

# caBIG™ Compatibility Guidelines

*The Cancer Biomedical Informatics Grid™  
Program*



**caBIG™** *cancer Biomedical  
Informatics Grid™*

an initiative of the National Cancer Institute

This is a U.S. Government work.

May 1, 2008

# Table of Contents

<b>Chapter 1 Introduction .....</b>	<b>1</b>
Purpose.....	1
caBIG™ .....	1
Levels of Maturity.....	1
Interoperability Definitions and Goals.....	2
Achieving Syntactic and Semantic Interoperability.....	3
caBIG Principles and Implications for Interoperability .....	4
<b>Chapter 2 Compatibility Matrix .....</b>	<b>5</b>
<b>Chapter 3 Compatibility Guideline Details.....</b>	<b>7</b>
Programming and Messaging Interfaces.....	7
Vocabularies/Terminologies and Ontologies .....	9
Data Elements .....	12
Information Models .....	13
<b>Appendix A Supplemental Resources.....</b>	<b>16</b>
Supplemental Specifications.....	16
Useful Links and Resources .....	16
<b>Appendix B Glossary.....</b>	<b>18</b>

---

# Chapter 1 Introduction

## Purpose

---

The purpose of this document is to provide the cancer Biomedical Informatics Grid™ (caBIG™)<sup>1</sup> community with compatibility guidelines for creating and adopting software systems that are syntactically and semantically interoperable. The guidance contained herein is intended to support the evaluation of existing systems and to inform the designs of new systems. This document focuses on issues related to the representation of, access to, and exchange among biomedical informatics resources. Requirements for integration and use of the caBIG standards management infrastructure are also addressed. However, with few exceptions, a particular technology implementation of a given system or tool is not specified.

These guidelines represent a synthesis of several sources of thought, experience, tools, and practice in the areas of information systems development, adoption of existing data standards, development of data standards when needed, and interoperability.

Contributing sources include: Cross-cutting and Domain Workspaces from caBIG; Model-Driven Architecture from the Object Management Group (OMG); the ISO/IEC 11179 standard for metadata registries; Health Level Seven Version 3 (HL7); Clinical Data Interchange Standards Consortium (CDISC); Open Grid Forum (OGF); Semantic Web from W3C; Web Service Resource Framework and Security Standards from Organization for the Advancement of Structured Information Standards (OASIS); caCORE from the NCI.

## caBIG™

---

caBIG is a voluntary network or “grid” of individuals and institutions that are working to create a highly interoperable environment to enable researchers to share cancer research data and software tools. The goal of the program is to speed the delivery of innovative approaches for the prevention, detection, and treatment of cancer. The infrastructure and tools created by caBIG also have broad utility outside the cancer community. caBIG is being developed under the leadership of the National Cancer Institute<sup>2</sup> (NCI), its Center for Bioinformatics<sup>3</sup> (NCICB), and the caBIG participants themselves.

## Levels of Maturity

---

The caBIG community has recognized that there can be differing degrees of interoperability between systems. These can be qualified in terms of four maturity levels:

- **Legacy** -Implies no interoperability with an external system or resource. A legacy system is a system that was designed prior to or without awareness

---

<sup>1</sup> caBIG: <http://cabig.nci.nih.gov/>

<sup>2</sup> NCI: <http://cancer.gov/>

<sup>3</sup> NCICB: <http://ncicb.nci.nih.gov/>

of the availability of these compatibility guidelines, and which may not meet any of the requirements for interoperability.

- **Bronze** - Classifies the minimum requirements that must be met to achieve a basic degree of interoperability.
- **Silver** - A rigorous set of requirements that, when met, significantly reduce the barrier for using a resource by a remote party who was not involved in the development of that resource.
- **Gold** - Extensions to the silver requirements aimed at standardization and harmonization, which when met, enable full syntactic and semantic interoperability of disparate systems.

## Interoperability Definitions and Goals

Interoperability can be defined as the ability of a system to access and use the parts of another system. The caBIG program has made interoperability between data and software components a primary strategic goal. These compatibility guidelines provide a high-level description of the decisions made to date with respect to requirements for interoperability. The cross-cutting Vocabulary/Common Data Elements (VCDE) and Architecture workspaces were created as part of the caBIG initiative to provide an ongoing forum and mechanism for defining and ensuring interoperability across caBIG technology and data products. The activities of these workspaces will result in more detailed standards, specifications, and requirements, ensuring that the program goals are met.

It is useful to consider the interoperability requirements for access independently from those for use, although of course they must be synthesized in the final implementation.

- *Access* requirements in caBIG include programmatic access to data and tools from software, not just interactive access from end-user interfaces. Given this requirement, the primary obstacle to accessing parts of another system is heterogeneity in the programming and messaging interface syntax across systems that have been developed by independent groups, if indeed these interfaces exist at all. The problem of access is therefore a problem of poor *syntactic* interoperability. Standardization of application programming and messaging interfaces is necessary to overcome obstacles to syntactic interoperability.
- *Use* of a resource demands more than just access. Scientific analysis and interpretation requires a deep understanding of the procedures, manipulations, and parameters that go into the creation of a data resource or tool. Given this requirement, the primary obstacle to using parts of another system is the ambiguity behind the origin and meaning of the data. The problem of usage is therefore a problem of poor or ambiguous *semantic* interoperability. Explicit descriptions and definitions of the contents and meanings of resources using a set of agreed-upon terms and definitions are necessary to overcome barriers to semantic interoperability.

The highest degree of interoperability is attained when access and use can be completely automated. To achieve this level of interoperability, programming and messaging interfaces must conform to standards that specify consistent syntax and

format across all systems in the federation. Furthermore, all data must be associated with metadata and terminology identifiers and codes that support computational aggregation and comparison of information that resides in separate resources.

