Problem Analysis on Data Quality Assessment Issues

ITQ #CRS 9062

Submitted to the TEM Alternatives Task Force

By

D.E. Moon, CDT Core Decision Technologies Inc.

Summary

This report emphasizes the elements of data quality, the inventory procedures that influence data quality, and a formal framework for evaluating the quality of data inputs to predictive ecosystem mapping (PEM). There are two reasons for this. The first is that following the definition of the elements of data quality for input to PEM, a quick review of potential input map products indicated that the quality of this data for PEM purposes is suspect. The second is that some users of potential input data do not understand the nature of the data nor its limitations for PEM. Given the potentially serious impact of these conditions, it was necessary to develop and explain the rationale for a rigorous, supportable framework for a formal evaluation of PEM input sources. This included:

1. Development of a detailed rationale for the elements of data quality,

2. Identification of the inventory procedures that influence data quality and an explanation of their impact,

3. Identification of the meta-data (documentation) needed to evaluate data quality,

4. Proposal of a framework and criteria for the evaluation of input data quality.

The framework is useful for evaluating both data inputs to predictive ecosystem mapping and for evaluating outputs from predictive ecosystem mapping. The framework was developed from the perspective of someone wishing to understand the fitness or utility of map information. It is developed with the knowledge that much of the meta-data about map products that could be used in evaluating data quality will be missing for most maps and that surrogates will have to be used. In all cases, the evaluation is dependent on the actual use to which the final product will be applied.

The framework is applicable to any thematic input or output map. It is applicable to a class of maps, an individual map, or portions of individual maps. The rating concentrates on the confidence with which the data can be used, and it incorporates both thematic and spatial criteria. The framework is developed in six stages.

Development of the Framework

Stage 1: The background, context, and scope of the report are defined.

Stage 2: The review of predictive ecosystem mapping, as applied in British Columbia, and the identification of the character and nature of the data elements or attributes used in the predictive ecosystem mapping process.

Stage 3: The identification and discussion of the elements of data quality relevant to predictive ecosystem mapping input and output products. These elements were identified as thematic and spatial in nature.

Thematic elements refer to the information conveyed by the map and it's associated reports and data. Thematic data may include forest cover, soils, terrain, *et cetera*. The thematic elements relate to thematic precision (exactness or resolution), thematic

accuracy (correctness within the defined precision), and the nature of thematic contents (what thematic information the map portrays).

Spatial elements are positional and topological in nature. Positional refers to the actual location of a point, line, or polygon on the map and in the real world. Topological refers to the relationship of points, lines and polygons to each other. They may be positionally incorrect but if they are in correct relation to each other (e.g., inside, outside, left, right, *et cetera*) they are topologically correct. Positional accuracy ensures topological accuracy but topological accuracy does not ensure positional accuracy. Positional accuracy for maps used in an overlay process is critical because inaccurate input maps produce incorrect combinations of properties in the overlay product. Incorrect inputs to predictive mapping produce incorrect interpretations even if the interpretation procedure is correct.

Stage 4: The identification and discussion of procedures, which will affect the quality of input and output data. The procedures and their possible impact on data quality are discussed.

Stage 5: The identification, documentation, and discussion of meta-data (data about the data) needed to adequately evaluate the data presented both for its quality and applicability to predictive ecosystem mapping.

Stage 6: The development, presentation, explanation, and limited application of a framework for evaluating input and output data quality.

Conclusions

Data quality issues, unrelated to the original purposes of potential PEM input maps, suggest serious problems with the application of existing inventory data for PEM. The evaluation of data input sources requires a tightly reasoned, widely accepted framework to ensure acceptance of potentially negative conclusions. The framework for data quality evaluation presented in this report proposes such a framework. It is, however, only a framework with suggested criteria. It is not proposed as a final procedure with definitive criteria nor has it been extensively tested. It does appear to have the following merits.

1. It provides a list of issues that should be addressed when determining the quality of data inputs or outputs related to predictive ecosystem mapping.

2. It provides a systematic process and suggests criteria for evaluating the quality of data inputs to or outputs from PEM.

3. It provides a framework for the creation, documentation, and implementation of data quality standards for predictive ecosystem mapping input and output products.

Recommendations

1. The framework should be tested and evaluated against a variety of possible PEM input sources and output products to:

    - identify and incorporate issues which may have been missed

- test the appropriateness of the concepts and criteria used

- refine the framework and criteria prior to presentation to a workshop involving practitioners of input mapping products, terrestrial ecosystem mapping products, and predictive ecosystem mapping products

2. A selection of potential data inputs should be evaluated for thematic and spatial accuracy in the field, the nature and magnitude of errors documented, and the implications of these errors on the quality of PEM output determined. The test areas should be stratified on results of the evaluation framework applied to existing data sources.

3. A workshop consisting of practitioners of input mapping products, terrestrial ecosystem mapping, and predictive ecosystem mapping should be sponsored to:

   - review and contribute to the framework

   - make recommendations on the adoption and use of the framework or a modified framework to establish and implement data quality standards for input and output products

4. If supported by the workshop, an expert committee should be struck to establish the criteria and ratings to be used as Resources Inventory Committee standards for the evaluation of input and output data quality related to terrestrial and predictive ecosystem mapping.

Table of Contents

Problem Analysis on Data Quality Assessment Issues for Predictive Ecosystem Mapping (PEM)

ITQ #CRS 9062

1. INTRODUCTION

1.1. CONTEXT

1.1.1. Program Context

This study is part of a larger initiative being undertaken by the TEM (Terrestrial Ecosystem Mapping) Alternatives Task Force.

The TEM (Terrestrial Ecosystem Mapping) Alternatives Task Force, with representation from the Ministry of Forests and the Ministry of Environment, Lands & Parks (Headquarters and Regions), and the forest industry was formed subsequent to the November 1997 release of the report "A Business Approach to Terrestrial Ecosystem Mapping (TEM) and TEM Alternatives" by the TEM and TEM Alternatives Project Steering Committee. The objective of the Task Force was to further the development, testing, and documentation of "alternatives to TEM", primarily the Predictive Ecosystem Mapping (PEM) approaches. These tasks were all viewed to be important and necessary precursors for the development of RIC standards regarding the use of predictive ecosystem mapping approaches.

The Task Force undertook an initiative (Evaluation, Role & Development of Standards for Predictive Ecosystem Approaches to Terrestrial Ecosystem Mapping) that is focused principally on Predictive Ecosystem Mapping approaches in terms of —

- their continued but more structured evaluation,
- their future role in the TEM inventory program, and
- the formation of TEM–PEM Principles & Transitionary RIC standards around the use of these methods.

Specific objectives for the initiative to the end of fiscal year 1998–99 and for early 1999-2000 comprised three main components as outlined below. Components A, B4 and C are being undertaken mainly by a Secretariat Team on behalf of the TEM Alternatives Task Force. Components studies B1, B2 and B3 will be coordinated by the Secretariat Team but be otherwise accomplished using expert consultants in the respective fields.

A. **Direction & Coordination** — to coordinate and create a forum for an integrated TEM and TEM Alternatives program by developing an overarching framework (i.e., a road map), establishing a set of guiding principles and preparing an evolving working "white paper". These coordinating functions and framework tools will provide overall direction and guidance for managing the TEM Alternatives inventory program during this transition phase.

B. **Detailed Supporting Studies** — to undertake detailed supporting studies, as set out in the above framework (A), that will help to establish a more thorough understanding and foundation for future TEM and TEM Alternatives inventory program. The detailed studies required are —

1) Situation Analysis of Knowledge Based Systems in the context of corporate information system directions,
2) Problem Analysis on Data Quality Assessment Issues: Inputs, Outputs and Reporting Requirements,
3) Problem Analysis on Inventory Reliability, Quality Control, Field and Knowledge Base Validation Procedures, and
4) Compendium of Predictive Ecosystem Mapping Case Studies.

C. **TEM–PEM Principles & Transitionary Standards** — to define a set of TEM–PEM Principles & Transitionary Standards in the context of an evolving and overarching framework (A) and based on the findings of the detailed studies (B1-B4). These principles and transitionary standards will aim to provide a framework for and clarity to the present transitionary circumstances while at the same time offer reasonable flexibility for evolution and innovation.

This study is item B2 above "Problem Analysis on Data Quality Assessment Issues: Inputs, Outputs and Reporting Requirements."

1.1.2. Predictive Ecosystem Mapping Context

More specifically, the context of the following discussion is as follows:

In evaluating other inventory data sources for input to a predictive ecosystem map, how do you determine the level of confidence that you should place in the input data or the output from a PEM based on this data?

Your evaluation should be based on the intended application(s) and the reliability requirements of the final product as dictated by its intended application and the consequences of error.

1.2. SCOPE

The distinction between TEM and PEM is somewhat arbitrary since TEM uses essentially the same knowledge base and input data as PEM, however the primary difference is how the knowledge and data are processed. All polygons that do not receive a ground verification of ecosystem attributes are the result of a predictive process. The principal distinction is that PEM applications utilize geographic information systems to support data integration and that the knowledge base used to predict ecosystem attributes is formalized in a digital knowledge base (whereas TEM processing is done cognitively and manually). The process of ecosystem mapping (TEM or PEM) has three principal components:

1. Data input
2. Data inference, and
3. Data output.

Component 1 (data input) is critical since the intent of PEM is to leverage existing data to reduce the costs of TEM. Since this is by definition “legacy data”, its basic character and quality cannot be changed and PEM must either accommodate its limitations or find alternative inputs.

