Tools for Defining Data

by Abdul-Malik Shakir

The key to using data dictionaries, data sets, and data models effectively lies in understanding their functions and characteristics. The author presents a comparison of each, including the parts they play in health information exchange.

You are unlikely to use an automobile to get from New York City to London—and you are unlikely to use a plane to get from your home to the corner market. Why? Because trains, planes, and automobiles have much in common, but they have many significant differences. Each one can each get you from point A to point B, but each has unique characteristics that make one more appropriate than the others for a particular use.

Similarly, data dictionaries, data sets, and data models have much in common, but many significant differences. Each can be used to define data, especially data that needs to get from point A to point B. This article describes each of these tools and the value they add to health information management. Specifically, we'll look at the role these tools play in the definition of data that is shared or exchanged between individuals, organizations, or clinical information systems, and highlight the unique characteristics that make one more appropriate than another for a particular use.

When data is shared or exchanged, suppliers and consumers of data should align their assumptions about how data is defined, what set of data is expected, and what semantic relationships are inherent within the data set. Performed effectively, this process of revealing and aligning assumptions sets the stage for successful transactions. In the management of health information, a diverse and often widely distributed collection of stakeholders share clinical data. Assumptions that are not aligned can lead to a wide spectrum of difficulties in health information exchange, ranging from provider underpayment to patient death.

The process of defining data sets and constructing data dictionaries or data models is a technological task usually conducted by information technology (IT) professionals. But the task cannot be completed without the involvement of subject matter experts. HIM professionals are often asked to participate in data dictionary or data model projects as subject matter experts. Therefore, it is important to understand the distinction between data dictionaries, data sets, and data models and their application as tools for health information management.

Starting on the Same Page: Data Dictionaries

Data dictionaries provide definitions for data elements that comprise a data set.

The need for a data dictionary and the process for developing one are best illustrated by this true story. In 1991, a large national health maintenance organization (HMO) was planning to create a disease management program for pediatric care. As part of the planning effort, the HMO decided to conduct a survey of its regional sites to determine the utilization pattern for pediatric care. It sent a questionnaire to each of the 12 regional offices asking these two questions:

1. How many pediatric members were enrolled as of year-end 1990?

2. How many pediatric visits took place in 1990?

On the surface, the questions seemed quite simple and appropriate. The HMO would determine the number of pediatric members and the number of pediatric visits by region. It could then compute pediatric utilization by region—and across the program as a whole. This data would then be used to determine a baseline for development of utilization management programs and would assist in comparative analysis of pediatric utilization across regions.

There was only one problem—the absence of common data definitions. Each of the regions operated somewhat autonomously and interpreted the request for information differently. As a result, the regional offices raised a number of questions and revealed numerous discrepancies in their interpretations of data definitions:

What is a pediatric member?

- a dependent member under the age of 18
- a dependent member under the age of 21
- a dependent child member, regardless of age
- a patient under the age of 18

What is a pediatric visit?

- a visit by a pediatric member
- a visit by a patient under the age of 18
- any visit to the pediatric department
- a visit with a pediatrician

Attempts to answer these questions only raised more questions. What is a member? What does it mean to be enrolled? What is a dependent? How is patient/member age calculated? What is a visit? What is a patient and how does one differ from a member? What is a department? What are the department types? What is a pediatrician?

A data dictionary provides answers to these types of questions. This story shows us how important it is that suppliers and consumers of data agree on data definitions before exchanging information. Had the regions not revealed their assumptions, the discrepancies in their interpretations of the data definitions might never have been recognized, and the organization would have unknowingly compared apples to oranges. In the long term, a business strategy with significant implications would have been based upon invalid information.

