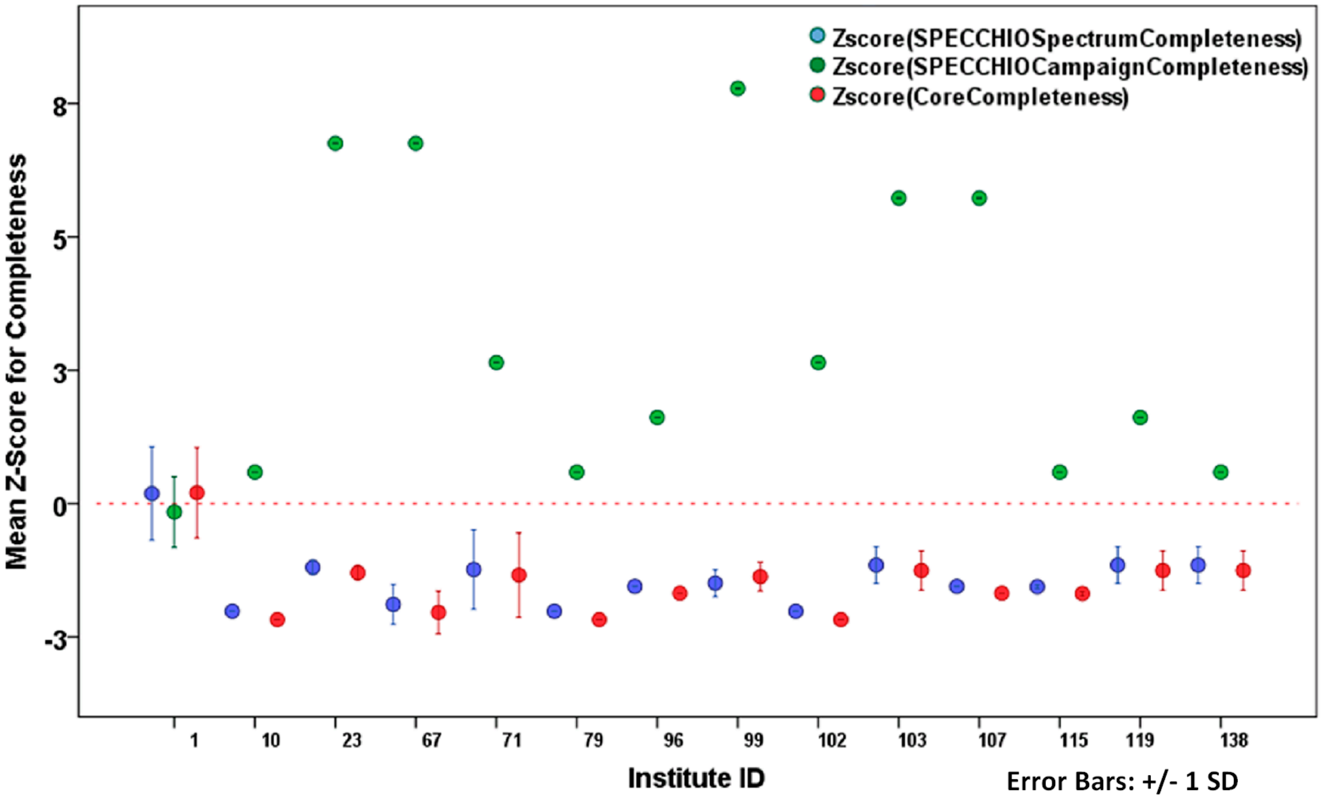

Figure 2. Mean Z-scores for completeness by institute.

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.

3.3.4. 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 2.2. 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).

3.3.5. Logical Consistency

Cumulative entropy analysis was performed to determine logical consistency within the USGS Spectral Library (please see the Methods section for a detailed explanation of the algorithm used). The metadata instance chosen for analysis was "sample description" and 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). Results are shown in Figure 3.

