“The Wonderful World of Define.xml

...... Practical Uses Today”

Mark Wheeldon, CEO, Formedix

DC User Group, Washington, 9th December 2008
Agenda

- Introduction to Formedix
- What is Define.xml?
- Features and Benefits
- Define – where can it be used?
- End to End Powered by CDISC
- Define driven Dataset Design Tools
- Downstream Transformations
- Conclusions
Introduction to Formedix

Background

- Involved with CDISC Standards since 2000
- CDISC Software Products
  - Formedix Origin™ - first ODM and define.xml study design tools
  - Formedix Transform™ - EDC and paper integration products
  - Formedix Submit™ – data transformation engine
- Must Work with Existing Processes and Tools
- CDISC Consultancy Services
  - Planning, preparation and implementation
  - Cross trained in all the CDISC metadata & data models
  - Technical Director won 2 awards from ODM team
- Optimize Study Design & Downstream Data Conversions
What is Define.xml?

Overview

• Based on the ODM model
• Not to be confused with Operational Data Model which describes the content and structure of a CRF/database
• ODM is hierarchical and arranged around subjects
• Define.xml
  – Describes the content & structure of datasets
  – “Mash-up” of ODM with extensions
• SDTM datasets are arranged around findings, interventions and events NOT around subjects
• Define is for metadata
  – No corresponding data model in XML – SAS transport
<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured</td>
<td>Not like Excel. Readily re-usable.</td>
</tr>
<tr>
<td>Tabular Model</td>
<td>With standardized content = SDTM. Otherwise describe any table structure. Proprietary, AdAM, Table shells etc.</td>
</tr>
</tbody>
</table>
Define.xml

The End to End CDISC Standard

- **Study Set-up**
  - Design of datasets
  - Establishment of libraries
- **Study Execution**
  - Automation of dataset conversions
  - Automation of Testing
    - "As specified" vs. "As built"
- **Submission**
  - Lifecycles of define.xml
  - Define.xml for submission in eCTD
During Study Setup

- Protocol to Submission Process
  - What are my findings, interventions and events?
  - Identifier, Topic, Qualifiers, Timing
  - Match with data collection instrument
- Define provides detailed human readable
  - Dataset Specification
  - Dataset Variable Listings
  - Codelists and valuelists
- Use Case
  - Communicating with a CRO
  - Communicating with a Partner
During Study Execution

- Exploit machine readable specification
- Automate transformations
  - Raw to dataset
  - Proprietary dataset to proprietary dataset
  - Proprietary dataset to SDTM datasets
- Use Case
  - Streaming partner to partner
  - Partner A conducting study
    Partner B submitting
  - Delivery of SDTM datasets to sponsors and/or regulators
## Overview of End to End Powered by CDISC

<table>
<thead>
<tr>
<th>Study Start-up</th>
<th>Study Conduct</th>
<th>Analysis &amp; Reporting</th>
<th>Submission</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>LIBRARIES IN ODM</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDASH</td>
<td>Terms</td>
<td>Therapeutic</td>
<td>_libversionlibversion</td>
</tr>
<tr>
<td>Produce Database Designs</td>
<td>ODM</td>
<td>EDC</td>
<td>Data Populated ODM</td>
</tr>
<tr>
<td>CRF/eCRF</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Produce Dataset Designs</td>
<td>Define.pdf</td>
<td>CROs</td>
<td>Converter Raw Data to SDTM Datasets</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proprietary</td>
<td>SDTM</td>
<td>AdAM</td>
<td>Define.pdf</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>LIBRARIES IN DEFINE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The Value of CDISC Standards
Single Study Time Savings

8.1 months time reduction per study

Before

After

CDISC Standards Impact

Gartner

BenchMark CDISC Standards Impact

Cycle Time in Months

Analysis/Reporting
Study Conduct
Study Start-Up
The Value of CDISC Standards
Single Study Operational Savings

Total Cost Savings
Per Study = $9.0 M

Assumes $37,000 average daily out-of-pocket development cost (Tufts CSDD 2005 estimate)
**Greatest Impact When Standards Implemented Early**

