Metadata and Data Harmonization

Daniel W. Gillman
Frank Farance
US Bureau of Labor Statistics[1]
Gillman.daniel@bls.gov
Farance Inc.
frank@farance.com

Introduction

Users of US federal statistical data often want to combine data from multiple sources to create new data sets. This need is independent of the fact that the US statistical system is divided among many different agencies. Rather, it is due to the inability of the agencies to produce data sets to match every need. There are just too many possible questions users want to answer, and resources across the agencies are limited. Instead, agencies produce data within the scopes of their subject matter areas, and users are required to find relevant data sets and combine them.

Finding relevant data sets can be very hard. Since the US federal statistical system is Balkanized, knowledge of which agency publishes data covering what subject matter over some area for a given time period is not widely known. The need to provide answers to these questions is why the study and implementation of metadata systems is currently a hot topic. However, for this paper, we will assume the user can find the relevant data. We are interested in exploring what is needed to combine data from multiple sources – i.e., data harmonization.

There are at least 2 important inter-agency efforts dedicated to helping users combine data from multiple sources. **Fedstats**[2] web site describes that it

> has been available to the public since **1997**, provides access to the full range of official statistical information produced by the Federal Government without having to know in advance which Federal agency produces which particular statistic. With convenient searching and linking capablilties to more than 100 agencies that provide data and trend information on such topics as economic and population trends, crime, education, health care, aviation safety, energy use, farm production and more, FedStats is your one location for access to the full breadth of Federal statistical information.

Now, the Administration started a new project called **Data.Gov**[3], and the Data.Gov web site describes it as a means

> to increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government. Although the initial launch of Data.gov provides a limited portion of the rich variety of Federal datasets presently available, we invite you to actively **participate** in shaping the future of Data.gov by suggesting additional datasets and site enhancements to provide seamless access and use of your Federal data. Visit today with us, but come back often. With your help, Data.gov will continue to grow and change in the weeks, months, and years ahead.

The aims of both of these projects are to provide a wide variety of data to the public without the need to know which agencies produced particular data sets. Moreover, and most relevant to this paper, the possibility of precisely identifying the most relevant data and providing the tools to combine or visualize them is animating new work. The **DataWeb**[4] system at the Census Bureau is designed to enable the user to combine data sets. Other tools are under development, spurred by the possibilities of Data.Gov.

[1] The opinions expressed in the paper are due to the authors and do not necessarily reflect the policies of the US Bureau of Labor Statistics

[2] www.fedstats.gov

[3] www.data.gov

[4] dataferrett.census.gov

The aim of this paper is to provide a framework for harmonizing data. It is based on work by the authors (Gillman and Farance, 2009) employing ideas from the theory of terminology of special languages (ISO, 2000) to defining and understanding data. The main emphasis of the paper is to look at the concepts underlying a datum. The paper begins with an overview of the theory of terminology for special languages. We then use this to define a datum, define metadata, and describe data harmonization. There are many ways to factor the concepts underlying data, so the need for an agreement as to how this should be done across the metadata describing data sets is a prerequisite. Laying out the concepts through a derivation chain then lets the user understand where and how data may be harmonized. We illustrate this with an example.

Overview of Terminology Theory

Properties and characteristics

A **property** is the result of a determination either directly or indirectly about some object. One form of determination is through observation – something humans perceive through their senses. Noticing the color of a person's eyes is an observation or direct determination of the eye color of that person. Another form of determination is through detection by an instrument. An oral thermometer is an instrument that detects internal body temperature of a person. Observing a reading on the thermometer is an indirect determination about the internal temperature of a person. The specific observed eye color and internal body temperature are properties of a person.

It is through properties that we are able to make distinctions between objects. For instance, one person may be 185 cm tall, have brown colored eyes and hair, and have medium brown colored skin. Another may be 170 cm tall, have blue colored eyes and blond hair, and have very light brown colored skin. The properties of each person serve to help distinguish between the two.

