The Importance of Establishing Common Methods and Terminologies in Data Mappings

Robert Cox
PEO STRI, PM ITTS/IMO
rob.m.cox@us.army.mil

Paul Dumanoir
Joint Training Integration & Evaluation Center
paul.dumanoir@us.army.mil

Louis Hembree
Naval Research Laboratory – Monterey
louis.hembree@nrlmry.navy.mil

Farid Mamaghani
SEDRIS
farid@sedris.org

Kevin Trott
Northrop Grumman Information Systems
kevin.trott@ngc.com

Michele Worley
SAIC
michele.l.worley@saic.com

Sponsored by Modeling and Simulation Coordination Office (M&S CO) under High level Task (HLT) S-C-3, Development and Maintenance of M&S Standards

Outline

- Introduction
- Data abstraction
- Why mapping between dictionaries is important
- Mapping approach
- Terminology for mapping
- Terminology for the analysis phase
- Mapping terminology for concept dictionaries
- Mapping terminology for data models
- Summary

Introduction

- A solid understanding of the semantics, organization, and constructs of data is critical to successful data interoperability.
- Conversion / translation software are built on such an understanding.
- A robust mapping between the data elements used in different systems is essential to achieving data interoperability.
- This mapping task is more complex and more critical in networked M&S applications, where diverse systems are involved.
  - Many models / simulations integrate data from a variety of sources.
  - Mapping the data from multiple sources to the internal representation of a system is a key part of the data integration process.
  - Data communication requirements demand a common data mapping.
- Establishing a common and consistent mapping terminology is critical to the mapping process.
- Lessons learned from applying mapping methodologies and terminologies to environmental data are presented. But the same issues and principles apply to non-environmental data mapping efforts.

Introduction (cont.)

- Classifying objects into categories is fundamental to human reasoning and communication.
  - The formal study of this topic dates back centuries, and plays a central role in philosophy, language, logic, mathematics, and many other areas.
- Classifying objects into categories usually depends on how the uses, functions, characteristics, and/or applications of those objects are viewed.
- This context-specific nature of categorization makes it difficult, if not impossible, to apply a single categorization for all purposes.
- How objects are organized within a given context can be completely different from how the same objects are thought of in a different context.
- Categorization is also critical to communication and interoperability between information system, and plays a key role in creating mappings that allow automated data translations between systems.
- Development and use of such mappings may apply at various levels of data abstraction.
- The data abstractions range from dictionaries, to data/information models, to physical data products, and any number of derivatives in-between these.

Data abstraction

- Several broad categories of data semantics and specifications exist.
  - **Each of these can be thought of as a model of the data at some level of abstraction.**
  - **The DoD Architecture Framework (DoDAF) version 2.0 defines similar artifacts.**
- ***Dictionaries*** – a collection of terms and their definitions used within a particular context.
  - DoDAF requires an Integrated Dictionary (AV-2) that defines the terms used in the architecture to ensure semantic understanding across the enterprise.
- ***Logical Data Models*** – defines the various kinds of classes (also known as concepts, items, objects, or entities) of interest within a domain, the attributes that describe those classes, and the relationships among those classes.
  - DoDAF requires a Logical Data Model (DIV-2) to document system data requirements and structural business process rules.
- ***Physical Data Models*** – adds the details of how information about each kind of object is to be stored, transmitted, and manipulated by hardware and software, including the data types and how each relationship and operation is to be implemented.
  - DoDAF requires a Physical Data Model (DIV-3) to specify how a Logical Data Model is to be implemented in terms of message formats, file structures, and physical database schemas.

Data abstraction (cont.)

- Dictionaries play a fundamental role in the development of specifications and the production of content for information systems.
  - For standards developed by the International Organization for Standardization (ISO) the Shorter Oxford English Dictionary (SOED) is the default dictionary for all terms that are used, but not explicitly defined.
- The terms defined in a dictionary denote concepts, including:
  - ***Objects*** – Also known as classes, entities, things, and, in the geospatial community, features.
    - These terms refer to objects of interest within the domain addressed by the dictionary.
    - Generally nouns, or noun phrases, used as subjects and/or direct or indirect objects.
  - ***Attributes*** – Also known as properties, characteristics, etc.
    - These terms are used in describing objects, either qualitatively or quantitatively.
    - They, or their values, are used as adjectives, or in other forms of descriptive phrases.
- **A dictionary is not a data model**.
  - However, its definitions may implicitly specify basic relationship information from which a data model can be developed.
  - Concept dictionaries can be used as building blocks in the development of data models for specific applications or data products.