## Achieving Syntactic and Semantic Interoperability

---

When considering how to overcome the obstacles to interoperability, the caBIG program members defined four areas of interoperability that must be addressed. One of the four areas addresses issues related to syntactic interoperability, the remaining three address issues related to semantic interoperability. The four areas are described as follows:

- **Programming and Messaging Interfaces.** Computer programs and the people who write them are able to access resources from other programs through programming and messaging interfaces. Each of these interfaces responds to a particular syntax for its communications. Agreement upon standards for these interfaces is necessary to overcome barriers to syntactic interoperability.
- **Vocabularies and Ontologies.** Biomedical information includes a substantial body of specialized concepts or meanings that are represented by terms. Agreement upon the basic concepts, terms, and definitions that are inherent in all biomedical information is essential for achieving semantic interoperability. Terminology development systems that use description logic are helpful tools for managing these concepts.
- **Common Data Elements.** Data that is collected on a given study or trial must be defined and described such that remote users of that data can understand what it means. These metadata descriptions are referred to as data elements, and are based on controlled terminologies. When many groups use the same Common Data Elements (CDEs), then larger-scale studies can be conceived, since consistency and comparability across sites, studies, and time becomes possible. CDEs are therefore critical constructs for semantic interoperability.
- **Information Models.** Individual types of data are rarely collected or presented in isolation. Rather, they are assembled into a contextual environment that includes closely and more distantly associated data and information. These associations and relationships can be represented in the form of an information model. These models convey both a human and a machine readable representation of the contextual environment of data in an information resource, and are important for achieving the highest degree of semantic interoperability.

## caBIG Principles and Implications for Interoperability

---

The caBIG program has defined several principles that have implications for interoperability and for the creation and dissemination of the compatibility guidelines themselves:

- **Open Source/Open Access.** Products that are funded by NCI in connection with the caBIG initiative must be made available under licenses that permit free use and redistribution by any party, whether commercial, academic, or non-profit. Note, however, that privately funded groups can develop interoperable systems and tools that meet caBIG compatibility requirements without necessarily providing the resulting products under an open source/open access license, as long as this development was not funded by the caBIG program. These compatibility guidelines are themselves a caBIG-funded product, and are therefore distributed as an open access document.
- **Open Development.** caBIG-funded activities must be conducted in open forums, with opportunity for observation, comment, and contribution by any interested and qualified member of the community. These caBIG compatibility guidelines have been formulated with public involvement, comment, and review, and therefore adhere to this principle.
- **Federated.** The caBIG program envisions a federation of cancer biomedical informatics resources rather than a single repository or hosting center. These caBIG compatibility guidelines have therefore been driven by the goal of enabling developers of independently managed information resources and tools to achieve interoperability with other systems not under their direct control.

## Chapter 2 Compatibility Matrix

The following table contains a Compatibility Matrix which summarizes the caBIG compatibility requirements in the four areas of interoperability, stratified by the four maturity levels. *Chapter 3, Compatibility Guideline Details*, on page 7, includes more detailed information about each area of interoperability.

<b>Maturity Model</b>	<b>Legacy</b>	<b>Bronze</b>	<b>Silver</b>	<b>Gold</b>
<b>Programming and Messaging Interfaces</b>	<ul style="list-style-type: none"> <li>- No programmatic interfaces to the system are available. Only local data files in a custom format can be read.</li> <li>- Data transfer mechanisms are implemented only on an ad hoc basis.</li> </ul>	<ul style="list-style-type: none"> <li>- Programmatic access to data from an external resource is possible.</li> </ul>	<ul style="list-style-type: none"> <li>- Well-described APIs approved by the caBIG Architecture workspace provide access to data in the form of data objects that are instances of classes represented by a registered domain model.</li> <li>- Electronic data formats corresponding to a registered domain model approved by the caBIG Architecture workspace are supported wherever messaging is indicated by the use cases.</li> <li>- Messaging protocols approved by the caBIG Architecture workspace are supported wherever messaging is indicated by the use cases.</li> </ul>	<ul style="list-style-type: none"> <li>- All features of silver, plus:</li> <li>- APIs are exposed as operations of a Grid service; Object-Oriented client APIs are available for invoking those operations.</li> <li>- Service operations use XML as data exchange format, and are invoked using standardized protocols and communication channels.</li> <li>- Services provide public access to caGrid standardized service metadata and have capability to register it with a caGrid Index Service.</li> <li>- Data-oriented services provide query access using the caGrid standardized query interface and language.</li> <li>- Secure services must use the caGrid standardized mechanisms for authentication, trust management, and communication channel protection.</li> </ul>
<b>Vocabularies / Terminologies &amp; Ontologies</b>	<ul style="list-style-type: none"> <li>- Free text or local vocabularies used throughout for data collection.</li> </ul>	<ul style="list-style-type: none"> <li>- Local vocabularies or publicly accessible controlled vocabularies are used</li> <li>- Vocabularies must include term names that meet caBIG VCDE workspace guidelines.</li> </ul>	<ul style="list-style-type: none"> <li>- Controlled vocabularies reviewed and approved by caBIG VCDE workspace for use in silver applications are used for all appropriate data collection fields and attributes of data objects.</li> <li>- Concept identification should be compatible with the caBIG Identifier and Resolution Scheme</li> <li>- Vocabularies must be used for their intended scope and purpose.</li> </ul>	<ul style="list-style-type: none"> <li>- All features of silver, plus:</li> <li>- Full adoption of caBIG vocabulary standards as approved by the VCDE workspace.</li> <li>- Concept identification in systems must use the caBIG Identifier and Resolution Scheme</li> <li>- Vocabularies must be accessed through a standard caGrid Vocabulary API.</li> </ul>
<b>Data Elements</b>	<ul style="list-style-type: none"> <li>- No structured metadata is recorded.</li> </ul>	<ul style="list-style-type: none"> <li>- Data element descriptions are</li> </ul>	<ul style="list-style-type: none"> <li>- Common Data Elements (CDEs) built from controlled terminologies and according to</li> </ul>	<ul style="list-style-type: none"> <li>- All features of silver, plus:</li> <li>- CDEs designated as caBIG Standards by the</li> </ul>

