Appendix C
Completeness, Consistency, and Integrity of the Data Model

Purpose

This appendix discusses how to assess the completeness, consistency, and integrity of the data model and metadata associated with the data model as part of data quality assessment for a database (as described in DQAF Measurement Types 1–5) (See Table C.1.). To be meaningful, data requires context. In databases, the data model, including its metadata, contributes significantly to this context. If information in the model is incomplete or inconsistent, it introduces risk. Data consumers may make poor decisions about which data to use, and technical processes may not function as expected.

Table C.1 Measurement Types Related to the Completeness, Consistency, Integrity of the Data Model

Number	Dimension of Quality	Measurement Type	Measurement Type Description
1	Completeness	Dataset completeness—sufficiency of meta and reference data	Assess the sufficiency and quality of metadata and reference data
2	Consistency	Consistent formatting within one field	Assess column properties and data for consistent formatting of data within a field
3	Integrity/Consistency	Consistent formatting cross-table	Assess column properties and data for consistent formatting of data within fields of the same type across a database
4	Consistency	Consistent use of default value in one field	Assess column properties and data for default value(s) assigned for each field that can be defaulted
5	Integrity/Consistency	Consistent use of default values, cross-table	Assess column properties and data for consistent default value for fields of the same data type across the database

A *data model* is a visual representation of data content and the relationships between data entities and attributes, created for purposes of understanding how data is or could be structured. Data models are tools for understanding data content, as well as for enabling the storage and access of data. Different types of data models (conceptual, logical, physical, and models of data consumer-facing views of the data) present different levels of abstraction.

The process of modeling involves decisions about how to represent concepts and relate them to each other. Modeling is rarely accomplished by one individual. Launching a large database usually requires a team of modelers, and because databases change over time, different people contribute to the model. Since the process involves many individuals and many decisions, it is difficult to be

consistent in all the details. However, unexpected differences in a data model can be confusing to data consumers, as well as to those responsible for managing data. Therefore, it is important to identify and remediate inconsistencies.

Process Input and Output

As input for this assessment, it is necessary to understand first which models exist and their relation to each other. For purposes of this discussion, we will describe assessing the physical data model of a large database. Since a data model is a visual representation and since the conventions of representation convey a significant amount of meaning, it is important to have a copy of the model in electronic or paper form. It is also necessary to have metadata from the model in a usable form. The assessment includes a set of comparisons between content in different fields. To be usable, metadata from the model should be in a spreadsheet or in database tables that can be queried and through which comparisons can be made. You will need output at the table, column, and relationship levels. Once it is in this form, metadata can be assessed using techniques similar to those used to assess the quality of any other data.

In planning for the assessment, you should determine which metadata attributes to focus on. Basic metadata includes definitions of entities and attributes represented in a database, their domains of valid values, along with details about their physical characteristics, such as data type and field length. Measurement Types #1–5 focus on overall completeness of metadata and reference data, as well as the consistency and integrity of data format and default values. Format includes data typing and field precision.

An additional step in planning is to identify metadata used to manage the database itself that can help you confirm what tables and columns are physically present in the database. Most databases include system catalog tables that database administrators (DBAs) and other technical staff use to ensure the database is functioning as expected.

Finally, you should have any documented standards used for modeling in order to assess the degree to which these have been followed. These include naming conventions, data-typing conventions, and assertions about particular kinds of data elements that should be treated consistently.

The goal of the assessment is to identify and document discrepancies from standards, internal inconsistencies, and any other findings that an analyst finds questionable, along with recommendations to resolve such discrepancies. Ideally, output should include additional assertions for standards that can be incorporated into the model.

High-Level Assessment

High-level assessment of the model determines how complete the metadata is. The following activities can be included in a high-level assessment of metadata sufficiency.

- Produce a listing of distinct entities in the database, based on the system catalog tables. Compare this output to the model to identify missing or redundant representations.
- Identify instances where reference or code tables define a domain of valid values for a core or fact table. Determine whether the degree to which reference tables represented or referenced in the model are present in the database.

