caBIG™ Compatibility Guidelines
Revision 2

July 7, 2005

The Cancer Biomedical Informatics Grid™ Program

TABLE OF CONTENTS

INTRODUCTION

Purpose

The purpose of this document is to provide the cancer Biomedical Informatics Grid™ (caBIG™) 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 between 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.

caBIG

caBIG is a voluntary network or 'grid' of individuals and institutions that are working to create a better environment for the sharing of 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, its Center for Bioinformatics, and the caBIG participants themselves.

Levels of Maturity

The caBIG community has recognized there can be differing degrees of interoperability between systems, and that these can be qualified in terms of maturity level. The caBIG Compatibility Guidelines are thus organized into four levels of maturity: Legacy, Bronze, Silver, and Gold.

- Legacy. Implies no interoperability with an external system or resource. A system that was designed without awareness of or prior to the availability of these compatibility guidelines, and which does 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 to use of a resource by a remote party who was not involved in the development of that resource.

- Gold. Currently being defined by caBIG. Is expected to provide for a formalized grid architecture and data standards that will enable standardized advertising, discovery, and use of all federated caBIG resources.

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 Architecture and Vocabulary/Common Data Elements Workspaces (VCDE) 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, thus ensuring that the program goals are met.

It is useful to consider the interoperability requirements for access independently from those for usage, though 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. Regularization 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 origins 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 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. Further, 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 arrived at four areas that need to 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:

- 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 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. When many groups use the same [common] data elements (CDEs), then larger-scale studies can be conceived, since consistency and comparability of 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 presented 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 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 being 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, not 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.

COMPATIBILITY MATRIX

This Compatibility Matrix table is a summary of caBIG compatibility requirements. Please refer to the body text of the following section for complete information on compatibility.

Maturity Model	Legacy	Bronze	Silver	Gold
Programming and Messaging Interfaces	*- No programmatic interfaces to the system are available. Only local data files in a custom format can be read* *- Data transfer mechanisms implemented only on an ad hoc basis*	*- Programmatic access to data from an external resource is possible.*	*- Well-described API's approved by the caBIG Architecture workspace provide access to data in the form of data objects that are instances of classes represented by a domain model* *- Electronic data formats reviewed and approved by the caBIG Architecture Workspace are supported for both input to and output from the system .* *- Messaging protocols approved by the caBIG Architecture Workspace are supported wherever messaging is indicated by the use cases.*	*- All features of Silver, plus:* *- Service-oriented components produce or use resources in the form of grid services that use XML as the primary interchange format.* *- Interoperable with caGrid data grid architecture being developed by caBIG Architecture Workspace- Other features to be determined by caBIG Architecture workspace*
Vocabularies / Terminologies & Ontologies	*- Free text used throughout for data collection*	*- Use of publicly accessible controlled vocabularies as well as local terminologies.* *- Terminologies must include definitions of terms that meet caBIG VCDE workspace guidelines*	*- Terminologies reviewed and validated by the caBIG VCDE Workspace used for all appropriate data collection fields and attributes of data objects.* *- Term definitions must meet VCDE Workspace guidelines.*	*- All features of Silver, plus:* *- Full adoption of caBIG terminology standards as approved by the VCDE workspace. Terminologies must be available through a caGrid service.*
Data Elements	*- No Structured metadata is recorded*	*- Data element descriptions are maintained with sufficient definitional depth to enable a subject matter expert to unambiguously interpret the contents of the resource without contacting the original investigator.* *- Data elements are built using controlled terminology* *- Metadata is stored and publicized in an electronic format that is separate from the resource that is being described..*	*- Common Data Elements (CDEs) built from controlled terminologies and according to practices validated by the VCDE workspace are used throughout.* *- CDEs are registered as ISO/IEC 11179 metadata components in the caBIG Context of the cancer Data Standards Repository (caDSR)*	*- All features of Silver, plus:* *- CDEs designated as caBIG Standards by the VCDE workspace are used* *- Metadata is advertised and discoverable via the caGrid services registry*
Information Models	*- No model describing the system is available in electronic format*	*- Diagrammatic representation of the information model is available in electronic format.*	*- 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.*	*- All features of Silver, plus:* *- Information models are harmonized across the caBIG Domain Workspaces*

COMPATIBILITY GUIDELINE DETAILS

Programming and Messaging Interfaces

The compatibility criterion of '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) will depend on the use cases and business requirements of the system being developed.

