Exposing Public Health Surveillance Data Using Existing Standards

Turbelin C, Boëlle PY

1) UMR S 707, INSERM, Pierre et Marie Curie University, Paris, France
2) APHP, Saint Antoine Hospital, Paris, France
Context

- Public Health Surveillance (PHS):
  - ongoing systematic data collection, analysis and interpretation followed by dissemination to stimulate public health action [1]

- An increasing number of data sources
  - Health care system (GPs, Hospital, ED)
  - Others sources (++)
    - Population-based surveillance, Search query logs (Google), social networks (Twitter), school absenteeism, climate data, ...
    - Useful for Emerging problems, reduce timeliness [2], complete existing surveillance systems [3]

- PHS data: data produced by PHS systems
  - Usually transformed data (aggregated, filtered,...)
    - Privacy issues, extract relevant information using statistical treatments, ...

Why sharing PHS data?

- An emerging research question & challenge
- Enable multi-source surveillance
  - Combine data sources to increase robustness or scope
- Evaluate & compare systems (improve Public Health)
- Requirement for publication?
- Sharing will need standards
  - Excel files, no metadata, few documentation, human based
- Example of ‘-omics’ domain
  - “Minimal Information About” standards family .. (~ 20 formats)
Source data: French Sentinelle Data

- PHS system based on general practitioners (1)
  - started in 1984, specific data collection.
- Publishing weekly estimation for 10 health conditions at national and regional levels (aggregated data only)

(1) http://www.sentiweb.fr
Existing standards

- Literature and web search to identify standards
  - Health Domain
    - HL7 based standards (health system related data)
    - Clinical Data Interoperability Standard Consortium (clinical trial)
  - Statistical data
    - Statistical Data and Metadata Exchange –Health Domain (SDMX-HD)
    - Data Document Initiative (DDI)
  - General purpose exchange protocol
    - OPeN Data Access Protocol (developed by NASA)
    - netCDF (UniData)
    - Google Dataset Publishing Language
    - OData
- Focused on general purpose and statistical standards
  - OData and SDMX-HD
SDMX - Health Domain standards

- Proposed by the WHO, specialization of SDMX standard
- Based on an information model SDMX-IM
  - Implementation of the metadata registry (ISO 11179)
  - Describes data elements using concept and value domain
  - Describes data organisation (Data Structure Definition)
- XML and EDI syntax

![Diagram showing relationships between MSD, DSD, Concept, Metadata Structure, and Data Structure]

- MSD: Metadata Structure Definition
  - Describes structure of Metadata Structure
  - Contains metadata
- DSD: Data Structure Definition
  - Describes structure of Data Structure
  - Contains data
SDMX-HD DataSet example (1)

GEOGRAPHICAL_PLACE_NAME
France, national level, All ages group

AGROUP
All ages group

TIME_PERIOD
Week

BOUNDARIES

FREQ
2008-W01
2008-W02
2008-W03

OBS_VALUE
2008-W01
167
144
190
262
237
287
424
380
468

INFORMATION
Influenza Like Illness, Incidence rate per 100,000 inhabitants

ORGANIZATION
French Sentinelles Network

All concepts were picked from SDMX-HD predefined common concepts

In Metadata:

HEALTH_EVENT*
Influenza like Illness, ...

INDICATOR_TYPE*
Incidence, incidence rate, ...