- If there are redundancies at the entity level, they are more likely to be related to reference data than to core data. Look across reference datasets to identify instances where similar concepts appear to be represented. Compare definitions and valid values to determine if there is, in fact, redundancy.
- At the metadata field level, identify any instances where metadata is missing from one of the fields being assessed. Put in SQL terms; query the metadata looking for WHERE a field is blank or NULL.
- Acting on high-level findings may simply involve adding missing entities or definitions to the model

Detailed Assessment

Detailed assessment requires profiling metadata to find discrepancies in how columns are named, defined, or typed. A few examples (from an actual production data warehouse) will illustrate the kinds of findings that show up in metadata.

Each column should be named and defined. Columns that represent the same concepts should be named and defined consistently. Table C.2 illustrates differences in naming conventions for the column business name and differences between descriptions for the attribute *address line one*. One set of business names contains the number 1, while the other spells out the word *one*. These differences demonstrate that even a less-than-controversial attribute can be defined inconsistently—and not defined particularly well. (What will address line one text be if it is not the mailing address?) For a generic attribute like the first line of an address, these differences may not pose a significant risk. Unless your organization depends on accurate address information, or if its address information has any complexity to it (if, for example, it contains addresses from more than one country); in such cases, the lack of clarity in these definitions could have negative consequences over time.

The same concepts should be represented in the same way across a database. Table C.3 illustrates inconsistencies with representation at the column name, valid values, and valid value definition levels. The concept represented in the database is whether or not a record is active. There are two ways of naming this concept: Active *Code* and Active *Indicator*. For Active Indicator, there are two sets of valid values. The first set, Y and N, make sense because the field is an indicator, and indicators should be used to answer simple yes/no questions. The second set (A for active, I for inactive, and U for unknown) makes sense in terms of the concept by not the field type.

Table C.2 Inconsistent Business Names and Definitions

Column Database Name	Column Business Name	Column Description
ADR_LN_1_TXT	Address Line 1 Text	The first line of the address for the account holder. This will be the mailing address if available.
ADR_LN_1_TXT	Address Line 1 Text	The first line of the address for the employer as specified by the postal office
ADR_LN_1_TXT	Address Line One Text	The first line of the address.
ADR_LN_1_TXT	Address Line One Text	User's business address 1.

Table C.3 Same Concept Represented and Defined Differently

Column Database Name	Column Valid Value Code	Column Valid Value Text
ACTV_CD	A	Active
ACTV_CD	I	Inactive
ACTV_IND	N	No—this report is no longer being run
ACTV_IND	Y	Yes—this is an active report
ACTV_IND	A	Active
ACTV_IND	I	Inactive
ACTV_IND	U	Unknown

Table C.4 Variations on a Concept Represented Consistently

Column Database Name	Column Valid Value Code
BIL_PROV_ZIP_CD	For valid values, see table ZIP_CODE.
CHK_ZIP_CD	For valid values, see table ZIP_CODE.
FEE_SCHED_ZIP_CD	For valid values, see table ZIP_CODE.
FEE_ZIP_CD	For valid values, see table ZIP_CODE.
LEG_ADR_ZIP_CD	For valid values, see table ZIP_CODE.
MBR_ZIP_CD	For valid values, see table ZIP_CODE.
NAT_FEE_SCHED_ZIP_CD	For valid values, see table ZIP_CODE.
OVTNS_ZIP_CD	For valid values, see table ZIP_CODE.
PAYE_ZIP_CD	For valid values, see table ZIP_CODE.
PCP_ZIP_CD	For valid values, see table ZIP_CODE.
PRI_NAT_FEE_SCHED_ZIP_CD	For valid values, see table ZIP_CODE.
PROV_ZIP_CD	For valid values, see table ZIP_CODE.
PUB_ADR_ZIP_CD	For valid values, see table ZIP_CODE.
SBSCR_ZIP_CD	For valid values, see table ZIP_CODE.
SEC_NAT_FEE_SCHED_ZIP_CD	For valid values, see table ZIP_CODE.
SRVC_PROV_ZIP_CD	For valid values, see table ZIP_CODE.
ZIP_CD	For valid values, see table ZIP_CODE.

