caCORE Training Workbook

Course 1010: caDSR Implementation of the ISO/IEC 11179 Metadata Standard

caCORE Training Website http://ncicb.nci.nih.gov/NCICB/training/cadsr_training
Help & Support ncicb@pop.nci.nih.gov (please include "caCORE Training" in the subject)

Revision History

Authors/Responsible Party	Revision Number	Page(s) Effected	Date
Jennifer Brush	Original	All	December 17, 2006
Jennifer Brush	1	All – template change	January 24, 2006
Becky Angeles	2	All – caCORE 3.1 Functionality	June 5, 2006
Becky Angeles	3	All – format change	September 28, 2006
Jennifer Brush	4	All – content review / update	November 15, 2006
Jennifer Brush	5	Update based on reviewer comments	December 20, 2006
Becky Angeles	6	Update based on reviewer comments	March 12, 2007
Becky Angeles	7		August 27, 2007
Becky Angeles	8	Review Questions	July 30, 2008

Table of Contents

1 INTRODUCTION 4

2 COURSE DETAILS 4

3 LESSON 1: DEFINING DATA AND METADATA 4

3.1 OBJECTIVES FOR THE LEARNER 4
3.2 DATA VS. METADATA 4

4 LESSON 2: ISO/IEC 11179 METADATA STANDARD 5

4.1 OBJECTIVES FOR THE LEARNER 5
4.2 PURPOSE OF THE ISO/IEC 11179 METADATA STANDARD 6
4.3 METADATA COMPONENTS 6
4.4 METADATA CONSTRAINTS 8

5 LESSON 3: METADATA IN THE CADSR 8

5.1 OBJECTIVES FOR THE LEARNER 8
5.2 ADMINISTERED ITEMS 8
5.3 METADATA ORGANIZATION 9
5.4 OWNING VS. USING METADATA 9
5.5 MONITORING CHANGES TO METADATA 10
5.6 METADATA UNIQUENESS 10
5.7 ISO/IEC 11179 METAMODEL 10
5.8 CADSR AND STANDARD TERMINOLOGIES 11
5.9 SEMANTIC INTEROPERABILITY 12

6 REVIEW QUESTIONS 12

7 REFERENCES AND RESOURCES 15

8 CONTACT INFORMATION 16

1 Introduction

Welcome to caCORE Training. The session is designed for caDSR Users and Metadata Consumers, including caBIG Developers and other participants.

In order to receive credit for mastery of this content, you will need to register for this course and complete the accompanying quiz in the caBIG Learning Management System.

To register, go here: http://ncicbtraining.nci.nih.gov/TP2005/tp2000web.dll/NCICBTraining

We want these sessions to be as effective as possible in meeting your needs so we ask that you complete the short training evaluation form (available when you register for a course) to share your feedback on the overall quality of the training process and materials.

2 Course Details

Course Category:	caCORE
Course Number:	1010
Course Title:	Introduction to the ISO/IEC 11179 Metadata Standard
Course Level:	Basic
Audience:	caDSR Users and Metadata Consumers

This course is divided into the following lessons:

- Lesson 1: Defining Data and Metadata
- Lesson 2: ISO/IEC 11179 Metadata Standard
- Lesson 3: Metadata in the caDSR

3 Lesson 1: Defining Data and Metadata

3.1 Objectives for the Learner

On completion of this lesson, you will be able to:

- Discuss the difference between data and metadata
- Discuss the difference between how data and metadata are stored

3.2 Data vs. Metadata

In order to talk about metadata and how it is formed, we first need to define the differences between data and metadata.

