Technical Overview of Version 1.0 of the CDISC ODM Model

CDISC Working Group History

The Clinical Data Interchange Standards Consortium (CDISC) originated as a "grass roots" effort in late 1997 with a vision to streamline clinical trials through data interchange standards for the biopharmaceutical industry. Two working groups were formed: the Data Modeling Group and the Glossary (Nomenclature) Group. In 1999 the data modeling group was split into two active working groups: one focusing on data and metadata standards to support electronic submissions (Submissions Data Standards-SDS and one focus on data and metadata standards to for collecting clinical data (Data Acquisition Standards-DAS Group, or DAIS Group, for Data Archive and Interchange Standards to support Data Acquisition).

In February 2000, CDISC became a non-profit organization supported by Corporate Sponsors, and CDISC was subsequently restructured to provide additional infrastructure and cohesiveness among the current and future CDISC working groups; the new structure is described in the CDISC Concept Paper (www.cdisc.org). As depicted in the Concept Paper, the original SDS and DAS/DAIS working groups will henceforth be referred to as the Submission Data Modeling (SDM) and Operational Data Modeling (ODM) working groups, respectively.

Background of the CDISC ODM Model

At the Annual DIA Meeting in June 1999, CDISC presented a metadata model to support electronic submissions. In conjunction with this event, the first meeting to address standards to support the acquisition of clinical data was held. Two models to support data acquisition for clinical research were shared openly by Phase Forward and PHT/Lincoln Technologies. The CDISC attendees were interested in further determining the implications and opportunities for achieving a common industry-wide data model to support data acquisition, and agreed to take action. At a face-to-face meeting in Chicago in September 1999, the two models were presented in more detail to this group.

The broader goals of this working group were stated as follows:

- Support interchange and archiving of data.
- Enable interchange between applications used in collecting, managing, analyzing and archiving.
- Enable full description of all data and meta-data required to produce regulatory submissions.
- Reduce costs of accumulation and conversion.

The meeting outcome was the formation of a Customer Requirements team and a Technical Analysis Team. The Customer Requirements Team was charged with surveying the potential 'customers' of the CDISC models and prioritizing the relevant requirements. The Technical Analysis Team was charged with further analyzing the two models and recommending next steps for CDISC towards developing an open, vendor-neutral data and metadata model to support data acquisition and all relevant regulatory requirements.

Results of the Customer Requirements Team

The Customer Requirements Team developed an initial survey to obtain feedback in prioritizing their efforts to achieve the DAS vision of developing standards to improve the process of acquiring and exchanging clinical trials information. Active members were R. Kush and R. Feller (NextPhase), J. Tunicliffe and L. Hauser (Quintiles) and S. Cassells (PPD Informatics). An initial survey was distributed at the DIA Electronic Data Capture Workshop in Durham, NC on November 9, 1999. Results from this survey were analyzed primarily as an exercise to determine where there was confusion on the questionnaire and how to design a superior questionnaire to prioritize the requirements. The survey was revised (formatting, evaluation scales, wording) and was distributed as a final version at the DIA Standards Workshop on November 30, 1999 in Washington, D.C. There were 48 respondents. The results were analyzed by R. Feller (NextPhase) and are summarized below.

Types of Organizations of Respondents: 30 pharmaceutical companies, 4 biotechnology companies, 4 contract research organizations, 2 academic institutions, 5 technology providers, 2 consulting companies, and 1 central laboratory.

Importance of Industry Standards: On a scale of 0-7, regarding the overall importance of standards for our industry, there were twenty-one respondents who expressed the highest level of importance (value of 7), twelve expressed a value of 6, eight expressed a value of 5 and seven did not respond to the question.

Timeframe of Availability: Two respondents stated that standards should be available in 6 months, fourteen responded 1 year, two responded 1.5 years, fourteen responded 2 years, four responded with 3 years, two with 5 years, one ASAP, one 'last year' and eight did not respond.

Priorities for Requirements for Industry Data Standards: Scores were from 0 (should not be considered) to 7 (should be given a very high priority). With 48 respondents, the highest potential total value for a given requirement would be 336.

Prioritized requirements were divided into four categories:

I. Data Standards to Facilitate Regulatory Submissions
II. Attributes of Data Standards

III. Standards to Facilitate Data Interchange
IV. Standards to Facilitate Metadata Interchange.

Of these, Category II had the highest-ranking requirements overall, followed by Categories III, I and IV in that order.