Again, for fields like this, there appears relatively little risk of data being misunderstood, so the differences are more an annoyance than a problem. Of more concern is what the presence of such inconsistency implies: a lack of attention to the details of managing metadata on which data consumers depend.

In contrast to the first two examples, Table C.4 illustrates consistency in metadata. The table contains the database names for a set of fields that contain ZIP code data. The database contains a ZIP code table. So the metadata consistently directs data consumers to that table to obtain valid ZIP codes.

Unfortunately, this level of consistency does not carry over to the description of the column. Table C.5 represents a set of definitions that are trying to say the same thing, but have minor

Table C.5 Inconsistent Definitions of the Same Concept

Column Database Name	Column Business Name	Column Description
MBR_ZIP_CD	Member Zip Code	Five-digit U.S. Postal ZIP Code of the Product Service Area.
MBR_ZIP_CD	Member Zip Code	The claimant or member's 5-digit U.S. Postal ZIP Code.
MBR_ZIP_CD	Member Zip Code	The number assigned by the U.S. Postal Service to a geographic area for the purposes of efficient mail sorting and delivery.
MBR_ZIP_CD	Member ZIP Code	The claimant or member's 5-digit U.S. postal code.
MBR_ZIP_CD	Member ZIP Code	The claimant or member's 5-digit U.S. Postal ZIP Code.
MBR_ZIP_CD	Member ZIP Code	The number assigned by the U.S. Postal Service to a geographic area for the purposes of efficient mail sorting and delivery.
MBR_ZIP_CD	MEMBER ZIPCODE	NULL
MBR_ZIP_CD	ZIP Code	The number assigned by the U.S. Postal Service to a geographic area for the purposes of efficient mail sorting and delivery.

Table C.6 Inconsistent Default Values and Data Types

Column Database Name	Column Default Value Text	Data Type Code
MBR_ZIP_CD	0	CHARACTER
MBR_ZIP_CD	0, 99999, Space	CHARACTER
MBR_ZIP_CD	99999	CHARACTER
MBR_ZIP_CD	No default value identified	CHARACTER
MBR_ZIP_CD	No default value identified	VARCHAR2
MBR_ZIP_CD	Space	CHARACTER
MBR_ZIP_CD	Spaces	CHARACTER

variations in wording and capitalization that make them look different in the metadata tables. Table C.6 includes analysis of the default values and data types associated with the metadata for the same column as it appears on different tables. In most cases, ZIP code is defined as a character field, but in at least one instance, it is defined as VARCHAR2. There are at least three functional default values for ZIP code: 0, 99999, and space. However, these are represented in several different ways (both as singular *space* and plural *spaces*, for example).

Address lines and ZIP codes are relatively straightforward to define, and most people using address data in an American context understand what they represent. The examples illustrate the kinds of variation that accrue in metadata as different people contribute to the data model and standards are not enforced.

Table C.7 Inconsistent Names and Definitions of the Same Concept

Column Database Name	Column Description
AGT_BRKR_TIN	The federal tax identification number (TIN) assigned to the agent or broker by the Internal Revenue Service.
ALT_PAYE_TIN	Par Providers only. The federal tax identification number (TIN) assigned by the Internal Revenue Service to the entity/person authorized to receive payments for services rendered by an individual provider.
BIL_PROV_TIN	The federal tax identification number (TIN) assigned to the provider or alternate payee by the Internal Revenue Service.
FED_TAX_ID_NBR	The 9-digit federal taxpayer identification number assigned to the Legal Entity or Customer.
FED_TAX_ID_NBR	The federal tax id of the provider.
FED_TIN	The 9-digit federal taxpayer identification number assigned to the Legal Entity or Customer.
PCP_TIN	The Tax ID of the member's assigned PCP at the time that the capitation was calculated.
PD_PROV_TIN	The pay-to TIN which the capitated contract uses for payment distribution.
PD_PROV_TIN	The federal tax identification number (TIN) assigned to the provider or alternate payee by the Internal Revenue Service.
PRI_PHYSN_TIN	The federal tax identification number (TIN) assigned to the Primary Physician by the Internal Revenue Service.
PROV_TIN	The federal tax identification number (TIN) assigned to the servicing provider or alternate payee by the Internal Revenue Service.
PROV_TIN	NULL
PROV_TIN	The federal tax identification number (TIN) assigned to the servicing provider or alternate payee by the Internal Revenue Service. Please note that the term *alternate payee* is not applicable to some sources.
SRVC_PROV_TIN	The federal tax identification number (TIN) assigned to the provider or alternate payee by the Internal Revenue Service.
SRVC_PROV_TIN	The federal tax identification number (TIN) assigned to the servicing provider or alternate payee by the Internal Revenue Service.
TIN	The federal tax identification number (TIN) assigned to the legal entity by the Internal Revenue Service.

