

Northwest Environmental Data Network (NED)

Draft: Recommendation to Improve Regional Consistency in Content and Use of Data Dictionaries

1.0 Introduction

This document is to introduce readers to the concept of the data-dictionary, identifies the importance of data dictionaries in the discovery and use of data, discusses the changing use of data dictionaries, provides a brief analysis of the divergence in data dictionary semantics and concludes with some recommendations for more consistent regional use of data dictionary semantics, definitions and formats.

2.0 Problem Statement

There is a regional need to share and understand (at least fishery, habitat and water) data across disparate databases. Not only is the data in many different databases, and of variable quality, but the underlying database documentation, including the data dictionaries, are developed using different terminologies, formats and contents. This creates additional problems for potential users of data, especially regional or landscape level data that must be ‘stitched’ together from multiple sources.

The development (and use) of a common set of data elements and formats for documenting database content and structures would help to make regional information systems more accessible. Tools may also be available to maintain a regional data dictionary repository as an organized on-line source, for example: table structures, collection protocols, data elements, and data element terms and definitions. This would reduce redundancy when new databases are created and improve understanding of the contents of existing databases.

3.0 Data dictionary – what is in it?

In its simplest form, the data dictionary is an organized collection of data element names and definitions, arranged in a table. A data dictionary may cover the whole organization, a part of the organization or a single database. More advanced data dictionaries contain database schema with reference keys, still more

advanced data dictionaries contain an entity-relationship model of the data elements. While the term "data element" is used here it is the same concept as "data object" or "object".

3.1 Data Element Definitions¹

- **Data Element Domain:** The context within which the element exists. For example information about a participant could include data element information about the participants address, phone number, title, and e-mail.
- **Data Element Number:** A unique number for the data element used in the technical documents.
- **Data Element Name:** Commonly agreed, unique data element name from the application domain. Also called an attribute.
- **Data Element Field name(s):** Field names are the names used for this element in computer programs and database schemas.
- **Data Element Definition:** Description of the meaning of the data element.
- **Data Element Unit of measure:** Scientific or other unit of measure that applies to the data
- **Data Element Precision:** The level to which the information will be reported (e.g. miles to 2 decimal places).
- **Data Element Data Type:** Data type (characters, numeric, etc.), size and, if needed, special representation.
- **Data Element Size:** The maximum field length as measured in characters and the number of decimal places that must be maintained in the database.
- **Data Element Field Constraints: Data Element is a required field (Y/N); Conditional field (c); or a null field:** Required fields (Y) must be populated. Conditional fields (C) must also be populated when another related field is populated (e.g. if a city name is required a Zip Code may also be required). "Not null" also describes fields that must contain data. "Null" means the data type is undefined (note: a null value is not the same as a blank or zero value).
- **Data Element Default Value:** A value that is predetermined -it may be fixed or a variable, like current date and time of the day.
- **Data Element Edit Mask:** An example of the actual data layout required (e.g. yyyy/mm/dd)
- **Data Element Business rules (Could include any of the material below)**

¹ Data dictionary definitions are modified from: Mattila, S. 2001. Tasks of the Database Administrator. University of Canberra. Published on www.

- **Data Element coding (allowed values) and intra-element validation details or reference to other documents:** Explanation of coding (code tables, etc.) and validation rules.
- **Related data elements:** List of closely related data element names when the relation is important.
- **Security classification of the data element:** Organization-specific security classification level or possible restrictions on use.
- **Database table references:** Reference to tables the element is used and the role of the element in each table. Special indication when the data element is a primary or secondary key for the table.
- **Definitions and references needed to understand the meaning of the data element:** Short application domain definitions and references to other documents needed to understand the meaning and use of the data element.
- **Source of the data in the data element:** Short description of where the data is coming from. Includes rules used in calculations producing the element value.
- **Validity dates for the data element definition:** Validity dates, start and possible end dates for when the element is or was used. There may be several time periods when the element has been used.
- **History references:** Date when the element was defined in present form, references to superseded data elements, etc.
- **External references:** References to books, other documents, laws, etc.
- **Version of the data element document:** Version number or other indicator. This may include formal version control or configuration management references.
- **Date of the data element document:** Written date of this version of the data element document.
- **Quality control references:** Organization-specific quality control endorsements, dates, etc.

3.2 Table Definitions

Table definitions may also be defined in a data dictionary. An example is provided below in TABLE I EXAMPLE OF A DATA DICTIONARY FOR A DATABASE TABLE.