A data dictionary should provide agreed-to definitions for each data element in the data set and include additional information needed to properly construct the data set. The additional information will document:

- *element group*—the grouping of data elements into data element groups. This includes logical groupings of data elements, such as the parts of a patient's name or address. The logical groupings are sometimes given names and are defined in the data dictionary in a manner similar to less complex data elements

- *element domain*—a specification of the allowable values for each data element. Allowable values are expressed as an enumerated list of values (e.g., gender :: M for male, F for female), a reference to an enumerated list of values (e.g., diagnosis :: SNOMED code), or a predicate expression describing the allowable values (e.g., age :: an integer from 0 to 103). The specification of allowable values may include constraints based upon the value of one or more other data elements in the data set (e.g., discharge date must be greater than or equal to admission date)

- *element conditionality*—a specification of conditional presence for each data element. These conditions range from mandatory (must have a value :: HICN number on Medicare claims) and optional (may be omitted :: patient middle name), to conditional (required to be valued under certain conditions :: if there is a value in discharge date, then there must also be a value in discharge disposition)

- *element default*—a specification of default values for omitted data element. A default value may be specified for optional data elements. This is a specification of what the recipient of the data set can assume to be the value for data elements not sent in the data set (e.g., for outpatient visits, if visit end date is omitted, it will be assumed to be the same as visit start date)

- *element derivation*—a specification of the derivation algorithm for derivable data elements. Derivable data elements may be as simple as counts, sums, differences, and averages, or may involve complex rules and procedures. Data sets containing derivable data elements will sometimes also include the raw data needed to perform the derivation for audit or data validation purposes

A data dictionary project involves collecting data definitions currently in use, assessing the disparity in data definitions, and reconciling the discrepancies. The resulting data dictionary generally contains technical, business, and clinical definitions for data and includes instructions for mapping the technical definition in a computer system to an agreed-to business/clinical definition for use in information collection, analysis, and dissemination. All parties affected by the project should be involved, including data entry and systems personnel, vendors, clinicians, and data analysts. A process for ongoing support and maintenance of the data dictionary must also be established as part of the project.

Developing a data dictionary is a critical step in the implementation of a computerized patient record system or clinical data warehouse. Data dictionaries are also produced by national standards organizations such as X12[1] and HL7[2] as part of specifying EDI data sets and clinical system interface specifications. Regulatory and accrediting bodies such as the Health Care Financing Administration and the Joint Commission on Accreditation of Healthcare Organizations also produce data dictionaries as part of their data set specifications.

A data dictionary is an essential component of any defined data set. A data set specification without an accompanying data dictionary is a formula for disaster. By the same token, a dictionary on its own is of little value unless it is used to provide definitions for data elements in one or more defined data sets.

Collections of Elements: Data Sets

A data set is a defined collection of data elements for a particular business or clinical purpose. The uses of healthcare information for business and clinical purposes are unlimited, so there is an equally unlimited set of definable data sets.

Software developers, regulators, accrediting agencies, government agencies, researchers, payer organizations, provider organizations, and standards bodies have all defined data sets in healthcare. Invariably, the content of these data sets overlaps. Sometimes the overlap is intentional; the developer of the data set may have adopted an existing data set and extended or modified it in some way. However, many times the overlap is unintentional and laden with inconsistencies.

Many attempts have been made to harmonize and reconcile inconsistencies among healthcare data sets. Consortiums are formed, regulations are enacted, and special interest groups are organized—all in the hope of gaining as much participation and consensus as possible among those who will be affected. Examples of these include the consortium for the specification of a public health data set,[3] the Health Insurance Portability and Accountability Act (HIPAA) electronic data interchange administrative transactions sets,[4] and the Health Plan Employer Data and Information Set (HEDIS).[5] Identification and definition of a common minimum data set is the proverbial Holy Grail of healthcare.

Data sets vary in size, complexity, and level of abstraction. They can be as simple as a single data element (e.g., a count of cases of death from cancer) or as complex as a healthcare claim. The level of abstraction in a data set can range from very specific (such as a case, patient, or particular service) to somewhat conceptual (such as a population, event, or disease).

