National Cancer Institute

CENDI/NKOS

New Dimensions in Knowledge Organization Systems

Semantic Interoperability in caBIG™

Leveraging Vocabulary,
Metadata Registries and Models

September 11, 2008

Denise Warzel
Associate Director, Core Infrastructure Program
NCI Center for Biomedical Informatics and Information Technology (CBIIT)

caBIG

National Cancer Institute

Agenda

- caBIG™ Semantic Interoperability Infrastructure
- Why bother?
- NCI's approach: Terminology + Metadata + Information Models

"The whole is bigger than the sum of the parts"

- Issues relative to implementation
  - (discussion)

caBIG

National Cancer Institute

Interoperability

- in·ter·op·er·a·bil·i·ty

  ability of a system...to use the parts or equipment of another system

  Source: Merriam-Webster web site

Syntactic interoperability

- interoperability

  ability of two or more systems or components to exchange information and

  to use the information that has been exchanged.

  Source: IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries, IEEE, 1990]

Semantic interoperability

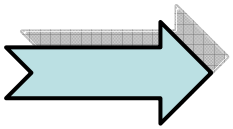
caBIG

National Cancer Institute

Why Bother?

- Simple programs = primary use of data within the immediate original context
  - spreadsheet/statistical packages
  - classic closed world RDBMS
  - file server/web server
  - E.g. Clinical Trial System, Patient Care, Image Analysis

- Enabling Discovery requires we get more out of data
  - *reuse of data outside of primary or original context*
  - integrate data from disparate sources
  - interoperability between systems
    - Computable Unambiguous meaning (semantics)
    - Computable Unambiguous syntax
      - Metadata Registries + Terminologies + Information Models



