NGDA Metadata Assessment Report, v1

(ISO 19139 Gap Analysis)

February 22, 2016

Produced for:

Produced by:

Team 

201 Loudoun Street SW
Leesburg, VA 20175

Contents

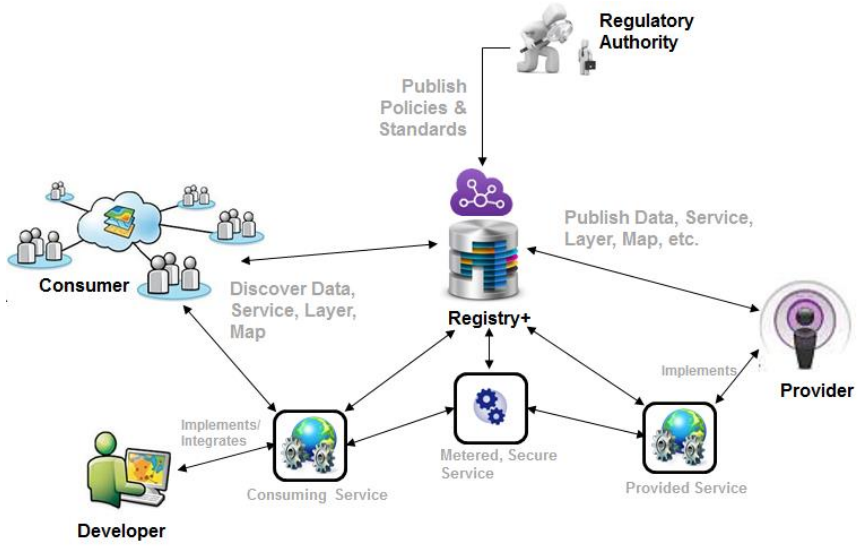
1. Overview

The National Geospatial Platform Initiative (NGPI, a.k.a. GeoPlatform) seeks to provide ready access to the nation's geospatial data and service assets. The nation's efforts to date have focused on making data assets available. A major part of this activity has focused on defining and publishing metadata for these assets. Strides have been made in adopting and employing metadata standards. Currently, the standard of choice for GeoPlatform is based on ISO 19139. While this standard adequately addresses many of the crucial requirements for finding and describing data, it falls short of meeting the needs of a seamless, service-centric experience through the GeoPlatform.

The main 2016 goal for GeoPlatform is to get National Geospatial Data Assets (NGDA) online and integrated with the platform. As such, one of the major technical objectives is to make sure that the GeoPlatform's new Registry+ capability (the hub of resource access and management) uses metadata that is both human-readable, so people understand its role and use, and machine-readable, so autonomous services understand its role and use. In a nutshell, we want to get beyond the point of just having a user read and interpret metadata. We want software to be able to read and interpret metadata. This requires formal semantics and enhanced metadata and service API standards.

A related technical objective is to not just build a register (catalog) for data assets, but also to build registers for services, layers and other first class GeoPlatform business objects. This is crucial to tackling service-enablement requirements and improving user productivity.

With these objectives in mind, we set about the task of auto-harvesting NGDA 19139s to build the means to automatically construct data, service and layer registers. In the process of doing so, we discovered the issues discussed in this document. We analyzed the issues to determine their impact on GeoPlatform experiences, and made recommendations concerning how to best address the issues. We also cited the benefits of addressing these issues.



2. Metadata Inconsistency and Semantic Issues

2.1 Inconsistent Resource Identification

Issue	**Inconsistent Resource Identification**
Description	There is no consistent way to define the identifiers for different resources (e.g. organizations, datasets, services, controlled vocabularies, etc.)
Why is it a problem?	Inability to link information and favor reusability. Resource information (concepts) are duplicated several times in different documents with variations of the same information. Update of information is more difficult to perform across all repositories. Need authoritative unambiguous references.
Recommendations	• Each resource should use a unique URI that is resolvable. • A policy needs to be put in place to manage the scheme of the different types of resources. • The maintenance of the information for each resolvable URI should be delegated to the authoritative party for the resource.
Benefits	A new policy to define URI Sets for US Government assets would provide a consistent means to make these trusted assets available for efficient, widespread discovery and re-use. This will encourage reuse and limit duplication. (Look at European Union policy.)

