P1752 Metadata Subgroup Group Meeting

Sponsored by IEEE Engineering in Medicine & Biology (EMB) Standards Committee

26 November 2019
Teleconference
Members/Attendance

- Subgroup chair: Ida Sim, Open mHealth / UCSF
- Subgroup secretary: Anand Nandugudi, U Memphis
- Call out your name in the following order if you’re here (so we can get familiar with your voice)
  - Pradeep Balachandran
  - Jakob Bardram
  - Daniela Brunner
  - Simona Carini
  - Paul Harris
  - Shivayogi Hiremath
  - Sean McConnell
  - Leonard Njeru Njiru
  - Henry Ogoe
  - Paul Petronelli
  - Udi Rubin
  - Anna T
  - Vishnu Ravi
Action Items
From Last Meeting
Action Items from Nov 5

• Sean McConnell, Jakob Bardram: Investigate further into DatapointID
• Ida: Data Absence, refine based on today’s discussion
• Paul Petronelli: Investigate more into source_creation_datatime
• Paul Harris, Simona Carini: Initial draft of the header schema(s) (Post call)
• Unassigned: Have more examples of datapoint and datapoint series for different types of data with the current header
Metadata Ecosystem: Runtime Metadata?
DataPoint Life Cycle Discussion
for IEEE 1752
Meta Data Group

P Petronelli
11/17/19

PALM Associates, Inc.
(408) 582-8056
palm@palmcorp.com
DataPoint Life Cycle

• Information or data seems to have a life cycle which could be reflected in the metadata.

• The information cycles through the following states:
  • State 1 is measurement or captured, recorded as capture time.
  • State 2 is entered when the data has been processed or calculated. This state may occur any time after captured and before archived states.
  • When communicated to another end point, the Transmission state is entered. Although valid, this state is not relevant to mobile health.
  • State 3 is entered when the information is recorded, stored or archived.

• Although useful in analysis, the ‘state’ is probably not to be elevated to be a metadata element.

• Prior art from FHIR has a similar status for ‘content’.

• Google Cloud Platform refers to a data life cycle which may be relevant.

• Copious references exist to data life cycle in the data management literature.

• Complex, industry wide accepted data points such as DICOM are available for analysis.
Data Point Time Marking Considerations

Types of events that reflect changes in state and can be recorded:

• Creation, or collection *time*
• Data reduction or *processed* time
• *Queued* somewhere in the network
• *Archived* in central storage
• *Effective Duration* or data validity is also a potential attribute
Consider DICOM message as a complex data point example

DICOM attributes that are relevant:

• Multiple time stamps
• *Effective life* of data
• *Time Offsets* for video imaging + frame display rate
• See companion presentation on the DICOM message
Definition of a DICOM Message

- **Digital Imaging and Communications in Medicine (DICOM)** is the standard for the communication and management of medical imaging information and related data. DICOM is most commonly used for storing and transmitting medical images enabling the integration of medical imaging devices such as scanners, servers, workstations, printers, network hardware, and picture archiving and communication systems (PACS) from multiple manufacturers. It has been widely adopted by hospitals, and is making inroads into smaller applications like dentists' and doctors' offices. [1]
Relevant Characteristics

- Complex data structure with a variety of data elements
- Fields are labeled to indicate type and format
- Intended for universal PACS (Picture Archiving and Communications System) systems. A long standing market, wide usage, extensive application to healthcare.
- Data includes complicated still images and video streams. May also contain multiple videos, stills, stills + analysis from sonographers. The imagery is complex including multi-spectral, Doppler images, ECG images, sonography images & ultrasound.
- Data elements include structured data
- For videos may include imagery offsets as discussed in the Meta Data group – see frame time vector. Tag: (0018, 1065)
- Data is a mix of private and clinical data
- Images come in a variety of formats grayscale, RGB, monochrome, YBR_FULL_422. Images are digitized and stored in a multidimensional pixel array.
### Relevant data elements