		<p>maintained with sufficient definitional depth to enable a subject matter expert to unambiguously interpret the contents of the resource without contacting the original investigator.</p> <ul style="list-style-type: none"> <li>- Data elements are built using controlled terminology.</li> <li>- Metadata is stored and publicized in an electronic format that is separate from the resource that is being described.</li> </ul>	<p>practices validated by the VCDE workspace are used throughout.</p> <ul style="list-style-type: none"> <li>- CDEs are registered as ISO/IEC 11179 metadata components in the caBIG Context of the cancer Data Standards Repository (caDSR).</li> <li>- Reuse of existing CDEs in the caDSR should be considered before any new data elements are created. In order of descending priority, Standard, highly re-used and released CDEs should be considered for reuse.</li> </ul>	<p>VCDE workspace must be used.as appropriate.</p> <ul style="list-style-type: none"> <li>- CDEs generated from the Backbone Model must be re-used as appropriate.</li> <li>- Data elements must be expressed in caGrid standard metadata format</li> <li>- Existing validated CDEs in the caDSR must be re-used or otherwise justified before any new data elements are created.</li> </ul>
<p><b>Information Models</b></p>	<ul style="list-style-type: none"> <li>- No model describing the system is available in electronic format.</li> </ul>	<ul style="list-style-type: none"> <li>- Diagrammatic representation of the information model is available in electronic format.</li> </ul>	<ul style="list-style-type: none"> <li>- Object-oriented domain information models are expressed in UML as class diagrams and as XMI files, and are reviewed and validated by the VCDE workspace.</li> <li>- The classes, attributes and relationships of the information model are registered in the caDSR and correspond to the CDEs used by the system</li> <li>- Classes and attributes must be semantically annotated using terms from a controlled vocabulary that has been approved by the VCDE workspace.</li> </ul>	<ul style="list-style-type: none"> <li>- All features of silver, plus:</li> <li>- Information models must be harmonized across the caBIG Domain workspaces</li> <li>-The Backbone Model must be used as a template for information modeling.</li> <li>- XML schemas must be bound to the classes in the information model and are registered to Global Model Exchange (GME) service.</li> <li>- Information model must be expressed in the caGrid standard metadata format.</li> </ul>

---

# Chapter 3 Compatibility Guideline Details

The subsections in this chapter reiterate the information in Chapter 2, the Compatibility Matrix, and explain the operability area in more detail by each level of maturity.

## Programming and Messaging Interfaces

---

The compatibility criteria for “Programming and Messaging Interfaces” addresses issues related to programmatic access to a resource, input and output formats, and messaging protocols. The applicability of automated messaging interfaces versus an application programming interface (API) depends on the use cases and business requirements of the system being developed.

- To achieve **bronze level** maturity, the resource should provide, at a minimum, programmatic access to data through a public, documented API. The API must be rich enough to provide for the basic query and retrieval of information. This requirement does not place a constraint on the specific technology used to create and propagate the API.
- Achieving **silver level** maturity is more demanding. Systems are formally defined as tools, client interfaces, or messaging interfaces; interfaces are further categorized as analytical or data-oriented. All silver compatible systems must have domain models constructed in the Unified Modeling Language (UML; see *Information Models* on page 13). The structure of these classes and attributes must correspond to CDEs registered in the caDSR (see *Data Elements* on page 12) and have data types approved by the caBIG VCDE/Architecture workspaces. Interfaces must be defined in UML and provide a well-documented public API that is based upon an object-oriented abstraction of the underlying data. The data itself must be in the form of data objects that are instances of classes in the model. For data-oriented interfaces, the abstraction layer must be derived from a domain information model that expresses the underlying information space, providing a query API that exposes the connectedness and navigability of the data objects<sup>4</sup>.

Wherever use cases indicate that a messaging system is warranted, a standards-based messaging protocol approved by the Architecture workspace must be used to exchange information. Wherever possible, standard data formats defined by the Architecture/VCDE workspaces should be reused. The data formats must correspond to the registered domain model, and the messaging interface must demonstrate an object-oriented abstraction. Silver compatible analytical tools must be able to read directly from silver compatible interfaces.

---

<sup>4</sup> Silver API White Paper and Checklist:

[https://gforge.nci.nih.gov/docman/index.php?group\\_id=233&selected\\_doc\\_group\\_id=1137&language\\_id=1](https://gforge.nci.nih.gov/docman/index.php?group_id=233&selected_doc_group_id=1137&language_id=1)

- Achieving **gold level** maturity focuses the broader silver level messaging requirements to create a unified service-oriented data and analytical grid. While this section provides a brief overview of the high-level requirements of systems participating in this grid, it does not provide exhaustive details. Further information can be found in the accompanying specification documents identified in the *Supplemental Resources* on page 16.

Gold level systems must expose programmatic access as operations on grid services adhering to the *Protocol* section of the *caGrid Specification Document*<sup>5</sup> in use by the production grid deployment.

While the grid is inherently a message-based infrastructure (requests and responses between endpoints identified by WS-Addressing), also specified in the *Protocol* section of the *caGrid Specification Document*, grid service operations must be presented through an object-oriented client API. Each such grid service must also meet several functional requirements. First, each service must publish appropriate service level metadata, which is standardized by caGrid and specified in the *Metadata* section of the *caGrid Specification Document*. This metadata must be publicly accessible (available without authorization) and associated with the service through standardized service operations, and the service must provide the capability to register to the caGrid Index Service as described in the *Metadata* section of the *caGrid Specification Document*<sup>5</sup>. This standardized metadata minimally details the semantics and syntax of the service, its data elements and information models, and the people and institutions responsible for it. The operations of the service must use XML representations of data objects, as described in the *Service Architecture* section of the *caGrid Specification Document*, that meet the silver compatibility criteria (published per the requirements of the *Information Models* on page 13), and be able to be invoked using standardized protocols and communication channels (namely SOAP over HTTP or HTTPS).

