Data Dictionaries: An Overview

by Philip G. Duffy

Healthcare is one of the most information-intensive industries in the economy but less automated than other industries such as banking and insurance. Modern healthcare delivery is governed by ad hoc teams of professionals assembled to address each case, which further complicates information automation. There is no lack of application systems to serve the healthcare industry, yet each of these appears to have been created in isolation with little regard for other systems. There are few data standards to facilitate communication between application systems, and those that exist are rather limited in scope. Healthcare organizations do not have the luxury of waiting for national data standards to emerge; they must begin by enforcing their own standards internally. In other words, they must create their own data dictionaries.

What Is a Data Dictionary?

There are many textbook definitions of a data dictionary, but most are unnecessarily technical. A data dictionary simply describes the data used by an organization. It is information about information in the same sense that a language dictionary like *Webster's* contains information about English. Just as a language has a certain structure (syntax), so too does an organization's information. This information structure tends to be described as a hierarchy. Individual data elements (fields) exist at the most detailed level. Data elements are aggregated into records of various types, e.g., a patient record, and records can be further aggregated into a larger entity called a database.

Notice from the above that the more specific the data entity, the easier it is to define. Healthcare organizations have little difficulty in understanding the meaning of a patient name or a medical record number, despite variability in the format of such numbers or names. It is more difficult to describe records and their relationships. Most healthcare professionals would acknowledge that it is important to have a master record for each patient who has had an encounter with a healthcare organization. They might also agree that it is important to create a record for each encounter, regardless of the name of that record. It becomes more arbitrary when considering individual events within an encounter such as admission, transfer, or discharge. Even professional system designers differ here. Some would create separate admission, transfer, and discharge records while others would simply collect admission and discharge information within the encounter record. Determining which records should be a part of which database is even more confusing. Unless the objective is to develop an application system rather than acquiring one from another source such as a vendor, healthcare organizations might be wise to bypass record and database definition and concentrate on individual data elements.

Why Develop a Data Dictionary?

Data consistency is a primary need of healthcare organizations and the principal goal of a data dictionary. Healthcare organizations rely on information consistency to care for patients and to report the results of that care. It is easy to achieve data consistency for a data element such as birthdate that is almost uniformly described. It is a little more difficult to describe a concept such as an adult, since such concepts tend to be both culture- and context-dependent. Data dictionaries must focus on these definitions.

In addition to data consistency, successful information system implementation is another reason for creating a data dictionary. While it is theoretically possible to create a data dictionary that applies to a manual system (called a passive data dictionary), that is rare in healthcare today. Modern information systems require data dictionaries as the first step in creating an automated information system (an active data dictionary). Ultimately, well-designed information systems contribute to more timely and appropriate patient care, improved staff efficiency and effectiveness, and improved management and planning.

Data Dictionaries and Coding Systems

At the data element level, data dictionaries are concerned primarily with two kinds of entities:

- Individual data elements, e.g., patient names and medical record numbers
- Coding schemes

As described above, other than the description of formatting rules -- e.g., a medical record number will be 10 characters in length and consist of all numeric characters -- there are few other attributes one might record for an individual data element. Coding schemes are a different matter. Consider a simple concept of race where clinical and external demands tend to collide. Current race coding schemes, for example, seldom identify a mixed race category, and yet that category is now significant and growing. Data dictionary projects must confront such contradictions and work out resolutions that satisfactorily meet all of the goals of the coding system.

Data dictionaries are not coding systems, but there is a close relationship between the two concepts. Every coding system relies upon a relatively simple data dictionary. The ICD coding system, for example, contains a code of a certain length and format, a code description, and some related pieces of information. Healthcare organizations employ hundreds of coding systems, some almost unconsciously (IP and OP for inpatient and outpatient respectively are codes in a coding system). The only linkage between a data dictionary and the coding system occurs when the data dictionary defines which codes or values are permissible. This is an important linkage for a healthcare organization, however. A data dictionary project typically expands to the consideration of related coding systems to eliminate code inconsistencies. These have arisen historically because code creation is done both externally and within multiple departments. Health information management departments and the patient accounting office have traditionally been challenged to maintain consistent coding systems.

Data Dictionary Media

In theory, data dictionaries could be compiled as a series of handwritten notes. A data dictionary in this form might be accurate and reasonable, but it is not likely to command the respect that it requires. While the mere participation in a data dictionary project may enlighten its participants, the real payoff comes when the document is formalized and accepted as the authoritative source by the organization creating it.