The more complex the data set, the more important it is that the parties who will be using it agree on data definitions and rules for the data set's construction. A data dictionary must accompany the data set. Because data elements are reused across data sets, it is ideal if elements in different data sets can be drawn from a common data dictionary. Of course, a common data dictionary for healthcare would contain literally thousands of data elements. The dictionary needs to be organized in a way that makes it relatively easy to find a desired data element and to detect unintended data redundancies. Data models have become a popular and effective means of meeting the requirements of a reusable data dictionary.

Giving a Context: Data Models

Data models provide a contextual framework and graphical representation that aid in the definition of data elements. In a data model, data elements are organized into classes, and each class represents some object in healthcare. An object can be a concept, person, place, thing, or event. Each class in the data model contains a description of the class of objects it represents. Each class has a set of properties that include the attributes of the class and the associations the class has to other classes in the model. (An attribute is some fact of interest about the class that could be carried in a data element of a data set.)

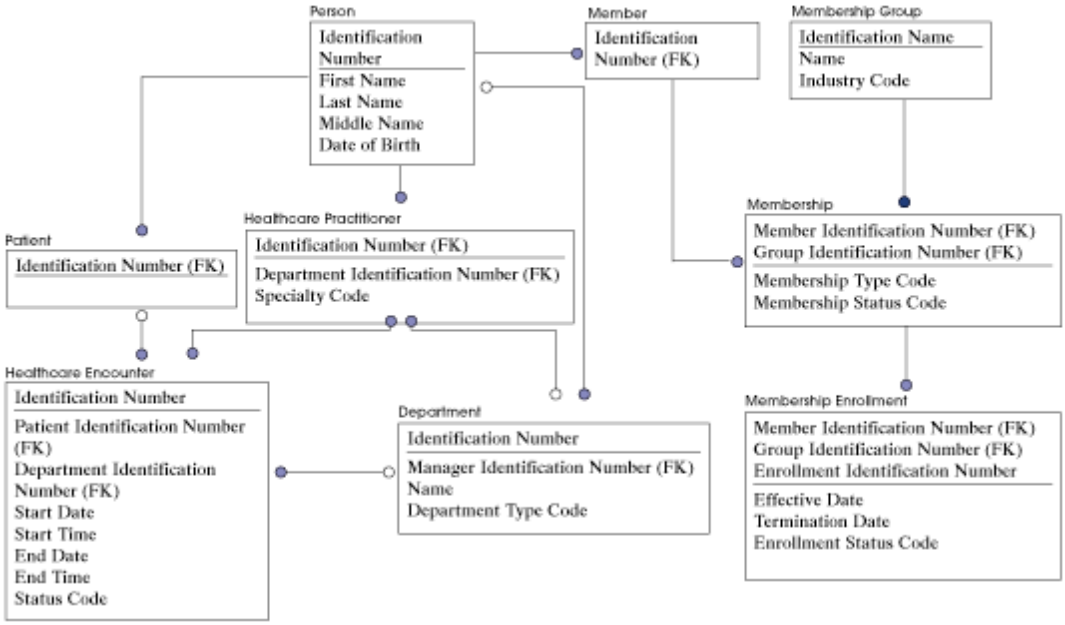
An example of a class is "person" representing individuals such as patients, members, or practitioners. Possible attributes of a "person" might be "name," "date of birth," or "gender." The person might be associated with a "department" class—such as

"person" works in or manages "department."

Data modeling adds a degree of rigor and formalism to the process of building data dictionaries and defining data sets. There are defined rules for constructing data models. IDEF[6] and UML[7] are two popular and widely used standards for data modeling. These standards provide a formal language for expressing and defining classes of concepts and their properties. This formalism is supported by data modeling tools that allow large data models to be developed by teams of modelers and subject matter experts. The data modeling tools provide support for textual data definitions and the graphical representation of those definitions.