caBIG

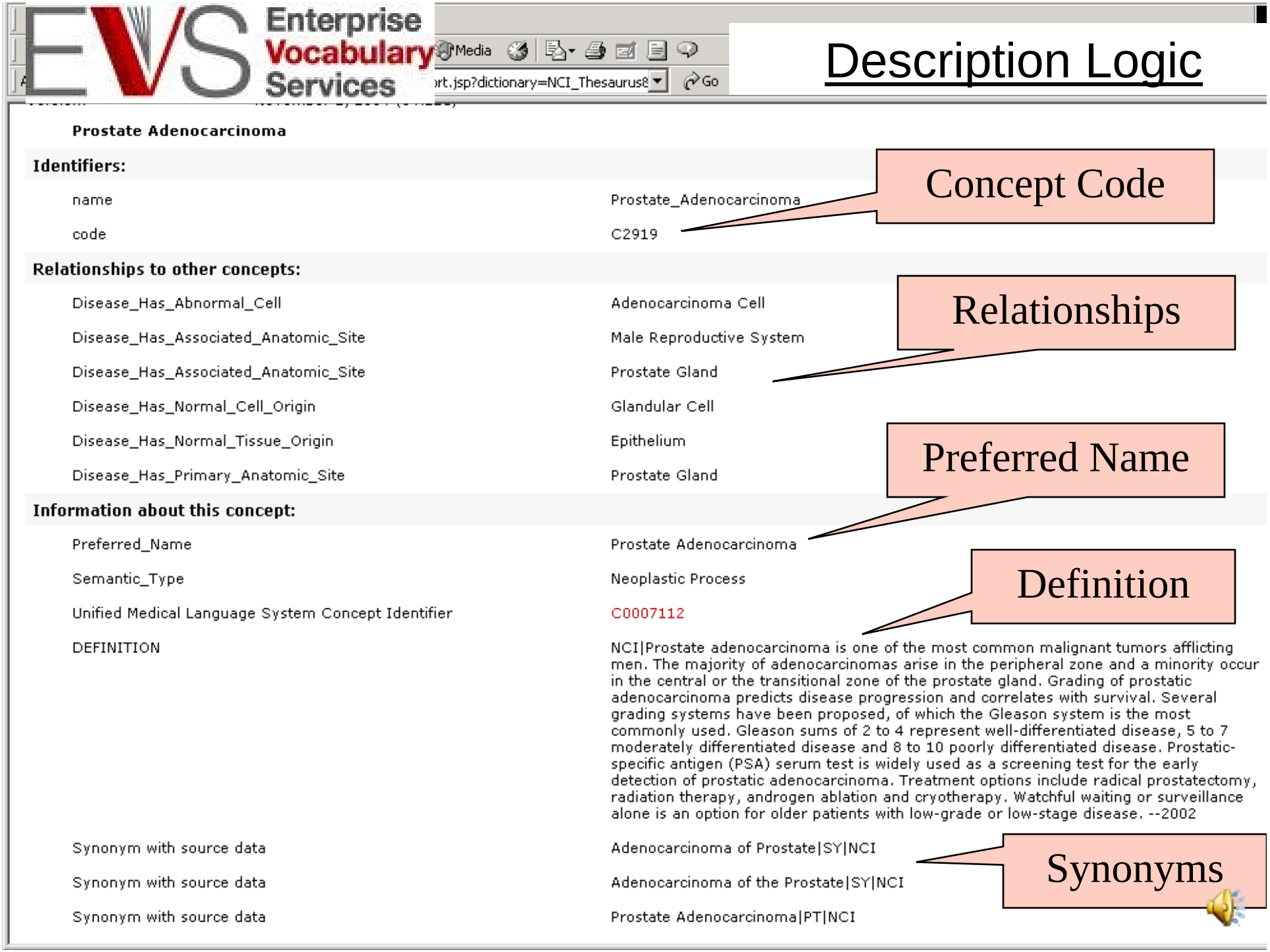
National Cancer Institute

caBIG™ Semantic Infrastructure

Information Models

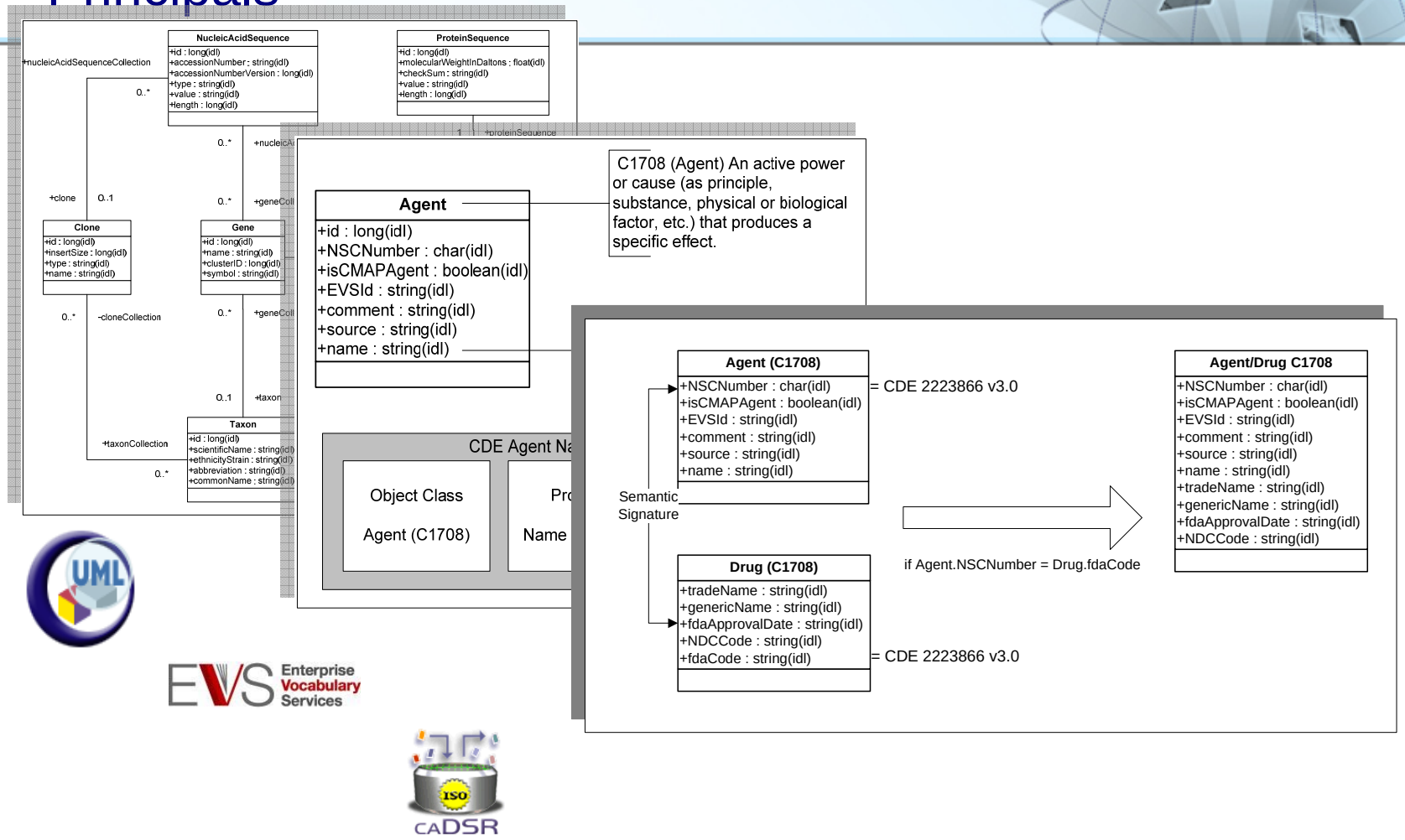
Data Elements

Enterprise Vocabulary


Enterprise Vocabulary Services
Media
Go
Description Logic
Prostate Adenocarcinoma
Identifiers:
name
Prostate_Adenocarcinoma
code
C2919
Concept Code
Relationships to other concepts:
Disease_Has_Abnormal_Cell
Adenocarcinoma Cell
Disease_Has_Associated_Anatomic_Site
Male Reproductive System
Disease_Has_Associated_Anatomic_Site
Prostate Gland
Relationships
Disease_Has_Normal_Cell_Origin
Glandular Cell
Disease_Has_Normal_Tissue_Origin
Epithelium
Disease_Has_Primary_Anatomic_Site
Prostate Gland
Information about this concept:
Preferred_Name
Prostate Adenocarcinoma
Preferred Name
Semantic_Type
Neoplastic Process
Unified Medical Language System Concept Identifier
C0007112
DEFINITION
NCI|Prostate adenocarcinoma is one of the most common malignant tumors afflicting men. The majority of adenocarcinomas arise in the peripheral zone and a minority occur in the central or the transitional zone of the prostate gland. Grading of prostatic adenocarcinoma predicts disease progression and correlates with survival. Several grading systems have been proposed, of which the Gleason system is the most commonly used. Gleason sums of 2 to 4 represent well-differentiated disease, 5 to 7 moderately differentiated disease and 8 to 10 poorly differentiated disease. Prostatic-specific antigen (PSA) serum test is widely used as a screening test for the early detection of prostatic adenocarcinoma. Treatment options include radical prostatectomy, radiation therapy, androgen ablation and cryotherapy. Watchful waiting or surveillance alone is an option for older patients with low-grade or low-stage disease. --2002
Definition
Synonym with source data
Adenocarcinoma of Prostate|SY|NCI
Synonym with source data
Adenocarcinoma of the Prostate|SY|NCI
Synonym with source data
Prostate Adenocarcinoma|PT|NCI
Synonyms

