



Tabular Data Submission Workgroup

In 2018, the Long-term Data Management (LTDM) Governance Board was formed to define, prioritize, and implement the Inventory & Monitoring Division's (IMD) plan to modernize Data Store. In its first year, the Governance Board developed a 10-year plan with three phases. Phase 1 focuses on defining, posting, and tracking data packages (datasets, metadata, digital object identifiers).

The Tabular Data Submission Workgroup was formed in September of 2018. The scope of the workgroup was to define initial policy and guidance for the submission of tabular data into the Data Store. The group's purpose was to establish the documentation and submission format of datasets so that the practice becomes standardized and supports long-term use. The following recommendations are the result of the workgroup's research and have been approved by the LTDM Governance Board for wider IMD consideration.

Data Package

Bundling essential components of a data submission into a package assures that the necessary information to assess and use the data always travels with the dataset. A complete data package consists of four components: machine-readable dataset files, metadata that describes the format and quality of the data files, a congruence report, and a package manifest (Figure 1). The congruence report is a file that records compliance with

and exceptions to metadata standards and policies and assesses agreement between the metadata and the data. The package manifest lists all files contained within the data package. Auto-generation of the congruence report, package manifest, and many components of the metadata would all be available under a full build-out of this concept, which should alleviate much of the burden to data producers (Figure 2).

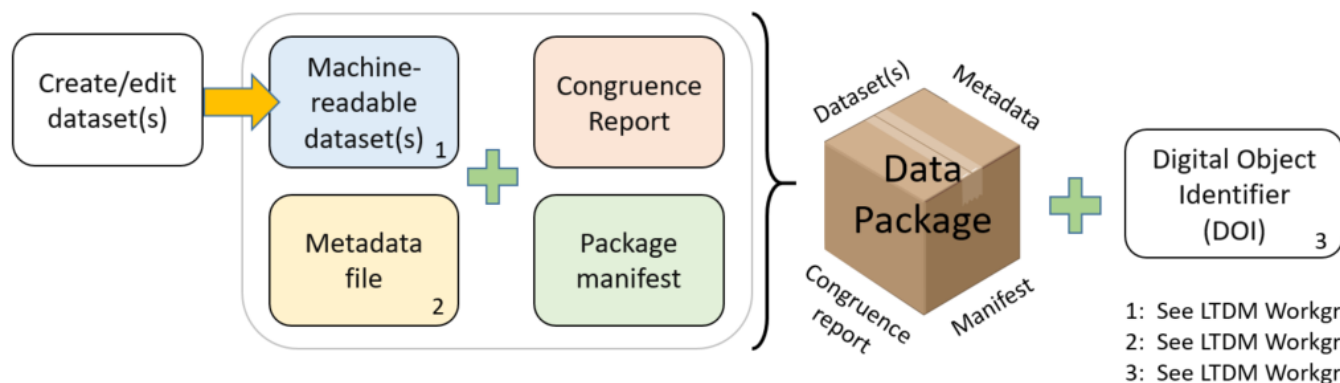


Figure 1. Data package components.

Acronyms

DOI = digital object identifier (see Terms for definition)
DQS = data quality standard
GLKN = Great Lakes Inventory & Monitoring Network
IMD = Inventory & Monitoring Division
IRMA = Integrated Resource Management Applications
LTDM = Long-term Data Management initiative
MBS = Managed Business Solutions
NCPN = Northern Colorado Plateau Inventory & Monitoring Network

NETN = Northeast Temperate Inventory & Monitoring Network
QAP = quality assurance plan
SFCN = South Florida/Caribbean Inventory & Monitoring Network
SOP = standard operating procedure
UMWN = Upper Midwest Inventory & Monitoring Network (fictional network used in an example)