Subcategories were then prioritized within each category as follows (actual scores in parentheses – potential of 336 if every respondent marked 7):

Standards to Facilitate Regulatory Submissions (Category I):

1) CDM data to electronic regulatory submissions (257)
2) AE systems to and/or from regulatory submissions (253)
3) EDC data to electronic regulatory submissions (212)

Attributes (Category II):

1) Uphold patient confidentiality and integrity (306)
2) Adhere to global regulatory guidance documentation (290)
3) Vendor neutral; application independent (284)
4) Support the cooperation among multiple systems and technologies (273)
5) Consistent with world wide web standards (254)
6) Support 'real-time' (vs. delayed) transfer (206)

Standards to Facilitate Data Interchange (Category III):

1) Laboratory data into CDM systems (272)
2) Data from CRO to sponsoring company (271)
3) Data from AE reporting systems to and/or from CDM systems (243)
4) Data from EDC applications to CDM systems (237)
5) Data between CDM systems of different vendors (233)
6) AE information from EDC applications to CDM systems (222)
7) Laboratory data into EDC systems (220)
8) Laboratory data into data warehouses (219)
9) ePatient Diaries/self-collection tools to CDM systems (214)
10) Transfer/archiving of audit trail information (211)

Standards to Facilitate Metadata Interchange (Category IV):

1) Ensure consistency between systems (236)
2) Support data warehousing applications (227)
3) Archiving trial data long-term for audit requirements (222)

Note that there were additional requirement that received lower ratings, but none had a total number less than 180, indicating that there were no listed requirements that should not be considered in the standards effort.

Results of the Technical Analysis Team

The DAS Technical Analysis Team (TAT) was created as a small, interdisciplinary team to examine two different XML-based data interchange models (which had been separately put forward by Phase Forward and by PHT/Lincoln Technologies), specifically to assess the feasibility of developing an integrated, single CDISC standard data and metadata model to support data acquisition. The original TAT members were representatives from Amgen (Ken Harter), Domain Pharma (Toby Bloom), IBM (Angel Diaz), Oracle Clinical (John Rees), Phase Forward (Jeff Klofft), PHT (Geoff Gordon), and PPD Informatics (Sally Cassells).

Although CDISC principles state that the data content, structure, and quality of the standard data models are of paramount importance in this endeavor -- *independent of implementation strategy and platform* -- the eXtensible Markup Language (XML) was identified as a key enabling software technology for this effort. XML is gaining wide acceptance as a data interchange framework in other industries, and is already beginning to be utilized by several vendors of clinical trials software products. XML supports the definition and representation of sophisticated data models in a consistent text-based (ASCII) format -- one that can be processed conveniently by a growing set of third-party tools. However, since XML is itself really a meta-language (a language for defining industry- or domain-specific sub-languages), to make the most effective use of XML for clinical data interchange, it is necessary for pharmaceutical industry participants to agree on a consistent specific modeling approach to representing the clinical trials data and metadata in XML.

The results of the initial TAT assessment were positive, with a recommendation that a consolidated CDISC model be developed to support clinical data acquisition. A high-level overview of the scope, requirements and model was presented by Technical Analysis Team (TAT) members at the DIA Workshop on *Achieving Data Standards for Clinical Development* in Washington, D.C., November 1999. (It was at this same conference where the attendees were surveyed on requirements prioritization by the Customer Requirements Team.)

The Technical Analysis Team continued to meet in December 1999 and January 2000 to develop an initial version of the XML “Document Type Definition” (DTD) that defined the specific consolidated CDISC data interchange model. That initial version was called DTD 0.8. To facilitate examination of the content of the model by pharmaceutical industry reviewers who are not closely familiar with the XML syntax, the technical analysis team also prepared documentation of the DTD in the form of linked HTML Web pages, making use of an adapted version of the public domain *dtd2html* program. The resulting documentation provides a convenient way to “navigate” the DTD tree in both the downward (to examine contained elements) and upward (to examine containing elements) directions. Each page of the documentation contains complete information on a given element, including its content (elements contained within it), its attributes, and comment texts both for the element as a whole and for its list of attributes.