National Cancer Institute

caGrid is based on Object Oriented Principals



caBIG Interoperability by Objects and CDE

National Cancer Institute

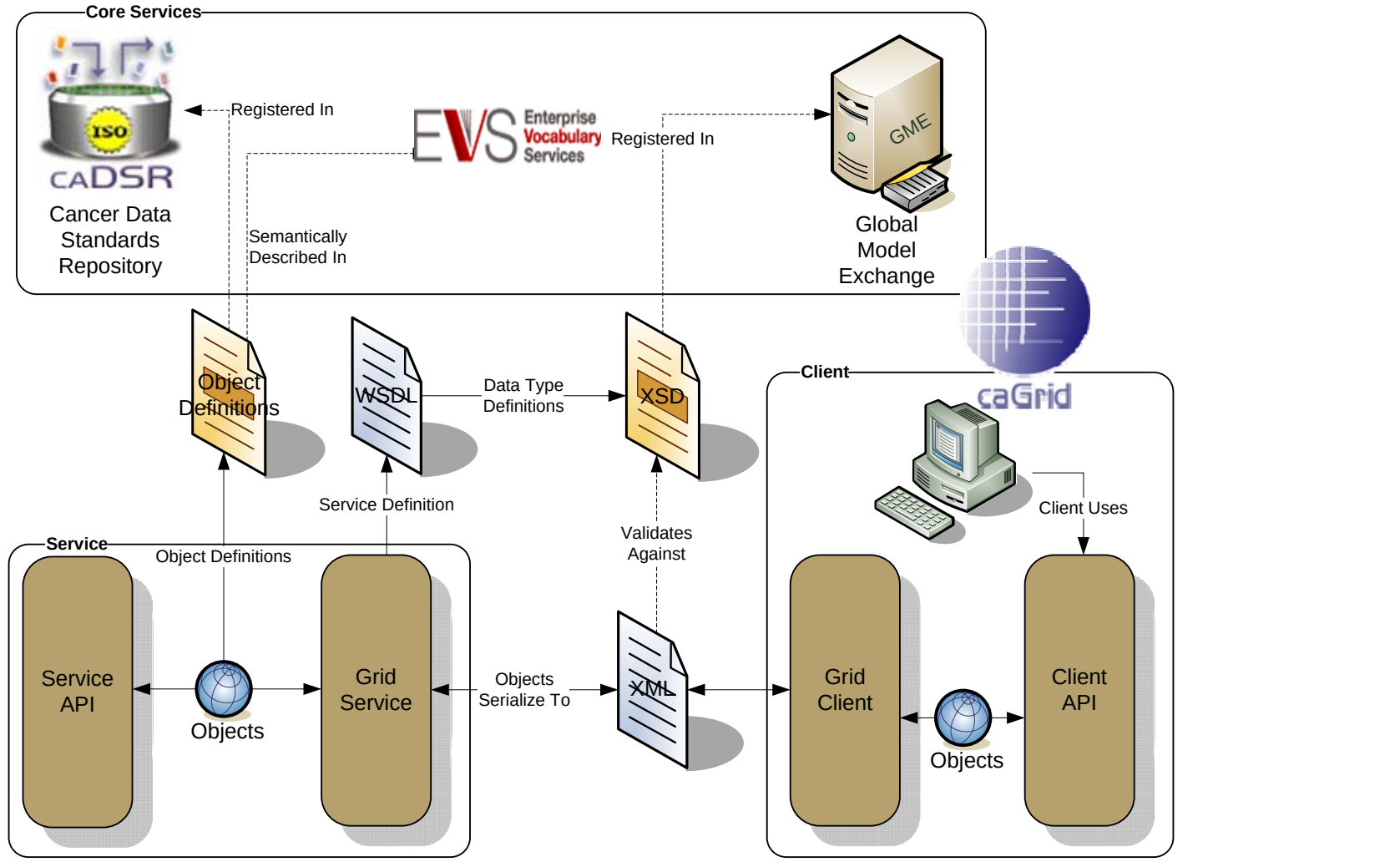
caBIG™ Community Involvement

- caGrid itself provides no real “data” or “analysis” to caBIG™; its the enabling infrastructure which allows the community to do so
- Community members add value to the grid as applications, services, and processes *(for example: shared workflows)*
  - caGrid provides the necessary core services, APIs, and tooling
- The real value of the grid comes from bringing this information to the end user
- Community members develop end user applications which consume of the resources provided by the grid