<table>
<thead>
<tr>
<th>FIELD</th>
<th>DEFINITION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Related</td>
<td></td>
</tr>
<tr>
<td>Study Date</td>
<td>Analysis/processed time</td>
</tr>
<tr>
<td>Current Date</td>
<td></td>
</tr>
<tr>
<td>Content Time</td>
<td></td>
</tr>
<tr>
<td>Frame Time Vector</td>
<td>Time offsets</td>
</tr>
<tr>
<td>Effective Duration</td>
<td></td>
</tr>
<tr>
<td>Identification</td>
<td></td>
</tr>
<tr>
<td>Study ID</td>
<td></td>
</tr>
<tr>
<td>Study Instance UID</td>
<td></td>
</tr>
<tr>
<td>Series instance UID</td>
<td></td>
</tr>
<tr>
<td>Data Content</td>
<td></td>
</tr>
<tr>
<td>Pixel Data</td>
<td>Raw data</td>
</tr>
<tr>
<td>Number of Frames</td>
<td></td>
</tr>
<tr>
<td>Recommended Display Frame Rate</td>
<td></td>
</tr>
<tr>
<td>Cine Rate</td>
<td>Analysis tool</td>
</tr>
</tbody>
</table>
### DICOM Example

<table>
<thead>
<tr>
<th>Tag</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0008, 0008)</td>
<td>Specific Character Set CS: 'ISO_IR 100',</td>
</tr>
<tr>
<td>(0008, 0000)</td>
<td>Image Type CS: ['DERIVED', 'PRIMARY', 'INTRACARDIAC']</td>
</tr>
<tr>
<td>(0008, 0016)</td>
<td>SOP Class UID UI: Ultrasound Multi-frame Image Storage</td>
</tr>
<tr>
<td>(0008, 0018)</td>
<td>SOP Instance UID UI: 0000.15190412837470432708817460473337863</td>
</tr>
<tr>
<td>(0008, 0020)</td>
<td>Study Date DA: '20090109'</td>
</tr>
<tr>
<td>(0008, 0023)</td>
<td>Content Date DA: '20090109'</td>
</tr>
<tr>
<td>(0008, 0030)</td>
<td>Study Time TM: '16:00:00'</td>
</tr>
<tr>
<td>(0008, 0032)</td>
<td>Patient ID LO: '123456'</td>
</tr>
<tr>
<td>(0008, 0034)</td>
<td>Accession Number SN: '0354550059371993'</td>
</tr>
<tr>
<td>(0008, 0046)</td>
<td>Modality CS: 'US'</td>
</tr>
<tr>
<td>(0008, 0070)</td>
<td>Manufacturer LO: '12345'</td>
</tr>
<tr>
<td>(0008, 0090)</td>
<td>Referring Physician's Name MN: 'John Doe'</td>
</tr>
<tr>
<td>(0008, 0099)</td>
<td>Study Description LO: 'ACQ_TTE'</td>
</tr>
<tr>
<td>(0008, 1023)</td>
<td>Procedure Code Sequence 1 item(s) ----</td>
</tr>
<tr>
<td>(0008, 1040)</td>
<td>Patient's Sex CS: 'M'</td>
</tr>
<tr>
<td>(0010, 0010)</td>
<td>Patient's Name PN:</td>
</tr>
<tr>
<td>(0010, 0020)</td>
<td>Patient ID LO:</td>
</tr>
<tr>
<td>(0010, 0040)</td>
<td>Patient's Age DA:</td>
</tr>
<tr>
<td>(0010, 0050)</td>
<td>Patient's Birth Date DA:</td>
</tr>
<tr>
<td>(0010, 0054)</td>
<td>Patient's Sex LO:</td>
</tr>
<tr>
<td>(0010, 0064)</td>
<td>Patient Identification Removed CS: 'YES!'</td>
</tr>
<tr>
<td>(0010, 0070)</td>
<td>Race LO: 'WHITE'</td>
</tr>
<tr>
<td>(0010, 0072)</td>
<td>Effective Duration DS: '2.432414'</td>
</tr>
<tr>
<td>(0010, 0080)</td>
<td>Frame Time Vector DS: ['0.5', '1.0', '1.5', '2.0', '2.5', '3.0', '3.5', '4.0', '4.5', '5.0']</td>
</tr>
<tr>
<td>(0010, 0082)</td>
<td>Frame Time Vector DS: ['1.0', '1.5', '2.0', '2.5', '3.0', '3.5', '4.0', '4.5', '5.0', '5.5', '6.0']</td>
</tr>
<tr>
<td>(0010, 0084)</td>
<td>Frame Time Vector DS: ['1.5', '2.0', '2.5', '3.0', '3.5', '4.0', '4.5', '5.0', '5.5', '6.0', '6.5']</td>
</tr>
<tr>
<td>(0010, 0086)</td>
<td>Frame Time Vector DS: ['2.0', '2.5', '3.0', '3.5', '4.0', '4.5', '5.0', '5.5', '6.0', '6.5', '7.0']</td>
</tr>
<tr>
<td>(0010, 0088)</td>
<td>Frame Time Vector DS: ['2.5', '3.0', '3.5', '4.0', '4.5', '5.0', '5.5', '6.0', '6.5', '7.0', '7.5']</td>
</tr>
<tr>
<td>(0010, 008A)</td>
<td>Frame Time Vector DS: ['3.0', '3.5', '4.0', '4.5', '5.0', '5.5', '6.0', '6.5', '7.0', '7.5', '8.0']</td>
</tr>
<tr>
<td>(0010, 0090)</td>
<td>Transducer Data LO: ['TS-2', 'UNUSED', 'UNUSED']</td>
</tr>
<tr>
<td>(0010, 0092)</td>
<td>Processing Function LO: 'CARD_ADULT'</td>
</tr>
</tbody>
</table>
Continued