As caGrid builds upon GSI, services that require security must use a standardized transport (HTTPS with support for X.509 proxy certificates) or message level (WS-Security or WS-SecureConversation) security mechanisms, as described in the *Security* section of the *caGrid Specification Document*. Additionally, a secure service must authenticate its clients using standardized mechanisms (X.509 proxy certificates), and have the capability to be part of the caBIG trust fabric (run with trusted service/host credentials and authenticate trusted user credentials), as described in the *Security* section of the *caGrid Specification Document*.

Data-providing services must at least provide query access in the form of the standardized query operation, as described in the *Data Service* section of the *caGrid Specification Document*, which specifies a common query language and processing faults. If a system has the need of uniquely identifying data on the grid, the caGrid Identifier Infrastructure, as described in the *Identifiers* section of the *caGrid Specification Document*, should be leveraged. This system is currently in the process of being developed and deployed; implementers needing to make use of this before its availability are encourage to review its design.

Because of the grid-oriented requirements for gold compatibility, creation of a gold compatible system from a silver compatible *messaging interface* (as opposed to an

---

<sup>5</sup> caGrid Specification Document: <http://gforge.nci.nih.gov/plugins/scm cvs/cvsweb.php/cagrid-1-0/Documentation/docs/specifications/caGrid-specification.doc?cvsroot=cagrid-1-0>

API) may require new components to be developed or replaced. That is, generally a silver system is wrapped, or exposed, as a grid service for gold compatibility. For this reason, new systems being developed with messaging components should be aware of the gold requirements, even if not currently targeting gold level compatibility.

Gold compatible tools must be able to receive data from gold compatible grid services. Wherever possible, the tools must leverage the discoverable nature of the grid; service endpoints, and specific data formats, should not be “hard-coded” in the system, but rather discovered from the caGrid production Index Service. Whenever data elements are presented to users, they should be presented using the semantics provided by the registered metadata of the common data element. The tool must be capable of invoking secure services using the gold security requirements defined above. Furthermore, in order to facilitate reuse and lessen the learning curve for users, core caGrid tools, services, and domain languages should be reused whenever possible (for example, DCQL for distributed query, BPEL for workflow, etc).

## Vocabularies/Terminologies and Ontologies

---

An important feature of modern terminology management is the recognition that the "concept" is the unit of semantic meaning, not simply the term or word. Concepts are described by preferred terms, synonyms, definitions and other properties. Given the diversity and overlap in meaning of terms in use, it is useful to use description logic to create and maintain concepts and to describe the relationships among concepts. These frameworks support the production of thesauri of non-redundant concepts that can be used to implement terminological and semantic consistency in data systems.

To be useful, a terminology must provide a clear textual definition of each term in the vocabulary, meet minimal levels of understandability, reproducibility, and usability (URU), provide adequate documentation, accessibility, and maintenance, and be free of serious intellectual property restrictions. As a vocabulary resource matures, it is expected that it will improve in all of these areas. Approval of a vocabulary by the VCDE workspace is contingent on meeting these criteria.

It is important to note that there are vocabularies whose use is mandated in certain settings (for example, to fulfill reporting requirements to a regulatory agency) or that are *de facto* community standards that will not meet the requirements of the caBIG compatibility guidelines. In these cases, the VCDE workspace is empowered to waive the requirements and will engage the owner/developer of the terminology in an effort to move the external vocabulary to the appropriate level of compliance.

Vocabularies represent a spectrum of semantic units including code sets, coding systems, controlled vocabularies, thesauri, taxonomies, and ontologies. For each type of semantic unit, caBIG provides varying guidelines and requirements based on principles of vocabulary and ontology best practices for the use in caBIG. Follow this caBIG<sup>6</sup> link for a variety of documents describing these principles.

---

<sup>6</sup> caBIG Vocabulary/Ontology best practices: <https://gforge.nci.nih.gov/projects/univocab/>

Concept identifiers and uniqueness are primarily managed by the ontology or vocabulary provider. The caBIG Concept Identifier on caGrid<sup>7,8</sup> is a common representation scheme for representing semantic classes (code schemes, vocabularies, taxonomies, or ontologies). The caBIG Identifier Scheme provides a syntax to utilize existing identifiers managed and maintained by vocabulary providers. The syntax represents a non-semantic identifier supporting machine interoperability and human readability to readily attribute authority, resource, and identifier. This system is currently in the process of being developed and deployed; implementers needing to make use of this before its availability are encouraged to review its design.

In addition to the caBIG Concept Identifier on caGrid, a global and common resolution mechanism is provided as part of the caGrid infrastructure. The resolution mechanism can resolve any caBIG Identifier to its common vocabulary metadata and attributes, such as description, definition, and provenance. There are fields for authority (where the terminology is from), the source (the terminology itself), source version, concept code and optionally, a concept version.

Vocabularies with ontological characteristics provide great potential for class or instance reasoning or inferring about data. As development matures, these additional capabilities will be standardized into core vocabulary and data service functionality on caGrid.

At the **bronze level** of maturity, the information resource or application utilizes public vocabularies in parts of the data collection and reporting process, but may supplement them with local vocabularies. All vocabularies, including those developed locally, should include descriptions of terms that are sufficient to distinguish the meaning of that term from other terms in the vocabulary. At a practical level, vocabularies at the bronze level should provide the essential characteristics of concept uniqueness and permanence. Term names should suggest the meaning and the intent of the term, as defined by the caBIG VCDE vocabulary criteria<sup>9</sup> and guidelines.