In February, the model was distributed to a multidisciplinary group of outside expert reviewers, and ten of these reviewers attended a meeting with the TAT in Boston on March 10, 2000 to provide comments and suggestions from their reviews. The model and the comments from the reviewers were presented at a CDISC meeting in conjunction with the DIA CDM Meeting in Philadelphia in March 2000. The model was then also posted openly on the CDISC website for others to review.

New members were added to the TAT following the external review. Specifically, Bob Lyons (NexTrials) and David Fram (Lincoln Technologies) played active roles beginning in March 2000. At a DAS (ODM) meeting in conjunction with the DIA Annual Conference in San Diego in June 2000, a decision was made to keep the existing TAT in place until they completed the development of several high-priority extensions to the DTD, particularly including support for an audit trail. The goal was to release Version 1.0 of the DTD by the end of the summer for external comment. This version will be revised after a period of industry comment, and expanded to address other issues such as query management and transactional data transfer in a future version to be released in 2001. Other ODM teams, including one to discuss the meaning and implications of "real-time" with respect to the model, and a Testing and Applications Group will also be formed within CDISC in the near future.

Technical Objectives for DTD 1.0

The overall goal of DTD 1.0 is to make available a first release of the definition of the CDISC DAS model, in order to support sponsors, vendors, and CRO's in the design of systems and processes around a standard interchange format.

The technical focus in the development of DTD 1.0 has been the definition of structures to represent the three major information components relating to a clinical trial:

- clinical study metadata (item definitions and protocol)
- clinical study administrative data (users and access privileges)
- clinical study data (complete record of patient data and audit trail)

This includes representation of metadata capable of supporting either direct electronic, or paper-based, data collection and capture of clinical data.

Several longer-term objectives of the CDISC DAS data model are not fully supported in the DTD 1.0 version. These additional capabilities, which will be developed in later versions of the standard, include:

- support for complex "use case" interoperability between systems (e.g., incremental data interchange, two-way data interchange)
- representation of queries and query resolution linked to clinical data
- support for long-term clinical data archive

Overview of the Model

The model provides a format for representing study metadata, study data and administrative data associated with a clinical trial. It represents only the data that would be transferred among different software systems during a trial, or archived after a trial. It need not represent any information internal to a single product, for example, information about how that data would be stored in a particular database.

An attempt has been made to make the format system-neutral. We have therefore attempted to avoid terminology that might be perceived as associated with a single system. Some of the terms in the model might therefore seem unfamiliar. We will explain the terminology as various parts of the model are explicated.

Major Sections of the Model

A top-level view of the model is shown in Figure 1 below. In this and the following diagrams, some aspects of the model (e.g., the attributes of XML elements) are not shown. The diagrams also depict the model as a single XML structure, rooted in Study, although the eventual DAS approach will permit the secure representation of the model in several different files, with the contents and links protected from tampering by digital signatures. The diagrams follow XML conventions in describing repeated and optional elements using the "*" (zero or more), "+" (one or more) and "?" (optional) prefixes.

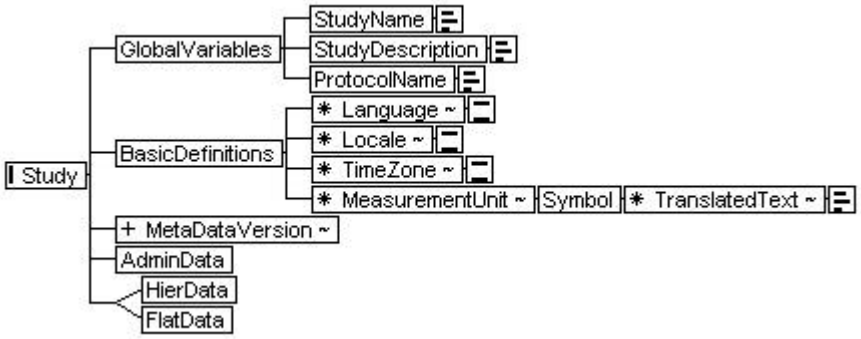

Figure 1 Major Sections of the Model

The model consists of four principal sections.

GlobalVariables contains descriptive information about the study as a whole. StudyName, StudyDescription, and ProtocolName are some examples of Global Variables.

