CHAPTER 5

Data Profiling

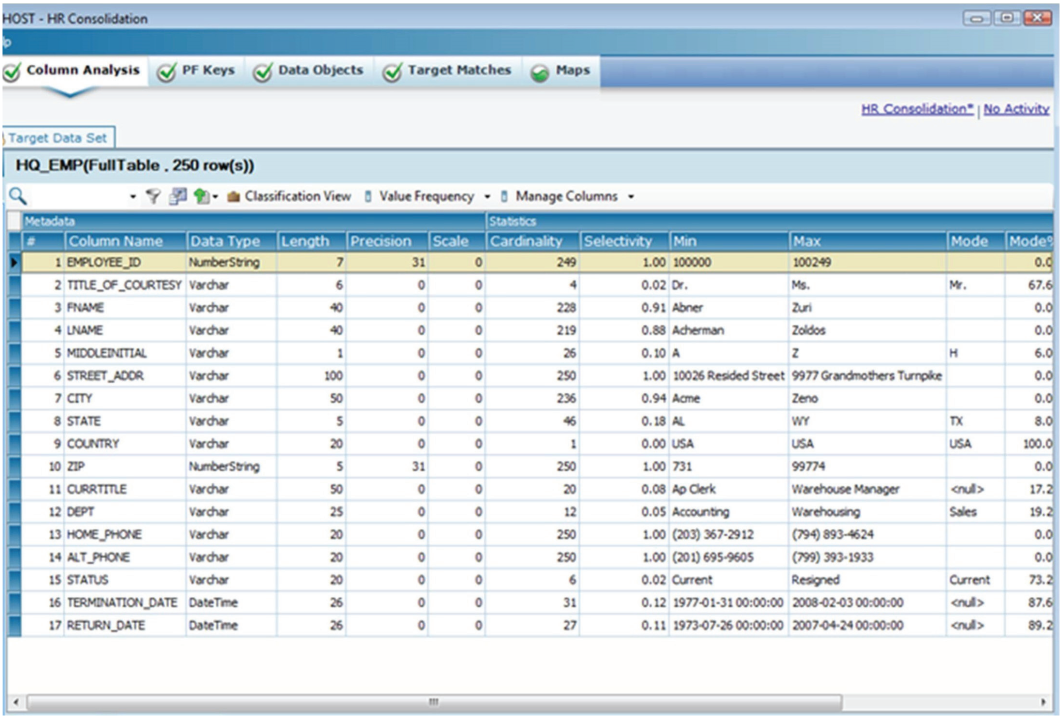
Data profiling is the process of understanding the data in a system, where it is located, and how it relates to other systems. This process includes developing a statistical analysis of the data such as data type, null percentages, and uniqueness. While there might be some nuances, we will use the terms "data profiling" and "data discovery" synonymously. In the absence of tools, data analysts have historically resorted to the use of SQL queries to discover and profile data. Data profiling tools can automate a number of tasks associated with data governance.

Conduct Column Analysis

The first step in any data profiling exercise is to conduct an analysis of the columns. In Figure 5.1, IBM InfoSphere Discovery displays a column analysis for the HQ_EMP table. The column analysis displays basic metadata about each column, as discussed below (not all metadata is shown in the screenshot):

- *#*—The sequence number of the column.
- *Column Name*—The name of the column as shown in the database table.
- *Data Type*—The data type, such as NumberString, Varchar, and DateTime. For example, the data type for the EMPLOYEE_ID column in Figure 5.1 is NumberString.
- *Length*—The defined length of the column. For example, the length of EMPLOYEE_ID is seven characters.

- *Precision*—The maximum number of digits that can be present in a number. For example, EMPLOYEE_ID can have a maximum of 31 digits.
- *Scale*—The maximum number of decimals after the decimal point. For example, EMPLOYEE_ID has zero digits after the decimal point.
- *Cardinality*—The number of unique values in a column. For example, FNAME and LNAME have 228 and 219 unique values, respectively.
- *Selectivity*—The degree of uniqueness of the values (including nulls) in the column, calculated as *Cardinality / (Row Count – Null Count)*. Selectivity is calculated on each column individually and is not the result of comparison to another column. This value is never greater than one.
- *Min*—The smallest or lowest value in the column, calculated numerically for numeric columns and alphabetically for other columns.
- *Max*—The largest or greatest value in the column, calculated numerically for numeric columns and alphabetically for other columns.
- *Mode*—The most common value in the column, not including null values. This value is calculated only if a particular value is displayed in more than five percent of the rows. In Figure 5.1, the mode for STATE is TX.
- *Mode%*—The number of times the mode (the most common value) is displayed in this column, as a percentage of all values in the column. For example, TX appears eight percent of the time in STATE.
- *Sparse*—Indicates whether the column is sparse, based on the Mode %. A sparse column contains mostly the same value except for a few exceptions.
- *Null Count*—The number of rows where the column value is null.
- *Blank Count*—The number of rows in the column that are blank (empty).