Data is a single item of factual information. For example, if a question on a form reads: “What is your age?” the data is the actual value recorded for an individual person:

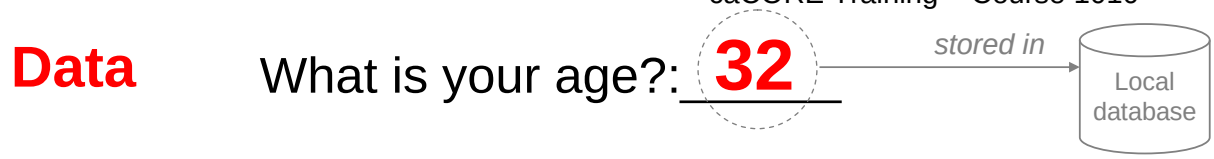


Figure 1 - Data Example

Metadata describes the meaning, content, quality, condition and other characteristics of a piece of data. For example, if a question on a form reads: "What is your age?" the metadata are the details describing what the value means and how the value(s) should be represented.

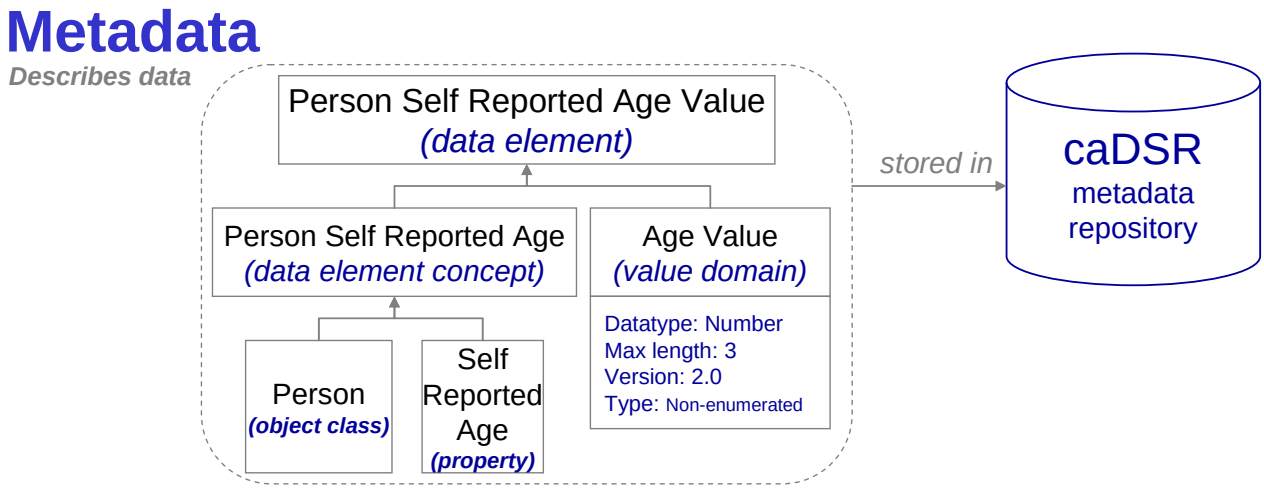


Figure 2 - Metadata Example

In the biomedical research community, data is stored either in a database or in hard copy on actual forms; metadata is stored in a metadata repository. For example, clinical trial data may be entered using electronic forms and stored in a database that is part of a clinical trial management system. Metadata is also stored in a database, but the metadata database (or metadata repository) is separate from the clinical trial data. NCI CBIIT metadata is stored in a metadata repository called the cancer Data Standards Repository (caDSR).

4 Lesson 2: ISO / IEC 11179 Metadata Standard

4.1 Objectives for the Learner

On completion of this lesson, you will be able to:

- Discuss the purpose and primary function of the ISO/IEC 11179 standard
- Describe the metadata components of the ISO/IEC 11179 standard
- List three constraints for forming metadata based on the ISO/IEC 11179 standard

4.2 Purpose of the ISO/IEC 11179 Metadata Standard

The caDSR is an implementation of the ISO/IEC 11179 metadata standard. The caDSR is a conforming implementation of the ISO/IOEC 11179 metadata standard with NCI extensions. This means that the caDSR has been designed to support the creation, maintenance and registration of metadata in accordance with the metadata standard. This also means the NCI has extended the standard to support use of NCI terminology to name and define its contents. The fundamental unit of metadata in the ISO/IEC 11179 Standard is called a Data Element.

