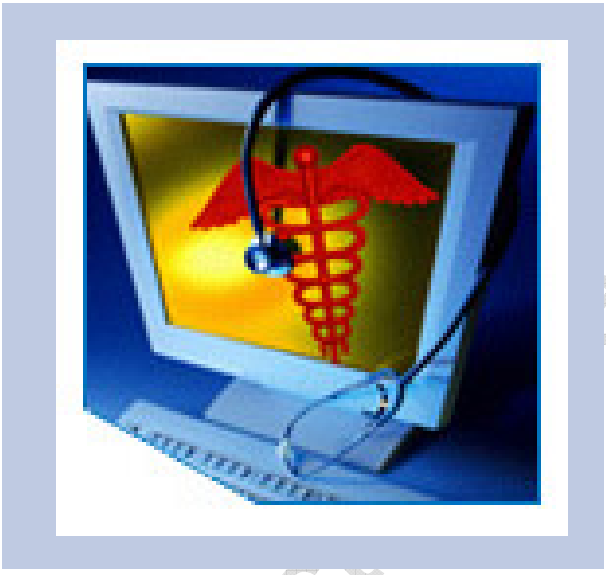
**August 19, 2006**
**Version 1.0**

# HITSP Interoperability Specification: Anonymize Component

## *HITSP/ISC- 25*



*Submitted to:*

**Healthcare Information Technology Standards Panel**

*Submitted by:*

**Biosurveillance Technical Committee**

## DOCUMENT CHANGE HISTORY

| Version Number | Description of Change | Name of Author | Date Published |
|---|---|---|---|
| 1.0 | Final draft | Biosurveillance Technical Committee | 16-August 2006 |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

# 1.0 FOREWORD

Healthcare Information Technology Standards Panel (HITSP) is a multi-stakeholder coordinating body designed to provide the process within which affected parties can identify, select, and harmonize standards for communicating healthcare information throughout the healthcare spectrum. HITSP functions as a partnership of the public and private sectors and operates with a neutral and inclusive governance model administered by the American National Standards Institute. The goal of the Panel is to:

- Facilitate the development of harmonized interoperability specifications and information policies, including SDO work products (e.g. standards, technical documents). These policies, profiles and work products are essential for establishing privacy, security and interoperability among healthcare software applications.
- Coordinate, as appropriate, with other national, regional and international groups addressing healthcare informatics to ensure that the resulting standards are globally relevant.
- Be use-case driven, utilize information from stakeholders and base its decisions on industry needs.

The HITSP shall serve the public good by working to ensure that the combined work of various healthcare information standards organizations supports interoperability, accurate use, access, privacy and security of shared health information.

In order to advance the goal of expanding harmonized interoperability specifications and information policies, HITSP was tasked with developing interoperability specifications for three main use case "breakthroughs areas" in which specific, near term value to the health care consumer could be realized. The harmonized use case areas are:

| | | |
|---|---|---|
| 1. | Biosurveillance | Transmit essential ambulatory care and emergency department visit, utilization, and lab result data from electronically enabled health care delivery and public health systems in standardized and anonymized format to authorized Public Health Agencies with less than one day lag time. |
| 2. | Consumer Empowerment | Allow consumers to establish and manage permissions access rights and informed consent for authorized and secure exchange, viewing, and querying of their linked patient registration summaries and medication histories between designated caregivers and other health professionals. |
| 3. | Electronic Health Record | Allow ordering clinicians to electronically access laboratory results, and allow non-ordering authorized clinicians to electronically access historical and other laboratory results for clinical care. |

The interoperability specification provide a detailed mapping of existing standards and specifications such as implementation guides, integration profiles to actions and actors that satisfy the requirements imposed by the relevant use cases. It identifies and constrains standards where necessary, and creates groupings of specific actions and actors to further describe the relevant contexts. Where gaps and overlaps are

identified, the interoperability specification provides recommendations and a roadmap for corrections to be made.

