The IEEE Metadata Standard for Supporting Big Data Management

Alex MH Kuo[1,2] (Ph.D)

[1] School of Health Information Science
University of Victoria, BC, Canada.
[2] CEDAR, School of Medicine
University of Stanford, USA.
Email: akuo@uvic.ca

Outline

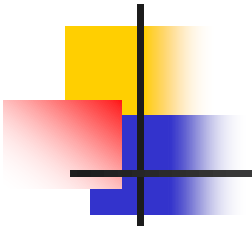
- Background
- Current status of metadata standards
- What we want to do?
- Discussion

1. Background

- The characteristic of Big Data
  - A collection of data so large, so complex, so distributed and growing so fast (or 6Vs- volume, variety, velocity, veracity, value, variability and veracity) that it becomes very difficult to maintain and analyze using commonly used database management systems and traditional data analysis applications.

Health Big data
- Structured EHR Data
- Unstructured Clinical Notes
- Medical Imaging Data
- Genetic Data
- Other Data (Epidemiology & Behavioral)

Other type Big data
- Transitional/Lab data



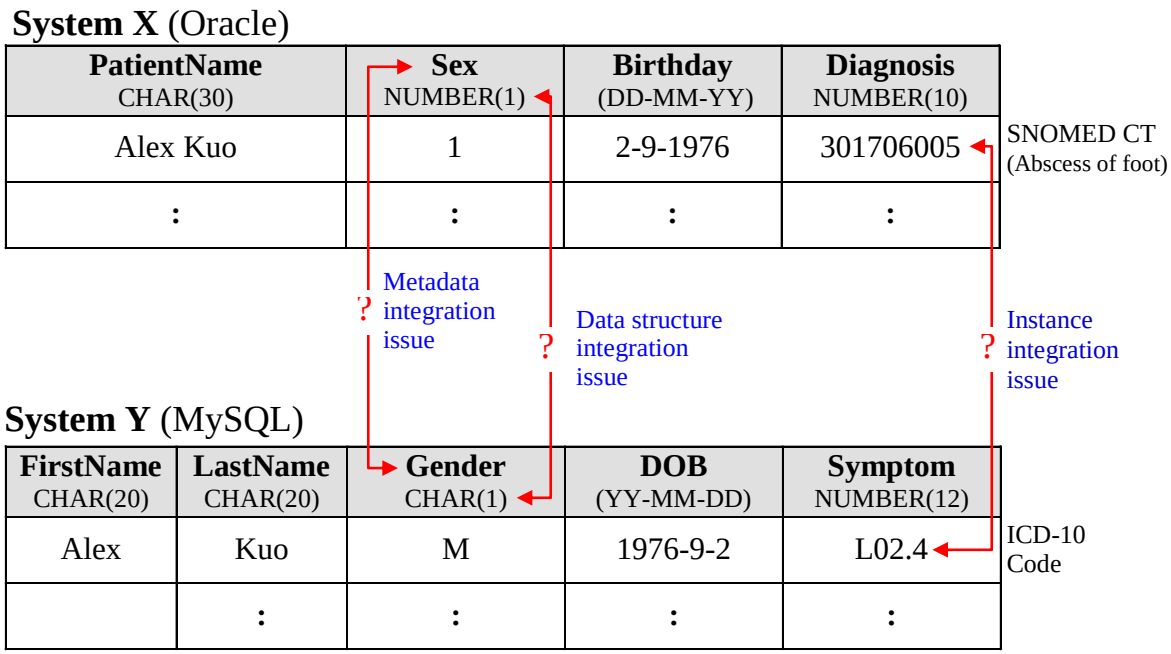
- The motivation
  - Big Data are not usable until they can be aggregated and integrated into a manner that computer can process to generate knowledge.
  - The challenges are
    - Where do I find the data I need?
    - How do I retrieve the data?
    - What is the data structure/format?
    - What are the data access (privacy/security) policies?
    - etc.

