INFORMATION TECHNOLOGY LABORATORY

Implementation Challenges and Standards Opportunity for FAIR Principles

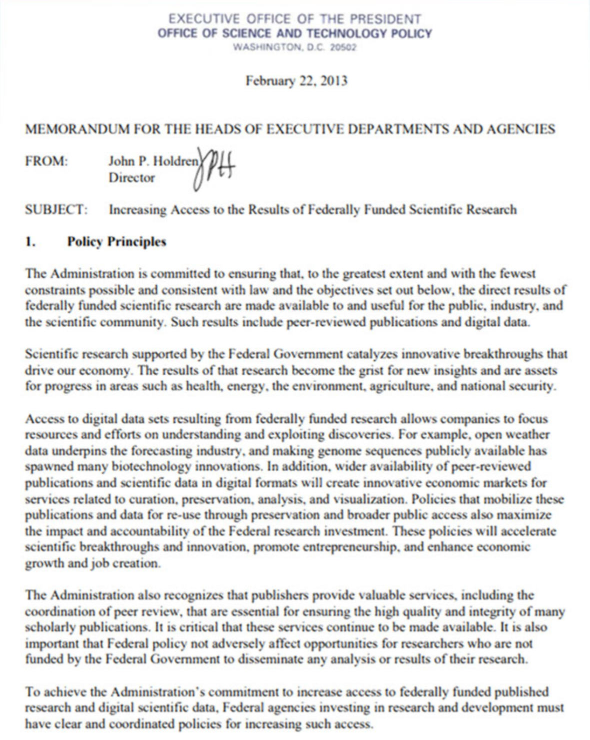
Wo Chang

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

P(A|B) = P(B|A) P(A) / P(B)
$i\hbar\frac{\partial\Psi}{\partial t} = \hat{H}\Psi(x,t)$

Open Data Initiative (2013)



EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

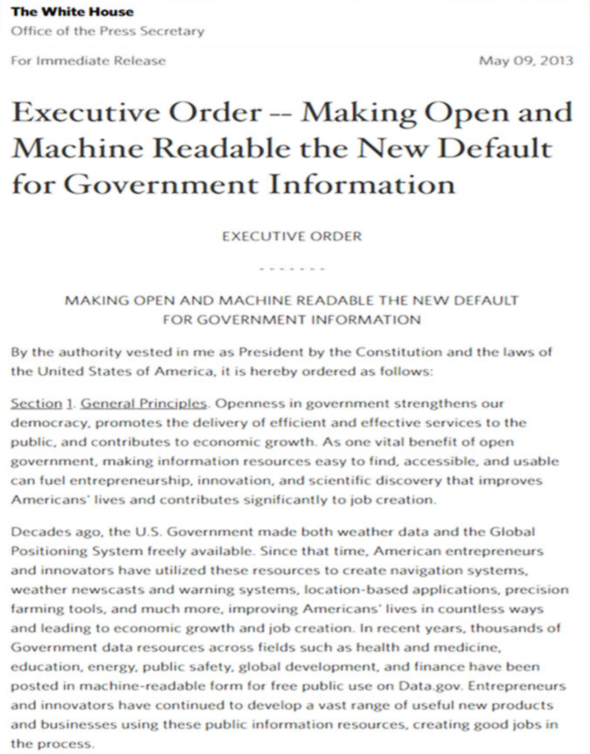
The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs that drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

Access to digital data sets resulting from federally funded research allows companies to focus resources and efforts on understanding and exploiting discoveries. For example, open weather data underpins the forecasting industry, and making genome sequences publicly available has spawned many biotechnology innovations. In addition, wider availability of peer-reviewed publications and scientific data in digital formats will create innovative economic markets for services related to curation, preservation, analysis, and visualization. Policies that mobilize these publications and data for re-use through preservation and broader public access also maximize the impact and accountability of the Federal research investment. These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation.

The Administration also recognizes that publishers provide valuable services, including the coordination of peer review, that are essential for ensuring the high quality and integrity of many scholarly publications. It is critical that these services continue to be made available. It is also important that Federal policy not adversely affect opportunities for researchers who are not funded by the Federal Government to disseminate any analysis or results of their research.

To achieve the Administration's commitment to increase access to federally funded published research and digital scientific data, Federal agencies investing in research and development must have clear and coordinated policies for increasing such access.



The White House
Office of the Press Secretary

For Immediate Release May 09, 2013

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

EXECUTIVE ORDER

- - - - - - -

MAKING OPEN AND MACHINE READABLE THE NEW DEFAULT
FOR GOVERNMENT INFORMATION

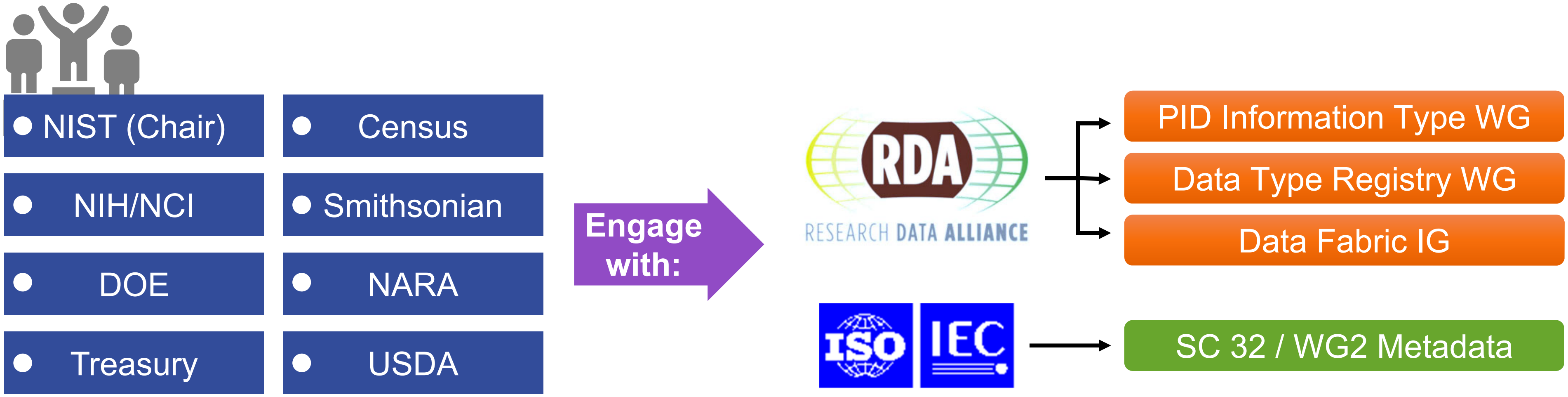
By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

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 newscasts 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 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.