Since the examples above use perceivable objects, it is important to note that conceivable objects have properties, too. For instance, consider the rational numbers "three and fourteen hundredths" and "negative seventeen". In the same way as with perceivable objects, properties of conceivable objects are the results of determinations about these objects. Here, the "sign" of the numbers is a property of them. The "sign" of 3.14 is positive, and the "sign" of -17 is negative.

A **characteristic** is a determinable. A determinable is something capable of being determined, definitely ascertained, or decided upon. Eye color, for instance, is a determinable. It is capable of being ascertained by looking into a person's eyes to determine their color. A property, on the other hand, is what gets determined. This is called a determinant. A determinant is an element that determines or identifies the nature of something. Blue is a determinant for eye color. So, a characteristic has the capacity for being determined (determinable), whereas the property is the result of a determination (determinant). Some characteristics of a person are height, eye color, hair color, and skin tone. Examples of corresponding properties, taken from the paragraph above in clause 4.3, are: height has the properties 185 cm and 170 cm; eye color has the properties brown and blue; hair color has the properties brown and blond; and skin tone has the properties medium brown and very light brown.
A set of properties corresponds to a characteristic. In examples 5 and 6 in clause 6.2, different sets of properties may correspond to the same characteristic, depending on needs. In addition, the same property may correspond to two characteristics. The following example illustrates this.

EXAMPLE 1: A property may correspond to two characteristics. Consider the following characteristics: height (of a person) and length of the diagonal (of a television screen). The property 60 inches (5 feet or about 152 cm) corresponds to both characteristics. Some people are 60 inches tall and some large widescreen television sets measure 60 inches diagonally across the screen.

The set of properties for a given characteristic may not always be the same. Consider marital status as a characteristic. Then, the set of properties might be single (not married) and married, or the set of properties could be single (never married), married, divorced, and widowed.

Concepts

A **concept** is a unit of thought differentiated by a set of characteristics. Consider the concept "person". The characteristics of a person include being designed to stand upright on 2 legs, ability to talk, age, marital status, and skin tone. There are many others.

Some characteristics are indispensable for understanding a concept. These are the **essential characteristics**. A **delimiting characteristic** is a characteristic used to distinguish it from a generic concept. For example, an essential characteristic of people is they are designed to stand and walk upright. This is also a delimiting characteristic since it distinguishes people from gorillas. The **intension** of a concept is the set of all characteristics which may differentiate the concept from others. The **extension** of a concept is the totality of objects to which a concept corresponds.

A **defining characteristic** is a characteristic which is representative of objects in the extension of a concept. A defining characteristic of people is that they stand and walk upright. Not every person is capable of walking and standing upright, even though they are designed that way. Paralyzed or injured people may not be able to stand.

Characteristics and properties are concepts in their own right. As concepts, each kind plays a role, and this is how the ideas are distinguished.

Example 2 illustrates the importance of establishing essential characteristics for a concept. In particular, the addition of a single characteristic may have profound influences on the objects in the extension of the concept. Adding or removing characteristics affects the meaning of a given concept, changing the concept itself. Thus, the extension would be expected to change.

EXAMPLE 2: The concept of planet was revised in 2006 by the International Astronomical Union. This revision resulted in the elimination of Pluto as one of the planets in the solar system. Pluto was long considered the ninth planet in the solar system, but some astronomers questioned this classification. Several properties Pluto possesses differ markedly from those of the other planets. Additionally, recent advances in astronomy - much better telescopes and vastly improved computation - showed there are many more celestial bodies that could be considered planets if Pluto remained one. Therefore, a concerted effort was made to define "planet" in a useful way.

The concept of a planet is now defined by these four essential characteristics: A planet is a celestial body that

1. Is in orbit around a star
2. Contains sufficient mass to maintain a nearly spherical shape due to its own gravity
3. Is not massive enough to cause thermonuclear fusion in its core
4. Has "cleared the neighborhood", i.e., become gravitationally dominant, so the only other bodies in its vicinity are its satellites

This fourth characteristic is what eliminated Pluto.