The purpose of the ISO/IEC 11179 standard is to define a framework and protocols for how such metadata can be specified, consistently maintained, and shared across diverse domains. The caDSR conforms to this standard; while it contains extensions developed specifically to support registration of forms used in clinical trials data management, usage of the caDSR is not limited to clinical applications.

In a typical data dictionary, a Data Element might consist of a name, data type, length and perhaps a definition. However, the ISO/IEC 11179 standard (http://www.iso.org) contains an extensive collection of carefully defined components that specify the kind and quality of metadata necessary to describe data. ISO/IEC 11179 separates the conceptual entity of a piece of data (called a Data Element Concept) from its physical representation in a database (called a Value Domain). ISO also specifies the management and administration of this metadata in a metadata registry.

4.3 Metadata Components

It might be easier to think of Data Elements as the fundamental units of data an organization manages. When an organization needs to transfer data to another organization, Data Elements are the units that make up the transaction sets. The transaction set has to include both the data (concept) and how it should be represented (value).

The following section describes the metadata components that make up a Data Element and how different metadata components can be reused and combined in different ways to create new metadata components.

A **Data Element** is formed when you combine a concept or idea (Data Element Concept) with how it should be represented (Value Domain). Before we get into the specifics of how a Data Element is formed, first we should examine the details for forming its two primary components: a Data Element Concept and a Value Domain.

A **Data Element Concept** is formed when you combine a general concept (Object Class) with a specific characteristic (Property) of that concept in order to differentiate it from other, similar concepts. Object Classes and Properties have relationships to one another.

> An **Object Class** is a set of ideas, abstractions or things in the real world that can be identified with explicit boundaries and meaning and whose properties and behaviors follow the same rules.

A **Property** is that is a peculiarity that is common to all members of an Object Class.

For example, the Object Class: "Person" can be combined with different Properties ("Name", "Address", or "Age") resulting in different Data Element Concepts. Each of the Data Element Concepts in the table below means something different.

Object Class	Property	Data Element Concept
Person	Name	Person Name
Person	Address	Person Address
Person	Age	Person Age

You can also further specify both components of a Data Element Concept by adding one or more **Qualifiers** to the Object Class, the Property or both:

Object Class Qualifier	Object Class	Property Qualifier	Property	Data Element Concept
	Person	First	Name	Person First Name
	Person	Last	Name	Person Last Name
Deceased	Person	First	Name	Deceased Person First Name
Deceased	Person	Last	Name	Deceased Person Last Name

Table 1. Data Element Concepts

A **Value Domain** contains the specifics about how you want to represent a Data Element Concept. A Value Domain is composed of a number of attributes including the Representation, Qualifiers and Valid Values (if enumerated).

For example, the representation term "code" can be combined with different qualifiers and lists of values to create different Value Domains.

Representation Qualifier(s)	Representation	Value Domain	Values/Value Meanings
Toxicity	Code	Toxicity Code	DR = Definite PO = Possible PR = Probable UL = Unlikely UR = Unrelated
Adverse Event	Code	Adverse Event Code	1 = not serious 2 = life threatening 3 = death 4 = disability 5 = hospitalized
Postal	Code	Postal Code	*n/a (this is a non-enumerated Value Domain)*

Table 2. Value Domains

In the ISO/IEC 11179 metadata standard, the Value Domain not only describe the list of values and how the data are represented, but also the data type and, if necessary, the unit of measure,

character set, minimum and maximum length, etc. allowing you to create very specific representations.

Now we put everything together to form a **Data Element**:

Data Element Concept		Value Domain	Data Element
Object Class	*Property*	*Representation*	
Person	Address	Postal Code	Person Address Postal Code

Table 3. Data Elements

4.4 Metadata Constraints

A critical notion in the ISO metadata standard is that any concept represented by a Data Element must have an explicit definition that is independent of any particular representation. To enforce this, the ISO metadata standard specifies constraints on how the metadata is formed. Following are some of the more critical constraints.