HOST - HR Consolidation

HQ_EMP(FullTable . 250 row(s))

Metadata						Statistics					
#	Column Name	Data Type	Length	Precision	Scale	Cardinality	Selectivity	Min	Max	Mode	Mode%
1	EMPLOYEE_ID	NumberString	7	31	0	249	1.00	100000	100249		0.0
2	TITLE_OF_COURTESY	Varchar	6	0	0	4	0.02	Dr.	Ms.	Mr.	67.6
3	FNAME	Varchar	40	0	0	228	0.91	Abner	Zuri		0.0
4	LNAME	Varchar	40	0	0	219	0.88	Acherman	Zoldos		0.0
5	MIDDLEINITIAL	Varchar	1	0	0	26	0.10	A	Z	H	6.0
6	STREET_ADDR	Varchar	100	0	0	250	1.00	10026 Resided Street	9977 Grandmothers Turnpike		0.0
7	CITY	Varchar	50	0	0	236	0.94	Acme	Zeno		0.0
8	STATE	Varchar	5	0	0	46	0.18	AL	WY	TX	8.0
9	COUNTRY	Varchar	20	0	0	1	0.00	USA	USA	USA	100.0
10	ZIP	NumberString	5	31	0	250	1.00	731	99774		0.0
11	CURRTITLE	Varchar	50	0	0	20	0.08	Ap Clerk	Warehouse Manager	<null>	17.2
12	DEPT	Varchar	25	0	0	12	0.05	Accounting	Warehousing	Sales	19.2
13	HOME_PHONE	Varchar	20	0	0	250	1.00	(203) 367-2912	(794) 893-4624		0.0
14	ALT_PHONE	Varchar	20	0	0	250	1.00	(201) 695-9605	(799) 393-1933		0.0
15	STATUS	Varchar	20	0	0	6	0.02	Current	Resigned	Current	73.2
16	TERMINATION_DATE	DateTime	26	0	0	31	0.12	1977-01-31 00:00:00	2008-02-03 00:00:00	<null>	87.6
17	RETURN_DATE	DateTime	26	0	0	27	0.11	1973-07-26 00:00:00	2007-04-24 00:00:00	<null>	89.2

Figure 5.1: IBM InfoSphere Discovery displays the column analysis for the HQ_EMP table.[1]

Discover the Values Distribution of a Column

Data discovery tools should also display the most frequent values of a specific column. As shown in Figure 5.2, Trillium TS Discovery displays the Values Distribution, which shows the top five values for the `Name` column. The names "Michelle," "Dorothy," "Joey," "Royson," and "Sunil" appear 28.571%, 14.286%, 14.286%, 14.286%, and 14.286% of the time, respectively.

[1] From the IBM Redbook *Metadata Management with IBM InfoSphere Information Server*, October 2011, Jackie Zhu et al.