(0010, 000E) Transducer Type CS: 'SECTOR PHASED'
(0020, 000D) Study Instance UID UI: 1999.2925656638492158573184233554223530770
(0020, 000E) Series Instance UID UI: 1999.119798386509532224918159120180429391951
(0020, 0010) Study ID SH: '11'
(0020, 0011) Series Number TS: '0'
(0020, 0013) Instance Number TS: '48'
(0020, 0020) Patient Orientation CS: '1'
(0020, 0010) Private Creator LO: 'VIRETIX'
(0021, 0011) Private Creator LO: 'MINTEX_GRAPHICS'
(0026, 0002) Samples per Pixel US: 3
(0026, 0004) Photometric Interpretation CS: 'YBR_FULL'
(0028, 0006) Planar Configuration US: 0
(0028, 0008) Number of Frames TS: '75'
(0029, 0009) Frame Increment Pointer AI: (0018, 1065)
(0029, 0010) Rows US: 600
(0029, 0011) Columns US: 800
(0028, 0014) Ultrasound Color Data Present US: 1
(0025, 0109) Site Allocated US: 5
(0025, 0102) Date Stored US: 5
(0025, 0101) High Bit US: 7
(0025, 0108) Pixel Representation US: 0
(0025, 0011) Burned In Annotation CS: 'NO'
(0025, 2100) Lossy Image Compression CS: '03'
(0025, 2114) Lossy Image Compression Method CS: 'JPEG_10018_1'
(0050, 000A) Study Status ID CS: 'READ'
(0025, 0009) Study Priority ID CS: 'R'
(0025, 0250) Performed Procedure Step Start Time TM: '092950.5'
(0026, 0260) Performed Protocol Code Sequence 1 list

m(s) ----
(0028, 0010) Code Value SH: '111'
(0028, 0012) Coding Scheme Designator SH: 'Syngo'
(0008, 0004) Code Meaning LO: 'TRANSVERSEECTOR ECHO'

(0040, 0010) Requested Procedure ID SH: '167617980'
(0008, 0010) Private Creator LO: 'Phillips US Imaging BD 113'
(0008, 0011) Private Creator LO: 'Phillips US Imaging BD 109'
(0026, 0010) Pixel Data OB: Array of 105120000 bytes
References