Component 2 (inference or prediction) must accommodate and be appropriate to both available data inputs and output requirements.

Component 3 (data output) is (or should be) defined by the application or processes to which the output will be applied.

This report will deal only with:

1) documentation (meta-data) needed to provide an assessment of data quality and utility for input to predictive ecosystem mapping

2) meta-data required to evaluate the inference system and

3) meta-data needed to provide an assessment of the quality and utility of predictive ecosystem mapping outputs for business applications

To the degree that meta-data gaps have been identified by current PEM activities, possible methods of bridging these gaps will be explored.

It will not address issues related to the quality or the utility of predictive methods, nor will it address the quality or the utility of specific data inputs or outputs. The reasons for this are straightforward. First, the size of the contract precludes this much larger scope, and second, predictive ecosystem mapping is in its early stages in BC. Only one system is ready for operational application and has yet to be fully evaluated for its efficacy as a TEM replacement or alternative. Any evaluation of specific data input needs would be premature and any evaluation of specific outputs without reference to specific goals, objectives, and criteria would be inappropriate.

The use of the term “quality” implies the character and nature of the data. It includes the principles, concepts, and procedures used in the collection, analysis, and storage of the data as well as its reliability within the context of the definitions used in the inventory.

1.3. Approach

Available documentation and reviews for six predictive ecosystem mapping projects (either recently completed or underway) in BC were reviewed to identify data inputs and approaches. In addition, digital approaches to terrain and landscape analysis that may be relevant to PEM but are not currently being used were reviewed for data requirements. Included in the review were materials provided by BCMoELP and BCFS, proposals submitted for FRBC funding, World Wide Web articles from public, NGO, and private sector organizations, materials from the UBC library, and telephone interviews.

After this initial review, the elements of data quality relevant to predictive ecosystem mapping inputs and outputs were identified and their relevance discussed.

Based on the elements of data quality identified at the previous stage, survey and inventory procedures that have an impact on data quality are identified and the nature of the impact reviewed.

The next step identified the meta-data (data about the data) necessary to evaluate PEM input and output map products. In particular, data about the nature of the things mapped, the pedigree or history of the data, spatial referencing, data collection and compilation procedures, and information management procedures.

Based on the preceding foundation, a framework for the evaluation of input and output data quality was developed and presented. The framework, its criteria, and its rating system are explained. The framework was applied to a hypothetical map as a preliminary test of its approach and concepts.

Finally, a summary identifying the strengths and limitations of the framework is presented together with recommendation for its testing and refinement.

2. PREDICTIVE ECOSYSTEM MAPPING IN BC

As noted in Section 1.2 - Scope, the distinction between Terrestrial Ecosystem Mapping (TEM) and Predictive Ecosystem Mapping (PEM) is somewhat arbitrary. There is a large element of inference and prediction in all forms of natural resource mapping and inventory. Both TEM and PEM input use primary, secondary, or even tertiary inferences to label or classify polygons. The principal distinction appears to be that PEM uses a greater degree of automation and that PEM relies more heavily on information which is the product of a previous and hence uncontrollable inference process and less heavily on primary data.

2.1. DATA INPUTS

With the exception of VRIEM and OASIS, all of the PEM approaches reviewed use similar input data sources. All rely heavily on geographic information systems to perform spatial analysis. They are used to generate slope, aspect, and in some cases slope position from digital elevation and/or to overlay some or all of BEC, FC1, VRI, Terrain, and Soil maps to produce resultant polygons and associated data. The resulting overlay polygons are then classified into TEM-like units.

2.2. PREDICTIVE OR INFERENCE PROCEDURES

Two principal approaches are being used for predictive mapping 1) human inferencing utilizing informal knowledge bases and 2) automated inferencing using formalized knowledge bases. The formalized knowledge bases utilize one or more of maximum likelihood estimation based on “belief matrixes” with look up tables and/or rule based inference. EcoGen processes the matrices using multiplicative and additive functions but the EcoGen program is evaluating the ‘best’ approach, based on pilot projects. One application – ELDAR – uses a combination of maximum likelihood and rule based inference.

3. MAPPING CONCEPTS

3.1. What makes a good map?

- It provides the required information
- at the appropriate precision of taxonomic classes, attribute values, and map resolution
- with the required level of taxonomic purity and attribute variability
- at appropriate levels of reliability.

3.2. Reliability

3.2.1. Required levels of precision and accuracy

Reliability is a relative and context-dependent concept. It has two components: precision (the exactness with which measurements are recorded or predictions made) and accuracy (the closeness of the presented value to the true value). Predictions of high precision will be more difficult to make with high accuracy than predictions of low precision.

The required level of precision and accuracy will be a function of the interpretive or predictive procedures used and a function of the use to which the map will be put. If the predictive models used show large responses to small changes in input values, the requirement for precision is high. If the predictive models show small responses to large changes in input values the requirement for precision is low.

The required level of accuracy will be a function of the consequences of predictive errors. If the consequences of predictive error are low, the need for accuracy is low. If the consequences of error are high, the need for accuracy is high.

3.2.2. Approaches to Reliability

Reliability may be measured as a degree of overall similarity between predicted and found classes. This is intuitively appealing but carries the possibility of missing large and potentially serious errors in individual parameters although the overall similarity is high.

Reliability may also be measured as errors in individual properties that may result in interpretive errors and consequences. This has strong conservative appeal because it will identify single significant errors but may cause otherwise well-mapped areas with high levels of similarity to look bad.

The choice of reliability approach should be dictated by proposed applications of the map, the interpretive procedures used, and the consequences of error in individual interpretations. It is possible to use both approaches to reliability in the same project.

3.3. Map-entities

There is much confusion with mapping terminology in the literature. This report will therefore use the following definitions:

Entity: a class or type of thing involved in the mapping process. An entity may be a site, site series, a soil, a polygon, a plant community, *et cetera*.

Instance: a specific occurrence or example of an entity (e.g., the Princeton soil association, site series SwAt – Step moss, polygon 192)

Map-feature: a point, line or polygon representing a site, linear feature, or area on a map.

Mapping-entities: any entities used in the mapping process but not necessarily labeled on the map. For example, a series forms part of the definition of a soil association that is being mapped or a stratum forms part of a geological formation that is being mapped. The series or strata are not labeled directly on a map, rather the association or formation that they comprise is labeled. While the series and strata are mapping-entities (entities used in the mapping process), they are not map-entities (represented as labels on the map).

Map-entities: mapping-entities used to label map features (digital or analogue). In the discussion above, both soil series and soil associations are mapping-entities but if the map features are labeled with soil associations, the association is the map-entity. If the map features are labeled with series then the series is the map-entity. Map features may be labeled with multiple map-entities. If map-entities consist of single mapping-entities, they are said to form simple map-units. If the map-entities consist of multiple mapping-entities, they are said to be compound or complex (see below). Both mapping-entities and map-entities can be simple, complex, or compound.

Simple entities: are entities which, for the purposes of the inventory, are not defined as being composed of other entities. For example, a soil series is an abstraction of multiple soil polypedons but, if reference to the individual polypedons is not part of the inventory definition, the series is a simple entity. A tree species has many occurrences or instances, but if reference to individual trees is not part of the inventory definition, the species is a simple entity.

Compound entities: are defined as consisting of two or more entities. The compound entity inherits the properties of its members or components. In compound map-entities, the definition includes a predictable or derivable pattern of the member entities comprising the compound entity. Knowledge of the pattern and distribution of component units within compound units cannot be recovered unless recorded as part of a formal map-entity definition. It may be annotated as part of the map-entity symbol, or recoverable as a description from the report. Soil catena and soil associations are examples of compound entities.

Compound entities may show first order through n-order compounding. A first order compound entity is composed of simple entities. A second order compound entity is composed of first order compound entities and so on. The essential distinction is that the pattern or distribution of all component entities of whatever order can be predicted or derived.

Complex entities: are defined as consisting of two or more entities. The complex entity inherits the properties of its members. Unlike compound entities, the definition of a complex entity does not include a predictable or derivable pattern of the member entities that would allow their specific location at a larger scale. For example, the merging of adjacent map-entities into a single polygon because they could not be mapped at the inventory scale. There is no way of predicting, based on landscape information, where in the polygon each mapping-entity occurs.

Multiplex entities: are entities consisting of two or more compound or complex entities for which the pattern of all member components cannot be predicted. An entity consisting of two or more compound entities is multiplex if the distribution of the member compound entities cannot be predicted or derived.

Map Legends: During the 1980's, there was much debate in the soils community over open versus closed legends. Terrain maps (open) and many soil maps (closed) of the 1980s represent the opposing camps. In an open legend, the map-entities and their modifiers that are found in a map polygon are listed in the polygon label with some indication of proportion or dominance (e.g., rCv/gMbv//Rs). In other words, the polygon label describes the polygon and the legend provides definitions for the components of the symbol. In a closed legend, the polygon label represents an abstraction of a number of similar polygons and the legend presents a description of this abstraction. These descriptions, of necessity, provide less precise estimates of proportional content than can be assigned to a polygon using an open legend and the information detail lost in the generalization process cannot be recovered.

Map Reliability: The term map reliability is widely used but its use is informal. In this paper, the use of the term reliability will be restricted to thematic and spatial accuracy and precision[1]. As such, reliability is one component of data quality. Whenever appropriate, the individual components of reliability will be discussed.

