IEEE SA
STANDARDS ASSOCIATION

INDUSTRY CONNECTIONS REPORT

BIG DATA GOVERNANCE AND METADATA MANAGEMENT: STANDARDS ROADMAP

Authored by the

IEEE Big Data Governance and Metadata Management
Industry Connections Activity

IEEE

3 Park Avenue | New York, NY 10016-5997 | USA

TRADEMARKS AND DISCLAIMERS

IEEE believes the information in this publication is accurate as of its publication date; such information is subject to change without notice. IEEE is not responsible for any inadvertent errors.

The ideas and proposals in this specification are the respective author's views and do not represent the views of the affiliated organization.

ACKNOWLEDGEMENTS

This document reflects the contributions and discussions by the membership of the IEEE Big Data Governance and Metadata Management Industry Connections Activity (BDGMM-IC), chaired by Wo Chang (National Institute of Standards and Technology, USA) and cochaired by Mahmoud Daneshmand (Stevens Institute of Technology, USA).

IEEE BDGMM-IC would like to acknowledge the following members for their contributions to this document:

Frederic Andres
National Institute of Informatics, Japan

Paventhan Arumugam
National Research and Education Network of India, India

Claire Austin
Environment & Climate Change, Canada

David Belanger
Stevens Institute of Technology, U.S.

Elizabeth Chang
University of Maryland, U.S.

Wo Chang
National Institute of Standards and Technology, U.S.

Paolo Cervolo
Università degli Studi di Milano, Italy

Marl Conrad
National Archives and Records Administration, U.S.

Mahmoud Daneshmand
Stevens Institute of Technology, U.S.

Kathleen M. Darcy
Uniformed Services University of the Health Sciences, U.S.

Santegeeta Dhamdhere
Modern College of Arts, Science and Commerce, India

Mohsen Farid
University of Derby, UK

Jane Greenberg
Drexel University, U.S.

Kathy Grise
IEEE Technical Activities

Keith Jeffery
Consulant

Rebecca Koskela
University of New Mexico, U.S.

Larry Lannom
Corporation for National Research Initiatives, U.S.

Yan Lu
National Institute of Standards and Technology, U.S.

Hiroshi Mano
Data Trading Alliance, Japan

Dhaivat Parikh
IBM, U.S.

Allison L. Powell
Corporation for National Research Initiatives (at time of contribution)
MITRE (current), U.S.

Ann Racuya-Robbins
World Knowledge Bank, U.S.

Luis E. Selva
U.S. Department of Veterans Affairs

Cherry Tom
IEEE Standards Association

Denise Warzel
National Institutes of Health, U.S.

Joan Woolery
IEEE Standards Association

The Institute of Electrical and Electronics Engineers, Inc. 3 Park Avenue, New York, NY 10016-5997, USA

PDF: STDVA24228 978-15-99-6787-2

NOTICE AND DISCLAIMER OF LIABILITY CONCERNING THE USE OF IEEE SA INDUSTRY CONNECTIONS DOCUMENTS

This IEEE Standards Association ("IEEE SA") Industry Connections publication ("Work") is not a consensus standard document. Specifically, this document is NOT AN IEEE STANDARD. Information contained in this Work has been created by, or obtained from, sources believed to be reliable, and reviewed by members of the IEEE SA Industry Connections activity that produced this Work. IEEE and the IEEE SA Industry Connections activity members expressly disclaim all warranties (express, implied, and statutory) related to this Work, including, but not limited to, the warranties of: merchantability; fitness for a particular purpose; non-infringement; quality, accuracy, effectiveness, currency, or completeness of the Work or content within the Work. In addition, IEEE and the IEEE SA Industry Connections activity members disclaim any and all conditions relating to: results; and workmanlike effort. This IEEE SA Industry Connections document is supplied "AS IS" and "WITH ALL FAULTS."

Although the IEEE SA Industry Connections activity members who have created this Work believe that the information and guidance given in this Work serve as an enhancement to users, all persons must rely upon their own skill and judgment when making use of it. IN NO EVENT SHALL IEEE OR IEEE SA INDUSTRY CONNECTIONS ACTIVITY MEMBERS BE LIABLE FOR ANY ERRORS OR OMISSIONS OR DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO: PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS WORK, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE AND REGARDLESS OF WHETHER SUCH DAMAGE WAS FORESEEABLE.

Further, information contained in this Work may be protected by intellectual property rights held by third parties or organizations, and the use of this information may require the user to negotiate with any such rights holders in order to legally acquire the rights to do so, and such rights holders may refuse to grant such rights. Attention is also called to the possibility that implementation of any or all of this Work may require use of subject matter covered by patent rights. By publication of this Work, no position is taken by the IEEE with respect to the existence or validity of any patent rights in connection therewith. The IEEE is not responsible for identifying patent rights for which a license may be required, or for conducting inquiries into the legal validity or scope of patents claims. Users are expressly advised that determination of the validity of any patent rights, and the risk of infringement of such rights, is entirely their own responsibility. No commitment to grant licenses under patent rights on a reasonable or non-discriminatory basis has been sought or received from any rights holder. The policies and procedures under which this document was created can be viewed at http://standards.ieee.org/about/sasb/iccom/.

This Work is published with the understanding that IEEE and the ICCom members are supplying information through this Work, not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought. IEEE is not responsible for the statements and opinions advanced in this Work.

TABLE OF CONTENTS

BIG DATA GOVERNANCE AND METADATA MANAGEMENT: STANDARDS ROADMAP

EXECUTIVE SUMMARY

Governance and metadata management poses unique challenges with regards to the Big Data paradigm shift. The governance lifecycle requires sustainability from creation, maintenance, depreciation, archiving, and deletion due to volume, velocity, and the variety of big data changes, and can be accumulated whether the data is at rest, in motion, or in transaction. Furthermore, metadata management must also consider the issues of security and privacy at the individual, organizational, and national levels.

From the new global Internet, Big Data economic opportunity in the Internet of Things, Smart Cities, and other emerging technologies and market trends, organizations considering implementing Big Data governance and metadata management assume there is a unified organizational buy-in with clear business missions and potential value propositions identified. Furthermore, we further assume that there is strong management support with well-defined roles and responsibilities between business processing units, enterprise architects, and application teams to drive the corporate visions forward. Therefore, this white paper solely focuses on what and how technologies and infrastructure can enable effective and proactive management for both governance management and metadata management.

From a business agnostic point of view, it is critical to have a standard reference architecture for Big Data Governance and Metadata Management that is scalable and can enable Findability, Accessibility, Interoperability, and Reusability between corporate heterogeneous datasets and repositories independent of various application domains without worrying about data source and structure.

The goal of Big Data Governance and Metadata Management (BDGMM) is to address Big Data characteristic challenges that would enable data integration/mashup among heterogeneous datasets from diversified domain repositories. This would allow data to be discoverable, accessible, and re-usable through a machine readable and actionable standard data infrastructure.

This document identifies the challenges and opportunities with the ever-increasing deluge of information and forms the scope of boundaries in Section 1. Section 2 surveys examples of how such data exploration could impact emerging vertical application domains such as IoT, social media, smart cities, smart manufactures, and 5G wireless networks. Section 3 provides a deeper understanding of the BDGMM challenges through a variety of case studies. Section 4 presents a set of extracted requirements from the identified case study challenges and aggregated them into five general categories of requirements. Section 5 lists the relevant standards activities that can potentially be used as enablers to support BDGMM. Section 6 constructs a potential reference architecture by analyzing common feature components from various governance and metadata management architectures, then performing standard technology gap analysis by comparing what the architecture feature components need and what existing standards can offer. Section 7 lists recommendations of possible standards development needs in order to support BDGMM.

1 INTRODUCTION

Corporate Governance is concerned with holding the balance between economic and social goals and between individual and communal goals. The corporate governance framework is there to encourage the efficient use of resources and equally to require accountability for the stewardship of those resources. The aim is to align as nearly as possible the interests of individuals, corporations and society.

—Sir Adrian Cadbury in "Global Corporate Governance Forum," World Bank, 2000 [1]

1.1 CHALLENGES AND OPPORTUNITIES

Technology propelled Industrial Revolution 1.0 with steam-powered factories in 1765; Industrial Revolution 2.0 with mass production and manufacturing in the 1870s; Industrial Revolution 3.0 with the rise of electronics and automation in 1969; and currently, Industrial Revolution 4.0 with the rapid development of "smart things" such as artificial intelligence in machine and deep learning, smart robots, smart devices, smart additive manufacturing, smart cities, etc. During each revolution, there were always challenges and opportunities for any organization.

As Industrial Revolution 4.0 is ushered into mainstream culture and produces an ever-increasing deluge of information, the collective sum of world data will grow from 33 zettabytes (ZB, 10^{21}) in 2018 to 175 ZB by 2025, with a compounded annual growth rate of 61% [2]. With such rich information resources, which can be likened to the "oil of the digital era," effective governance and management can bring about economic growth and a strong competitive edge in the marketplace.

1.2 SCOPE OF THIS PAPER

This document focuses on forming a community of interest from industry, academia, and government, intending to develop a standards roadmap for Big Data Governance and Metadata Management (BDGMM). The approach includes the following:

- Review BDGMM-related technology trends, use cases, general requirements, and reference architecture;
- Gain an understanding of what standards are available or under development that may apply to BDGMM;
- Perform standards, gap analysis, and document the findings; and
- Document vision and recommendations for future BDGMM standards activities that could have a significant industry impact.

Within the multitude of best practices and standards applicable to BDGMM-related technology, this document focuses on approaches that: (1) apply to situations encountered in BDGMM; (2) explore best BDGMM architectures that may be nonexistent, and (3) facilitate addressing BDGMM industry use cases' needs.

2 DATA EXPLOSION

This section explores potential data growth in certain industry sectors that could cause major disruptions to society due to data-related explosive trends. They include device platforms, diverse domain applications with different communication protocols, file formats, and data types. Organizations have a deeper understanding of each of their dynamic challenging environments so they can better govern and manage such digital assets efficiently and effectively.

2.1 INTERNET OF THINGS (IOT)

The Internet of Things impacts all parts of our society. To name a few simple scenarios—in healthcare, IoT enables patients to receive remote, realtime health monitoring, staff and equipment tracking, and enhanced chronic disease and drug management; for logistics, IoT enables realtime tracking, monitoring, cargo coordination, transportation, quantity consumption and refill, to product safety control with self-regulated sensors for controlling/maintaining temperature, humidity, moisture, and lighting.

IoT device growth trend shows that for the connected device, there were approximately 15.4 billion in 2015 growing to 30.7 billion in 2020, and 75.4 billion by 2025, a 50% growth. For wearable devices, there were about 28.3 million units sold in 2016 to 82.5 million units in 2020, a 31% growth [3].

2.2 SOCIAL MEDIA

The rise of social media platforms nearly two decades ago allowed communities to network, share, and discuss information in realtime. The increase in Internet speed and bandwidth, as well as the proliferation of wireless personal devices like smartphones and tablets, has enabled nearly 42% of the total world population to share and access one or more realtime streams like texting, audio and movie clips, and taking and sharing personal photos simultaneously. When major outbreaks called for requiring stay-at-home orders, web conferencing applications became a necessary communication tool for small (a few), medium (a few hundred), and large groups (a few thousand) of people to virtually meet for causal gatherings, formal seminars, and major events, respectively.

As of January 2018, in a total population (TP) of 7.593 billion people, the growth rate per year of Internet users was about 7% (53% TP), social media users was around 13% (42% TP) – or about 11 new users per second, and mobile phone users was roughly 4% (68% TP) [3].

2.3 SMART CITIES

The aims for smart cities are to improve municipal service management, public safety, connected transportation infrastructure, and healthy living environment. Utilizing arrays of diverse sensors with edging and cloud computing would enable efficient and effective management in the facilitation of traffic, transportation, utility and water usage, waste disposal, and smart parking, among a number of areas that economize resources. Scenarios include street traffic safety between vehicles to vehicles or between vehicles and pedestrians in order to avoid collisions; smart parking for locating the closest available parking space using realtime parking maps; smart transportation to alert passengers of arrival/departure time; online payment for public transportation; realtime monitoring and alerting system about air pollution and ground-level zone level; smart street lighting to switch on/off based on object motion detection or turning lights brighter for pedestrians crossing during the night; etc.