# 2.0 INTRODUCTION

This document, the HITSP Biosurveillance Interoperability Specification (BSV IS), defines specific implementations of established standards. These are intended to achieve integration goals that promote appropriate exchange of biosurveillance information to coordinate the optimal detection, event monitoring, and event management among health care providers, public health authorities, resource managers, and the public. This initial specification is scoped to address the population of the Biosurveillance Information System. The HITSP Biosurveillance Interoperability Specification identifies a subset of the functional components of the healthcare enterprises and health information networks called HITSP actors, and specifies their interactions in terms of a set of coordinated, standards-based transactions.
The other domains within the HITSP initiative also produce Interoperability Specifications within their respective areas that together form the HITSP Interoperability Specification. These areas include:

- HITSP Electronic Health Record
- HITSP Consumer Empowerment
- HITSP Biosurveillance
- HITSP Chronic Care

## 2.1 OVERVIEW

The Interoperability Specification focuses on a set of constrained standards for information interchange that address the core requirements of the Use Case described herein. It does not define all functions, constructs and standards necessary to implement a conforming system in a real world environment. In particular, an implementer must provide the technical infrastructure and security framework necessary to support operations in accordance with law, regulation, best practices and business agreements.

This document refers to 2006 cycle of the HITSP Biosurveillance initiative. It defines the component specification that provides the ability to Anonymize patient identifiable information.

| Related Documents | Document Description | Document Name and Location |
|---|---|---|
| HITSP/IS-01 | HITSP Interoperability Specification Biosurveillance | ISTP_HITSP_02_v0.6_2006_Biosurveillance Interoperability Specification |

## 2.2 AUDIENCE

The interoperability specification is designed to be used by analysts who need to understand the interoperability requirements for the described use case, and by implementers working to develop interoperable applications. Understanding and using the relevant interoperability set of specifications is a key requirement for establishing interoperability compliance.

## 2.3 TERMS AND DEFINITIONS

The definitions used for the purposes of this document can be found in the glossary. Refer to glossary in the appendix.

## 2.4 CONVENTIONS

This specification uses the following to convey the full descriptions and usage of standards:

**UML sequence and activity diagrams**

In these diagrams, the actors and transactions are highlighted within the framework of the specific scenario or context. The actors involved in the specified use-scenario or context are mapped out, and the interactions between each action and actor for a particular context, and the flow of data are provided through the use of arrows. Diagrams are named according to the section in which they reside, and will use the following naming convention:

Figure <section number>-<consecutive number for the diagram, e.g. 1, 2, 3, etc.>. <Short name/description of diagram>. For example, a diagram residing in section 3.1.3 showing the Actor Interactions for the Send Lab Results transaction package is named:

Figure 3.1.3-1. Send Lab Results Transaction Package

**Tables**

Tables are used to indicate standards categorizations, as well as dependencies and constraints between constructs. Tables are named according to the section in which they reside, and will use the following naming convention:

Table <section number>-<consecutive number for the table, e.g. 1, 2, 3, etc.>. <Short name/description of table>. For example, a table residing in section 2.7.1 showing the Dependencies between the transactions for the Send Lab Results transaction package is named:

Table 2.7.1-1. Send Lab Results Transaction Package dependencies

**References**

When references are made to another section within an Interoperability Specification a section number is used by itself. When references are made to other constructs that are related to the Interoperability Specification, such as Transaction Packages, Components or Composite Standards, the HITSP document short name and section number are displayed as follows:

<HITSP Document short name or Composite Standard Short Name>-<Volume Number>: <section number>

where:

<HITSP document short name> is a short designator for the construct (e.g. HITSP/ISTP-013)

<Composite Standard Short Name> is a short designator for the composite standard (e.g. IHE-ITI TF)

<Volume Number> is the applicable volume within the given composite standard (e.g. 1)

<section number> is the applicable section number (e.g. 3.1)

For example: HITSP/ISTP-013: 3.1 refers to Section 3.1 in the Interoperability Specification for a Transaction Package, IHE-ITI TF-2: 4.33 refers to Section 4.33 in volume 2 of the IHE IT Infrastructure Technical Framework.

**Reproductions**

Where large sections of composite standards or base standards are reproduced within a HITSP specification, the reproduced sections are cited with introductory text containing the reference information for the composite or base standard. In addition, the beginning and ending of the reproduced text are respectively shown using a beginning statement:

The text for the <composite or base standard name> specification begins here:

And an ending statement:

The text for the <composite or base standard name> ends here.

## 2.5 COMMENTS

To submit comments for this interoperability specification, please download the Comment Submission sheet from the HITSP site at www.hitsp.org and provide all relevant information, and then email the completed document to hitspcomments@ansi.org. Comments are consolidated periodically and sent to the Technical Committees for review.