A single Data Element Concept may be associated with many Data Elements.
You can reuse Data Element Concepts in combination with different Value Domains to form unique Data Elements.

A single Value Domain may be associated with many Data Elements.
You can reuse Value Domains in combination with different Data Element Concepts to form unique Data Elements.

Each Data Element Concept + Value Domain pair is Unique.
It is not possible to reuse the same Data Element Concept + Value Domain pair to define a new Data Element. A Data Element specifies exactly one Data Element Concept and one Value Domain.

5 Lesson 3: Metadata in the caDSR

5.1 Objectives for the Learner

On completion of this lesson, you will be able to:

- Define administered items
- Describe how metadata is organized and managed in the caDSR
- Discuss the difference between owning metadata and designating metadata for reuse
- Identify the tool used to monitor changes to caDSR metadata
- Identify the two attributes of an Administered Item that comprise its unique identifier
- Discuss the use of standard terminologies in creating caDSR metadata

5.2 Administered Items

An administered item is literally a component for which administrative information must be recorded. It may be a Data Element itself or one of its associated components (Representation, Value Domain, Data Element Concept, Conceptual Domain, Object Class, or Property). Following is a list of attributes that are common to all administered items.

- Context
- Public ID
- Version
- Name (LongName)
- Short Name
- Definition
- Workflow Status

Many administered items have attributes that are unique to their administered item type. For example, a Value Domain also has attributes related to representation such as its Data Type, Unit of Measure, Minimum Length, Maximum Length, etc.

5.3 Metadata Organization

The Cancer Data Standards Repository (caDSR) is a metadata repository at NCI CBIIT. In the caDSR, metadata is organized by Contexts and managed by Context Administrators. NCI CBIIT has created a number of interactive tools to help Context Administrators create, manage and maintain Data Elements in the caDSR. In this section, we will use some of the caDSR tools to illustrate how metadata is organized in the caDSR.

First let us talk about the notion of Contexts in the caDSR. Following is a screen shot from one of the caDSR tools, the Common Data Element (CDE) Browser (http://cdebrowser.nci.nih.gov). On the left hand side of the screen, you'll see a list of folders. These are the caDSR Contexts. Each Context owns (maintains and administers) their metadata. You can select the Context name to display all the Data Elements in that particular Context. Alternatively, you can search for Data Elements by entering one or more search terms (or a Public ID) in the text field on the right hand side of the screen.

