SC32 WG2 Metadata Standards Tutorial

Metadata Registries
and Big Data

WG2 N1945

June 9, 2014 Beijing, China

WG2 Viewpoint

Big Data magnifies the existing challenges and issues of managing and interpreting data.

Primary Scope of WG2
Metadata relevant to Big Data

Scope:

- Standards for data management and interchange
  - Within and among local and distributed information systems environments

Goals:

- Facilitate interoperability
- Facilitate discovery, transformation, analysis and data integration

Approach:

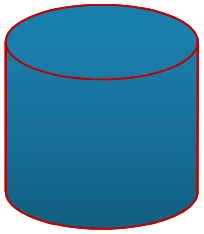
- Standardized data management facilities
  - Metadata Registration and management, naming conventions
- Structured semantics and syntax, use of ontologies and terminology to define meaning
- Reference models and frameworks
  - 11179 framework for registering and managing metadata
    - Metadata describing the meaning and constrains of data fields
      - "Data Elements"
  - 19763 framework for registering models and mappings between models
    - Ontologies, Role/Goal (actors), Processes, Services, Information Models, Form Design, mappings
  - 24707 common specification for logic languages
- Technical reports to support implementation
  - 20943 MDR consistency
  - 20944 MDR interoperability
  - Other technical reports

Big Data Stakeholders*

- Government
- Commercial (Manufacturing and Distribution?)
- Defense
- Healthcare
- Deep Learning
- EcoSystem for Research
- Astronomy and Physics
- Earth and Environment
- Polar Sciences
- Energy

* from JTC1 N0030

Metadata for Big Data



Big Data

What is the data about?
Where is it?
What is the structure?
What do the data values mean?
How was it created?
Who is responsible for it?
Is it suitable for use with program X?
How “big” is it?
How often is it updated?
When was it last updated?
Are there any publications related to the data?
Who are the users of the data?
Are there any services that can process the data?
Where are those services located?
What standards were used to produce or validate the data?

Gaps Identified by JTC1

1. Definitions, Vocabulary and Reference Architectures (e.g. system, data, platforms, online/offline, etc.) [for Big Data]

    19763-3 Registration of Ontologies

2. Specifications and standardization of metadata including data provenance

    11179 MDR

5. Domain-specific language and semantics of eventual consistency [for specific industry domains]

    19763-3 Registration of Ontologies, 19763-12 Information Models, 19763-10 Mappings

7. General and domain specific ontologies and taxonomies for describing data semantics

    – 11179-3 Data Element Concepts, Value Domains; 19763-12 Information Models, 19763-13 Form Designs, 19763-3 Ontologies, 19763-5 Process registration, 19763-7 Service, 19763-8 Role/Goal (Actor)

Gaps Identified by JTC1

9. Remote, distributed, and federated analytics (taking the analytics to the data) including data and processing resource discovery

- 19763-3 Ontologies, 19763-5 Process registration, 19763-7 Service, 19763-8 Role/Goal (Actor), 19763-9 On Demand Model selection

10. Data sharing and exchange

- *This is the main objective for 11179 and 19763 standards*

12. Data analysis and mining

- *The use of structured metadata allow logic languages to reason over data. 11179 and 19763 models facilitate use of structured metadata.*

Volume

- The amount of data is sufficiently large to require special considerations.
- Large dataset, large individual data, high dimensionality, large memory requirements for analyzing the data
- *Metadata for addressing above issues:*
  - *New project split: 11179-7 Registration of Dataset Metadata (metadta*

Variety

- Variety of data represented in different formats (json, xml, sql, etc) – metadata is not relevant to different formats
- Variety between database structures for representing the data for similar universes of discourse
- Variety between information models for representing similar universes of discourse (different semantics and different way to represent semantics)
- Variety different terminology used in the data e.g. data = population of a city
  - "Soft Patterns" dependent on usage
  - What does it mean to be a city? How is the population defined/established/ what is meant by population?
  - Different meaning for the same term
    - different context or universe of discourse for the same
    - the term "city" can be a political entity, versus "city" geographic location
- Variety of data structure: structured (data represented in tables or objects e.g. SQL), unstructured (data that is a blob with not tags or external data model describing it, e.g. photo and freeform text), semi-structured (eg data that does not necessarily fully conform with an external data model, and/or may have tags embedded in the data, e.g. XML documents, JSON)
- *Metadata standards help address above issue*
  - *19763-12 registered Information models*
  - *11179-3 registered data element*
  - *19763-3 registered ontologies*
  - *24707 Common Logic expressions of registered semantics (computable) and mappings between semantics/ontologies*

Velocity

- Frequent changes: Where stale data is inappropriate; this refers to use cases where there is a high frequency of data updates e.g. continuous monitoring of environmental measurements or human data streams
- Rapid Response: A need for rapid response to changes/input data
- *There are no existing relevant WG2 Metadata standards addressing these aspects of Velocity.*

Veracity

- Uncertainty due to incompleteness, inconsistency, ambiguity
- Authenticity/truthfulness of the source of the data
- Accuracy and correctness of the data
- *Metadata standards help address above issue*
  - *11179-3 registered data elements*
  - *24707 Common Logic to determine inconsistencies in the data and incompleteness*

Value

- Data has perceived or quantifiable benefit to the organization
- Need to determine whether or not the data has value to the organization
- *Metadata standards that help address above isseus.*
  - *RGPS registers the role and goal, and the process used to create the data*
  - *19763-5 registered Process models*
  - *19763-7 registered Service models*
  - *19763-8 registered Role/Goal models*
  - *19763-3 registered Ontology concepts*
  - *19763-12 registered Information model*
  - *19763-10 registered mappings between models*
  - *19763-13 registered Form Designs*
  - *11179-3 registered data element*

Summary of WG2 Relevance to Big Data

- Understanding Big Data still requires an understanding of the meaning, structure and of data
- Existing WG2 standards help address big data metadata issues.
- WG2 standards can be extended to meet gaps
- *Two new areas:*
  - *Dataset metadata*
  - *Provenance*

ISO/IEC 11179 & 19763 families

- ISO/IEC 11179 specifies a Metadata Registry and associated procedures
  - Specifies metadata required to describe specific metadata items, such as *data elements*, *data element concepts*, *conceptual domains*, *value domains* and *concept systems such as classification schemes*;
- ISO/IEC 19763 specifies metamodels for registering various types of models and model mappings. It uses the registration procedures of ISO/IEC 11179.
  - Specifies metadata required to describe *Ontologies, Information model, form designs, process models, service models, role and goal models*

Questions and Discussion

???