## 2.6 COPYRIGHT PERMISSIONS



# 3.0 STANDARDS REFERENCES

The Biosurveillance Technical Committee (Bio TC) has focused its work around an analysis of the Harmonized Biosurveillance Use Case provided by the American Health Information Community (AHIC). This work has also been informed by the proceedings of the AHIC Biosurveillance Workgroup Data Steering Committee (BDSC).

The Biosurveillance TC has selected standards first in accordance with HITSP Tier 1 and Tier 2 processes. The TC worked with USHIK to evaluate the metadata and repository for use in standards selection using demographic and encounter data as a test case. The results and the resource will be used in extension of this interoperability specification to additional domains and clinical data information exchange standards.

This TC has selected standards with more options than might otherwise be defined between communication partners. As Biosurveillance is based upon secondary use of clinical data, the processes and data capture options are somewhat opportunistic, and associated data mining processes have more latitude in translation and data preparation processes. Since it is important to maximize the data sources to contribute data to the biosurveillance information system, information exchange selections include options for data capture from both legacy environments and emerging environments. Vocabulary, message, and content standards have been selected in consideration of providing the most comprehensive, machine processable fulfillment of the data requirements provided by the AHIC Biosurveillance Data Steering Committee.

## 3.1 LIST OF BASE STANDARDS

Table 3.1-1 documents standards that are referenced by this component.

| Context Standards | |
|---|---|
| Standard | Description |
| None | |
| **Information Interchange Standards** | |
| Standard | Description |
| None | |
| **Terminology Standards** | |
| Standard | Description |
| None | |
| **Security Standards** | |
| Standard | Description |
| ISO/DTS 25237 | Health Informatics – Pseudonymization |
| DICOM Supplement 55 | Attribute Level Confidentiality (including De-identification) |
| **Identifier Standards** | |
| Standard | Description |
| None | |
| **Functionality and Process/Process and Workflow Standards** | |
| Standard | Description |
| ISO/DTS 25237 | Health Informatics – Pseudonymization |
| **Legislative Standards** | |
| Standard | Description |
| HIPAA | Reference: NIH Publication Number 003-5388. 'Protecting Personal Health Information in Research: Understanding the HIPAA Privacy Rule.' |
| **Other Standards** | |
| Standard | Description |
| None | |

Table 3.1-1 List of Base Standards

### 3.2 LIST OF COMPOSITE STANDARDS

None

## 4.0 COMPONENT

Components define atomic constructs used to support an information exchange or to meet an infrastructure requirement (e.g., security, logging/audit). This is accomplished by:

*(a)* Referencing one or more underlying standards, and
*(b)* Specifying constraints and other rules for using the standards

### 4.1 CONTEXT OVERVIEW

The HIPAA regulation in 45 CFR 164.519(b) states, “a covered entity may use or disclose protected health information to the extent that such use or disclosure is required by law and the use or disclosure complies with and is limited to the relevant requirements of such law”

The HITSP Biosurveillance TC interprets the rule to permit covered entities to disclose protected health information without authorization for specified public health purposes. However, this permissiveness is not a “blank check” to request any and all data. In practice, public health supports such data requests with rational supporting justification. This commonly takes the form of state or local legislation declaring the types of data, or specific data elements required for public health. The HITSP Biosurveillance TC recommends that all implementations undergo legal review to ensure compliance with local, state, and federal regulations. The TC supports further harmonization of policy and practices for more uniform biosurveillance data exchange.

Disclosure of patient identifiable data to public health authorities in the context of reportable conditions monitoring is routine; this disclosure is based upon the need to monitor and manage known public health threats. Biosurveillance systems collect a broad variety of healthcare data that may go beyond capturing data to support assessment of known threats. As such, the TC supports the use of anonymization and pseudonymization approaches to protect individual privacy and confidentiality. This Component specifies anonymization protections for such data collection.

HIPAA defines 18 data elements that must be removed from personal health records in order for those records to be considered anonymized. The AHIC Biosurveillance Data Steering Committee has defined some demographic data elements of interest that need to be retained in order to accurately evaluate the data to detect potential threats to public health. This component specification specifies removal and aggregation requirements for data variables submitted to a Biosurveillance Information System (BIS).

ISO DTS25237 defines the following levels concepts with respect to anonymity.

Level 1 anonymity: Removal of clearly identifying data.

A first, intuitive level of anonymity can be achieved by applying rules of thumb. This method is usually implicitly understood when de-identifying data is discussed. In many contexts, this first level of anonymity may provide a sufficient guarantee.