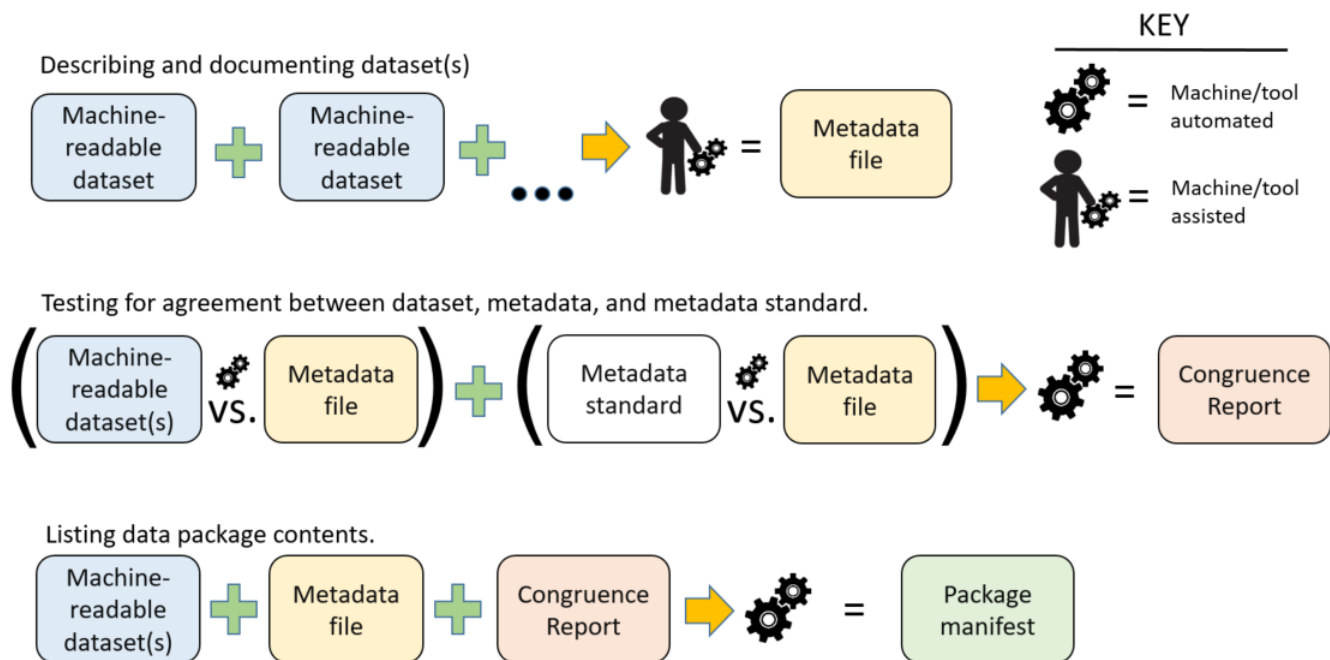


Figure 2. How data package components are created.

Data Track

A data track is a grouping of datasets associated with a monitoring effort or a project.

For example, each IMD monitoring effort would likely have two ongoing data tracks associated with it: (1) the complete data; and (2) the public data, which may have personally identifiable information (PII) and other protected records removed or otherwise treated (e.g., fuzzed or generalized) for distribution and extraneous information removed. If the complete data and the public data are identical, these data tracks may be one and the same.

There may be other terminal data tracks associated with an IMD monitoring effort, such as for a publication, report, or some other purpose, which have an addition of data and then are closed (Figure 3).

Within a data track, there should not be more than one active record for a data point (measurement in time and space). For example, the measurement of streamflow at 6:42pm at the Beaver Creek stream gage should consist of a single value. If there are two candidate values, ambiguity is introduced.