BasicDefinitions contain definitions of information that is not likely to vary between studies. These are terms needed to define other elements within the XML DTD. They can actually be thought of as independent of the particular study, but since Study is the top-

level container we define in the DTD, these definitions are included here. As an example, TimeZone is defined here, so that every place that a time zone needs to be referenced within the study can refer to this same set of definitions. One or more TimeZones may be defined in BasicDefinitions, as indicated by the "*" preceding TimeZone. To see the attributes to be associated with each TimeZone defined, consult the full DTD definition.

The next section of the DTD is MetaDataVersion. This is the section that contains the metadata definition for the study. More details on the metadata definition are provided in the Metadata section below.

AdminData includes information about the users of the system, the clinical sites involved in the study, and associated security information.

The clinical data itself is represented either as HierData or as FlatData. The model provides alternate representations to satisfy different requirements for handling the clinical data. For example, the hierarchical model may be more intuitive for a complete archive of the data, but the flat form might be more useful for small incremental updates to data. This is described in greater detail in the Data section below.

MetaData Section of the Model

This section of the model describes study metadata. The metadata includes the *definitions* of the study data, but not the data itself. The metadata describes the structure and information in a CRF book. That includes some global information about the study, such as the study name. It also includes definitions of the visits to be scheduled within the protocol; the forms associated with each visit; and the information to be collected on each form.

If two systems are to exchange data during a clinical trial, then the metadata information may be exchanged once at the beginning of the trial, and then exchanged again each time the protocol is updated. There is no need to send the metadata each time the clinical data for a given visit is sent, for example. The overall structure of the data is the same for every patient, even though the actual data values are different for each patient, and each visit. Of course, that means that the structure must take into account the presence of repeating pages, repeating visits, and repeating groups of items. As an example, metadata exchange might be used when integrating an EDC tool with a Clinical Data Management System (CDMS) tool. Providing for metadata exchange allows the two tools to share a single design for a study protocol, so that work need not be duplicated.

The MetaData section of the model is shown in Figure 2. The metadata includes six major components:

- *Items* – clinical data items, analogous to fields or variables
- *ItemGroups* – groups of Items, analogous to datasets, panels or tables
- *Forms* – clinical data collected and signed together, analogous to pages

- *StudyEvents* – significant planned or unplanned timepoints, analogous to visits
- *Protocol* – overall data collection framework for the study
- *Presentation* – details of how data is presented and collected in forms

An Item represents an individual clinical data item, including name and datatype, along with optional associated Codelists, MeasurementUnits, and RangeCheck information. Items also include an associated label or question. Codelists, questions, and messages provide for translation into multiple languages (through the TranslatedText element that identifies Language and Locale).

An ItemGroup is a closely related set of items that is generally analyzed together. All items in an ItemGroup will have the same "repeating" behavior (for example, each row might represent a patient, a patient visit, an occurrence, etc.). In some CDM systems, ItemGroups are referred to as "panels" or "tables".

A Form is analogous to a page in a paper CRF book. It is a data capture unit logically contained within a StudyEvent, and generally collects a set of logically and temporally related information together with an electronic signature. A Form also contains an ArchiveLayout element which defines the visual appearance of the corresponding data collection form by reference to a PDF file.

A StudyEvent represents a timepoint at which clinical data is collected. In a typical outpatient study, a StudyEvent is associated with the information collected during a patient visit. StudyEvents are intended also to be usable for collection of information that is not strictly aligned with visits (such as an adverse event or a patient diary entry).

The StudyProtocol links elements of the data dictionary into the logical structure in which data are collected. The Protocol element contains a list of StudyEvents, each of which contains Forms representing one or more ItemGroups. In DTD 1.0, the StudyProtocol does not attempt to define the time schedule for, or "repeat structure" of, these elements.

The Presentation is left undefined as a vendor-specific element. It is intended to represent, at the StudyEvent, Form, ItemGroup, and Item levels, information related to forms display and data collection. For example, the graphical user interface control (e.g., text box, check box, etc.) to be used for an Item might be defined at the Item level, and the visual layout of the Items on the screen at the ItemGroup and Form levels.

