IEEE STANDARDS ASSOCIATION

IEEE

P1752 Metadata Subgroup Group Meeting

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

17 December 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
  - Christina Caraballo
  - Simona Carini
  - Paul Harris
  - Shivayogi Hiremath
  - Sean McConnell
  - Leonard Njeru Njiru
  - Henry Ogoe
  - Paul Petronelli
  - Udi Rubin
  - Anna T

Action Items
From Last Meeting

Action Items from Nov 26

- F/U with Brian Page on DatapointID [Ida, Sean]
- Summarize runtime metadata approach [Ida]
- Initial draft of the header schema(s) [Paul H, Simona]
- AMA BP use case example [Ida]

Metadata Ecosystem: Runtime Metadata?

Metadata Landscape

Open mHealth

data and metadata standards

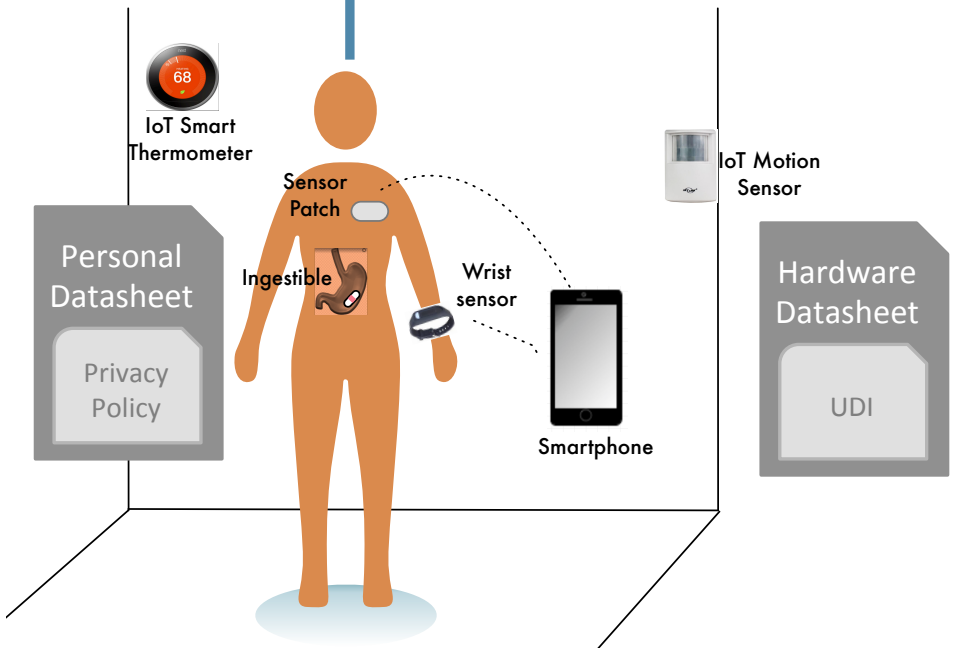
Software Datasheet

Dataset & ML Details

Data stream(s) → Firmware/ algorithm →

```
{
  Header/min metadata
}
{
  Body: datapoint
}
```

