National Park Service
U.S. Department of the Interior

Natural Resource Stewardship and Science

Inventory and Monitoring Division Database Standards

September 11, 2015

Natural Resource Report NPS/NRSS/NRR—2015/1035

ON THE COVER
Photograph of worker bees constructing new honeycomb nest chambers and filling them with honey
Photograph provided by Washington State University website

Inventory and Monitoring Division Database Standards

September 11, 2015

Natural Resource Report NPS/NRSS/NRR—2015/1035

Brent Frakes, Simon Kingston, Margaret Beer

National Park Service
Inventory and Monitoring Division
1201 Oakridge Drive, Suite 150
Fort Collins, Colorado 80525

September 2015

U.S. Department of the Interior
National Park Service
Natural Resource Stewardship and Science
Fort Collins, Colorado

The National Park Service, Natural Resource Stewardship and Science office in Fort Collins, Colorado, publishes a range of reports that address natural resource topics. These reports are of interest and applicability to a broad audience in the National Park Service and others in natural resource management, including scientists, conservation and environmental constituencies, and the public.

The Natural Resource Report Series is used to disseminate comprehensive information and analysis about natural resources and related topics concerning lands managed by the National Park Service. The series supports the advancement of science, informed decision-making, and the achievement of the National Park Service mission. The series also provides a forum for presenting more lengthy results that may not be accepted by publications with page limitations.

All manuscripts in the series receive the appropriate level of peer review to ensure that the information is scientifically credible, technically accurate, appropriately written for the intended audience, and designed and published in a professional manner.

This report received informal peer review by subject-matter experts, including Inventory and Monitoring Division (IMD) network data managers, and IMD Central Office database managers, developers and architects.

Views, statements, findings, conclusions, recommendations, and data in this report do not necessarily reflect views and policies of the National Park Service, U.S. Department of the Interior. Mention of trade names or commercial products does not constitute endorsement or recommendation for use by the U.S. Government.

This report is available in digital format from the Natural Resource Publications Management website (http://www.nature.nps.gov/publications/nrpm/). To receive this report in a format optimized for screen readers, please email irma@nps.gov.

Please cite this publication as:

Frakes, B., S. Kingston, and M. Beer. 2015. Inventory and Monitoring Division database standards: September 11, 2015. Natural Resource Report NPS/NRSS/NRR—2015/1035. National Park Service, Fort Collins, Colorado.

NPS 909/129822, September 2015

Contents

Figures

Page

Tables

Page

Acknowledgments

Our sincere thanks to the many people who contributed ideas, advice, expertise, and time to this project. Data managers and others from Inventory and Monitoring networks across the country reviewed and commented on multiple drafts, including H. Bayley, J. Boetsch, K. Bonebrake, N. Bywater, B. Campbell, P. Capece, T. Colson, R. Frith, M. Hart, B. Johnson, S. Miller, J. Patterson, G. Rowell, T. Shepherd, D. Skidds, A. Snyder, K. Straub, and H. Thomas.

J. DeVivo, M. Tucker, and D. Kocol provided guidance and valuable advice, and database developers B. Seidle and M. DiSibio kept us on track. We thank L. Nelson for her geospatial and database expertise and both L. Poulin and M. Landgraf for their reality checks. Finally, we thank K. Gallo for her leadership of the Inventory and Monitoring Division.

Inventory and Monitoring Division Database Standards

Version 1.0

9/11/2015

Change Log

Date	Version	Change
9/11/2015	Version 1.0	Initial standards published

Purpose

The purpose of this document is to define and establish a common core of database standards for the National Park Service Inventory and Monitoring Division (NPS-IMD). These standards are meant to promote a new level of data integrity and consistency program-wide, and establish the foundation for greater data sharing and interoperability. These standards also include database documentation standards, which are essential for ensuring, over the long term, the proper use and interpretation of data.

These standards apply to all transactional master databases (Figure 1) developed or revised by the NPS-IMD networks or central office. They are strongly recommended for databases used for field data collection, wherever practical. They do not necessarily apply to databases optimized for reporting. The expectation is that networks will ensure that data collected in non-compliant databases are portable to compliant structures. Given the number of databases that currently do not conform, these standards apply to databases under development or undergoing significant revision. While these standards apply specifically to NPS-IMD, every attempt has been made to make them broadly applicable across the National Park Service. Adoption of these standards by other divisions will facilitate better data management, and will help establish long-term consistency among all NPS databases.

