TSMAD26/DIPWG5-11.7E

Paper for consideration by TSMAD

Uniform Resource Identifiers for S-100

Submitted by:	Jeppesen
Executive Summary:	This paper proposes that S-100 provide for the use of uniform resource identifiers as standard identifiers for S-100 artifacts, data objects, and resources.
Related Documents:	(1) S-100 Ed. 1.0.0; (2) S-100 change proposal
Related Projects:	(1) S-100

1 Introduction

S-100 product specifications and e-navigation data products and applications are likely to need persistent identifiers for different kinds of items, including definitions and services as well as data – catalogue entries, schemas, code lists, data services, and documents, as well as geospatial feature data. Information may be of various types (feature object data, information object, feature or information class specification from a feature catalogue, external web page, etc.) and formats (ISO 8211, GML, TIF, XML, etc.), and be available from various data stores and services – database, XML file, web page, web service, etc. S-100 Edition 1.0.0 does not define object identifiers though S-57 ENCs use a feature object identifier and other S-10x data products will probably use equivalents. This paper proposes that S-100 add provisions facilitating the use of uniform resource identifiers (URIs) as globally unique, persistent identifiers that enable different S-100 data products to harmonize data and maintain its validity through production, distribution, and use.

2 Terms and Abbreviations

DOI	Digital Object Identifier
GML	Geography Markup Language
IETF	Internet Engineering Task Force (organization)
INSPIRE	Infrastructure for Spatial Information in the European Community
OGC	Open Geospatial Consortium
OWL	Web Ontology Language
RDF	Resource Description Framework
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
W3C	World Wide Web Consortium
XML	Extensible Markup Language

3 References

IN.D.2.5: D2.5: Generic Conceptual Model, Version 3.4rc2. INSPIRE draft document D2.5_v3.4rc3, 05 April 2012.

IN.D2.7: D2.7: Guidelines for the encoding of spatial data, Version 3.3rc2. INSPIRE draft document D2.7_v3.3rc2, 15 June 2012.

ISO 26234: Information and documentation – Digital Object Identifier system. (2012).

RFC2141: URN Syntax. R. Moats, 1997. URL: http://www.rfc-editor.org/info/rfc2141.

RFC3986: Uniform Resource Identifier (URI): Generic Syntax. T. Berners-Lee, R. Fielding, L. Masinter. Internet Standard 66, IETF, 2005. URL: http://www.rfc-editor.org/info/std66.

S-100: Universal Hydrographic Data Model, Edition 1.0.0, January 2010.

S-100-INF: IHO S-100 – the Universal Hydrographic Data Model. R. Ward and B. Greenslade; IHB, January 2011.URL: http://www.iho.int/iho_pubs/standard/S-100/S-100_Info.htm.

TSMAD25-4.3.2: Revisions and Extensions to S-100 Edition 1.0.0, 2013. Paper at TSMAD 25, Jan. 2013. URL: http://www.iho.int/mtg_docs/com_wg/TSMAD/TSMAD25/TSMAD25-4.3.2_S-100_GapsandExtensions-Final.pdf.

4 Discussion

4.1 Uniform resource identifiers, uniform resource locators, and uniform resource names

Uniform resource identifiers (URIs) were originally developed for the World Wide Web but are also used in other software technology. A URI is a string of characters used to identify an abstract or physical resource, and conforming to the generic syntax and semantics defined in [RFC3986]. A "resource" is defined, somewhat circularly, as anything that can be identified, named, or addressed, for example web pages, web services, human beings, and corporations [RFC 3986]. A resource is not necessarily accessible via the Internet.

Uniform resource locators (URLs) are URIs which include addressability information[1] and thereby carry the connotation of acting as locators (e.g. http://www.iho.int/ both identifies the IHO home page and locates the page via the normal meaning and implementation of the HTTP network protocol). Uniform resource names (URN) need not include addressing (e.g., URNs such as *urn:ogc:def:crs:EPSG::4326*[2]). The syntax for URLs and URNs is similar but URNs must begin with "urn:" followed by a namespace identifier, and obey slightly different lexical constraints [RFC2141].