[1] Throughout this paper the term precision will be used in the non-statistical sense of exactness of or how fine is the resolution with which a feature or property is described. While in some degree related, the statistical definition refers to the standard error or repeatability of measurement not the exactness or resolution of the measurement. Some statistical textbooks use the terms precision and reliability as synonyms. That is not our usage.

4. ELEMENTS OF DATA QUALITY

For the purposes of this report, data quality attributes refer 1) to attributes of data that will influence the suitability of input data sources for predictive ecosystem mapping and 2) attributes of predictive ecosystem mapping outputs for business applications. These attributes of data quality will deal with three basic components: 1) content, 2) precision, and 3) accuracy. The three concepts are applied to both thematic and spatial issues.

4.1. Thematic Issues

Thematic issues are related to the entities and to the information types and content describing the entities mapped. In particular, they are related to the concepts used in the definition and the mapping of entities and to the concepts and principles underlying the synthesis and generalization used in the preparation of the data input sources.

Categorical generalization is used 1) to extrapolate site specific data to a polygon map, 2) to display mapping individuals which cannot be portrayed at the scale of presentation, or 3) to portray patterns and relationships between mapping-entities which would not be evident from the display of individual mapping individuals. An example of categorical generalization is the mapping of soil associations. A soil association is a group of soil series (up to five) which occur in a recurring topographically determined pattern on the landscape (conceptually similar to the catena). The association may be mapped because individual series cannot be delineated at the scale of presentation, because the boundaries between individual series are too indistinct to map, or because the relationship between the soil members is important to interpretations.

4.1.1. Thematic Precision

Thematic precision[2] refers to the exactness with which mapping-entities are defined. For example, a soil entity may be defined as having imperfect drainage. This would be a more precise definition than if it were defined as having drainage ranging from moderately well to poor. Precision may refer to:

1. the exactness with which attributes defining mapping-entities are defined
2. the exactness of the attributes associated with a mapping-entity when it is found in a specific polygon as opposed to its general definition
3. the definitional exactness of the composition and distribution of mapping individuals in a compound map-entity
4. the exactness of composition and distribution of mapping individuals in a specific polygon labeled as a compound map-entity

[2] Note again our use of precision to mean exactness of or degree of resolution of measurement or description.

4.1.2. Thematic Accuracy

Thematic accuracy refers to the correctness of polygon labeling and is distinguished from but related to thematic precision. In simple terms, a polygon is correctly labeled if the attributes of the polygon fall within the defined attribute ranges of the map-unit and its components. There is generally an inverse relationship between thematic precision and thematic accuracy in mapping projects. The following results from two 1: 20,000 soil inventories indicate the potential magnitude of the problem. Table 1 shows the average polygon area occupied by property errors large enough to cause a significant interpretive error. Table 2 shows the average polygon area occupied by at least one of five property errors large enough to cause significant interpretive errors.

- **Table 1 Estimate and 95% Confidence Interval for the area of Dissimilar and Contrasting Properties**

Property	Area	Confidence Interval
CF	36%	29 - 44
Depth	32%	24 - 41
Drainage	19%	11 - 26
Surf Text	13%	7 - 19
Slope	8%	4 - 11

- **Table 2 Estimate and 95% Confidence Interval for the area of defined error classes**

Class	Area	Confidence Interval
Correct	17%	8 – 27
Similar	18%	9 – 21
Dissimilar	7%	2 – 12
Contrasting	61%	48 - 73
Diss+Contr	68%	57 - 78

4.1.3. Thematic Content

Thematic utility is influenced by the thematic concepts and content being portrayed in map form. At the highest level, thematic content identifies the resource being inventoried (e.g., forest cover, soil, and terrain). However, other elements of thematic content will influence thematic utility. For example, thematic utility will also be influenced by the landscape and resource concepts embodied in the map. In addition to identifying the resource type, thematic content should also identify the nature of the mapping-entities and map-entities used in the inventory together with any additional map symbol components which modify or further characterize the mapping and map-entities. The distribution, pattern, and relationships of mapping-entities within a compound map-unit can have a significant impact on interpretations and if so should be identified.

One practical aspect of the thematic utility of a map-unit is the interpretation of compound units. A soil map-entity may be a complex of three soil associations. Each soil association may be comprised of as many as five series. Therefore, a map polygon may have up to fifteen soil series present, each of which may have a significantly different interpretation for some application. Clearly, the map-unit is inappropriate to site-specific interpretation if the mapping individuals (soil series) have different interpretations.

4.2. Spatial Elements

4.2.1. Positional Precision

Positional precision refers to the exactness with which a location can be determined. Location specifications to degrees, minutes, and seconds to three decimal places are more precise than degrees, minutes, and seconds to one decimal place. With the widespread use of geographic information systems, there is a tendency to confuse the positional precision possible within the geographic information system and the positional precision at which the original input data was compiled. Geographic information systems can and do store location data to within millimeters of real word location coordinates but the actual input data may have had a compilation precision of only tens of kilometers.

4.2.2. Positional Accuracy

Positional accuracy refers to the degree to which map coordinates correspond to the real world coordinates of features shown on the map. As with thematic accuracy and precision, increasing the precision claimed reduces the magnitude of location error that will result in a defined error. Positional accuracy is often stated as the probability of a map feature being represented within a specified distance. For example US National Map Accuracy Standards for 1: 24 000 maps is a 90^{th} percentile error of approximately 12 meters.

A sometimes-overlooked influence on positional accuracy is the process of “cartographic enhancement”. Cartographic enhancement refers to the techniques used to portray thematic information that is either too detailed or too complex for presentation at the chosen scale of presentation. A simple example is the shifting of roads away from river features so that both may be visually distinct at the scale of presentation. While topologically correct (i.e., the road is on the correct side of the river), the presentation is spatially incorrect (i.e., the map distance between the road and river is incorrect). A full discussion of cartographic enhancement techniques is beyond the scope of this paper but it can introduce significant positional, distance, and area errors. A number of biological and land use inventories have been compiled against base maps that had taken significant cartographic license. The author is aware of two research projects where paper road maps were digitized for use as a base map and the thematic data then merged with TRIM.

4.2.3. Topological (Referential) Accuracy

Topological accuracy refers to the properties of points, lines, or polygons that are not normally affected by changes in size, shape or absolute position. For example, although a point is not located at the correct coordinates, if it is located within the correct polygon it is topologically correct with reference to the polygon. Topological attributes are always with reference to two or more features. Of particular concern are the topological attributes of inside, outside, left, right, contiguous, congruent, connected. Topological accuracy is normally much greater than positional accuracy and because maps are frequently used as a means of locating oneself in the field by reference to topological relationships, positional inaccuracies often go unnoticed. However, during the spatial overlay process, positional inaccuracy leads to incorrect attribution of the polygons resulting from the overlay process.

4.2.4. Spatial Utility

Spatial utility will be determined by how well the positional accuracy and precision and by how well the topological accuracy meet the needs of the inventory. Positional accuracy and precision will have a significant impact on the registration of multiple map themes. If positional accuracy and precision are inadequate, the overlay process will produce a significant number and area of artifact polygons that have no real world expression.

5. PROCEDURAL ISSUES INFLUENCING DATA QUALITY

The inventory and compilation procedures followed in the creation of land based resource surveys will strongly influence the quality of the data. Of particular concern are the survey and sampling procedures, correlation (quality control and quality assurance) procedures, and data management procedures.

5.1. SURVEY AND SAMPLING PROCEDURES

Survey and sampling procedures may range from formal grid surveys, where sample points are located at predetermined intervals on a predetermined grid and boundaries between units interpolated between sample points, through stratified random sampling where samples are located at random within predefined strata, to free survey with selective or modal sampling where boundaries are delineated and samples selected to represent the central concept of the map-unit being characterized. Hybrid procedures that mix free survey with systematic or stratified random sampling are also used.

5.1.1. Boundary location accuracy and precision

Free survey; because it delineates boundaries, will presumably produce higher accuracy boundaries than systematic sampling with interpolation. However, the entities being mapped, the imagery used, and the field procedures followed will strongly influence the accuracy of boundary location. Field traverses designed to verify boundaries and then characterize or confirm polygon content will produce more accurate boundaries than field traverses designed only to characterize or confirm polygon content.

Boundary location precision will be influenced by the scale and resolution of the base map and imagery used in the mapping process. Accuracy will be most strongly influenced by the nature of the entities mapped and the field procedures used. With the exception of bedrock geology and more recent "bio"-terrain maps, few land resource maps emphasize boundary validation or provide any indication of boundary accuracy or precision. Those that do should have cartographic conventions, field procedures, and data records to support the distinction.

5.1.2. Determination of thematic content

The issue of whether free survey with modal sampling produces more or less accurate thematic content than more statistically rigorous approaches has been the topic of much animated discussion in the BC natural resources inventory field. Despite the energy expended, neither side has changed its position. What can be stated with confidence is that statistically based sampling can be used to characterize the expected range of conditions and attach confidence levels to these estimates while free survey with selective sampling cannot. What can also be stated with confidence is that, if practiced well, selective or modal sampling will produce much narrower and potentially more misleading estimates of homogeneity than statistically based approaches.