Smart cities continue to grow [4] globally and ranges from implementing intelligent, energy-efficient lighting for

all public roads, solar panels for thousands of buildings' rooftops, sustainable airports powered only with renewable energy, sensor-based systems to operate automatic street and building lighting with waste management and security, electric vehicles and eventual self-driving vehicles.

2.4 SMART MANUFACTURING

Smart manufacturing enables instant decision-making based on collective sensors data, predictive modeling and realtime analysis from all these devices. It also helps to move static operations into dynamic, reliable, scalable, secure, and on-demand production. With Industrial Revolution 4.0, technology connectivity, human and machine collaborations, and unprecedented access of contextualized data and models provide many benefits; these include: (a) enable the manufacturing process to influence design decisions and avoid designs sent to production, (b) additive manufacturing like 3D printing of bio replacement body parts like human tissue and organs [5] to 3Dirgio, the world's largest 3D printed boat [6], a record 25-foot long and 5,000-pound sailing vessel.

Data-driven smart manufacturing [7] will continue to improve in the following areas:

a) Growing physical and digital sources like robotics tools and autonomous robots, AI and cognitive systems, augmented reality, etc.

b) Reducing machine downtime reductions in maintenance planning by 20–50%, lowering total maintenance costs by 5–10%, and increasing equipment productivity and availability by 10–20%.

2.5 5G WIRELESS NETWORK

The goal of high bandwidth wireless networks is to provide compatible high-quality and reliable services like wired communication. Improvements from 4G (1 Gbps, high-definition mobile TV, video conferencing) to 5G (35 Gbps, device to device communication) would enable many new applications such as high-quality audiovisual immersive entertainment, connected vehicles to share road situations and driving conditions, public safety and infrastructure in managing cities resources, smart manufacturing in human and machine partnership, machine remote control for heavy machinery to avoid hazardous environments, realtime high resolution video surveillance for monitoring intrusions or break-ins, telemedicine for physical therapy via AR/VR, remote precision surgery, first responder and rescue for rough terrains, smart agriculture with faster connectivity and more sensors to track data points on rainfall, water content, nutrients in the soil, ground temperature, and many others.

5G wireless will continue to grow and impact many areas including (a) network coverage for 40% of the world by 2024, handling 25% of all mobile traffic data, (b) supporting millions of devices per square mile, and (c) under optimal conditions, data latency will range between 1 to 4 milliseconds [8].

2.6 HEALTHCARE

Healthcare systems now need to focus on patient experiences in terms of evidence based personalized and value-driven outcomes while demanding privacy and security of the data.

Healthcare is facing major challenges to address the issues of limited medical resources in term of equipment, logistics and facilities, shortages of qualified medical providers, population increase, chronic diseases and aging population, endemics, predicting and managing pandemics and addressing mental health as an integral component of medical health among others. With the introduction of mobile and wearables such as mobile phones and watches, etc. it has become apparent that the data they generate have become vital in giving insights into monitoring patients' conditions and the progress in response to treatments. Additionally, privacy, security and the ability to control fraud and keeping the cost of healthcare provisioning must also all be well managed.

The challenges presented above introduce other challenges, i.e., how we design data systems specific to healthcare, in terms of data collection, preprocessing, curation and definition of requirements for data storage, communication and interoperability. Healthcare data forms vary greatly in terms of the data sources, i.e., lab results, textual reports, multimodal images (e.g., x-ray, MRI, ultrasound, etc.) handwritten reports, financial and insurance reports, just to mention a few. With widespread increase in the use of Internet of Things (IoT) devices [9] the amounts of data they produce, it has become evident that new paradigms are needed to address the volume, the velocity and the variety in order to address the personalized value-added outcomes. Since 2016, organizations are facing the massive increase of close to 900% in healthcare data generation leading over 8.0 petabytes only in 2018 [10].

Major issues are the lack of protocols that related producers and consumers of data within the medical data and the rapid growth in data production [11]. Additionally, we have entered the era of do-it-yourself (DIY) healthcare, where patients will benefit from the ability to monitor vitals for minor as well as major conditions with synchronized and asynchronized communication with medical providers to allow them to monitor and rapidly respond to situations where immediate care is required.

3 BDGMM CASE STUDY

A *case study* is a typical application that provides high-level general background information. However, it is important to study how its unique characteristics and requirements compare to other case studies across fields. In order to develop a consensus list of requirements across all stakeholders, BDGMM began by collecting various case studies.

The following information is presented for each case study:

- Background: high-level description that may include architecture diagram.
- Big Data Governance Management Challenges: identify Big Data governance management issues.
- Big Data Metadata Management Challenges: identify Big Data metadata management issues.
- Big Data Analytics Challenges: identify Big Data analytics issues.
- Data Characteristics: file format, archival data or streaming data, etc.

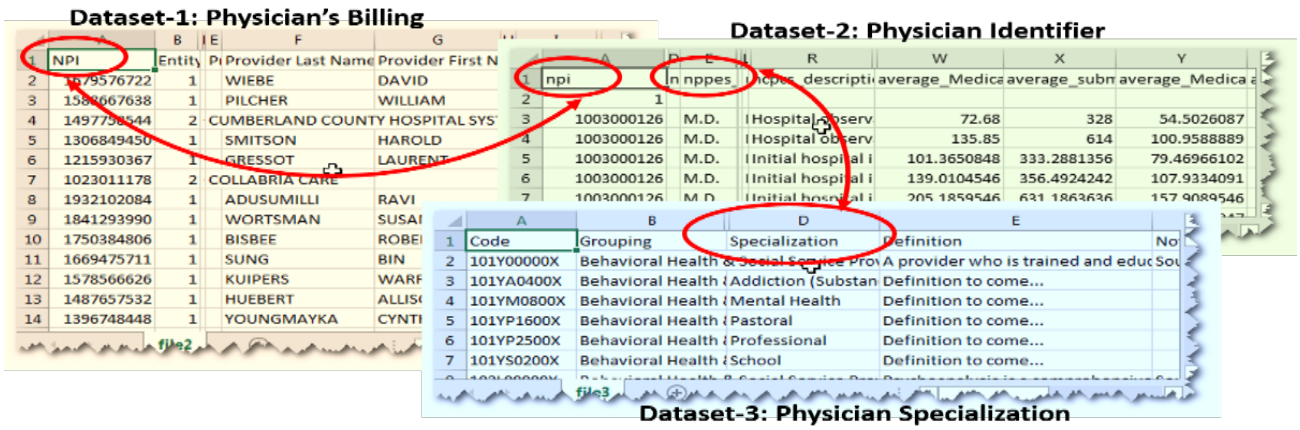
Case studies presented in this section were reproduced according to original submission—original content has not been modified. Specific vendor solutions/technologies and other standards references mentioned in these case studies were included for the purpose of understanding different aspects of requirements. However, the listing of these solutions and technologies does not constitute endorsement from BDGMM. The front matter (page 3) contains a general disclaimer.

3.1 CASE STUDY #1: BIG DATA ANALYTICS FOR HEALTHCARE FRAUD DETECTION

BACKGROUND

Large quantities of healthcare data are produced continually and stored in different databases as shown in Figure 1. With widespread adoption of electronic health records, the amount of available data has increased exponentially. Nevertheless, healthcare providers have been slow to leverage that vast amount of data for improving the healthcare system or for improving efficiency and reducing overall healthcare costs.

FIGURE 1: SAMPLE OF HEATH CARE DATA



Healthcare data has immense potential to innovate healthcare delivery in the U.S. and inform healthcare providers about the most efficient and effective treatments. Value-based healthcare programs will provide incentives to both healthcare providers and insurers to explore new ways to leverage healthcare data to measure the quality and efficiency of care.

In the U.S., an approximate $75B to $265B is lost each year to healthcare fraud [12]—these criminal schemes cannot be ignored. Healthcare providers must develop automated systems to identify fraudulent activities, waste, and abuse in order to reduce its harmful impact on businesses.

BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

Different organizations have different governance management practices. Mechanisms and rights for accessing healthcare data from such diversified organizations can be challenging.

BIG DATA METADATA MANAGEMENT CHALLENGES

Healthcare data comes from various organizations (e.g., healthcare providers and insurance agencies) and therefore their metadata tagging protocols (naming, structuring, meaning, etc.) are different. Furthermore, the unit of measurement on data fields can vary. For example, patient temperature is recorded in Fahrenheit units in the U.S.; whereas, Celsius is used in the UK.

BIG DATA MASHUP CHALLENGES

Working with multiple datasets with different file formats (csv, etc.) from different repositories (file-based, RESTful API, etc.) across different organizations and/or national boundaries; being able to align or link data fields in order to expand/discover new knowledge at the machine level without humans in the loop.

BIG DATA ANALYTICS CHALLENGES

Identifying and applying appropriate tools for statistical analysis, machine learning, and visualization for predictive modelling in order to detect irregularities and prevent healthcare payment fraud.

DATA CHARACTERISTICS

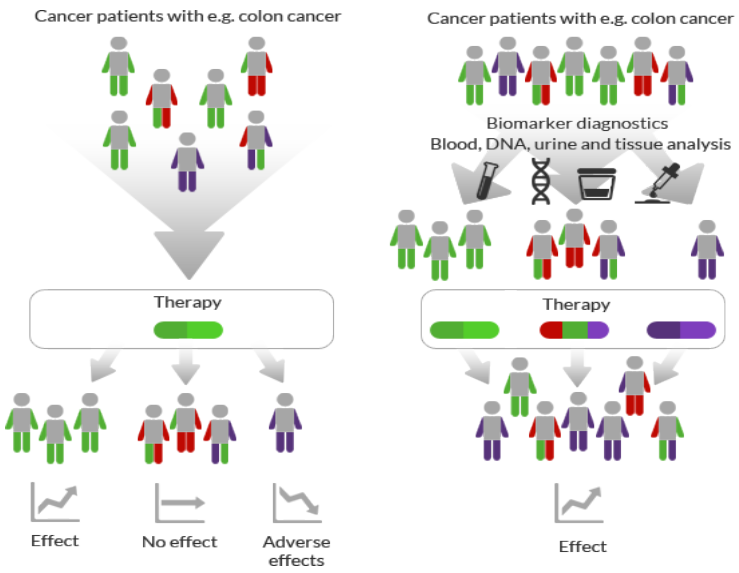
ASCII, CSV files, file-based

3.2 CASE STUDY #2: PERSONALIZED MEDICINE FOR DRUG TARGETING IN PROSTATE CANCER PATIENTS

BACKGROUND

Personalized medicine is the act of tailoring chemotherapy or drugs based on a patient's specific set of DNA or genes. When a person is diagnosed with cancer, a variety of tests are performed (blood, DNA, urine, or tissue analysis) as shown in Figure 2, giving physicians a snapshot into that patient's unique set of DNA. This information allows for "smart" prescribing of medications that complement a patient's signature genetic background and achieve therapeutic response.

FIGURE 2: SAMPLE OF PERSONALIZED MEDICINE



NCI's Genomic Data Commons (GDC) data portal is a huge data repository of more than 32,000 patient cases, and includes clinical data, treatment data, biopsy results, 22,000+ genes, as well as a whole host of other information. This allows accessibility to other researchers who want to uncover new biomarkers, find correlations between genes and survival, or delve deeper into new novel topics of particular interest.

BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

Publicly-available genomic data and their associated clinical information are generated from multiple organizations, each with their own methodology of data collection and dissemination. Mechanisms and rights for accessing genomic data from such diversified organizations can be challenging.

BIG DATA METADATA MANAGEMENT CHALLENGES

Genomic data from sequencing or analytical platforms have different metadata structures between different data hubs. Understanding semantic meaning and syntactic structures of these data fields can be challenging.

BIG DATA MASHUP CHALLENGES

While NCI GDC genomic data are well curated with different levels of description, applying appropriate linkage between data fields without humans in loop would be challenging.

BIG DATA ANALYTICS CHALLENGES

Aside from applying appropriate statistical analytical tools to find significant p-values, without specific parameters, combining datasets to do discovery will be challenging and often times a SME is necessary.

DATA CHARACTERISTICS

ASCII, TSV or JSON, file-based.