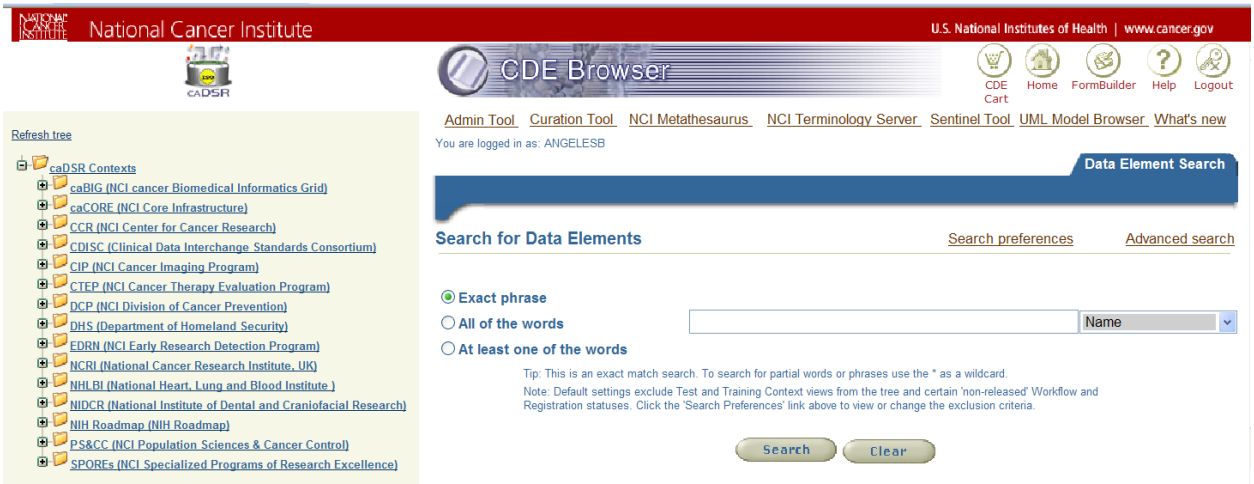
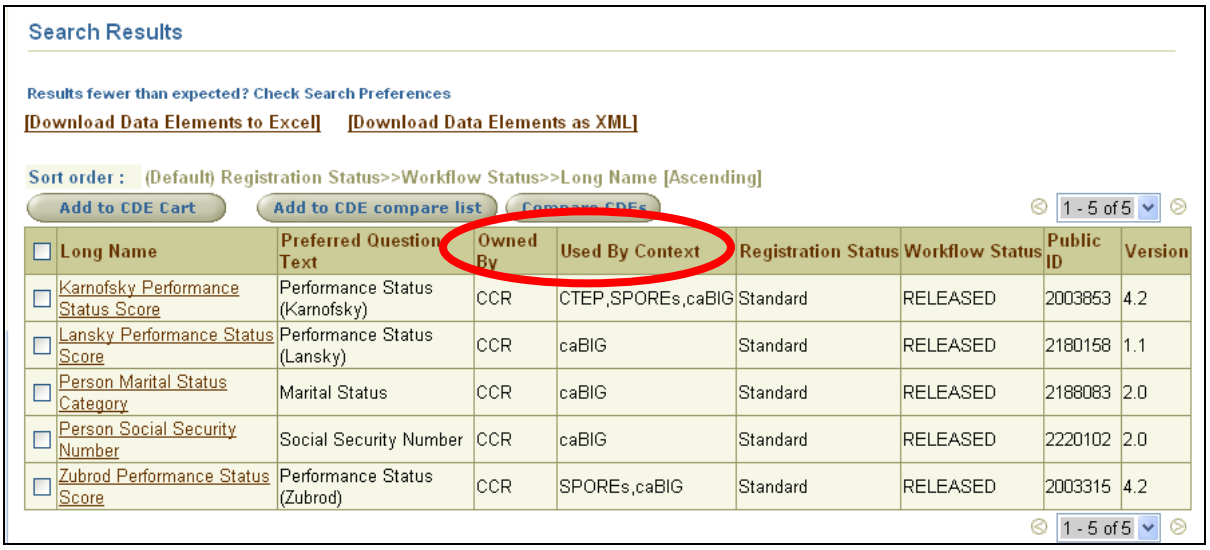


Figure 3 - CDE Browser Interface

5.4 Owning vs. Using Metadata

The caDSR supports the notion of "owning" metadata vs. "using" metadata. Owners create, modify, and version Data Elements. Contexts may also share their metadata with other

Contexts, or make use of metadata in other Contexts themselves. A Context Administrator can designate a Data Element for "reuse" within their Context. The screenshot below illustrates a set of Data Elements that are "Owned By" the CCR Context, but have also been designated for reuse ("Used By") other Contexts.



Long Name	Preferred Question Text	Owned By	Used By Context	Registration Status	Workflow Status	Public ID	Version
Karnofsky Performance Status Score	Performance Status (Karnofsky)	CCR	CTEP,SPOREs,caBIG	Standard	RELEASED	2003853	4.2
Lansky Performance Status Score	Performance Status (Lansky)	CCR	caBIG	Standard	RELEASED	2180158	1.1
Person Marital Status Category	Marital Status	CCR	caBIG	Standard	RELEASED	2188083	2.0
Person Social Security Number	Social Security Number	CCR	caBIG	Standard	RELEASED	2220102	2.0
Zubrod Performance Status Score	Performance Status (Zubrod)	CCR	SPOREs,caBIG	Standard	RELEASED	2003315	4.2