The **silver level** of maturity introduces the requirement for review and approval of vocabularies by the caBIG VCDE workspace<sup>10</sup>. Local or private vocabularies that are not available to the caBIG community may not be implemented. The NCI Enterprise Vocabulary Services (EVS) provides a management system for approved terminologies, but note that not all EVS-hosted terminologies have necessarily been reviewed and approved for caBIG. The VCDE workspace will use the criteria described above (understandability, reproducibility, usability, documentation, accessibility, maintenance, and intellectual property) to determine whether a

---

<sup>7</sup> caBIG Concept Identifier on caGrid:

[https://gforge.nci.nih.gov/docman/index.php?group\\_id=306&selected\\_doc\\_group\\_id=1451&language\\_id=1](https://gforge.nci.nih.gov/docman/index.php?group_id=306&selected_doc_group_id=1451&language_id=1)

<sup>8</sup> caBIG Concept Identifier on caGrid:

[https://gforge.nci.nih.gov/docman/view.php/306/7679/20070807\\_Arch\\_VCDE\\_F2F\\_ConceptID\\_caGrid.ppt](https://gforge.nci.nih.gov/docman/view.php/306/7679/20070807_Arch_VCDE_F2F_ConceptID_caGrid.ppt)

<sup>9</sup> caBIG vocabulary criteria: <http://gforge.nci.nih.gov/frs/download.php/1148/EVRC-criteria-1.0.pdf>

<sup>10</sup> caBIG VCDE Criteria:

[https://gforge.nci.nih.gov/docman/index.php?group\\_id=233&selected\\_doc\\_group\\_id=1136&language\\_id=1](https://gforge.nci.nih.gov/docman/index.php?group_id=233&selected_doc_group_id=1136&language_id=1)

vocabulary should be approved as described in the caBIG VCDE Vocabulary Criteria.

Core metadata about the vocabulary will be created and made available by the vocabulary submitter and/or reviewers as part of the review process, including scope and intended use of the vocabulary. Silver level maturity assures that vocabularies are used for their intended purpose and scope. As indicated above, the VCDE workspace may accept a vocabulary that meets these requirements for most of its terms, that is required for mandated reporting, or that has a clear plan for meeting these requirements. All data collection fields and attributes of data objects in silver level applications must use only approved vocabularies, as appropriate.

The **gold level** of maturity for an application or information resource is similar to the silver level, but with the added requirements that registered standards approved for caBIG-wide usage are implemented wherever they are available and that the vocabulary is accessible through a standard caGrid Vocabulary API. This is currently in the process of being developed and deployed; implementers needing to make use of this before its availability are encouraged to review its design. Gold compatible systems will reference and use vocabularies approved by the VCDE WS for use by gold systems. The following criteria apply to all vocabularies/ontologies approved for use by gold level systems:

- They are discoverable and accessible via standard caGrid services.
- Concepts identifiers of vocabularies are persisted and resolved based on the caBIG™ Identifier and Resolution Scheme<sup>11</sup>.

Given the dynamic nature of scientific research, terminology standards for caBIG are expected to grow and evolve as the scope of the program grows. Therefore, the enhancement and extension of currently available vocabulary resources will be an ongoing activity. Vocabularies used by applications will have differing semantic depth and breath based on their purpose. Software developers must be aware of these strengths and weaknesses of vocabularies in the context needed for the application/system.

New vocabulary resources will be considered in the context of the scope covered and in relation to their intended purpose. For new vocabularies with overlap in scope, appropriate concept translations will be required to ensure long term consistency and comparability of data. These mappings may be provided either through use of the NCI Metathesaurus, UMLS or a vocabulary to vocabulary mapping.

The primary goal of gold compatible as it relates to vocabularies is that caGrid data and analytic services will be able to leverage semantic relationships present in the vocabularies and ontologies. These relationships will aid in inferring, mapping, and including use cases for supporting cancer research.

Gold level systems will be fully enabled with implementation of the caBIG Identifier and Resolution Scheme for semantic classes used in data and analytic services. This system is currently in the process of being developed and deployed;

---

<sup>11</sup> caBIG Identifier and Resolution Scheme:

[https://gforge.nci.nih.gov/docman/index.php?group\\_id=306&selected\\_doc\\_group\\_id=1451&language\\_id=1](https://gforge.nci.nih.gov/docman/index.php?group_id=306&selected_doc_group_id=1451&language_id=1)

implementers needing to make use of this before its availability are encourage to review its design.

## Data Elements

---

While controlled terminology sources provide the semantic "raw material" for interoperability, they are stand-alone, independent resources that do not describe any particular data system. Developers of data management systems must separately characterize the contents of the actual system by mapping the data fields to structured metadata, or data elements. This requirement for documenting the metadata only covers attributes exposed as part of the system's public APIs or messaging interfaces, not all the internal features of lower layers. The public interfaces are the access points for the resources, and the output from these interfaces is what will be supplied to the next step of the information flow during the execution of a given use case.

A Data Element (DE) is a unit of metadata that describes the concept meaning behind a given datum that is collected. Common Data Elements (CDEs) provide a means toward semantic continuity and data comparability across information systems over time. The controlled terminology assigned to each object class and property must be synonymous (consistent) with the developer-derived definition of the class and attribute of an UML model. It is critical that the semantic annotation, which forms the basis of the data element, completely and accurately define the data that the data element is supposed to represent. CDEs help solve problems of ambiguity by providing precise definitions of data fields and types sufficient to unambiguously characterize the specific meaning of any particular datum collected in an information system. CDEs ultimately save analysis time by minimizing the need to reverse engineer meaning from data and also by enabling consistent data collection across locations in large multi-site investigations. The caBIG VCDE workspace has adopted a series of processes and best practices for the construction of well-formed Data Elements.

It is worth noting that there are two major mechanisms for creating CDEs in the caDSR.

1. Constructing individual data elements and their associated components by trained metadata curators using editing tools that operate directly on the caDSR.
2. Deriving data elements from an information model properly constructed in UML, where each data element corresponds to a tuple of class, attribute, and value domain. Such models developed by caBIG projects must be registered into the caDSR.

**Bronze level** systems have their metadata structured in an electronic format that details the specification of each data element that is in the system. These metadata are constructed from selected controlled terminology sources and include sufficient descriptive information to enable a subject matter expert to interpret the contents of the system without having to contact the original investigator. The metadata are exposed in a publicly accessible electronic resource that is distinct from the information system itself.