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)
- Census
- NIH/NCI
- Smithsonian
- DOE
- NARA
- Treasury
- USDA

Engage with:

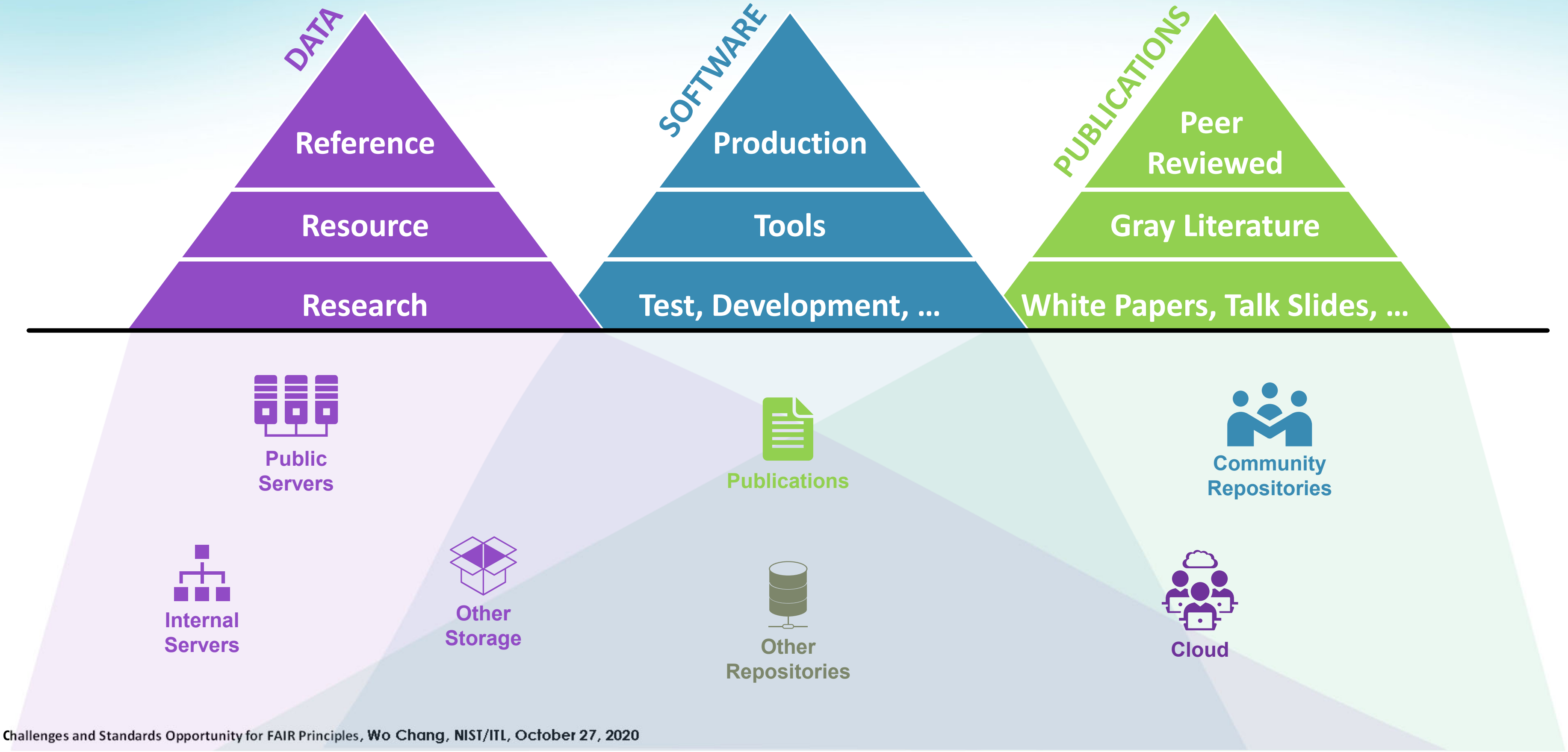
RDA RESEARCH DATA ALLIANCE

- PID Information Type WG
- Data Type Registry WG
- Data Fabric IG

ISO IEC

- SC 32 / WG2 Metadata

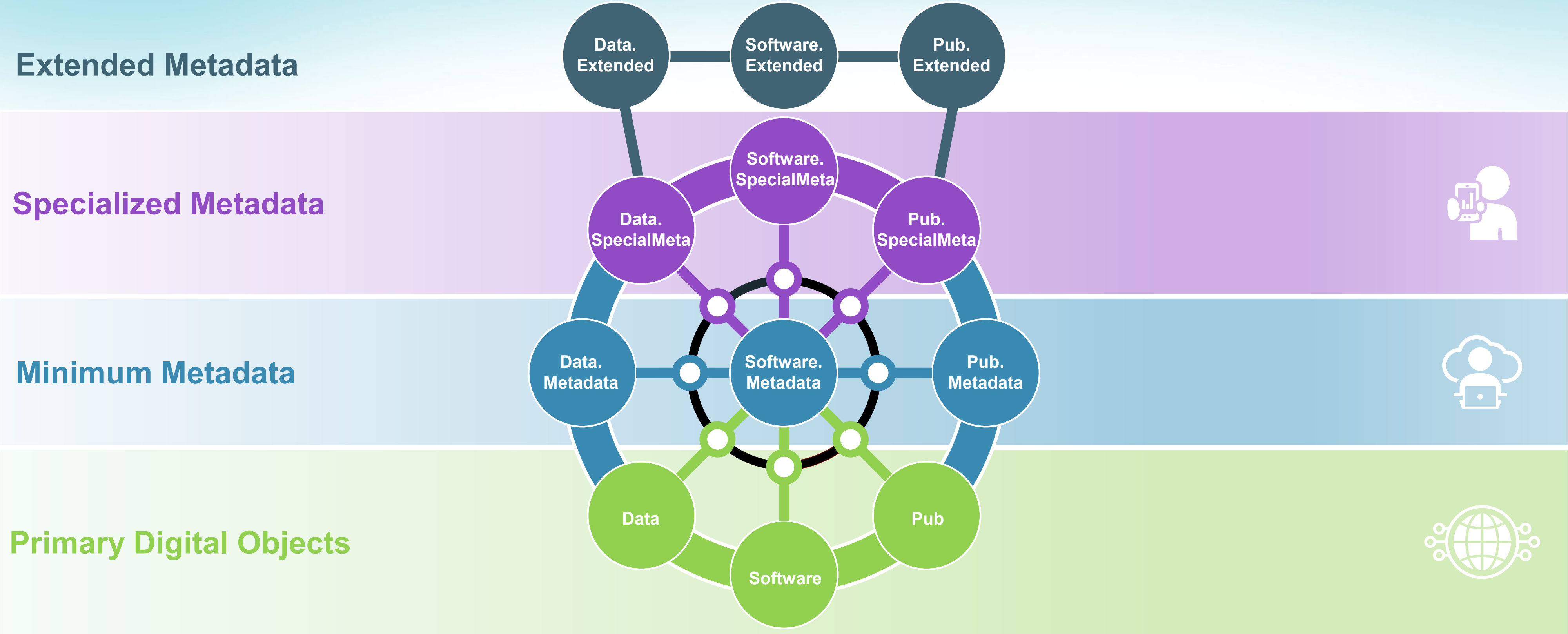
Challenging Problem: Concept Model



ITL INFORMATION TECHNOLOGY LABORATORY