A simple study question

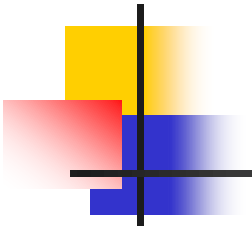
- How the gene ‘myosin light chain 2’ was associated with chamber type hypertrophic cardiomyopathy? The similarity relative to a subset of the genes’ features? The potential connections between pairs of genes.
- Where does he find the relevant documents?
  - Retrieve the documents from well known databases such as NCBI Gene and PubMed.
- Other data sources? Can we use a single search engine to retrieve all collections?

- What is metadata?
  - Metadata is "data about data". It is descriptive information about a particular dataset, object or resource, including how it is formatted, and when and by whom it is collected.
  - Metadata examples

- Metadata integration challenges
  - Each system use its own metadata to describe data.



2. Metadata Standards



- Metadata Standards.
  - Standardized annotations can help organize electronic resources, facilitate legacy resource integration, and support archiving and preservation.
  - Some important standards
    - Dublin Core (DC),
    - Metadata Encoding and Transmission Standard (METS),
    - IEEE Learning Object Metadata (LOM)
    - ISO/IEC 11179
    - Encoded Archival Description (EAD)
    - Machine Readable Cataloguing 21 (MARC 21)

- Metadata standards cross-walking issues.
  - Different metadata standards serve distinct needs and communities.

Table 1. Example of Metadata Crosswalk Mapping

	Dublin Core	EAD	MARC 21
Title Element	Title	<titleproper>	245 00$a (Title Statement/Title proper)
Author Element	Creator	<author>	700 1#$a (Added Entry--Personal Name) (with $e=author) 720$a (Added Entry–Uncontrolled Name/Name) (with $e=author)
Date Created Element	Date.Created	<unitdate>	260 ##$c (Date of publication, distribution, etc.)

(Borrow from "Understanding Metadata", NISO Press, 2004. http://www.niso.org/publications/press/UnderstandingMetadata.pdf)