A **general concept** is a concept which corresponds to two or more objects which form a group by reason of common properties. An example is the concept "planets in our solar system". An **individual concept** is a concept which corresponds to exactly one object. An example is the concept "Saturn". In other words, a general concept may have more than one object in its extension, and an individual concept must have exactly one object in its extension. Note, a concept might be so defined that there exists only one object in its extension even though the possibility for more exists. This is still a general concept. For example, the notion "all planets with one moon" is a general concept. There is one known planet with one moon – Earth – but the possibility there are more cannot be ruled out.

Signifiers, labels, and designations

A **signifier** is a concept whose extension contains only perceivable objects. An object in the extension of a signifier is a **token**. For instance, the objects **5** and **5** are both tokens of "the numeral five", a signifier.

A signifier has the potential to refer to an object. In this case, the referring signifier is a label. If that object is a concept, then the referring signifier is a designation. A **label** is a representation of an object by a signifier which denotes it. For instance, the token "ISO/IEC 11179-4" is a label for this International Standard. A **designation** is a representation of a concept by a signifier which denotes it. For instance, the token "apple" is an English word designating the concept "the fleshy usually rounded red, yellow, or green edible pome fruit of a usually cultivated tree (genus *Malus*) of the rose family" [5]. The token "M" might designate that a person is married, as recorded in some database. A non-empty set of designations is called a **term set** (or designation set).

Definitions

A **definition** is representation of a concept by a descriptive statement which serves to differentiate it from related concepts. There are 2 kinds of definitions. An **intensional definition** is a definition which describes the intension of a concept by stating the superordinate concept and the delimiting characteristics. The definition of delimiting characteristic in clause 3.6 is an example of an intensional definition. An **extensional definition** is a description of a concept by enumerating all of its subordinate concepts under one criterion of subdivision. The definition of relation in clause 3.23 is an example of an extensional definition.

NOTE: Both kinds of definitions usually depend on knowing the definitions of other concepts in order to fully understand the concept under study.

Data as Terminology

Values

A fundamental requirement for data is that they can be copied. In Information Technology, the need for copying happens all the time in data processing. The only way to know a datum has been faithfully copied is to compare the copy to the original. The comparison determines whether equality is satisfied (and the copy is faithful).

A concept may have an equality notion defined for it. Usually, the same equality notion is defined for a set of values. This means that if two people say they have the same concept, a determination of equality between them can be made. This operation may be different depending on the situation. In fact, more than one measure of equality can be defined for any given concept. See Examples 3 and 4. A **value** is a concept with a notion of equality defined.

EXAMPLE 3: Consider the natural number "seventeen". It is a concept, and its extension is all situations of 17 objects. Equality is defined as it is commonly understood for natural numbers. Another way to define equality for natural numbers, including "seventeen", is to ask if the number is even or odd. In this situation, all odd numbers are equal, and all even numbers are equal.

EXAMPLE 4: M for married, as in some particular person is a married. Married is a value, since marriage is a social and legal status controlled by the state. Equality may be determined by referencing the meaning in common law.

Datum

A **datum** is a designation of a value in some context (Farance and Gillman, 2006). In context, a datum is a designation of a value representing a class in a partition[6] of the extension of a concept, where the partition of the extension is defined for some characteristic of that concept. Each class of the partition is a value corresponding to a property of the characteristic. Each class is a concept (a value), so the signifier representing the class is a designation. In the finite case, usually called categorical data, the partition is often called a classification, e.g., marital status categories.

[5] Taken from Merriam-Webster on-line dictionary at http://www.merriam-webster.com/

[6] A partition is a non-empty set of mutually exclusive and exhaustive subsets of some other set. The number of subsets is not necessarily finite.

Sometimes when using the idea of values, people refer to a signifier as the value, but this is incorrect: the value refers to the concept portion of the datum and not its signifier. For example, one can easily discuss the notion of the value of seventeenness, a *number* (a concept), independent of any particular signifier, e.g., a *numeral* (a kind of signifier that designates numbers).

