NED

NORTHWEST ENVIRONMENTAL DATA-NETWORK

Best Practices for Data Dictionary Definitions and Usage

Version 1.1 2006-11-14

1.0 Introduction

This document introduces readers to the importance of a data dictionary as a critical part of a metadata record or repository, describes how data dictionaries help with the discovery and use of data, provides a brief comparison of differences in usage in published data dictionaries and concludes with Recommendations of Data Dictionary Data Element Terms and Definitions. Data dictionaries come in many different forms and sizes. The document has been reviewed and approved by the Northwest Environmental Data Network (NED) Steering Committee. Stewart Toshach is the author.

A data dictionary is a part of the metadata that is used to understand the data and the databases that contain it. The data dictionary identifies data elements and their attributes including names, definitions and units of measure and other information. Often they are organized as a table. The focus here is on the need to adopt and support more consistent use of data dictionary elements and terminology as a part of improving metadata.

While dictionaries have long been used to communicate the meaning of written and oral language their use to communicate meaning of numerical language is less widespread, despite the similarity of the problem.

> *"When I took the first survey of my undertaking, I found our speech copious without order, and energetick without rules: wherever I turned my view, there was perplexity to be disentangled, and confusion to be regulated; choice was to be made out of boundless variety, without any established principle of selection; adulterations were to be detected, without a settled test of purity; and modes of expression to be rejected or received, without the suffrages of any writers of classical reputation of acknowledged authority."*[1] Preface to Dictionary of the English Language, Samuel Johnson, 1755

2.0 Importance of Data dictionaries

There is a regional need to share and understand (at least) fishery, habitat and water data across geographic boundaries and in disparate databases. Not only is existing regional data held in many different databases, and of variable quality, but the underlying database documentation, including the data dictionaries, and metadata 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.

[1] Preface to Dictionary of the English Language, Samuel Johnson, 1755

The development (and use) of a consistent set of data elements and formats for documenting database content and structures would help to make regional information systems more accessible. Metadata Repository 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. The Bureau of Reclamation has developed a prototype "Protocol Builder" that, pending testing and web enablement, may also function as a regional data dictionary. The use of common data dictionary terminology and the availability of a regional data dictionary would reduce redundancy when new databases are created and improve understanding of the contents of existing databases.

According to the International Standards Organization (ISO)[2] *"The increased use of data processing and electronic data interchange heavily relies on accurate, reliable, controllable, and verifiable data recorded in databases. One of the prerequisites for a correct and proper use and interpretation of data is that both users and owners of data have a common understanding of the meaning and descriptive characteristics (e.g., representation) of that data. To guarantee this shared view, a number of basic attributes has to be defined*."

The purpose of this paper is to accelerate the development of shared definitions for and usage of data dictionary elements, as a part of metadata in order to improve regional data management.

[2] International Standards Organization, 2004. Information Technology Parts 1-6 (2nd Edition) http://www.iso.org/

3.0 Data dictionary – what can be in it?

A simple data dictionary is an organized collection of data element names and definitions, arranged in a table. It may describe all the data holdings of an organization, a part of the holdings or a single database. More advanced data dictionaries can contain database schema with reference keys and entity relationship diagrams[3]. The following sections, 3.1 through 3.5, provide a description of the possible contents of data dictionaries.

3.1 Descriptions of Data Elements

- **Data Element Domain:** The context within which the data element exists. For example information about a participant (the domain) 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 technical documents.
- **Data Element Name:** Commonly agreed, unique data element name.
- **Data Element Field Name(s):** Names used for this data 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 value.
- **Data Element Value:** The reported value.
- **Data Element Precision:** The level to which the data element value will be reported (e.g. miles to 2 decimal places).
- **Data Element Data Type:** Data type (characters, numeric, etc.), size and, if needed, any special representation that applies to the data element.
  - **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).

[3] Data dictionary lists are modified from: Mattila, S. 2001. Tasks of the Database Administrator. University of Canberra. Published on www.

    - **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):**
    - **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 relationship is important.
    - **Security classification of the data element:** Organization-specific security classification level or possible restrictions on use.
    - **Database table references:** Reference to tables where the element is used and the role of the element in each table. 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 data element value.
    - **Validity dates for the data element definition:** Validity dates, start and possible end dates for when the data element is or was used. There may be several time periods when the data element has been used.
    - **History references:** Date when the data 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. Tables typically include a collection of data elements to demonstrate how the data are arranged and related to other data elements. 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**

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 are 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 a database analysis and design tool. It lists data elements, attributes of the elements and relationships amongst the elements and tables shown in graphical form.