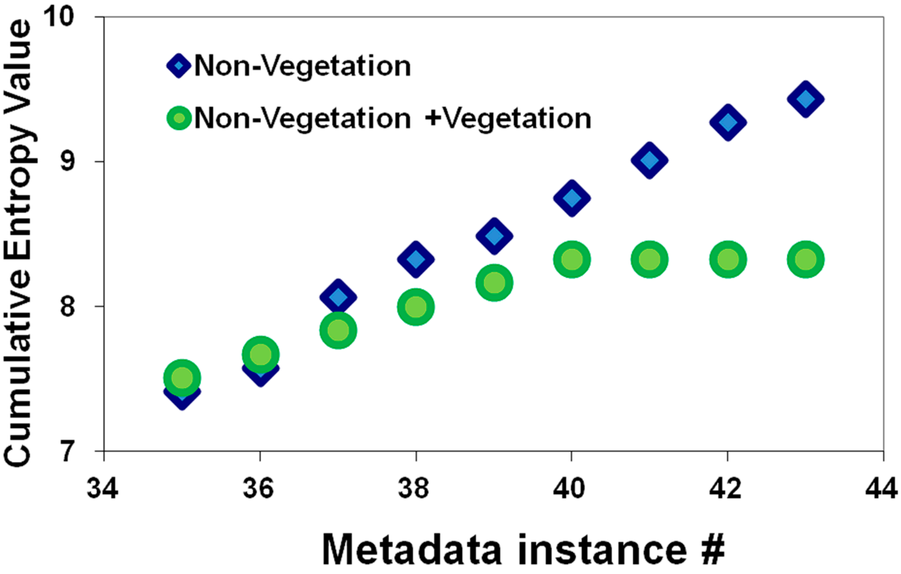

Figure 3. Cumulative entropy for non-vegetation and mixed groups.

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 producing 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.

4. Conclusions

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 the core metadataset. 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 a data user to assess whether a given dataset is suitable for their purpose.

Metadata in 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 databases fail to comply completely with their internal metadata policies, with 59.3% compliance with the campaign-level and 51.7% compliance with spectrum-level completeness for SPECCHIO, and an average of 72% compliance for samples in the USGS Spectral Library. There are no metadata quality parameters in either database, 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.

Overall, the fact that publicly available field spectroscopy datasets are underperforming on these quality and completeness measures prevents current and future data users from having the confidence that the metadata available allows them to make informed decisions about the suitability of a given dataset for a particular application. If field spectroscopy metadata is to be implemented in large-scale data sharing systems, the field spectroscopy community must be proactive in improving existing metadata policies. The results presented in this paper serve as indicators for areas of focus.

The methods and algorithms used in these test cases for quality and completeness assessment can be applied to 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 those areas where users are not educated on which metadata are critical, and in identifying systematic problems with metadata policies. The metadata quality and completeness measures presented here can also be easily implemented for wide-scale assessment of metadatasets. They were developed with a focus on the users' needs in terms of discovering metadata and assessing it as suitable for their purposes, an underlying principle currently lacking in existing metadata standards [37]. Adopting these metadata quality and completeness measures as a standard can be of great value to the field spectroscopy community. They are built on the foundation of a metadataset established by the field spectroscopy community and have incorporated additional elements of metadata quality parameters that serve to enhance the discoverability and interoperability of datasets. These metadata quality and completeness measures, if standardized, would encourage diligence on the part of data creators to produce high-quality metadata and also tailor such measures to align with their discipline-specific requirements [8].

Given that the spectral libraries examined in this paper 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.

Acknowledgments

The authors gratefully acknowledge the encouragement, constructive criticism, and input of the manuscript reviewers and editors.

Author Contributions

Barbara Rasaiah designed the field spectroscopy metadata quality and completeness measures, performed data preparation and data cleaning, formulated and executed data analysis, and performed statistical analysis and interpretation of results. Simon Jones, Chris Bellman, and Tim Malthus helped provide the conceptual framework for the research, and provided guidance on the research methodology and editorial input for the article. Andreas Hueni provided the SPECCHIO dataset and clarified the SPECCHIO metadata schema and content.

Appendix

Appendix A

SPECCHIO Metadata Parameters not Mapped to the Core Metadataset

There are metadata fields defined within SPECCHIO that do not exist within the core metadataset and therefore were not mapped. 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*. Metadata fields denoted by * 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.