As an example of level 1 anonymity, the HIPAA rule is given. The HIPAA rules require that for data to be considered de-identified, the following elements should be removed from the data:

- Names (individual, employer, relatives, etc.)
- Address (street, city, county, precinct, zip code – initial 3 digits if geographic unit contains less than 20,000 people, or any other geographical codes)
- Telephone and Fax numbers
- Social Security numbers
- Dates (except for years)
    - Birth date
    - Admission date
    - Discharge date
    - Date of death
    - Ages >89 and all elements of dates indicative of such age (except that such age and elements may be aggregated into a category “Age >90”)
- E-mail addresses
- Health Plan Beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle Identifiers and Serial numbers (e.g., VINs, License Plate numbers)
- Device Identifiers and Serial Numbers
- Web Universal Resource Locators (URLs)
- Internet Protocol (IP) address numbers
- Biometric Identifiers (e.g. finger or voice prints)
- Full face photographic images) and any comparable images
- Any other unique identifying number, characteristic, or code

<u>Level 2 anonymity: Static, model based re-identification risk analysis</u>

The second level of anonymity takes into account the global data model and the data flows inside the model. This level includes a static risk analysis that checks for re-identification vulnerabilities by different actors. This level may e.g. include the removal of absolute time references. A reference time marker “T” is defined as e.g. the admission of a patient for an episode of care and other events, e.g. discharge is expressed with reference to this time marker.

Level 2 anonymity issues with free-form text:

Free text and privacy are not very compatible since the notion of “free” denotes the opposite of structured. In IT terminology, the notions of “data” and “information” are treated separately. Structured data gives some indication of what information can be expected where. It is then up to re-identification risk analysis to make assumptions about what can lead to (unacceptable) identification risks, ranging from simple rules of thumb as specified in HIPAA up to analysis of populated databases and inference deductions. In “free text” as opposed to “structured”, there is no way to begin automated analysis for privacy purposes with guaranteed outcome (and the derived liabilities). “Free” and “Structured” are not necessarily black or white concepts. E.g. the presence and position of an information item in a free text document may not be predictable but when it is present, it can be deducted from a pattern, e.g. a sentence like ‘the patient had

complaints about ….." or "patient <name> was discharged at …" Simple pattern parsing or enhanced NLP can deduct structure in those cases, but perhaps not for the whole text. The notion "free" is more connected to unpredictability of presence or position of information elements. Structure is obtained by the ability to extract data elements either through fixed position, delimiters or tags. Even then, a user may e.g. put a patient number where a diagnosis should be put, but the certainty about the content is higher in structured documents. There can be a discussion on how unstructured "free text " is. Policies could e.g. define that the free text part shall not contain directly identifiable information such as patient numbers, names, or (CFR rule of thumb items such as defined in HIPAA). Parsing and natural language processing could be applied to separate directly identifying items (e.g. numbers with a certain length, structure or preamble). In some cases, the free text originates from structured text (e.g. an automated letter of discharge from a hospital generated from the hospital's HIS). This makes it easier for the parsing or NLP.

Ultimately, the primary de-identification decision is to:

- Single out what according to your policy (and desired level of identifiably) is identifiable information.
- Delete what you don't need.
- Keep together (in the so called payload) what is considered according to the policy as non-identifiable.

This is never a black and white decision, hence the need of clearer definition into levels and reference in policies. Depending on the ability to single out identifiable information (and thus to structure information), free text makes that zone very grey. The structuring this discussion should be interpreted with respect to privacy: what can lead to identification and what will not.

A hospital policy could define that investigators cannot put id information into the free text part and define what is meant by identifiable. From a privacy point of view this turns it into structured data with the payload containing free text. The liability for privacy violations is shifted towards the editor of the free text to stick to the agreed policies. From a privacy point of view the baseline on deciding if text is free is the following:

- Parts (possibly) containing identification are known
- Parties denoted as non-identifying should at least not contain nominative information.
- Hybrid situations are possible where e.g. the part with the ID is structured but the rest unstructured.

### 4.1.1 CONTEXTUAL CONSTRAINTS

This component is constrained to address the Biosurveillance AHIC Data variables subject to identification risk. With the exception of the data variables described below, all identifiers from the list of identifiable data variables defined by HIPAA that must be removed to accomplish de-identification are expected to be removed from the documents and messages that are communicated to the BIS.