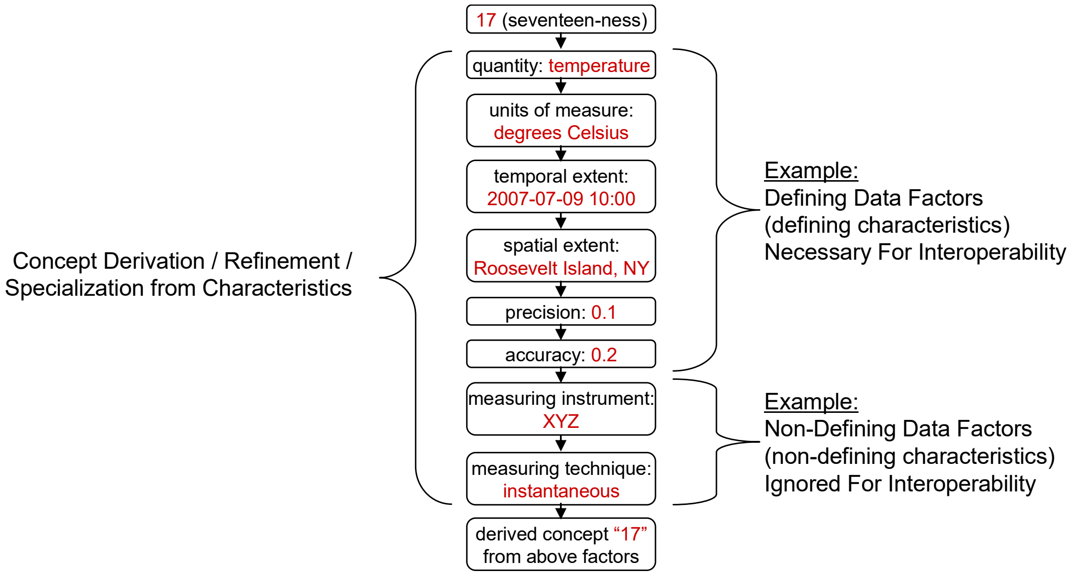
The following 3 examples show simple descriptions of data in context:

EXAMPLE 5: Example of a partition of people based on marital status
Concept = people of the UK
Characteristic = marital status
Partition = {single, married, divorced, widowed}, where "single" means never married and the rest correspond to their usual meanings. The signs S, M, D, and W designate these concepts, respectively.

EXAMPLE 6: Second example of a partition of people based on marital status
Concept = people of the UK
Characteristic = marital status
Partition = {single, married}, where "single" means not married and married takes its usual meaning. The signs S and M designate these concepts, respectively. The purpose of the example is to show that more than one partition may apply to a characteristic of a concept.

EXAMPLE 7: Example of a partition of gambling casino games based on probability of winning
Concept = gambling casino games
Characteristic = probability of winning
Partition = {x | 0 < x ≤1} (the set of all numbers, x, such that x is greater then zero and less than or equal to one), where x is a probability. The signs are the numeric strings that designate the numbers, to some agreed upon precision, fixing the lengths of the strings.

The following diagram shows a stack of concepts that comprise a composite concept for a hypothetical datum:


17 (seventeen-ness)
quantity: temperature
units of measure: degrees Celsius
temporal extent: 2007-07-09 10:00
spatial extent: Roosevelt Island, NY
precision: 0.1
accuracy: 0.2
measuring instrument: XYZ
measuring technique: instantaneous
derived concept "17" from above factors
Concept Derivation / Refinement / Specialization from Characteristics
Example:
Defining Data Factors
(defining characteristics)
Necessary For Interoperability
Example:
Non-Defining Data Factors
(non-defining characteristics)
Ignored For Interoperability

In words, this datum is “the temperature on Roosevelt Island at 10:00am on 9 July 2007 as measured by instrument XYZ with precision 0.1 and accuracy 0.2 is 17 degrees Celsius”. The features in red are particular to this datum "17". The generic concept of the number 17 (the top concept defined by the characteristic of "seventeen-ness") is specialized by additional concepts that produce the derived concept of "17" (the bottom concept). This is an example of a derivation chain.

