Implementation Challenges and Standards Opportunity for FAIR Principles

Wo Chang
Digital Data Advisor
ISO/IEC JTC 1/SC42/WG2 Big Data, Convenor
IEEE Big Data Governance and Metadata Management WG, Chair
wchang@nist.gov

October 27, 2020
Open Data Initiative (2013)

Executive Order — Making Open and Machine Readable the New Default for Government Information

Section 1. General Principles. Openness in government strengthens our democracy, promotes the delivery of efficient and effective services to the public, and contributes to economic growth. As one vital benefit of open government, making information resources easy to find, accessible, and usable can fuel entrepreneurship, innovation, and scientific discovery that improves Americans’ lives and contributes significantly to job creation.

Decades ago, the U.S. Government made both weather data and the Global Positioning System freely available. Since that time, American entrepreneurs and innovators have utilized these resources to create navigation systems, weather forecast and warning systems, location-based applications, precision farming tools, and much more, improving Americans’ lives in countless ways and leading to economic growth and job creation. In recent years, thousands of Government data resources across many fields, such as health and medicine, education, energy, public safety, global development, and finance have been posted in machine-readable form for free public use on Data.gov. Entrepreneurs and innovators have continued to develop a vast range of useful new products and businesses using these public information resources, creating good jobs in the process.

Implementation Challenges and Standards Opportunity for FAIR Principles, Wo Chang, NIST/ITL, October 27, 2020
Interagency Technical Advisory Group (Dec. 2013)

To provide a forum for Federal agency coordination on operational requirements and insights on how to maximize access to scientific and technical data. Members are Federal employees participating in their individual capacity as subject matter experts and providing their own perspectives from a range of agency and entity settings including:

- NIST (Chair)
- NIH/NCI
- DOE
- Treasury
- Census
- Smithsonian
- NARA
- USDA

Engage with:
- PID Information Type WG
- Data Type Registry WG
- Data Fabric IG
- SC 32 / WG2 Metadata

Implementation Challenges and Standards Opportunity for FAIR Principles, Wo Chang, NIST/ITL, October 27, 2020
Challenging Problem: Concept Model

DATA
- Reference
- Resource
- Research

SOFTWARE
- Production
- Tools
- Test, Development, ...

PUBLICATIONS
- Peer Reviewed
- Gray Literature
- White Papers, Talk Slides, ...

Repositories
- Public Servers
- Internal Servers
- Other Storage
- Other Repositories
- Cloud

Implementation Challenges and Standards Opportunity for FAIR Principles, Wo Chang, NIST/ITL, October 27, 2020
Challenging Problem: Logical Model

Extended Metadata

Specialized Metadata

Minimum Metadata

Primary Digital Objects
NIST Common Access Platform (CAP, 2014)

Goal:
Develop an interoperable data infrastructure that is scalable to enable automatic data mashups between heterogeneous datasets from various domains without worrying about the data source and structure.

Approach:
Provide basic data infrastructure using persistent identifiers to enable:
* Standard metadata registry for data discovery using a machine-readable format
* Standard data type registry that enables data consumption using a machine actionable format

(without standard data type registry, the data is not easily interoperable and re-usable)
NIST CAP Architecture

Data Consumer
- Supplier
- Enricher
- Aggregator
- Developer

Metadata Registry Service

Type Registry Service

Other Fed. MR
Harvest
Harvest
Harvest

NIST Fed. MR
CKAN
DSpace
Fedora

Types
Properties

Shared Understanding

Specialized Knowledge

FAIR
Findability
Discover
Accessibility
Discover
Interoperability
Map
Reusability
Access

Implementation Challenges and Standards Opportunity for FAIR Principles, Wo Chang, NIST/ITL, October 27, 2020
Implementation Challenges
Use Case: FEMA Emergency Preparedness

In anticipation of a flood FEMA seeks to understand whether or not a State will require assistance, and what kind of resources will be needed.