Metadata can be "versioned" to support revisions in the study definition. DTD 1.0 accommodates incrementally deployed mid-study changes, such that multiple versions may be used simultaneously (perhaps due to delays in IRB approvals at different sites). An individual Study document may contain multiple MetaDataVersions, reflecting one or more mid-stream study design changes. The earliest version contains a full set of metadata, and each subsequent version contains only the necessary modified or newly-added metadata elements, along with new parent elements to complete the hierarchy.

This approach is used to allow the older versions of the protocol to remain intact (supporting clinical data collected against that version), and simultaneously provide a compact way to represent changes. Each collected piece of data contains keys (reference identifiers) that unambiguously specify which metadata version was actually used to collect it.

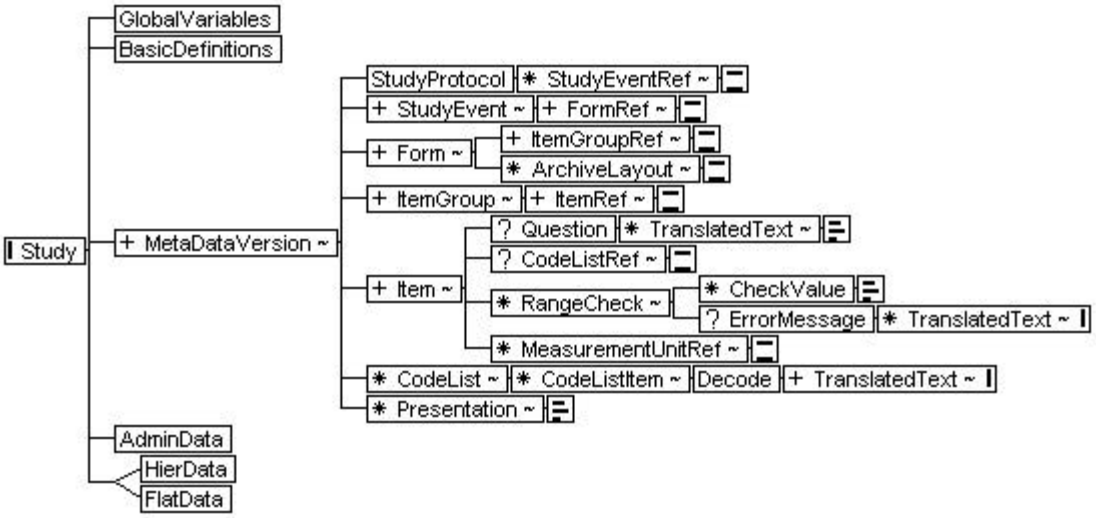

Figure 2 Metadata Elements

Figure 2. Metadata Elements

Administrative Section of the Model

The administrative section of the model is shown in Figure 3. Principal elements of the administrative section include system users, locations, and signature definitions.

System users (e.g., data entry personnel, CRA's, project managers) are defined together with identifying, contact, and electronic signature information. Subsequent to DTD 1.0, the definition of user will be extended to include an explicit role which will be linked to access permission indicators.

A set of "Locations" is also defined to represent the set of sites and other administrative locations defined in a study. In Version 1.0, explicit linking between users and locations is left as a vendor extension. The intersection of role-related permissions and defined locations will make it possible to represent data access controlled by form and by patient group.

Signature definitions (referenced from the clinical data) are also defined as a component of the administrative data. These definitions encapsulate the information necessary to describe the meaning of an electronic signature (e.g., original entry, modification, review, approval). The details of electronic signature support are defined through vendor

extension. An example of an electronic signature implementation might include digital signatures at the Form (page) level to provide a tight cryptographic connection between data values and the signer's identity, preventing unauthorized modification of data even when the XML data file is accessible. These signatures could be linked to a hierarchy of certificates, each certificate itself containing a signature manifest (fingerprint) that establishes it as issued by a more trusted authority.