2.2 Inconsistent Use of Resolvable URI

Issue	**Inconsistent Use of Resolvable URI**
Description	Identifiers used in the 19139 document are often internal (e.g., a primary key in store implementation) and not accessible as unambiguous web resources.
Why is it a problem?	The lack of consistent machine-resolvable URIs impedes interoperability and limits automation (concepts must be grounded with unambiguous meaning for services to interpret and respond). Grounded URIs will also help humans better understand important concepts.
Recommendations	• Make links resolvable and semantically-grounded URIs with the right information to support human and machine exploitation (for controlled vocabularies, licenses, organizations, etc.) • Make the information accessible both for human consumption (HTML) and machine-understanding (Linked Data).
Benefits	• Enables the exploration of a “unified knowledge graph” that links and describes resources. Allows users to search, discover and navigate through “GeoPlatform Concept Space”, whereupon each concept is resolvable to a grounded (unambiguous) resource for consistent human and machine understanding. • Enables the decentralization of the management of resources, where each agency is in charge of their resources and associated concept space.

2.3 Lack Multilingual Support

Issue	**Lack Multilingual Support**
Description	The current standard does not enable the support of translations of human readable text in multiple languages. Language is handled at document level, not field level.
Why is it a problem?	There is no framework in place to handle translations at the field level for the ISO metadata.

Recommendations	Favor an implementation that natively provides multilingual support (such as Linked data) or provide guidelines for how to handle multiple languages (e.g., through JSON protocols).
Benefits	• Enable crowdsourcing of translations to authorized parties. • Enable access to US data using different languages.

2.4 Lack URL-based External Resource Descriptions

Issue	**Lack URL-basedExternal Resource Descriptions**
Description	A number of properties are referring to external resources (homepage, landing page, online resource for contact, page about document, reference to metadata document). Standards such as POD model these resources using a simple URL assigned to a property. This helps the user understand the role and meaning of the URL.
Why is it a problem?	External resources modeled as URL reduces the need for capturing additional inconsistent information, and leads to improved grounded meanings for metadata properties. Grounds the role and meaning of the external (auxiliary) resource in the context of a given resource.
Recommendations	• Model external resources as objects when their role is ambiguous. • If the property referring to a resource URL is unambiguous (homepage), use URL directly.
Benefits	• Ability to extend the description of external resources in the future • Provide more contextual information for external resources

2.5 Invalid XLinks

Issue	**Invalid XLinks**
Description	For some of the ISO 19139, xlink:href are not valid URLs (example #FS Lower 48)
Why is it a problem?	The ISO 19139 documents with invalid Xlink references do not validate with XML schema validator.
Recommendations	• Comply to standard XML Schema for xlink:href using URL
Benefits	Correct validation of ISO 19139

2.6 Lack Controlled Vocabulary Management

Issue	**Lack Controlled Vocabulary Management**
Description	• Controlled vocabularies are not made publicly available or resolvable (e.g., where is the National Map Theme Thesaurus?) • Lack unique identifier for controlled vocabulary (e.g., GCMD, Global Change Master Directory) • Lack unique identifier for keyword concepts (e.g., Paris, France) • Duplication of concepts (keywords) from different taxonomies, e.g., National Map Theme Thesaurus contains “Elevation” and NGDA Portfolio Theme refers to it as “Elevation Theme”. Are they the same concept and meaning? • Tendency to use alternative spellings for same concept (e.g., US and United States)

Why is it a problem?	• Not machine-readable, so automation is hindered • Can't perform enhanced semantic search • Lack consistent use of concepts (keywords) across 19139s • Ambiguity in the meaning of concepts (lack of grounded concepts)
Recommendations	• Define concepts in SKOS encoding with unique identifiers that are resolvable • Group alternate labels or translations under same concept • Provide SKOS mappings to other vocabularies to enable semantic search across GeoPlatform concept space • Make controlled vocabularies publicly available and uniquely identified with a resolvable URL.
Benefits	• Favor reusability of controlled vocabularies • Less verbose document • Unambiguous interpretation of key concepts • Inference enabled by using standard SKOS semantics (semantic search) • Enables multilingual search by concept