5.2. Correlation (Quality Control / Quality Assurance)

Correlation normally refers to on-going quality control and quality assurance. It does not usually but can include formal procedures for assessing the reliability (accuracy and precision) on completion of the inventory. Formal correlation procedures can improve data quality and, if the procedure is formally documented and the results of the procedures recorded for each inventory, it can be used as an indicator of survey reliability. Correlation procedures can address the following data-quality issues.

5.2.1. Taxonomy

The correlator confirms the accuracy of class designations for the taxonomy or classification used in the inventory. This can be accomplished in two ways. The most common is for the correlation staff to visit pre-selected (by the surveyor or the correlation staff), pre-prepared sites to either decide ambiguous sites or to validate the surveyor's assignments.

5.2.2. Attributes

The correlator confirms the accuracy of attribute values recorded by confirming appropriate measurement techniques, field inference procedures (e.g., hand texturing and drainage assignment, nutrient regime assignment), and polygon composition.

5.2.3. Survey and sampling procedures

The correlator confirms conformance to specified survey and sampling procedures including ground and air verification design, sample selection for modal or selective sampling, conformance to inspection strategy and sample location for statistical sampling.

5.2.4. Documentation

Confirmation of conformance to documentation standards, including meta-data, inventory and sampling procedures, traverses, site inspections, sample locations, analytical procedures, imagery, base maps, thematic inputs, etc.

5.3. Data / Information Management

5.3.1. Data Documentation

Documentation / meta-data should be available for the inventory's supporting data. This should include the entities or things about which data is recorded (e.g., map-entities, mapping-entities, agencies, polygons, sites, *et cetera*), the relationships between entities, and the attributes of the entities and relations. This should include entity definitions, relation descriptions, attribute definitions, attribute domains, method and precision of analysis, and units of measure.

5.3.2. Data Integrity

Meta-data on the completeness, consistency, and currency of the data and information should be provided. This should include business rules governing data integrity (e.g., a site must be located in one and only one polygon) and the implementation of these business rules (e.g., manual validation, database integrity constraints, programmatic validation routines).

5.3.3. Input Constraints

Any constraints imposed on field data recording, including class limits, category limitations, and number of map-unit components should be documented.

5.4. Data Processing Issues

5.4.1. Map Medium

The map medium identifies the material on which working maps or images are presented. For example, blue print copies of the base used in the field will have significant spatial distortions as a result of stretching and irregular image transfer during the reproduction process. The map will suffer distortion during field use and during the compilation process in response to changes in temperature and relative humidity. Plastic or Mylar bases will suffer less distortion but will still suffer distortions due to changes in temperature during the compilation procedure.

5.4.2. Spatial Overlay

Spatial overlay refers to the superposition of two or more thematic maps to produce a resultant map in which boundaries from each contributing map are used to form polygons in the resultant map. Each polygon in the resultant map carries the identifier of each contributing thematic polygon. If spatial errors exist in either or both of the maps used in the overlay, the resulting polygons and their attributes will be artifacts of the overlay process and will not represent reality. The process is mechanical and is uninfluenced by the thematic content of the contributing maps.

Positional errors: If one or more of the maps contributing to the resultant map has positional errors then all resultants of the map or maps will inherit those positional errors.

Positional errors may arise from at least four sources:

- Errors in the base map used for mono-restitution. Mono-restitution is commonly used to refer to moving delineated features from aerial photography to a base map. More correctly, it refers to the process of converting spatial information to a common coordinate system. Errors may result from three possible sources:
  - the master base map compilation process
  - the reproduction process (e.g., printing distortions)
  - shrink / swell during mono-restitution
- Errors in the mono-restitution process
- Errors in determining the location during the inventory process
- Errors resulting from the sliver removal process used to reconcile apparent small polygon artifacts resulting from the overlay process

Topological errors: As a result of positional errors in one or more of the maps contributing to the overlay, the mono-restitution process may generate topological inconsistencies between the contributing maps.

5.4.3. Thematic Overlay

Thematic overlay refers to the combination of thematic information associated with each input polygon contributing to the resultant polygon in a spatial overlay.

Characterization errors: If contributing polygons have errors in the characterization of the feature being transferred, the error will be propagated to all of the resultants to which it contributes.

Reconciliation errors: The overlay of compound map-units on compound map-units presents the problem of reconciling which of the possible thematic combinations actually occur in the resultant. For example, the simple overlay of a compound polygon A^7B^3 with X^5Y^5 produces the following possible proportions for each thematic resultant:

- AX from 20% to 50%
- AY from 20% to 50%
- BX from 0% to 30%
- BY from 0% to 30%

There is no *a priori* method of determining the actual proportional representation unless some combinations can be shown impossible.

5.4.4. Combined Spatial and Thematic Errors

The combination of spatial and thematic overlay errors with compound map-units can result in compounding or reducing the errors associated each of the spatial or thematic errors. Depending on the nature of the compound map-unit and the distribution of components in the compound unit, the resultant polygons may have none, one, or more of the components of the contributing polygon represented in the resultant. Positional errors create artifact polygons that imply thematic combinations that do not exist.

The 1:20,000 soil maps available from the Salmon Arm Forest District represent the kind of issues that can arise. These maps are digitized from 1:50 000 manuscript maps (compiled on an unidentified base but probably 1:50 000 NTS topographic map sheets). Since the publication scale was 1:100 000, it is unlikely that positional precision and accuracy were commensurate with even 1:50 000 scale. The lack of documentation on the base map, positional accuracy of the base or thematic line transfer, or the method of conversion to TRIM bases make the positional accuracy suspect and its estimation impossible. The map-entity of the original survey was a polygon composed of from 1 to 3 named soil associations. Each association could have from 2 to 5 soil series as members, and the polygon could contain up to 30% unidentified inclusions. The resulting 1: 20 000 polygons were labeled with a maximum of three soil series or phases and estimates of the proportion of the polygon occupied by each series or phase was adjusted to sum to 100%. The nature of the original map-units and the lack of documentation on criteria for the selection of 3 series from a possible 15 series and up to 30% unidentified inclusions makes an estimation of thematic reliability impossible.

6. META-DATA REQUIREMENTS FOR DATA QUALITY

The following discussion deals only with meta-data useful to characterize data quality for input to predictive ecosystem mapping or to evaluate the quality of ecosystem mapping outputs for business applications. Additional meta-data elements may be useful for other applications and some of the elements discussed below will have other applications but these are beyond the scope of this report. This section will refer to entity and attribute definitions, pedigree, spatial referencing, inventory procedures, and data management and integrity.

The meta-data elements are grouped into five categories: 1) entity and attribute definitions, 2) pedigree, 3) spatial reference, 4) inventory procedures, and 5) data integrity. Entity and attribute definition is necessary to determine the relevance and appropriateness of the concepts and entities used in the inventory as inputs to predictive ecosystem mapping. It is also necessary to determine whether the appropriate attributes are available in the appropriate form, and measured by the appropriate methods. The second, pedigree, identifies the currency, the status, and the authority responsible for the inventory data. The third, spatial reference information is necessary to determining the accuracy and precision of the input data and to the reconciliation of different spatial data sets. The fourth, inventory procedures will provide an indication on the nature of boundaries and their accuracy and precision. In the absence of reliability measures, this information may be used to infer levels of reliability. The fifth, database information, identifies the contents of the database, quality assurance / quality control procedures used to ensure the integrity of stored data, and identifies data management constraints which may influence the reliability of data recorded. Each category and element will be described and its relevance to TEM or PEM discussed.

6.1. ENTITY AND ATTRIBUTE DEFINITIONS

Most common errors in the use of existing land based resource inventories arise from a lack of understanding of the concepts and characteristics underlying the mapping and map-entities portrayed by the inventory. The following meta-data set should provide sufficient information about the nature of the entities portrayed in the map for a knowledgeable user to accommodate limitations in the database or to apply procedures to adapt the data to the user's present need.

6.1.1. Entities

6.1.1.1. Entity types

All entity types used in the survey should be defined and described. Typical entity types used in land based resource inventories include: Sample site, inspection site, traverse, soil series, soil association, terrain unit, site series, map, map delineation, *et cetera*.)

6.1.1.2. Entity descriptions

The concept embodied in the entity should be described. For example, a forest stand is an entity used in forest inventory maps, landforms in terrain mapping, soils and soil associations in soil mapping. Compound entities like soil associations are often defined as recurring patterns or distributions of component members. Soil associations and BEC site series represent significantly different concepts of the landscape relationships embodied in each. FC1 polygons are compound entities composed of component instances of the entity tree species.

6.1.1.3. Entity definitions and limits

Both simple and compound entities will have attributes and criteria used to identify the type and class of entity and will have attributes used for distinguishing the boundaries of instances of entities. These attributes and criteria will be values or value ranges, business rules, or relationship rules that can be used to process the attribute values of an instance to produce a boundary or classification. These business or relationship rules should be documented as should the attributes and limits used.

6.1.1.4. Entity Relationships

Relationships define how entities are related to each other. Some examples of the many relationship types are:

- Associative – relationships identify a simple association of two entities. For example, site series may be associated with soil series. There is no implied precedence, membership, or organizational structure.
- Hierarchical – relationships define a strict membership. Classifications and organizational structures are common examples of hierarchical relationships. A common rule defining a hierarchical relationship is that a child must have one and only one parent but a parent can have any number of children.
- Definitional – definitional relationships identify how the members of a compound entity define the compound entity.