80% savings when standards implemented in study start-up

- **Study Start-up (80%)**
  - Study design
  - Protocol development
  - CRF development
  - DB Structure/validation
  - Edit Checks/validation
  - LAB/ECG specs
  - Site/PI Identification
  - Training of team/sites
  - Randomization plan
  - Test article prep
  - Statistical analysis plan
  - Analysis table shells

- **Study Conduct (40%)**
  - Data acquisition
  - Data exchange
  - SD verification
  - Site monitoring/audits
  - Transfer of lab/ECG data
  - Site audits
  - Database QA and lock
  - Analysis programming
  - Initial stat tables
  - Study closeout/archive

- **Analysis & Reporting (50%)**
  - Data analysis
  - Safety assessment
  - Analysis table preparation
  - Clinical assessments
  - Report generation

- **Submission**
  - ISS/ISE preparation
  - Clinical-Statistical integrated report
  - Listings, tabulations and datasets
  - eCTD file structure modules (2-5)

*Subject participation time is excluded*

Source: Gartner Group, Sept 2006
A Define.xml Powered Dataset
Design Tool Overview

- Requires no knowledge of XML or standard
- Supports all features of SDTM Model
  - At domain level
    - Findings, Interventions, Events, SP Domains
  - At variable level
    - Value Lists
    - Supplemental Qualifiers
- Design any type of dataset structure
  - SDTM, ADaM, Proprietary, SEND, Tables
- Templates
- Libraries – drag and drop reuse
- Visualizations
- Extension capability
No more hand coding of this ...
A Dedicated Define.xml Design Environment

### Define Project: Submission Automation

**Collection:** CDISC SDS 3.1.1

- Events
  - Medical History [MH]
  - Adverse Events [AE]
- Interventions
- Findings
- Vital Signs [VS]
- Inclusion/Exclusion Exceptions
- Special Purpose
  - Comments [CO]
  - Demographics [DM]

### Define View

**Define Project**

**Collection:** CDISC SDS 3.1.1

- Events
  - Adverse Events [AE]
  - Interventions
  - Findings
- Vital Signs [VS]
- Inclusion/Exclusion Exceptions
- Special Purpose
  - Comments [CO]
  - Demographics [DM]

### Library View

**Events**

- Interventions
  - Exposure [EX]
  - Concomitant Medications
- Findings
- Special Purpose
- Relationships
  - Controlled Terms