Life Cycle

Collected/Sensed/Calculated

Processed

Queued

Archived
## State Definitions

<table>
<thead>
<tr>
<th>STATE</th>
<th>DEFINITION</th>
<th>ALTERNATE NAMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creation Time</td>
<td>Data captured by sensor</td>
<td>Capture time</td>
</tr>
<tr>
<td>Analysis Time</td>
<td>Data analyzed by algorithm or person</td>
<td>Processes time</td>
</tr>
<tr>
<td>Queued Time</td>
<td>Data stored interim somewhere in the network</td>
<td>May be multiple interim storages</td>
</tr>
<tr>
<td>Archived Time</td>
<td>Data securely stored</td>
<td>Immutable time.</td>
</tr>
<tr>
<td>Effective Duration</td>
<td>Data acquired over a time span</td>
<td>May be valid lifetime? I.e. after this time frame data validity is questionable</td>
</tr>
</tbody>
</table>
IEEE 1752 Context
Configuration Parameters

• For Physical Activity & Mobility Schema, “kcal_burned”
  • Simple formula is Energy expenditure (calories/minute) = .0175 x MET (from a reference table) x weight (in kilograms). This formula can be in a static software datasheet.
  • The weight is a parameter value that is an input to the algorithm at runtime that generates the observation/datapoint payload

• Are runtime parameter values metadata that belong in the header?
  • Where does this stop? An algorithm can take “big data” as input.
  • this parameter value can be interrogated at the system level? Like whether a datapoint is “absent” under irregular data acquisition
Datapoint ID
Datapoint versus Datapoint series: IDs

• Schema can be used for instances of arrays of observations (i.e. a series) not only a single datapoint
• Metadata must be identical for every data point in the series.
• Is a unique ID assigned to the Datapoint or each observation in the Datapoint series?

JSON arrays are ordered
Unique ID Options

• UUID (16 bytes; 32-char string)
  • At least 5 different standard versions, some including timestamp and MAC address.
  • Another implementation is GUID, which is still RFC 4122 compliant from Micro$oft.
  • It seems that v-5 is frequently preferred, since it uses SHA-1.
  • Can include a hashed namespace, which could perhaps help with Datapoint series.
    Example: AA97B177-9383-4934-8543-0F91A7A02836

• ULID (16 bytes; 26-char string)
  Example: 01BX5ZZKBKACTAV9WEVGEMMVS0

• Autoincrement-type IDs (often 8 bytes; integer)
  Example: 18446744073709551615
Considerations/Principles?

• How important is security, and the chance (however small) of being able to guess a key / ID?

• This extends to any need for lack of duplication or to avoid potential collisions –

• To what extent does the ID need to be unique, e.g., as a key, for merging data across files or datasets?

• What about the need for sorting the IDs, and time involved in storage/retrieval?

• Do the IDs (ever) need to be URL-safe?
ULID Approach

• 128-bit compatibility with UUID
• $1.21 \times 10^{24}$ unique ULIDs per millisecond
• Lexicographically sortable!
• Canonically encoded as a 26 character string, as opposed to the 36 character UUID
• Uses Crockford’s base32 for better efficiency and readability (5 bits per character)
• Case insensitive
• No special characters (URL safe)
• Monotonic sort order (correctly detects and handles the same millisecond)
Data Absence
Data Absence

• Absence is no data value of any sort available even though a value was expected, e.g.,
  • regular sampling of a data series
  • data collection was intended (e.g., per protocol)

• If data of any sort is available, then data is not absent
  • e.g., person did not respond to EMA when triggered: the response is absent, but there is a time stamp of initiation so the EMA can be labeled “missed”
  • data may be of poor quality, “poor” being with respect to a specific use
    • Insufficient data
    • Unusable data, e.g.,
      • Motion above a threshold
      • Sensor not worn
  • data may be obfuscated for privacy or other reasons and may or may not be labelled as such
Proposal for Metadata Minimum

• Objective is to support *identification* of absent data
  • Able only to support identification of absence of regularly acquired data
    • via acquisition_rate in Acquisition metadata, used only for datapoint series
    • Absence of irregularly acquired data is discoverable through system interrogation
  
• Will not represent absent data, i.e., we will not have datapoints with metadata but no data
  
• Will not capture reason for data absence or removal of data due to lack of usability
  • Non-trivial to diagnose, no one taxonomy meets all needs
  • If reason was due to privacy withhold, knowing that is a privacy leak
Data Absence – Proposed Approach