6.1.1.5. Entity / Relationship Attributes

The attributes used to characterize instances of entities or relationships and the attributes used to define classes of entities should be documented. Attributes may describe instances of entities or instances of relationships between entities. For example, in a FC1 polygon, the order of species listed implies cover dominance. The attribute dominance is an attribute of the relationship between species and polygon. In other words the species when it occurs in a specific polygon. The site value, unless explicitly specified, could characterize: 1) an entity (the polygon irrespective of species), 2) a relationship (the dominant species in the polygon or any combination of the species listed in the polygon), or 3) a reference species which may or may not be present in the polygon. Another example of an attribute of a relationship is the probability of two entities occurring together. The value assigned to the probability may differ depending on the direction of the relationship. For example, the probability of a site series A occurring on Soil X may be 50% but the probability of Soil X supporting site series A may be 70%.

Attributes that describe an entity or relationship can be usefully grouped into four kinds.

1. Definitive – attribute values or ranges of attribute values which define the entity and its class and/or which define the boundaries of an instance of the entity.

2. Diagnostic – attribute values or ranges of attribute values which are used to infer the entity type and its class and/or that are used to infer the boundaries of an instance of the entity.

3. Associative - attribute values or ranges of attribute values which are normally associated with the class of an entity but which are neither diagnostic nor definitive.

4. Accidental - attribute values or ranges of attribute values which are generally specific to an instance of a class and cannot be inferred from class membership. Modifiers, site symbols, and annotations are generally used to identify important accidental attribute values.

6.1.2. Attributes

Attributes used to describe or characterize entities or instances of entities should be fully defined. A full definition includes:

1. Definition and description – the formal definition of the attribute and a general description of the attribute. For example, pH: the negative log of the hydrogen ion concentration. A measure of the acidity or alkalinity of a material.

2. Method – the method of measurement or analysis. It is critical to define the method of measurement or analysis. The same attribute may be measured or analyzed in a number of ways with each method giving a different value. Using the example of pH, three of the methods for measuring the pH of a soil sample are litmus paper, a Helige triplex colour indicator, and a glass electrode. In addition to the indicator or measuring device being used, the method of sample preparation may also vary from moistening the sample through a 1:1 soil-water paste to a 1:5 soil-water paste. The nine possible variants will commonly give nine significantly different results. The reason for the multiple methods is that the answers given are appropriate to different interpretations. The appropriate use of the data value cannot be evaluated unless the full method is known.

3. Unit of measurement – the units of measurement used to specify the attribute value e.g., centimeters, microns, *et cetera*. In some cases, measurement will be unitless. In other cases, there may be multiple units of measure in common use. As with methods, the unit of measure must be known if the value is to be used appropriately. In some cases, the specification of method will define units of measure but in others, significant errors can result if the wrong units are assumed.

4. Precision – the exactness with which a value is measured and recorded.

5. Statistic – data reported to describe an entity or relationship may be a single observation, a list, the mean value of multiple observations, a measure of variability (e.g., standard deviation, maximum, minimum, skewness, kurtosis, *et cetera*). If any

statistic other than observation is reported, the number of observations on which the statistic is based should be reported. In addition, the sample strategy should also be reported (see Section 6.4 Sampling Design and Method).

6. Domain – the valid range of values that an attribute can take. The domain may be defined for the attribute in general (e.g., pH dimensionless values from 0 – 14) or for the value associated with a class (e.g., pH dimensionless values from 3.5 – 5.5)

6.2. Pedigree

Pedigree refers to the history of and the organization(s) responsible for the data.

Citation – supports a search for supporting information, may indicate peer review, and identifies the individuals responsible for producing the data.

Responsible Agency – may provide an indication of supporting materials such as standards and methods manuals,

Purpose of Inventory – may indicate the relevance of the inventory to potential applications not originally intended. The data element should include goals, objectives, interpretive applications, the level of planning or decision support for which it was developed.

Period of content – indicates the currency of the information content.

Period of compilation - provides an indication about methods, procedures, and inputs extant at the time of the inventory.

Status – indicates the stage of completion, degree of correlation, integrity checking, revision status, and availability.

Use constraints – indicate acceptable uses of the information. These may be recommendations from the responsible agency, recommendations from an external agency, or disclaimers of responsibility.

Native data set environment – identifies the computer systems, operating systems, and enabling technologies used in the compilation and maintenance of data. This may indicate know software problems, data format constraints, or data integrity issues.

6.3. Spatial Reference Information

Base Map – it has been said that the three most important things to know about spatial data integration are base maps, base maps, and base maps. The base map provides the spatial features used to orient and locate point, boundary, and feature locations. It is also used to determine the coordinates used to merge multiple themes. The spatial precision and accuracy of the base map(s) places an upper limit on the spatial precision and accuracy of the thematic information presented on the base map. Relevant information about the base map is:

    - Compiling Agency – may provide reference to the base map and compilation method used if this information is not available from the map. It may identify QA/QC procedures and standards.
    - Year of compilation – may be necessary to identify which base maps and procedures were being used by the compiling agency at the time of compilation.
    - Projection – is needed to adjust maps to a common projection.
    - Ellipsoid – is needed to adjust maps to a common ellipsoid.
    - Year and Number of Datum points – indicate the relationship between surveyed ground control points and the degree of coordinate inference required between datum points. Year identifies the procedure used to establish the datum points used for positional reference and the density.
    - Ground control (feature density and distribution) – significantly influences the accuracy and precision of the mono-restitution procedure if ortho-corrected imagery is not being used. The mono-restitution process involves the reconciliation of ground control features (hydrography, roads, *et cetera*) common to the base map and the interpreted images. The greater the density of ground control features, the less interpolation will be required in the location of features. The density and nature of the ground control features will also significantly influence the ability of the surveyor to locate map features on the imagery or maps being used.
    - Compilation method – will influence reliability. If the compiling agency does not report minimum standards for positional accuracy and precision, reasonable estimates may be inferred from the compilation method. Some plotters used to transfer features from imagery to base map adjust for camera tilt and relief, others for one or the other, and some neither. Many inventories conducted during and prior to the 1980's used an epidiascope that corrected for neither camera tilt nor relief. These plots contain serious positional errors especially in mountainous terrain. Other projects used the Kail radial planimetric plotter. This plotter corrected for relief but not camera tilt. Despite the topographic correction, significant errors still result in highly dissected, high relief terrain.
    - Compilation quality control / quality assurance procedures, criteria, and results – are correlated with overall data quality and indicate the level of understanding of the compiling agency.
- Point and vector object or raster object data structures – are needed to import digital data if there is no conversion utility between the source system and the importing system.
- Horizontal and vertical coordinate systems – are needed to convert or merge multiple data sources.

The issue of base map quality should not be underestimated. In the 1980s, when evaluating the overlay of a 1:15K forest cover map and a 1:20K terrain map, the BC Land Resource Unit of Agriculture Canada identified location errors of 200 – 400 meters. The errors resulted from the combination of spatial errors in the base maps and errors in the mono-restitution process when the maps were merged to a common base.

Input maps – Automated ecosystem mapping generally uses multiple input maps to compile and integrate the information needed to predict ecosystems. The quality of the final product will be dependent on both the quality of the input map (which cannot be improved) and the process of merging and reconciling spatial discrepancies in the multiple maps. In the BC context, this will generally involve merging and reconciling spatial data sources to the TRIM base map.

The crudest approach to this task is to identify control points common to both maps. One of many possible mathematical functions is applied to adjust the coordinate values from one map to conform to the chosen base map. This approach is less than ideal because the only points for which the required adjustment can be known with certainty are the selected control points. In addition, the adjustment of coordinates between control points is based a mathematical interpolation function which may not reflect the required adjustment. Furthermore, there is no certainty that any or all of the control points on the map being adjusted were positionally correct. While not ideal, this may be the only available procedure if appropriate meta-data on the input maps is unavailable.

Ideally, the maps to be reconciled should be adjusted to a common projection and ellipsoid before the reconciliation process. The spatial quality of input maps can be evaluated or inferred from the meta-data.

The same criteria applied to evaluating base maps should also be applied to evaluating input maps.

Remote Imagery – the characteristics of the remote imagery used to delineate thematic content will have a significant impact on the accuracy and precision of boundary and feature location. It will also have a significant impact on the reliability of thematic inferences. Meta-data on the imagery used should include the following.

- Time period – will indicate currency of the land use and vegetation cover inferences
- Cloud cover – will indicate the portion of the area for which visible coverage was available and the degree to which inferences may be confounded by cloud shadow.
- Scale – or more appropriately resolution indicates the level of detail observable and hence the information available for inference. This will influence tree species identification, mensuration, and other inference procedures.
- Medium – (prints versus transparencies) influences the level of detail recoverable. Transmitted light (positive or negative) images display much greater tonal range and hence more interpretive information than reflected light images (prints).
- Spectral characteristics – influence the type of information emphasized. Some thematic content is more easily inferred from different spectral contents. Tree species identification is easier with colour than panchromatic black and white and disease is often more easily inferred from infra red imagery *et cetera*.
- Positional control – (the positional accuracy of the photo centre, the camera tilt, the skew of the camera relative to the flight line, the focal length of the camera, and the elevation from which the image is taken) are needed for accurate transfer to an ortho-base. The accuracy and precision with which these attributes are controlled should be reported.
- Ground control features – (the nature, frequency, and density of ground control features such as hydrographic features, roads, peaks, *et cetera*) will influence spatial accuracy in two ways. It will influence the reliability with which the surveyors can relate their position on the ground to their location on the air photograph and it will influence the accuracy of boundary transfer to the base map during the compilation process. The fewer ground control features common to both photo and base map, the greater is the interpolation of position required.