### AE, Events, CDISC SDS Version 3.1.1, One record per adverse event per subject, Tabulation

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>Type</th>
<th>Controlled</th>
<th>Origin</th>
<th>Role</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>STUDYID</td>
<td>Study Identifier</td>
<td>text</td>
<td></td>
<td>CRF</td>
<td>Identifier</td>
<td>Context data</td>
</tr>
<tr>
<td>DOMAIN</td>
<td>Domain Abbreviation</td>
<td>text*</td>
<td></td>
<td>CL_DOMAINS</td>
<td>Identifier</td>
<td>Context data</td>
</tr>
<tr>
<td>USUBJUID</td>
<td>Unique Subject Identifier</td>
<td>text</td>
<td></td>
<td>Sponsor</td>
<td>Identifier</td>
<td>Context data</td>
</tr>
<tr>
<td>AESEQ</td>
<td>Sequence Number</td>
<td>integer</td>
<td></td>
<td>CRF or ...</td>
<td>Identifier</td>
<td>Based on the posit...</td>
</tr>
<tr>
<td>AETERM</td>
<td>Reported Term for the...</td>
<td>text</td>
<td></td>
<td>CRF</td>
<td>Topic</td>
<td>IAE_TERM</td>
</tr>
<tr>
<td>AEDEC...</td>
<td>Dictionary Derived Term</td>
<td>text*</td>
<td></td>
<td>Derived</td>
<td>Synony...</td>
<td>&quot;(no decoding)&quot;</td>
</tr>
<tr>
<td>AESEV</td>
<td>Severity/Intensity</td>
<td>text*</td>
<td></td>
<td>CRF</td>
<td>Record G...</td>
<td>IAE.SEV</td>
</tr>
<tr>
<td>AESER</td>
<td>Serious Event</td>
<td>text*</td>
<td></td>
<td>CRF or ...</td>
<td>Record G...</td>
<td>IAE.SER</td>
</tr>
<tr>
<td>AEACN</td>
<td>Action Taken with Study...</td>
<td>text*</td>
<td></td>
<td>CRF</td>
<td>Record G...</td>
<td></td>
</tr>
<tr>
<td>AEACN...</td>
<td>Other Action Taken</td>
<td>text</td>
<td></td>
<td>CRF</td>
<td>Record G...</td>
<td>CLmap of IAE.CONTRT</td>
</tr>
<tr>
<td>AEREL</td>
<td>Causality</td>
<td>text*</td>
<td></td>
<td>CRF</td>
<td>Result Q...</td>
<td>IAE. REL</td>
</tr>
<tr>
<td>AEOUT</td>
<td>Outcome of Adverse Event</td>
<td>text*</td>
<td></td>
<td>CRF</td>
<td>Result Q...</td>
<td>IAE.OUT</td>
</tr>
<tr>
<td>AESTD...</td>
<td>Start Date/Time of Adverse Event</td>
<td>date...</td>
<td></td>
<td>CRF or ...</td>
<td>Timing</td>
<td>IAE.STDTDC</td>
</tr>
<tr>
<td>AENED...</td>
<td>End Date/Time of Adverse Event</td>
<td>date...</td>
<td></td>
<td>CRF or ...</td>
<td>Timing</td>
<td>IAE.ENDTDC</td>
</tr>
<tr>
<td>AEENRF</td>
<td>End Relative to Reference</td>
<td>text*</td>
<td></td>
<td>CL.....RF</td>
<td>Derived</td>
<td>Timing</td>
</tr>
</tbody>
</table>
Template-based dataset creation
  - Four broad classes of templates supported

SDTM Template
  - Domains built from these can choose from the superset of all available GOC SDTM variables
  - Supports SDTM 1.0 and 1.1

SDTM-IG Template
  - Domains built using these can chose only from the variables defined CDISC IG
  - Enforces required, expected and permissable rules.
  - Supports SDS 3.1 and 3.1.1

Upgrade path
  - CDISC publish a new domain or new variable
  - Ship a new template
  - Supports all the new versions and the old versions
  - No need to revalidate core application
Exploiting Define.xml Benefits

Templates (2)

• User Templates
  – Just as CDISC has its core SDTM-IG for “Common to All”
  – You create new domains for each new Therapeutic Class
  – Have your own IG template for each Therapeutic Class
  – Templates will have their own core set of variables with their own rules
    • For required, expected, permissable variables
    • Controlled terms

• New way of managing your domains – not just the two letter abbreviations

• Enterprise Standards Maintenance
Exploiting Define.xml Benefits
Templates (3)

• Proprietary Domain Templates
  – Supporting existing structures & CDISC
  – Same principles as User Templates but for proprietary CRT structures
  – No adherence to CDISC at all – variables can be called anything you like
  – Templates could follow your in-house rules

• Collections
  – Domains are arranged in collections according to the template type that created them
  – Multiple collections are allowed
    • IG domains with SDTM domains
    • Proprietary domains
    • ADaM – together with SDTM or alone
Collection of Domains from Various Templates