**Silver level** maturity is again more rigorous, but as such provides for a much higher degree of semantic interoperability, including the provision for computational

aggregation and comparison of data. CDEs constructed according to best practices defined by the caBIG VCDE workspace must be used. These CDEs are all registered in the caBIG Context of the NCI cancer Data Standards Repository (caDSR)<sup>12</sup>, an implementation of the ISO/IEC11179<sup>13</sup> standard for metadata registries. Reuse of existing CDEs in the caDSR must be considered before any new data elements are created. This could take the form of partial CDE reuse, which includes data element concept, object class and property reuse, with or without concomitant value domain reuse. All new CDEs are subject to review and validation by VCDE workspace-determined processes before being made publicly available in the caDSR.

**Gold level** requirements for data elements are an extension of the silver level specification, with added requirements for usage of appropriate CDEs that have been approved as standards for caBIG-wide usage. In addition, highly re-used CDEs, including those derived from the Backbone Model (see *Information Models* on page 13), must be re-used as appropriate. Justification is required if these high impact CDEs are not re-used. Developers are expected to review caBIG CDE standards, and caBIG will provide pointers to necessary resources. Examples of highly reused CDEs, one or more of which are pervasive through many developer projects, include those where the class represents the patient or participant, the biospecimen, and genes and gene products. Note that all of these are modeled in the Backbone Model generated by the VCDE Large Scale Harmonization Group<sup>14</sup>. Reuse of high impact CDEs that function as touch points among models would greatly increase the semantic interoperability of caGrid applications.

The Gold level requirements for *Programming and Messaging Interfaces* (see page 7) specify that service level metadata be made available in caGrid standard metadata format. The Data Elements used by the service as part of its operations must be fully described in this metadata to facilitate effective discovery, advertisement and interoperability.

Additional requirements for data provenance will be addressed in future releases.

## Information Models

---

Data Elements are precise specifications of individual types of data that are collected during a research study or using measurement technologies. However, scientific interpretation relies on the placement of data elements into the broader semantic context of an information model. Therefore, in order to attain the highest degree of semantic interoperability, data must be expressed in the context of such a model.

The benefits of using a standard modeling language are significant. UML is derived from a structured meta-model, and therefore all UML models share a common parental meta-structure. This trait allows for programmatic access to the models themselves, a feature that is leveraged when models are registered into the caDSR.

---

<sup>12</sup> NCI Cancer Data Standards Repository:

[http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore\\_overview/cadsr](http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr)

<sup>13</sup> ISO/IEC 11179 Standard for Metadata Registries: <http://metadata-standards.org/11179>

<sup>14</sup> VCDE Large Scale Harmonization Group:

[https://gforge.nci.nih.gov/frs/download.php/1651/caBIG\\_Model\\_Harmonization\\_whitepaper.doc](https://gforge.nci.nih.gov/frs/download.php/1651/caBIG_Model_Harmonization_whitepaper.doc)

The common meta-model also enables software code to be automatically generated from the models, a key benefit of the model-driven architectural paradigm espoused by the Object Management Group<sup>15</sup> (OMG). The Model driven Architecture (MDA) paradigm and its implementation using UML are adopted in caBIG. In order to assist developers in constructing information models, two relevant resources exist:

1. The Backbone Model which can be used as a template for constructing an information models.
2. UML model representations of caBIG™ CDE standards are also available<sup>16</sup>.

Using MDA and UML, a number of caBIG silver requirements for programming interfaces can be satisfied by automatically generating model-driven middleware code. Refer to the silver API white paper and checklist to review the criteria for silver programming interfaces<sup>17</sup>.

The **bronze level** requirement for an information model is simply that the person or organization requesting the review provides in an electronic format a diagrammatic representation of the information structure being produced by the system.

**Silver level** maturity requires the use of the industry-standard modeling language, UML, to create domain models that describe the content of a system. UML class diagrams that illustrate the data classes, attributes, and relationships are required. Using other aspects of UML modeling is encouraged as a best practice in development methodology, but is not central to the issue of semantic interoperability. Class diagrams should conform to the naming conventions<sup>18</sup> and terminology standards<sup>19</sup> prescribed by the caBIG program. UML models must be fully annotated with class and attribute definitions, and with associated terminology concept codes. UML model classes and attributes are semantically annotated with controlled terminology that must be synonymous (consistent) with the developer-derived definitions. The models must be provided in XML Metadata Interchange (XMI) format in addition to any diagrammatic representations. Upon review and validation through processes determined by VCDE workspace, models must be registered into the caDSR<sup>10</sup>.

**Gold level** maturity for Information Models will involve an added degree of harmonization across caBIG domains. The goal of the harmonization effort is to maximize the interoperability between caBIG applications and workspaces. This harmonization effort consists of three processes:

1. Creation of a caBIG underspecified domain model referred to as the Backbone Model<sup>20</sup>.
2. Evolution of current caBIG systems in relation to the Backbone Model.

<sup>15</sup> OMG: <http://www.omg.org/>

<sup>16</sup> CDE Standards with UML snippets: <https://gforge.nci.nih.gov/projects/cdestandards/>

<sup>17</sup> Silver API White Paper and Checklist:

[https://gforge.nci.nih.gov/docman/index.php?group\\_id=233&selected\\_doc\\_group\\_id=1137&language\\_id=1](https://gforge.nci.nih.gov/docman/index.php?group_id=233&selected_doc_group_id=1137&language_id=1)

<sup>18</sup> caCORE SDK Programmers Guide: <http://ncicb.nci.nih.gov/NCICB/infrastructure/cacoresdk>

<sup>19</sup> Vocabulary Standards Gforge Site: <https://gforge.nci.nih.gov/projects/vocabstandard/>

<sup>20</sup> Backbone Model:

[https://gforge.nci.nih.gov/frs/download.php/1651/caBIG\\_Model\\_Harmonization\\_whitepaper.doc](https://gforge.nci.nih.gov/frs/download.php/1651/caBIG_Model_Harmonization_whitepaper.doc)