* new concepts
SDMX-Health Domain: DataSet example

Concepts:
- AGROUP
- ORGANIZATION
- FREQ
- INDICATOR
- BOUNDARIES
- GEOGRAPHIC_PLACE_NAME
- TIME_PERIOD
- OBS_VALUE

Data Structure Definition

Compact DataSet Structure

Data Structure (Compact DataSet)
XML representation

```xml
<?xml version="1.0" encoding="UTF-8" ?>
<CompactData>
  <DataSet CALENDAR_TYPE="ISO8601" ORGANIZATION="SENTIWEB"
           FREQ="W" AGROUP="/ALL" VALUE_TYPE="1">
    <Series INDICATOR="31" GEOGRAPHIC_PLACE_NAME="FR" BOUNDARIES="2">
      <Obs TIME_PERIOD="2012-W11" OBS_VALUE="128945"/>
      <Obs TIME_PERIOD="2012-W10" OBS_VALUE="174704"/>
      <Obs TIME_PERIOD="2012-W09" OBS_VALUE="225648"/>
    </Series>
    <Series INDICATOR="3TI" GEOGRAPHIC_PLACE_NAME="FR" BOUNDARIES="2">
      <Obs TIME_PERIOD="2012-W11" OBS_VALUE="203"/>
      <Obs TIME_PERIOD="2012-W10" OBS_VALUE="275"/>
      <Obs TIME_PERIOD="2012-W09" OBS_VALUE="355"/>
    </Series>
  </DataSet>
</CompactData>
```
Exposing data using SDMX-HD

- Get and visualize SDMX-HD XML files
- Visualize data (from SDMX-HD files)
- Upload your SDMX-HD package

SDMX-HD Data Structure implementation

SDMX-HD Data visualization
OData Protocol

- Generic data access protocol
  - Proposed by Microsoft Corp., proposed at OASIS
  - Based on web standards: HTTP, AtomPub, JSON
- Web-service based access to a database
  - REST architecture: URI is the query
    - identify and access data using the URI
- Data are feed of Collections of Typed entries (EntityType)
  - Instance of EntityType (class)
  - Described in an Entity Data Model (~ relational DB)
- Rich ecosystem
OData Service overview

Service Document

List of available entitySets in the Service

Metadata Document

EntitySet Document

CSDL : Conceptual Schema Definition Language
Exposing OData service

- Browse OData webservice (XML or JSON)
- Visualize Entity Data Model and time series from the service
# Standards comparison

<table>
<thead>
<tr>
<th>Properties</th>
<th>SDMX-HD</th>
<th>OData</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standardization</strong></td>
<td>ISO Technical Standard</td>
<td>Based on W3C standards</td>
</tr>
<tr>
<td><strong>Information Model</strong></td>
<td>SDMX-IM</td>
<td>Entity Data Model</td>
</tr>
<tr>
<td><strong>Syntax</strong></td>
<td>XML, EDI</td>
<td>XML (AtomPub), JSON</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Complex</td>
<td>Simple</td>
</tr>
<tr>
<td></td>
<td>Any object, hierarchical</td>
<td>(relational, in the data model)</td>
</tr>
<tr>
<td><strong>Common Concepts</strong></td>
<td>Cross-domain concepts</td>
<td>No</td>
</tr>
<tr>
<td><strong>Encourage harmonization</strong></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Sharing model</strong></td>
<td>Centralized (registry) or federated</td>
<td>Decentralized</td>
</tr>
<tr>
<td><strong>Domain specific implementation</strong></td>
<td>Yes, proposal (SDMX-HD)</td>
<td>No</td>
</tr>
<tr>
<td><strong>Archiving feature</strong></td>
<td>Yes (SDMX-HD standalone packages)</td>
<td>No</td>
</tr>
<tr>
<td><strong>Authoring Tools</strong></td>
<td>Several</td>
<td>Several</td>
</tr>
<tr>
<td><strong>End-user tools</strong></td>
<td>Few Libraries for java, .NET</td>
<td>Many Libraries for &gt;10 languages</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Excel plugin</td>
</tr>
</tbody>
</table>
Summary & Perspectives

• Limits
  • Simple data model
  • Open Data perspective (not designed for a specific usage)
  • SDMX-HD sustainability

• Two kinds of standards
  • SDMX-HD => Interoperability, cost of implementation
  • OData => Accessibility, everyday tools integration.

• Existing initiatives to standardize Public Health data
  • PHDSC (EHR-Public Health) (Orlova, 2004)
  • Population Health Record (popHR) (Friedman, 2010)

• Web Semantic & Linked Data
  • Epidemic Marketplace (modeling project) : NERO ontology
  • W3C Data Cube Vocabulary, based on SDMX (2013)