The use of HL7 Clinical Document Architecture schema to define a data warehouse dimensional model for secondary purposes

Fabrizio Pecoraro, Daniela Luzi, Fabrizio L. Ricci

Institute for Research on Population and Social Policies
National Research Council, Rome – Italy
Background #1

Integration of data provided by heterogeneous information systems developed for different specialties and purposes and by different organizations in a Data Warehouse for secondary purposes.

This makes it necessary to develop specific tools to improve interoperability.
Integration of information system has been addressed in healthcare to improve semantic interoperability

Data modelling

HL7 represents one of the main candidates for the exchange of information among different stakeholders mainly used for patient’s care delivery purposes (i.e. primary uses)

In our vision,

HL7 standards and in particular the Clinical Document Architecture (CDA) can be the basis to define a common schema to represent clinical information in a data warehouse
Objective

Propose a semi-automatic approach to extract information from XML document structured using the CDA standard and transform them to be stored in a data warehouse.

Adopt the conceptual framework that maps the HL7 CDA components (HL7 Hierarchy, Act) with the Dimensional Model concepts (Fact, Dimension) (F. Pecoraro, D. Luzi, F.L. Ricci, 2015) based on the dimensional model lifecycle proposed by Kimball (R. Kimball, M. Ross, 2009).
Map the HL7 CDA components (HL7 Hierarchy, CDA Backbone) with the dimensional model concepts (Fact, Dimension)
Mapping procedure: Dimensional model lifecycle

1. Choose the business process
2. Declare the grain
3. Identify the Fact & Measures
4. Identify the Dimensions
5. Refine the dimensional model
Identify the Fact

Identify a CDA concept that represents a "measurement of the healthcare business process".

Acts that define the CDA Backbone are suitable candidates to identify Fact.

An action performed to determine an answer or a result value.
Identify the Measures

Attributes of the chosen Act are used to identify measures of the Fact table.

- `value` and `interpretationCode` attributes that represent a quantitative and qualitative description of the event observed.
Identify Dimensions

Dimensions are determined by answering the following questions:

- Who participates (person): patient, provider, responsible parties
- What is studied (fact): encounter, hospitalization, adverse reaction
- When has been performed (time): date, year, month, week-day
- Where has been placed (location): facility, hospital, patient’s home
- Why has been performed (reason): pathology, adverse reaction
- How has been measured (manner): visit, hospitalization

(Zachman framework, Inmon et al., 1997)

Our approach,

Identify CDA concepts able to answer these questions

(CDA Act classes, HL7 Hierarchies, Classes and Attributes)
Identify Dimensions: CDA Act classes

Healthcare service provided over a period of time

ClinicalDocument and Section can be included in the model to provide low-level of details additional information.
Identify Dimensions: HL7 Hierarchies

Performer (who): the physician(s) that carried out the clinical event

RecordTarget (who): represents the medical record that this document belongs to

Both are related with the “scoper” Entity Organization through a Role.
This class can be used to identify where the observation has been placed
Participant (who):  
- Authenticator: A verifier who legally authenticates the accuracy of an act (who)  
- Consultant: An advisor participating in the service by performing evaluations and making recommendations  
- Responsible party: the provider (person or organization) who has primary responsibility for the act
Specimen (what):

- part of some entity, typically the subject, that is the target of focused laboratory, radiology or other observations
- used when observations are made against some substances or objects that are taken or derived from the subject
Other Dimensions: HL7 Hierarchies

• location (where): healthcare facility where the event occurred

• consumable (what): substance taken up or consumed as part of the administration

• product (what): a material target that is brought forth (e.g. dispensed) in the service

• author (who): the humans and/or machines that authored the document

• informant (who): is a person that provides relevant information, such as the parent of a comatose patient who describes the patient’s behavior prior to the onset of coma
Identify Dimensions: Act attributes

- **code (what):** particular kind of event/act (e.g. hemoglobin, white blood cell, platelets, electrolytes)
- **effectiveTime (when):** time when the event has been performed
- **statusCode (how):** state of the clinical statement (e.g. completed, cancelled, active, nullified)
- **methodCode (how):** technique used to make an observation (e.g. ultrasound, x-ray, computed axial tomography)

**Degenerate Dimensions**

dimension keys in the fact class that is not related to a dimension table
Refine the Dimensional Model

The design of a dimensional model based on the CDA elements results in a high-level normalized data model.

This representation is typically adopted in transactional database where an high volume of transactions (insert, update, delete) is performed.