Figure 4 - Owning vs. Using (Designating) Data Elements

5.5 Monitoring Changes to Metadata

The caDSR Sentinel Tool (http://cadsrsentinel.nci.nih.gov) allows you to create and manage Alert Definitions for caDSR metadata. Alert Definitions are a set of rules that are periodically evaluated against the caDSR. If the conditions in those rules are met, you will be sent notification (via email) in the form of a hyperlink to an activity report that specifies the changes that have taken place. The caDSR Sentinel Tool provides users with the ability to monitor all changes to administered items, including Data Elements, Data Element Concepts and Value Domains. Reports may be set to run automatically daily, weekly, monthly or on demand.

5.6 Metadata Uniqueness

The unique identifier for all administered items in the caDSR is the Public ID and Version. There is also a unique identification sequence number for each version of an administered item that can be used if you are accessing caDSR metadata programmatically.

5.7 ISO / IEC 11179 Metamodel

The ISO/IEC 11179 defines a number of fields and relationships for metadata registries including a detailed metamodel for defining and registering administered items. The primary component of the ISO/IEC 11179 metamodel is a Data Element. The diagram below illustrates how the various components of the model relate to one another. Each box in the diagram contains the name of the ISO/IEC 11179 administered item type, and underneath an example of metadata describing the various parts of a Data Element.

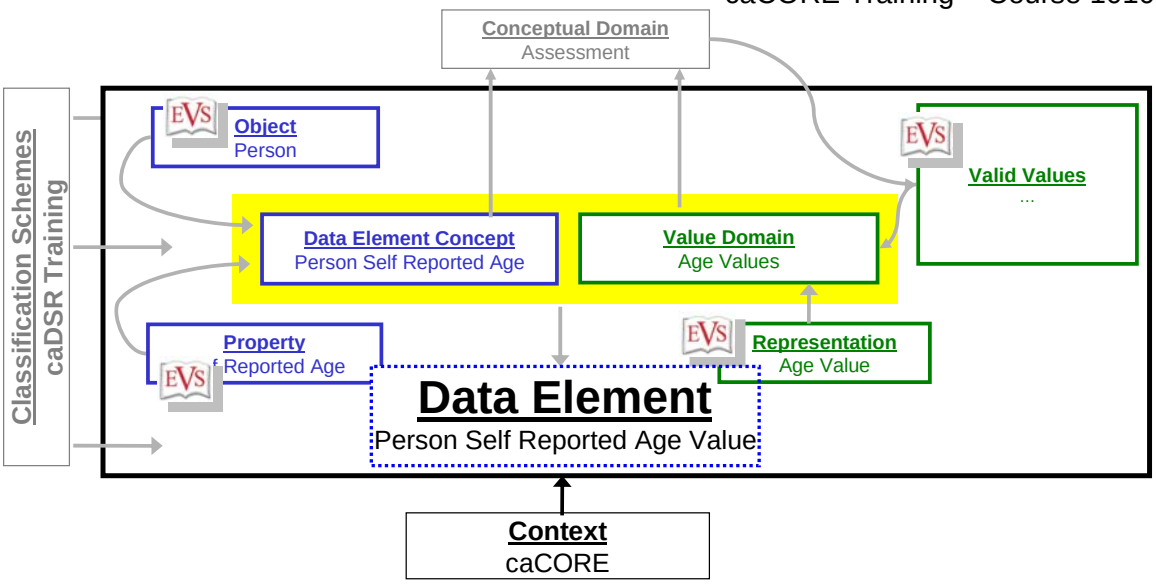


Figure 5. caDSR's Implementation of ISO

5.8 caDSR and Standard Terminologies

caDSR administered items are backed by the use of externally defined terminologies and controlled vocabularies. NCI CBIIT bases its data semantics on controlled terminology supplied by the NCI Enterprise Vocabulary Services (EVS) Project (http://evs.nci.nih.gov/). EVS is a set of services and resources that address NCI's needs for controlled vocabulary. Within the NCI CBIIT, EVS provides the semantics used to create caDSR metadata.