TABLE I EXAMPLE OF A DATA DICTIONARY FOR A DATABASE TABLE				
Column Name	Optional	Format	Length	Description
ACRONYM	Y	VARCHAR2	10	Agency acronym
AGENCY	N	VARCHAR2	100	Full Name of Agency reporting release
AGENCYID	N	NUMBER	22	System Generated Sequence Number
DATAENTRYID	N	VARCHAR2	30	User ID of person who entered information
DATAENTRY_DATE	N	DATE	7	Data on which this information is added to the system
PARENT_AGENCY	Y	NUMBER	22	System Generated Sequence Number
TYPEID	Y	NUMBER	22	System Generated Unique Identifier
UPDATEID	Y	VARCHAR2	30	User ID of person who updated the information
UPDDATE	Y	DATE	7	Data on which the information update is done

Data dictionary information about database tables can include the following:

- **Table name**
- **Table owner or database name**
- **List of data element (column) names and details**
- **Key order for all the elements, which are possible keys**
- **Possible information on indexes**
- **Possible information on table organization**
 Technical table organization, like hash, heap, B+ -tree, AVL -tree, ISAM, etc. may be in the table definition.
- **Duplicate rows allowed or not allowed**
- **Possible detailed data element list with complete data element definitions**
- **Possible data on the current contents of the table**
 The size of the table and similar site-specific information may be kept with the table definition.
- **Security classification of the table**
 Security classification of the table is usually same or higher than its elements. However, there may be views accessing parts of the table with lower security.

3.3 Database Schema

The database schema is usually a graphical presentation of the whole database. Tables are connected with external keys and key columns. When accessing data from several tables, database schema will be needed in order to find joining data elements and in complex cases to find proper intermediate tables. Some database products use the schema to join the tables automatically, for example XML exchange formats.

3.4 Entity-relationship Model of Data

The entity-relationship model is database analysis and design tool. It lists real-life application entities, attributes of entities and relationships amongst entities. The type of each relationship is also indicated. Entity-relationship model is represented in graphical form.

3.5 Database Security Model

Database security model associates users, groups of users or applications (programs) with database access rights.

4.0 Using Data Dictionaries

There are many different ways to create and use data dictionaries. Different data elements and different definitions for those data elements abound. For example, the **TABLE II COMPARISON OF DATA-DICTIONARY TERMINOLOGY** (below) is a comparison of how different and often confusing terminology is used in data dictionaries. The sample includes data dictionary elements for some Pacific Northwest and other (available) on-line data dictionaries. The purpose of the comparison is to begin to identify a common set of semantics for consistent data dictionary terminology. The top row of the table, together with the definitions, is a first level-of-effort to define a core set of elements and definitions.

Traditionally data dictionaries have been of primary interest to database developers, where interest has focused on a subset of the core data elements: usually the data element number, the data field name, the type of data, the size, edit masks and constraints.

However, there is now more interest from data providers and users (e.g. collectors and analysts) in directly inputting data to and accessing data from databases. Along with this there is an increased need for database transparency. What is the data structure? What are the table structures? Where do I see my data? What does my data mean? What does their data mean?

While the sub-set above is sufficient for creating data storage, it is insufficient for providers of data who need to know more about how the data is defined or data analysts who need to know more about its meaning. These user group efforts are improved if the data dictionary contains additional information and (of course) if it is available. Brackett ² refers to these broadly different types of dictionary information as follows:

“technical data resource data...what is meant to build manage and maintain databases. They include things like physical data names and structures, data types and formats and etc” and, “semantic data resource data...that help people understand the data resource and use that resource to support business activities...they include things like primary data names, data definitions, logical data structure, and so on” (my underlining).

A data dictionary acts as a single repository for the various data that will be stored in a database or databases. It provides for efficiencies in database development, for example, the developers of data entry devices and the database can create code simultaneously, more efficient data structures can be developed and synonymous and potentially duplicative data and data-tables can be identified.

Others have also recognized the importance of conventions and standards for data element language. ISO 11179 is a global standard³ that provides guidelines for standardizing and registering data elements. It consists of:

Specification and Standardization of Data Elements

Classification of Data Elements

Basic Attributes of Data Elements

Rules for Data Definitions Naming and Identification of Data Elements

Registering and Storing Data Elements

² Brackett, M.H. 2000. Data Resource Quality. Addison –Wesley Information Technology Series.

³ <http://www.iso.ch/iso/en/ISOOnline.frontpage>

Some existing efforts, for example the Grants.gov Data Modeling and Analysis Effort and the ESRI GIS Portal Toolkit are compliant with (at least) parts of ISO 11179.

In summary, data discovery and sharing is improved when users have access to both the technical and semantic data necessary to understand the underlying information system definitions and assumptions. Data dictionaries provide a window to the contents of databases that help begin the process of identifying the degree of similarity across databases. When data is both understandable and highly comparable there is more potential for data integration. Consistent use of common data dictionary elements would help to increase data transparency for data developers and users. ISO 11179 (and other comparable systems) should be reviewed in detail to identify its potential for use by NED and others in the NW Region.