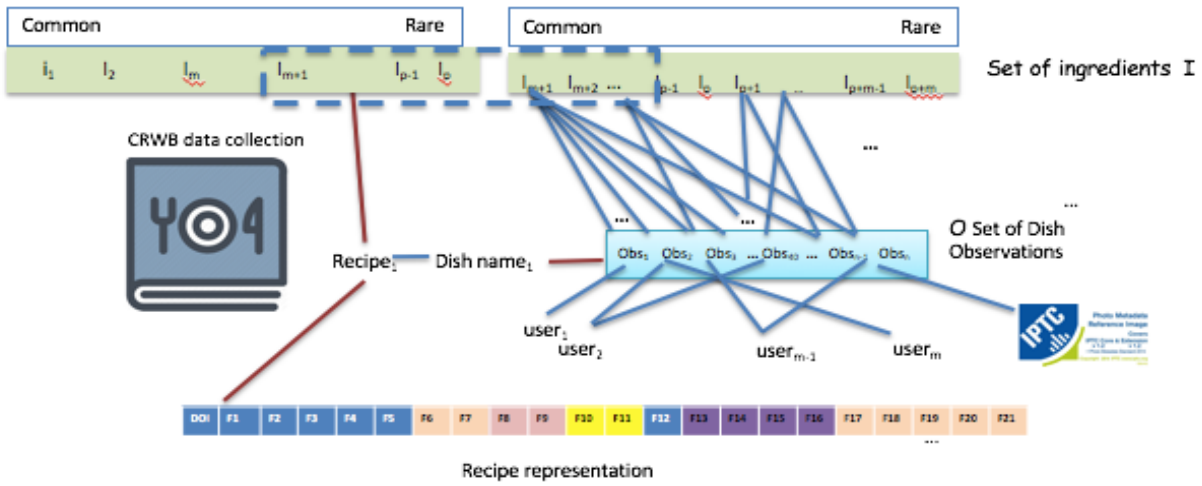
3.3 CASE STUDY #3: INTELLIGENT FOOD AND COOKING RECIPE

BACKGROUND

Food science is the study of the physical, biological, and chemical makeup of food and the concepts underlying food processing. Recipes are perhaps the simplest examples of the connection between food science and cooking: Cooking is Chemistry and Physics as there is a fusion of cooking and science—ingredients and measurements, instructions and written documentation, all designed to lead anyone to a specific, repeatable outcome that someone else has perfected.

CRWB data portal [13],[14] is a linked open data repository containing a record of 60 years of recipes and a CRWBvoc set of ontologies including an ingredients ontology, recipe taxonomy, a cooking process-centric ontology, and a food tasting ontology as shown in Figure 3. This allows accessibility to other researchers who want to investigate several fields (dish recipe generation, cooking execution plan optimization, recipe DNA coding, recipe similarity), find correlations between cooking the recipe, nutrition issues, and food tasting, or look into whatever topic interests them.

FIGURE 3: FOOD ONTOLOGIES



BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

Publicly-available cooking recipe data and their associated nutritional and health-related information are generated from multiple organizations, each with their own methodology of data collection and dissemination. Mechanisms and rights for accessing enhanced cooking recipe data from such diversified organizations can be challenging.

BIG DATA METADATA MANAGEMENT CHALLENGES

Cooking recipe collections as raw data or as linked open data have different metadata structures between different data hubs. Promoting interoperable semantic meaning and syntactic structures of these data fields can be challenging.

BIG DATA MASHUP CHALLENGES

CRWB data will facilitate Big Data Mashup including IoT mashup tools and the integration of heterogeneous cooking recipes and applications from multiple sources including IoT for healthcare and other research purposes.

BIG DATA ANALYTICS CHALLENGES

Aside from applying appropriate classifier tools and finding similarities between recipes and combinations of ingredients, generating new recipes as healthy and gastronomical discoveries will be challenging.

DATA CHARACTERISTICS

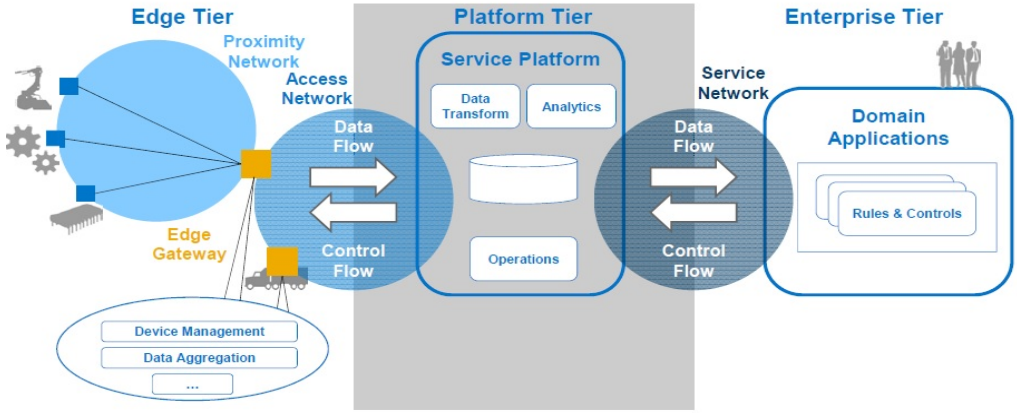
CSV or JSON .format

3.4 CASE STUDY #4: INTERNET OF THINGS (IOT)

BACKGROUND

The IEC defines Internet of Things as "an infrastructure of interconnected objects, people, systems and information resources together with intelligent services to allow them to process information of the physical and the virtual world and react."[15] To illustrate the IoT system, the following three-tiered architecture is shown in Figure 4 comprising of edge tier, platform tier, and enterprise tier. The edge tier includes IoT physical devices (e.g., sensors, actuators, controllers, tags, gateways) and IoT edge devices (e.g., network components and IoT gateways used for connectivity and communication). IoT platform enables collection of data from heterogeneous sources and transforms them towards intelligent decisions. Many domain-specific applications can be built in the enterprise tier supporting intelligent control actions. Various IoT application use cases fall into three application domains—Industrial, Consumer, and Public Sector. Industrial IoT applications such as smart grid and predictive maintenance requires realtime data collection with high reliability, while Consumer applications such as smart home and wearables can work with best-effort network. The Public Sector IoT system includes various smart city and social sensing applications involving multiple stakeholder public sector agencies who exchange data and rely on a different level of mashups to enable smart services.

FIGURE 4: THREE TIER IOT ARCHITECTURE PROPOSED BY INDUSTRIAL INTERNET CONSORTIUM



BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

Managing of data from heterogeneous data sources like various sensors, video cameras in both discrete and streaming mode, and providing a data access mechanism between diverse sets of applications is a challenge. Additionally, the data across various IoT devices and vendors needs to be normalized before they can be used for analysis. Some data access requests from Industrial applications will have to be met in realtime.

BIG DATA METADATA MANAGEMENT CHALLENGES

For providing effective data exchange between entities, IoT data needs to be catalogued with all of the relevant information—owner of the data, its components, formats, authorized users, whether data source is stationary or mobile, anonymization requirement, etc.

BIG DATA MASHUP CHALLENGES

Standardized interfaces for data sharing, discovering, and interoperability of information mashups.

BIG DATA ANALYTICS CHALLENGES

Placement of the analytics functionality in the IoT network—in-device, edge or in the cloud—depending on both historical and realtime data.

DATA CHARACTERISTICS

Audio, video, text, numeric, both discrete and streaming.

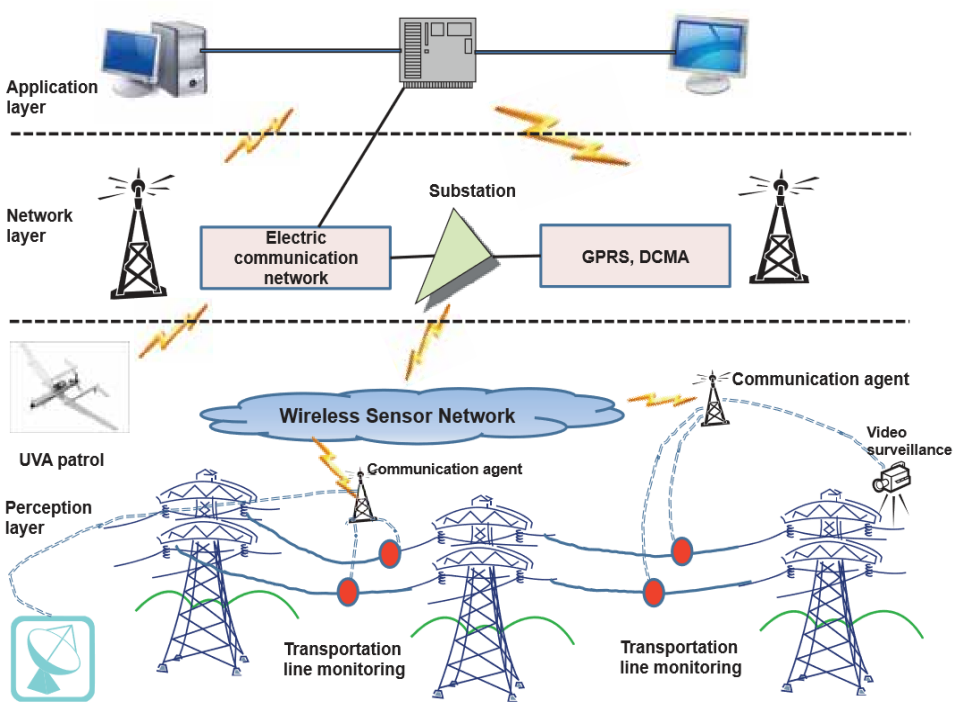
3.5 CASE STUDY #5: IOT SENSOR NETWORK

BACKGROUND

IoT Sensor Network can be defined as uniquely identifiable objects and their virtual representations in an "internet-like" structure (Ashon, 1999 - http://www.rfi djournal.com/articles/view?4986). Concept and research in IoT sensor networks are almost as old as the remote, electronic sensor development. Initially, it was used to gather battlefield intelligence and situation awareness in realtime.

At a high level, the three-tiered architecture for smart IoT Sensor Network is illustrated in Figure 5 (taken from IEC White Paper on IoT Sensor Networks [16]). The lowest level is data and information ingestion layer, it comprises largely of IoT sensor systems that are made of many heterogeneous sensors across the geographic span, including moving the sensor (e.g., mounted on UAV) that have the ability to form local sensor network and communicate within the local mesh. The second layer is the network communications layer. This layer is concerned with moving a large volume of appropriate data to the central data repository using various communication protocols/mechanisms (mobile, satellite, WiFi, etc.) The third layer is data, analytics, and access platform. At a high level, the platform has a realtime component that primarily deals with information aggregation, curation and storage/archival as well as 'data science' components that deal with business intelligence and decision optimization based on historical data. This also includes user interface and information dissemination interface.

FIGURE 5: FOOD ONTOLOGIES



Many domain-specific applications can be built with intelligent sensor networks supporting industry, smarter cities, and defense. The primary use cases for Sensor Network umbrella fall into the following categories:

- Industrial—Equipment Monitoring, Process Monitoring, Quality Control
- Infrastructure—Condition Monitoring of Roads, Bridges, Dams, Power Transmission, etc.
- Smarter Cities—Air and Water Quality, Traffic Monitoring and Congestion Control, Public Safety and Crowd Control, etc.

BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

The data governance and management challenges in Sensor Network use cases vary by how wide and heterogeneous the network is. Some of the key challenges regardless of scope and scales are listed as follows:

1) Ownership of data—The various sensors participating in-network might be owned by different entities and so the question of who owns the data aggregated over the network is an issue.

2) Ownership of analytics—Moreover, the result of the analytics on this aggregated data shared among the network participants is going to be complicated.

3) Adverse impact with aggregation—It is easy to control and confirm that any individual sensor is not collecting information that can affect the individual adversely. However, if the analytics on the shared data created with Sensor Network results in personally-identifiable information (PII) or other results that can affect individuals adversely, it would be very hard to control the outcome.

BIG DATA METADATA MANAGEMENT CHALLENGES

The data sources for an IoT Sensor Network are varied and include multiple contributors and users—primarily sensors measuring physical quantities (temp., flow, vibrations, speed, etc.) as well as locations, images, and video and audio signatures. The naming and definition of data entities are different for the same concept. Moreover, a data representation of the same sounding entity might be different depending on the purpose and technologies involved (e.g., vehicle identifier—is it license plate or VIN?). Also, the methodology to measure or capture the quantity might be different leading to different accuracy and tolerance for the measurement (e.g., laser vs. ultrasound to capture speed) and the unit of measure varies across boundaries (e.g., lb vs. kg of CO_2 emission equivalent).