Variables contained in the AE Domain selected
### Select Variables

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>Type</th>
<th>Role</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>LNID</td>
<td>Study Identifier</td>
<td>text</td>
<td>Identifier</td>
<td>Unique identifier for a study within the submission</td>
</tr>
<tr>
<td>DS</td>
<td></td>
<td></td>
<td>Identifier</td>
<td>Two-character abbreviation for the domain most relevant to the observation.</td>
</tr>
<tr>
<td>Notes</td>
<td>Unique identifier for a study within the submission</td>
<td></td>
<td>Identifier</td>
<td>Unique subject identifier within the submission.</td>
</tr>
<tr>
<td>LDSEQ</td>
<td>Sequence Number</td>
<td>integer</td>
<td>Identifier</td>
<td>Sequence number given to ensure uniqueness within a dataset for a subject. Can be...</td>
</tr>
<tr>
<td>LDGRPID</td>
<td>Group ID</td>
<td>text</td>
<td>Identifier</td>
<td>Used to tie together a block of related records in a single domain to support relationships within...</td>
</tr>
<tr>
<td>LDSTID</td>
<td>Specimen ID</td>
<td>text</td>
<td>Identifier</td>
<td>Internal or external specimen identifier. Example: Specimen ID.</td>
</tr>
<tr>
<td>LDSPID</td>
<td>Sponsor ID</td>
<td>text</td>
<td>Identifier</td>
<td>Optional Sponsor-defined reference number. Perhaps pre-printed on the CRF as an explicit ID...</td>
</tr>
<tr>
<td>LBTESTCD</td>
<td>LAB Test or Examination Short Name</td>
<td>text</td>
<td>Topic</td>
<td>Short name of the measurement, test, or examination described in LBTEST. It can be...</td>
</tr>
<tr>
<td>LBTEST</td>
<td>LAB Test or Examination Name</td>
<td>text</td>
<td>Synonym Qualifier</td>
<td>Verbatim name of the test or examination used to obtain the measurement or finding...</td>
</tr>
<tr>
<td>LBSCAT</td>
<td>Category for Lab Test</td>
<td>text</td>
<td>Grouping Qualifier</td>
<td>Used to define a category of related records. Examples: such as HEMATOLOGY, URINALYSIS, ...</td>
</tr>
<tr>
<td>LBSCAT</td>
<td>Subcategory for Lab Test</td>
<td>text</td>
<td>Grouping Qualifier</td>
<td>A further categorization of a test category such as DIFFERENTIAL, COAGULATION, LIVER FUNCTION, ...</td>
</tr>
<tr>
<td>LDORRES</td>
<td>Result or Finding in Original Units</td>
<td>text</td>
<td>Result Qualifier</td>
<td>Result of the measurement or finding as originally received or collected.</td>
</tr>
<tr>
<td>LDORRESU</td>
<td>Original Units</td>
<td>text</td>
<td>Variable Qualifier</td>
<td>Original units in which the data were collected. The unit for LDORRES. Example: mL.</td>
</tr>
<tr>
<td>LDORRLO</td>
<td>Reference Range Lower Limit In Orig Unit</td>
<td>text</td>
<td>Variable Qualifier</td>
<td>Lower end of reference range for continuous measurements in original units. Should be populated in the...</td>
</tr>
<tr>
<td>LDORRHI</td>
<td>Reference Range Upper Limit In Orig Unit</td>
<td>text</td>
<td>Variable Qualifier</td>
<td>Upper end of reference range for continuous measurements in original units. Should be populated in the...</td>
</tr>
<tr>
<td>LBSTRESC</td>
<td>Character Result/Finding in Std Format</td>
<td>text</td>
<td>Result Qualifier</td>
<td>Contains the result value for all findings, copied or derived from LDORRES in a standard format,...</td>
</tr>
<tr>
<td>LBSTNRC</td>
<td>Reference Range for Cha...</td>
<td></td>
<td>Variable Qualifier</td>
<td>For normal range values that are character in ordinal scale (e.g., &quot;NEGATIVE TO TRACE&quot;, 1 to +...</td>
</tr>
<tr>
<td>LBSTRESN</td>
<td>Numeric Result/...</td>
<td></td>
<td>Result Qualifier</td>
<td>Used for continuous or numeric results or findings in standard format; copied in numeric format...</td>
</tr>
<tr>
<td>LBSTRESU</td>
<td>Standard Units</td>
<td>text</td>
<td>Variable Qualifier</td>
<td>Standardized unit used for LBSTRESC or LBSTRESN.</td>
</tr>
<tr>
<td>LBSTRLO</td>
<td>Reference Range Lower Limit-Stnd Units</td>
<td>text</td>
<td>Variable Qualifier</td>
<td>Lower end of reference range for continuous measurements in standardized units. Should be populated in the...</td>
</tr>
<tr>
<td>LBSTRHI</td>
<td>Reference Range Upper Limit-Stnd Units</td>
<td>text</td>
<td>Variable Qualifier</td>
<td>Upper end of reference range for continuous measurements in standardized units. Should be populated in the...</td>
</tr>
<tr>
<td>LDRIFND</td>
<td>Reference Range Indicator</td>
<td>text</td>
<td>Variable Qualifier</td>
<td>Indicates where value falls with respect to reference range defined by LDORRLO and LDORRHI.</td>
</tr>
<tr>
<td>LDSTAT</td>
<td>Lab Status</td>
<td>text</td>
<td>Result Qualifier</td>
<td>Used to indicate exam not done. Should be null if a result exists in LDORRES.</td>
</tr>
<tr>
<td>LDRIFND</td>
<td>Reference Test Unit Name</td>
<td>text</td>
<td>Result Qualifier</td>
<td>Describes and measurement or test was not performed such as RR/INR, PO/HEMT, SGPT, SGOT, ...</td>
</tr>
</tbody>
</table>
Libraries of Reusable Dataset Designs