Analysts must rapidly identify and assemble data for presentation in the dashboard.
NIST Efforts to Develop CAP Arch. For Interoperable Data Structure

Metadata Catalog/Registry

- Data.gov (OMB Common Core 1.1)
  - OMB CC 1.1
    - Catalog #n Fields
    - Dataset #n Fields
      - dataElements
      - Distribution #n Fields

- Census American Community Survey
  - PA Johnstown Population 65+
    - dataTypePID
dataSourceAPI

- USDA SNAP Retailers
  - SNAP Retailer PA Johnstown
    - dataTypePID
dataSourceAPI

- NOAA Flood
  - Flood Area PA Johnstown
    - dataTypePID
dataSourceAPI

Data Type Registry

- Census ACS Population
  - Description
    - Type: FIPS code
    - Concepts: geoMap in LatLong

- USDA SNAP Retailer Location
  - Description
    - Type: Postal Address
    - Concepts: geoMap in LatLon

- NOAA Flood Area Polygon
  - Description
    - Type: LatLong
    - Concepts: geoMap in LatLong

Apply partial ISO 11179 to map multiple concepts

User

Implementation Challenges and Standards Opportunity for FAIR Principles, Wo Chang, NIST/ITL, October 27, 2020
What level of granularity should be exposed when complex/composite datatypes exist?

Parameterized, summarized, or query-based data sets are more challenging than static ones to work with and describe using data types.

How should the needed metadata be created and by whom?

Specificity vs. re-usability of data type
Standards Opportunity
**Goal:**

To enable data integration/mashup among heterogeneous datasets from diversified domain repositories to make data discoverable, accessible, and usable through a machine-readable and actionable standard data infrastructure.

Big data provides key characteristics in *Volume*, *Velocity*, *Variety*, and *Variability*, commonly referred to as the Vs of Big Data. BDGMM is focusing on data from a single source or *Varieties* of data from multiple sources to create an integrated data source for analytics and AI machine learning consumption.

From the new global Internet Big Data economy opportunity in Internet of Things, Smart Cities, and other emerging technical and market trends, it is critical to have a standard reference architecture for Big Data Governance and Metadata Management to support the scalable FAIR (Findability, Accessibility, Interoperability, Reusability) foundation principles.
IEEE Big Data Governance and Metadata Mgt (BMGMM) WG
https://standards.ieee.org/project/2957.html#Working (Sept. 2020)

Approach:

Apply metadata for scalable and machine actionable to
**FAIR** (Findability, Accessibility, Interoperability, Reusability) principles

**BMGMM**

- **Scalable**: Utilizes machine-readable and machine actionable formats
- **Governance Metadata-based Metadata**: Provides authoritative, control, and shared decision making over the management of data assets
- Provides PID-based standard data infrastructure to enable:
  * Catalog Registry for data discovery using a machine-readable format
  * Data Type Registry for data consumption using a machine actionable format
IEEE BDGMM Roadmap White Paper (July 2020)

1 Introduction
1.1 Challenges and opportunities
1.2 Scope of this White Paper
2 Data Explosion
2.1 Internet of Things (IoT)
2.2 Social Media
2.3 Smart Cities
2.4 Smart Manufacturing
2.5 5G Wireless Network
3 BDGMM Case Study
3.1 Case Study #1: Big Data Analytics for Healthcare Fraud Detection
3.2 Case Study #2: Personalized Medicine for Drug Targeting in Prostate Cancer Patients
3.3 Case Study #3: Intelligent Food and Cooking Recipe
3.4 Case Study #4: Internet of Things (IoT)
3.5 Case Study #5: IoT Sensor Network
3.6 Case Study #6: Smart Cities
3.7 Case Study #7: Social Media
3.8 Case Study #8: Analysis of application logs in outsourced scenarios.
3.9 Case Study #9: Data Integration and Management for Additive Manufacturing
IEEE BDGMM Roadmap White Paper (July 2020)

4. BDGMM Technical Requirements
   4.1 Governance Requirements (GR)
   4.2 Metadata Requirements (MR)
   4.3 Data Mashup Requirements (DMR)
   4.4 Analytics Requirements (AR)
   4.5 Data Characteristic Requirements (DCR)
5. Relevant Standardization Activities
6. Standard Technology Gap Analysis
7. Recommendation Standardization Areas and Issues to IEEE-SA
   7.1 Big Data Governance Management
   7.2 Big Data Metadata Management
   7.3 Big Data Integration Framework
   7.4 Persistent Identifier Framework
8. References

Potential Standard BDGMM Reference Architecture

User/System: Application/Service

Open Authentication Encrypted

(1) Discover (3) Access (2) Map

Federated Catalog Registry

- Catalog Metadata Registry (CMR)
  - Core
  - Domain
  - Relation
  - Model
  - Dictionary
  - Schema

Repositories (repo1, repo2, ..., repoN)

Federated Data Types Registry

- Types Metadata Registry (TMR)
  - Derived Types
  - Derived Properties
  - Basic Types
  - Basic Properties

API
Catalog Registry

Metadata for that content is generated in the Repository and pushed, via a Registration Service, into Metadata Registry, creating a digital object.