BIG DATA MASHUP CHALLENGES

Data Mashup challenges for IoT network are very similar to the challenges of IoT sensor that are stand-alone. Working with multiple formats of input data, working with multiple representations of the same data, inconsistent formatting and nuances of human-generated data (e.g., voice).

On top of these usual and similar challenges, the IoT sensor network poses the following two additional complications:

1) Due to the massive amount of data, a lot of processing and mashing has to happen at the edge (as opposed to central data store) and only a selected amount of data should be brought back into the central data store.

2) Beyond the data formats and representation, different technologies and communication protocols involved in sensor-to-sensor communication alter the raw data transmission rates and formats (REST APIs, mobile protocols, and IoT devices communicating over proprietary protocols), posing additional issues with data mashup.

BIG DATA ANALYTICS CHALLENGES

Other than algorithmic complexity and issues relating to metadata and normalization, the following are key analytics challenges:

1) The amount of data involved is massive and growing at a fast rate (this is projected to be in the ZB range soon). Without automation, it will not be humanly possible to analyze and create analytical models with this amount of data.

2) If the network of sensors and machines involved in the process start making autonomous decisions based on analytics, the liability and legality of such a situation and the role of the machine "owner/operator" vs. the "autonomous decision power of the machine" could pose thorny issues.

DATA CHARACTERISTICS

Multiple formats—Structured (CSV), Semi-Structured (JSON, XML) and Unstructured (Video, Voice, Image, TXT).

3.6 CASE STUDY #6: SMART CITIES

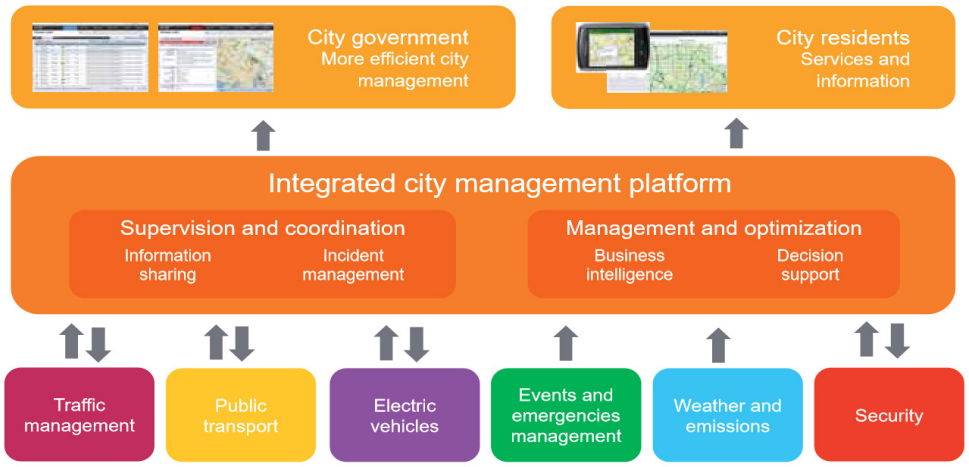
BACKGROUND

Smart Cities is a term that falls into the category coined by a famous U.S. Judge as "can't define it but I know it when I see it." Cities have been a hub of economic and cultural development but the increasing pace of urbanization all around the world is putting a lot of pressure on urban development. The use of technology on urban development was highlighted under the term "smart city" coined in early 1990 (Gibson et al., 1992 [17]). Among the various definitions, the one offered in Caragliu et al., 2011 [18] is most closely related to the pressure due to urbanization, where it defines that a city is smart when the aim of investing in cyberinfrastructure (ICT—Information Communication Technology) is to encourage sustainable economic growth, a better quality of life for citizens, and efficient management of natural resources (air, water, and energy).

At a high level, the three-tiered architecture for smarter cities is illustrated in Figure 6 (taken from IEC White Paper on Smart Cities [19]). The lowest level is the data and information ingestion and communication layer, it comprises largely of IoT systems with some integration of Social Media and other information sources.

The second layer is a platform that consists of multiple data stores and processor nodes divided among many different agencies and domain-specific applications. At a high level, the platform has a realtime component that primarily deals with information sharing between parties and efficient management of immediate actions or incidents as well as "data science" components that deal with business intelligence and decision optimization based on historical data. The third layer consists of user interface and information dissemination—at a high level, one set of interfaces for government and other city management agencies for 360 views of city operations and another set of interfaces for citizens, businesses, and other stakeholders for relevant and timely information related to services.

FIGURE 6: SMART CITIES



Many domain-specific applications can be built in the integrated city management platform tier to support intelligent service provision and incident response. The primary use cases for a Smart Cities umbrella fall into the following categories:

- Efficient Delivery of Citizen Services—Education, Utilities, Traffic and Congestion Management, etc.
- Incident Management—Police, Firefighter, Crowd Control, etc.
- Environment Impact—Carbon Emission, Air Quality, Water Quality, etc.

BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

Data governance and management challenges in Smarter Cities' use cases vary by how widely adopted the use case is within each Smarter City umbrella area. Some key challenges regardless of scope and scales are as follows:

1) Management of data from heterogeneous data sources like various sensors, video cameras, social media, and local/national agencies DB for both discrete and streaming mode as well as historical data aggregations is challenging.
2) Since the various applications have different levels of rights/authorization to various types of data sources, managing entitlement as well as data masking or aggregation required for different applications is another challenge.
3) Data collected in such a framework can potentially have characteristics that are covered by PII, HIPPA or GDPR and other similar regulations. Even when not covered by such regulations, the data can be very potent in identifying social, financial, political, health, etc. characteristics of individuals and if misused can cause great harm.

BIG DATA METADATA MANAGEMENT CHALLENGES

The data sources for a Smarter Cities framework are varied and include multiple contributors and users, primarily government or quasi-government agencies. The naming and definition of data entities are different for the same concept. Moreover, data representation of a similar sounding entity might be different depending on the purpose and technologies involved (e.g., vehicle identifier—is it license plate or VIN?). Additionally, the methodology to measure or capture a quantity might be different leading to differences in accuracy and tolerance for the measurement (e.g., laser vs. ultrasound to capture speed), and the unit of measurement may vary across boundaries (e.g., lb vs. kg of CO_2 emission equivalent).

BIG DATA MASHUP CHALLENGES

Working with multiple formats of input data, working with multiple representations of the same data, different technologies and communication protocols involved (REST APIs vs. CSV and mobile protocols vs. IoT devices communicating over proprietary protocols), inconsistent formatting, and nuances of human-generated data (e.g., voice).

BIG DATA ANALYTICS CHALLENGES

Other than algorithmic complexity and issues relating to metadata and normalization, the following are key analytics challenges:

1) Making sure that judicious use of algorithms produces results that are unbiased and do not violate legal requirements of the various overlapping jurisdictions
2) Realtime decision-making is separated and narrower in scope compared to long-term analysis and its impact on policymaking
3) The data samples are representative to generalize the use of results across the stakeholder base
4) The objective functions represent the interest of all stakeholders

DATA CHARACTERISTICS

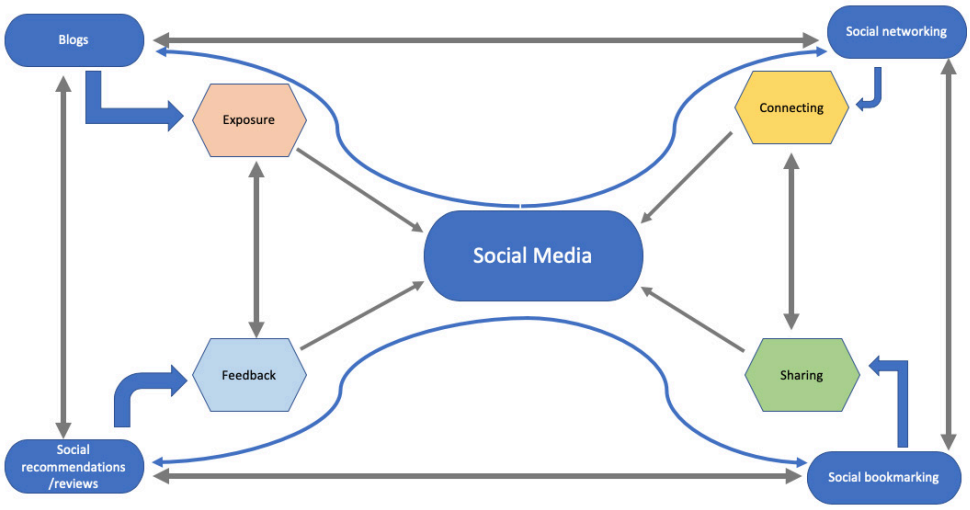
Multiple formats—Structured (CSV, Relational Databases), Semi-Structured (JSON, XML), and Unstructured (Video, Voice, Image, TXT).

3.7 CASE STUDY #7: SOCIAL MEDIA

BACKGROUND

A social media ecosystem and model are composed of four waves (see Figure 7): (1) an exposure wave, (2) a social networking wave, (3) a feedback wave, and (4) a sharing wave. It is the consequence of users' behavior and expectations. The behaviors of users impact the amount of content named social big data, their added value, and related quality. Users' expectation impacts the number of players. Currently, the social big data is mostly covered by blogs (Google), social networking (e.g., Facebook, Twitter), social bookmarking, and social recommendations/reviews (Quora, Pinterest, TripAdvisor). The social media ecosystem has been evolving to support increasing demands of social users with Web 2.0 (e.g., semantic Web) and Web 3.0.

FIGURE 7: DESCRIPTION OF THE SOCIAL MEDIA MODEL



BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

The primary challenge is controlling the message. Social media moves rapidly and systems and procedures are needed to respond to social media in realtime and avoid legal exposure from improper communications. A governance and risk management strategy are key to being able to be nimble and responsive, but also to manage legal and business risks.

BIG DATA METADATA MANAGEMENT CHALLENGES

Social media data sources are varied and include multiple contributors and users. It influences various business aspects, including human resources. We can classify big data as personal and professional social media data. It impacts metadata and its management. In addition, there are several legal challenges [20] that affect metadata management. Those challenges include social media and law, the public order in a virtual space, and private law responses to social media.

BIG DATA MASHUP CHALLENGES

Social media presents a way to discover, report, and share different types of events. Social media can be considered as a dynamic source of information that enables individuals to stay informed of real-world events. In this use case, the focus is on the detection of events from social media [21], [1]. Challenges mainly include the noise of data within social media and the semantic detection of dimensions such as topic, time, and location [22].

BIG DATA ANALYTICS CHALLENGES

Social media big data analytics process involves four distinct steps: (1) data discovery, (2) collection, (3) preparation, and (4) analysis. While there is a great deal of literature on the challenges and difficulties involving specific data analysis methods [23], there hardly exists research on the stages of data discovery, collection, and preparation. The volume scale of social media big data is the first challenge to impact the stages of data discovery, collection, and preparation.

DATA CHARACTERISTICS

Streaming text, images, audios, and videos.