6.4. Inventory Procedures

In the absence of formal determinations of map reliability, some inferences about the expected levels of accuracy and precision can be made from the procedures used to conduct the inventory.

Delineation method and criteria – two approaches can be taken to delineating areas as map-entities. One uses recognizable features on imagery, which are assumed to correspond to changes in one or more of the criteria for recognizing the map-entity (e.g., the factors influencing vegetation or soil development, changes in material deposition, *et cetera*). These delineations are then classified. An alternative procedure conducts ground traverses to identify entities and boundaries and then identify corresponding features on the imagery to delineate entities. In either case, the image criteria and the change in properties that they are assumed to represent should be documented.

Validation method and criteria – The method of validating inferred entities and their boundaries should be documented. Possible methods of are field traverses, point inspections, aerial inspection, road traverse, *et cetera*. In addition to the method, the criteria used for verification, and the validation design (selective, modal, stratified random *et cetera*), and the locations of points and traverses should be documented. Ideally, the results of the validation and any resulting changes to the map based on the validation results should be recorded.

Sampling design and method – sampling is distinguished from validation as the procedures used to characterize the nature and variability of the entities mapped. As noted in Section 5.1.2, many inventories use selective or modal sampling that should show narrower ranges of properties than statistically based procedures. The sampling design and selection criteria should be documented, as should the location of all sample points. In the case of selective or modal sampling, the selection criteria and in the case of stratified sampling the stratification criteria should be documented.

In addition to documenting the design, the sampling method should also be documented. Examples may be for vegetation, full relevé versus dominant species and for soils, interval versus horizon sampling. Documentation should consist of either project specific documentation or the published procedures that were followed.

Correlation Procedures – the correlation procedures followed in the inventory indicate the nature and level of quality assurance and quality control practiced. Correlation procedures are used to ensure conformance to specified inventory standards and procedures. Unless formal reliability estimates are part of the correlation procedure, the correlation results will only indicate that standards and procedures were conformed to at an acceptable level. The determination of accuracy and precision can be done by statistically sampling examples of correlated inventories to determine the range in reliability for surveys conforming to the correlation procedure.

Documentation of correlation procedures include:

- Taxonomy – the taxonomy / classification and version being correlated, the criteria and procedure followed, and the correlation results
- Attributes – the attributes correlated, the correlation criteria and procedure followed, and the results
- Survey and Sampling design – the survey and sampling design used in the inventory, the correlation criteria and procedures followed, and the results
- Documentation – documentation correlated, the criteria and correlation procedures followed and the results

6.5. Database Information

Most database discussions refer to digital or electronic databases versus analogue or paper databases. The distinction is arbitrary. The appropriate use of data requires a full understanding of the data including its content and limitations. The following discussion identifies the meta-data that should be available to support the appropriate use of any database, digital or analogue.

Data Dictionary – at a minimum, the data dictionary should record the name, location, and format of the data elements stored in the database. The utility of the data dictionary for determining data quality can be significantly increased with the inclusion of the additional meta-data identified in Section 6.1.2.

Logical consistency – meta-data should include documentation of any rules or validation procedures used to identify or prevent logical inconsistencies. For example, attribute values for an entity that are inconsistent with its class designation or improbable associations such as a fen on a well drained soil. It should also identify the nature, frequency, and occurrence of known inconsistencies.

Completeness – meta-data should report on the completeness of the data (e.g., the number or proportion of data records with values for core, highly desirable, or optional data values). It may also be useful to associate with each data record a value for completeness of data for various applications.

Pedigree – the pedigree of attribute data is analogous to the pedigree described in Section 6.2. In this case, pedigree applies to individual records and should include one or more of:

- date of last update or modification,
- reason for update or modification,
- agency or person responsible for the update,
- status of validation or integrity checking,
- *et cetera*.

7. EVALUATION FRAMEWORK

7.1. Concept

The framework is not constrained by scale or map type. It is designed to be applicable to a class of input maps, individual input maps, individual or groups of map-entities in individual maps or groups of maps, or to individual areas of individual maps.

The evaluation framework uses two evaluation matrixes, one for thematic quality and one for spatial quality. Each matrix will have rows representing data elements and columns representing a "quality rating" assigned to the data element for the data source being evaluated.

Both matrixes use an overall multiplicative rating system. The multiplicative approach was adopted because some information elements are essential to the evaluation. If an essential element is missing, it is not possible to provide a valid assessment of the data. Therefore, if a zero value is assigned to any of these elements, any multiplicative operation using that element will receive a zero rating irrespective of the values assigned to other information elements.

7.2. Evaluation Criteria

The following section will discuss evaluation criteria for assessing data quality. It will discuss the criteria, their purpose, and their interpretation. This report does not assign values to the various criteria for three reasons. The first is that it is beyond the scope of the contract. The second is that it would be overly presumptive of the author, and the third is that the values should change with the input data being assessed and the proposed application of the PEM being produced.

7.2.1. Thematic Quality

7.2.1.1. Entity Types

The utility of an input map for predictive ecosystem mapping will depend on how well the mapping-entities in the polygons of one theme can be related to the mapping-entities of congruent or partly congruent polygons in another contributing theme. Simple map-units present little problem since only one mapping-entity need be considered. First-order compound map-units present a greater problem. The compilation process must determine which mapping-entities from one theme correspond to which mapping-entities of the congruent polygon in another theme. This may be inferred if the pattern and distribution of the mapping-entities within the map-entities are similar or if there is a known correlation between mapping-entities of the different themes. As map-entities progress through second and n-order compound map-entities to complex map-entities, the correspondence becomes increasing difficult to resolve.

The entity type sets an upper limit to the quality of an input map source. Other thematic ratings will be between 0 and 1 and the entity type value will be multiplied by the other ratings constraining the rating to a maximum of that assigned to the entity type.

The framework assigns the highest utility value to simple map-entities and the lowest to n-order and complex polygons. The progression is probably non-linear. The example given in Section 7.5 shows a non-linear relationship.

7.2.1.2. Attribute Definition

The meaning of an attribute value is not clear unless the attribute value is fully defined. Section 6.1.2 discusses the full set of features required to define an attribute. As discussed in Section 6.1.1.5 attributes may describe both entities and relationships between entities. Attributes such as rooting depth or texture are intuitively straightforward. Attributes of relationships can be as well. For example, fidelity is a possible attribute of the relationship between thematic entities (e.g., Soil X is developed in Terrain class rCv 80% of the time). The distinction is that the attribute is shared by two entities. When reported, attributes of relationships should provide the same meta-data as attributes of entities.

7.2.1.2.1. Method

As discussed in Section 6.1.2, the value of a property assigned to an instance must be qualified by the method used to measure or analyze the sample. The value assigned to method in the framework ranges from 0 to 1 with 1 being fully qualified and 0 being so unqualified as to be of no use. For example, Site index is referenced to at a specific age (generally 50 or 100 years). If the site index is not qualified, a number from 0 to 1 should be assigned to represent the confidence with which the reference year can be inferred or the impact that errors in the possible ranges could have on use of the value.

7.2.1.2.2. Units of measurement

Units should receive a value from 0 to 1. A value of zero should be assigned to units if they are undefined, cannot be inferred with any confidence, and potential errors in assumed units will invalidate any interpretations made from this value. A value of 1 should be assigned if the units are fully defined. Otherwise, a value between 0 and 1, reflecting the confidence with which the units of measurement can be determined.

7.2.1.2.3. Status

Status reflects the degree of completeness of the data set. A value from 0 to 1 should be assigned to reflect the proportion of entities for which the specified attribute is available.

7.2.1.2.4. Characterization

Characterization refers to the aspect of population distribution being described and to how the distribution is derived. The examples offered below are not intended as recommendations for appropriate statistics to be used. Rather, they represent the reality of existing databases. These examples, presented in possible order of increasing value are:

Observation: a single observed, measured, or analyzed value.

Estimated value: a value based on expert opinion, extrapolation from another area, or informal observation.

Estimated variability: the range or variability of the attribute value as estimated by expert opinion, extrapolation from another area, or informal observation.

Mean: a value determined as the sum of observed values divided by the number of observed values (the arithmetic mean) reported without sample size.

Mean and Standard Deviation: the arithmetic mean and standard deviation reported without sample size.

Distribution: the arithmetic mean, standard deviation, and either the number of samples or the standard error of the arithmetic mean.

7.2.1.2.5. Currency

Currency refers to the degree to which the data reflects current conditions. This is especially important for themes such as land use, forest cover, range condition but can also be important in areas prone to active processes such as stream bed migration or mass wasting. As with other meta-data elements, the value should range from 0 to 1 depending on the utility of the information for the proposed use.

7.2.1.2.6. Domain

A domain is a set of business validation rules, formal constraints, and other properties that apply to an attribute or group of attributes. For example a list of valid values, a range, a qualified list or range, or any combination of these. The existence of these domain elements provides criteria by which to determine if the data is consistent with the domain. The lack of domain definitions while decreasing confidence in the quality of the data does not decrease it to 0. Therefore, the value assigned for domain documentation should range from greater than 0 to 1.