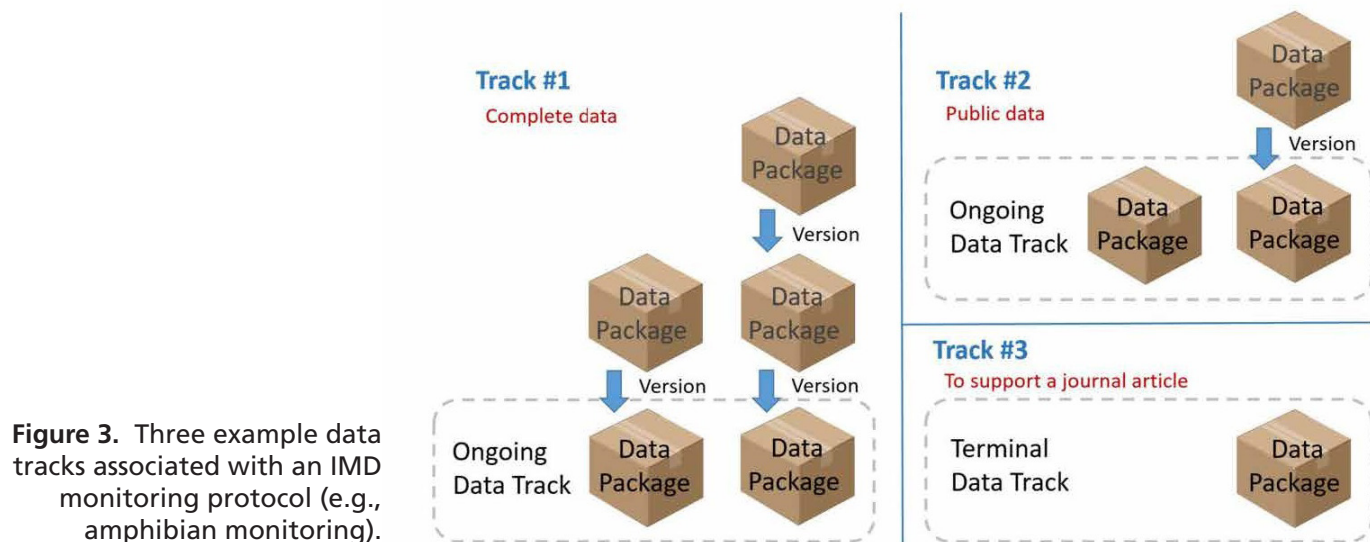


Figure 3. Three example data tracks associated with an IMD monitoring protocol (e.g., amphibian monitoring).

Data Submission Types

Submission of datasets to Data Store will occur as data packages associated with a data track. Addition of a new dataset or a version of an existing Dataset will be accepted as long as the data track does not contain record conflicts (Figure 4; record conflicts shown in green). Once uploaded to Data Store, data packages (and therefore datasets) can exist in only one state—active or inactive.

For example, the Upper Midwest Network (UMWN) uploads a data package containing amphibian monitoring data for 2016 and 2017 and associates it with a

data track for that protocol. The next year, UMWN uploads a data package containing only corrected data for 2017 as an addition to that same data track. Since both data packages need to remain active to provide access to the complete data track (2016 and 2017 records), this addition results in conflicting records for 2017 and is prohibited (see Figure 4, Row 5). The correct procedure in this case would be for UMWN to upload to this data track a data package including both 2016 and corrected 2017 records as a complete replacement of the first data package (see Figure 4, row 2).

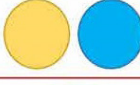
	Submission type	Existing Dataset	+	Submitted Dataset	=	Resulting Data Track	Action taken
Accepted	All new data	None	+		=		Add new Data Store reference and relate to Data Track
	Complete replacement		+		=		Version to existing Data Store reference
	Complete replacement and addition		+		=		Version to existing Data Store reference
	Addition only		+		=		Add new Data Store reference and relate to Data Track
Prohibited	Incomplete replacement		+		=		Prohibited submission type. Creates conflicting records within the same Data Track
	Incomplete replacement and addition		+		=		

Figure 4. Accepted and prohibited data submission types within a data track.