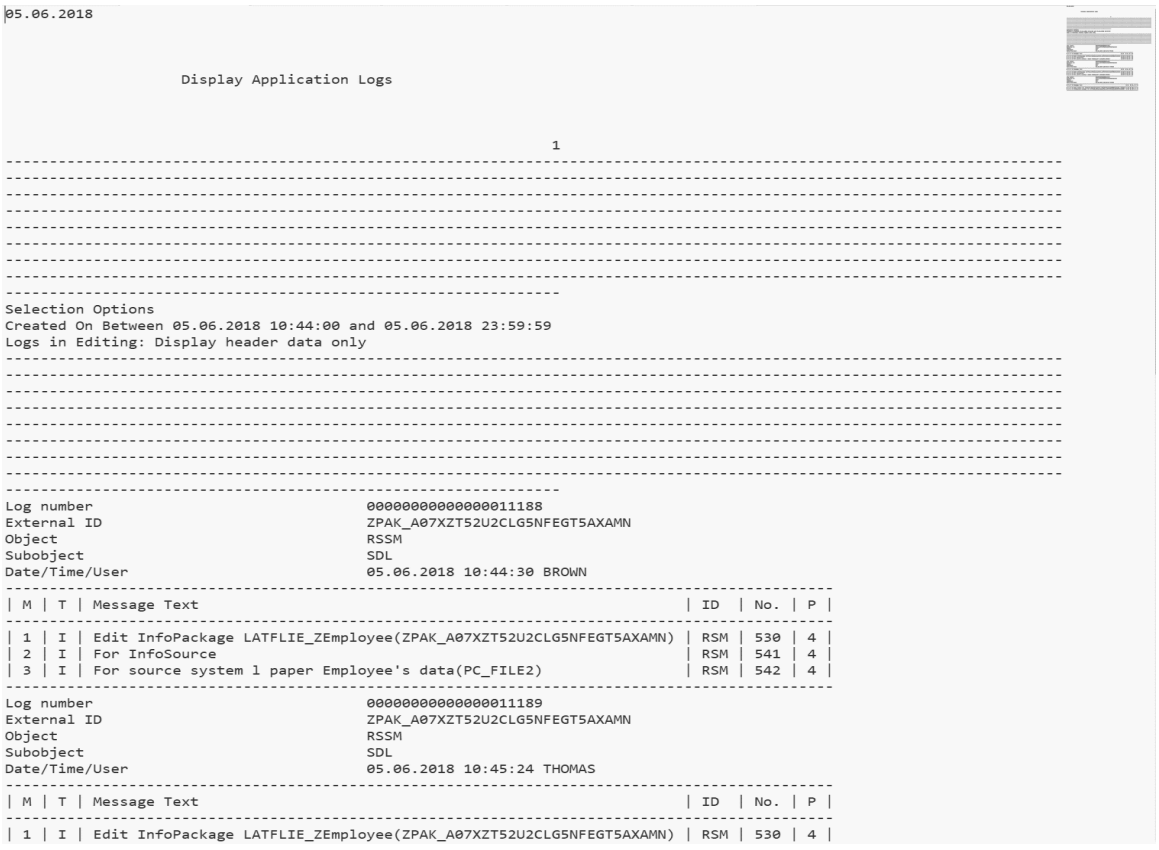
3.8 CASE STUDY #8: ANALYSIS OF APPLICATION LOGS IN OUTSOURCED SCENARIOS

BACKGROUND

Given the strategic relevance of ICT infrastructures and the ever more stringent constraints of regulatory audits, the detection of security incidents from application logs is a more and more relevant topic for any organization.

The problem can be described with a simple example. Bob is leading the monitoring department of an organization requiring auditing of its application logs. His goal is detecting and acting on security incidents of a Business Information Warehouse, connected to an ERP system. He receives application logs from the system and wants to get an alert when a security incident occurs. Figure 8 shows raw logs coming from a pilot run internally to the TOREADOR project [24] by SAP.

FIGURE 8: RAW APPLICATION LOGS FROM SAP



```
05.06.2018

                    Display Application Logs


                                                                1
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
-------------------------------------------------------------
Selection Options
Created On Between 05.06.2018 10:44:00 and 05.06.2018 23:59:59
Logs in Editing: Display header data only
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
--------------------------------------------------------------------------------------------------------------------------------
-----------------------------------------------------------
Log number                          00000000000000011188
External ID                         ZPAK_A07XZT52U2CLG5NFEGT5AXAMN
Object                              RSSM
Subobject                           SDL
Date/Time/User                      05.06.2018 10:44:30 BROWN
-------------------------------------------------------------------------------------------
| M | T | Message Text                                                        | ID  | No. | P |
-------------------------------------------------------------------------------------------
| 1 | I | Edit InfoPackage LATFLIE_ZEmployee(ZPAK_A07XZT52U2CLG5NFEGT5AXAMN) | RSM | 530 | 4 |
| 2 | I | For InfoSource                                                      | RSM | 541 | 4 |
| 3 | I | For source system l paper Employee's data(PC_FILE2)                 | RSM | 542 | 4 |
-------------------------------------------------------------------------------------------
Log number                          00000000000000011189
External ID                         ZPAK_A07XZT52U2CLG5NFEGT5AXAMN
Object                              RSSM
Subobject                           SDL
Date/Time/User                      05.06.2018 10:45:24 THOMAS
-------------------------------------------------------------------------------------------
| M | T | Message Text                                                        | ID  | No. | P |
-------------------------------------------------------------------------------------------
| 1 | I | Edit InfoPackage LATFLIE_ZEmployee(ZPAK_A07XZT52U2CLG5NFEGT5AXAMN) | RSM | 530 | 4 |
```

Alice is leading the AI department of a company offering Big Data Analytics-as-a-service. Alice can help Bob but she has to train a machine-learning algorithm. At the same time, to train the model, Alice needs access to the data, which Bob cannot provide.

They then settle on the following approach: Bob will provide anonymized data to Alice so that Alice can train a classifier. Alice will train a classifier on the anonymized data, with the intent to deploy the trained classifier in a platform where Bob can consume it directly.

There are two application-specific security problems they want to tackle.

The first one is about privilege abuse: by definition, administrators of a Business Information Warehouse can access all content, including sensitive data. It is not possible from the log to deduce if the currently logged-in user is legitimate to execute the action, but it is possible to mark as malicious users usually performing administrative actions when they perform a read action on a sensitive table. This incident shall be reported as an occurrence of a 'nosy admin'.

The second one is about privilege escalation: non-administrator users are by definition not authorized to perform any administrative task; such executions should be blocked by the access control mechanism of the Information Warehouse. However, it might be possible for a user to manage to gain the extra privilege through an unknown attack path (maybe by convincing an administrator to add his user as a substitute, or by discovering and exploiting a zero-day attack). Since these unknown attack vectors cannot be prevented, Alice plans on relying on detection.

The initial step for Alice is to apply unsupervised learning in order to create the training dataset. Using clustering she can group similar cases and label them using her expertise. The number of clusters is much smaller than the actual dataset size, allowing to considerable time to be saved.

Before training the classifier, Alice needs to prepare the sample data she received by running multiple preparation steps. Once the classifier performance is considered satisfactory, Alice can publish the trained model. Now Bob can productively use the service pipeline designed by Alice in a platform he controls.

BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

The challenges related to governance are mainly dependent on the ownership of data, as customers may be interested in outsourcing the audit of application logs but at the same time, the risk of data leakage can limit their availability. Making application logs auditable even when anonymized or obfuscated is then a crucial challenge to support the proposed scenario.

Dividing the auditing task into a training and an alerting (basically, a classification) stage, makes it possible to deploy them on infrastructures subject to different controls, limiting the risk of leakage. However, the service provider may also be concerned about ownership of analytics or even by the ownership of specific boosting strategies incarnated in a pipeline of services and analytics. The deployment procedure is then expected to be modular, to adapt to the specific needs of a customer, but not necessarily transparent to all the actors involved.

BIG DATA METADATA MANAGEMENT CHALLENGES

Addressing the proposed case study by a modular solution metadata is of paramount importance. It is necessary to disambiguate the different columns of a record, their confidence level, their format and eventually their linkage conditions.

BIG DATA MASHUP CHALLENGES

Data mashup is relevant in this case study as multiple log files may be integrated during an audit. A domain-specific language for addressing data integration by a fast configuration is, for example, an important challenge to highlight.

BIG DATA ANALYTICS CHALLENGES

The proposed case study underlines that metadata must support the reuse of data analytics and services. Consider for example that a new request emerges for the organization concerned with auditing. For example, Bob is now interested in dormant accounts, i.e., accounts that are not being used for a long duration and which suddenly become active. Detecting them may require deriving a new feature from the dataset, for example, an "elapsed time" column, which will contain, for each entry, how many seconds passed since the same user performed an action. This implies a new data preparation must be included in the pipeline; however, the rest of the pipeline is not modified. A good set of metadata can allow deploying the new pipeline simply by reusing the previous set of specifications with the required addition or removal of services. This requires the system to support a representation of Big Data pipelines to support their reproducibility and verifiability.

DATA CHARACTERISTICS:

Log files mainly in ASCII text

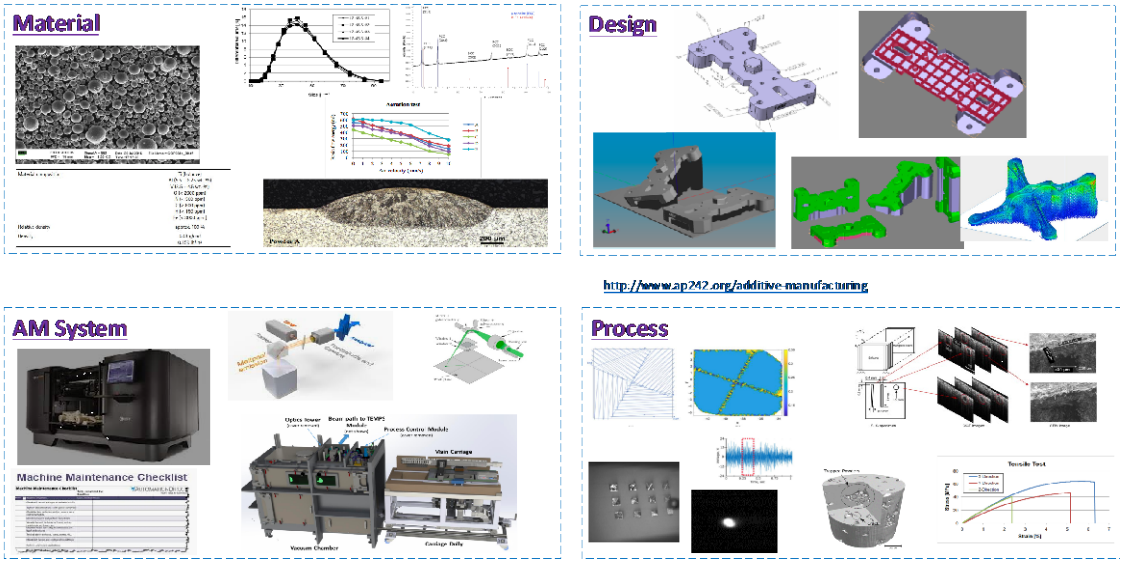
3.9 CASE STUDY #9: DATA INTEGRATION AND MANAGEMENT FOR ADDITIVE MANUFACTURING

BACKGROUND

Additive manufacturing (AM) processes build parts layer-by-layer directly from 3D models. Compared to traditional manufacturing processes where objects are shaped or cut out of blocks of solid materials with well-understood material properties, AM enables the fabrication of complex heterogeneous parts on demand. The advantages of AM make it an attractive alternative for high-value, low-volume production.

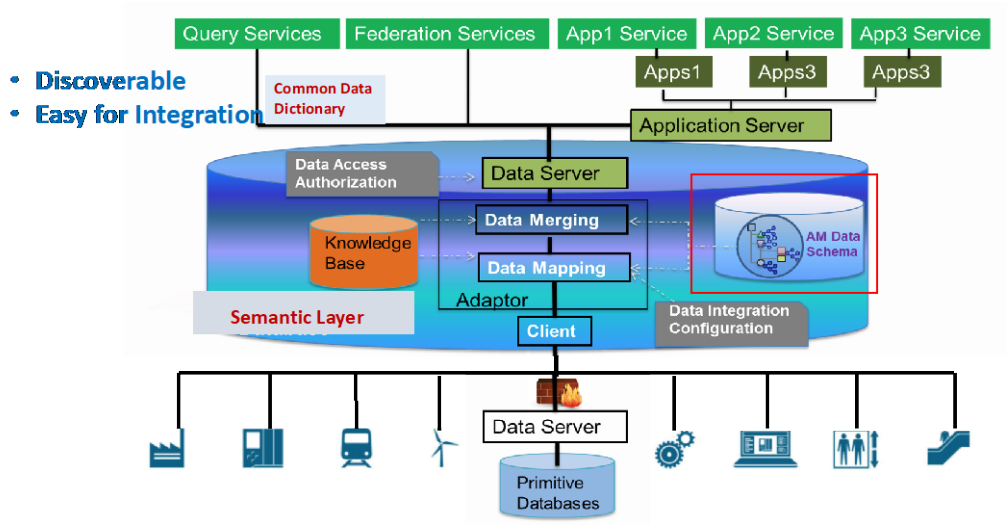
A large amount of data are generated from Additive Manufacturing (AM) lifecycle and value chain activities. Figure 9 shows the broad landscape for AM data, covering four various domains—material, machine, design, and process. These data, appropriately integrated, play a critical role in streamlining the AM development process, from design to fabrication and part certification. AM data, especially those generated from in-situ monitoring and ex-situ inspections, embodies all the 4 V's characteristics of Big Data—volume, velocity, variety, and veracity. For example, the amount of data produced is estimated at 3000 TB for qualification for an additively built aircraft component. Moreover, in-situ process monitoring during a powder bed fusion process can produce data at GB/sec. These big data are usually collected, archived, and analyzed for AM process understanding, process variation management, and part qualification. In order to accelerate AM development and deployment, better data models and best practices for data integration and management are needed for more effective and efficient curation, sharing, processing, and use of AM data for fast knowledge discovery.