Data abstraction (cont.)

- Definitions in a dictionary usually follow a pattern: they relate a concept to a more general concept, and then specify what differentiates this concept from all others within that more general concept.
  - ***"A barn is an agricultural building designed to house animals and related equipment"***
  - "Building" is the more general concept.
  - "Agricultural" and "designed to house animals and related equipment" specify how a "barn" is different from other buildings, with respect to its form, function, and use.
- Once a data model has been developed, the collection of terms (classes or entities, and associated attributes) used in that data model is called a ***data dictionary***.
  - The term "data dictionary" is **often incorrectly used** to refer to concept dictionaries, catalogs, feature/attribute lists, and other dictionaries of terms.
- In ***concept dictionaries***, attributes are defined generically, independent of how they may be used to describe specific objects.
  - Concepts such as length or color can be used to describe many different types of objects.
  - Defining attributes generically facilitates their consistent use with different objects.
- In contrast, ***data models*** define specific pairings between objects & attributes.
  - Data models provide additional constraints to **meet specific application or product needs**.
  - **Data dictionaries contain only those terms required for the associated data model**.

Why mapping between dictionaries is important

- Examples from 3 different dictionaries, that share a similar lineage but have evolved differently over time, will be used to illustrate this.
- The Defence Geospatial Information Working Group (DGIWG) ***Feature and Attribute Coding Catalogue (FACC)***, the basis for traditional NGA products (note the use of the term "catalogue" in its title).
  - Earlier versions related attributes to features, but in the final version (Edition 2.1, Sep 2000, and subsequent baseline maintenance releases) these relationships were dropped.
- The ***Environmental Data Coding Specification (EDCS)***, ISO/IEC 18025, one of the SEDRIS technology components.
  - Lessons learned from FACC were instrumental in the development of EDCS; however, the scope and level of detail of EDCS was broader than FACC.
  - EDCS is composed of 9 related concept dictionaries, and introduced refinements in the logical decomposition of concept definitions such as: **separation of units of measure and scale from attributes** and placing their definitions in their own dictionaries; **explicitly relating a given concept to other concepts** in EDCS; **providing citations and references** for concept definitions; and providing content **extensibility through an online registry**.
- The ***DGIWG Feature Data Dictionary (DFDD)*** is the successor to FACC.
  - DFDD is derived from both FACC and EDCS, incorporating concepts such as **separating units of measure from definitions**, as well as utilizing an **online registry**.
  - DFDD does not explicitly relate concept definitions.

Why mapping between dictionaries is important: Examples

- **FACC 2.1 – Building** (AL015):

  A relatively permanent structure, roofed and usually walled and designed for some particular use.

  - FACC did not define specific kinds of buildings
  - Instead, it provided an attribute called Building Function Category (BFC): *Type or purpose of the building*, with a list of coded values. One of these values was BFC 125, Barn/Machinery Shed.

- **EDCS – BUILDING**:

  A fixed, relatively permanent <STRUCTURE> with a <ROOF> and usually with <WALL>(s) that is designed for use and occupancy by <HUMAN>s; a building.

  - EDCS also defines a BARN: *A <FARM_BUILDING> that is used to store hay, grain, and implements and/or to house <NON_HUMAN_ANIMAL>s; a barn [SOED, "barn", A.1] [SOED, "barn", A.2].*
  - FARM_BUILDING, used in the definition of BARN, is defined as: *A <BUILDING> located on a <FARM>.*
  - Similar to FACC, EDCS includes a BUILDING_FUNCTION attribute, defined as *The function of a <BUILDING>*, which includes a value BARN.

Why mapping between dictionaries is important: Examples (cont.)