NIST National Institute of Standards and Technology U.S. Department of Commerce

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.

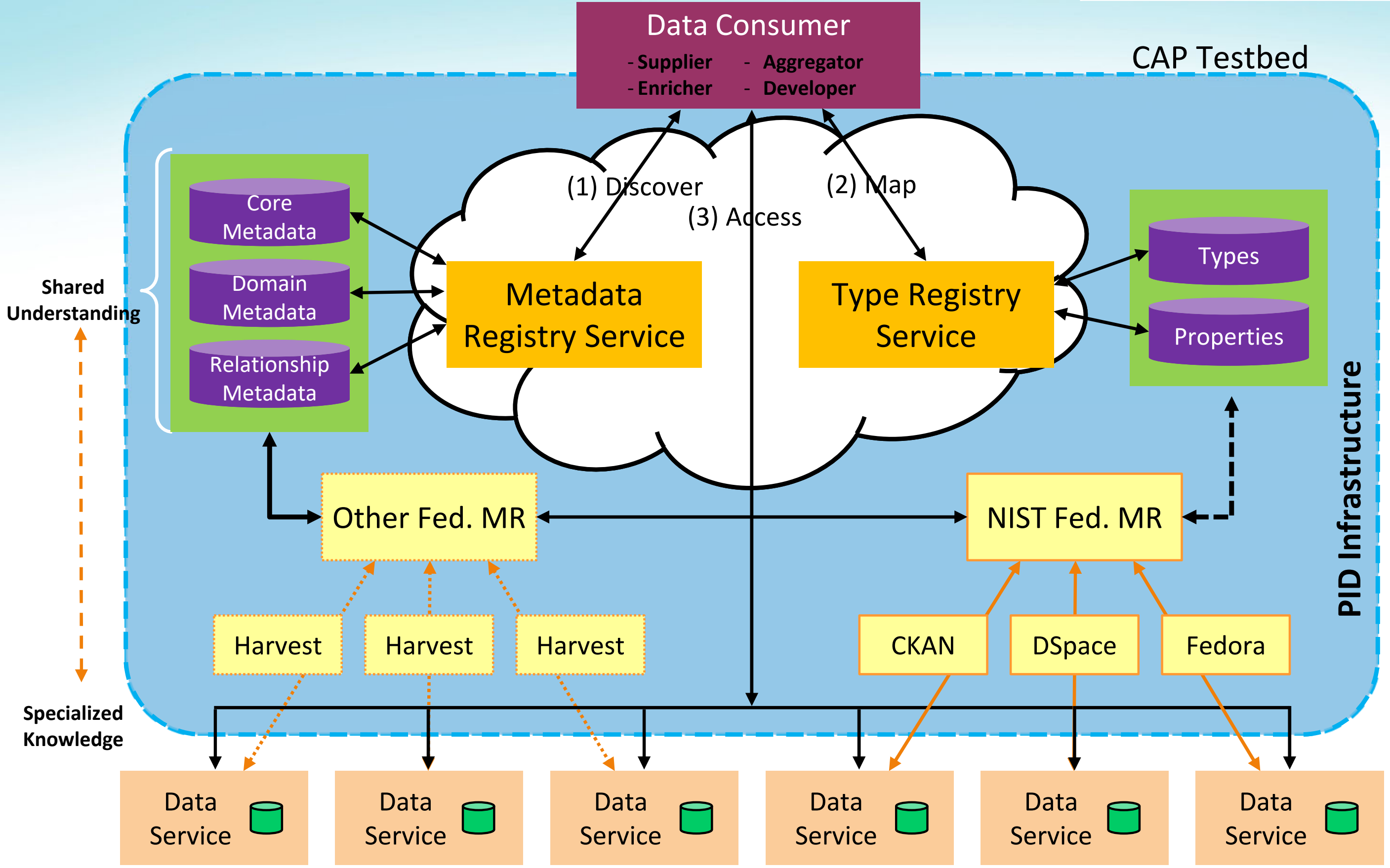
Approach:

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**

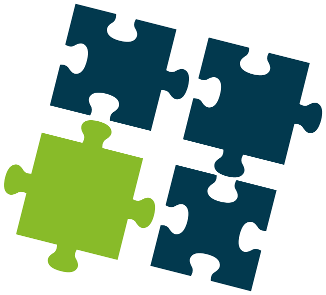
(without standard data type registry, the data is not easily interoperable and re-usable)

FAIR	CAP
Findability	Discover
Accessibility	Discover
Interoperability	Map
Reusability	Access