- Libraries of Re-usable Dataset Design Content
  - Machine readable libraries of any dataset structure
  - Proprietary datasets, SDTM, ADaM, Table Shells
- Use Define.xml to establish submission and analysis standards
  - Proprietary dataset standards (define.xml)
  - Pure CDISC SDTM & AdAM (define.xml)
  - “Common to All” Dataset standards = SDTM IG in define.xml
  - “Therapeutic specific” datasets = taken from SDTM variable superset
- Converters get you there quickly ....
  - SAS.xml, Excel, existing application metadata to
- Dataset design tools exploit XML style sheet technologies
  - One piece of metadata = multiple views
  - Tabular view of define.xml
  - Pilot used one type of stylesheet
- Dataset design tools should have an extensible visualization framework
  - Can support any format and any style
- Requires no re-validation of core product
- Client specific visualizations possible
  - By variable
  - Domains by alphabetical order
- Define.xml and Define.pdf in harmony
  - Leverages extensions and visualization framework
  - PDF capability for design tools
  - TOC, bookmarking, hyperlinks ….
Use of Extensions

- CDISC allows extensions to define.xml
- Fully CDISC compliant
- Extension framework built in - “Custom attributes”
- No revalidation of core application is required
- Extensions are possible at any level in the model
- Variety of uses
  - Templates
  - Presentation
    - Font, margins, TOC style ....
  - Company specific
    - Submission name, type, drug program
  - Advanced uses – machine readable mappings
# Typical Clinical Data Transformations

<table>
<thead>
<tr>
<th>ODM</th>
<th>CRTDDS (define.xml)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Operational DB</td>
<td>Pre CRT</td>
</tr>
<tr>
<td></td>
<td>Analysis</td>
</tr>
</tbody>
</table>

- **ODM**
  - 3/2 part & non-ISO 8601 date/time
  - Numerically coded text values
  - Text & numeric result values
  - Administrative/system variables
  - Tables, Panels, Screens, Pages
  - Non-standard variable names
  - Horizontal data structures
  - Repeating-group data structures

- **CRTDDS (define.xml)**
  - ISO 8601 date/time
  - Decoded text values
  - Text result values
  - SDTM administrative variables
  - SDTM domain & variable names
  - Vertical data structures

---

April 3-5, 2006

Oncology Clinical Trials: Strategies to Accelerate Development and Decrease the Cost of Oncology Clinical Trials
• Formal Specifications do not Exist for the Entire Study Set-up Process
• Raw Data ➔ Proprietary ➔ CDISC Standard Submission Datasets
• Downstream dataset design/specification not addressed at Study Set-up leading to a specification “gap”
Define.xml Metadata Driven Extract Transfer and Load (ETL) Process

- Define the Source (ODM or CRTDDDS)
  - Raw data – use ODM metadata
  - Your Proprietary or Legacy datasets (CRTDDDS)