***AHIC Identifiers***

Table 4.1.1-1 illustrates patient identifying data elements are defined for the Biosurveillance Data Dictionary subject to Level 1 Anonymity concerns:

| AHIC Data Variable | HIPAA Concern |
|---|---|
| Data Linker | Any other unique identifying number, characteristic, or code |
| Encounter date/time | Dates |
| Date of Birth | Dates |
| Age | Aggregate to >89 where age is >89 |
| Gender | Aggregate: Utilize only gender specifications of M/F/U |
| Zip | Aggregate to – initial 3 digits if geographic unit if Zip region contains less than 20,000 people |

Table 4.1.1-1 Patient Identifying Level 1Data Elements

Table 4.1.1-2 illustrates patient identifying data elements are defined for the Biosurveillance Data Dictionary subject to Level 2 Anonymity concerns:

| AHIC Data Variable Likely to be in the form of Freeform Text |
|---|
| Chief Complaint |
| Nurse / Triage Note |
| Test interpretation |
| Susceptibility Test interpretation |
| AHIC Data Variables Subject to Re-Identification Risk through Combination with other fields |
| Facility Code |
| Diagnosis code |
| Deceased date |
| Lab Result |

Table 4.1.1-2 Patient Identifying Level 2 Data Elements

### 4.1.2 TECHNICAL ACTORS

No technical actors are contained in this document. Application technical actors are described in higher-level specifications that incorporate this component.

## 4.2 INFORMATION INTERCHANGE COMPONENTS: RULES FOR IMPLEMENTING

The following sections document the content of the Anonymize component. It provides the basics elements and secondary standards that are supported by this component and the constraints that are being placed on those standards.

### 4.2.1 PROCESS PRE-CONDITIONS

It is a pre-condition that a message or document is generated and formatted in accordance with HITSP-defined constructs that will be sent to the BIS. It is also a pre-condition that the pseudo-identifier is already assigned in accordance with the HITSP Biosurveillance Pseudonymization Transaction specification.

### 4.2.2 PROCESS POST-CONDITIONS

The post condition is that the prepared data is rendered anonymous.

### 4.2.3 DATA FLOWS

None

### 4.2.4 DATA MAPPING

Different jurisdictions and stakeholders will have different requirements and agreements that may not require full anonymization of these data elements (See Context Overview in section 4.1 of this document).

#### 4.2.4.1 Level 1 Anonymity Considerations

To be compliant with full de-identification, all patient identifying information specified by HIPAA must be removed from the message or document to be submitted to the BIS with the exception of the data variables below. These variables would be subject to the protections specified in Table 4.2.4.2-1.

| AHIC Data Variable | Protection |
|---|---|
| Data Linker | Pseudonymized in accordance with the HITSP Biosurveillance pseudonymization transaction. Where linking across organizations is not of interest to the BIS, this may alternatively use a randomized data linker assigned by the local organization |
| Encounter date/time | Aggregate to: Month/Year only |
| Date of Birth | Aggregate to: Month/Year only |
| Age | Age >89 group |
| Gender | Use Administrative Sex, mapped to M/F/U |
| Zip | Address (street, city, county, precinct, zip code – initial 3 digits if geographic unit contains less than 20,000 people, or any other geographical codes) |
| State | NONE |

Table 4.2.4.1-1 Level 1 Patient Data Elements

#### 4.2.4.2 Level 2 Anonymity Considerations:

This section describes the Level 2 Anonymity considerations pertain to the data elements within the AHIC Data Steering Committee Data Dictionary.

Inference Risk Mitigations:

Freeform data poses a privacy and confidentiality risk because freeform text can contain identifiable information. If one would like to get value out of that data, then they need to develop methods to extract codified information from that. To be compliant with full de-identification, this approach should be applied to freeform text within the message or document to be submitted to the BIS. Based upon the AHIC Data Steering Committee Data Dictionary, the variables in Table 4.2.4.1-2 would be subject to such protections:

| AHIC Data Variable Likely to be in the form of Freeform Text |
|---|
| Chief Complaint |
| Nurse / Triage Note |
| Test interpretation |
| Susceptibility Test interpretation |
| Date and Time of illness onset |

Table 4.2.4.1-2 Risk Mitigation Data Elements

For the 2006 HITSP cycle, no stipulation is made in this specification as to the algorithms or process by which the codification is accomplished. Future specifications may address this area further, but until then, the approach is left to the implementer.