At a minimum, a data dictionary should be compiled as a word processing document to facilitate changes and to reflect a certain structure. Minimally the structure of the data dictionary would include the data element name, its description, and format. A better organization of the data dictionary would impose summary categories such as patient-related, encounter-related, physician-related, etc. Any exceptional information about the data element should also be included. For example, does the patient name mean the current legal name that excludes maiden names and aliases?

While a word processing document is a big step up from handwritten notes, it is a long way, in terms of utility, from the use of a database management system. The best database management systems have rich data description capabilities, and they enforce a description discipline that is not present in a data dictionary created with a word processor. A sophisticated database management system can offer the following features for dictionary creation:

- A structure of forms that prompt the user for data dictionary elements
- A mechanism for describing attributes of data elements such as the data type (integer, decimal, character) and the number of occurrences of the data element (monthly statistics for a year would have 12 occurrences)
- Labels to be used in reports and screens
- On-screen data entry help messages
- Edits that always apply to the data item (e.g., an admission date should never be greater than the corresponding discharge date), and error messages if those edits are violated
- Programs that should be executed before or after entering certain information in a screen (triggers). For example, a trigger program could be initiated for an attending physician's entry. That trigger program would first determine if the physician was on staff and the physician's admitting privileges not suspended

Types of Database Management Systems

Historically, a limited number of files were created to support a relatively restricted application system such as patient accounting. Many of these original systems were designed as encounter-based systems. That meant that a patient's name,

address, and other demographics were repeated every time the patient had an encounter with the healthcare organization. As system designers learned more about the healthcare environment, they created additional records such as a lifetime patient record that eliminated some data repetition. To completely eliminate unnecessary data repetition in a file system (a process called normalization by system designers) required the construction and processing of very complex file structures. Application developers turned to specialists to provide the generic tools (database management systems) to create these.

Fundamentally, a database management system is software that enables the creation of a complex file system that includes many record types. The database management system also provides the generic software that allows application programs to create, read, write, and delete records in a database. Database management systems have been in existence since the late 1960s, but they did not dominate the information systems industry until relational database management systems became commercially available in the early 1980s. A relational database management system is a collection of records of different types that are connected together by primary index keys. Relational databases have the following advantages that are generally not found in earlier database management systems:

- They are intuitive to a nontechnician. Many nontechnical people are surprised by the enthusiasm over relational databases because they had always assumed this is the way information has always been stored on a computer system. To the contrary, information is actually stored according to physical entities such as disks, tracks, and sectors. All database management systems hide this physical layer from application developers, but relational database management systems do this in a particularly natural way that facilitates communication between the technician and the nontechnical user
- They are particularly adept at supporting ad hoc reporting. Older file systems frustrated users. Users knew information had been collected in the system, yet they had to plead with an overwhelmed information systems department to produce custom and ad hoc reports. With the advent of report writers for relational database management systems, many users are able to generate ad hoc and custom reports themselves
- They are commercially available. Previous database management software had been costly and available from limited sources. Relational database management systems, on the other hand, are more reasonably priced and widely available

Today, most new application systems are developed with a relational database, but already there are signs that another type of database management system may either replace the relational database or become a part of it-the object-oriented database management system.

Distinguishing them from relational databases, object-oriented database management systems segregate entities into classes and subclasses that "inherit" characteristics from the parent class. For example, an object-oriented system might be based on persons as the super class with subclasses of patients, physicians, employees, visitors, and volunteers. All persons have common attributes such as age, height, weight, and a principal address. These attributes are inherited by all subclasses of persons, including patients and physicians. Both patients and physicians have distinguishing characteristics that are defined in an object-oriented data dictionary. Physicians, for example, have professional licenses that are not required to be a patient. The value of this classification scheme is that a single program can be written to maintain all of the common attributes of all subclasses of persons. This is the foundation of software component reuse.

Data Dictionaries -- Present and Future

By viewing data dictionaries on a time line that spans from 1960s to the present, it is possible to make some reasonably accurate predictions. It is important to note that healthcare has traditionally lagged behind the economy's other information-intensive industries.

Sample of Database Management	
PC-based	**Departmental and Enterprise-wide**
Access	Informix
Clarion	Oracle
Paradox	PROGRESS SQL Server Sybase

While relational database management systems have become significant for new development in healthcare, healthcare continues to be dominated by major systems that employ prerelational file systems. Protected within the walls of healthcare, many assume that nothing will change. But forces like managed care are already pounding those walls, because the technology required to appropriately support managed care cannot be found in old file systems. The newer technology will overwhelm the old, more likely sooner than later.