3. Evaluation of new application models in relation to Backbone Model prior to system implementation.

The Backbone Model is constructed from existing caBIG compatible applications. There is a graceful evolution from current system models to a harmonized model based on the Backbone Model and commonly used information objects within caBIG. This is a consensus driven process coordinated by both the VCDE and Architecture workspaces and is implemented with an awareness of caBIG's fundamental principle of federation. To facilitate the adoption of existing caBIG information objects, new systems must be reviewed during their initial UML modeling process. This will ensure that whole classes from other caBIG-compatible UML models are reused if appropriate. The use of whole classes is to support interoperability of analytical services. The review is to be performed in relation to the Backbone Model (which will be used to identify common information objects and CDEs); it will result in the maximal reuse of information objects and CDEs. The reuse of CDEs at gold level is consistent with processes of reuse at the silver level.

Gold level requirements for *Programming and Messaging Interfaces* (see page 7) specify that service level metadata be made available. The Information Model used by the service as part of its operations, or exposed through query operations, must be fully described in this metadata to facilitate effective discovery, advertisement, and interoperability.

Gold level requirements for *Programming and Messaging Interfaces* (see page 7) requirements specify that XML must be used as the data exchange language of the Information Models and that those XML formats must have a canonical representation and be defined as XML Schemas<sup>21</sup>. The XML Schemas must be published to the caGrid production Global Model Exchange (GME), and have a formal binding to their corresponding Information Model, as described in the *Service Architecture* section of the *caGrid Specification Document*.

---

<sup>21</sup> XML Schemas: <http://www.w3.org/XML/Schema>

# Appendix A Supplemental Resources

## Supplemental Specifications

---

- Grid Specifications: Details the relevant specifications and policies associated with the current production Grid environment of caBIG. While caGrid tooling obviates the need for most developers to have a detailed understanding of these specifications, they are provided as reference when said tooling may not be able to be leveraged.  
<http://gforge.nci.nih.gov/plugins/scmvs/cvsweb.php/cagrid-1-0/Documentation/docs/specifications/caGrid-specification.doc?cvsroot=cagrid-1-0>
- Silver API Checklist and White Paper  
[https://gforge.nci.nih.gov/docman/index.php?group\\_id=233&selected\\_doc\\_group\\_id=1137&language\\_id=1](https://gforge.nci.nih.gov/docman/index.php?group_id=233&selected_doc_group_id=1137&language_id=1)
- Silver Information Model, CDE and Vocabulary Checklist  
[https://gforge.nci.nih.gov/docman/index.php?group\\_id=233&selected\\_doc\\_group\\_id=1136&language\\_id=1](https://gforge.nci.nih.gov/docman/index.php?group_id=233&selected_doc_group_id=1136&language_id=1)
- caBIG VCDE Vocabulary Criteria:  
<https://gforge.nci.nih.gov/projects/vocabcriteria/>  
<http://gforge.nci.nih.gov/frs/download.php/1148/EVRC-criteria-1.0.pdf>
- Crosscutting Model Harmonization White Paper:  
[https://gforge.nci.nih.gov/frs/download.php/1651/caBIG\\_Model\\_Harmonization\\_whitepaper.doc](https://gforge.nci.nih.gov/frs/download.php/1651/caBIG_Model_Harmonization_whitepaper.doc)
- caBIG Concept Identifier and Resolution Scheme for Vocabulary Resources  
[https://gforge.nci.nih.gov/docman/index.php?group\\_id=306](https://gforge.nci.nih.gov/docman/index.php?group_id=306)  
[https://gforge.nci.nih.gov/docman/view.php/306/7679/20070807\\_Arch\\_VCD\\_E\\_F2F\\_ConceptID\\_caGrid.ppt](https://gforge.nci.nih.gov/docman/view.php/306/7679/20070807_Arch_VCD_E_F2F_ConceptID_caGrid.ppt)  
[https://gforge.nci.nih.gov/docman/view.php/281/8003/20070823\\_VCDE\\_WS\\_Minutes.doc](https://gforge.nci.nih.gov/docman/view.php/281/8003/20070823_VCDE_WS_Minutes.doc)

## Useful Links and Resources

---

- caBIG Architecture Workspace: <http://cabig.nci.nih.gov/workspaces/Architecture>. Forum for discussing, prototyping and defining caBIG architectural standards, interoperability technologies, and engineering best practices.
  - caGrid Page: <https://cabig.nci.nih.gov/workspaces/Architecture/caGrid/>
  - caGrid Wiki: <http://www.cagrid.org>
- caBIG VCDE Workspace: <http://cabig.nci.nih.gov/workspaces/VCDE>. Forum for establishing and reviewing the use of caBIG data standards.