The Catalog Registry provides Information Management and Discovery Services for users.
Client (process or people) encounter data of an unknown type

Resolved the Type to Type Registry

Response includes type definitions, relationships, properties, and possibly service pointers. Response can be used locally for processing, or, optionally

Typed data or reference to typed data can be sent to service provider
BDGMM Workflow

Datasets

BDGMM System

Catalog Registry

Type Registry

Consumer

Service Provider

BDGMM System

Dataset #4
Definition

PID for Dataset #4

Title

Description

Tags

Last Update

Publisher

Contact Name

Contact Email

Unique Identifier

Public Access Level

Download URL

Database API

1. Register

2a. Search

2b. Return

3a. Access

3b. Service

4a. Manual/Import JSON

Manually/Import JSON

Type Lookup

Access

Type Lookup

Manually/Import JSON

Return

Type Lookup

Access

Type Lookup

Manually/Import JSON

Type Lookup

Access

Manually/Import JSON

Type Lookup

Access

Manually/Import JSON

Type Lookup

Access

Manually/Import JSON

Type Lookup

Access

Manually/Import JSON

Type Lookup

Access

Manually/Import JSON

Type Lookup

Access

Manually/Import JSON

Type Lookup

Access
Operational Model:
Interested parties could deploy federated Metadata Registries and Data Type Registries

BDGMM Framework
- Portals/Apps
- Discover
- Access
- Map

Operational Model:
Interested parties could deploy federated Metadata Registries and Data Type Registries
Recommendations for Standards Development

Big Data Governance Management

Description: Streamlining governance of IT and data are essential for organizations to meet the challenges of the digital era and whoever could govern and manage such resources effectively can reduce the organization’s burden and maximize the customers’ needs. Key recommendations may include the following:

- Adopt/develop standard interface for human readable and machine-actionable to access corporate data catalog that provides detail description and linkage to datasets and their usage.
- Utilize best practices standard networking protocols to support open and multi-levels of security for accessing datasets for (a) end-to-end over the net, (b) at repository, (c) at dataset, (d) at data record/element, etc.
- Adopt/develop extensible PID with scalable resolver to handle massive PID resolution.
- Adopt/develop revision control on datasets with backward and forward compatibility.
Recommendations for Standards Development

Big Data Metadata Management

Description: Supporting diversified metadata schemas and models for various datasets would be essential to organizations to meet the ever-growing customers’ needs and whoever could manage these metadata cohesively across all datasets can reduce corporate burden. In addition, providing computable object workflow functionality between data elements of various datasets would be a great additional service to customers for monitoring events, trigger certain conditions, etc. Key recommendations may include the following:

- Utilize best practices standard metadata as much as possible to capture precise description, data types, properties, unit of measurement, characteristics, etc. for given data elements.
- Adopt/develop standard federated metadata registries to support catalogs and types registries.
- Adopt/develop standard interface to support online data element definition.
- Adopt/develop standard computable object workflow functionality to trigger certain conditions including privacy and ethical issues in datasets.
Recommendations for Standards Development

Big Data Integration Framework

Description: Supporting data integration or data mashup among heterogeneous datasets would be critical for analytics to discover new patterns or knowledge and whoever could manage these rich resources effectively would gain much insights into better decision making. Key recommendations may include the following:

- Adopt/develop standard interface to access data at record level regardless of data at rest or in motion (streaming) from public or secured repositories.
- Adopt/develop standard scalable metadata model to map individual data model across heterogeneous datasets from multiple data sources.
Recommendations for Standards Development

Persistent Identifier Framework

Description: Tagging datasets as persistent identifier (PID) at any level (dataset itself, data record, data element, data type, data property, etc.) would be essential in enabling Findability, Accessibility, Interoperability, and Reusability. Having a standard PID framework would enable interoperability among all heterogeneous datasets across all data repositories. Key recommendations may include the following:

- Adopt/develop standard PID framework that provides organizational namespace with flexible and extensible structure to meet organizational needs.
- Adopt/develop scalable PID resolver to handle massive PID resolution in a millisecond time interval.
Questions?
Please contact: wchang@nist.gov