Table C.7 illustrates a different sort of problem. It contains definitions of the Federal Tax Identification Number, or TIN. TIN is a nine-digit number assigned to businesses by the Internal Revenue Service so that the IRS can collect taxes. Like Social Security Number (SSN), it is also used by other organizations as a means to identify specific entities. In health care data, it is one of several identifiers that can help associate records related to the same provider. As was true of the examples in Tables C.2 and C.5, the definitions of TIN are presented inconsistently. What is more important in Table C.7, though, is additional information about the business relationships that is embedded in the definitions—for example, the assertion that alternate payee TIN (ALT_PAYE_TIN) applies only to participating providers and only to specific data sources. These assertions are not directly related to the definition of TIN. But they appear important enough that they should be captured elsewhere in the

Table C.8 Contradictory Definitions

Column Database Name	Column Description
ALT_ID	A randomly assigned 11-digit identifier that can be used to identify data at a subscriber or an employee level.
ALT_ID	A unique identification number, other than SSN, assigned by the source to identify the subscriber.
ALT_ID	For Source A this is a randomly assigned 11-digit identifier that can be used to identify data at a subscriber or an employee level. For Source B this field will be populated with the new MEMBER_ID value if the Member row was once built with the old SSN based member id's.
ALT_ID	Social Security Number of the subscriber of the policy which the card requestor is also part of

metadata rather than at the level of the individual column definition. The example points to the ways that metadata itself can benefit from being organized relationally. Whether it applies to a billing provider, servicing provider, or agent/broker, a TIN is still a tax identification number. The concept defining should be the same for all instances. However, there should also be a place to capture additional details related to the specific instances of TIN, if, as in the case of Alternate Payee, these details have an effect on how data consumers might understand the data.

Another serious problem is illustrated in Table C.8, where we find contradictory definitions of a field with the same name, Alternate Identifier. In one definition, Alternate Identifier is specifically defined as NOT a Social Security Number ("A unique identification number, other than SSN, assigned by the source to identify the subscriber."). In another it is specifically identified as a Social Security Number ("Social Security Number of the subscriber of the policy which the card requestor is also part of"), and in a third, it is defined as replacing a Social Security Number.

Based on the name, most people would assume that Alternate Identifier is the same attribute, regardless of where it appears in the database. However, the definitions imply that it represents different concepts on different tables. In order to understand the risk associated with such an attribute, it would be necessary to confirm exactly what is being populated in each field and, more importantly, to determine how data consumers use the data.

Quality of Definitions

Definitions are critical to data management and to data quality, yet historically, organizations have paid very little attention to their role in enabling data management (Chisholm, 2010). One of the most important aspects of assessing data quality is having access to clear, comprehensible data definitions. Assessing the completeness, consistency, and integrity of the data model also includes assessing the quality of definitions for concepts, entities, and attributes.

A definition is a statement or an explanation that gives the meaning of a word. ISO 11179, the Metadata Registry Standard, defines *definition* as "a representation of a concept by a descriptive statement which serves to differentiate it from related concepts." Definitions should take a particular form,

stating the term to be defined, the broader class it belongs to, and the features that distinguish the term from others in its class. Definitions can be expanded to provide information about the concept being defined. For example, they can provide examples of the term in use, identify synonyms for the term, or even describe what the term does not mean (Inmon, O'Neil, and Fryman, 2008).