7.2.1.3. Attribute Reliability

The definition of an attribute identifies what the attribute means; reliability identifies how well it was mapped. The reliability of an inventory may be reported, or it may be estimated from inventory and measurement methods.

7.2.1.3.1. Reported Reliability

Reliability estimate: A principal element of concern in evaluating the quality of input and output information is the reliability (accuracy and precision) of the data and information. Estimates of reliability included with the inventory may be taken at face value for the purposes of this evaluation. The estimate should be recorded as a value from 0 to 1. If reported as a percentage value, it is a simple conversion from percentage. If reported as a qualitative estimate (e.g., high, medium, low, *et cetera*), it should be converted to the same (all be it inappropriately precise) scale. The reason for taking the estimate at face value is that it will be qualified by the method of estimation below.

- **Table 3: Reliability Methods**

Informal		A	
Correlated		B	
	Taxonomy		B1
	Attributes		B2
	Content		B3
	Boundaries		B4
Statistical		C	
	Transect		C1
	Proportional		C2

Reliability estimation method: As with any attribute, the method of estimation or measurement should influence our confidence in the estimate. A value between 0 and 1 will be assigned to the estimation method depending on the estimation procedure followed. This value will be multiplied by the reliability estimate above to provide the reliability estimate. Three basic reliability methods are recognized in Table 3. These are Informal A, Correlated B, and Statistical C.

Informal: the reliability estimate is not supported by formal procedures or criteria. This will normally be the lowest value. In the case of informal or undocumented reliability methods, the estimate will be ignored and a maximum value of .3 for detailed and semi-detailed and 0.5 for reconnaissance will be assigned to the reliability rating. While these ratings may seem counter intuitive, they are based on empirical data showing the reliability of reconnaissance maps to be higher than detailed maps. This can be explained by the fact that reconnaissance maps use more generalized concepts and are more heavily based on landscape patterns than on site properties. The detailed maps, whose reliability results were reported in Tables 1 and 2, had reported reliability estimates of 95% but measured reliabilities of less than 30 percent.

Correlated: the reliability estimate is based on formal correlation procedures. The procedures may include taxonomic B1, attribute B2, content B3, and/or boundary correlation B4. The numeric assignments for each of the correlation components should be additive (summing to a maximum of the value assigned to B correlated reliability). Estimates based on correlation procedures will normally be intermediate values.

Statistical: the reliability estimate is based on a formal statistical procedure. The procedures used should be assigned independent ratings. For example, fixed interval sampling along a transect bisecting a polygon does not have sampling probability proportional to area and may introduce systematic error. Sample size being equal, sampling with probability of being sampled proportional to area will provide a better estimate or test of composition than fixed interval sampling along a transect.

7.2.1.3.2. Method Based Reliability Estimates

If there are no estimates of reliability provided in the associated report or with the maps, estimates may be derived from the method of measurement or observation. It is beyond the scope of this report to provide reliability estimates for each method of measurement or observation used. However, the number of attributes and methods is limited and estimates can be assigned based on empirical data or expert consensus. For critical attributes, statistical determinations should be made.

7.2.1.4. Attribute Rating

The final attribute rating is determined as the multiplication of individual ratings.

Rating = Method * Units * Status * Statistic * Currency * Domain * Reliability Rating

The individual attributes may be multiplied by an importance rating and then summed and averaged to produce an overall attribute class rating. Either attribute class ratings or a weighted-average of class ratings multiplied by the map-entity rating will produce a map rating.

As noted earlier, the multiplicative method can yield a maximum value of 1 and minimum value of 0. If any component of the evaluation receives a 0, the overall rating is 0.

7.2.2. Spatial Quality Criteria

The spatial quality criteria refer exclusively to the reliability of coordinate data. It does not evaluate and is not influenced by the thematic content of the input map.

7.2.2.1. Documentation

Documentation refers to the identification of the projection system, spheroid, and geographic datum used in compilation of the base map. While not essential, this documentation greatly facilitates the integration of multiple spatial data sources into a single spatial data set. Because it may be possible to salvage undocumented spatial data through control point registration and "rubber sheeting", the evaluator may choose to use a value of greater than zero to avoid an overall zero rating.

7.2.2.2. Precision / Resolution

Precision is applied to both digital or analogue maps and remote images. It refers to the exactness of positional identification (e.g., plus or minus 10 meters). Resolution applies to remote imagery and refers to the fineness of detail that can be distinguished on the image. The value assigned should range from 0 to 1, and should reflect the suitability of the resolution or precision to the project requirements (e.g., is the resolution adequate to identify tree species if the project is to map forest cover).

7.2.2.3. Accuracy

Accuracy is applied to both maps and ortho-images and it refers to the correctness with which the true location of the feature is identified by the coordinates taken from the map or image. As noted earlier, there is an inverse relationship such that, as claimed precision is decreased, it is easier to obtain high levels of accuracy.

7.2.2.4. Ground Control

Ground control features have a significant impact on both the ability of the surveyors to relate their position on the ground to a position on the map and to the ability of the compiler to transfer image features to the digital or analogue base map. Other than that the value should range from 0 to 1 for reasons discussed earlier, this report does not propose a specific procedure for evaluating ground control features. However, such a procedure should consider the permanence, density, and distribution of ground control features common to both the imagery and the base map being used to compile the spatial data.

7.2.2.5. Working Medium

Working medium identifies the material on which working maps or images are presented. The rating, ranging from 0 to 1 should evaluate the appropriateness of the medium to the application. As with other criteria, it is inappropriate to assign a value to each possible medium in this report because the importance of these criteria will be project and application specific. It is, however, possible to rank medium in terms of the stability during the reproduction and compilation processes. For example, the following small sample of possible media is presented in increasing order of stability:

Photocopy < printed map < plastic base < temperature stable Mylar

7.2.2.6. Compilation method

The compilation procedures used for thematic map production should be documented and rated for positional accuracy by appropriate specialists. As with other attributes, the rating should range from 0 to 1 with 0 being unacceptable for the purposes of the survey and 1 being fully acceptable for the purposes of the survey. The following is only a partial list ranked in increasing order of normal spatial reliability of the product.

Visual < Pantograph < Epidiascope < Kail Potter < Zoom transfer stereoscope

7.2.2.7. Control Feature Congruity

As noted earlier, the source location errors can range from compilation base map errors through transfer errors to reproduction errors. Control feature congruity can be used to determine the magnitude and, if present the systematic nature of the discrepancies between a thematic input map and the project base map being used (in most cases TRIM). The evaluation is based on the assumption that the base map will not be altered and assumed correct for the purposes of reconciliation.

The control feature contiguity criteria should be developed in two stages.

Stage 1: Systematic errors, if present, should be identified and corrected. It may also be used to identify unknown meta-data for the base map on which the thematic input was compiled. For example, comparison of ground control features on an unknown projection, spheroid, or year of datum to the same features on a base of known projection, spheroid, and datum may be used to infer the unknown projection, spheroid, and datum.

Stage 2: Once this is done the magnitude of non-systematic errors should be identified and converted to a 0 – 1 scale where 0 is unusable and 1 is fully suited to inventory needs. The magnitude of acceptable spatial errors will be dictated by project needs (the need for topological versus location accuracy and the number and size of artifact resultants). However, a generic overall rating may be developed based on the magnitude of expected location error and the size and shape of the polygons to be merged (e.g., the portion of polygons expected to be congruent with their true position after merging with the base map).

The method for determining this criteria needs to be developed but should be based on an estimate of initial non-systematic error and the probable errors remaining after a "rubber sheeting" adjustment.

7.2.2.8. Rating

As with the thematic criteria, the final spatial rating is multiplicative.

Rating = Documentation * Accuracy * Precision * Ground Control * Medium * Compilation * Congruity

7.3. THEMATIC EVALUATION MATRIX

Table 4 is a template for the evaluation of thematic map data used as input to predictive ecosystem mapping or the evaluation of predictive ecosystem map output.

- **Table 4: Thematic Framework Template**

Thematic Content													
Entity Type		x											
									Reliability				
									Reported		Method	Rating	Attribute Rating
	Wt	Attribute	Method	Units	Status	Statistic	Currency	Domain	Estimate	Method	Rating		
	y		a	b	c	d	e	f	g	h	i	j	
Attribute Class 1	0-1	Attribute 1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	0-1	g*h or i	a*b*c*d*e*f*j
		...	...	...	...	...	...	...	...	...	...		...
		Attribute class rating = (Sum of class ratings/number of attributes) * Wt =											Class Rating
Attribute Class ...	0-1	...	...	...	...	...	...	...	...	...	...		...
			...	...	...	...	...	...	...	...	...		...
		Attribute class rating = (Sum of class ratings/number of attributes) * Wt =											---
Map Rating = Sum of Class Ratings * Entity Type Rating =													Map Rating

Step 1: Evaluate the map-entity type. The map-entity type (Simple, Compound, Complex) is assigned a rating depending on the intended application. This rating, “x”, on the template provides the upper limit on quality of the input map data.

Step 2: Attributes may be grouped into classes with each class given a weighting value commensurate with its importance to the inventory need. The sum of class weights should equal 1.

Step 3: The attributes are evaluated as described in Section 7.2.1.2 to produce an attribute rating. Reliability is a special subset in which the reliability estimate is multiplied by the rating assigned to the method. This product then provides the reliability rating used in the attribute rating calculation.

Step 4: Each attribute class recieves a class rating by summing the attributes, dividing by the number of attributes evaluated and multiplying by the class weight.