NIST CAP Architecture



Implementation Challenges

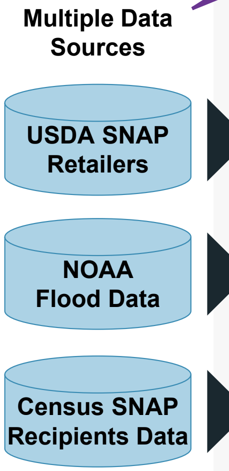


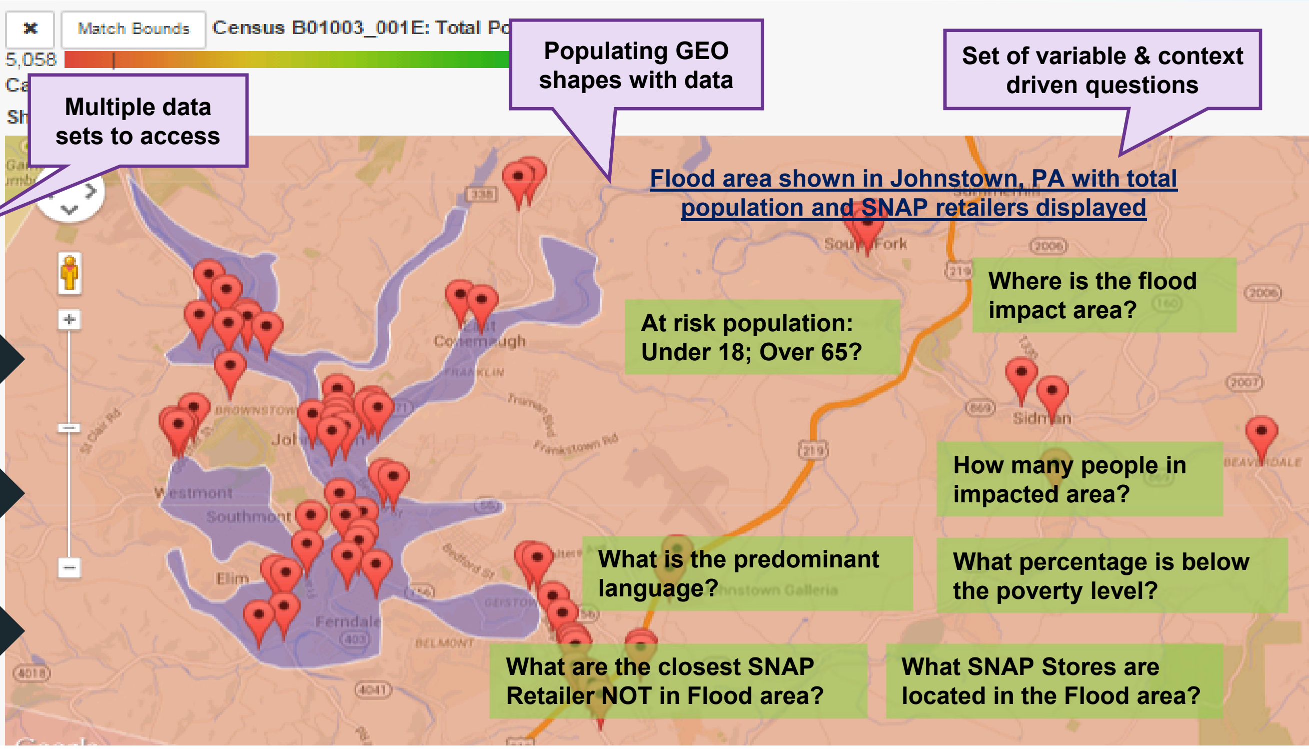
FindIT ConnectIT Prototype – Based on NIST CAP Arch.

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





Multiple data sets to access

Populating GEO shapes with data

Set of variable & context driven questions

Flood area shown in Johnstown, PA with total population and SNAP retailers displayed

Where is the flood impact area?

At risk population: Under 18; Over 65?

How many people in impacted area?

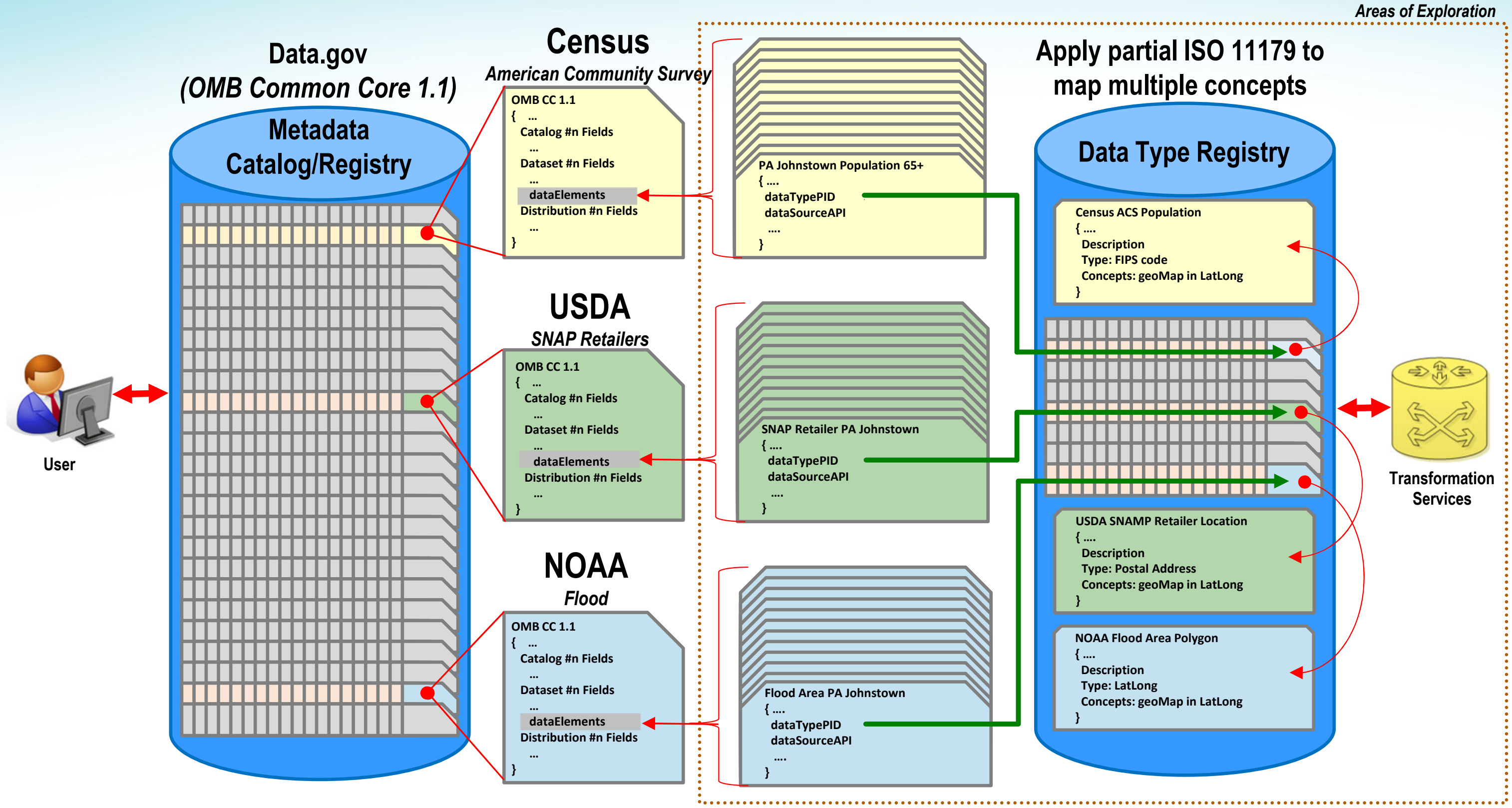
What is the predominant language?