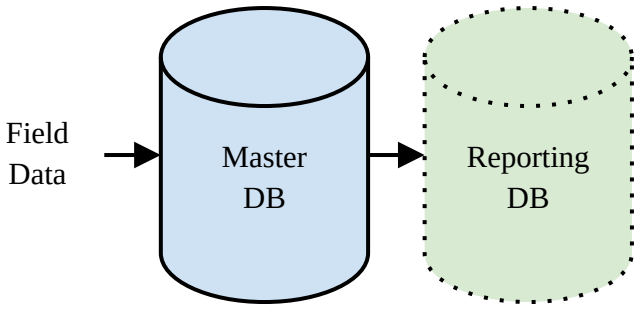

Figure 1. Conceptual model showing data flow from initial collection through reporting.

These standards are very specific and focus only on table design and structure. It is assumed networks will continue to use other good practices including proper planning, documentation and testing throughout the data management lifecycle.

Approach

For the first 15 years of the NPS-IMD, the Natural Resource Database Template (NRDT) was used as the model and standard for network database development. As uses and applications of I&M data evolve and database capabilities improve, the concept of a standard database model needs to be replaced by common database standards and best practices. These standards and practices allow flexibility in database modeling, yet will ensure the database integrity and consistency needed to meet increasing requirements for broad data sharing and dissemination.

Whenever possible and appropriate, these standards are goal-oriented. Focusing on the outcome of a standard (e.g., data integrity), as opposed to the process (e.g., following a specific template), is a more flexible strategy for producing a successful database design. Implementations are likely to differ depending on the data being modeled and the type of platform (i.e., MS Access or SQL-Server). To ensure the standards are usable and practical, we have attempted to keep them general, but with some notable exceptions. Furthermore, we have tried to strike the right balance of being thorough, yet not overly detailed.

In cases where third party systems need to be used (e.g., ESRI geodatabases, AQUARIUS, and EQuIS), we recognize that they will have their own set of standards which may conflict with ours (e.g., use of Globally Unique Identifiers [GUIDs[1]], field naming and definitions, primary keys, compound keys, referential integrity). Our standards should be applied if or where possible, and 3rd-party databases should be evaluated in light of these standards with careful considerations of the associated risks. This document can be a tool to help identify benefits and risks of 3rd-party databases.

In addition to the standards we have added a “Best Practices” section. This section addresses the differences between MS Access and SQL Server, which are the current accepted solutions for databases within NPS-IMD. Best Practices also includes recommendations that, while not to the level of a standard, are strongly encouraged.

Finally, it is worth mentioning that these standards come with the understanding that database design is as much art as science. There are never absolutes to database design and while exceptions are expected, they should be justified and documented. A database is a model of reality, which can be complex and messy. With any database design, the art comes into balancing three conflicting constraints of simplicity, accuracy, and completeness.

---

[1]One benefit of GUID primary key is that it can facilitate merging data from a field collection database into a master database. However, GUIDs can lead to large indexes and can often be difficult to deal with in application interfaces. Ultimately, your decision to use GUIDs must carefully balance these considerations.

Governance

These standards will be updated when necessary to reflect lessons learned, technological changes, and evolving programmatic goals and policies. These standards will be updated no more than quarterly and will be reviewed at least annually by the central office data management leadership in consultation with network data managers. All changes will be tracked in the "Change Log" section of this document.

Standards

Table 1. General Database Standards

ID	Topic	Standard	Justification and Notes
1.1	3rd Normal Form and Cardinality	Data are managed in a relational database that is in at least 3rd normal form. All tables have the cardinality to efficiently and correctly model the information collected for the protocol or project.	Data integrity
1.2	Referential Integrity	Referential integrity is enforced among all core and lookup tables. Rare exceptions include systems/version, staging, or temporary tables. It is the job of a database to enforce integrity, not the application.	Data integrity
1.3	Circular References	References among tables are not circular	Data integrity; an example would be the case where Table A refers to Table B, Table B refers to Table C, and Table C refers to Table A. This is distinct from self-joins, in which a column in a table references another column in the same table (e.g., an Employee table may have a SupervisorID column that points to the employee who is the supervisor of the current employee)
1.4	Data Stand Alone	The database is self-documenting and internally consistent. Validation, integrity and documentation are independent of the user interface (UI), workflows, and reports. If a constraint/validation can be created in the database, do it. If referential integrity can be created in the database that correctly models the objects/entities in your protocol, do it. If descriptions of database objects can be embedded in the database, do it.	Data longevity; proper application and interpretation.