Cloud or Edge: Runtime acquisition and privacy metadata



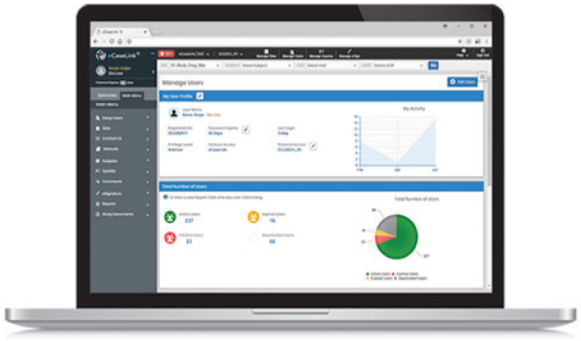
Product Datasheet

Privacy Policy

Electronic Health Record

Study Protocol

Data Sharing Policy

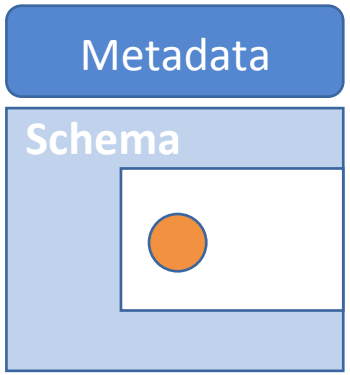


Clinical Trial Management System

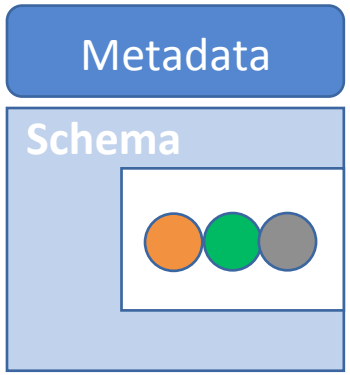
Datapoint ID

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?



Datapoint



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

ULID Approach

- 128-bit compatibility with UUID
- 1.21e+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)

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?

Data Absence

Data Absence – Proposed Approach

With sampling/acquisition rate and no offsets

```
{
    "header": {
        "id": "123e4567-e89b-12d3-a456-426655440000",
        ”acquisition_rate”: {
              ”value”: 1/60,
            ”unit”: “hz”
        },
        [...]
    },
    "body": {
        "stress_values": [
            {
                "probability": 0.75
                 "start_date_time": "2019-08-01T07:00:00Z",
                 ”duration”: {
                    ”value”: 1,
                    ”unit”: “min”
                 }
            },
            {
                "probability": 0.85
                "start_date_time": "2019-08-01T07:01:00Z",
                 ”duration”: {
                    ”value”: 1,
                    ”unit”: “min”
                 }
            },
            {
                "probability": 0.80
                "start_date_time": "2019-08-01T07:03:00Z",
                 ”duration”: {
                    ”value”: 1,
                    ”unit”: “min”
                 }
            }
        ]
    }
}
```

2019-08-01T07:02:00Z” to 2019-08-01T07:02:59Z” is missing

- 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

Drafting Metadata Schema

Draft header.json

Draft schema.json

“uri” and “uri-reference”

- "uri": A universal resource identifier (URI), according to RFC3986.
- "uri-reference": A URI Reference (either a URI or a relative-reference), according to RFC3986, section 4.1.
- "iri": The internationalized equivalent of a “uri”, according to RFC3987.
- "iri-reference": The internationalized equivalent of a “uri-reference”, according to RFC3987
- https://json-schema.org/understanding-json-schema/reference/string.html#format

$id

- The $id property is a URI that:
  - declares a unique identifier for the schema
  - declares a base URI against which $ref URIs are resolved
  - https://json-schema.org/understanding-json-schema/structuring.html#the-id-property
- Currently not used

Mininum Metadata: Proposal

Metadata Elements: Datapoint

Needs	Property (bold = required)	Example
Which datapoint is this?	**UUID (datapoint, datapoint series?)**	**Generate using RFC 4122 approach**
What does this value represent?	**schema ID and schema metadata**	**Pointer to the stress datapoint schema**
When is the effective time of this data?	[in the datapoint itself]	

Metadata Elements: Acquisition

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

Metadata Elements: Source

Needs	Properties (bold = required)	Example
What firmware/algorithm? What hardware? What app/ product? Which person? Which study?	Pointer(s) to *Software Datasheet, Hardware Datasheet (UDI), Product Datasheet, Personal Datasheet (User ID), Study Datasheet (Study ID)*	Datasheet type {software, hardware, product, personal, study} Pointer: URI

Future Work

Items from Schema Review Calls

- Filtering
- Flag identifying raw data

Outstanding Items

- Datapoint UUID – Sean, Jakob
- Source_creation_datetime – Paul P
- Draft metadata sample data examples

Future Meetings

Upcoming Meetings

- Metadata WG
  - Tuesday, January 7 or January 14: **9:00 – 10:00** AM Pacific

Adjournment