Evaluation Criteria for Transport Formats

Definitions

- Encapsulated Data - implies that the transmitted data is manipulated in a non-destructive manner with the necessary metadata in a header of a given package for sender and receiver to understand and process.
- Data Provenance - refers to the ability to trace and verify the origin of data, as well as how and by what systems it has been altered since its origination.

Data File Format

- File Size
  - Definition: For a given dataset the comparative size of the transport package (in bytes)
- Compression
  - Definition: Does the transport format support compression and decompression of encapsulated data using standard open compression formats?
- Encryption
  - Definition: Does the transport format support the encryption of encapsulated data using industry standard algorithms (including PKI)?
- Digital Signature
  - Definition: The transport format will support the application of one or more digital signatures on encapsulated dataset
- Data integrity
  - Definition: The transport format will support a hash or checksum function to mitigate unexpected data changes
- Schema driven
  - Definition: The transport format should support a schema to ensure that a data transport file will be well-formed and valid.
- Well defined Metadata
  - Definition: The transport format will support a set of well-defined metadata tags that allow effective communication of encapsulated data between sender and receiver.
  - Examples: Encryption used, record number, subject UUID, etc.
    - A use case for this would be partitioning study datasets for a into subject transfers and having enough metadata to reconstitute the original study
    - Sending partial datasets for a subject
    - Incremental or cumulative data transfers
- Wide Payload Support
  - Definition: The transport format may support transfer of a wide range of well-defined payloads over and above data currently well-described using tabular data structures.
  - Examples:
    - Image data, DICOM data, WAVEform, RDF [Ask Armando]
    - Protocol (electronic)
    - Statistical Analysis Plan (electronic)
- Relationship Data
  - Definition: The transport format will support meaningful relationships between data.
  - Example: Replace RELREC with metadata laden links for relationship between clinical observations and histopathology findings.
- Partial Data Transfers
  - Definition: The transfer format should support the transmission of subsets of data in a meaningful fashion.
  - Examples: (this should be linked with the well defined metadata)
    - Transmitting data on a subject level
    - Transmitting all data for a given time period across multiple subjects on request
    - Transmitting incremental datasets
- Must be an Open Standard
  - Definition: The full transport format specification is freely available, well documented and allowed for free use without license. All supporting materials (eg schemas, documents) will be available without cost.
- Should support multibyte character encodings
  - Definition: The transport format supports the fidelity of captured source data in transmission without requiring translation or transcoding. The encoding of a transport file should be declared by the format. Restrict support to UTF-8 encoding.
  - Example:
    - Should support submissions in Kanji for Japanese Studies
- Audit records
  - Definition: The transport format should support the transport of audit data/metadata.
  - Example:
    - The CRF-level audit trail should be able to be transported as part of an end-to-end submission
    - Something similar to the capability present in the ODM
- Traceability and Provenance
  - Definition: The transport format should support the transport of traceability data and metadata to establish data provenance.
  - Example:
    - For a given data value in a submission analysis dataset it will be possible to trace back to the original source of data.
- Transmit data and metadata
  - Definition: It will be possible to transfer both data and metadata in the same transport file.
  - Example:
    - In a given data transfer incorporate both the metadata and data, and link from data elements to corresponding metadata