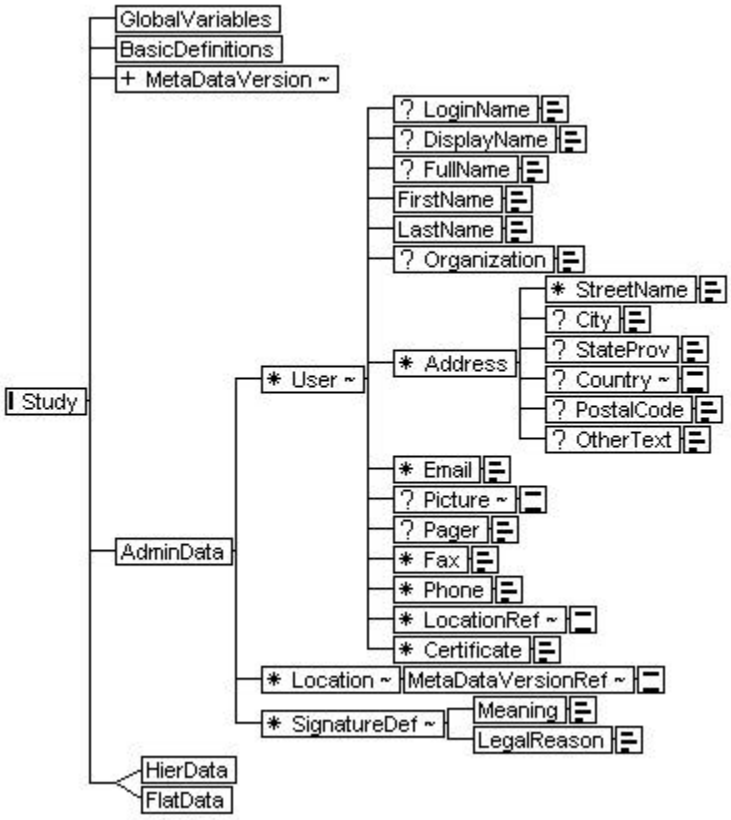

Figure 3 Administrative Elements

Clinical Data Section of the Model

Clinical data is represented according to a logical hierarchy (StudyEventData, FormData, ItemGroupData, ItemData) that parallels the metadata hierarchy (StudyEvent, Form, ItemGroup, Item).

The model allows for two different structures for representing that hierarchy. The two are called Hierarchical Data, and Flat Data. In the hierarchical data structure, the model parallels the structure for metadata. Each data element is nested, in the XML structure, within the container to which it logically belongs. In the flat data model, each item exists independently within the XML document and, in this case, the hierarchy is represented

by a set of keys on each element which reference the containing elements at each level of the hierarchy.

The hierarchical model is shown below in Figure 4a and 4b:

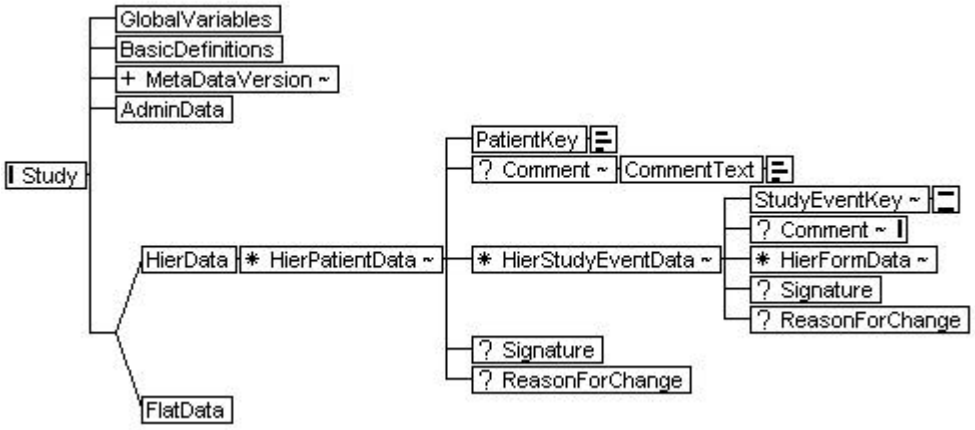

Figure 4a Hierarchical Model of Data Elements (Study through HierFormData)