What percentage is below the poverty level?

What are the closest SNAP Retailer NOT in Flood area?

What SNAP Stores are located in the Flood area?

NIST Efforts to Develop CAP Arch. For Interoperable Data Structure



Areas of Exploration

User

Data.gov
(OMB Common Core 1.1)

Metadata Catalog/Registry

Census
American Community Survey

OMB CC 1.1
{ …
Catalog #n Fields
…
Dataset #n Fields
…
dataElements
Distribution #n Fields
…
}

PA Johnstown Population 65+
{ ….
dataTypePID
dataSourceAPI
….
}

USDA
SNAP Retailers

OMB CC 1.1
{ …
Catalog #n Fields
…
Dataset #n Fields
…
dataElements
Distribution #n Fields
…
}

SNAP Retailer PA Johnstown
{ ….
dataTypePID
dataSourceAPI
….
}

NOAA
Flood

OMB CC 1.1
{ …
Catalog #n Fields
…
Dataset #n Fields
…
dataElements
Distribution #n Fields
…
}

Flood Area PA Johnstown
{ ….
dataTypePID
dataSourceAPI
….
}

Apply partial ISO 11179 to map multiple concepts

Data Type Registry

Census ACS Population
{ ….
Description
Type: FIPS code
Concepts: geoMap in LatLong
}

USDA SNAMP Retailer Location
{ ….
Description
Type: Postal Address
Concepts: geoMap in LatLong
}

NOAA Flood Area Polygon
{ ….
Description
Type: LatLong
Concepts: geoMap in LatLong
}

Transformation Services

1

Discover (Findability and Accessibility) is doable but mapping (Interoperability) different information models is very hard (e.g., the way geographic information was encoded and captured by Census vs USDA). Once mapping can be done correctly, access (reusability) is doable. We only hard coded the mapping but we need structure and automatic mapping.

7

Others…

6

How should the needed metadata be created and by whom?

Lessons Learned: very hard problem

2

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

3

How to represent complex/composite data types so the semantics and technical details can be queried by automated search engines?

4

What level of granularity should be exposed when complex/composite datatypes exist?

5

Specificity vs. re-usability of data type

Standards Opportunity

IEEE Big Data Governance and Metadata Mgt (BMGMM) WG

https://standards.ieee.org/project/2957.html#Working (Sept. 2020)

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 *V*s 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.

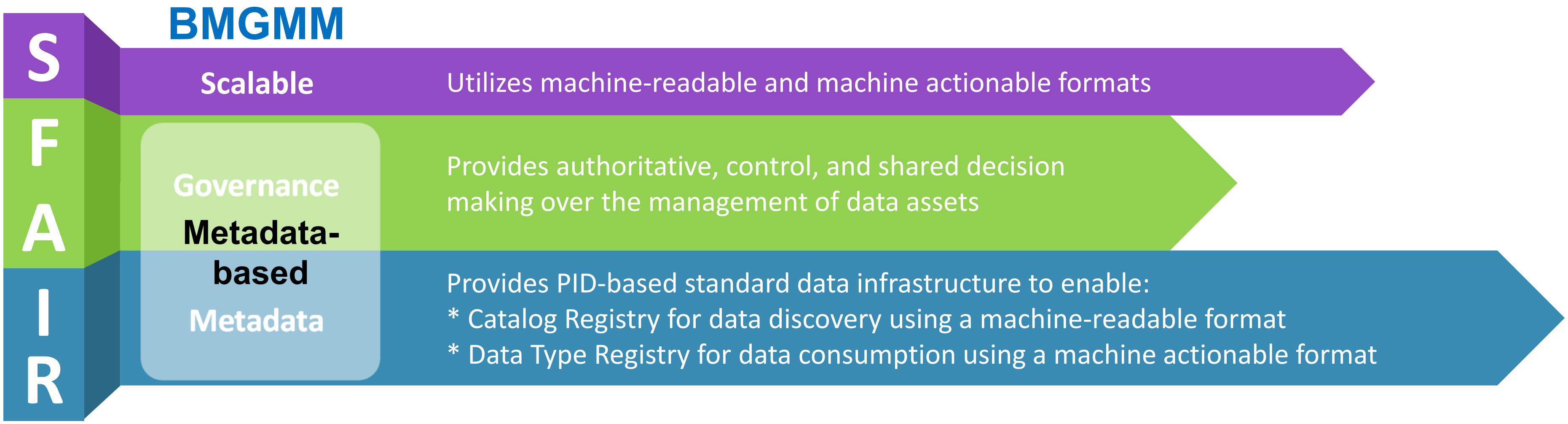
From the new global Internet Big Data economy opportunity in Internet of Things, Smart Cities, and other emerging technical and market trends, it is critical to have a standard reference architecture for Big Data Governance and Metadata Management to support the scalable FAIR (Findability, Accessibility, 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