- **DFDD – Building** (AL013):

  A free-standing self-supporting construction that is roofed, usually walled, and is intended for human occupancy (for example: a place of work or recreation) and/or habitation.

  - DFDD also defines the feature concept Barn (AJ085) : *A roofed farm building designed for sheltering harvested crops (for example: hay), livestock (for example: cattle), and/or farm machinery (for example: tractors and plows).*
  - DFDD does not include a general "building function" attribute;
  - Instead, it provides a collection of more specific "Facility Type" attributes, including Agricultural Facility Type, which has values that include Barn and Farm Building.

Why mapping between dictionaries is important

- It is clear that different dictionaries, even those that share a common heritage, vary significantly in how they deal with hierarchical concepts.
  - In some cases, **feature concepts are defined at multiple levels of specialization**; in other cases, **attributes are used to further specialize a feature concept**.
  - It is not uncommon for these two approaches to be combined within a single dictionary.
- In M&S applications, data is received from legacy sources (such as those products based on FACC) or new sources (such as those based on DFDD).
  - It is important to **provide a consistent and common mapping approach and terminology** to capture which concept (or concept combination) in a given source dictionary can or should map to which concept (or concept combination) in a target dictionary.
  - The mapping product must **be clear in its terms and semantics of the mapping**, and **can be provided as a software library** that can be easily incorporated into a converter or translation application.
- Because many data products can use the same dictionary, the designer of data conversion applications can **start with existing dictionary mappings**, and **extend to address the data structure and data organization mappings**.

Mapping approach

- Begin by **analyzing a single concept entry in the source dictionary** and **determine if an equivalent concept exists in the destination dictionary**, either as another single entry or as a combination of several entries.
  - Same approach is often used in mapping data models, but in addition it may be necessary to combine multiple concepts to meet specific data model requirements.
- Since in practice data translation (or movement) is in one direction, mapping from a source concept to a destination concept is a single one-way mapping.
  - Mapping from the destination concept back to the source concept is considered to be another, separate, one-way mapping.
  - **A complete two-way mapping is composed of two one-way mappings**.
- Whether a mapping is for concept dictionaries, data dictionaries or data models, at the end either a mapping for a given concept exists, or it doesn't.
  - Particularly **important to application designers** when using concept dictionary mappings, since **partial or potential mappings are not useful, unless specific conditions on how such mappings are to be applied** can be established.
  - During the development of a mapping, some instances may be marked as "unresolved" (with appropriate explanation) until they can be resolved in the final product.

Mapping approach (cont.)

- During the development of a mapping, especially during the analysis phase, it is often possible to map a given source concept to multiple semantically equal destination concepts.
  - In such cases, the **final mapping product should identify only the best, most logical, and most practical of those mappings**.
- A number of mapping types and subtypes exist. These usually involve additional information, special conditions, or identify a collection of concepts in order to provide the same semantic in the destination.
- In addition, explanation, rationale, or analysis information may be provided to help the end user (as well as the reviewers, during the development of mappings) to better understand why a certain type of mapping has been designated for a given concept.
  - Such supplemental information **can be one of several categories of rationale or analysis** and should be **captured in a separate field, adjacent to the mapping type**.

Terminology for mappings

- **Different terminologies** are needed **for different stages or categories of mappings**.
- To establish a mapping for environmental data, a developer may produce **mapping products between concept dictionaries or between data models** (or specific data products).
- During the development of such mapping products there is considerable analysis that will take place to search and analyze the concepts in both the source and destination material.
- Important to be able to **identify, through a shorthand notation**, the **type of analysis** as well as the **type of mapping** associated with a given mapping.
- Description of **specific notation**, **definition of types**, and examples of **how** different categories and **types of terminology can be applied** are provided in the following slides.
- These include terminology used for:
  - **The analysis phase**
  - **Mapping concept dictionaries**
  - **Mapping data models (or data products)**

Terminology for the analysis phase

- In analysis phase, appropriate **terminology**, and **shorthand identifiers**, are needed **to concisely express the rationale for** a given mapping.
  - **Analysis terminology is distinct from the mapping terminology** and should be captured in a separate field associated with a given mapping.
  - Even after mapping completion, the analysis information, along with comments, can be helpful in understanding why the specified mapping was chosen.
  - Such information does not have to be included with the mapping software library that will be used to look up the mapping for a given concept, but is useful in retaining a trace of the rationale for future revisions or reviews.
- **Analysis terminology** is used to describe **the relationship between the source and destination concepts**. Examples include:
  - **Aggregate-component** relationship, and the inverse,
  - **Specific-to-general** relationship, and the inverse,
  - Source concept **is identical** to destination concept,
  - Source concept **is equal**,
  - Source concept **does not have an equal**,
  - Source concept **cannot be mapped**.