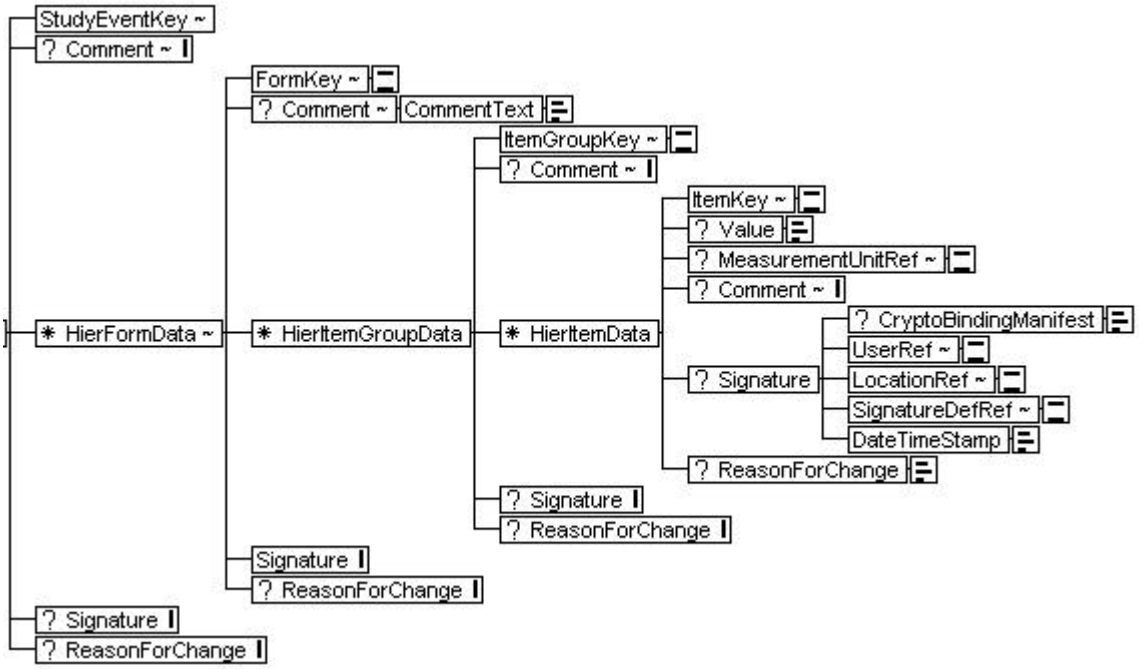

Figure 4b Hierarchical Model of Data Elements (continued from HierFormData)

The hierarchy shows PatientData, StudyEventData (information related to a visit), FormData, ItemGroupData, and finally ItemData.

An actual clinical data value would be stored in the Value element, shown in the DTD diagram above on the far right. The ItemData also includes the unit of measurement for the data. At each level, there can be a signature, a comment and/or a reason for change.

Patients are represented in the model through the HierPatientData elements, which contain basic identifying information about patients (patient ID, sex, date of birth, etc.) and a key to permit clinical data to be unambiguously associated with a patient.

Signature information (representing the electronic signing of completed forms) is potentially stored at the PatientData, StudyEventData, FormData, ItemGroupData, or ItemData levels. It can include optional cryptographic binding between user (digital certificate) and data (CryptoBindingManifest). The signature information also includes the signing user (UserRef), the location (LocationRef), the meaning of the signature (SignatureDefRef), and the date and time of signing.

A set of pre-defined keys define how the clinical data values fit within the overall study definition. These include:

- a patient key (PatientKey)
- a pair of keys at the StudyEventData (e.g., visit) level defining a specific StudyEventV instance (StudyEventID), and a possible repeat of that instance (StudyEventRepeatKey)
- a pair of keys at the FormData (e.g., page) level defining a specific Form instance (FormID), and a possible repeat of that instance (FormRepeatKey)
- a pair of keys at the ItemGroupData (e.g., panel) level defining a specific ItemGroup (ItemGroupReference), and a possible repeat of that instance (ItemGroupRepeatKey)
- a reference to a specific item (ItemRef)

In the hierarchical data model, the hierarchy is represented within the XML, and the element at each level contains its own key, so that it can be referenced from elsewhere. The keys for containing elements can be determined by the structure.

When stored in the flat format, each element must contain the full set of keys.
The flat model is shown in Figure 5 below. In this model, PatientData, StudyEventData, FormData, ItemGroupData, and ItemData appear parallel to each other in the model. The structure and nesting of the information is represented by the set of keys. For example, each FlatItemData element contains, not only the key for FlatItemData, but also the keys for every containing element in the hierarchy. The data itself is represented in the same way in both models. Values are the same, as are the measurement units, which both appear in the FlatItemData element here. Signatures, reason for change, etc. are the same in both models.

The hierarchical format might be most appropriate for the representation of a complete study in an archive, and the flat format for incremental data transfer.

