Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

PROCEEDINGS OF THE 6th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP

Framework for data element standardization

Dagobert Soergel
College of Library and Information Services
University of Maryland, College Park, MD 20742

This paper describes a method for the definition of data elements in CASE (Computer-Assisted Systems/Software Engineering) systems, data element dictionaries, and data element repositories. The proposed method derives its power from its simplicity and its use of classification and knowledge representation techniques. It is based on the entity-relationship approach. A data element definition consists of a frame which gives the underlying relationship type, the focal entity type (about which the data element gives data), the related entity type (the nature of the data values), and a rule for determining the data value. These frames form a hierarchy which is based on a hierarchy of entity types (or ontology) and a hierarchy of relationship types.

This paper describes a method or data model for the definition of data elements in CASE (Computer-Assisted Systems/Software Engineering) systems, data element dictionaries, and data element repositories. Data element definition, an essential part of system documentation, is important for many purposes:

- In system development: Saving work through reusing existing definitions; assuring data compatibility with existing systems where appropriate and feasible; serving as external memory and communication device for the development team.

- In data interchange: Making sure that data from system A are indeed the data needed in system B and supporting conversion where necessary.

- In system documentation for users: Finding systems that contain the data of interest; showing the user clearly the nature of the data in a system.

Data element definitions must facilitate retrieval of data elements and must make relationships between data elements explicit and clearly visible. They must show the scope of a data element; for example, there is a data element *height* that applies to any physical object, and a more specialized data element *height of horse*.

The proposed method for data element definition fulfills these requirements; it derives its power from its simplicity and its use of classification and knowledge representation techniques. It is based on the entity-relationship approach and owes much to the way data are defined in relational databases. A data element makes an assertion about a focal entity (or object) as being in a relationship with some other entity, possibly in the framework of a relationship with multiple arguments (or slots). Put differently, a data element gives the value of a property of the focal entity. Two examples of **data element definition frames** illustrate the proposed approach by way of introduction:

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

DE2 **Average height of a horse breed**

Relationship:	Entity/1 **has-height** length-measure/2
Focal entity:	Horse (taxon)/1
Representation of values:	Taxonomy of the Amer. Horse Breeding Soc.
Restrictions:	Horse taxon is breed
Related entity:	Length-measure/2
Nature of value:	Average for taxon
Representation of values:	Hands
Restrictions:	3 - 9 hands
Rule:	With horse standing, measure distance from ground to withers (shoulders); average for suitable sample

Note: /1, /2, etc. denotes the argument (slot) number in the underlying relationship. Usually the focal entity is /1 (for an example where it is /2 see DE8 in the appendix).

DE7 **Course offering attended by student**

Relationship:	Person/1 **attends** course-offering/2 **with** grade/3
Focal entity:	Person (individual)/1
Representation of values:	Social Security Number
Restrictions:	Enrolled at educational institution X
Related entity:	Course offering (individual)/2
Nature of value:	Individual measure
Representation of values:	Course number + section number + semester
Restrictions:	Course offering scheduled by educ. institution X
Rule:	Course offering value from registration procedure.

These examples show the essential components of a data element definition:

the underlying relationship type;

the type of the focal entity (the entity about which a statement is made) and its argument/slot number in the relationship;

the type of the related entity and its argument/slot number in the relationship.

There are other important components including but not limited to: Data element name, narrative description, single vs multiple values, required or not, administrative matters (who defined the data element when, where is it used, who is authorized to change the definition), but we will focus strictly on the essential conceptual elements.

A framework for data element definition, then, consists of

1 A classification (taxonomy) of entity types / object classes

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

2 A classification (taxonomy) of relationship types

3 A classification of data elements, organized in a frame hierarchy

We will discuss each of these in turn.

1 A classification (taxonomy) of entity types / object classes (ontology)

Entity types are used to indicate the scope of the focal entity whose property is at issue and to indicate the type or range of values the related entity (or property of the focal entity) can take. The box shows some excerpts from a classification of entity types.