Table 2. Table Organization and Naming

ID	Topic	Standard	Justification and Notes
2.1	Table Naming	Table names are meaningful, unambiguous, consistent in case, and compatible with MS Access and MS SQL Server	Clarity, consistency and compatibility; documentation. (See *Best Practices #4* for further suggestions.)
2.2	Table Description	Every table has a clear and meaningful description. When possible, the description relates back to the object or concept defined in the monitoring protocol or other project specifications.	Documentation and data integrity; not needed if table name is obvious and unambiguous
2.3	Grouping Tables	If it is necessary to group tables (e.g., grouping all lookup tables together), the mechanism for grouping is consistent	Consistency and interpretability (See *Best Practices #6* for further suggestions.)
2.4	Primary Keys	Every table has a single-column primary key to uniquely identify each row in the table. Exceptions include tables used for staging, importing, batching, and join tables. The data type for the primary key is specified in section 3.1. See also Standard 2.8 which discusses unique records.	Consistency and efficiency of identifying table rows and joining tables. If an auto-incremented ID is used, it is strongly recommended to add a unique index on the candidate key.
2.5	Reserved Words	Reserved words are avoided. See (http://msdn.microsoft.com/en-us/library/ms189822.aspx)	Reserved words can have conflicting meanings and functions depending on software used, and their use may create errors.
2.6	Joining Table Names	In cases where a table functions as a many-to-many join between two other tables, the name of the joining table will include the names of the respective joined tables. Order preference is first for the Data Table and then the Lookup/Enum Table. For example, a many-to-many join table between table Species (a lookup list of all species) and table Plot (a table containing all of the plot information) will be named PlotSpecies. Where both tables are lookups or both tables are data, the order of preference is left up to the designer.	Consistency and readability Unless there is an English word that appropriately describes the join. For example, a table that joins *Question* to *Answer* might be named *Solution* instead of *QuestionAnswer*. A table that joins *User* to *Unit* could be called *UnitPointOfContact* instead of *UserUnit*.
2.7	External Databases	Except for standardized lookup tables, data tables have no relational dependencies with tables external to the database instance or container.	Data integrity, data longevity
2.8	Unique Records	Tables have constraints or indexes on one or more data columns to enforce unique records.	Data integrity - prevention of duplicate records
2.9	Distinct Lookup Tables	Lookup values, (i.e., enumerated values to enforce domains) will only be managed in separate tables in the database. Lookup values have clear and unambiguous	Lookup tables serve as data dictionaries; ensures the database is self-

ID	Topic	Standard	Justification and Notes
		descriptions.	documenting
2.10	Lookup Tables Not Shared	Lookup, or enumerated, tables convey one purpose and are not shared by multiple tables unless their interpretation is entirely independent of all related foreign key tables. Lookup tables are designed with an understanding of the data types and how the data will be accessed.	Avoid shared master lookup tables where some values may not be valid for all tables.
2.11	Standardized Lookup Tables	Standardized lookup tables (Appendix A) should be used, as needed.	Standardized lookup tables are those that: (a) have values whose interpretation is entirely independent of the foreign tables; and, (b) are supported programmatically, thereby minimizing the risk of data loss.

Table 3. Column Organization and Naming

ID	Topic	Standard	Justification and Notes
3.1	Primary Key Name and Type	The first column of every table is the primary key with a column name of “ID.” Primary keys are auto-incrementing, and are the smallest possible integer data type to accommodate the expected number of records. For clarity and consistency, candidate or natural keys are not used as primary keys. Likewise, GUIDs and rich text (i.e., compound data rich keys) will not be used as primary keys. For 1:1 relationships, the pattern will be applied consistently. To enforce a 1:1 relationship, the foreign key column ‘constraint’ in the related table will be set to unique.	Consistency, cross-platform compatibility, and efficiency in joining tables.
3.2	Candidate Key	If a natural or candidate key is used, it is required and unique	Consistency, data integrity
3.3	Foreign Key Name	Foreign key columns are named [PrimaryTableName]ID	Consistency and clarity
3.4	Column Names	Column names are consistent in case, descriptive, unambiguous, and singular. Acronyms are not used. Where appropriate, column names relate directly back to the objects or concepts defined in the monitoring protocol.	Consistency and clarity; documentation
3.5	Identical Table and Column	Column names do not repeat the table name.	Minimize confusion as to which object is being referenced