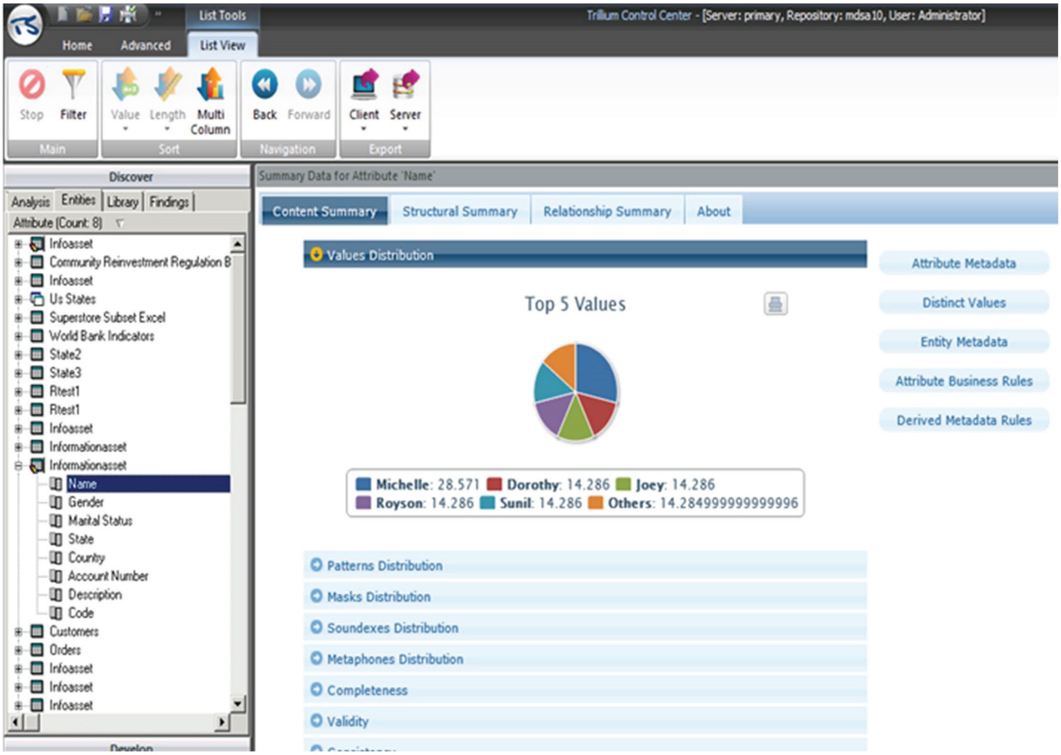


Figure 5.2: The Values Distribution for the Name column in Trillium TS Discovery.

Discover the Patterns Distribution of a Column

Data discovery tools should also display the patterns distribution of a specific column. As shown in Figure 5.3, Trillium TS Discovery displays the Patterns Distribution, which shows the top five patterns for the Name column. The most common patterns are alphanumeric six characters and alphanumeric eight characters, each with 28.571% of the records. These are followed by alphanumeric four, alphanumeric five, and alphanumeric seven, each with 14.286% of the records.