TABLE II COMPARISON OF DATA-DICTIONARY TERMINOLOGY													
Data Dictionary Name	Data Element Domain Name (Object Class)	Data Element Number (for reference in data model)	Data Element Name (Attribute)	Data Element Field Name	Data Element Definition	Data Element Unit of Measure (uom)	Data Element Precision	Data Element Data Type and Size (decimals)	Data Element Size (Max Width)	Data Element Field Constraints Required Field, Y/N/Conditional, Null	Data Element Default Value	Data Element Edit mask (e.g. of actual layout)	Data Element Business Rules
PCSRF			Description		Definition	Units of measure	Included with uom		max	Active, required			Optional comments
John Day Data Dictionary ⁴	Domain (General Characteristic)		Attribute		Description	Units	Precision	Var.txt, date, Floating point, Date, Time, etc					Comment
PNWODE			Data Element		Description	Data Type (format)				Required Y/N/conditional			Business rules
SteamNet ⁵ (exchange formats)				Field Name	Field Description			Type -Integ., Date, Char.)	Max. Width	Req (Y/N)			Codes conventions
Software Project Management for Dummies		Number		Data element				Type	Size	Edits/Validations (Null ⁶ , Not Null, Optional)		Edit mask	
Australian National Health Data Dictionary	Domain		Data Element Type		definition	Unit of Measure		Data Type	Size				Context
University of South Carolina (Inst. Planning & Assessment)			Descriptor	Variable Name	Definition			Field Attribute				Values	source
USEPA (CERCLIS) USC			Common Name	System Name	Definition			Data Type, Length		Required			Values (codes)
Grants.gov XML schema		Number	Name		Description			Number	Min & Max	Required Y/N, duplicates Y/N			comments

⁴ Spatial Dynamics, 3/29/2004

⁵ StreamNet Exchange Format Documentation Version - 98.2 July, 1998

5.0 Recommendation for NED

Review potential use of ISO/IEC 11179 guidance for potential use by NED

Work with regional partners (e.g. PNW-RGIC, PNAMP and Agencies) to improve, develop and promote more consistent use of regional data dictionary element language and core data dictionary content.

TABLE III POSSIBLE NED DATA DICTIONARY DATA ELEMENT TERMS AND DEFINITIONS (below) is proposed as a starting point for consideration and comparison to ISO/IEC 11179, or other standards that can be identified.

Next steps:

Evaluate this draft against conventions and standards – in particular ISO/IEC 11179 Conventions and Standards – a global standard providing guidelines for standardizing and registering data elements.

Determine whether the ISO/IEC 11179 framework provides a workable solution and modify this document accordingly.

Work to adopt data dictionary element terms and definitions for use by NED and promotion to other NED partners.

TABLE III. POSSIBLE NED DATA DICTIONARY DATA ELEMENT TERMS AND DEFINITIONS.	
<p>The data dictionary, at its simplest, is an organized collection of the data elements and the details of these data elements associated with defined topical domains. Where possible the name of the data dictionary should be unique and it should always include the version number and the date of the version.</p> <p>For example: the John Day Data Dictionary identifies in detail, the discrete data elements that will be used for the collection and subsequent management of data using certain data collection and data sampling protocols.</p>	
DATA ELEMENT TERMS	DATA ELEMENT DEFINITIONS
Data Element Domain Name	A data content topic, for example, a named data collection protocol – EMAP. Note there may be multiple domains or sub domains within a particular data dictionary.
Data Element Number (for reference in data model)	Data element number is used in the technical documents.
Data Element Name	Commonly agreed, unique data element name. Also called attribute. Note: there are likely to be multiple data element names for a particular domain.
Data Element Field Name	The name used for this element in computer programs and database schemas. It may be an

	abbreviation of the Data Element Name (eg. Cellular Phone Number might be assigned a field name of Cell_Ph_No).
Data Element Definition	Description of the element in the application domain.
Data Element Unit of Measure (uom)	Scientific or other unit of measure that applies to the data.
Data Element Precision	The level to which the information will be reported (e.g. Miles to 2 decimal places).
Data Element Data Type	The type of data (e.g. Characters, Numeric, Alpha-numeric, date, list, floating point)
Data Element Size and Decimalization	The maximum field length that will be accepted by the database together with any decimal points (e.g. 30(2) refers to a field length of 30 with 2 decimal points).
Field Constraints: Data Element is a required field (Y/N); Conditional Field (C); or a “null” field	Required fields (Y) must be populated. Conditional fields (C) must be populated when another related field is populated (e.g. if a city name is required a zip code may also be required). “Not null” also describes fields that must contain data. “Null” means the data type is undefined (note: a null value is not the same as a blank or zero value).
Default Value	A value that is predetermined. It may be fixed or a variable, like current date and time of the day.
Edit Mask (e.g. of actual layout)	An example of the actual data layout required, (e.g. yyyy/mm/dd).
Data Business Rules	There are often the rules that define how data would be managed within an information system (e.g. data could be coded (1=adult, 2=parr, 3=juveniles) and these codes would be included in the data dictionary for use by developers and users. Other business rules could define how the database would be operated for example how rights to create, read, update or delete records are assigned if they are needed.

Additional References:

For various examples of current data dictionaries readily available via the www see the following:

National Health Service data dictionary (UK).
http://www.nhsia.nhs.uk/datastandards/pages/dd_m.asp

Survey of Labour and Income Dynamics Electronic Data Dictionary (SLID) (Canada).
<http://www.statcan.ca/english/SLID/diction.htm>

EPA (CERCLA) Data Dictionary. <http://www.epa.gov/superfund/sites/ded/>

For a straightforward example of the importance of the data dictionary to data discovery and sharing look at the following site:

http://www.grants.gov/assets/Grants.gov_XMLSchema-Implementation-Guide.doc

Add others here PNWQDE?