Step 5: An overall rating is calculated by summing the class ratings and multiplying by the entity rating.

7.4. SPATIAL EVALUATION MATRIX

Table 5 is a template for the evaluation of spatial map data and photo imagery used as input to PEM or the evaluation of PEM output.

- **Table 5: Spatial Template**

Spatial Reference							
	Documentation	Accuracy	Precision/ Resolution	Ground Control	Compilation Method	Ground Control Congruity	Rating
	a	b	c	d	e	f	g
Thematic Base	.5-1	0-1	0-1	0-1	0-1	0-1	a.b.c.d.e.f
Imagery	...	...	...	...	...	...	...

- **Table 6 : Hypothetical Thematic Application to a Soil Map**

Thematic Content													
Entity Type	3												
		Attribute	Attribute Definition						Reliability				Attribute Rating
			Method	Units	Status	Statistic	Currency	Domain	Reported		Method Based	Rating	
									Estimate	Prec			
Map entity (Polygon)		slope	0.8	1.0	1.0	0.3	1.0	0.9	-	-	0.3	0.3	0.06
		Surf Expr	0.8	1.0	1.0	1.0	1.0	1.0			1.0	0.9	0.72
												Class Rating =	0.39
Mapping entities (components)		texture											
		drainage											
		pH											
		C											
		Base Sat											
		C/N											
												Class Rating =	
Relationships													
member to polygon		proportion											
		fidelity											
members to member		topology											
		fidelity											
member to site series		fidelity											
member to FC class		fidelity											
member to Terrain class		fidelity											
member to veg type		fidelity											
												Class Rating =	
												Map Rating =	

The evaluation procedure is straightforward. Depending on how well it meets project needs, each attribute is assigned a rating ranging from 0 – 1. The overall rating is the product of the values identified below.

Documentation*Accuracy*Precision*GroundControl*Compilation*ControlCongruity.

7.5. Example Application

Table 6 presents the evaluation of two attributes of a **hypothetical** 1: 100 000 soil map. It is being evaluated for input to a 1: 20 000 predictive ecosystem map. The evaluation is not intended to portray the quality of existing soil maps nor to identify the methods used to produce any given soil map. The attributes evaluated and the values assigned to the criteria are for illustrative purposes only.

The following narrative will discuss the reasoning behind each assigned value in the framework.

Step 1: Evaluate the map-entity type. Map-entity type is a polygon labeled with a 1st order compound entity with associated polygon characteristics. It is assigned a value of 3 because the distribution of components on the landscape can be inferred and logical inferences about associated vegetation and terrain can be made.

Step 2: Identify attributes and attribute classes to be evaluated. In this example, three attribute classes are recognized.

Map-entity (polygon): The attributes of the polygon are grouped under attributes of the map-entity. They characterize the polygon as a whole, not the members of the 1st order compound entities being mapped and can be assumed to represent any or all areas of the polygon. As such, they represent a simple entity.

Mapping-entity (component): Attributes of the mapping-entities (component soils) are grouped under mapping-entity attributes. These ratings are for the attributes of the component not the distribution of components in the polygon.

Relationships: Relationships recognized in this example include the relationship of a mapping-entity (component) to the polygon it is found in. Attributes of this relationship are the proportion of the polygon occupied by the component and the fidelity with which the component is found within a polygon so labeled. Other relationships are the relationships between the soil components and other map themes (e.g., soil to terrain class, soil to site series, *et cetera*).

Step 3 Evaluate attributes

Step 3a: Slope

Slope is considered an important input datum for the estimation of sustainability and hazard evaluation.

Method: Assigned value = 0.8. The method of slope determination was unreported. However, from the normal methods during the period of the inventory it was probably interpolation from a 1:50 000 NTS topographic map.

Units: Assigned value = 1.0. The legend identifies slope recorded as percentage. It is fully qualified and there for receives full points.

Status: Assigned value = 1.0. All polygons have a slope estimate assigned and the project is considered complete.

Statistic: Assigned value = 0.3. There is no reported statistic. A range of slopes is reported for many polygons but there is no basis to infer whether the values represent a maximum and minimum or dominant slope classes on the landscape. Based on probable methods of measurement and the potential difference between minimum and maximum reported values versus dominant slope classes the statistic is assigned 0.3.

Currency: Assigned value = 1.0. Slope is a relatively stable attribute over the period of decades and active processes that may cause short-term change are recognized on the map symbol.

Domain: Assigned value = 0.9. The domain of slope is largely defined by the unit of measurement but slopes in excess of 100% are possible and the inventory provides no upper limit on expected slopes for the area to provide a validation check for recording or transcription errors. We infer these from the nature of the study area but recognize some potential for error.

Reliability: Assigned value = 0.3. The survey reported an estimated reliability of high (no numerical estimate). There was no reported method for determining reliability therefore method is assigned 0. This would result in a reliability rating of 0 that probably undervalues the inventory. The probable method of estimating slope was measurement of contours on a 1: 50 000 NTS map. By its nature, this scale presents generalized topographic information where extremes of slope and large changes over short distances are not recognized. Since the scale of the intended PEM is 1: 20,000, there is a scale mismatch. Combined with the probable need to recognize multiple site series in a single soil polygon, the slope estimate is of low value.

Rating: The attribute rating for slope and the soil map is 0.06 of a possible 1.0. This low value indicates that slope data from the soil map is of very low quality and is inappropriate as a source of data for 1: 20 000 predictive ecosystem mapping. This very low rating will influence both the attribute class rating and the overall map rating quite dramatically. It is appropriate to find an alternative source of input data for slope and delete slope as an attribute to be compiled from soil maps. However, it will be left as an attribute to be evaluated for the purposes of this exercise.

Step 3b: Surface expression

Method: Assigned value = 0.8. Method is not identified but surveys of this era generally used air photo interpretation either 1: 30,680 or 1: 63,360 panchromatic black and white aerial photography augmented by ground traverses.

Units: Assigned value = 1.0. Surface expression is a unitless attribute, therefore units may be considered fully qualified.

Status: Assigned value = 1.0. The survey is complete and all polygons are labeled.

Statistic: Assigned value = 1.0. There is no reported statistic. However based on the assumed method, the value is an observation of the entire polygon. The alternative statistics would be unlikely to provide additional information.

Currency: Assigned value = 1.0. Surface expression is a very stable attribute and therefore surface expression can be considered current.

Domain: Assigned value = 1.0. The domain for surface expression is specified by reference to the Canadian System for Soil Classification class limits and codes.

Reliability: Assigned value = 0.9. No reliability estimates are provided, and no correlation procedures used to QA/QC surface expression were in place or cited for this inventory. Therefore, reliability is inferred from the method used that in this case is considered very reliable if based on appropriate scale imagery and if class definitions are interpreted correctly.

Rating: Calculated value = 0.72. This is a reasonably high value indicated that the soil map is an appropriate source of information for surface expression.

Step 3c: Calculate the attribute class rating by adding the attribute ratings and dividing by the number of attributes being evaluated in the class. Using both slope and surface expression, the class rating is 0.39. If the soil map were being used only to compile surface expression, the class would receive a much more respectable 0.72 out of a possible 1.

Step 4: Calculate the Map rating. The map rating can be calculated as the sum of attribute class ratings divided by the number of classes or by summing the attribute ratings and dividing by the number of attributes evaluated. The class rating and the map rating will probably provide some overall guideline to map reliability but in reality, the important criteria are the values of each attribute for which the map is the best compilation source.

8. CONCLUSION AND RECOMMENDATIONS

8.1. CONCLUSIONS

Data quality issues, unrelated to the original purposes of potential PEM input maps, suggest serious problems with the application existing inventory data for PEM. The evaluation of potential data input sources requires a tightly reasoned, widely accepted framework to ensure acceptance of potentially negative conclusions. The framework for data quality evaluation presented in this report proposes such a framework. It is only a framework with suggested criteria. It is not proposed as a final procedure with definitive criteria nor has it been extensively tested. It does appear to have the following merits.

1. It provides a list of issues that should be addressed when determining the quality of data inputs or outputs related to predictive ecosystem mapping.

2. It provides a systematic process and suggests criteria for evaluating the quality of data inputs to or outputs from PEM.

3. It provides a framework for the creation, documentation, and implementation of data quality standards for predictive ecosystem mapping input and output products.

8.2. RECOMMENDATIONS

1. The framework should be tested and evaluated against a variety of possible PEM input sources and output products to:

    - identify and incorporate issues which may have been missed

    - test the appropriateness of the concepts and criteria used

    - refine the framework and criteria prior to presentation to a workshop involving practitioners of input mapping products, terrestrial ecosystem mapping products, and predictive ecosystem mapping products

2. A selection of potential data inputs should be evaluated for thematic and spatial accuracy in the field, the nature and magnitude of errors documented, and the implications of these errors on the quality of PEM output determined. The test areas should be stratified on results of the evaluation framework applied to existing data sources.

3. A workshop consisting of practitioners of input mapping products, terrestrial ecosystem mapping, and predictive ecosystem mapping should be sponsored to:

    - review and contribute to the framework

    - make recommendations on the adoption and use of the framework or a modified framework to establish and implement data quality standards for input and output products

4. If supported by the workshop, an expert committee should be struck to establish the criteria and ratings to be used as RIC standards for the evaluation of input and output data quality related to terrestrial and predictive ecosystem mapping.