Figure 1 is an example of a graphical representation of a data model using the IDEF notation. The rectangular boxes represent classes. The class name appears just above the box and the class attributes appear within the box. The lines connecting classes depict the class associations.

figure 1—data model graphical expression example



Classes within a data model are an effective means of capturing useful, normalized groupings of data elements. These data element groups can then be reused in multiple data sets. Data models also capture reusable data element definitions and domain specifications. Capturing element conditionality, defaults, and derivation specification in the data model is best delayed until using the data model to define data sets. These specifications should be added to the data model's data dictionary as an extension of the model. There is significant advantage to keeping derived data elements out of the base data model and out of the graphical expression. The key advantage is the reduction of redundant data definitions.

A data model has essentially two data dictionaries: a core data dictionary that contains non-derived, unique, atomic data elements, and an extension to the core data dictionary that contains derived, qualified, and composite data elements. The rules of data modeling require data elements (class attributes) to convey concepts that are unique within the scope of the model and are not an encapsulation of significant subordinate concepts. These rules of data modeling, known as normalization, enable the construction of a data model that is resilient, extensible, and applicable to multiple uses. The data elements that are the core of the data model conform to these rules of normalization.

Data elements included in the extended data dictionary are not constrained by the rules of normalization. Derived data elements used in data sets are defined in the extended data dictionary, along with their rules for derivation (e.g., patient age :: encounter start date minus person date of birth plus one). Alias names for data elements may also be defined in the extended data dictionary as equivalent to data elements in the core data dictionary (e.g., admission date :: encounter start date). The description for the alias should also include text that explains when the alias is to be used (e.g., admission date is to be used as the alias for encounter start date when the encounter is an inpatient stay).

Qualified data elements convey concepts that are a subset of a concept in the core data dictionary. The name of a qualified data element contains a qualifier word preceding the name of a core data element. The qualifier word declares the subset. For

example, the core data dictionary might include the data element "last name" in the "person" class, and the extended data dictionary might include "patient last name" and "physician last name" as qualified versions of the "last name" data element.

The extended data dictionary might also include data elements that are composites of data elements included in the core data dictionary. For example, the core data dictionary might include the data elements "enrollment effective date" and "enrollment termination date" while the extended data dictionary includes the composite data element "enrollment period."

Contents of the extended data dictionary are best defined during the process of building a specification for a data set. The data sets derived from a data model will have their entire contents drawn from the data model data dictionaries.

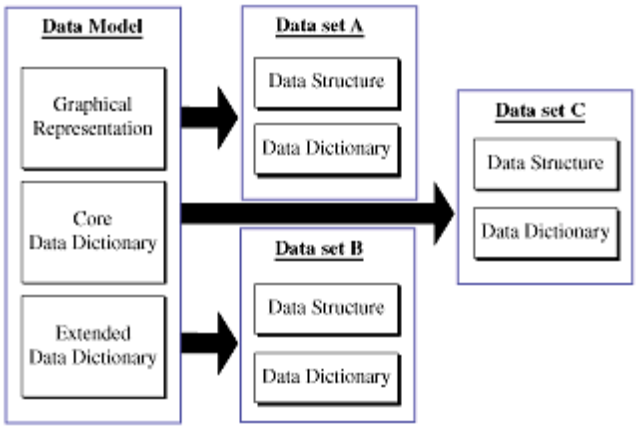
If the data set requires a data element not yet included in the data model data dictionary, it is to be added to the appropriate dictionary so that it is available for reuse in subsequently defined data sets. If the new data element is added to the core data dictionary, it must also be added to the data model's graphical representation. The more the data model is used as the source of data elements for data sets, the more complete it becomes and the more useful it is as a tool for ensuring consistency among multiple data sets.

Conclusion

The features and advantages of data dictionaries, data sets, and data models complement each other and, when used together, enhance their effectiveness in creating well-defined, semantically rich, reusable data elements.