- Define the Destination (CRTDDDS)
  - Any destination dataset format is supported (your datasets/SDTM)
  - Destination dataset shell description held within CRTDDDS
  - Enables multiple types and combinations of transformations
    - Raw Data ➔ Proprietary/Legacy ➔ SDTM today’s reality
    - Legacy ➔ SDTM legacy conversions
    - Raw Data ➔ SDTM ➔ Proprietary future vision
    - Raw Data ➔ SDTM future vision

- Have Your Cake and Eat It !!!
  - Working with existing “Inward Facing” and “Regulator Friendly CDISC”
Define.xml Metadata Driven Extract Transfer and Load (ETL) Process

• Define the Wiring/Mapping (Mapping Metadata)
  – Needed why? We have a wiring problem ....
  – Extended define.xml or XML language
  – Modelled from destination back to source
  – Destination data comes from
    • ODM, in-house datasets & SDTM data and metadata

• Automatic Data Transformations
  – Today manual raw data ➔ SDTM since no machine readable metadata
  – Metadata drives extract, transfer and load process = no manual coding
  – Engine uses ODM, define.xml, and ODM data

• Convert to proprietary ETL metadata to drive any ETL engine
Ultimate Flexibility
Study by Study Variability with No Coding

• Metadata controls all study to study variability
  – New define.xmls and mapping metadata
  – Study to study effort minimized by library re-use

• Transform Engine remains static
  – Structural Transformations
    – One Source ➔ Multiple Destinations
    – Multiple Sources ➔ One Destination
  – Variable Level Transformations
    – Derivations (Function calls)
    – Hard-coded strings
    – Conditional mappings (If, Then, Else)
    – Decodes
    – Codelist mappings and more ….
Downstream Raw to Submission Dataflow With Standards

Protocol Available → Interpret into Dataset Designs using Libraries

- Automatically Create Proprietary Dataset Design
- Automatically Create CDISC Dataset Design
- Automatically Create Analysis Dataset Design

Machine Readable Dataset Designs

Converter - Raw Data to Proprietary & CDISC Datasets

Machine Readable Mappings

- Standard CDISC Datasets
- Proprietary Datasets

Analysis Dataset Creation → Submission

- MachineReadable Dataset Designs
- Raw Data
- <Mapping 1>
- <Mapping 2>
- <Mapping 3>
- <Mapping 4>
- <Mapping 5>
- <Mapping 6>
- <Mapping 7>
- <Mapping 8>

- MachineReadable Dataset Designs

- Raw Data

- <Mapping 1>
- <Mapping 2>
- <Mapping 3>
- <Mapping 4>
- <Mapping 5>
- <Mapping 6>
- <Mapping 7>
- <Mapping 8>

- Standard CDISC Datasets
- Proprietary Datasets

- Analysis Dataset Creation
- Submission

Graphical representation of the dataflow process involving raw data interpretation, dataset creation, and submission.
End to End Clinical Trial Process With Standards

1. Study Starts
   - Protocol Available
   - CRF/Clinical Database Design
   - CRF Layout/ Data Capture System Build
   - Dataset Designs

2. Study Setup
   - 68% decrease
   - 55% decrease

3. Study Execution
   - Raw Data to Proprietary Datasets
   - Proprietary Datasets to Standard CDISC Datasets

4. Downstream Raw to Submission Dataflow
   - 76% decrease

5. Study Ends
   - Analysis Dataset Creation
   - Submission

Automation Introduced

Automation Introduced
Conclusions
Define.xml Delivers Today

• Save time and money throughout your clinical trial process
• Study Set-up
  – Perfect specification format for tabular structures
  – Between sponsor-CRO and development partners
  – Libraries – 20-80% tabular content reused
• Increased Quality -Validation and Testing Automation
• Downstream Raw Data to Submission Dataflow Optimisation
  – Increased quality by producing dataset and mapping specification
  – Conservative four fold reduction in time for in-house datasets
  – Gave SDTM capabilities that were not previously available