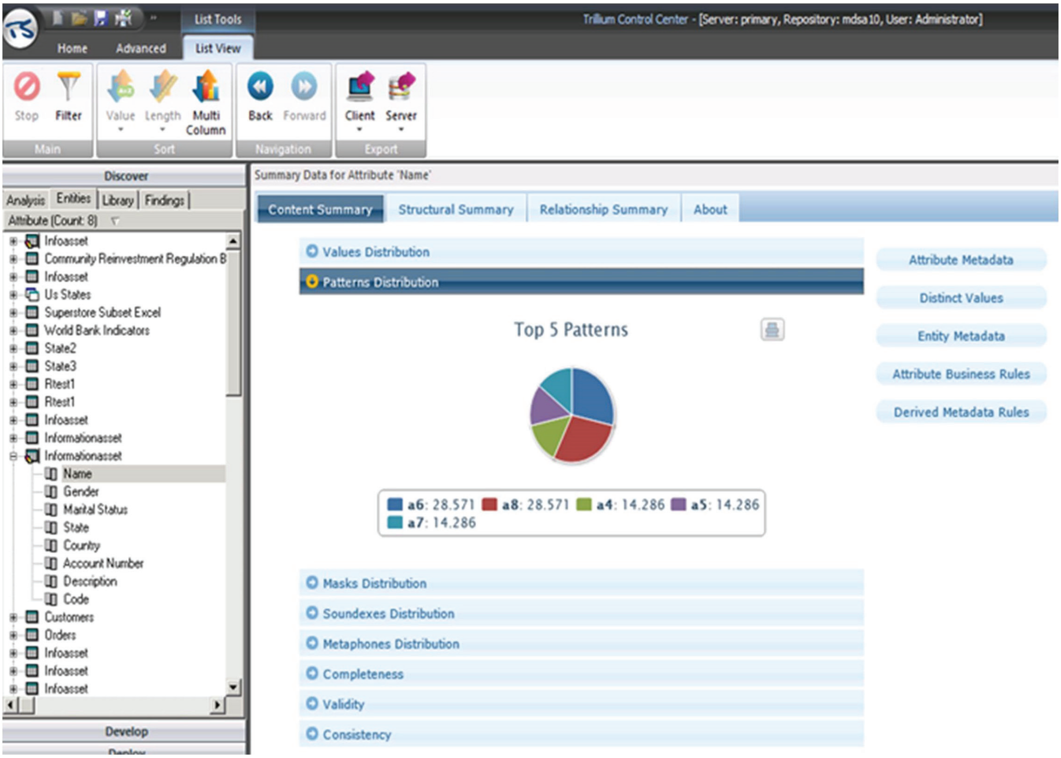
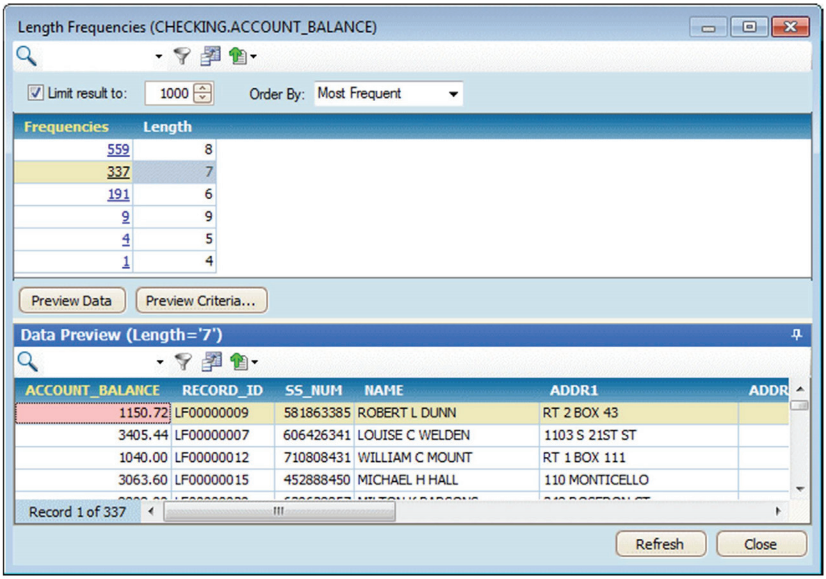


Figure 5.3: The Patterns Distribution for the Name column in Trillium TS Discovery.

Discover the Length Frequencies of a Column

Data discovery should also display the length frequencies of columns. In Figure 5.4, IBM InfoSphere Discovery displays the length frequencies for CHECKING.ACCOUNT_BALANCE. For example, values with a length of eight and seven occur 559 and 337 times, respectively. The bottom of the screen shows a preview of the rows where the length of CHECKING.ACCOUNT_BALANCE is seven.



Length Frequencies (CHECKING.ACCOUNT_BALANCE)

☑ Limit result to: 1000

Order By: Most Frequent

Frequencies	Length
559	8
337	7
191	6
9	9
4	5
1	4

Data Preview (Length='7')

ACCOUNT_BALANCE	RECORD_ID	SS_NUM	NAME	ADDR1	ADDR
1150.72	LF00000009	581863385	ROBERT L DUNN	RT 2 BOX 43	
3405.44	LF00000007	606426341	LOUISE C WELDEN	1103 S 21ST ST	
1040.00	LF00000012	710808431	WILLIAM C MOUNT	RT 1 BOX 111	
3063.60	LF00000015	452888450	MICHAEL H HALL	110 MONTICELLO	