3.5 Database Security Model

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

4.0 Using Data Dictionaries

Many different approaches are used to create and use data dictionaries, and this is a part of the problem. Different use, semantics and definitions of data dictionary terminology abound. TABLE II, shows **DIVERGENCE IN TERMINOLOGY USED TO DESCRIBE DATA-DICTIONARY ELEMENTS** (below), shows how different and often confusing terms and definitions are used in data dictionaries. The sample includes data dictionary elements for some Pacific Northwest and a few other data dictionaries available on-line. Many data bases have been developed without formal data dictionaries at all, so an important challenge is not just how to identify and describe data elements but to increase the use of documentation.

Table III identifies **RECOMMENDED DATA DICTIONARY DATA ELEMENT TERMS AND DEFINITIONS**. The focus on data elements and their attributes recognizes their critical importance in information technology development and data sharing. The use of defined elements provides a first step to understanding data. The terms are consistent with ISO and Federal Geographic Data Committee (FGDC) metadata reporting guidance or requirements. The FGDC are currently working with the ISO to ensure consistency across FGDC and ISO standards.

Traditionally, data dictionaries have been of primary interest to database developers, with the interest 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.

Now there is 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 interest 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? Where is the data located? When was the data last updated?

While the sub-set of data dictionary components referred to above is sufficient for creating simple 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 the meaning of the data. These user-group efforts are improved if the data dictionary contains additional information and (of course) if it is available. Brackett [4] who has written extensively on data quality 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 describes the various data that will be stored in a database or databases. It provides for efficiencies in database development, for example, allowing developers of data entry devices and the

[4] Brackett, M.H. 2000. Data Resource Quality. Addison –Wesley Information Technology Series.

database to simultaneously create code. More efficient data structures can be developed while synonymous and potentially duplicative data or data-tables can be identified and understood.

There is international recognition of the importance of conventions and standards for data element language. ISO 11179 is a global standard[5] (where the USA is a partner) to provide 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

Existing NED efforts, for example, the ESRI GIS Portal Toolkit, are compliant with 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.

[5] http://www.iso.ch/iso/en/ISOOnline.frontpage

TABLE II - DIVERGENCE IN TERMINOLOGY USED TO DESCRIBE DATA-DICTIONARY ELEMENTS

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. actual layout)	Data Element Business Rules
PCSRF			Description		Definition	UOM	Included with UOM		Max	Active, required			Optional comments
John Day Data Dictionary[6]	Domain (General Characteristic)		Attribute		Description	Units	Precision	Var.txt, date, Floating point,Date,Time, etc					Comment
PNWQDE			Data Element		Description	Data Type (format)				Required Y/N/conditional			Business rules
StreamNet[7] (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[8], Not Null, Optional)		Edit mask	
Australian National Health Data Dictionary	Domain		Data Element Type		Definition	UOM		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

[6] Spatial Dynamics, 3/29/2004

[7] StreamNet Exchange Format Documentation Version - 98.2 July, 1998

5.0 Recommendation for NED

1.0 Adopt, for NED use, a short list of ISO 11179 compliant data dictionary data element terms and definitions as a part of needed Metadata. See Table III below for a recommended list.

2.0 Work with regional partners (e.g. PNW-RGIC, PNAMP, Agencies and other entities) to promote the consistent use of these Data Dictionary Element Terms and Definitions.

TABLE III. Recommended Data Dictionary Data Element Terms and Definitions.

The data dictionary, at its simplest, is an organized collection of the data elements and the details of these data elements as associated with defined topical domains. Where possible the name of the data dictionary should be unique and it should always include a version number and the date of the version.

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)	A number associated with the data element name for use in technical documents.
Data Element Name	Commonly agreed, unique data element name. Note: there are likely to be multiple data element names for a particular domain.
Data Element Field Name	The name used for this data element in computer programs and database schemas. It is often an abbreviation of the Date Element Name (eg. Cellular Phone Number might be assigned a field name of Cell_Ph_No).
Data Element Definition	Description of the meaning of the data element
Data Element Unit of Measure (uom)	Scientific or other unit of measure that applies to the data element.
Data Element Precision	The level to which the data will be reported, eg 1 mile plus or minus .001 mile.
Data Element Value	The reported data.
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. Fish data could be coded (1=adult, 2=parr, 3=juveniles) and these codes would then be included in the data dictionary for use by developers and users. Other business rules, for example how rights to create, read, update or delete records are assigned if they are needed.