Conversely, analytical processing is characterized by a low volume of transactions (insert) with complex queries to be executed.
Refine the Dimensional Model

Resolving attributes’ Data Types

**Observation**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>classCode: CS [1..1]</td>
<td>M = OBS</td>
</tr>
<tr>
<td>code: CD [0..1]</td>
<td>U &lt; CD:ActCode</td>
</tr>
<tr>
<td>derivationExpr: [0..1]</td>
<td>U</td>
</tr>
<tr>
<td>effectiveTime: [0..1]</td>
<td>U</td>
</tr>
<tr>
<td>id: [0..*]</td>
<td>U</td>
</tr>
<tr>
<td>interpretationCode: [0..*]</td>
<td>U &lt; CD:Observation</td>
</tr>
<tr>
<td>languageCode: CD [0..1]</td>
<td>U &lt; CD:HumanLang</td>
</tr>
</tbody>
</table>

**SET<CE>**

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>code:</td>
<td>string</td>
</tr>
<tr>
<td>codeSystem:</td>
<td>uid</td>
</tr>
<tr>
<td>codeSystemName:</td>
<td>string</td>
</tr>
<tr>
<td>codeSystemVersion:</td>
<td>string</td>
</tr>
<tr>
<td>displayName:</td>
<td>string</td>
</tr>
</tbody>
</table>

**Solution #1**

A separate table is created with the five elements as independent attributes.

The table is linked with a one-to-many relationship with the Observation to store multiple valued attribute.

**Solution #2**

The five elements are included in the Observation class as independent attributes.

Only the first entry of the interpretationCode is collected.
1) Generation of the transformation template based on the conceptual framework

2) Transformation of the CDA document to a XML document

3) Loading of the XML document in the data Warehouse
Implementation of the Conceptual Framework

1) Generation of the transformation template based on the conceptual framework

2) Transformation of the CDA document to a XML document

3) Loading of the XML document in the data Warehouse
1. Find all nodes that match with the class chosen as Fact of the dimensional model

2. For each node the XML tree is navigated in both directions:
   1. parent-to-child and
   2. child-to-parent
Each child of the Fact class is included as child of the Fact element along with its children and attributes.
Each ancestor of the Fact class is included as child of the Fact element along with its children and attributes.
Transformation Template (XSLT) document
The Act class Observation has been chosen to represent the Fact table of the dimensional model.
<observation classCode="OBS" moodCode="EVN">
  <!-- children and attributes of the original observation -->
  <id root="107c2dc0-67a5-11db-bd13-0800200c9a66"/>
  <observation-code-bridge>
    <code-code>4548-4</code-code>
    <code-displayName>Hgb Alc MFr Bld</code-displayName>
    <code-codeSystem>2.16.840.1.113883.6.1</code-codeSystem>
    <code-codeSystemName>LOINC</code-codeSystemName>
  </observation-code-bridge>
  <observation-value-bridge>
    <value-type>PQ</value-type>
    <value-value>7</value-value>
    <value-unit>%</value-unit>
  </observation-value-bridge>
  <effectiveTime-value>20100725100000</effectiveTime-value>
  <observation-interpretationCode-bridge>
    <interpretationCode-code>N</interpretationCode-code>
    <interpretationCode-codeSystem>2.16.840.1.113883.5.83</interpretationCode-codeSystem>
  </observation-interpretationCode-bridge>
  <referenceRange>
    <observationRange>
      <value>
        <low-value>4</low-value>
        <low-unit>%</low-unit>
        <high-value>4</high-value>
        <high-unit>%</high-unit>
      </value>
    </observationRange>
  </referenceRange>
  <!-- ancestors of the original observation -->
  <ClinicalDocument>
    <!-- attributes and children of this element -->
  </ClinicalDocument>
  <Section>
    <!-- attributes and children of this element -->
  </Section>
  <recordTarget>
    <!-- recordTarget hierarchy -->
  </recordTarget>
  <documentationOf>
    <serviceEvent classCode="PCPR">
      <performer typeCode="PRF">
        <!-- performed hierarchy -->
      </performer>
    </serviceEvent>
  </documentationOf>
</observation>
<ClinicalDocument>
   <!-- CDA Header -->
   <recordTarget>
      <!-- recordTarget hierarchy -->
   </recordTarget>
   <documentationOf>
      <serviceEvent classCode="PCPR"/>
      <performer typeCode="PRF"/>
      <!-- performed hierarchy -->
   </performer>
   <observation classCode="OBS" moodCode="EVN">
      <!-- id root = "107c2dc0-67a5-11db-bd13-0800200c9a66" -->
      <code type="CE" code="4548-4" displayName="Hgb Alc MFr Bld"
            codeSystem="2.16.840.1.113883.6.1" codeSystemName="LOINC"/>
      <effectiveTime value="20100725100000"/>
      <value type="PQ" value="6.6" unit="%"/>
      <interpretationCode code="N" codeSystem="2.16.840.1.113883.5.83"/>
      <referenceRange>
         <observationRange>
            <value xsi:type="IVL_PQ">
               <low value="4" unit="%"/>
               <high value="5.6" unit="%"/>
            </value>
         </observationRange>
      </referenceRange>
   </observation>
   <!-- /referenceRange -->
   <!-- /observation -->
   <!-- /component -->
   <!-- /organizer -->
   <!-- /entry -->
   <!-- /section -->
   <!-- /component -->
   <!-- /structuredBody -->
   <!-- /component -->
</ClinicalDocument>
<ClinicalDocument>
  <!-- CDA Header -->
  <recordTarget>
    <!-- recordTarget hierarchy -->
  </recordTarget>
  <documentationOf>
    <serviceEvent classCode="PCPR">
      <performer typeCode="PRF">
        <!-- performed hierarchy -->
      </performer>
    </serviceEvent>
  </documentationOf>
  <!-- CDA Body -->
  <component>
    <structuredBody>
      <component>
        <section>
          <entry typeCode="DRIV">
            <organizer classCode=
            <component>
<!-- ancestors of the original observation -->
<ClinicalDocument>
  <!-- attributes and children of this element -->
</ClinicalDocument>