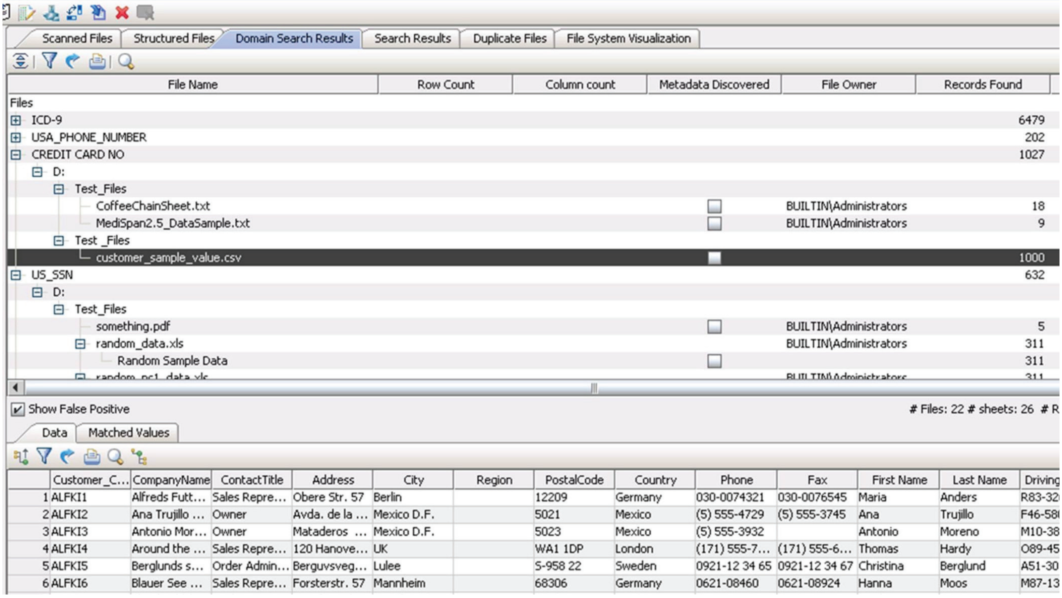
Record 1 of 337

Figure 5.4: Length frequencies in IBM InfoSphere Discovery.[2]

Discover Hidden Sensitive Data

Data discovery tools can also discover hidden sensitive data, which is a specific form of pattern matching. The sensitive nature of the data might not be reflected in column or table names. For example, U.S. Social Security numbers might be hidden in a field called EMP_NUM. Figure 5.5 shows that credit card numbers have been discovered by the Global IDs Profiler within the CoffeeChainSheet.txt, MedSpan2.5_DataSample.txt, and customer_sample_value.csv data sources.

[2] From the IBM Redbook *Metadata Management with IBM InfoSphere Information Server*, October 2011, Jackie Zhu et al.



File Name	Row Count	Column count	Metadata Discovered	File Owner	Records Found
Files					
⊞ ICD-9					6479
⊞ USA_PHONE_NUMBER					202
⊟ CREDIT CARD NO					1027
⊟ D:					
⊟ Test_Files					
CoffeeChainSheet.txt			☐	BUILTIN\Administrators	18
MediSpan2.5_DataSample.txt			☐	BUILTIN\Administrators	9
⊟ Test _Files					
customer_sample_value.csv			☐		1000
⊟ US_SSN					632
⊟ D:					
⊟ Test_Files					
something.pdf			☐	BUILTIN\Administrators	5
⊟ random_data.xls				BUILTIN\Administrators	311
Random Sample Data			☐		311
⊟ random_pc1_data.xls				BUILTIN\Administrators	311

☑ Show False Positive

Files: 22 # sheets: 26 # R

	Customer_C...	CompanyName	ContactTitle	Address	City	Region	PostalCode	Country	Phone	Fax	First Name	Last Name	Driving
1	ALFKI1	Alfreds Futt...	Sales Repre...	Obere Str. 57	Berlin		12209	Germany	030-0074321	030-0076545	Maria	Anders	R83-32
2	ALFKI2	Ana Trujillo ...	Owner	Avda. de la ...	Mexico D.F.		5021	Mexico	(5) 555-4729	(5) 555-3745	Ana	Trujillo	F46-58
3	ALFKI3	Antonio Mor...	Owner	Mataderos ...	Mexico D.F.		5023	Mexico	(5) 555-3932		Antonio	Moreno	M10-38
4	ALFKI4	Around the ...	Sales Repre...	120 Hanove...	UK		WA1 1DP	London	(171) 555-7...	(171) 555-6...	Thomas	Hardy	O89-45
5	ALFKI5	Berglunds s...	Order Admin...	Berguvsveg...	Lulee		S-958 22	Sweden	0921-12 34 65	0921-12 34 67	Christina	Berglund	A51-30
6	ALFKI6	Blauer See ...	Sales Repre...	Forsterstr. 57	Mannheim		68306	Germany	0621-08460	0621-08924	Hanna	Moos	M87-13

Figure 5.5: The Global IDs Profiler discovers credit card numbers within multiple data sources.

Discover Values with Similar Sounds in a Column

The data discovery tool should also discover column values with similar sounds. As shown in Figure 5.6, the Soundexes Distribution in Trillium TS Discovery shows the Name values with a similar sound that are grouped together as soundexes when the data is analyzed. The soundex is based on the first four values of every attribute. Trillium TS Discovery also checks for the Metaphone, which is based on the entire attribute value, helping to check for misspellings and data discrepancies.