S F A I R

Scalable — Utilizes machine-readable and machine actionable formats

Governance — Provides authoritative, control, and shared decision making over the management of data assets

Metadata-based

Metadata — 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)

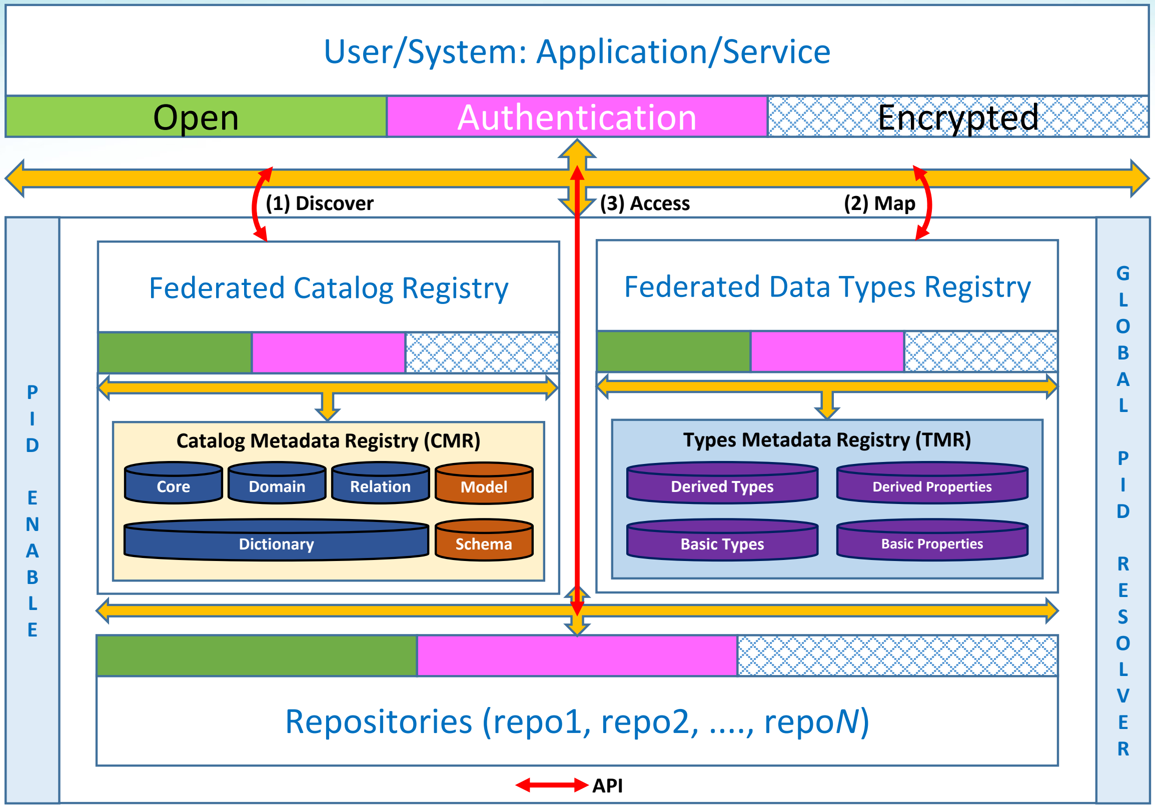
1 Introduction
1.1 Challenges and opportunities
1.2 Scope of this White Paper
2 Data Explosion
2.1 Internet of Things (IoT)
2.2 Social Media
2.3 Smart Cities
2.4 Smart Manufacturing
2.5 5G Wireless Network
3 BDGMM Case Study
3.1 Case Study #1: Big Data Analytics for Healthcare Fraud Detection
3.2 Case 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)

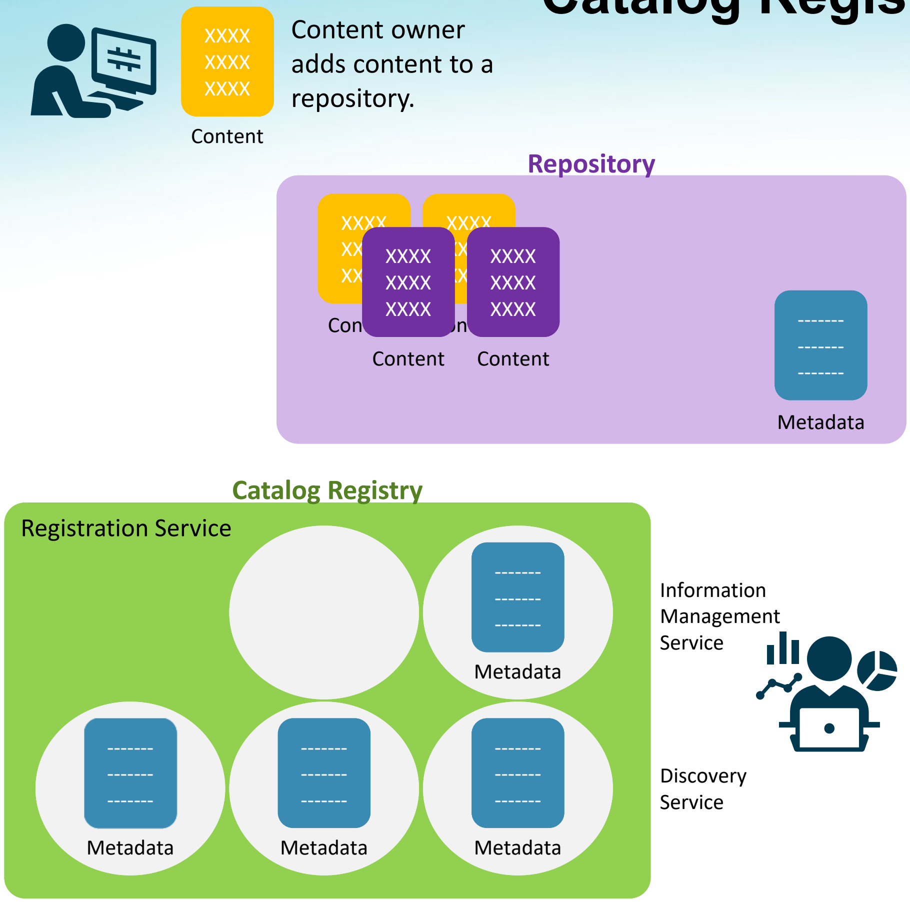
4 BDGMM Technical Requirements
4.1 Governance Requirements (GR)
4.2 Metadata Requirements (MR)
4.3 Data Mashup Requirements (DMR)
4.4 Analytics Requirements (AR)
4.5 Data Characteristic Requirements (DCR)
5 Relevant Standardization Activities
6 Standard Technology Gap Analysis
7 Recommendation Standardization 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