2.7 Limited Number of Keywords Types

Issue	**Limited Number of Keyword Types**
Description	The list of keyword types in ISO 19115 is limited to a few categories (discipline, strata, topic, place, temporal).
Why is it a problem?	Inability to accommodate new types of concepts such as audience, function, subject, topic, etc.
Recommendations	• Provide a mechanism to extend the list of keyword types in ISO 19115 using SKOS controlled vocabularies • Make the keyword types controlled vocabulary, which are uniquely identified and resolvable • Refer to the keyword type by resolvable URL
Benefits	• Provide an extensibility mechanism to accommodate other types of concepts (Audience, Function, Purpose, etc.). • Favor reusability of keyword types

2.8 Keyword Labeling Inconsistencies

Issue	**Keyword Labeling Inconsistencies**
Description	In some instances, multiple labels are encoded in one keyword (e.g., list of all US states for one keyword).
Why is it a problem?	While this fine for doing lexical-based text search, it is not sufficient in support of semantic search, where each concept must be grounded to a unique meaning.
Recommendations	• Each keyword should refer to one concept only • Use URI to refer to concept, in addition to label
Benefits	• Less verbose document • Inference enabled by using standard SKOS semantics

2.9 Lack Authority for Controlled Vocabularies

Issue	Lack Authority for Controlled Vocabularies
Description	The ISO 19139 uses the list of topic categories in the standard ISO 19115. There is a SKOS encoding available in the European Registry located at: http://inspire.ec.europa.eu/metadata-codelist/TopicCategory.
Why is it a problem?	Need an ISO or US host for our controlled vocabularies. Do we want the Europeans controlling the registration and maintenance of the vocabulary?
Recommendations	• GeoPlatform needs a registry for controlled vocabularies that are reusable across government agencies. • FGDC could host controlled vocabularies encoded in SKOS (currently only a GML document is hosted by Inspire folks). • The mapping to Registry+ uses the European Registry URI (above) to reference dcat:theme. This should be hosted at GeoPlatform.
Benefits	• The authority defining the standard maintains the taxonomy[1]

2.10 Inconsistent Place Name Encoding

Issue	Inconsistent Place Name Encoding
Description	ISO 19139 uses a keyword to define place names, and references a thesaurus name that is not accessible online. There is no consistent way to define place names and resolve ambiguities.
Why is it a problem?	The place name can be ambiguous as there are many places with the same names (e.g. Leesburg, FL versus Leesburg, VA)
Recommendations	• Use unique resolvable identifier (URI) to define place name along with a human readable name • Provide human readable page for place name URI and Linked Data representation, with partonomy[2] relationships, i.e., a semantic gazetteer • Use gazetteers where each unique place name has a resolvable URI • Use well known gazetteers (e.g., Geonames, GNIS)
Benefits	• Each place is unique • Reusability of place names from well-known gazetteers • Spatial reasoning is enabled by partonomy relationships • Remove redundancy, ambiguity and resulting user confusion • Favor reusability of authoritative gazetteers

2.11 Inconsistent Contact Point Encoding

Issue	Inconsistent Contact Point Encoding
Description	Contact Point in ISO 19139 is not systematically encoded in the document. The individual's name is required in POD but is not always present in the ISO document. A generic email reference for the contact role is sometimes used.
Why is it a problem?	When a problem is present in the metadata, a contact point with email should be available for expedient resolution of issues.
Recommendations	• Enforce Contact Point for every Resource with email and role name and individual name. • Email associated with contact point should be assigned to a role, not a specific individual.
Benefits	The use of a generic role-based email for the contact will smoothly handle staff changes.

[1] Taxonomy – a categorization of objects or entities that is based on discrete sets of classes, subclasses and types
[2] Partonomy - a type of hierarchy that deals with part–whole relationships

2.12 Responsible Party Sometimes Lacks Role

Issue	**Responsible Party Sometimes Lacks Role**
Description	Some responsible parties are published without a role, while the standard indicates that the role is mandatory.
Why is it a problem?	Without role, we are unable to understand the role of each party for a data source.
Recommendations	Enforce role in ISO 10139 for each responsible party.
Benefits	Unambiguous role of each party.