Harmonization

Metadata

Data used to describe some object is called **metadata**. The loose definition “data about data” is inadequate, because descriptive data for any kind of object has the same characteristics as data about data. Also, the use of the word “about” is imprecise, so the general definition is more useful.

The essential characteristics of metadata include the following:

- They are data
- They describe some object

For example, if ***P*** is data and if the symbol ***P*→*Q*** represents the descriptive relationship meaning that ***P*** describes ***Q***, then ***P*** is metadata for ***Q***. If there is no relationship from ***P*** to ***Q***, then ***P*** is no longer metadata, i.e., ***P*** is merely data, because being metadata requires that data be in a descriptive relationship to some object. Stated differently, ***P*** becomes metadata once its descriptive relationship to ***Q*** is established. It is impossible to determine, independent of context and relationships, that any data is also metadata.

Attributes

Objects are described through the properties they have. Each property of the object is the determinant for some characteristic, and no two properties of the object are associated with the same characteristic.

Any concept can be **reified** – turned into data – by writing down a definition, description, or even a designation of it. Since properties are concepts and they describe objects, then reified properties are metadata. An **attribute** is a concept in the role of a characteristic whose properties are reified.

EXAMPLE 8: This paper you are reading is an object. One characteristic of the paper is its author(s). In this case, Gillman and Farance are the authors. So, the text “Gillman and Farance” is the data that reifies the property of the characteristic author(s).

Factoring

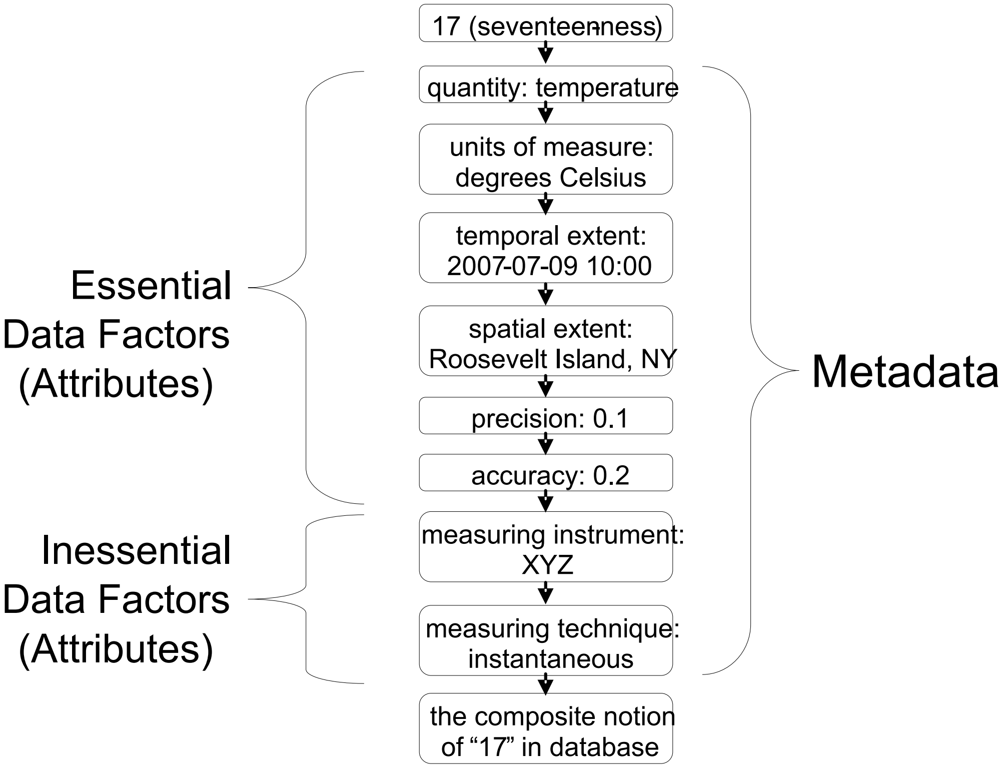
Given the myriad attributes that may be used to describe some data, choices about which to use have to be made. The process of identifying the attributes is called **factoring**. Derivation chains based on different factorings will be different and difficult to compare.