Appendix B

USGS Library Metadata Parameters not Mapped to the Core Metadataset

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

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Rasaiah, B.A.; Jones, S.D.; Bellman, C.; Malthus, T.J. Critical metadata for spectroscopy field campaign. *Remote Sens.* **2014**, *6*, 3662–3680.
2. Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. Building better hyperspectral datasets: The fundamental role of metadata protocols in hyperspectral field campaigns. In Proceedings of the Surveying & Spatial Sciences Conference, Wellington, New Zealand, 21–25 November 2011.

3. Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. Designing a robust hyperspectral dataset: The fundamental role of metadata protocols in hyperspectral field campaigns. In Proceedings of the GSR 1, Melbourne, Australia, 28–30 November 2011.
4. Rasaiah, B.; Malthus, T.; Jones, S.D.; Bellman, C. Critical metadata protocols in hyperspectral field campaigns for building robust hyperspectral datasets. In Proceedings of the XXII ISPRS Congress, Melbourne, Australia, 26 August–1 September 2012.
5. Rasaiah, B.; Jones, S.D.; Chisholm, L.; Hueni, A.; Bellman, C.; Malthus, T.J.; Chisholm, L.; Gamon, J.; Huete, A.; Ong, C.; *et al*. Approaches to establishing a metadata standard for field spectroscopy datasets. In Proceedings of IGARSS 2013, Melbourne, Australia, 21–26 July 2013.
6. Gamon, J.A.; Rahman, A.F.; Dungan, J.L.; Schildhauer, M.; Huemmrich, K.F. Spectral Network (SpecNet)—What is it and why do we need it? *Remote Sens. Environ.* **2006**, *103*, 227–235.
7. Jung, A.; Götze, C.; Glässer, C. Overview of experimental setups in spectroscopic laboratory measurements–the SpecTour Project. *Photogramm.-Fernerkund.-Geoinf.* **2012**, *4*, 433–442.
8. Dor, E.B.; Ong, C.; Lau, I.C. Reflectance measurements of soils in the laboratory: Standards and protocols. *Geoderma* **2015**, *245*, 112–124.
9. FGDC. Content Standard for Digital Geospatial Metadata: Extensions for Remote Sensing Metadata. Available online: http://www.fgdc.gov/standards/projects/FGDC-standards-projects/csdgm_rs_ex/MetadataRemoteSensingExtens.pdf (accessed on 24 August 2013).
10. ISO. Project Information--Fact Sheet 19113: 19113 Geographic Information—Quality Principle. Available online: http://www.isotc211.org/Outreach/Overview/Factsheet_19113.pdf (accessed on 13 October 2013).
11. ISO. ISO 19113:2002: Geographic Information--Quality Principles. Available online: http://www.iso.org/iso/catalogue_detail.htm?csnumber=26018 (accessed on 13 October 2013).
12. ISO. ISO/WD 19159 Geographic Information–Calibration and validation of remote sensing imagery sensors and data, 2011. Lysaker: ISO/TC 211 Secreteriat. Available online: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_tc_browse.htm?commid=54904 (accessed on 13 October 2013).
13. ANZLIC. ANZLIC Metadata Profile Version 1.1. Available online: http://spatial.gov.au/sites/default/files/legacy/osdm.gov.au/Metadata/ANZLIC%2BMetadata%2BProfile/default.html (accessed on 27 August 2013).
14. INSPIRE. INSPIRE Metadata Implementing Rules: Technical Guidelines Based on ENISO 19915 and ENISO 19119. Available online: http://inspire.jrc.ec.europa.eu/documents/Metadata/INSPIRE_MD_IR_and_ISO_v1_2_20100616.pdf (accessed on 17 May 2010).
15. Margaritopoulos, T.; Margaritopoulos, M.; Mavridis, I.; Manitsaris, A. A conceptual framework for metadata quality assessment. In Proceedings of the International Conference on Dublin Core and Metadata Applications, Berlin, Germany, 22–26 September 2008.
16. Stvilia, B.; Gasser, L.; Twidale, M.B.; Shreeves, S.L.; Cole, T.W. Metadata quality for federated collections. In Proceedings of the Ninth International Conference on Information Quality, Cambridge, MA, USA, 5–7 November 2004.
17. Park, J.R. Metadata quality in digital repositories: A survey of the current state of the art. *Cat. Classif. Quart.* **2009**, *47*, 213–228.