ID	Topic	Standard	Justification and Notes
	Names		
3.6	Column Names Representing Unit of Measure	If a column represents data in a specific unit of measure, the unit of measure is a suffix to the column name (e.g., precipitation_in). Unit may be abbreviated except where the interpretation is ambiguous. Time is not treated as a unit except when explicitly tracking a unit of time (e.g., ElapsedTime_minutes)	Consistency, clarity and data integrity, documentation
3.7	Column Description	An unambiguous and meaningful description accompanies each column for every table.	Documentation, data longevity
3.8	Missing Values	The strategy to manage missing and blank values conforms to the quality control procedures as prescribed in the associated Quality Assurance Plan of the protocol or project.	Nulls can be ambiguous since they may indicate no observation, inability to read field sheet, a value not within the desired domain, or failure to transcribe/import data. Nulls need to be flagged, notated or corrected.
3.9	Geospatial Units and Datum	Geospatial units will be represented using either decimal degrees or UTM coordinates, or a combination of the two as long as both aren't used to represent a single point. If latitude and longitude are represented as decimal degrees, longitude should be negative in the Western Hemisphere. The Datum of NAD83 is represented with its own column and should include the realization (e.g. NAD83(CORS96), NAD83 (CORS2011)) when known. If UTM coordinates are used, Easting and Northing will be managed in separate columns. Columns for UTM zone and datum (i.e., NAD 83 and preferably including realization) must be included. If known, a column for the type of GPS unit used. Example: Garmin, ArcPad, Trimble, etc.	Consistency; storing locational data as standard coordinates (Lat/Long and UTM) NAD83 is a federal standard Datum realizations update occasionally
3.10	Lookup Table Columns	Lookup tables follow a common pattern for column names and data types.	Consistency
3.11	Constraints	Column domains (i.e,, the acceptable values for a column) and other constraints are enforced whenever possible to prevent erroneous values.	Data integrity
3.12	Default	Data columns do not have defaults.	Data integrity. Audit or system columns

ID	Topic	Standard	Justification and Notes
	Values		(e.g., CreatedDate) may have default values where appropriate, but data columns (e.g., WaterTemperature_C) do not have default values.
3.13	Zero Length Strings	Zero length strings are not allowed.	Data integrity
3.14	Data Types and Column Sizing	The smallest data type that accurately and efficiently represents the data is used.	Data integrity. Use date (SQL Server) if only a date is being captured. Use date/time if a date is being captured in Access or if time is included (both). Do not use a text data type for dates. If decimals are not needed for numeric values, use an integer data type of the appropriate size. Match the domain of the data you want to store with the data type that best fits, and then add a constraint to get down to the exact domain desired.

Table 4. Data Quality and Line of Sight[2]

ID	Topic	Standard	Justification and Notes
4.1	Protocol Traceability	Every data observation is unambiguously traceable to a specific version of a monitoring protocol, a quality assurance plan (QAP), and suite of standard operating procedures (SOPs). A Reference Code will point to the publicly accessible files in the IRMA Data Store.	Documentation; data longevity.
4.2	Protocol /Project Specific QA/QC Flags	Where appropriate for interpretation, data will be qualified with one or more flags or notations based on SOPs, protocols or QAPs.	
4.3	Certification	Every data observation has an associated QA/QC processing level (e.g., raw, provisional, certified) based on an associated quality assurance plan.	Documentation; ensure audience-appropriate release of data

[2] Guidance and examples of implementation for both data quality and line of sight are forthcoming.

ID	Topic	Standard	Justification and Notes
4.4	Sensitivity	There is an explicit strategy for flagging and managing sensitive records. Strategy includes data tables as well as any related queries or views, and allows clear identification[3] of all sensitive records.	Prevent inadvertent release of sensitive data
4.5	Tracking Changes to Data	There is a clear and consistent mechanism to track changes, including deletions, to data. This includes when, why, and who made the change.	Ensure that Quality Control procedures can be used to improve overall data integrity. Documentation. There will be further guidance on what types of changes must be tracked and what is exempted.

Best Practices

The following table represents the current set of best practices recommended by NPS-IMD. While these practices are encouraged, they are not currently enforced. Best practices should be followed when they holistically optimize the integrity, consistency, clarity and inter-operability of the database.