To achieve Bronze compatibility, the resource should provide, at a minimum, programmatic access to data through a public, documented API. The API needs to 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.

Silver-level compatibility is more demanding. Data-oriented systems must provide a well-documented public API that is based upon an object-oriented abstraction of the underlying data. This abstraction layer must be derived from a domain information model constructed in the Unified Modeling Language (UML; see *Information Models* below). Data must be returned in the form of data objects that are instances of classes in the model. Data formats must conform to standards set by the caBIG workspace with which the resource is aligned. Wherever use cases indicate a messaging system is warranted, a standards-based messaging protocol approved by the caBIG Architecture Workspace is used to exchange information. Silver-level analytical tools and client applications must be able to read directly from these caBIG-compatible interfaces.

Gold-level Programming and Messaging Interfaces are currently being defined by caBIG participants in the Architecture Workspace. Several decisions have been made to date: The Gold architecture will include a service-oriented data and analytical service grid with standardized service advertising and discovery features; service APIs will communicate using a specified XML syntax, and will return results as data objects that have been serialized into XML; an identifier system for data objects will be implemented across all grid data services; a grid-level security strategy will be implemented to allow for access control. The grid architecture is currently being developed by the caGrid project team in the caBIG Architecture Workspace.

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 either meeting these criteria, or a demonstration (satisfactory to the workspace) that the vocabulary is moving in the direction of meeting the 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.

At the Bronze level of maturity, the information resource utilizes public controlled vocabularies in parts of the data collection and reporting process, but may supplement them with local vocabularies. All terminologies, including those developed locally, should include definitions of terms that are sufficient to distinguish the meaning of that concept from other concepts in the terminology (a 'working definition'). At a practical level these definitions must meet the following criteria:

1. They are stated in the singular
2. They describe what the concept is, not just what it is not
3. They are stated as a descriptive phrase or sentence(s)
4. They contain only commonly understood abbreviations that are themselves defined in the terminology
5. They can be expressed without embedding definitions of other concepts (i.e. any other concepts that are required must also exist in the terminology)
6. They must not involve circular reasoning (i.e. they do not use the term in its definition)

Silver-level maturity introduces the requirement for review and approval of terminologies by the caBIG VCDE workspace. Local or private terminologies 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 if a vocabulary should be approved. Definitions in silver level vocabularies should:

1. Describe the essential nature of the concept

2. Be concise, precise and unambiguous
3. Be expressed without embedding rationale, functional usage, or procedural information
4. Use the same terminology and consistent logical structure to describe similar concepts.
5. The presence of description logic relationships to other concepts in the vocabulary may be leveraged to produce the English language definition. In any case the English language definition of the concept must not conflict with description logic relationships asserted about the concept.

As indicated above, the VCDE workspace may (at its discretion) accept a vocabulary that meets these requirements for most of its terms, or that has a clear plan for meeting these requirements.

Gold compatibility is similar to Silver, but with the added requirements that registered standards approved for caBIG-wide usage are implemented wherever they are available and that the terminology be accessible through a caGrid vocabulary service.

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 terminology sources is anticipated to be an ongoing activity.

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 fed into the next step of the information flow during the execution of a given use case.

A Data Element is a unit of metadata that describes the concept behind a given datum that is collected. Common Data Elements (CDEs) provide a means toward semantic continuity and data comparability across studies over time. 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 a research study. 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.