2.13 Lack Machine-Readable Way of Encoding Responsible Party Role

Issue	**Lack Machine-Readable Way of Encoding Responsible Party Role**
Description	ISO 19139 defines a well-defined taxonomy for Responsible Party roles (e.g., Publisher, etc). 19139 refers to a URL through an Xpointer to a GML document which contains roles and many other concepts.
Why is it a problem?	• Information conveyed in a GML document cannot be interpreted automatically; Need custom code to interpret the XML schema. • Xpointer URL cannot be used in the context of Linked Data. • In order to understand the meaning of a role, an unambiguous machine-readable description and human-readable page needs to be provided for each role.
Recommendations	Encode the role taxonomy in SKOS (machine-readable) and use resolvable URI for roles.
Benefits	Machine and human can understand the unambiguous meaning of the concept.

2.14 Lack of Support for Organization Hierarchy

Issue	**Lack of Support for Organization Hierarchy**
Description	ISO 19139 does not provide support for subOrganizationOf property (recommended by Project Open Data).
Why is it a problem?	• Difficult to understand the hierarchy of organizations • Search within hierarchy of organization is broken.
Recommendations	• Add a suborganization property to Registry+ • Make the organization resolvable to a URL that provides a machine-readable definition of the organization
Benefits	When search of resources is performed for a given organization, the hierarchy can be leveraged to search within suborganizations too (using transitive inferencing).

3. Service Access Issues

3.1 Inconsistent Use of onlineResource Information (Ambiguous Service Identification and Distribution)

Issue	**Inconsistent Use of onlineResource Information (Ambiguous Service Identification and Distribution)**
Description	In some documents, the link to services and distribution (zip files) are put in responsible party contact information (onlineResource), instead of TransferOptions in Distribution, or in ServiceIdentification
Why is it a problem?	The semantic of the onlineResource in ContactInfo is misused.
Recommendations	• Enforce a consistent way to encode distribution and service description • Clarify the role of onlineResource in ContactInfo
Benefits	Consistency of description of services and distributions in ISO 19139

3.2 Lack Standard Way to Reference Service API

Issue	**Lack Standard Way to Reference Service API**
Description	Lack standard way to refer to applicable services API standard, e.g., WMS, WFS, ArcREST
Why is it a problem?	There is no systematic and unambiguous way to identify web services standards. Version of standard is often not clear (OGC:WMS). Smart software agents, assisted by people, need to resolve spec confusion.
Recommendations	• Service API should reference authoritative spec URI to remove any ambiguity. • Make the URI of the referred standard resolvable (example: http://www.opengis.net/spec/wms/1.3)
Benefits	Proper classification of service standards, disambiguation, and autonomous operations support

3.3 Need Machine-Readable Service API Specification

Issue	**Need Machine-Readable Service API Specification**
Description	Lack of best practices to refer to machine-readable API specification (RAML, ALPS, Swagger, WSDL, etc.)
Why is it a problem	• The standard is not up to date with the best practices currently used in the industry, i.e., REST based API with machine-readable API specifications. • Specifications are defined as free text… not suitable for machine.
Recommendations	Registry+ accommodates machine-readable API Document.
Benefits	Integration to service API can be automated.

3.4 Inconsistent Service URL

Issue	**Inconsistent Service URL**
Description	The access URL for a service is not consistently encoded. For example in a WMS, some URIs point to a GetCapabilities endpoint, while others point to the base URL of the service
Why is it a problem?	There is no systematic way to access the service endpoint for a given service. Software agents must analyze the URL to get a normalized form, and this can lead to invalid endpoint and disruption of user activity.
Recommendations	• Use the base URI for a service • Provide reference to a machine-readable API document.
Benefits	Systematic access to service endpoint.

3.5 Insufficient Service Metadata

Issue	**Insufficient Service Metadata**
Description	The service description associated with Datasets have very little metadata, usually limited to an accessURL and format.
Why is it a problem	There is not enough metadata to enable the discovery of services and access to key service resources (e.g., layers for a WMS). The service identification information are sometimes too abstract for service operations to be leveraged by modern tools at runtime.
Recommendations	• Use the base URI for a service • Define a rich metadata model for services and coupled resources • Provide reference to a machine-readable API document or standard
Benefits	Enable the discovery and invocation of services in automated way.