Because there are re-identification risks identified within the AHIC Data Set in combination with other fields, the information resource must have access restricted to authorized persons contractually bound or otherwise bound (and subject to sanction) to use the resource for specified purposes. This component specification recommends that access be restricted to authorized public health authorities and infection control clinical staff associated with healthcare provider organizations.

| AHIC Data Variables Subject to Re-Identification Risk through Combination with other fields |
|---|
| Facility Code |
| Diagnosis code |
| Deceased date |
| Lab Result |

Table 4.2.4.1-2 Re-Identification Risk Data Elements

For the 2005/2006 HITSP cycle, no stipulation is made in this specification with respect to access control except for the inherent mechanism provided in the functional flow scenarios described in the HITSP Biosurveillance Interoperability Specification. Future specifications may address this area further, but until then, the approach is left to the implementer.

### 4.2.5 MINIMUM DATA-SET

None

### 4.2.6 DATA PRE-CONDITIONS

It is a pre-condition that a message or document is generated with data elements subject to the constraints of this construct that will be sent to the BIS.

#### 4.2.7 DATA POST-CONDITIONS

Data is rendered anonymous.

#### 4.2.8 ADDITIONAL SPECIFICATIONS

None

## 5.0 CONSTRAINTS FOR REUSE

Any further use beyond those uses defined in the AHIC Biosurveillance use case should undergo a privacy risk assessment and assert mitigating privacy protection measures.

## 6.0 APPENDIX

This appendix provides additional detail not included in the other parts of the specification, but that are supportive of the specification.

### 6.1 HITSP HARMONIZATION FRAMEWORK

There are several constructs that are being used to define the interoperability specification, with each level providing more granularity to the standards applicable for fulfillment of the Use Case. The table below describes the current framework within which the interoperability specification is being built, the relationships between each construct, and further illustrative examples.

| | CONSTRUCT | DEFINITION | EXAMPLE | RULES |
|---|---|---|---|---|
| 1 | Use Case Harmonization Request | Defines business/functional requirements and specifies the relevant context | ONC Harmonized ONC Harmonized EHR Use Case | |
| 2 | Interoperability Specification | Models the business/functional requirements, identifies technical/system requirements to meet the specified use-case, and then identifies how to use one or more standards to meet the use-case | HITSP EHR Interoperability Specification | Based on UML diagram to identify actors and actions<br>Sets context<br>Testable functional requirements<br>Identifies transaction(s) or packages of transactions |
| 3 | Transaction Package | Defines how two or more transactions are used to support a stand-alone information exchange within a defined context between two or more systems | Record Locator Service, Entity Identification Service | Thin context and functional requirements<br>Testable<br><br>Based on analysis of like actors, context and content harmonized across the transactions<br>May be fulfilled by one or more complex standards<br><br>Expresses constraints on how the transactions are used together |
| 4 | Transaction | Logical grouping of actions, including necessary content and context, that must all succeed or fail as a group. | Query lab result, Send lab result | Fulfills all actions between two systems that meet one or more functional requirements<br>Testable<br><br>Expresses constraints on how the |

| | CONSTRUCT | DEFINITION | EXAMPLE | RULES |
|---|---|---|---|---|
| | | | | components and/or standards are used together |
| 5 | Component | An atomic construct used to support an information interchange or to meet an infrastructure requirement (e.g., security, logging/audit) | Lab result message,<br>Lab result context | Typically will use one "primary" standard and may have other "secondary" standards<br><br>May express constraints on how the standards are used |
| 6 | Base Standard | A standard capable of fulfilling a discrete function within a single category produced and maintained by a single standards organization. | Messaging standard, Security standard, Code set. | Per HITSP definition the term "standard" refers to (and is not limited to):<br>–Specifications<br>–Implementation Guides<br>–Code Sets<br>–Terminologies<br>–Integration Profiles |
| 7 | Composite Standard | Grouping of coordinated base standards, often from multiple standards organizations, maintained by a single organization. In HITSP, it can serve as a component, transaction or transaction package functional requirements. | Integration profiles<br>Implementation guides<br>Health transaction services | Per HITSP Definition |

## 6.2 GLOSSARY

The HITSP glossary that spans all the specifications can be found in the following folder on the HITSP site:
http://publicaa.ansi.org/sites/apdl/Documents/Forms/AllItems.aspx?RootFolder=http%3a%2f%2fpublicaa%2eansi%2eorg%2fsites%2fapdl%2fDocuments%2fStandards%20Activities%2fHealthcare%20Informatics%20Technology%20Standards%20Panel