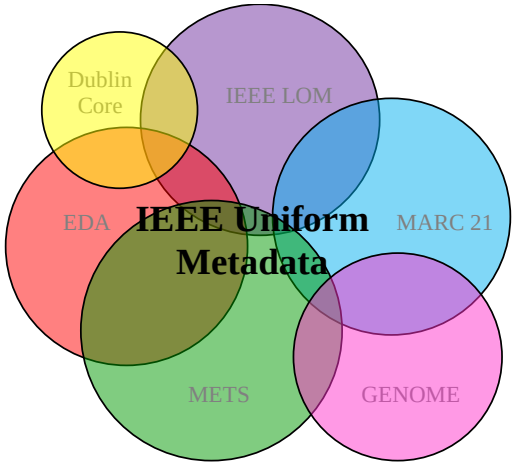
Stanford Library

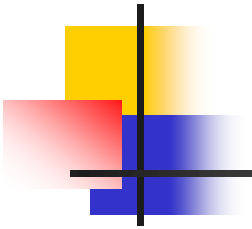
Congress Library

UC Berkeley Library

3. What we want to do?

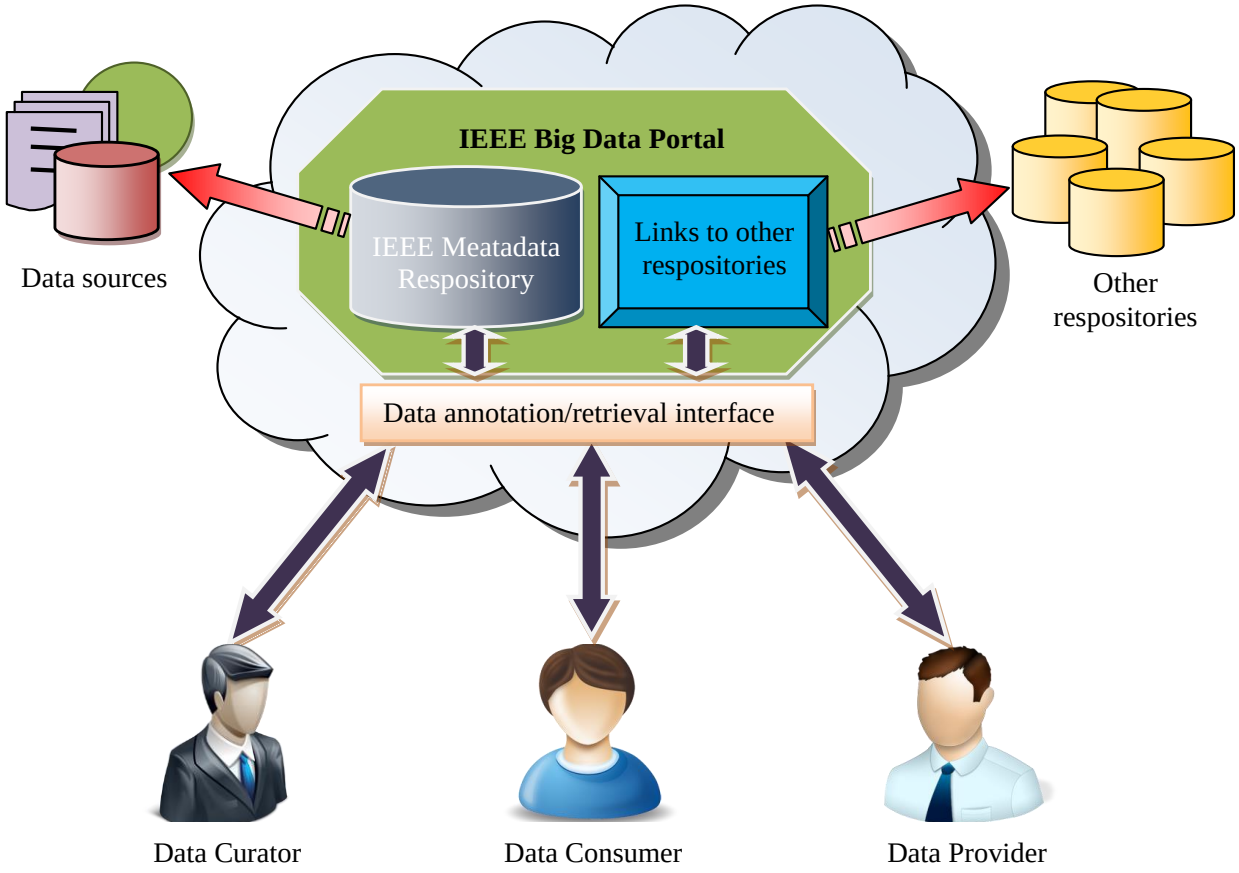
- Form a study/research group
  - We plan to form a research group to study on where there is a need and opportunity for developing a metadata standard for Big Data management.
  - The uniform metadata (UM)
    - It is a set of elements to comprehensively and expressively describe data sources. It is a set of "common concept" metadata extracted from commonly used metadata standards
    - Example