If we think of a characteristic as an attribute before its properties are identified and reified, then a metadata model is a set of characteristics and its defined properties. That is, if we know the properties of a characteristic will be reified, and that data used to describe the objects are those properties, then the characteristics and properties are a template, or metadata model, for metadata.

Harmonization

Semantic interoperability is achieved when an understanding of a datum in one system can be translated into the same meaning in another system. A sufficient condition for this is when two systems share the metadata model. Then, the data in one system, which is the reification of attributes of objects, can be translated into the reification of the same attributes of those objects in the other system. This preserves meaning.

EXAMPLE 9: This is an example of factoring and interoperability:
Step #1: Factor the data description according to terminological principles. Express the description as a reification of identified attributes. For the datum described as “On 2007-07-09 at 10:00 on Roosevelt Island, NY, USA, the temperature, in Celsius, measured by thermometer XYZ, to 0.1 degrees precision, with accuracy 0.2, is: 17”, the attributes shown below are identified.
Step #2: Express the attributes in the list as shown.
Step #3: Determine which attributes are essential for comparison with data from another source. Harmonization is dependent upon these essential characteristics. In this case, the following attributes are essential: temporal extent, spatial extent, precision, accuracy, units of measure, and quantity.


17 (seventeenness)
quantity: temperature
units of measure: degrees Celsius
temporal extent: 2007-07-09 10:00
spatial extent: Roosevelt Island, NY
precision: 0.1
accuracy: 0.2
measuring instrument: XYZ
measuring technique: instantaneous
the composite notion of “17” in database
Essential Data Factors (Attributes)
Inessential Data Factors (Attributes)
Metadata

To make this work, attributes may be structured linearly, hierarchically, or in some other grouping. The choice is one that allows the user to make comparisons easily. By understanding the meaning of data (collectively or in individually), there can be agreement upon harmonization.

Disagreement
What has not been discussed so far is what happens when the factors are the same but the attributes don’t have the same data. There are 3 possibilities:

1 The 2 concept sets are generalized to a concept set that subsumes both
2 The 2 concept sets are specialized to a concept set that both subsume
3 A new concept set is produced that both are associated with

An example of the first kind is the following:
Take the case of marital status, where one set of properties is single (not married), married – living together, married – not living together and another set of properties is single (never married), married, widowed, divorced. Then a harmonized set would be the generalization to single (not married), married. Here, married – not living together includes legally separated, and widowed and divorced are included as not married.

An example of the second kind is the following:
Take 2 data sets in which one set contains data on adults (18 and older) and the other on females. Then a harmonized set will be on females 18 and older.

An example of the third kind is as follows:
One set is data on people who are employed, and the other is data on employers. Then a new set contains data on the employer – employee relationship (maybe called employments).

Further research will is required to make this more precise.

Conclusion

This paper contains a framework for harmonizing data, typically from different sources. We structured this using the theory of terminology for special languages. After describing the basic theory, we showed that data are terminological constructs, and metadata are just data used to describe some object. From this descriptive relation, we derived the idea of an attribute, which is the unit in which a description is factored. By illustrating the factorization in a derivation chain, it is easy to compare descriptions across sources.

A sufficient condition for harmonization is based on maintaining the same metadata model for each data set. The model provides the factors used in describing a datum, and the same model across systems allows for easy comparisons. Three cases of disagreements between attributes were described.

References

Farance, F. and Gillman, D. (2006). The Nature of Data. Working Paper #12 in *Proceedings of the UNECE Workshop on Statistical Metadata*, Geneva, Switzerland

Gillman, D. and Farance, F. (2009). Data and Metadata from the Terminological Perspective. In *Proceedings of the Joint Statistical Meetings 2009*, Washington, DC, USA

ISO (2000). ISO 704: *Principles for terminology*. Geneva: International Organization for Standardization