3.6 Ambiguous Pairing of Format and OnlineResource

Issue	**Ambiguous Pairing of Format and OnlineResourc**
Description	The ISO standard decouples Format and OnlineResource. One format can have more than one online resource URL.
Why is it a problem	Having multiple URLs for a format is ambiguous and not friendly to machine or user.
Recommendations	Enforce parity of OnlineResource with format as best practice.
Benefits	Unambiguous pairing of format with online resource.

3.7 Lack Distinction between Download Format and Service API

Issue	**Lack Distinction between Download Format and Service API**
Description	The ISO standard does not clearly distinguish download file format versus service API in Dataset distribution.

Why is it a problem	Classification of services versus downloads is difficult…. and not friendly to machine or user.
Recommendations	• Improve standard and best practices to make clear distinction between service and download. • Provide a rich description of services.
Benefits	• Better classification of different distributions of datasets. • Support autonomous operations

3.8 Inconsistent Service Format Description

Issue	**Inconsistent Service Format Description**
Description	There is no consistent way to define format of services (OGC:WMS). Usage of mime type is not consistent in the standard. Formats are described mostly for human consumption, not machine consumption.
Why is it a problem?	Inconsistency of format description makes it difficult for software agents to access data in an automated way
Recommendations	• Use of standard URI for referring to standard service API • Use of MIME type from IANA to refer to representation formats.
Benefits	Enable automation, content negotiation and service selections based on controlled vocabularies.

3.9 Insufficient Map Layer Description

Issue	**Insufficient Map Layer Description**
Description	The ISO standard does not provide enough information to map a dataset to a layer in a map service (WMS, ArcREST). Often multiple layers are provided by the map service and there is no deterministic way to find out which one corresponds to the dataset.
Why is it a problem?	Traceability from dataset to map layer is broken. Layer metadata is lacking to support GeoPlatform search, discovery and proper use.
Recommendations	• Define a richer description of services/layers and make this available through Registry+ • Define a new standard to describe layer metadata, with commensurate best practices and policies.
Benefits	Support vastly better resource search, map building and other GeoPlatform experiences.

3.10 Data-Centric Approach Limits Autonomous Service Operations

Issue	**Data-Centric Approach Limits Autonomous Service Operations**
Description	A data-centric approach to metadata schema standardization is taken. This employs a syntactic approach, which imposes strict (inflexible) adherence to the standard. Semantic integrity suffers.
Why is it a problem?	• Data schemas have limited expressiveness. • Data schema captures only the syntactic and structural constraints of data model, but does

	not provide machine-readable conceptual model and business rules. Implementers are required to hardcode the rules with the risk of having different interpretations of a written specification of the rules. • Evolution of the domain model and associated software is difficult when using data-centric approach due to the fact that business rules and semantics of the data model are hardcoded into services (a brittle approach). Any new changes in the standard require update of software. Very often evolution of the data model requires building consensus and standardization, which can be a very lengthy process. • Integration and interoperability with other domains is difficult due to the heterogeneity of data schemas and business models, the lack of common protocols and machine-readable conceptual model and business rules.
Recommendations	• Use a semantic-based approach to embrace the heterogeneity of domain models by providing a common formal, sharable framework mechanism to easily extend resource metamodels for accommodating specific needs, while benefiting from core model elements, patterns and best practices. The extensions can be done in a decentralized way without breaking existing infrastructure. • Use of Linked Data standards (such as OWL) provides a standard-based mechanism to ground formal concepts for improved machine-readable operations. • Use ontologies to provide a framework to extend metamodel and associated vocabularies to accommodate the needs of different domains.
Benefits	• Decentralized extension of the Registry+ model • Handles specialized resource model profile to accommodate model specifics • Shareable and machine-readable model and business rules • Exchangeable machine-readable rules and conceptual models, which allows greater automation and reduction of code • Unambiguous interpretation of domain model • Cost reduction in software updates • Software adapts and evolves to accommodate domain model changes with no or minimal impact on existing code base • Decentralized and organic evolution of domain model • Software can adapt quickly to model and business rule changes.