The generic URI syntax consists of a hierarchical sequence of components. A typical form is:

<scheme>://<authority>/<path>/<query>#<fragment>

e.g., http://registry.iho.int/hydro/feature/AnchorageArea as an identifier for the current definition of the AnchorageArea concept in the feature concept dictionary of the GI registry (note this is different from the URL locating the definition, which is currently http://registry.iho.int/s100_gi_registry/FeatureConceptDics/feature%20frames/fdd_current_detail.php?recordID=2002653&register=20&status=2);

and urn:iho:101.1.0.550.377889955 as an identifier for a specific anchorage area feature produced by NOAA (producer code 550) conforming to version 1.0 of the S-101 product specification (an HTTP URI can also be defined and used, e.g., http://location.noaa.gov/101/1/0/377889955 but would carry the connotation of an addressable Internet resource, i.e., a web address that is expected to reply with some kind of relevant information when queried with the HTTP protocol).

URLs and URNs are used in different ways, not just as locations of resources, e.g., XML schemas use them to specify namespaces.

4.2 The case for URIs as persistent identifiers

The argument for uniform resource identifiers in S-100 data consists of two parts: (a) the need for persistent global identifiers, and (b) the case for using URIs for these identifiers.

Persistent global identifiers are needed in order to maintain data object identity as data objects pass through the data chain, are stored in different data stores, transformed to different formats, and re-purposed for different domains. The same chunk of information may be present in different data stores in different formats (ISO 8211, XML, relational database record, etc.). Using a single identifier for the same chunk of data in all formats and stores will obviously help harmonization, validation, and tracking of data across multiple application domains and at different places in the data supply chain. Similarly for data integration, especially references to features in a different data product and data set from the referring feature, require persistent identity. For example the use of ENC feature spatial objects for locating nautical publications information, such as the regulations applicable to marine protected areas or passing rules applying in specified parts of a maintained navigable channel.

[1] E.g., the "http" part indicates that applications searching for the resource use a specific networking protocol to fetch the resource from the internet host named "www.iho.int."

[2] Defined by the OGC as a name identifying the EPSG:4326 CRS.

Guidelines prepared by the INSPIRE project strongly recommend unique identifiers for spatial objects where references from other spatial objects are expected [IN.D.2.5].

Several types of data objects in the IHO domain already use global persistent identifiers: geographic features are identified by a feature identifier (FID), and in the ISO 8211 encoding by the feature object identifier field (FOID), which uses the FID; navigation aids in a List of Lights are assigned a light list number unique within the country, some also have an "international number"; and the producer codes assigned to ENC producers. Given the wide use of global persistent identifiers in probable S-10x data products, it is necessary to harmonize the structure of persistent global identifiers while still allowing each application domains to define rules for generating a domain-specific component of identifiers (such as the S-57 "agency/feature ID/feature ID subdivision" rule for geographic features).

A common scheme for global identifiers is needed because if every domain or data product defines its own unique architecture for identifiers, it adds avoidable complexity and potential failure points to data formats, applications, and production tools.

Using uniform resource identifiers (as opposed to some other scheme for structuring identifiers) is, at the most basic level, a method of specifying the identification scheme and resolution protocol as part of the identifier, in a widely-known form. For example, instead of identifying an spatial feature by its "long name" or combination of identification number, producer, and feature subdivision, (e.g., LNAM=022613FD373E0F6C; or FIDN=335361854, AGEN=550, FIDS=3948), an identifier such as "urn:iho:101.1.0.550.335361854" identifies the same feature, includes information about the product specification and version ("101.1.0" for S-101 version 1.0) and other specifications can use identifiers in similar formats, e.g., "urn:iho:102.1.3.180.98767" for a high density bathymetry product specification version 1.3 feature with ID 98767 produced by BSH (producer code 180).

Using uniform resource identifiers as persistent global identifiers is also consistent with common practice in spatial data infrastructure, publishing, and web technology:

- The most recent draft INSPIRE guidelines recommend the use of URIs as "persistent, location-independent resource identifiers" for the benefit of the "larger community of users of spatial data and the developers of software components" [IN.D2.7]. URIs are also recommended for code list implementations [IN.D.2.5 §G.9].
- XML schemas use URIs/URNs as namespace identifiers.
- GML uses URI/URNs to identify CRSs.
- Academic research papers are identified by Digital Object Identifiers (DOIs) [ISO26324] which can be used with the "DOI handle" system to locate any paper given its DOI. For example, the identifier 10.1016/j.chb.2011.02.004 can be used with the DOI resolution implementation at http://dx.doi.org or they can be combined into a single URL http://dx.doi.org/10.1016/j.chb.2011.02.004 to reach the article by T. Ryan and S. Xenos titled "Who Uses Facebook..." published in Volume 27, Issue 5 of the Elsevier journal "Computers in Human Behaviour" (to see how it works, follow the links given in this paragraph or paste them into your browser's address box).
- Advanced Web technology, including W3C standards in the Linked Data initiative, e.g., RDF (Resource Description Framework) and OWL (Web Ontology Language) use URIs as identifiers.

Using URI/URN in S-100 data is a foundation for advanced applications using those technologies.

4.3 Proposal and implementation

INSPIRE defines a new base type *Identifier* [IN.D.2.5], defined as an external unique identifier. The type is a class with attributes *namespace*, *version*, and *localId* (local ID is an identifier unique within the namespace, meaning that no other feature has the same local ID). Identifiers are strongly recommended except where the underlying datasets do not maintain them or no requirement to identify or reference the object exists. This proposal uses similar criteria and allows product specifications more latitude with the specification and bindings of namespace and local identifier.

Proposal: Feature and information instances must have global persistent identifiers, except that identifiers need not be defined where the physical realities dictate otherwise or it is known that a reference to the object will not be needed, even from an external dataset conforming to another product specification. For example, identifiers need not be defined for cartographic objects.

It is not a requirement that complete persistent global identifiers be stored or transferred along with object data. Combining a persistent unique local identifier with a namespace normally suffices for a persistent global identifier. (The persistent unique local identifier need be unique only within the namespace, e.g., the Feature ID number (FIDN) in S-57 is unique only within a country.) The name space and rule for constructing a global unique identifier from local ID and namespace must obviously be provided in the product specification.

To reduce data storage, the namespace might be given in dataset metadata instead of every record. Obviously this depends on records in a single dataset being in the same namespace.

If datasets can mix and match objects from different namespaces then the namespace too must be an attribute of the data object, along with the localID. The namespace could be an optional attribute and storage efficiency tactics can be used, for example:

(1) Leverage the fact that exchange sets consist of different datasets - use a different dataset for each namespace, storing the namespace once in each dataset's metadata;

(2) Provide a value for the namespace attribute of a data object if and only if it is different from the namespace in the metadata;

(3) Subdivide datasets according to the namespaces of the object instances.

Global identifiers can also be used for documents and registry artefacts, including:

- feature concept dictionary entries for features and attributes;
- feature and portrayal catalogues (identify the XML file which is the catalogue and the documentation);
- entries for feature and information classes in feature catalogues (also for production tools).

Since local identifiers already exist for the above ("code" in the registry model), it is only necessary that namespaces be defined and the identifier construction rules described. Global identifiers for catalogues, etc., are potentially useful for production tools and advanced W3C technology (e.g., SKOS – Simple Knowledge Organization System).

4.4 Use cases

A. Nautical publications data

A national hydrographic office distributes S-101 exchange sets as well as nautical publications datasets for marine protected areas (MPA datasets). The S-101 data contain the usual ENC features, including *RestrictedArea* features. The MPA data contain geographic features and information objects. Some of the information objects are *Regulations* and *ContactDetails* instances that are associated with *RestrictedArea* and *SeaArea* features included in S-101 data. The *Regulations* objects encode national regulations applicable to all marine protected areas for the country and also specific rules pertaining to individual protected areas. The *ContactDetails* objects contain contact details for national and local environmental authorities administering the protected areas. Since the *RestrictedArea* features are already in the relevant S-101 datasets, the hydrographic office prefers to identify the areas to which each regulation applies by specifying the object identifiers in the MPA dataset, rather than duplicating the spatial primitives in both datasets.