With sampling/acquisition rate and no offsets

- Explicit about start time
- Explicit about duration of effective time
  - can be represented using duration or start and end times
- Given the expected acquisition rate, it can be inferred that a value is missing

2019-08-01T07:02:00Z to 2019-08-01T07:02:59Z is missing
Drafting Metadata Schema
Timeline Proposal for Draft Review (1)

- October 15 WG call
  - INTRO: draft standard document; beginning of review period, comments due Oct 31st in iMeet Central
    - By Oct 31: distribute 1st batch of sleep schemas and the PA schema via iMeet; with comments from WG members due in iMeet Nov 19
  - By Oct 31: distribute 1st batch of sleep schemas and the PA schema via iMeet; with comments from WG members due in iMeet Nov 19

- November 5 WG call
  - DISCUSSION: draft standard document
  - INTRO: Quantitative sleep schemas (7-8) + physical activity schema, beginning of review period, comments due Nov 19
    - By Nov 14: distribute 2nd version of draft standard document
    - By Nov 21: distribute 2nd batch of quantitative sleep schemas via iMeet; comments from WG members due in iMeet Dec 13

- November 26 WG call
  - DISCUSSION: 2nd version of draft standard
  - DISCUSSION: 1st batch of quantitative sleep schemas (7-8) + physical activity schema
  - INTRO: 2ND batch of quantitative sleep schemas + qualitative sleep schemas, beginning of review period, comments due Dec 13
    - By Dec 6: distribute 2nd version of quantitative sleep schemas (7-8) + physical activity schema
    - By Dec 12: distribute metadata schema via iMeet; comments from WG members due in iMeet Jan 7
    - By Dec 13: “final” draft standard ready
Minimum Metadata: Proposal
## Metadata Elements: Datapoint

<table>
<thead>
<tr>
<th>Needs</th>
<th>Property (bold = required)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Which datapoint is this?</td>
<td>UUID (datapoint, datapoint series?)</td>
<td>Generate using RFC 4122 approach</td>
</tr>
<tr>
<td>What does this value represent?</td>
<td>schema ID and schema metadata</td>
<td>Pointer to the stress datapoint schema</td>
</tr>
<tr>
<td>When is the effective time of this data?</td>
<td>[in the datapoint itself]</td>
<td></td>
</tr>
</tbody>
</table>
## Metadata Elements: Acquisition

<table>
<thead>
<tr>
<th>Needs</th>
<th>Properties (bold = required)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>When was this datapoint first created at the (sensor) source? Recorded or packaged time.</td>
<td><strong>source_creation_datetime</strong> date-time schema represents a point in time (ISO8601). Timezone is UTC unless otherwise specified</td>
<td>2019-08-01T07:01:00Z</td>
</tr>
<tr>
<td>Was the datapoint sensed or self-reported?</td>
<td><strong>modality</strong></td>
<td>sensed</td>
</tr>
</tbody>
</table>
| If data was acquired with a periodic rate, what was the rate?         | **acquisition_rate**                                                                          | Value : 100
Unit : Hz.                                                            |
## Metadata Elements: Source

<table>
<thead>
<tr>
<th>Needs</th>
<th>Properties (bold = required)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>What firmware/algorithm?</td>
<td>Pointer(s) to <em>Software Datasheet</em>, <em>Hardware Datasheet (UDI)</em>, <em>Product Datasheet</em>, <em>Personal Datasheet (User ID)</em>, <em>Study Datasheet (Study ID)</em></td>
<td>Datasheet type {software, hardware, product, personal, study} Pointer: URI</td>
</tr>
<tr>
<td>What hardware?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>What app/product?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Which person?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Which study?</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Future Work
Outstanding Items

• Datapoint UUID – Sean, Jakob
• Source_creation_datetime – Paul P
• Draft metadata schema and other examples
• AMA Blood Pressure use case to test
Future Meetings
Upcoming Meetings

• Metadata WG
  • Tuesday, December 17: 9:00 – 10:00 AM Pacific
Adjournment