Value

- Costs of adoption
  - Definition: The transport format should represent a net positive return on investment for adoption
- Resource costs - cognitive load for personnel
  - Definition: The transport format should be sufficiently familiar to not require large costs of training and utilisation
  - Example:
    - The transport format should support transport of tabular datasets
    - The transport format should be simple to build (e.g. PROC ALTRANS)
- Resource costs - storage/transport
  - Definition: The choice of the new transport format should not incur large increases in costs for processing, sending and storing data held in the format.
  - Example:
    - An substantial increase in file size would increase costs of hard drive space and bandwidth for transmitting.
    - Complex encryption mechanisms might incur a processing cost for unencrypting at each stage of data creation and review
- Resource costs - software
  - Definition: The adoption of the alternative transport format will not require a large capital outlay for software to build and manipulate the data format. It should be supportable using existing data management systems
  - Examples:
    - PROC_XPORT -> PROC_XPT++
    - Compatible with ODM systems (e.g. XML based or similar)
- Cost of Format adoption for generation and processing of clinical data.
  - Definition: The time taken to get a submission to regulators and for regulators to be able to initiate and complete review should not be impacted by the adoption of the new transport format.
  - Example:
    - There will be a minimal cost in time for generation of data in the new format, relative to the existing standard.
- Value of adoption of new transport format
  - Definition: Time to review of submission should decrease because of better expressivity and improved quality of datasets
  - Example:
    - The format should support self-validation for identification of common submission issues
    - Time spent recreating full context datasets should decrease
- Validation of capability of new format
  - Definition: Tools exist that are capable of validating the content of transport files against CDISC implementation guide rules. These rules include data format standard rules and data domain context rules. Any new transport format would need tools to product similar validation.
  - Example:
    - Value not found in non-extensible code list
    - Missing data for --STRESC when --ORRES is provided.

Content

- Definition: Changes to the content model that will deliver benefits for adopters.
- Able to represent relationships in the data without requiring duplication within a single data transfer
  - Definition: The ability to indicate relationships between elements within a encapsulated dataset. The relationship should also be able to be annotated (e.g. reason for ascribing relationship)
  - Examples:
    - Represent the causality for a given concomitant medication with respective to one or more adverse events (and vice versa)
    - Actions taken on Adverse event, for example hospitalisation
    - Findings About about an result or intervention related to the observations incurred
    - Refine model to avoid duplication of data, context, metadata
- Able to represent relationships to external resources
  - Definition: It will be possible to link encapsulated content to external resources such as standard controlled terminology
  - Examples:
    - Link to Controlled Terminology using resource URI
- Tabular Data Representation
  - Definition: Encapsulated content should support tabular representations of data
  - Example:
    - It will be possible to represent legacy datasets.
    - Transform data into tabular data structure.
- No field width restrictions
  - Definition: The transport format will support arbitrary width fields. Format should allow declaration of width for the purposes of content validation.
  - Example:
    - Data should not need to be truncated for transport
    - Data should only occupy as much space as needed (not fixed width)
- More discrete datatype definitions
  - Definition: Transfer format will support additional datatypes than existing than CHAR/NUM, eg XML Schema Definitions.
  - Examples:
    - Date
    - Time
    - Datetime
    - Datetime with timezone
    - Integer
    - Float
    - Bool
- Transactional Data Model
  - Definition: The content model will support the expression of transactional data for a data submission if requested.
  - Examples:
    - Reflect changes to data to reflect findings of a data safety monitoring board
- null Flavour support
  - Definition: The content model should support something similar to the null flavour in ISO21090 datetypes
  - Examples:
    - A missing value should have a qualifier to indicate reason for absences (eg not given, refused)
    - This is currently absent from the SDTM model

Compatibility/Extensibility

- Backward compatibility
  - Definition: New transport format will be capable of being transformed to and from existing transport format
  - Examples:
    - Decompose defined data types to CHAR/NUM
    - Truncate variable length fields to fixed length fields and SUPPQUAL
    - Translate discrete relationships to RELREC where possible
  - Note that this would not accommodate loss through UTF-8 -> US ASCII
- Compatibility with existing Health data standards
  - Definition: It should be be compatible with existing standard healthcare formats
  - Examples:
    - Transform to and from ODM (including dataset-XML)
    - Transform to and from HL7 C-CDA
    - Transform to and from BIMO (?)
- Projected Lifespan of Standard Support
  - Definition: The transport format should be supported by a non-commercial industry body with a mandate for a minimum length of time of full support for the transport format. This may depend on the age of the existing standard
  - Example:
    - Consider CDA vs FHIR, will both standards continue to exist in active development or will one supplant the other? Will the standard owner continue to maintain support and development?
- Extensibility
  - Definition: It should be able to accommodate new content requirements easily, cost-effectively, and retain backwards compatibility (i.e. no or minimal need to modify data management tools or processes). This implies support for namespaces.
  - Example:
    - Addition of custom attributes peculiar to a system adopting the standard
    - Systems naive to an extension will not be affected by use of extension