Figure 2 illustrates how multiple data sets can be defined from a common data model. The graphical portion of the data model provides a visual aid to documenting and validating the intrinsic interrelationships that exist among data elements from which contents of the data sets will be drawn. The data model's core data dictionary provides descriptive text for non-derived, unique, atomic data elements. The model's extended data dictionary contains the derived, qualified, and composite data elements and their definitions.

figure 2—multiple data sets derived from a common data model



The data set structures are a particular ordered subset of data elements drawn from the core and extended data model data dictionaries. The data set data dictionary contains the data definitions for the subset of data elements in the corresponding data structure. The data set data dictionary may extend the definition of the data element to include additional data sets with specific constraints such as conditionality, defaults, and allowable values. These additional constraints must not be in conflict with constraints defined in the data model data dictionary. Data set defined constraints govern a particular application of the data element and do not alter the semantics of the data element in any way.

The data model-driven approach to constructing data dictionaries and data sets is an effective way to achieve consistency in data element and data set definitions. The HL7 version 3 message development methodology is an example of this approach. X12 is in the early stages of implementing a similar methodology for the development of its EDI standards. The two organizations have even discussed the possibility of collaborating on development of a common data model for use in defining ANSI standard data sets for use in health data interchange. Participation and support from HIM professionals in the efforts of these two organizations is a critical success factor for achieving uniform health data standards.

Notes

1. X12 is an American National Standards Institute (ANSI) accredited standards committee that develops, maintains, interprets, publishes, and promotes the proper use of American National and UN/EDIFACT International Electronic Data Interchange Standards. For more information, visit the Data Interchange Standards Association Web site at www.disa.org.

2. Health Level Seven (HL7) is an American National Standards Institute (ANSI) accredited standards developing organization that develops specifications, the most widely used being a messaging standard that enables disparate healthcare applications to exchange key sets of clinical and administrative data. For more information on HL7, visit their Web site at www.hl7.org.

3. The Public Health Data Standards Consortium was established to organize the public health and health services research communities on data standards issues. The consortium will expand public health involvement in existing health data standards and content organizations and facilitate the development of new public health data standards. For more information on the Public Health Data Standards Consortium, go to www.lewin.com/hipaa/consorti.htm.

4. The Administrative Simplification provisions of the Health Insurance Portability and Accountability Act of 1996 (HIPAA) are intended to reduce the costs and administrative burdens of healthcare by making possible the standardized, electronic transmission of many administrative and financial transactions that are currently carried out manually on paper. For more information on HIPAA, go to http://aspe.hhs.gov/admnsimp/.

5. HEDIS is a set of standardized performance measures designed to ensure that purchasers and consumers have the information they need to reliably compare the performance of managed health care plans. The performance measures in HEDIS are related to many significant public health issues such as cancer, heart disease, smoking, asthma, and diabetes. For more information on HEDIS, go to www.ncqa.org.

6. IDEF is a technique for constructing models to support activities such as process function analysis/ reengineering and information specification. Models constructed using IDEF are characteristically less susceptible to reader misinterpretation than those built using tools with a less systematic set of rules. IDEF0 is used to perform functional decomposition delineating inputs, controls, outputs, and mechanisms interrelating the functions. IDEF1X is used to define the entities that take part in such processes, providing linkages through verb phrases as well as inheritance. For more information on IDEF, go to http://joy.gsfc.nasa.gov/MSEE/idef.htm.

7. Unified Modeling Language (UML) is a standard notation for the modeling of real-world objects as a first step in developing an object-oriented program. Among the concepts of modeling that UML specifies how to describe are: class (of objects), object, association, responsibility, activity, interface, use case, package, sequence, collaboration, and state. For more information on UML, go to www.whatis.com/uml.htm.

Abdul-Malik Shakir *is a senior advisor with The Huntington Group, a consulting services unit of IDX. He has more than 23 years of experience in information technology, with special emphasis on data modeling in healthcare and the application of data modeling to healthcare information management. He is a board member of HL7 and past chair of the X12 Healthcare Business and Information modeling task group (X12N-TG3). He can be reached at Abdul-Malik_Shakir@IDX.com.*

Article citation:
Shakir, Abdul-Malik. "Tools for Defining Data." *Journal of AHIMA* 70, no.8 (1999): 48-53.