FIGURE 9: AM DATA LANDSCAPE



Standards are the enabler for effective AM data integration and federation. Two-stage approaches are identified, as shown in Figure 10. First, build a common data dictionary and semantics for easy discovery of the existing AM data. Then develop a common information model by leveraging existing data standards and/or neutral format(s) for data exchange. The common data model and neutral data exchange formats will enable a virtual collaborative AM data repository(ies) and common interfaces for data access.

FIGURE 10: ADDITIVE MANUFACTURING DATA INTEGRATION



BIG DATA GOVERNANCE MANAGEMENT CHALLENGES

There are no established AM data governance management practices. However, both data quality and data traceability are critical in data-driven decision making for AM. In addition, the cost of generating AM data is high. However, the mechanisms and standards for AM data sharing are not available. Data security and privacy considerations are also important factors that affect the scale of AM data exchange.

BIG DATA METADATA MANAGEMENT CHALLENGES

AM covers a broad range of activities, from material development and characterization, to process control and certification, to part design, fabrication, testing, and qualification. While the data used and generated from most of the activities are similar to those of the traditional manufacturing systems, there are also many greenfield activities, for example, AM material reuse, function material design and the corresponding complex process settings, multi-modal and multi-scale high speed in-process monitoring, as well as highly demanded microstructure characterization and non-destructive evaluation for part inspection. There are big gaps in meta information modeling for these data, as well as the mechanism for big dataset tagging and naming necessary for data sharing and integration among the AM stakeholders.

BIG DATA MASHUP CHALLENGES

The broad scope of AM data suggests overwhelmingly different data formats. Although STL is a commonly used data format for part design, various CAD formats exist for both concept design and detail design. For process control, proprietary data formats dominate the market. AM process setting, process monitoring, and build pedigree can be captured in excel sheets, word document, pdf files, relational database or NoSQL systems. In addition, NDE equipment usually provides multiple formats for data exporting. Wrangling with various types of data, multiple representations of the same type of data, and proprietary data formats is a great challenge for AM data use. Mashup with text-based data, which is frequently used to capture manual tasks performed during the AM development process, is another big challenge.

BIG DATA ANALYTICS CHALLENGES

The significant AM big data analytics challenges lie in the following three areas:

1) Data registration—When AM data are used for process understanding, process control, and part qualification, all of the data should be registered correctly using a common coordinate system. This is very challenging considering the high spatial resolution (e.g., in microns) of the data generated from varying setup of in-situ and ex-situ measurement devices.

2) Multi-rate and multi-scale data fusion—In-situ and ex-situ AM data are usually generated using multi-modal sensing techniques focusing on different spatial scales, from micron to tens of centimeters, and acquired at various sampling rates, from several HZs to several million Hz. Fusing multi-rate and multi-scale data is a challenge engineering task for data-driven AM decision making.

3) Semantic data analytics—Experiments are always very expensive for AM. Using physics-based modeling to assist AM big data analytics is a viable path. At this stage, both research and best practices are needed in this area for reliable deployments of semantic big data analytics results in AM production.

DATA CHARACTERISTICS

Multiple formats—Structured (CSV, 3D models, Relational Databases), Semi-Structured (JSON, XML) and Unstructured (Video, Image, TXT).

4 BDGMM TECHNICAL REQUIREMENTS

GOVERNANCE MANAGEMENT REQUIREMENTS

Governance represents a collective agreement in routines, specified actions, decision making, and plans. Gartner provides an insightful definition of "information governance" as "the specification of decision rights and an accountability framework to ensure appropriate behavior in the valuation, creation, storage, use, archiving and deletion of information." [25] Governance, in terms of top-level oversight, usually involves a group of members with representative authority, forming a committee or board, and often one person is known as the chair. Governance Management Requirements in Big Data is a combination of such governance strategies, depending on environment and agency requirements. Agencies may follow these practices to organize and manage their big data.

METADATA MANAGEMENT REQUIREMENTS

Metadata requires the management of a series of intersecting standards. These include 1) data structure standards, 2) content value standards, 3) communication standards, and 4) syntax standards. With the exception of syntax, these different aspects of metadata needing to be managed are designed and endorsed by separate bodies, but they are often single metadata renderings with big data.

With the big data governance and metadata challenges captured in Section 3, this following subsection will study each case study and evaluate their requirements in the areas of (a) Big Data requirements, (b) Big Data governance management Requirements, (c) Big Data metadata management, and (d) Big Data infrastructure requirements. Appendix A lists each case study with its extracted requirements. Once detailed analysis is performed, aggregation of case study-specific requirements allows for the formation of more generalized requirements. These generalized requirements are listed next.

There were twenty-six (out of 87+ specific requirements) general requirements in five categories as shown in the following subsections.

4.1 GOVERNANCE REQUIREMENTS (GR)

GR1 (Access Rights)	a) Access Rights – needs to support signing Data Release Agreement before access datasets. (M96, M97, M117, M118, M134)
GR2 (Access Method/Protocol)	a) Needs to support httpd-based API like RESTful, SOAP, etc. (M96, M117, M118) b) Needs to support web portal navigation to access appropriate datasets (M97) c) Needs to support standard interfaces based on open networking protocols to access IoT sensors data (M112)
GR4 (Revision Control)	a) Needs to support revision control (M96, M117, M118, M134)

4.2 METADATA REQUIREMENTS (MR)

MR1 (Metadata Schema/Model)	a) Needs to support any schema/model (M96, M97, M112, M117, M118) b) Needs to support additive schema - graph-based data model (M134)
MR2 (Schema Data Elements)	a) Needs to support online data elements definition (M96, M97, M117, M118, M134)

MR3 (Schema Data Values)	a) Needs to support any data unit measurements (M96, M97, M117, M118)
MR4 (Derived Data)	a) Needs to support labeling of raw data for advanced analysis (M112)
MR5 (Context Data)	a) Needs to support linking data with different contexts can enable the discovery of new patterns and relationship (M112, M134)

4.3 DATA MASHUP REQUIREMENTS (DMR)

DMR1 (Mashup by dataField/ID/etc.)	a) Needs to support data dictionary for any given data field (M96, M97, M117, M118) b) Needs to support unique ID for each unique data field (M96, M97, M116, M118) c) Needs to support cross-reference between datasets using unique ID (M96, M97, M117, M118, M134)
DMR2 (Data source)	a) Needs to support any streaming data source (M96, M97, M112, M134, b) Needs to support any heterogeneous datasets (M96, M97)
DMR3 (Access Method/Protocol)	a) Needs to support vendor-independent interfaces to access device information (M112)
DMR4 (Language)	a) Needs to support integration language to aggregate data (M134)

4.4 ANALYTICS REQUIREMENTS (AR)

AR1 (tools/utilities)	a) Needs to support deployable analytics tools (M96, M97, M112, M117, M118, M134) b) Needs to support the integration of 3rd parties of system utilities and tools (M96, M97, M112, M117, M118)
AR2 (App context, semantics)	a) Needs to support the right context and semantics depending on the IoT application (M112, M134)
AR3 (Consistency)	a) Needs to support consistency between analytics and metadata can be checked (M134)

4.5 DATA CHARACTERISTIC REQUIREMENTS (DCR)

DCR1 (File Format)	a) Needs to support diverse file format like ASCII, (M96, M97, M112, M117, M118, M134)
DCR2 (Data Structure)	a) Needs to support CSV, TSV, JSON file structure (M96, M97, M117, M118, M134) b) Needs to support graph-based data model like JSON-LD, RDF (M134)
DCR3 (Data Storage)	a) Needs to support local file-based and streaming system (M96, M97, M117, M118) b) Needs to support cloud storage (M112)

5 RELEVANT STANDARDIZATION ACTIVITIES

Advance information technologies for Big Data, governance management, metadata management, enterprise management, and other related management are being developed rapidly. This section tries to identify the sample of the related standard technologies that can potentially contribute and enable effective and proactive Big Data governance and metadata management at the system architectural level that is scalable for Findability, Accessibility, Interoperability, and Reusability between corporate heterogeneous datasets and repositories independent from various application domains without worrying about data source and structure.

This section describes related standardization activities with governance management and metadata management from the Big Data perspective in an effort to identify standards gaps. The current content is based on an informal literature search by BDGMM members and contributions from other SDOs. Specific BDGMM related standards are being developed by a variety of well-established SDOs and industry consortia as outlined in Table 1. The following sections provide a sample of additional details of activities by organizations that relate to BDGMM.

TABLE 1: SAMPLE POTENTIAL SDOS WORK RELATED TO BDGMM

SDO/Consortium	Interests area on Standardization	Main deliverables
ISO/IEC JTC 1/SC 32 [26]	Data management and interchange, including database languages, multimedia object management, metadata management, and e-Business.	e-Business standards, including role negotiation; metadata repositories, model specification, metamodel definitions; SQL; and object libraries and application packages built on (using) SQL.
ISO/IEC JTC 1/SC 40 [27]	IT Service Management and IT Governance, including IT activities such as audit, digital forensics, governance, risk management, outsourcing, service operations, and service maintenance.	Governance of IT, maintenance and development of ISO/IEC 20000 on service management, IT-enabled services/business process outsourcing, and IT service management of infrastructure.
ISO TC 20/SC 13 [28]	Space data and information transfer systems, including gather best practices by the major space agencies of the world to develop common solutions to the operation of space data systems.	Develops both the technical and the institutional framework for international interoperability to facilitate appropriate cross-support opportunities of space data systems.
Dublin Core Metadata Initiative (DCMI) [29]	The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. The name “Dublin” is due to its origin at a 1995 invitational workshop in Dublin, Ohio; “core” because its elements are broad and generic, usable for describing a wide range of resources.	The terms in DCMI vocabularies are intended to be used in combination with terms from other, compatible vocabularies in the context of application profiles and based on the DCMI Abstract Model [DCAM].
W3C Data Catalog Vocabulary (DCAT) [30]	DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web.	Describes datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs.
Common European Research Information Format (CERIF) [31]	Promotes cooperation within and shares knowledge among the research information community and interoperability of research	Provides the framework for the flow of information between a broad variety of stakeholders: researchers, research

Table continues

SDO/Consortium	Interests area on Standardization	Main deliverables
	information through CERIF, the Common European Research Information Format.	managers, and administrators, research councils, research funders, entrepreneurs, and technology transfer organizations.

5.1 ISO/IEC JTC 1/SC 32/WG2 METADATA MANAGEMENT

ISO/IEC JTC 1/SC 32, titled "Data management and interchange," currently WG2 on Metadata Standards has focused on three major areas: (1) Specification of generic classes of metadata and frameworks for representing the meaning and syntax of data, including metamodels, ontologies, processes, services, and behavior and mappings between them, (2) Specification of facilities to manage metadata including registries and repositories, and (3) Specification of facilities to enable electronic metadata exchange over the Internet, within the cloud, and other information technology telecommunications avenues. Relevant projects that are potentially beneficial to BDGMM include the following:

- ISO/IEC 11179:2019 Metadata registries (MDR)—A framework for registering and managing metadata about Datasets
  - 11179-2:2019 [32] Part 2, Classifications: describes the registration of classification schemes and using them to classify registered items in an MDR. Any metadata item can be made a Classifiable_Item so it can be classified, which can include object classes, properties, representations, conceptual domains, value domains, data element concepts and data elements themselves.
  - 11179-3:2013 [33] Part 3, Registry Metamodel and basic attributes: specifies the structure of a metadata registry in the form of a conceptual data model, which includes basic attributes that are required to describe metadata items.
  - 11179-3:2019* Part 3, Registry Metamodel—Core model: specifies the structure of a metadata registry in the form of a conceptual data model.
  - 11179:7:2019* Part 7, Metamodel for dataset registration: provides a specification in which metadata that describes datasets, collections of data available for access or download in one or more formats, can be registered.
- ISO/IEC 21838:2019, Top-level ontologies (TLO)—A framework and specification and standardization of top-level ontologies
  - 21838-1* Part 1, Requirements: specifies the requirements of the relationship between top-level ontology and domain ontologies along with the formulation of definitions and axions in ontologies at lower levels.
  - 21838-2* Part 2, Basic Formal Ontology (BFO)—provides definitions of its terms and relational expressions and formalizations in Web Ontology Language (OWL) and in Common Logic (CL).
- ISO/IEC TR 19583, Concepts and usage of metadata
  - 19583-1* Part 1, Metadata concepts: provides the means for understanding the concept of metadata, explains the kind and quality of metadata necessary to describe data, and specifies the management of that metadata in a metadata registry (MDR).