ISO 11179 provides guidance for what constitutes a good definition. It should:

- Be stated in the singular.
- State what the concept is, not just what it is not.
- Be stated as a descriptive phrase or sentence.
- Contain only common abbreviations.
- State the essential meaning of the concept.

Definitions should also be precise and unambiguous and be understandable on their own. They should not embed rationale, functional usage, or procedural information. They should not be circular.[1] A circular definition defines the thing in terms of itself. For example, defining an *address type code* as "a code that describes address types" is circular.

As is apparent from many discussions on data governance, most organizations do not invest time in developing robust definitions of key terms that can be used across the enterprise.[2] (Unfortunately, this means that they spend a lot of time producing somewhat adequate but largely inconsistent definitions that are used in silos within the enterprise. And they often repeat this process every few years.) Why is this the case? I think there are at least three contributing factors. First, in our everyday lives, we don't spend a lot of time defining terms, and in most situations, we can communicate reasonably well without doing so. So we do not always recognize the need for establishing clear definitions of terms. But data is different. It always represents something else, and given the amount of data in most organizations, there is great risk involved in using data that is not adequately defined (Chisholm, 2010). Next, the conditions under which many organizations try to define terms are not conducive to success. Not everyone is skilled at formulating definitions, but projects and governance efforts are frequently organized around pulling together large groups of subject matter experts to write definitions. Much time is consumed in wordsmithing but the quality of the end products does not reflect the investment. The politics of such efforts makes the process worse. Most people do not want to seem stupid. So if they read a definition that does not quite make sense, they have one of two reactions. Either they go along with it because it sort of makes sense and they get the basic idea, or they completely revise it, bringing the whole committee back to square one. Finally, very few organizations actually manage their definitions in a way that enables people to have one source for all terms. So there is a burgeoning pile of glossaries and dictionaries, but not a consistent, reliable source of meaning.

In this situation, many organizations turn to technology, seeking a metadata "solution." This reaction makes sense. It is kind of fun to test out a new metadata solution. It is less so to clean up or try to fill out the missing pieces of any existing one. Technology by itself is not the answer. Business metadata requires human thought. Organizations should hire writers to write definitions, rather than expecting SMEs to be writers. Assessing the quality of definitions involves risking looking dumb.

[1] See Chisholm (2010), particularly Chapters 8 and 9, for additional characteristics of good definitions in the context of data management.

[2] See the conference programs for 2009–2012 DGIQ conferences.

It takes someone who is willing to say, "I do not understand what this means" to identify metadata that is not working. Improving definitions requires someone with writing skills who can put the right words in the right places so that months or years hence, another person can understand the meaning of a term.[3]

Summary

This appendix reviewed actions that can be taken to assess the condition of the data model and the definitions at its heart. While some of these actions can involve measurements (for example, identifying the percentage of columns that are missing definitions or the number of inconsistent representations of the same concept, etc.), the purpose of this assessment is to understand the general condition of the model and its metadata and identify actions to improve both. Such actions can include the formulation of standards to be applied to the model or to the revision of definitions.

The first time such an assessment is conducted it will be time-consuming, especially if the model has not been maintained or if time was not taken initially to develop high-quality definitions. Metadata should be periodically reassessed to ensure that it is maintained in conjunction with the growth and evolution of the database. Since metadata is required for other measurements, it should always be an object of scrutiny as part of those measurements. Recommendations after the first assessment should include approaches for long-term management of this critical information. Management of definitions requires involvement from business SMEs who can provide expertise and insight on content. Such involvement has positive repercussions. SMEs who contribute to metadata maintenance gain knowledge of the model and can help others use its metadata effectively.

[3]For a great book on the drama of dictionaries and a study in the power of effective project management, see *The Professor and the Madman: A Tale of Murder, Insanity, and the Making of The Oxford English Dictionary* by Simon Winchester.