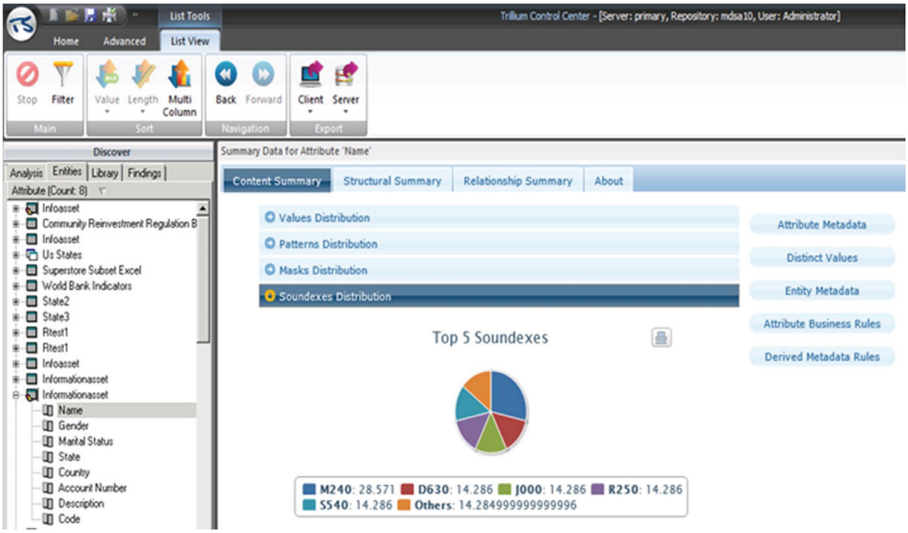


Figure 5.6: The Soundexes Distribution for the Name column in Trillium TS Discovery.

Agree on the Data Quality Dimensions for the Data Governance Program

According to DAMA UK in "The Six Primary Dimensions for Data Quality Assessment" (October 2013), a data quality dimension is a recognized term used by data management professionals to describe a characteristic, attribute, or facet of data that can be measured or assessed against defined standards in order to determine the quality of data. The next step is to select the data quality dimensions that need to be measured in the scorecard. Although there are no industry-standard definitions for data quality dimensions, Table 5.1 lists some data quality dimensions and the way they are used in business rules.

Table 5.1: Data Quality Dimensions

Data Quality Dimension	Definition	Sample Business Rule
1. Completeness	The degree to which data elements are populated	Customer phone number should not be null or blank.
2. Conformity	The degree to which data elements correspond to expected formats or valid values or ranges of values	State should be from the agreed upon list of code values for state. Phone number should be in the format NNN-NNN-NNNN. Further, "000-000-0000," "111-111-1111," and "999-999-9999" are not allowed.

Table 5.1: Data Quality Dimensions (continued)

Data Quality Dimension	Definition	Sample Business Rule
3. Consistency	The degree of relational integrity between data elements and other data elements	Minors should have guardians. The insurance policy expiration date should be greater than or equal to the policy effective date.
4. Synchronization	The degree to which data elements are consistent from one data store to the next	The transaction database should only contain orders for customer records that already exist in the master data hub. A student's date of birth has the same value and format in the school register as that stored within the student database.[3]
5. Uniqueness	The degree to which data elements are unique within a data store	U.S. Social Security numbers should not be repeated in the customer database. Employee IDs should not be repeated in the employee database.
6. Timeliness	The degree to which data is available on a timely basis	Emergency contacts should be entered into the system within two days of being provided by the employee. For example, Tina Jones provides details of an updated emergency contact number on June 1, 2013, which is then entered into the student database by the administration team on June 4, 2013. This indicates a delay of three days.[3]
7. Accuracy	The degree to which data elements are accurate	Email addresses should be validated by customer service every six months. Mail items should not be returned by the United States Postal Service as undeliverable.

Develop Business Rules Relating to the Data Quality Dimensions

The data governance team should document business rules in the business glossary. These business rules should relate to only one data quality dimension. This is important because

[3] From the IBM Redbook *Metadata Management with IBM InfoSphere Information Server*, October 2011, Jackie Zhu et al.