Terms

congruence report

Assessment of completeness/correctness of metadata vs. metadata standards and entity (dataset) and attribute (dataset columns) metadata vs. actual dataset entities and attributes (see Figures 1 and 2).

dataset

A collection of data. Every column of the dataset represents a particular variable, and each row corresponds to a given member of the dataset in question. The dataset lists values for each of the variables, such as height and weight of an object, for each member of the dataset. For the purposes of this discussion, we also imply that datasets are machine-readable and not in a proprietary (e.g., Microsoft Access relational database file) format.

data package

An assemblage of metadata, one or more machine-readable dataset files, a package manifest, and a congruence report (see Figures 1 and 2).

data track

A category/grouping for data packages associated with a monitoring effort or a project that relates the data packages to each other (see Figure 3).

digital object identifier (doi)

A unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to its location on the Internet.

package manifest

A component of a data package that documents the contents of the data package (see Figures 1 and 2).

Changes to Records

It is expected that changes that have been made to data records will be documented in the Dataset itself (i.e. change tracking) following the guidance provided by [Certification Guidelines for Inventory and Monitoring Data Products](#).

Changes in Data Structure

Metadata should clearly and completely document the data structure of the datasets in the data package. The specifics of how the data structure of a dataset has changed from a previous version does not need to be separately documented, as comparing sequential versions of a dataset's metadata will identify any change. In the future we anticipate there will be tools to analyze metadata and datasets and automatically report any differences in data structure.

Relationships Between Data Package Versions

Once uploaded to Data Store a data package cannot be changed, but can be replaced. Therefore, if any components of a data package need to be updated the data package must be replaced with a new version containing the updated components. Data Store's versioning tools will ensure relationships (i.e. provenance) are maintained between the original data package and any of its subsequent versions (Figure 4) and will take care of changing the status of the replaced data package to inactive.

Workgroup Members

Mark Hart, *GLKN Data Manager*

Simon Kingston, *Central Office Data Manager*

Al Kirschbaum, *GLKN Remote Sensing Specialist*

Dan Kocol, *MBS Project Manager*

Adam Kozlowski, *NETN Data Manager*

Judd Patterson, *SFCN Data Manager and Acting National Data Manager*

David Thoma, *NCPN Landscape Ecologist*

Next Steps

The work group's recommendation is to take an iterative approach to implementing changes to systems so that small sets of new features can be delivered and reviewed interactively with users, allowing adaptation to changing needs (e.g., findings of subsequent workgroups) and priorities.

Items expected to be implemented in the next two years:

- Package metadata file
 - Metadata standard issued (required and recommended fields)
 - Implementation of data packages will need to align with required fields as defined by metadata standards, and will need to adjust as requirements change.
 - Templates for data package metadata file
- Package manifest
 - Tools to facilitate the creation of a manifest given a set of data package contents
- Congruence report
 - Implement congruence checker that inspects submitted data for compliance and matches to Metadata submission
- Required updates to IRMA applications - Data Store/Protocol Tracking
 - Data tracks associated with protocols/monitoring efforts and other types of containers (e.g., projects)
 - Immutable data packages with active/inactive flags
 - Currently holdings can be replaced without versioning
 - Improve connectivity/relationship between Protocol Tracking and Data Store applications to avoid resubmission of protocol information and use a single source for management of relationship between protocol and associated dataset.
 - Require that a submitted IMD monitoring protocol data package is related to an IMD protocol's data track
 - Enhance both Data Store and Protocol Tracking to show relationships and information about the data package and protocol.
 - Enhance both Data Store and Protocol Tracking to support upload of a data package and process as appropriate (e.g., assign DOI, define specific reference type).
 - Enforce the use of Data Store versioning for IMD protocol data package versioning.
 - Report combined protocol information from Data Store and Protocol Tracking both internally and to the public.
 - Data Store assesses and reports on data package completeness (has all the required component files) during submission process.
- Shift organization from posting of databases to posting datasets

Suggestions and considerations for the next 2-5 years:

- Automate metadata creation based on file submission
- Ingest and vend data from datasets as service(s)

For more information, see our [InsideNPS page](#) and [SharePoint Site](#) with updates, documents, meeting notes, and presentations.