The NCI EVS produces two terminology products: the NCI Thesaurus (NCIt) a cancer focused terminology, and the NCI Metathesaurus (Meta) a mapping between many terminologies. The NCI Thesaurus is the preferred terminology for creating caDSR metadata. Architecturally the EVS terminology servers are a caCORE component. The caCORE client side API provides an open, publicly available unified interface to both of the EVS server environments. Applications that rely on EVS for terminology support make use of the caCORE API.

NCI CBIIT's clinical trials informatics efforts draw on EVS whenever domain terminology is required. This reliance on controlled vocabulary ensures that term meanings remain clear and can be related to other terms with similar concepts.

NCI Thesaurus has four features which provide the foundation of semantic integration for the caDSR and caCORE:

1. Concept codes that exhibit concept orientation
2. Permanence and non-semantic construction
3. Preferred terms for each concept
4. English language definitions and synonyms.

Of these the concept code is the most important, as it provides unambiguous linkage to a specific meaning and is machine interpretable. EVS Vocabularies can be searched for primary and qualifier concepts that can be used to create both the names and definitions for Object Class, Property, and Representation Terms. These are the building blocks for Data Element

Concepts and Value Domains. EVS Vocabularies are also searched when creating Permissible Values and Value Meanings as well as when creating the definitions for any of the administered items.

EVS vocabulary services are accessed via APIs (application programming interfaces) to provide touch points during caDSR metadata creation. The result is a collection of administered items that are bound to immutable concept codes. These touch points, denoted by the EVS logos in Figure 5, are implemented at the Object Class, Property, Representation Term, and Valid Value levels of the metadata model.

5.9 Semantic Interoperability

Semantic interoperability is the ability to exchange data in such a way that the precise meaning of the data is readily accessible and the data itself can be translated by any system into a form that it understands. The metadata registry allows a community, such as the cancer research and clinical trials communities, to register Data Elements from different sources; data collection forms, forms, legacy data repositories or applications, to name a few.

The DEC conveys the semantics of the Data Element with very precise detail. It is through the reuse of Data Element Concepts, and the semantics described by their metadata, that Semantic Interoperability is achieved.

The semantics of the DEC are based upon the Object Class and Property. In ISO, the Object Class is defined in unambiguous terms through related concepts. Concepts are unique, immutable, vocabulary identifiers drawn from EVS. At the DEC level in the caDSR, this unambiguous notion further is taken further by tying concepts to Properties and Qualifiers. Properties are what humans use to describe or distinguish objects from one another. Qualifiers are used to differentiate between names (i.e. Object Class and Property combinations) that would otherwise be the same.

Similarly, the Value Domains' semantics are unambiguously represented by a set of attributes, and in caDSR, the values in an enumerated Value Domain are linked to EVS concepts. Linking the names and definitions for the caDSR metadata to externally controlled vocabularies through concept identifiers, in particular the cancer terminology in the NCI Thesaurus enhances the overall semantics. The EVS concept identifiers can be used to navigate back to and leverage the source ontology.

6 Review Questions

1. A question on a form asks "What state were you born in?" and you answered "Texas." Your answer is:
    a. Data
    b. Metadata

2. A question on a website asks "What city were you born in?" The website limits your answer to 30 characters or less. This would be considered:
    a. Data
    b. Metadata

3. NCICB Metadata is stored in a:
    a. Electronic Dictionary
    b. Metadata Repository
    c. Clinical Trial Management System

4. NCICB's Metadata Repository is the:
    a. Enterprise Vocabulary Services (EVS)
    b. National Cancer Institute Center for Bioinformatics (NCICB)
    c. Cancer Data Standards Repository (caDSR)
    d. International Organization of Standardization (ISO)