caBIG

National Cancer Institute

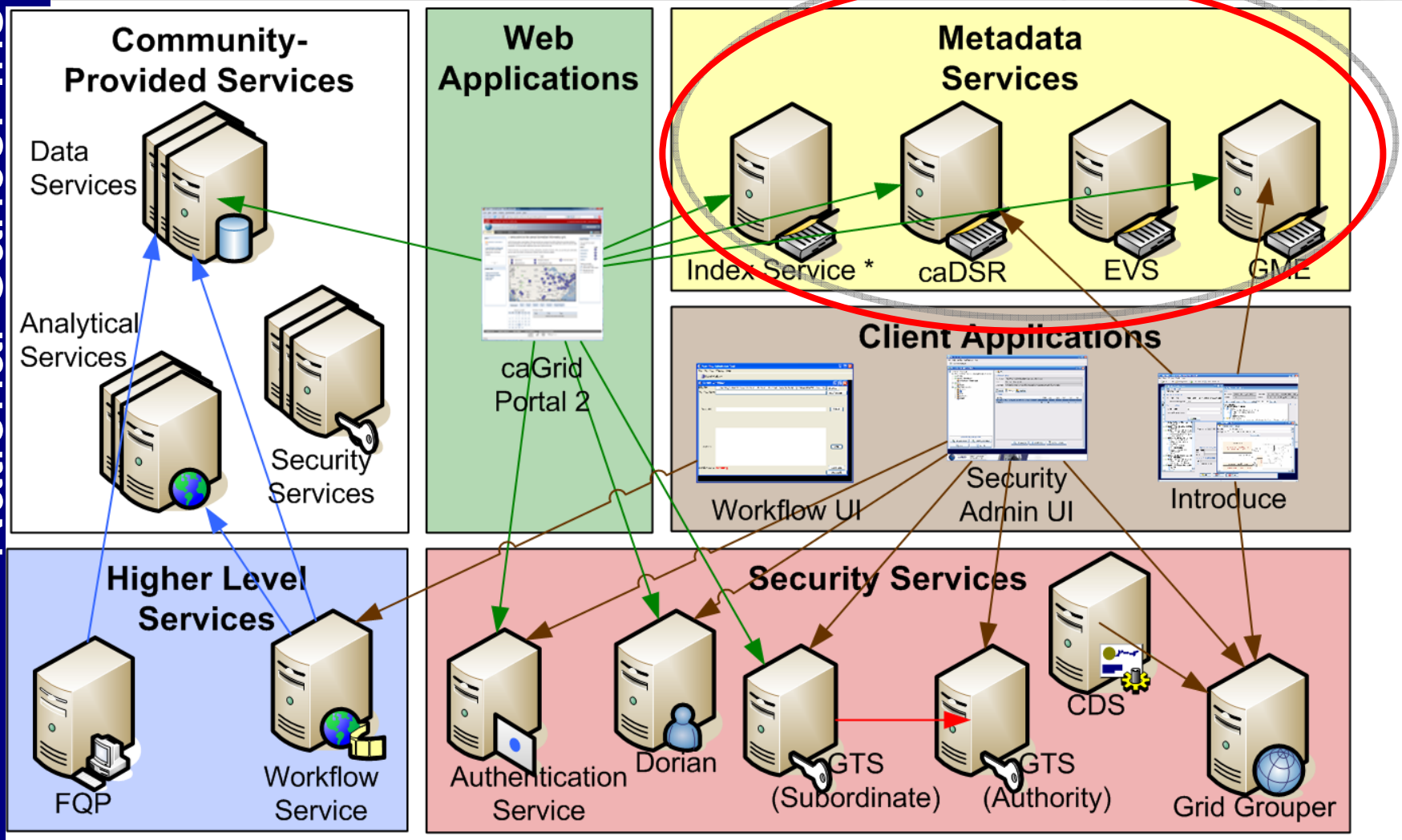
Semantics Artifacts in caGrid



caBIG

National Cancer Institute

caGrid Production Environment

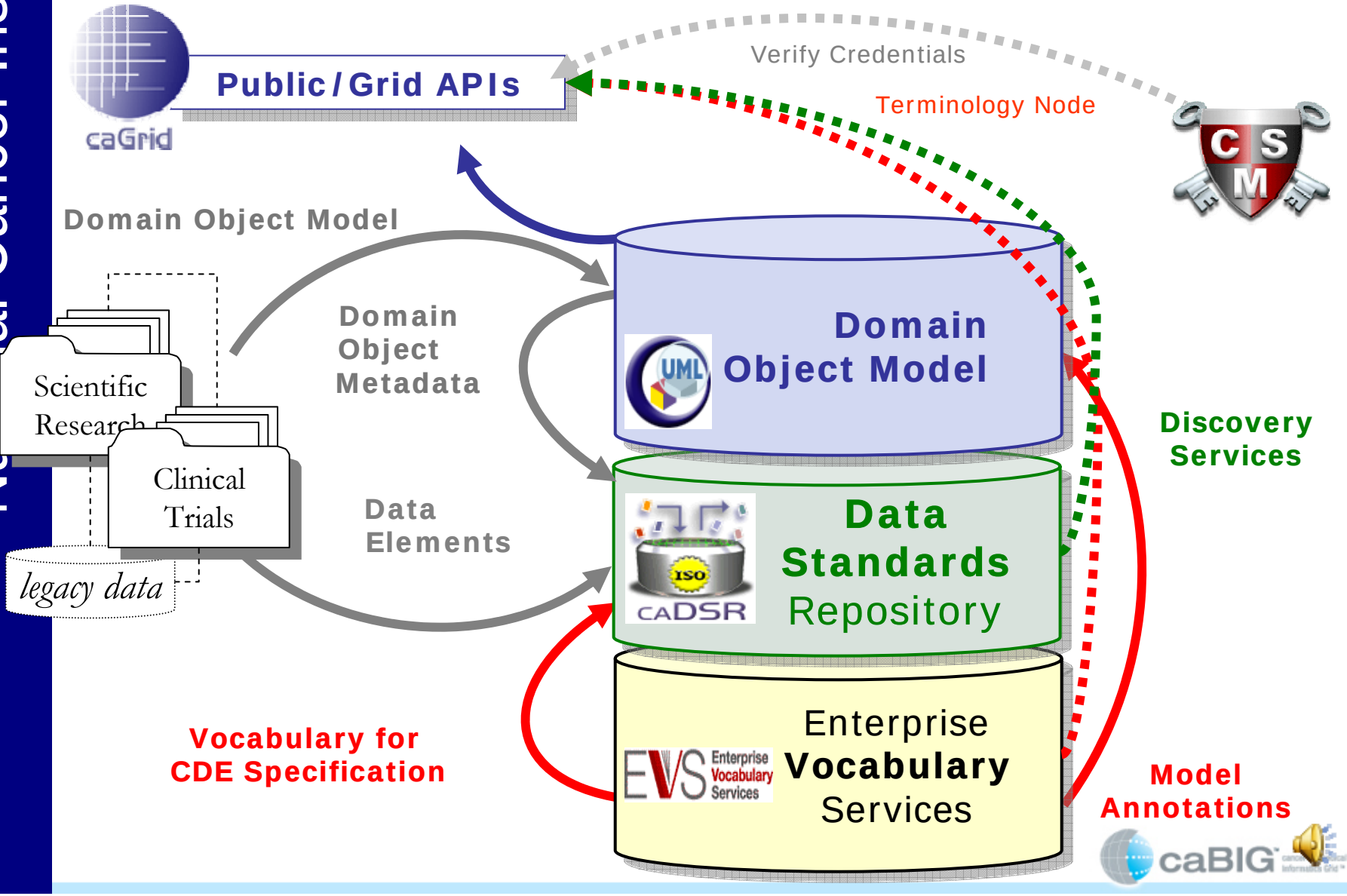


*All Services Register with the Index Service

caBIG

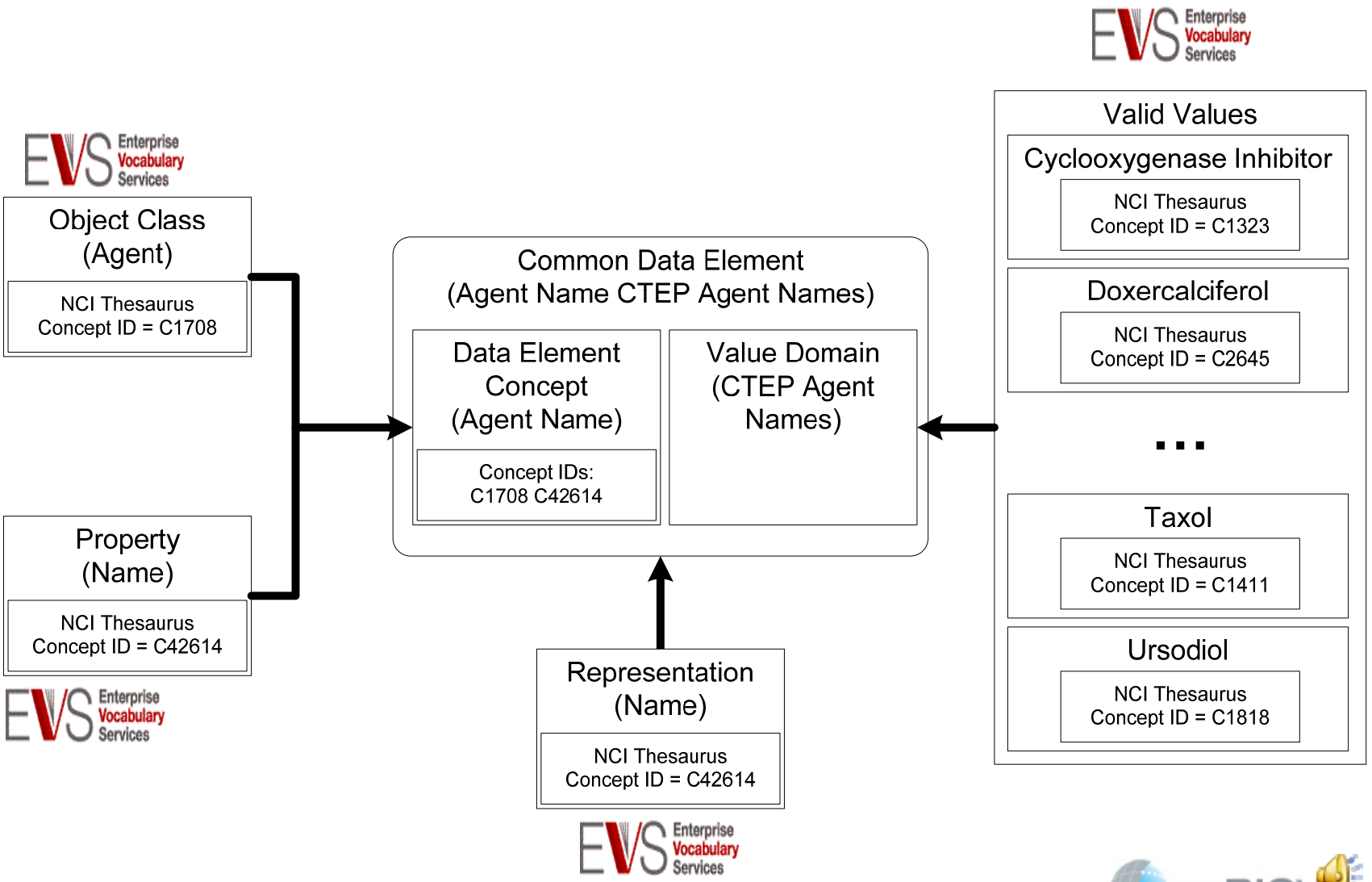