- 19583-2* Part 2, Metadata usage: describes a framework for the provision of guidance on the implementation and use of the registries specified in ISO/IEC 11179, information technology—Metadata registries (MDR), and ISO/IEC 19763, Information technology—Metamodel framework for interoperability (MFI).

- ISO/IEC 19763, Metamodel framework for interoperability (MFI)—A framework for registering models and mappings between ontologies

  - 19763-1:2015 [34] Part 1, Framework—provides an overview of the underlying concepts, the overall architecture and the requirements for the development of other standards within the MFI family are described.

  - 19763-3:2010 [35] Part 3, Metamodel for ontologies registration—specifies a metamodel that provides a facility to register administrative and evolution information related to ontologies, independent of the languages in which they are expressed. A registry that conforms to this document, together with repositories that contain actual ontologies, makes it possible for users to gain the benefits of interoperation among application systems based on ontologies.

- ISO/IEC 11404:2007 [36], General purpose datatypes (GPD)—specifies a collection of datatypes commonly occurring in programming languages and software interfaces including both primitive and non-primitive datatypes, in the sense of being wholly or partly defined in terms of other datatypes.

Note: "*" = under development

5.2 ISO/IEC JTC 1/SC 40/WG1 GOVERNANCE OF INFORMATION TECHNOLOGY

ISO/IEC JTC 1/SC 40, titled "IT service management and IT governance," currently WG1 leads the development of standards, tools, frameworks, best practices and related documents on the governance of information technology. This includes providing guidance on the inter-relationships between organizations, stakeholders, and information technology [WG 1 N7 Recommendation on ToR]. Relevant projects that are potentially beneficial to BDGMM include the following:

- ISO/IEC 38500:2015, Governance of IT for the organization—Guiding principles for members of governing bodies of organizations on the effective, efficient, and acceptable use of information technology (IT) within their organizations.

  - ISO/IEC 38505-1:2017 [37] Part 1, Application of ISO/IEC 38500 to the governance of data—applies to the governance of the current and future use of data that is created, collected, stored or controlled by IT systems, and impacts the management processes and decisions relating to data.

  - ISO/IEC 38505-2 [38] Part 2, Implications of ISO/IEC 38505-1 for data management—identifies the information that a governing body requires in order to evaluate and direct the strategies and policies relating to a data-driven business and the capabilities and potential of measurement systems that can be used to monitor the performance of data and its uses.

5.3 ISO TC20/SC13 SPACE DATA AND INFORMATION TRANSFER SYSTEMS

Main contributions are from the Consultative Committee for Space Data Systems (CCSDS) [39], which was formed in 1982 with the goal of gathering best practices by the major space agencies of the world and developing a common solution to the operation of space data systems. While the CCSDS is concerned primarily with space data, the work of ISO TC20/SC13 is applicable well beyond the space data community. The National Archives and

Records Administration and other digital cultural organizations also participate in the group. Much of the work is focused on long-term (long enough to be concerned about obsolescence and usability) preservation and use of information, and interoperability between data repositories, data producers, and their users. Relevant projects include the following:

- ISO 14721 [40], Reference Model for an Open Archival Information System (OAIS) "The explosive growth of information in digital forms has posed a severe challenge not only for traditional Archives and their information providers, but also for many other organizations in the government, commercial and non-profit sectors. These organizations are finding, or will find, that they need to take on the information preservation functions typically associated with traditional Archives because digital information is easily lost or corrupted. The pace of technology evolution is causing some hardware and software systems to become obsolete in a matter of a few years, and these changes can put severe pressure on the ability of the related data structures or formats to continue effective representation of the full information desired. Because much of the supporting information necessary to preserve this information is more easily available or only available at the time when the original information is produced, these organizations need to be active participants in the Long-Term Preservation effort, and they need to follow the principles espoused in this OAIS reference model to ensure that the information can be preserved for the Long-Term. Participation in these efforts will minimize the lifecycle costs and enable effective Long-Term Preservation of the information." This reference model:
  - Provides a framework for the understanding and increased awareness of archival concepts needed for long-term digital information preservation and access;
  - Provides the concepts needed by non-archival organizations to be effective participants in the preservation process;
  - Provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives;
  - Provides a framework for describing and comparing different Long-Term Preservation strategies and techniques;
  - Provides a basis for comparing the data models of digital information preserved by archives and for discussing how data models and the underlying information may change over time;
  - Provides a framework that may be expanded by other efforts to cover long- term preservation of information that is NOT in digital form (e.g., physical media and physical samples);
  - Expands consensus on the elements and processes for Long Term digital information preservation and access, and promotes a larger market that vendors can support;
  - Guides the identification and production of OAIS-related standards [40].
- ISO 16363, Audit and Certification of Trustworthy Digital Repositories (TDR). The OAIS Reference Model was adopted by many data repositories and they began to assert they were "OAIS-compliant" and thus trusted or trustworthy digital repositories. At the time ISO 14721 was first developed, there was no standard to assess compliance with the Reference Model. ISO 16363 was developed to fill that gap. In addition to providing for the audit and certification of TDRs, the standard can serve as a roadmap for developing the policies, procedures, staffing, and infrastructure for standing up a TDR that is compliant with the OAIS Reference Model [41].
- ISO 20652, Producer-Archive Interface Methodology Abstract Standard (PAIMAS) [42]. "This

Recommended Practice identifies, defines, and provides structure to the relationships and interactions between an information Producer and an Archive. This Recommended Practice defines the methodology for the structure of actions that are required from the initial time of contact between the Producer and the Archive until the objects of information are received and validated by the Archive. These actions cover the first stage of the Ingest Process as defined in the Open Information System (OAIS) Reference Model." [3]

- ISO 20104 [43], Producer-Archive Interface Specification (PAIS). "This Recommended Standard provides the abstract syntax and an XML implementation of descriptions of data to be sent to an archive. These descriptions are negotiated agreements between the data Producer and the Archive and facilitate the production of agreed data by the Producer and validation of received data by the Archive. The Recommended Standard includes an abstract syntax for describing how these data will be aggregated into packages for transmission and one concrete implementation for the packages based on the XML Formatted Data Unit (XFDU) standard." [44]

5.4 DUBLIN CORE

This standard emerged as an attempt to produce a general metadata standard for describing webpages. Originally in 1995, Dublin Core (DC) [29], had 13 elements (attributes) that were later extended to 15 in 1998 and again as Qualified DC to 18 including Audience, Provenance, and Rights Holder. DC was initially based on text and HTML but evolved to include the concept of namespaces for elements (with approved terms for the semantics of element values) coincident with the move to Qualified DC and towards using XML. Later the community realized that relationships among elements were important and an RDF version was proposed. However, the major volume of DC metadata is still in HTML format and so the benefits of using namespaces—and later relationships—are not realized. Indeed, this is the major criticism of DC: it lacks referential integrity and functional integrity. The former problem means that it is hard to disambiguate element values in repeating groups. An example of the latter is 'contributor' regarded as an attribute of the digital object being described; whereas, in fact, the contributor is a digital object in its own right and so has a relationship with the digital object being described.

5.5 W3C DATA CATALOG VOCABULARY

The original Data Catalog Vocabulary (DCAT) was developed at the Digital Enterprise Research Institute (DERI), refined by the eGov Interest Group [46], and then finally standardized in 2014 by the Government Linked Data (GLD) [47] Working Group leading to W3C recommendation. It is based on Dublin Core but adopts linked data principles with a schema including links between a dataset and a distribution (of that dataset—i.e., a replicate or version), between a dataset and a catalog and also between a dataset and an agent (person or organization). The use of relationships or links partially avoids the problems with Dublin Core—but not completely.

5.6 COMMON EUROPEAN RESEARCH INFORMATION FORMAT

Common European Research Information Format (CERIF) [31] is a European Union Recommendation to Member States (i.e., a standard). CERIF91 (1987–1990) was quite like the later Dublin Core (late 1990s) but was tested and found to be inadequate. CERIF2000 (1997–1999) used full enhanced entity-relationship (EER) modelling with Base entities related by Linking entities with role and temporal interval (i.e., decorated first-order logic). In this way, it preserves referential and functional integrity. In 2002, the European Council requested euroCRIS to maintain, develop and promote CERIF. There are commercial CERIF systems, two of which were bought by Elsevier and Thomson-Reuters to include CERIF in their products.

5.7 DIGITAL OBJECT ARCHITECTURES

The Digital Object (DO) Architecture is a non-proprietary open architecture that manages information in digital form in a network environment, including in particular the Internet. A key concept in this architecture is the notion of a digital object. A digital object comprises one or more sequences of bits, or a set of sequences, incorporating a work or a portion of work or other information in which a party has rights or interests, or in which there is value, and structured in a way that is interpretable by one or more computational facilities in a network. Each DO has an associated unique persistent identifier, and can be signed and validated using PKI or other cryptographic methods.

What became known as the DO Architecture had its roots in work at Corporation for National Research Initiatives (CNRI) on mobile program technology, known as Knowbots, in the 1980s; and the overall architecture is described in a number of documents [48],[49],[50],[51]. A formal standard based on the DO Architecture was approved as ITU-T Recommendation X.1255 (9/2013) [52]. For purposes of X.1255, the term digital entity (DE) is substantially similar to concept of a digital object; and the approach has been summarized as follows:

Digital entities are the core element around which all other components and services are built and managed. Digital entities do not replace existing formats and data structures, but instead provide a common means for representing these formats and structures, allowing them to be uniformly interpreted and thus moveable in and out of various heterogeneous information systems and across changes in systems over time. This model, though simple at its core, is non-trivial in its detailed implementation, and includes a protocol for interacting with DEs through repositories.

The unique persistent identifier of a digital object (or digital entity) is the essential fixed attribute of the object. An identifier registration and resolution system, which resolves identifiers to current state information, e.g., current location(s), access controls and validation, about the entities, is one of the basic components of the architecture. Other required components are repositories that manage storage of digital objects and provide access to them based on their identifiers, and registries for discovering identifiers for sought after digital objects. All three of these components have been implemented and are in current use. For example, a reference implementation of the identifier/resolution system, known as the Handle System, was created by CNRI and now operates under the overall administration of the DONA Foundation based in Switzerland. The DOI System makes global use of the handle technology to manage published information. Many other applications of registries and repositories exist.

While a thorough discussion of implementation progress and remaining research issues lies outside the scope of the current document, two issues merit mention. The first is typing and type registries. Type/value pairs are attributes within the architecture; and digital information held in type/value pairs is interpreted according to their types. The types are also digital objects, and their identifiers resolve to type records held in type registries. Clients encountering new types can resolve their identifiers to acquire information about the newly encountered type and react accordingly. This level of typing has been the subject of multiple Research Data Alliance (RDA) Working Groups, and is now under consideration by the joint ISO/IEC Joint Technical Committee SC32, which is focused on data management and interchange.

Finally, the digital object interface protocol (DOIP), with a new implementation about to be released, provides a high-level persistent means of accessing digital objects. This protocol provides a way to access objects by use of their identifiers and enables an extensible set of operations, a few of which are mandatory, e.g., return a list of possible operations for the given object.

An operation on a digital entity involves the following elements:

- EntityID: the identifier of the digital entity requesting invocation of the operation;
- TargetEntityID: the identifier of the digital entity to be operated upon;

- OperationID: the identifier that specifies the operation to be performed;
- Input: a sequence of bits containing the input to the operation, including any parameters, content or other information; and
- Output: a sequence of bits containing the output of the operation, including any content or other information.

5.8 PERSISTENT IDENTIFIERS

The overarching goal of BDGMM is to maximize access to data across heterogeneous repositories while still adhering to protect the confidentiality and personal privacy. The objective is to improve the ability to locate and access digital assets such digital data, software, and publications while enabling proper long-term stewardship of these assets by optimizing archival functionality, and (where appropriate) leveraging existing institutional repositories, public and academic archives, as well as community and discipline-based repositories of scientific and technical data, software, and publications.