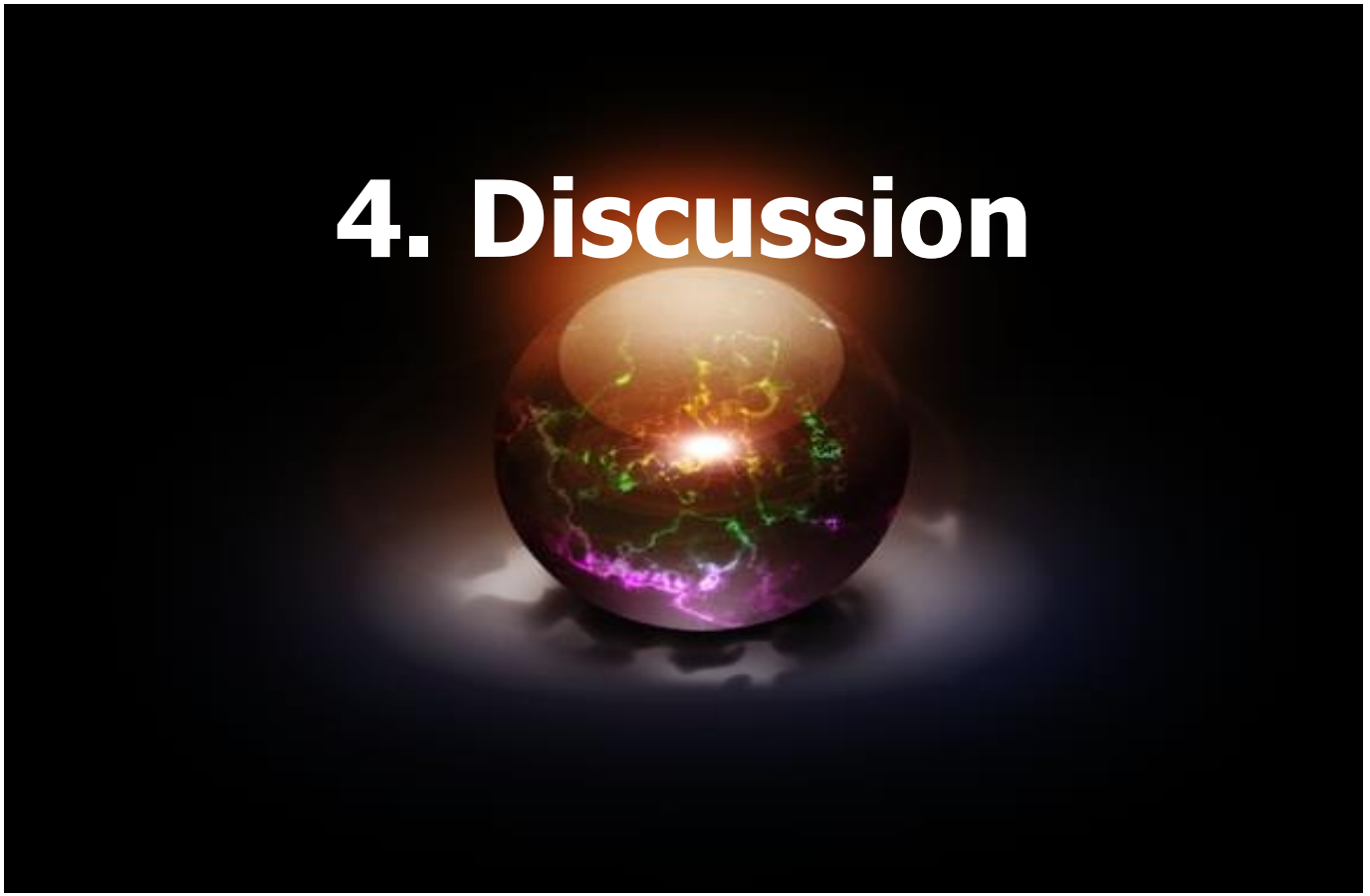


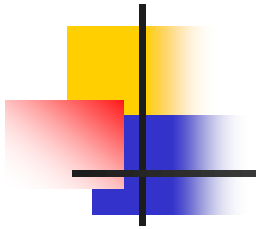
- Develop a proof-of-concept portal
  - The portal will be used to facilitate easy data annotation, discovery and interpretation by using the uniform metadata standard.
  - The repository was used to store descriptive information about digital resources, NOT the original data source.