Analysis terminology (cont.)

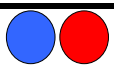
- **Given a pair of definitions** in two different dictionaries, **two sets of real object instances should exist**, conforming to each of those two definitions.
  - For example, given the respective EDCS and DFDD definitions of "Barn", a set of real (farm/agricultural) buildings exist that conform to each of those definitions.
- One way of determining the relationship between the two concepts is to **consider the relationship between those two sets of object instances**.
- There are **only five possible relationships between those two sets**. **Which relationship applies** in a given case can be **determined by** asking **three yes-or-no questions**:

  Q1) Are there instances that conform to both definitions?

  Q2) Are there instances that conform to the first definition, but not to the second definition?

  Q3) Are there instances that conform to the second definition, but not to the first definition?

The possible relationships between the two concepts is given in this table:

Q1	Q2	Q3	Result	Graphic
Y	N	N	Concepts are identical	
Y	Y	N	First concept includes second concept	
Y	N	Y	Second concept includes first concept	
Y	Y	Y	Concepts overlap	
N	-	-	Concepts are disjoint	

Mapping terminology for concept dictionaries

- Mapping terminology identifies how a given concept in the source dictionary maps to a particular concept in the destination dictionary.
  - Within a dictionary of concepts it is only necessary that definitions be unique and unambiguous, they need not be "normalized" such that no two concepts overlap.
- Mapping terminology examples include:
  - The source concept **maps to a specific destination** concept;
  - **No mapping exists** for a given concept;
  - There is a mapping, however the main **concept in the destination dictionary is qualified by one or more attributes from the destination dictionary**;
  - There is a mapping between attributes, but **a data type change is required**;
  - There is a mapping, but **a change in unit of measure** is needed;
  - There is **an interim mapping** that requires **additional determination at data conversion**;
  - There is an attribute mapping, but the concept's **enumerants are specifically split into multiple attribute-enumerant combination** concepts in the destination dictionary
- Terminologies used in the development of concept dictionary mapping products may also be used in developing mappings between data models.

Mapping terminology for data models

- Mapping between data models (or data products) requires additional terminology that is unique to complexities associated with mappings between data models (or products).
  - **Start with the concept dictionary mappings** (if applicable), then **apply additional constraints imposed by** the source and/or destination **data models** (or product).
- Examples of specific data model mapping terminology include:
  - The **source concept is qualified** with other concepts, and the combination has an equal (single) concept in the destination dictionary;
  - The **source concept is qualified** with other concepts in the source dictionary, has an equal in the **destination concept qualified** with other concepts in the destination dictionary;
  - The source concept may be **mapped, if a given condition is met** in the source data.

Summary

- Enabling **interoperability** between M&S systems, and **data integration** from multiple sources, **requires a consistent and common approach to converting the data**.
- Data providers and system developers often have unique or different methodologies for defining the content of their data:
  - Based on **concept dictionaries used in a variety of data models**, or
  - Specific **data dictionaries for** particular **data models or data products**.
- Therefore, having a **consistent methodology and terminology for** providing **data mappings,** especially **for automated data conversions**, is critical.
- **Establishment and use** of a common and consistent **mapping terminology and methodology** is a significant **factor in increasing the interoperability** of systems and applications, and **reducing the development cost of converters**.
- **Approach** presented **is the result of prior and on-going work** in developing mappings between FACC and EDCS, DFDD and EDCS, and NGA's NFDD (which is based on, but not identical to, DFDD) and EDCS.
- ***The same principles are applicable to other (non-environmental) data interoperability challenges in M&S applications.***

Questions ?