National Cancer Institute

Building caBIG™ Compatible Systems



National Cancer Institute

The ISO 11179 Model and Terminology Linkage in caDSR

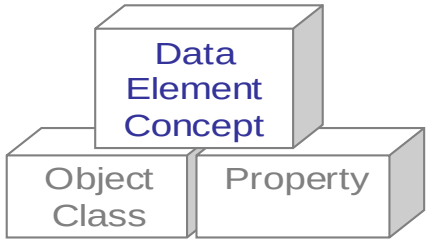


National Cancer Institute

ISO 11179 'Grammar'

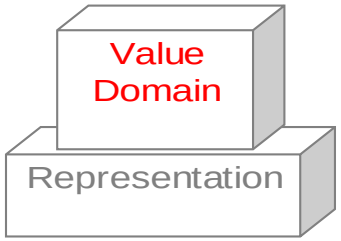
What is it? + *How do you want to represent it?* = *Common Data Element*

FDA Agent Name + Agent Name Text = FDA Agent Name Text



FDA Agent Name

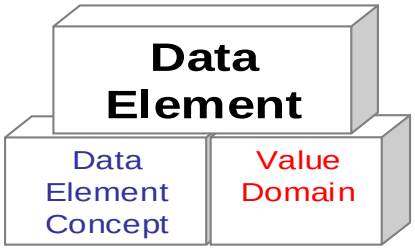
\+



Sting, Max=40

Anethole Trithione
Cyclooxygenase Inhibitor
Ginger
Green Tea
Iloprost
Taxol
...
Ursodiol