The architecture of the metadata portal



4. Discussion

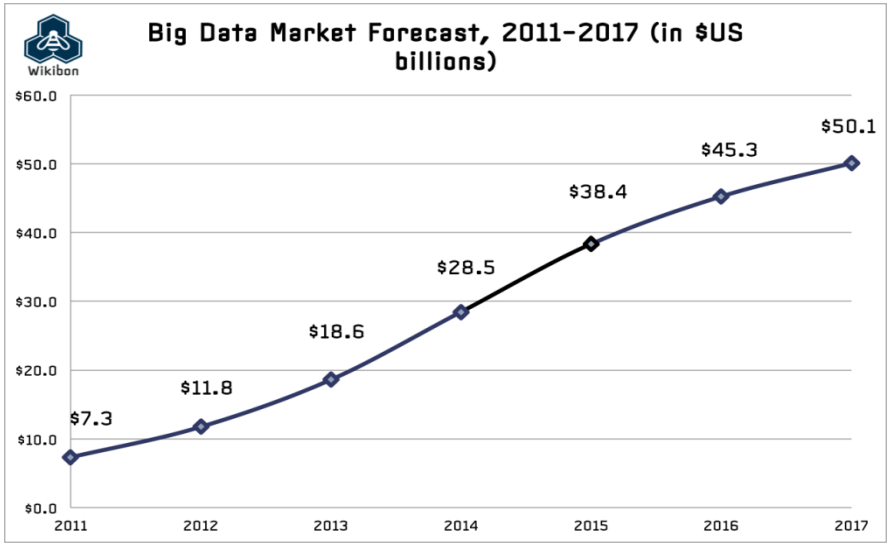


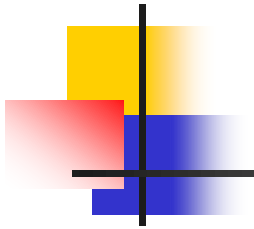


- **Distinct identity**
  - The uniform metadata (UM) is different from other existed standards;
  - The uniform metadata is flexible to adapt a new application domain;
  - A metadata repository and retrieval portal will be developed as a proof-of-concept.

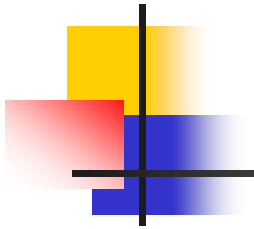
- **Opportunities**
  - The set of uniform metadata has high potential to be developed as an international standard.
  - What does this effort benefit to Academic Institutes and Industries?
    - Academic institutes can enhance research collaborations and avoid duplication of efforts;
    - Industries can develop commercial tools based on the new metadata standard for (big) data annotation, retrieval and analysis.

- **Big Data market potential**
  - International Data Corporation (IDC): The market will reach $125 billion worldwide in 2015.
  - Wikibon: Vendor revenue will reach $50 billion in year 2017.





- **Potential sponsoring groups**
  - The IEEE Standards Association (IEEE-SA)
  - Canada Infoway (https://www.infoway-inforoute.ca/)
  - Stanford Center for expanded data annotation and retrieval (http://med.stanford.edu/cedar.html)
  - Oxford Biosharing (University of Oxford e-Research, https://www.biosharing.org/)



- **Other competitors**
  - ISO/IEC JTC 1/SC 32
  - W3C
    - Community and Business Groups;
    - Semantic Web for Health Care and the Life Sciences Interest Group.
  - NIH BD2K metadata working Group
  - Others (e.g. Open Data initiatives, Digital Library projects)

- **What are our next steps?**
  - Understand the standardization process,
  - Apply funding,
  - Find right people,
  - Go …

Q&A

Two metadata examples

Metadata Data

ID	FirstName	LastName	Sex	Age	Symptom
P45	Li	Lo	Male	47	S87

Metadata Data

#,ID,title,gsm,series_id,gpl,status,submission_date,last_update_date,type,source_name_ch1,organism_ch1,characteristics_ch1,molecule_ch1,label_ch1,treatment_protocol_ch1,extract_protocol_ch1,label_protocol_ch1,source_name_ch2,organism_ch2,characteristics_ch2,molecule_ch2,label_ch2,treatment_protocol_ch2,extract_protocol_ch2,label_protocol_ch2,hyb_protocol,description,data_processing,contact,supplementary_file,data_row_count,channel_count,

1,1,Foreskin Fibroblasts,GSM1,GSE506, GPL4,Public on Sep 28 2000,28/09/2000,19/11/2008, SAGE,mRNA of untreated foreskin fibroblasts, Homo sapiens,NA,total RNA,NA,NA,NA,NA,NA, NA,NA,NA,NA,NA,NA,NA,NA,"Primary human foreskin fibroblasts at passages 15-18, cultured in Eagles Minimal Essential Medium (MEM) supplemented with penicillin and 10% fetal calf serum (FCS).; ..... **xxx.csv**