Bronze-level systems have their metadata structured into an electronic format that details the specification of each data element that is in the system. These metadata are constructed from the 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 is once 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. Common Data Elements 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), an implementation of the ISO/IEC11179 standard for metadata registries. Reuse of existing validated CDEs in the caDSR must be considered before any new data elements are created. All new CDEs are subject to review and validation by the VCDE workspace before they are deployed.

It is worth noting that there are two major mechanisms for creating CDEs in the caDSR. Editing tools that operate directly on the caDSR can be used by trained metadata curators to construct individual data elements and their associated components. The other alternative is to derive data elements from an information model properly constructed in UML. Such models can be submitted by caBIG projects for loading into the caDSR.

Gold requirements for data elements will likely be an extension of the Silver specification, with added requirements for usage of specific CDEs that have been approved as standards for caBIG-wide usage. Additional requirements for advertisement of service descriptions and data provenance in the caGrid architecture are also anticipated.

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 a broader semantic context, 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 Bronze-level requirement for an information model is quite modest. A diagrammatic representation of the information structure that is being produced by a system is necessary, and must be available in an electronic format.

Silver-level compatibility requires the use of the industry-standard modeling language, UML, to create domain models that describe the content of the 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 must conform to the naming conventions and terminology standards prescribed by the

caBIG program. UML models must be fully annotated with class and attribute definitions, and with associated terminology concept codes. The models must be provided in XML Metadata Interchange (XMI) format in addition to any diagrammatic representations. Upon review and validation by the VCDE workspace, models can be submitted for registration and loading into the caDSR.

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 loaded into the caDSR. 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 and adopted in caBIG. In this way, caBIG Silver requirements for Programming Interfaces can be satisfied by automatically generating model-driven middleware code.

Gold requirements for Information Models will likely involve an added degree of harmonization across caBIG domains.

ABOUT THE CABIG COMPATIBILITY GUIDELINES

These Guidelines represent a synthesis of several sources of thought, experience, tools, and practice in the areas of information systems development, data standards development and adoption, and interoperability. These sources include: Cross-cutting and Domain Workspaces from caBIG; Model-Driven Architecture from the Object Management Group; the ISO/IEC 11179 standard for metadata registries; Health Level Seven Version 3; Semantic Web from W3C; caCORE from the NCI.

Changes in Revision 2 as compared to the last revision include the following:

- Example system architecture diagrams that were intended to illustrate possible ways to deploy caBIG-compatible systems proved confusing, and distracted from the main theme of syntactic and semantic interoperability. These diagrams have been removed.
- "Interface Integration" has been renamed "Programming and Messaging Interfaces" to improve the clarity and precision of this label.
- Use of Common Data Elements registered in the caBIG Context of the caDSR is now required for Silver-level compatibility.
- Data Elements with sufficient definitional information to enable a subject matter expert to unambiguously interpret the contents of the resource are now required for Bronze-level compatibility.
- Explanatory information has been revised and reorganized according to the four areas of compatibility rather than by Bronze-Silver-Gold classification.
- Requirements for concept definitions in vocabulary sources have been enhanced and clarified.
- Initial features of the anticipated caGrid service-oriented architecture are described.

Revision 2 of the caBIG Compatibility Guidelines was edited by Peter Covitz, National Cancer Institute Center for Bioinformatics.

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.
- caBIG VCDE Workspace: http://cabig.nci.nih.gov/workspaces/VCDE. Forum for establishing and reviewing the use of caBIG data standards.
- NCI Center for Bioinformatics Core Infrastructure: http://ncicb.nci.nih.gov/core. Home of caCORE, NCI's information technologies and services for semantics and data management.
- Cancer Data Standards Repository: http://ncicb.nci.nih.gov/core/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 the 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/core/SDK. Developer tools that assist with the creation of a caCORE-like system that meets caBIG Silver-level compatibility guidelines.
- Model-Driven Architecture: http://www.omg.org/mda
- Introduction to Unified Modeling Language: 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