5. The _____ metadata standard specifies the management and administration of metadata in a metadata registry.
    a. ISO/IEC 11179
    b. caBIO
    c. caCORE

6. Select the best answer: The NCI Thesaurus has which of the following features to provide the foundation of semantic integration for the caDSR and caCORE:
    a. Concept codes that exhibit concept orientation
    b. Permanence and non-semantic construction
    c. Preferred terms for each concept
    d. English language definitions and synonyms
    e. All of the Above

7. A Data Element is formed when you combine a Data Element Concept with a _____.
    a. Value Domain
    b. Object Class
    c. Property
    d. Conceptual Domain

8. A Data Element Concept is formed when you combine an _____ and Property.
    a. Object Class
    b. Data Element
    c. Value Domain

9. For the Data Element Concept "Deceased Person First Name", the word "First" is a(n):
    a. Object Class Qualifier
    b. Object Class
    c. Property Qualifier
    d. Property

10. For the Value Domain “Adverse Event Type”, what word conveys the Representation?
    a. Adverse
    b. Event
    c. Type

11. The Data Element “Person Last Name Text” is formed using the Data Element Concept “Person Last Name”. What is the Object Class?
    a. Name
    b. Last Name
    c. Person
    d. Person Last Name

12. The Data Element “Person Last Name Text” is formed using the Data Element Concept “Person Last Name”. What is the Property?
    a. Name
    b. Last Name
    c. Person Last Name

13. The unique identifiers for administered items in the caDSR are the _____ and Version Number.
    a. Public ID
    b. Concept Code
    c. Value Domain

14. In the caDSR, EVS concepts are tied to the Object Class, _____, Representation and Values.
    a. Property
    b. Conceptual Domain
    c. Classification Scheme

15. Different combinations of Data Element Concepts and Value Domains may be reused to create many _____.
    a. Data Elements
    b. Object Classes
    c. Properties

16. A _____ is a unique, immutable, vocabulary identifier from EVS.
    a. Data Element
    b. Concept Code
    c. Public ID

17. Which caDSR tool would you use to easily search for all the metadata in the caBIG Context?
    a. CDE Browser
    b. caDSR Sentinel Tool

c. Semantic Integration Workbench
d. CDE Curation Tool

18. Which caDSR tool would you use to monitor changes to your metadata in the caDSR?
    a. CDE Browser
    b. caDSR Sentinel Tool
    c. Semantic Integration Workbench
    d. CDE Curation Tool

19. _____ is the ability to exchange data in such a way that the precise meaning of the data is readily accessible and the data itself can be translated by any system into a form that it understands.
    a. Semantic Interoperability
    b. Curation

20. The NCI _____ is the preferred terminology for creating caDSR metadata.
    a. Thesaurus
    b. Metathesaurus

7 References and Resources

Below is a list of links to documentation used to create this session and of recommended reading materials.

caDSR Homepage:
- http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr

caDSR Training Home Page
- http://ncicb.nci.nih.gov/NCICB/training/cadsr_training

caCORE Developer's Guide:
- https://gforge.nci.nih.gov/docman/view.php/148/8650/caCORE%20SDK%204.0%20Developer's%20Guide_101007.pdf

caCORE Programmer's Guide:
- ftp://ftp1.nci.nih.gov/pub/cacore/SDK/v3.2.1/caCORE_SDK_3.2.1_Programmers_Guide.pdf

caDSR Business Rules
- http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr/business_rules/

caDSR_Users ListServ Subscription:
- https://list.nih.gov/archives/cadsr_training-l.html
- http://list.nih.gov

Send Request for caDSR Account to:
- ncicb@pop.nci.nih.gov

8 Contact Information

NCICBIIT Application Support:		ncicb@pop.nci.nih.gov
Instructor:	Jennifer Brush	brushj@mail.nih.gov
NCICBIIT Liaison:	Dianne Reeves	reevesd@mail.nih.gov