While the data dictionary set of capabilities, as described above, is only one part of a database management system, it is the foundation. By knowing where commercial data dictionaries are going, one can predict where larger database management systems are going.

It is clear that data dictionaries will become more object-oriented, because object orientation is virtually mandatory for building reusable software components. It is less clear whether object orientation will require the replacement of relational databases. For most healthcare professionals, this is a battle that the technologists can wage without their involvement.

On the other hand, the world of reusable software components will be more demanding on healthcare. Most healthcare organizations are struggling to create data dictionaries at the level of the more concrete relational database management system. It is not likely that many will be able to make the leap to an object-oriented data dictionary without professional assistance. Healthcare professionals who are comfortable with patient and encounter records may find it more difficult to deal with abstract persons who are patients in one context and physicians in another. While these abstractions correspond to real life, they are more difficult to conceptualize.

The scope of data dictionaries undoubtedly will expand. Traditional data dictionaries described records that were capable of collecting patient, encounter, and other information. Future data dictionaries are likely to describe medical image objects, documents, photographs, sound clips, and other nontraditional data. In addition, the traditional concept of an application is likely to further erode as health professions press for more integrated, and therefore more consistent, data. In most cases it will not be possible to limit one's thinking to the boundaries of an application. Instead an enterprise-wide view of data and information system processing will be required.

These predictions suggest a growing gap between the requirements of building a data dictionary and the resources available to do the job. The commercial world tends to bridge these gaps as profitable market opportunities. Therefore it is reasonable to predict that specialized vendors will emerge that can create the foundation for all information systems, along with the necessary data dictionaries. Although vendors, both present and future, may relieve healthcare organizations of some of the effort in creating data dictionaries, they do not relieve the organization of its responsibility in establishing them.

Reports from the Field

The following comments on data dictionaries appeared on HIM-L, an Internet discussion group on health information management topics.

We are just starting to compile a data dictionary. It seems like an overwhelming task! We will be using our SPARCS (state database to which we submit data) data dictionary as a foundation. Our first initiative is to define the elements in our base table files and append to the SPARCS data element description.

-Polly Nolan, RRA

Our data dictionary effort began with the coding/abstracting module with our IS. This module was chosen because the majority of reports used for administration/marketing and medical staff are generated from this module. As part of an in-depth study I was preparing, it became apparent that the data fields had subjective definitions due to a variety of reasons-staff turnover, system updates, changed mapping, etc. . . .Each data field was reviewed with mapping identified and definitions determined. Uses of the data were also determined . . .A hard copy manual was created with hopes to put online look-up help this coming year. The Quality Improvement Council has recommended that additional modules follow the same process, starting with the admissions module since a majority of their entry maps to the coding/abstracting module . . .

-Susan Nelson, RRA

During the past year, I have coordinated a group, within our state mental health system, to develop a data dictionary. The dictionary is a resource for 13 facilities plus a central billing office. Our system has components for HIM, admitting, patient relations, and accounting.

-Becky Brown, ART

How to Begin a Data Dictionary

Data dictionaries share two characteristics with cathedrals -- they are large and complex structures, and they never seem to be completed. It is no wonder that healthcare organizations are reluctant to initiate a data dictionary project. And yet, as with most major undertakings, the data dictionary begins with relatively simple steps:

1. **Define the scope of the project.** Will the scope be a single application, a set of clinical applications, or all clinical applications?
2. **Determine if the results of data dictionary building should be associated with automated system development or acquisition.** If the data dictionary is to be active, the data dictionary capabilities of the development system can be used. Otherwise, select any low-cost database management system with a robust set of data dictionary capabilities to create a passive data dictionary.
3. **Determine who should oversee the project and who should participate.**
4. **Learn from the experiences of others.** Use the Internet to locate that experience. (See the sidebar "Reports from the Field.")
5. **Define the entities that are most important at your healthcare organization.** It is at this step that a healthcare organization must simply jump in and start describing its information structure.

Philip G. Duffy *is president of Electronic Cottage Associates, West Chester, PA. Readers may reach Phil Duffy at his e-mail address: eleccott@pond.com*

Article citation:
Duffy, Philip G. "Data Dictionaries: an Overview." *Journal of AHIMA* 68, no.2 (1997): 30-34.