PDF https://standards.ieee.org/content/dam/ieee-standards/standards/web/governance/iccom/bdgmm-standards-roadmap-2020.pdf

Potential Standard BDGMM Reference Architecture


User/System: Application/Service
Open
Authentication
Encrypted
(1) Discover
(3) Access
(2) Map
PID ENABLE
Federated Catalog Registry
Catalog Metadata Registry (CMR)
Core
Domain
Relation
Model
Dictionary
Schema
Federated Data Types Registry
Types Metadata Registry (TMR)
Derived Types
Derived Properties
Basic Types
Basic Properties
GLOBAL PID RESOLVER
Repositories (repo1, repo2, …., repoN)
API

Catalog Registry



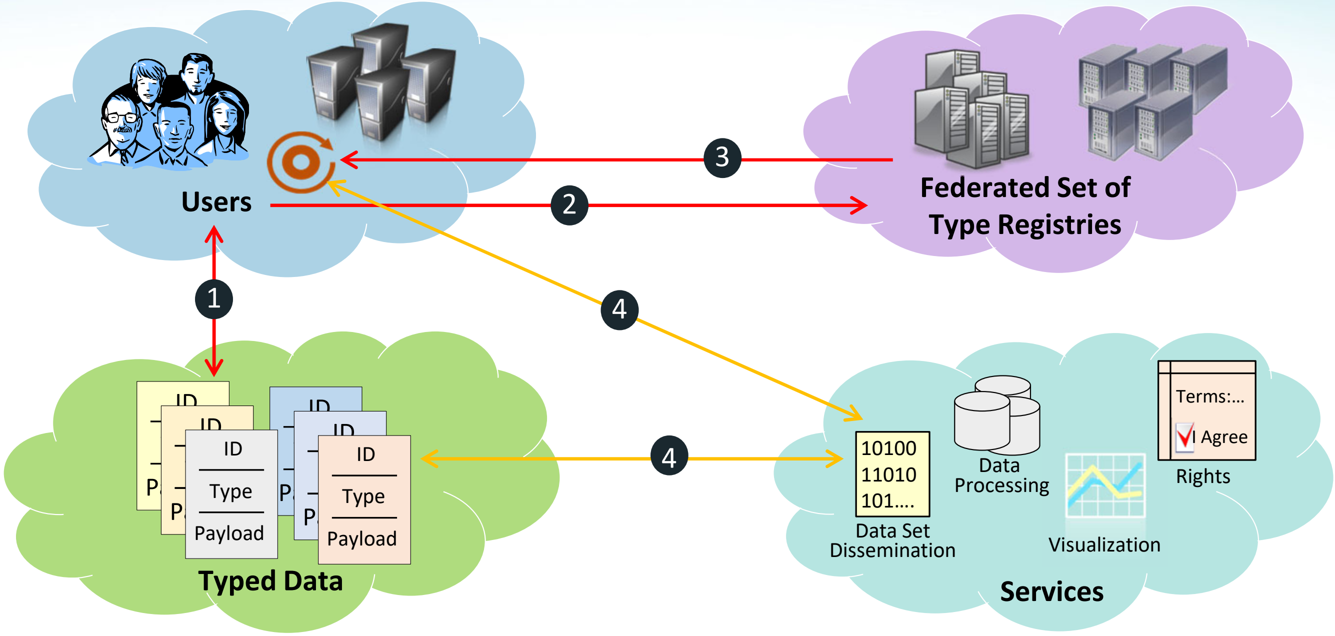
Content owner adds content to a repository.

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.