The existence of two models for representing the clinical data values implies that any tool supporting the standard must be able to read both formats, although each tool is required to write to only one of the formats.

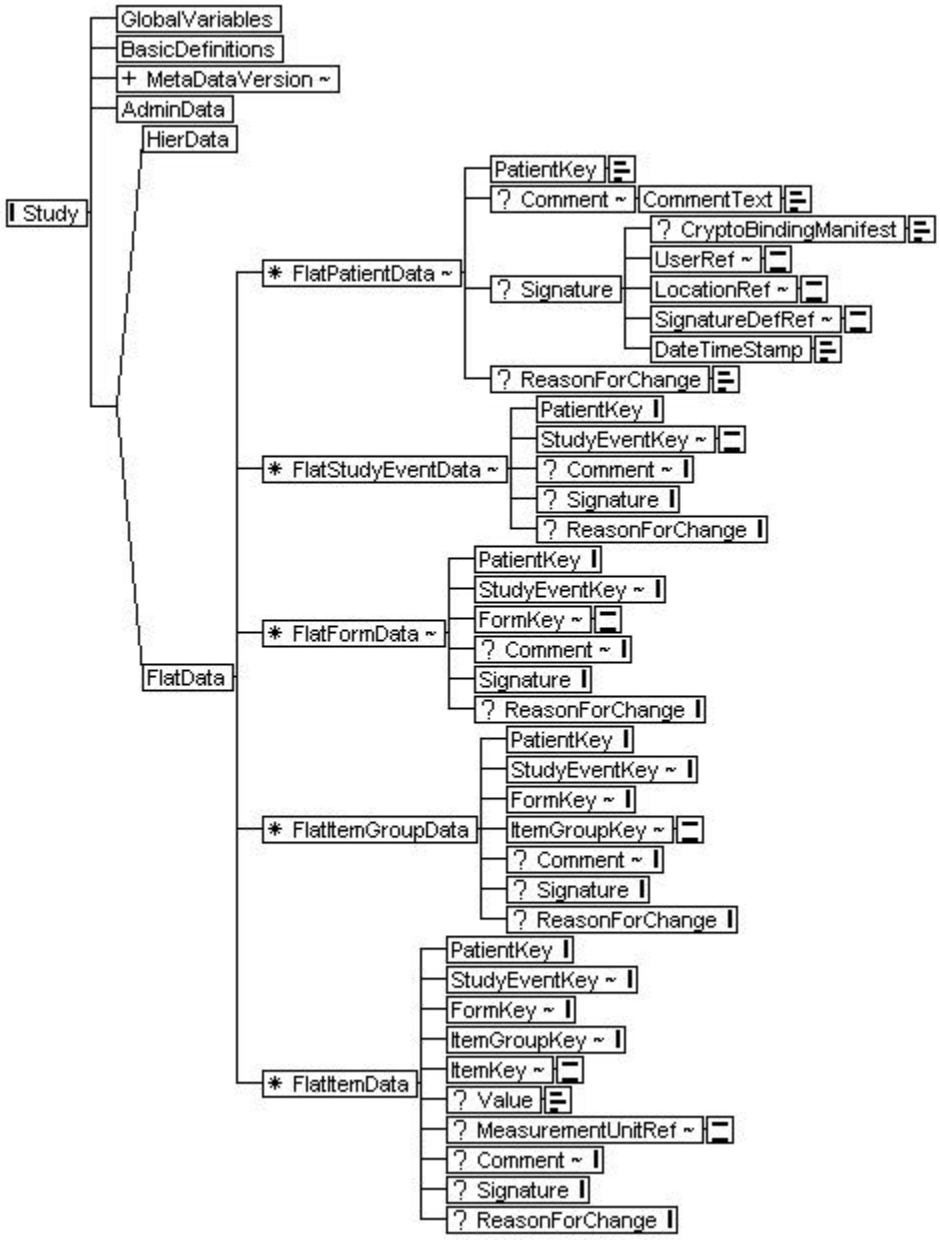

Figure 5 Flat Model for Clinical Data

Support of an audit trail is provided through preservation in the data structure (both in the flat and in the hierarchical representation) of prior versions of data elements, along with associated timestamps and signatures. Modified elements are linked to prior versions through the keys, with the order of change specified through the signature timestamps.