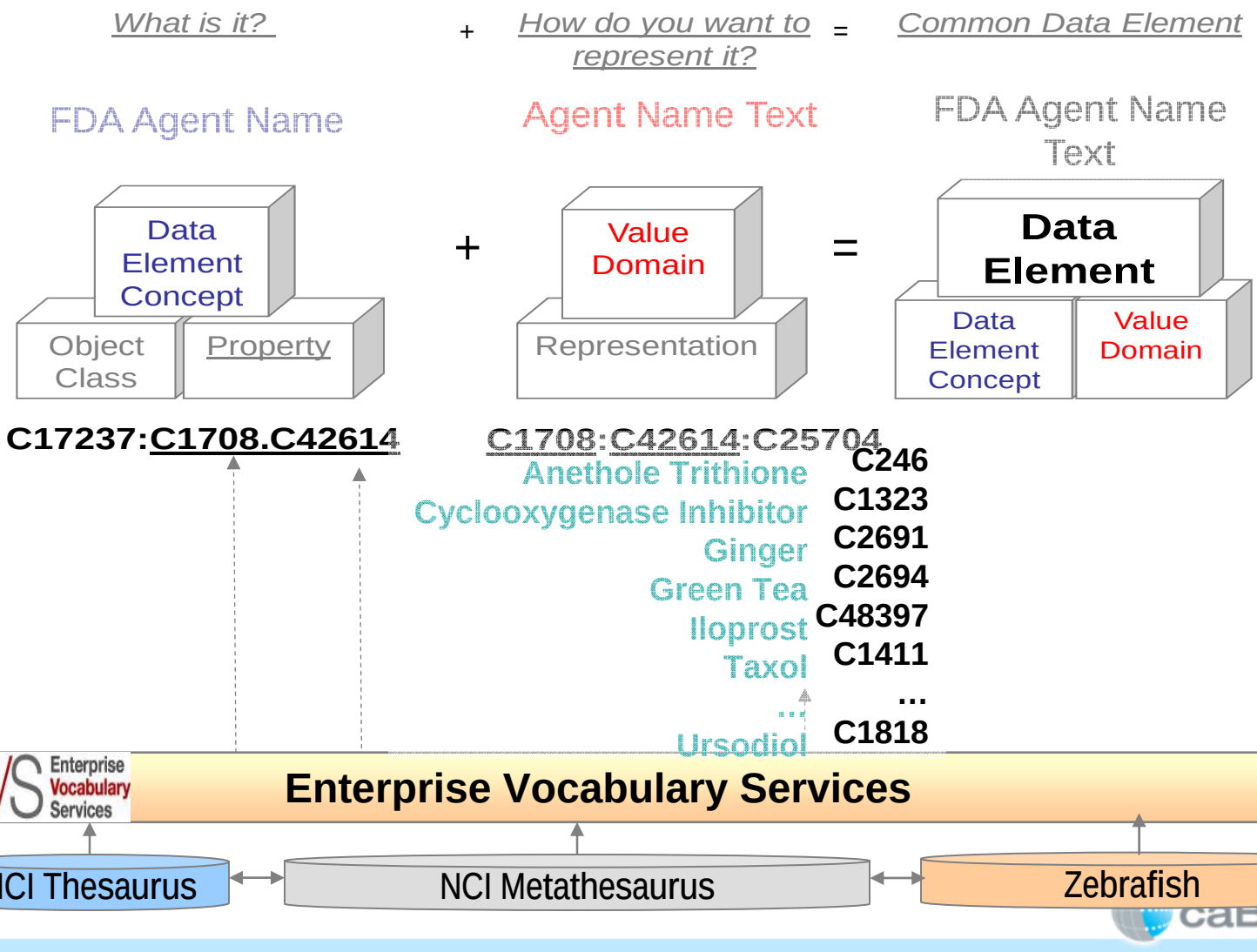
=



caBIG

National Cancer Institute

NCI Extension: 11179 Grammar + Concepts



National Cancer Institute

No Controlled Terminology? No computable interoperability

- Systems cannot automatically exchange or use information if they have use incompatible codes or tokens to store the data
- Linkage of metadata to terminology concept assures consistent meaning across the enterprise
- Metadata registry information can enable automation mappings/transformations between different tokens for the same code

caBIG

National Cancer Institute

Challenges

- Challenges relative to implementation when building disparate systems that still can interoperate
  - Information Modeling vs Domain Modeling (representing context)
  - Resolution of concepts from different terminologies – are they the same concept or different concept?
  - Many vocabularies aren't available programmatically
  - Many vocabularies don't contain identifiers
  - Vocabulary concepts usually aren't versioned
  - Vocabulary vs Metadata? (e.g. code sets, permissible value sets) - people get confused

caBIG