Table 5. Database Best Practices

BP#	Topic	Best Practice	Justification and Notes	Applies to
1	Default Columns for Lookup Table	Lookup tables follow the pattern: ID, Code, Label, Summary. All columns are required and all columns except Summary must be unique.	Consistency and clarity Flexible for multiple levels of readability and use	Both
2	Lookup Table - Column Specification for SQL-Server	Lookup tables have the following columns: ID TINYINT IDENTITY(1,1) NOT NULL PRIMARY KEY CLUSTERED, Code VARCHAR(n) NOT NULL CONSTRAINT [ConstraintName] UNIQUE, Label VARCHAR(20) NOT NULL CONSTRAINT [ConstraintName] UNIQUE, Summary VARCHAR(200) NOT NULL	Consistency	SQL-Server
3	Primary Key Type	All primary keys are defined as ID [TINYINT/SMALLINT/INT/BIGINT[4]] IDENTITY(1,1) NOT NULL PRIMARY KEY CLUSTERED	Efficiency and Consistency	SQL Server

[3] Separate standards for reporting will specifically address reporting of sensitive data.

[4] Always pick the smallest integer possible that meets the existing set of requirements

BP#	Topic	Best Practice	Justification and Notes	Applies to
4	Table Naming	Table names are singular, not plural. Ambiguous names (e.g., data, information) are avoided. Special characters, and spaces, are not used. Numbers are not used in a leading position. Do not shorten names if not necessary. Use full word if it adds clarity, especially for downstream users (e.g., *UnapprovedSpecies* instead of *SpTemp)*.	Clarity and documentation	Both
5	Pascal Case	Table and column names use PascalCase. Underscores are acceptable when they add clarity.	Clarity and consistency	Both
6	Grouping Tables/Table Name Prefixes	If it is necessary to group tables by function or type, schemas are used. Do not add prefixes (e.g., tbl, tlu, etc.) or suffixes to the table names.	Tables can be re-grouped without forcing the renaming of the table names. Using schemas is also easier to read and adds consistency.	SQL Server
7	Grouping Tables/Table Name Prefixes	Prefixes may be necessary in order to group lookup tables separately from data tables in Access, due to a lack of schemas or other mechanisms for grouping tables.	There is value in being able to separate lookup tables from data tables when reviewing a database.	MS Access
8	Column Order	Column order is: Primary Key Required Foreign Keys Nullable Foreign Keys All data columns Audit columns	Consistency and clarity	Both
9	Key Names	These recommendations apply to the names of the key objects in SQL Server and index objects in MS Access. They do not apply to the names of the columns involved in the keys. Primary key: PK_[NameOfTable] Foreign key: FK_[NameOfTable]_[NameOfForeignKeyColumn]	Example primary key names: PK_Location PK_ProgramPersonnel Example foreign key names: FK_ProgramPersonnelID FK_ProgramPersonnel_ProgramID	Both
10	Constraint Names	The following naming conventions are recommended for SQL Server constraints: Check constraint: CK_[TableName]_[ColumnName]_[RuleEnforced]	Example check constraint names: CK_PoolStage_Stage_meters_range	SQL Server

BP#	Topic	Best Practice	Justification and Notes	Applies to
		Default constraint: DF_[TableName]_[ColumnName]	CK_IncubatorRun_Comments_DisallowZeroLength Example default constraint name: DF_PoolStage_DateCreated	
11	Index Names	The following naming conventions are recommended for indexes: Primary key index: PK_[TableName] Unique index: UN_[TableName]_[OptionalColumnName(s)] NonUnique index: IX_[TableName]_[OptionalColumnName(s)]	Example unique index name: UN_WaterQuantityEvent UN_State_Code	Both

Appendix A. Standardized Lookup Table

The following sources are actively managed by NPS-IMD REST services.

Topic	Location(s)
NPS Units	http://irmaservices.nps.gov/v2/rest/unit/collections Note: that JSON is also available: http://irmaservices.nps.gov/v2/rest/unit/collections?format=json
NPSpecies	Contact NPSpecies Data Manager
Taxonomy	Contact Taxonomy Data Manager

Current Taxonomy Data Manager: Simon Kingston

Current NPSpecies and Unit Data Manager: Alison Loar

Recommended Reading

Davidson, Louis and Moss, Jessica (2012). Pro SQL Server 2012 Relational Database Design and Implementation. Apress Publishing, 784 pages.

Sheldon, Robert (2015). How to get Database Design Horribly Wrong. https://www.simple-talk.com/sql/database-administration/how-to-get-database-design-horribly-wrong/, August 2015.

The Department of the Interior protects and manages the nation's natural resources and cultural heritage; provides scientific and other information about those resources; and honors its special responsibilities to American Indians, Alaska Natives, and affiliated Island Communities.

NPS 909/129822, September 2015

National Park Service
U.S. Department of the Interior

Natural Resource Stewardship and Science
1201 Oakridge Drive, Suite 150
Fort Collins, CO 80525

www.nature.nps.gov