<Section>
  <!-- attributes and children of this element -->
</Section>

<recordTarget>
  <!-- recordTarget hierarchy -->
</recordTarget>

<documentationOf>
  <serviceEvent classCode="PCPR">
    <performer typeCode="PRF">
      <!-- performed hierarchy -->
    </performer>
  </serviceEvent>
</documentationOf>

</observation>

</root>
Implementation of the Conceptual Framework

1) Generation of the transformation template based on the conceptual framework

2) Transformation of the CDA document to a XML document

3) Loading of the XML document in the data Warehouse
XML native data warehouse

- **Fact.xml**
  - `<root>`
  - `<Observation>`

- **Attributes of the Fact**
  - @
  - ...
  - @

- **Attributes to identify the relevant dimensions**
  - *

- **recordTarget.xml**
  - `<root>`
  - `<recordTarget>`
  - *

All the facts are stored in a XML with additional attributes to identify the relevant dimensions.

Each dimension is stored in a specific XML.
First results

• Source: set of 682 CCD documents provided within the repository of sample available on the web. 
  [https://github.com/chb/sample_ccdas](https://github.com/chb/sample_ccdas)

• Business process analyzed: quality of care of patients with diabetes
  • Process indicator: Proportion of patients who have carried out at least three check of hemoglobin glycated in the last year
  • Outcome (intermediate) indicator: Proportion of patients who have the value of hemoglobin glycated under 7%
  • Outcome (final) indicator: incidence of diabetic retinopathy
Conclusions

The use of HL7 CDA elements to define a dimensional model has many advantages:

- Makes the integration of heterogeneous systems more robust limiting the ambiguity in the semantics of messages and avoiding the proliferation of dialects

- Makes data transformation a simplified straightforward process thus reducing the resources to be invested to implement ETL tools

- Enhances the extensibility of the data warehouse architecture facilitating the integration of other systems in the future
Issues

Issues related to the adoption of CDA

- Data representation: incorrect information, terminology misuse, inappropriate identifiers, inclusion/omission of optional elements, data redundancy
  
  \[(D’Amore et al., 2011 and 2014).\]

- Data structures: data types to represent information as well as coding vocabularies can vary among countries and structures
  
  \[(http://www.ringholm.com/docs/03020_en_HL7_CDA_common_issues_error.htm)\]

- Although the CDA has applied a set of templates to develop a document, a set of “local representations” can be identified depending on the implementation rules adopted
  
  \[(Dolin et al., 2001 & Flores, 2010)\]

- Identification of patients and other actors (e.g. providers) can vary among hospitals/regions/countries considering local representations.
Issues related to the adoption of CDA for secondary purposes

- An event (e.g. lab exam) can be captured in many documents using different representations (redundancy).
- Variability in the structure of documents with particular attention on the Act relationship (CDA backbone).
- Automatic comparison of variables measured using different methods and procedures.
- Secondary use of data is based only considering the structured information and not the narrative part.
- Information are analysed using “semi-automatic” software, so that the error and heterogeneity listed in the previous slide would not be easily detected.
Future works

The feasibility of this framework should be demonstrated on the basis of:

- Documents provided by different organizations
- Different types of document capturing the care pathway of an individual, for instance the prescription, the booking and the execution of a specific exam.

DATA! DATA!! DATA!!!
Thank you for your attention

Fabrizio Pecoraro
f.pecoraro@irppps.cnr.it