Tangible object
- Natural object
  - Living organism
    - Animal
      - equus
        - equus caballus
          - Horse breed X
    - Person
- Man-made object
  - Automobile
    - Honda Civic Wagon
      - Honda Civic Wagon 1989

Legal-entity-or-system (persons, organizations, and systems that can act in a similar fashion)

Date

Color

Length-measure

Course-offering

Grade

Two points need further clarification. The first is the distinction between individuals, such as an individual horse, and taxa, such as *animal* or the genus *equus*. We will use the notation *animal (individual)* to define a data element that pertains to any individual entity that is an animal and *animal (taxon)* to define a data element that pertains to any taxon that falls under animal in the classification. Thus we make one classification do double duty. Note the difference between making a statement (through a defined data element) about an individual horse and making a statement about the taxon *equus caballus*.

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

The next point is the distinction (if any) between **entity type** and **object class**. An individual horse belongs to (is a member of) the entity type / object class *equus caballus* and also belongs to (is a value of) the entity type / object class *living organism*. One might want to fix (somewhat arbitrarily) some level of generality for entity type; for example, one might admit *living organism* as an entity type and consider everything below an object class. A precise distinction between entity type and object class is not needed for this framework.

For each entity type, the ontology must specify the various ways of identifying and representing values. This includes:

- Indicating units of measure (each identified by a code) for quantitative (interval or ratio scale) variables, and conversion rules from one unit to another.

- Indicating the several formats (each identified by a code) in which values such a date or a personal name could be written, with transformation rules from one to the other.

- Referring to one or more lists of values (represented through terms or codes) for the entity type (or members of the object class), such as taxonomies of living organisms or color classifications. One list could be designated as the preferred standard for that entity type. Database systems should use the values from one of these lists or link their values to one of the lists. When a term or code is used, it should always be qualified by indicating the list it came from (adapting the system used in biology, where the name of a taxon is always qualified by the author who assigned the name). A code should be assigned to each source for this purpose. The system should allow for, but not require, establishing concordances from one list to another.

Any ontology (classification of entity types) could be used for data element definition. Many such ontologies have been developed by philosophers, by researchers in artificial intelligence who need an ontology for structuring knowledge bases and rules for knowledge processing, and by linguists who need to categorize terms (or rather the designated concepts) as a basis for formulating rules for natural language processing; thus ontologies are often developed in conjunction with large dictionary projects. However, broad use of data element definitions would be facilitated by agreement on a common ontology that may need to be developed to greater levels of detail in specialized user communities. Developing such a common ontology (on the basis of the widely overlapping ontologies that exist) and extending it through specialization as needed presents a challenge for classification research.

2 A classification (taxonomy) of relationship types

A relationship type is used to indicate the nature of the data given by the data element, i.e., the mature of the relationship between the focal entity and the related entity. Some examples from a classification of relationship types.

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

entity/1 **has-dimension** length-measure/2

entity/1 **has-height** length-measure/2

entity/1 **has-width** length-measure/2

entity/1 **has-turn-radius** length-measure/2

entity/1 **has-color** color/2

entity/1 **made-on** date/2

entity/1 **made-by** legal-entity-or-system/2

entity/1 **authored-by** legal-entity-or-system/2

entity/1 **edited-by** legal-entity-or-system/2

person/1 **attends** course-offering/2 **with** grade/3

The underlying relationship type represents the core meaning of a data element. A classification of relationship types would thus provide semantic organization to the vast number of data elements and bring related data elements together. Developing such a classification presents an even greater challenge for classification research since there are only few and limited classifications to start from.

3 A classification of data elements, organized in a frame hierarchy with hierarchical inheritance.

As the introductory examples and the examples in the appendix show, we propose to represent each data element through a frame with the slots shown in outline form in the box below and with explanations in the next box.