B. Real-time data service for docking information

An oil terminal provides a 'docking data web service' which provides ships approaching berths with precise and up-to-date hydrographic information including quay dimensions surveyed to a precision higher than ENC requirements, obstructions, water depths alongside, and the availability, capability level, and operational status of support equipment like laser range finders. The web service supports all quays in the terminal, and uses URI identifiers to identify the equipment, sensors, and services at any specific berth. The ship's docking application uses the identifiers to obtain only information about the services and dimensions relevant to the docking plan. For example, up to date information about dimensions and depth alongside Berths B84-B87 in the Port of Long Beach might be obtained by means of the (hypothetical) URLs http://terminals.polb.org/101/1/0/335427431 and http://terminals.polb.org/101/1/0/335424838 which combine components of the identifiers for the relevant *ShorelineConstruction* and *Berths* data objects in the S-101 standard to make URIs which can be resolved by the hypothetical web service at the Internet host terminals.polb.org maintained by the port to provide real-time docking information.

4.5 New convenience data types

Multiple S-1xx product specifications are likely to use URI attributes with different meanings. The SNPWG data model defines the attribute *internetAddress* (Internet address / ADRNET) intended for web site addresses. For the convenience of product specification writers, consideration should be given to defining S-100 data types for URI, URN, and URL as derived data types of CharacterString, with W3C standards as normative references for their syntax.

Note that miscellaneous attributes in S-100 are already defined as being of type URL, e.g., Part 6, Table 6-25; Part 11, Table 11-2; and Part 11, Table 11-8. The proposal will formally introduce this type.

4.6 URI format identifiers and S-101

The S-101 product specification has the necessary information in metadata and feature records and needs at most a formal description of the construction of URIs from available information. Clause 4.4 of S-101 (Phase 4) requires unique universal feature identifiers. (This does not include information types, which means that persistent global identifiers cannot be constructed for information objects in S-101 data.) The ISO 8211 encoding in S-101 (Phase 4) defines the **Product identifier** and **Product edition** subfields of **Data set identification field** (**DSID**):

Product identifier	PRSP	"INT.IHO.S-101.1.0"	A()	Unique identifier for the data product as specified in the product specification
Product edition	PRED	"1.0"	A()	Edition of the product specification

Given values as in the Phase 4 draft together with the contents of **Feature Object Identifier (FOID),** it contains the information needed to construct a URI identifier for ENC objects, e.g., the identifier http://location.noaa.gov/s-101/1/0/335424838 can be constructed from the **Product identifier** sub-field in the S-101 **DSID** and the individual object's **AGEN** code (550) and FIDN (335424838), with some string operations and table lookups.

Incidentally, using http://s-101.iho.int as prefix instead of http://location.noaa.gov (and others) would be equivalent to the DOI handle system and DOI resolution protocol mentioned earlier. It would implicitly commit the IHO to maintaining a resolution service like the DOI resolution service at http://dx.doi.org.

5 Conclusion

Providing for persistent global identifiers in S-100 allows S-100 products and tools to standardize identifier de-referencing, define simple but flexible identification schemes, create globally unique identifiers, accommodate versioning, and make S-1xx artifacts and data products compatible with off-the-shelf technology.

6 Recommendations

1. Update S-100 to recommend persistent global identifiers in S-100 data products as proposed in Section 4.3.
2. Update S-100 to add convenience types for URI, URL, and URN.

The details of proposed changes to S-100 are in the accompanying redline markup documents.

7 Justification and Impacts

Justification: The proposed update to S-100 facilitates efficient object lifecycle and provenance management, information sharing, and data integration in and across multiple S-1xx data products. It aligns product specifications with current technology and standards and the INSPIRE initiative. Implementing it at the S-100 level facilitates harmonization of different product specifications. URIs, URLs, and URNs have been standardized by the IETF and W3C and are likely to be used in different domains. Defining types for them at the S-100 level will avoid the need to define them in multiple product specifications.

Impacts: Product specification writers will need to define namespaces, and include requirements and rules for local and/or global identifiers. Object identifiers will be required only of products and object types where an external identifier for a data object is considered necessary by specification developers or the GI registry control body.

If registry artefacts are identified by persistent URIs the control body must define and publish an appropriate URI scheme and the registry owner and manager ensure the URIs resolve to appropriate resources (e.g., XML catalogues, definitions). This responsibility may need to be mentioned in S-99.

8 Actions Requested of TSMAD

TSMAD is invited to:

- agree to this proposal for persistent global identifiers;
- agree to the proposal to use URIs as such persistent global identifiers;
- agree to the addition of the proposed data types to S-100;
- review and adopt the accompanying S-100 changes giving effect to these proposals.