18. NISO. A Framework of Guidance for Building Good Digital Collections. Available online: http://www.niso.org/publications/rp/framework3.pdf (accessed on 7 January 2014).
19. Bruce, T.R.; Hillmann, D.I. The continuum of metadata quality: Defining, expressing, exploiting. In *Metadata in Practice*; Hillmann, D., Westbrooks, E., Eds.; American Libray Association Editions: Chicago, IL, USA, 2004; pp. 238–256;
20. Stvilia, B.; Gasser, L.; Twidale, M.B.; Smith, L.C. A framework for information quality assessment. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58*, 1720–1733.
21. Ochoa, X.; Duval, E. Automatic evaluation of metadata quality in digital repositories. *Int. J. Digit. Libr.* **2009**, *10*, 67–91.
22. Liolios, K.; Schriml, L.; Hirschman, L.; Pagani, I.; Nosrat, B.; Sterk, P.; Field, D. The Metadata Coverage Index (MCI): A standardized metric for quantifying database metadata richness. *Stand. Genomic Sci.* **2012**, *6*, 438–447.
23. Currier, S.; Barton, J.; O'Beirne, R.; Ryan, B. Quality assurance for digital learning object repositories: Issues for the metadata creation process. *Res. Learn. Technol.* **2004**, *12*, 5–20.
24. Goovaerts, M.; Leinders, D. Metadata quality evaluation of a repository based on a sample technique. In *Metadata and Semantics Research*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 181–189.
25. ANDS. Metadata. Available online: http://ands.org.au/guides/metadata-awareness.html (accessed on 7 January 2014).
26. USGS. USGS Digital Spectral Library splib06a. Available online: http://speclab.cr.usgs.gov/spectral.lib06/ds231/index.html (accessed on 26 October 2013).
27. Hueni, A.; Nieke, J.; Schopfer, J.; Kneubuehler, J.; Itten, K.I. The spectral database SPECCHIO for improved long-term usability and data sharing. *Comput. Geosci.* **2009**, *35*, 557–565.
28. SAS Institute. Data Mining and the Case for Sampling. Available online: 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 (accessed on 15 October 2013).
29. Khandar, P.V.; Dani, S.V. Knowledge discovery and sampling techniques with data mining for identifying trends in data sets. *Int. J. Comput. Sci. Eng.* **2011**, *2011*, 7–11.
30. Linting, M.; Meulman. J.J.; Groenen, P.J.; van der Koojj, A.J. Nonlinear principal component analysis: Introduction and application. *Psychol. Methods* **2007**, *12*, 36–58.
31. Meulman, J.J.L.; Heiser, W.J. IBM SPSS Categories 21. Available online: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/21.0/en/client/Manuals/IBM_SPSS_Categories.pdf (accessed on 14 January 2014).
32. Starkweather, J. RSS SPSS Short Course Module 9 Categorical PCA. Available online: http://www.unt.edu/rss/class/Jon/SPSS_SC/Module9/M9_CATPCA/SPSS_M9_CATPCA.htm (accessed on 30 August 2013).
33. Kaiser, H.F. The application of electronic computers to factor analysis. *Educ. Psychol. Meas.* **1960**, *20*, 141–151.
34. Simon, S.P. Mean: Checks for Data Quality Using Metadata. Available online: http://www.pmean.com/08/CheckMetadata.html (accessed on 26 October 2013).

35. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: Hoboken, NJ, USA, 1991; pp. 112–113.
36. Hueni, A. SPECCHIO User Guide V. 2.1.2. Available online: http://specchio.ch/user_guides.php (accessed on 23 March 2012).
37. Goodchild, M.F. Beyond metadata: Towards user-centric description of data quality. In Proceedings of the 2007 International Symposium on Spatial Data Quality, Enschede, The Netherlands, 13–15 June 2007.