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

Relationship type

Focal entity type / object class

- Further restrictions on permissible entity values
- Representation chosen for entity values

Related entity type / object class

- Nature of value
- Further restrictions on permissible entity values
- Representation chosen for entity values

Rule

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

Relationship type

The relationship type underlying the data element, a binary or n-ary relationship.

Focal entity type / object class

This slot defines the scope of the data element. For economy of representation, the argument number of this entity type in the relationship, really a separate slot in the data element frame, is shown after /.

Further restrictions on permissible entity values

Further limitation of the scope, beyond that implied by the focal entity type.

Representation chosen for entity values

How are the entity values expressed? For example, if the entity type is *living organism (taxon)*, from which taxonomy of living organisms are the values drawn? If the values are not drawn from any of the lists given in the entity type / object class taxonomy, enumerate the list of values used. Relate each value to the corresponding value from a standard list. If the entity type is *equus caballus (individual)*, how are the individual horses identified? If the entity type is *length-measure*, what unit of measure is used?

Related entity type / object class

This slot defines the range of possible values of the related entity, i.e., of the property of the focal entity represented by this data element. Again, the argument number in the underlying relationship is shown after /.

Nature of value

This includes indications such as individual measurement, design specification, average value for all the members of a taxon, and others. Individual measurement is understood very broadly; the color of an individual car would be an individual measurement, so would be the course offering an individual person is attending.

Further restrictions on permissible entity values (as above)

Representation chosen for entity values (as above)

Rule

A rule that defines precisely how the value of the related entity type is determined, for example a rule for measurement or a formula for computing the value from other data elements.

The frames form a hierarchy of data element definitions which results form regular frame hierarchy rules (such as specializing a frame that does not give a unit of measurement by filling

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

that slot) and from the hierarchies of entity types and relationship types. A general data element definition, for example a data element "average height" or "average height of mammal taxon" could be used for horse breeds if all the slot fillers in the general definition were valid for horses. But horse height is measured in hands, not meters, and a different rule on how to measure height applies. The data element definition must be modified accordingly. Still, this is less work than defining a data element from scratch. Also, the differences between the general data element "average height of mammal taxon" and the more specific data element "average height of horse breed" are made explicit.

The examples do not include slots for broader and narrower frames; the relationships in the frame hierarchy can be inferred from the values of the slots given. In a full-featured frame hierarchy, one would introduce such slots and omit from a narrow frame the slots that can be inferred through hierarchical inheritance from the frame above.

Creation of these frames is easy. Retrieval involves simply inserting known slot fillers into a frame template. Retrieval based on the frame hierarchy returns the corresponding data element or the nearest broader data element which can be used as is or modified.

Acknowledgments

The ideas presented in this paper arose from and were further refined by discussions with ANSI committee X3L8 in an attempt to devise a method of data element definition that provides a simpler and more powerful alternative to the scheme considered by the committee. The committee materials and discussions posed the problem and contributed instructive examples.

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

Appendix. Examples of data element definitions

DE 1 **Height of any tangible object**	
Relationship:	Entity/1 **has-height** length-measure/2
Focal entity:	Tangible object (individual)/1
Restrictions:	No further restrictions
Representation of values:	No list given
Related entity:	Length-measure (individual)/2
Nature of value:	Individual measure
Restrictions:	No further restrictions
Representation of values:	Any unit, default meters
Rule:	Place object in its normal orientation with respect to the ground, measure distance from ground to the object point farthest away from the ground.

Note: /1, /2, etc. denotes the argument (slot) number in the underlying relationship. Usually the focal entity is /1 (for an example where it is /2 see DE8).