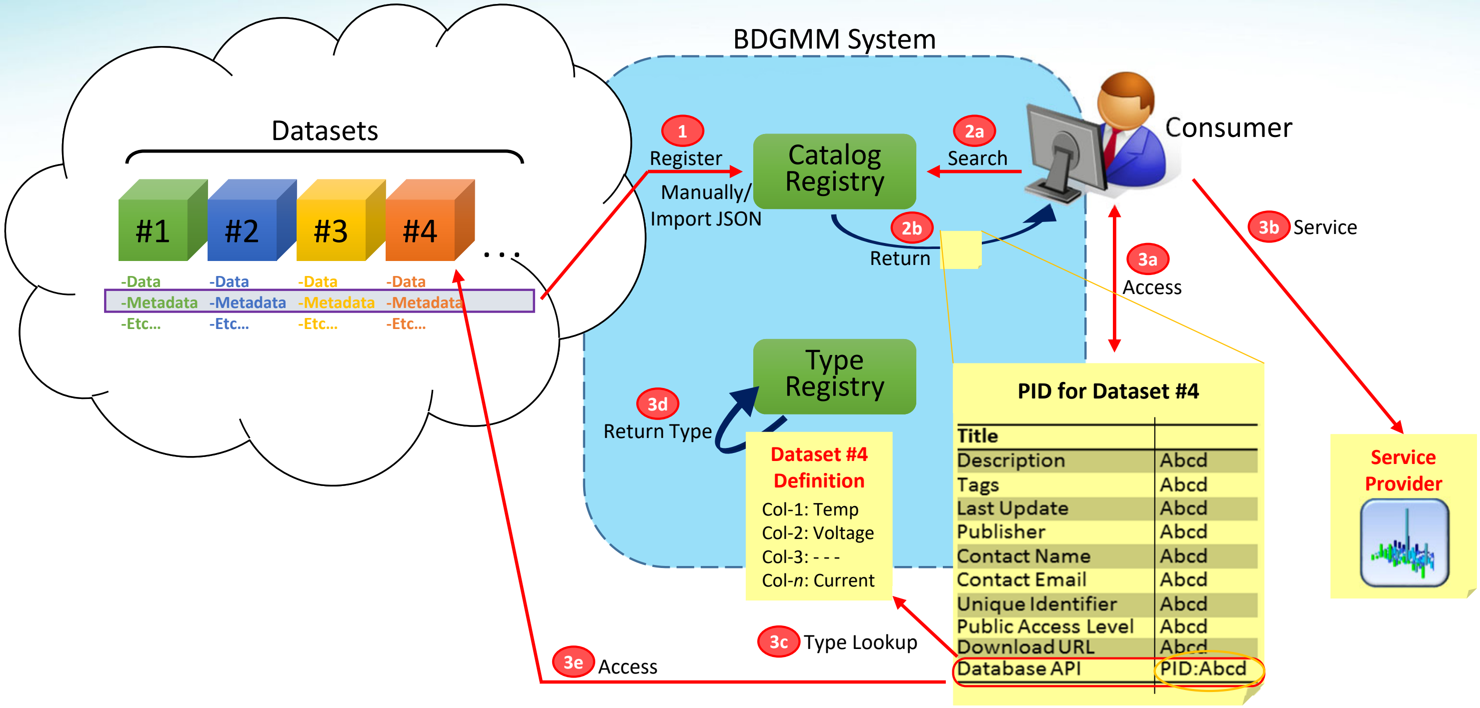
Data Types Registry

1. Client (process or people) encounter data of an unknown type
2. Resolved the Type to Type Registry
3. Response includes type definitions, relationships, properties, and possibly service pointers. Response can be used locally for processing, or, optionally
4. Typed data or reference to typed data can be sent to service provider



Source: Corporation for National Research Initiatives

BDGMM Workflow



BDGMM System

Datasets

#1 #2 #3 #4 . . .

-Data -Metadata -Etc...

-Data -Metadata -Etc...

-Data -Metadata -Etc...

-Data -Metadata -Etc...

1 Register Manually/ Import JSON

Catalog Registry

2a Search

Consumer

2b Return

3a Access

3b Service

Type Registry

3d Return Type

Dataset #4 Definition
Col-1: Temp
Col-2: Voltage
Col-3: - - -
Col-*n*: Current

3c Type Lookup

3e Access

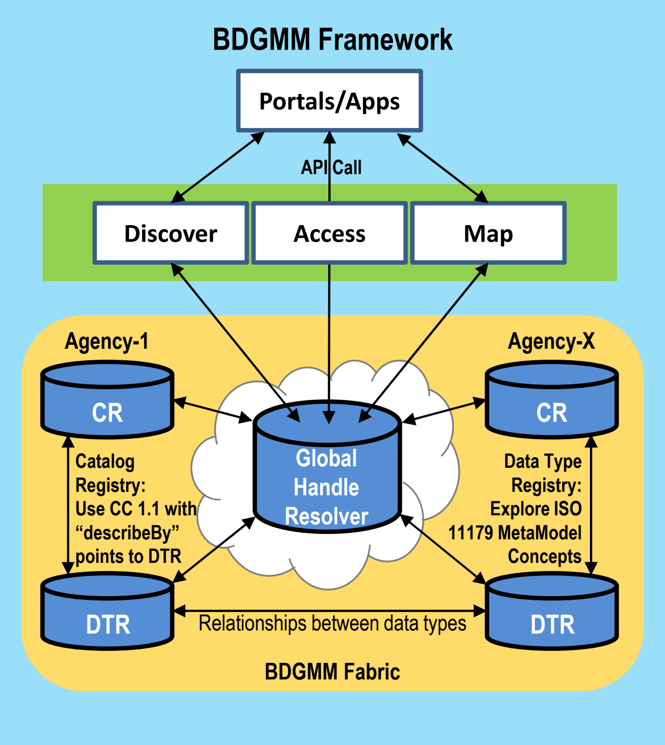
Service Provider

PID for Dataset #4

Title	
Description	Abcd
Tags	Abcd
Last Update	Abcd
Publisher	Abcd
Contact Name	Abcd
Contact Email	Abcd
Unique Identifier	Abcd
Public Access Level	Abcd
Download URL	Abcd
Database API	PID:Abcd

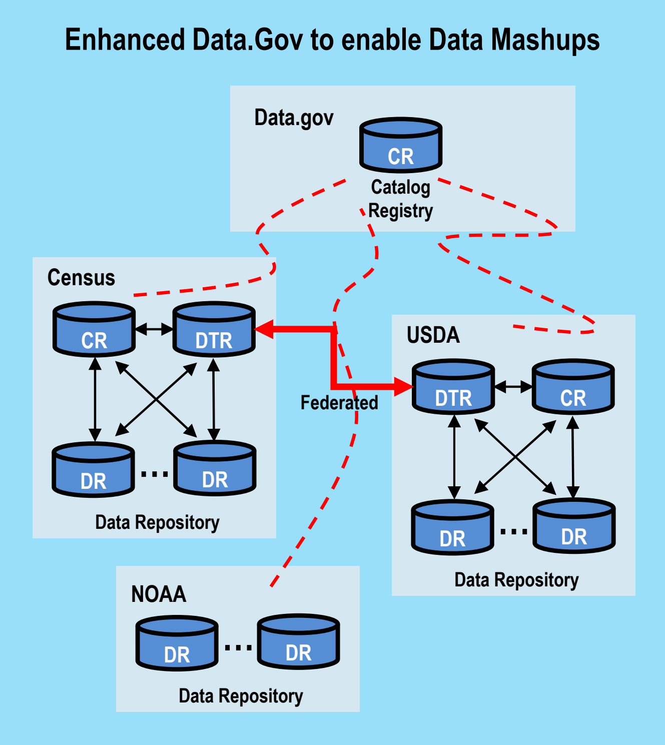
Operational Model:

Interested parties could deploy federated Metadata Registries and Data Type Registries



BDGMM Framework

Portals/Apps
API Call
Discover
Access
Map
Agency-1
Agency-X
CR
Global Handle Resolver
CR
Catalog Registry: Use CC 1.1 with “describeBy” points to DTR
Data Type Registry: Explore ISO 11179 MetaModel Concepts
DTR
Relationships between data types
DTR
BDGMM Fabric



Enhanced Data.Gov to enable Data Mashups

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