A simple example

Table 1. Example of Metadata Crosswalk Mapping

	Dublin Core	EAD	MARC 21
Title Element	Title	<titleproper>	245 00$a (Title Statement/Title proper)
Author Element	Creator	<author>	700 1#$a (Added Entry--Personal Name) (with $e=author) 720$a (Added Entry–Uncontrolled Name/Name) (with $e=author)
Date Created Element	Date.Created	<unitdate>	260 ##$c (Date of publication, distribution, etc.)

Topic (UM1001)

Owner (UM1002)

Publication_Date (UM1003)

(Borrow from "Understanding Metadata", NISO Press, 2004. http://www.niso.org/publications/press/UnderstandingMetadata.pdf)

Example:

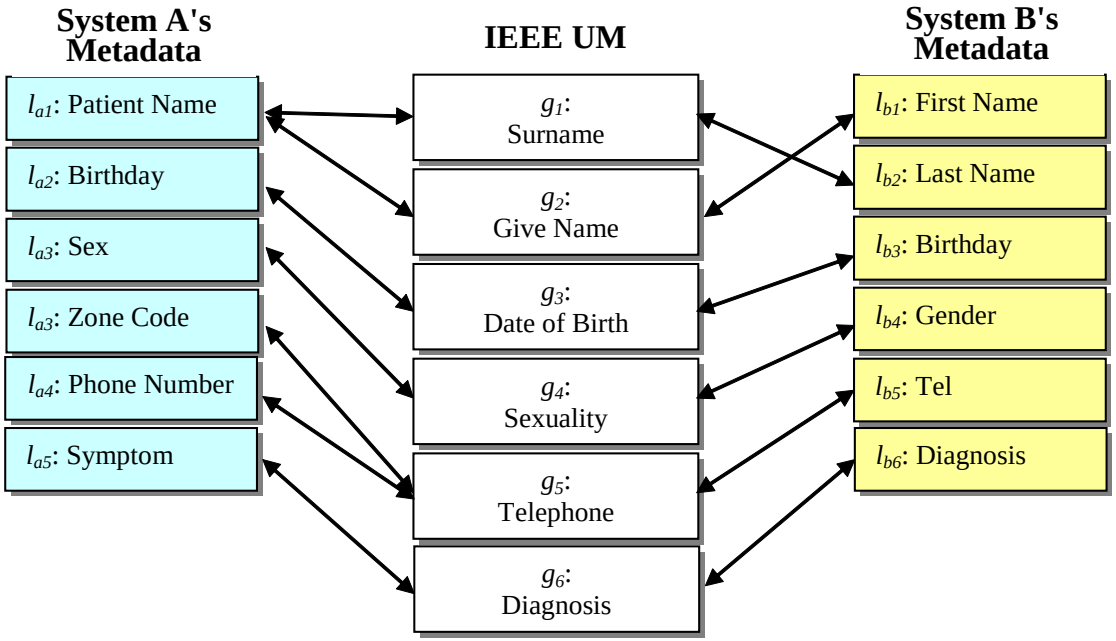
Patient Name = (Surname + Given Name)

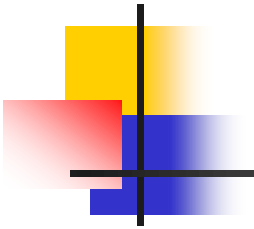
Birthday = Date of Birth

Sex = Sexuality

(Zone Code + Phone Number) = Telephone

Symptom = Diagnosis





- The process of defining the UM:
  1. collecting and examining existed metadata standards,
  2. Classifying metadata into categories based on their similarities,
  3. Drafting BMS format proposals by the research team,
  4. Meeting with standard development organization (SDO) or domain experts to form a consensus,
  5. Encoding the UM schema with reusable components.