DE2 **Height of a horse**	
Relationship:	Entity/1 **has-height** length-measure/2
Focal entity:	equus caballus (individual)
Argument number:	1
Restrictions:	No further restrictions
Representation of values	Breeder ID + number given by breeder
Related entity:	Length-measure (individual)
Argument number:	2
Nature of value:	Individual measure
Restrictions:	3 hands to 9 hands (the outside limits) (Note: If the breed of each horse is given in the database, and if the database also gives for each breed a range of heights, than the restriction here can be specified as the range for the breed to which the horse belongs)
Representation of values:	Hands
Rule:	With horse standing, measure distance from ground to withers (shoulders)

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

DE3 **Turn radius of an automobile make and model**	
Relationship:	Entity/1 **has-turn-radius** length-measure/2
Focal entity:	Automobile (taxon)/1
Restrictions:	Automobile taxon is make and model
Representation of values	X catalog of all automobiles produced in the US
Related entity:	Length-measure (individual)/2
Nature of value:	Design specification
Restrictions:	5m - 15m
Representation of values:	meters
Rule:	(Rule how to measure the turn radius)

DE4	**Color list of an automobile make and model**
Relationship:	entity/1 has-color-list color-list/2
Focal entity:	Automobile (taxon)/1
Restrictions:	Automobile taxon is make and model
Representation of values:	X catalog of all automobiles produced in the US
Related entity:	color-list (individual)/2 (An enumerated list of the colors in which the automobiles of the make and model come, using the manufacturer's color names, preferably relating them to the corresponding names from the preferred color list)
Nature of value:	Design specification
Restrictions:	
Representation of values:	Some identifying number for color lists
Rule:	(A rule on how to find the manufacturer's colors)

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

DE5 **Make and model of an automobile**	
Relationship:	Entity/1 **belongs-to** entity/2
Focal entity:	Automobile (individual)/1
Restrictions:	No further restrictions
Representation of values	Manufacturer and serial number
Related entity:	Automobile (taxon)/2
Nature of value:	Individual measure
Restrictions:	Automobile taxon is make and model
Representation of values	X catalog of all automobiles produced in the US
Rule:	(No rule needed)

DE6 **Color of an automobile**	
Relationship:	entity/1 **has-color** color/2
Focal entity:	Automobile (individual)/1
Restrictions:	No further restrictions
Representation of values	Manufacturer and serial number
Related entity:	Color (individual)/2
Nature of value:	Individual measure
Restrictions:	Color values restricted to those found in the list given for the automobile make and model to which the automobile belongs, unless repainted
Representation of values:	Color name taken form color list valid for the automobile make and model
Rule:	Determine color from visual inspection. (Another rule may be: Copy color from manufacturer's shipping record)

ISSN: 2324-9773

Soergel, D. (1995). Framework for data element standardization. 6th ASIS SIG/CR Classification Research Workshop, 133-144. doi:10.7152/acro.v6i1.12666

DE7 **Course offering attended by student** (on student's transcript

Relationship:	Person/1 **attends** course-offering/2 **with** grade/3
Focal entity:	Person (individual)/1
Restrictions:	Enrolled at educational institution X
Representation of values:	Social Security Number
Related entity:	Course offering (individual)/2
Nature of value:	Individual measure
Restrictions:	Course offering scheduled by educational institution X
Representation of values:	Course number + section number + semester
Rule:	Value determined through registration procedure.

DE8 **Student attending course offering** (for class list)

Relationship:	Person/1 **attends** course-offering/2 **with** grade/3
Focal entity:	Course offering (individual)/2
Restrictions:	Course offering scheduled by educational institution X
Representation of values:	Course number + section number + semester
Related entity:	Person (individual)/1
Nature of value:	Individual measure
Restrictions:	Enrolled at educational institution X
Representation of values:	Social Security Number
Rule:	Value determined through registration procedure.

DE9 **Grade received by student in course offering**

Relationship:	Person **attends** course-offering **with** grade
Focal entity:	Person (individual)/1
Restrictions:	Enrolled at educational institution X
Representation of values:	Social Security Number
Related entity:	Grade/3
Nature of value:	Individual measure
Restrictions:	Grade value allowed by educational institution X
Representation of values:	Letter with + or -
Rule:	Grade value is submitted by instructor

ISSN: 2324-9773