- caBIG Guides to Mentors: <https://gforge.nci.nih.gov/projects/guide/> . Forum contains specific compatibility criteria and requirements, as well as guidance on compatibility issues.
- NCI Center for Bioinformatics Core Infrastructure: <http://ncicb.nci.nih.gov/NCICB/infrastructure> . Home of caCORE, NCI's information technologies and services for semantics and data management.
- Cancer Data Standards Repository: [http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore\\_overview/cadsr](http://ncicb.nci.nih.gov/NCICB/infrastructure/cacore_overview/cadsr) . Provides metadata registration and management services; the caCORE component that hosts common data elements.
- Common Data Element Browser: <http://cdebrowser.nci.nih.gov>. Web application that provides CDE search, browse and retrieval capabilities.
- NCI Enterprise Vocabulary Services: <http://ncicb.nci.nih.gov/core/EVS>. Provides terminology management and development services to the cancer community, and also a component of the caCORE architecture. Jointly managed by the NCI Center for Bioinformatics and Office of Communications.
- NCI Terminology Browser: <http://nciterms.nci.nih.gov>. Web application that provides browse and search capabilities for NCI Thesaurus and other terminologies.
- NCI Metathesaurus Browser: <http://ncimeta.nci.nih.gov>. Web application that provides browse and search capabilities for NCI Metathesaurus.
- caCORE Software Development Kit: <http://ncicb.nci.nih.gov/NCICB/infrastructure/cacoresdk> . Developer tools that assist with the creation of a caCORE-like system that meets caBIG silver level compatibility guidelines.
- Semantic Integration Workbench (SIW): <http://cadsrsiw.nci.nih.gov> . The SIW is designed to facilitate and streamline the process of semantic integration—how UML metadata is mapped to EVS concept codes.
- Model-Driven Architecture: <http://www.omg.org/mda>
- Introduction to Unified Modeling Language: [http://www.omg.org/gettingstarted/what\\_is\\_uml.htm](http://www.omg.org/gettingstarted/what_is_uml.htm).
- ISO/IEC 11179 standard for Metadata Registries: <http://metadata-standards.org/11179>
- Health Level Seven: <http://www.hl7.org>
- Semantic Web: <http://www.w3.org/2001/sw>
- Simple Knowledge Organization System (SKOS): <http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/>
- Dublin Core: <http://dublincore.org/>
- OASIS: <http://www.oasis-open.org/home/index.php>

## Appendix B Glossary

<b>Term</b>	<b>Description</b>
BPEL	Business Process Execution Language. A standard orchestration-based business process modeling language that is written in XML and is executable by a BPEL engine. <a href="http://www-128.ibm.com/developerworks/library/specification/ws-bpel/">http://www-128.ibm.com/developerworks/library/specification/ws-bpel/</a>
Concept ID on caGrid	Currently, there is no one standard for representing codified content within caBIG data services. The best practice is the use of a universal unique identifier or composite identifier made up of concept code, coding scheme, and coding scheme version. Identifier use requires additional review for guaranteeing content for future uses. In the case of CDEs, NCI Thesaurus use of NCI concept code alone provides a reasonable guarantee of concept uniqueness and meaning over time. However, as the use of federated vocabularies grows, the working group recommends that all caGrid services using vocabulary content use bigIDs or composite identifier in the form of triplet. The approved recommendations of the WG are currently being implemented by the Architecture WS. <a href="https://gforge.nci.nih.gov/docman/view.php/306/7679/20070807_Arch_VCDE_F2F_ConceptID_caGrid.ppt">https://gforge.nci.nih.gov/docman/view.php/306/7679/20070807_Arch_VCDE_F2F_ConceptID_caGrid.ppt</a>
CQL	caGrid/Common Query Language. A caGrid-defined XML language used to express single data service queries. It uses a declarative approach to describe desired data by identifying the nature of the instance data with respect to its containing UML information model. That is, a query can be seen as identifying a class in a UML model, and restricting its instances to those that meet criteria defined over that class's UML attributes and UML associations. <a href="http://www.cagrid.org/mwiki/index.php?title=Data_Services:CQL">http://www.cagrid.org/mwiki/index.php?title=Data_Services:CQL</a>
DCQL	Distributed caGrid/Common Query Language. A caGrid-defined XML language used to express federated queries. It is an extension to CQL to express such concepts as joins, aggregations, and target services.
GSI	Grid Security Infrastructure. The Globus Grid Security Infrastructure is the underlying security architecture used on the grid, which is based on public key cryptography. <a href="http://www.globus.org/security/overview.html">http://www.globus.org/security/overview.html</a>
HTTP	HyperText Transfer Protocol. Standard application level protocol used for exchanging files on the World Wide Web. <a href="http://www.w3.org/Protocols/rfc2616/rfc2616.html">http://www.w3.org/Protocols/rfc2616/rfc2616.html</a>
HTTPS	HyperText Transport Protocol Secure. A standard URI scheme used to indicate a secure HTTP connection. <a href="http://tools.ietf.org/html/rfc2818">http://tools.ietf.org/html/rfc2818</a>
SOAP	Simple Object Access Protocol. A standard protocol for exchanging XML-based messages over computer networks, normally using HTTP.

<b>Term</b>	<b>Description</b>
	<a href="http://www.w3.org/TR/soap/">http://www.w3.org/TR/soap/</a>
WS-Addressing	Web Services Addressing. A standard defining XML elements to identify Web service endpoints and to secure end-to-end endpoint identification in messages. <a href="http://www.w3.org/Submission/ws-addressing/">http://www.w3.org/Submission/ws-addressing/</a>
WS-Security	Web Services Security. A standard describing enhancements to SOAP messaging to provide quality of protection through message integrity, message confidentiality, and single message authentication. <a href="http://www-128.ibm.com/developerworks/library/specification/ws-secure/">http://www-128.ibm.com/developerworks/library/specification/ws-secure/</a>
WS-SecureConversation	Web Services Secure Conversation. A standard built on top of the WS-Security and WS-Policy models to provide secure communication between services. <a href="http://www.ibm.com/developerworks/webservices/library/specification/ws-secon/">http://www.ibm.com/developerworks/webservices/library/specification/ws-secon/</a>
X.509	A standard format for describing digital certificates. <a href="http://www.ietf.org/rfc/rfc2459.txt">http://www.ietf.org/rfc/rfc2459.txt</a>
XMI	XML Metadata Interchange. An OMG standard for exchanging metadata information via XML. The most common use of XMI is as an interchange format for UML models. <a href="http://www.omg.org/technology/documents/formal/xmi.htm">http://www.omg.org/technology/documents/formal/xmi.htm</a>
XML	EXtensible Markup Language. An open standard for describing data from the W3C. <a href="http://www.w3.org/XML/">http://www.w3.org/XML/</a>
XML Schema	A formal description of a type of XML document, typically expressed in terms of constraints on the structure and content of documents of that type, above and beyond the basic syntax constraints imposed by XML itself. XML Schemas are themselves XML documents. <a href="http://www.w3.org/XML/Schema">http://www.w3.org/XML/Schema</a>