For the new global Internet, Big Data economy opportunity in Internet of Things, Smart Cities, and other emerging technologies and market trends, it is critical to have a standard data infrastructure for Big Data that is scalable and can apply the FAIR (Findability, Accessibility, Interoperability, and Reusability) data principle between heterogeneous datasets from various domains without worrying about data source and structure.

A very important component as part of the standard data infrastructure is the definition of new Persistent Identifier (PID) types. PIDs such as Digital Object Identifiers (DOIs) are already widely used on the Internet as durable, long-lasting references to digital objects such as publications or datasets. An obvious application of PIDs in this context is to use them to store a digital object's location and state information and other complex core metadata. In this way, the new PID types can serve to hold a combination of administration, specialized, and/or extension metadata. Other functional information, such as the properties and state of a repository or the types of access protocols it supports, can also be stored in these higher layers of PIDs.

Because the PIDs are themselves digital objects, they can be stored in specialized repositories, similar to metadata registries that can also expose services to digital object users and search portals. In this role, the PID types and the registries that manage them can be viewed as an abstraction layer in the system architecture and could be implemented as middleware designed to optimize federated search, assist with access control, and speed the generation of cross-repository inventories. This setting can enable data integration/mashup among heterogeneous datasets from diversified domain repositories and make data discoverable, accessible, and usable through a machine readable and actionable standard data infrastructure.

6 STANDARD TECHNOLOGY GAP ANALYSIS

A standard technology roadmap for Big Data governance and metadata management is critical to any enterprise organization. The roadmap explores the overall direction of how to cohesively manage the ever-increasing deluge of data produced from an organization's diverse products and services. The goal is to provide effective and efficient technical approaches on how to meet current needs while providing opportunity growth for the future.

From the new global Internet, Big Data economy opportunities from the emerging technologies and market trends, organizations considering implementing Big Data governance and metadata management assume there is a unified organizational buy-in with clear business missions and potential value propositions identified. Furthermore, the assumption is that there is strong management support with well-defined roles and responsibilities between business processing units, enterprise architects, and application teams to drive the corporate visions forward.

This section provides gap analysis from the standard technologies and infrastructure perspective and utilizes information gathered on data trend in Section 2 "Data Explosion," detailed Big Data governance and metadata management challenges in Section 3 "BDGMM Case Study," specific and general technical requirements in Section 4 "BDGMM Technical Requirements," insights and understanding on challenging problems and solution approaches in Section 5 "BDGMM Lessons Learned from Workshops, Hackathons, and Invited Speakers," and identifies existing available standards in Section 6 "Relevant Standardization Activities."

The next subsections will focus on the analysis of the following:

a) Analyzing general data trends and their unique characteristics

b) Analyzing specific feature components from BDGMM activities

c) Identifying the general framework addressing BDGMM current and future needs

d) Identifying existing standards supporting BDGMM general framework

6.1 ANALYZING GENERAL DATA TREND AND THEIR UNIQUE CHARACTERISTICS

Section 2 clearly shows tremendous data growth from social media communications to massive smartphone deployment, and from IoT smart devices to smart cities. The collective sum of world data will grow from 33 zettabytes (ZB, 10^{21}) in 2018 to 175 ZB by 2025 [53]. Effective governance, management, and analysis of such rich information resources would reduce organizations' burdens and maximize customers' needs.

The dependence of big data analytics and AI machine learning and deep learning on massive quality data for training purposes is matched by Big Data, which has the means to provide quality data especially when dealing with Big Data characteristics in Volume, Velocity, and Varieties of data from multiple sources to create an integrated data source—an all-important step for legacy analytics and AI consumption. Understanding the future data landscape is an essential requirement for the development of potential BDGMM reference architecture, which would better serve and support future scalable big data analytics and AI challenges.

6.2 ANALYZING SPECIFIC FEATURE COMPONENTS FROM BDGMM ACTIVITIES

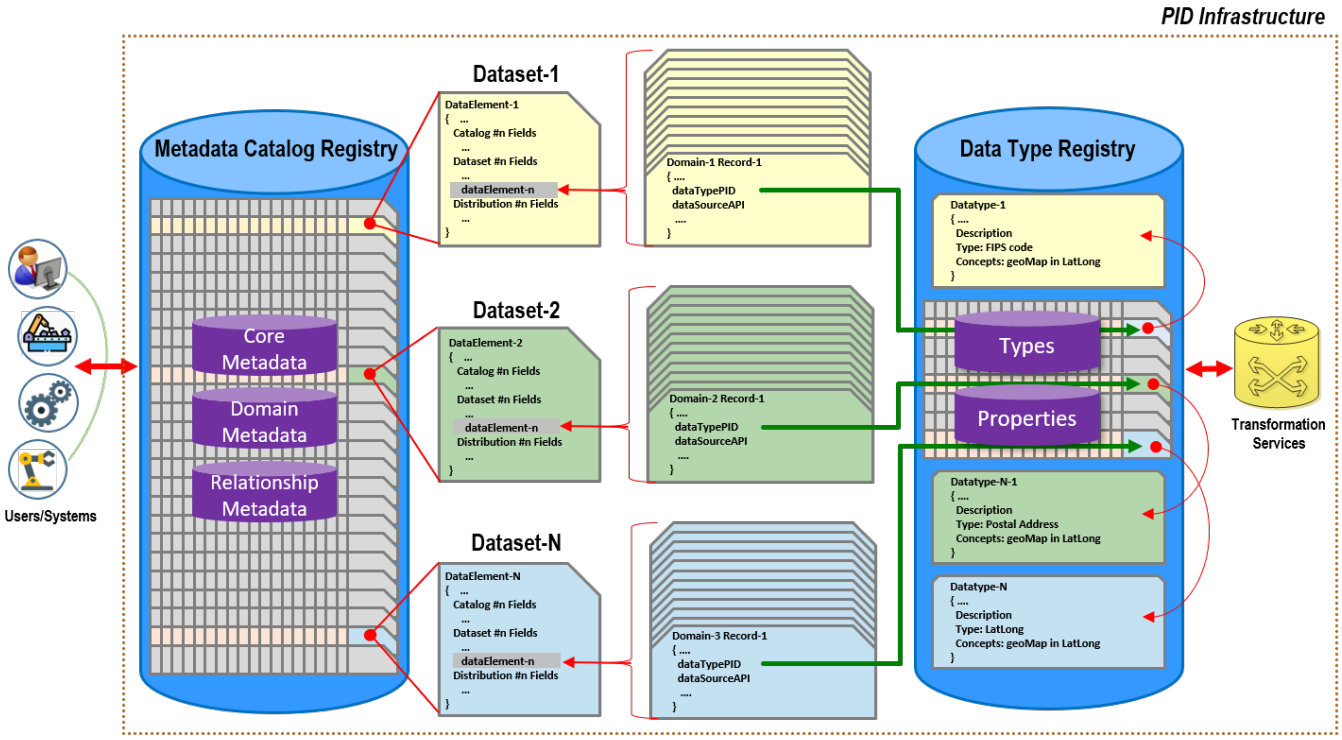
From its inception, main BDGMM activities included holding regular meetings, inviting speakers, hosting workshops and hackathons for aggregating architecture approaches and solutions for the BDGMM Roadmap White Paper development. BDGMM fully recognizes that the lessons learned from these activities do not cover the gamut of popular and available Big Data governance and metadata management. However, they serve the sample space and provide valuable insight and requirements to assess different BDGMM challenge problems.

The following architectures and approaches were gathered from BDGMM invited speakers. They provide high-level descriptions of each architecture.

6.2.1 NIST COMMON ACCESS PLATFORM

The fundamental goal of the Common Access Platform (CAP) shown in Figure 11, [an effort conceived by the National Institute of Standards and Technology (NIST)], is to enable the development of a data interoperability infrastructure that would allow two or more heterogeneous datasets, or databases, to be "mashed-up" without extensive effort on the part of the data user. Currently, seamlessly "mashing-up" or "fusing" diverse datasets is a meticulous and time-consuming endeavor. The CAP approach focuses on achieving data interoperability by utilizing persistent identifiers (PIDs) to enable (1) a standard metadata registry for data discovery using a machine-readable format, (2) a standard data type registry for data consumption using a machine-actionable format, and (3) standard end-point services to convert data values between different types. By making data interoperable, data scientists and other users can focus on the substance of a problem they are trying to solve. This will allow them quickly unlock deep insights and benefits from their data analysis.

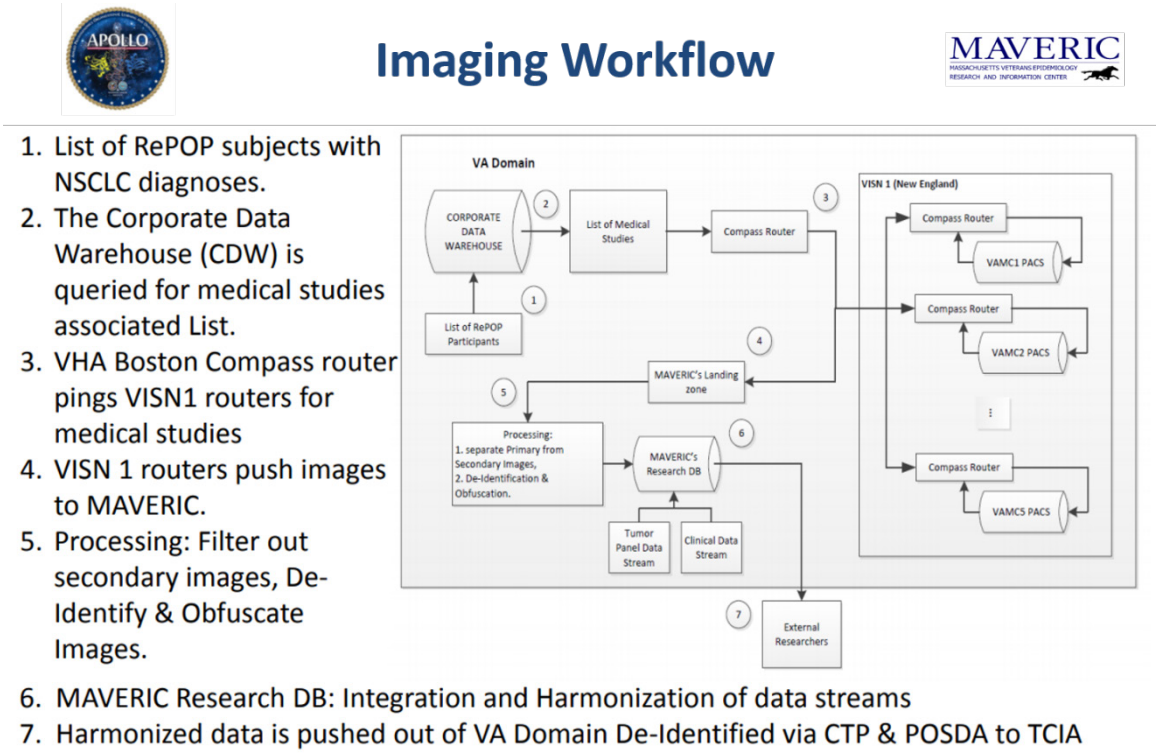
FIGURE 11: NIST COMMON ACCESS PLATFORM ARCHITECTURE



6.2.2 APOLLO VA IMAGING

Objective of APOLLO (Applied Proteogenomics OrganizationaL Learning and Outcomes) is to correlate all genomic, proteomic, and clinical data with imaging data with the focus on precision medicine or targeted medicine. Three major developments were launched. First, Precision Oncology Program (POP) (March 2015), the Department of Veterans Affairs (VA) program focused initially on lung cancer. It was designed to seamlessly merge traditional clinical activities with a systematic approach to exploiting potential breakthroughs in genomic medicine and generating credible evidence in real world settings and in real time. Second, Apollo (July 2016), inspired by Moonshot, where a coalition was formed between VA, the Department of Defense (DoD), and the National Cancer Institute (NCI) to help cancer patients by enabling their oncologist to more rapidly and accurately identify drug treatments based on the patient's unique proteomic profile. Third, Research POP (RePOP) (July 2016), the research arm of POP consists of Veterans who agreed to share their medical records (Clinical, Imaging, Genomic, etc.) within and outside the VA for the purpose of finding the cure for cancer. Veterans Heath Admin (VHA) consists of 8,000,000 Veterans, 160 VAMC, 800 clinics, 135 nursing homes. It also has backbone operational infrastructure of Veterans Information Systems and Technology Architecture (VistA). Figure 12 shows the APOLLO imaging workflow and a list of activities.

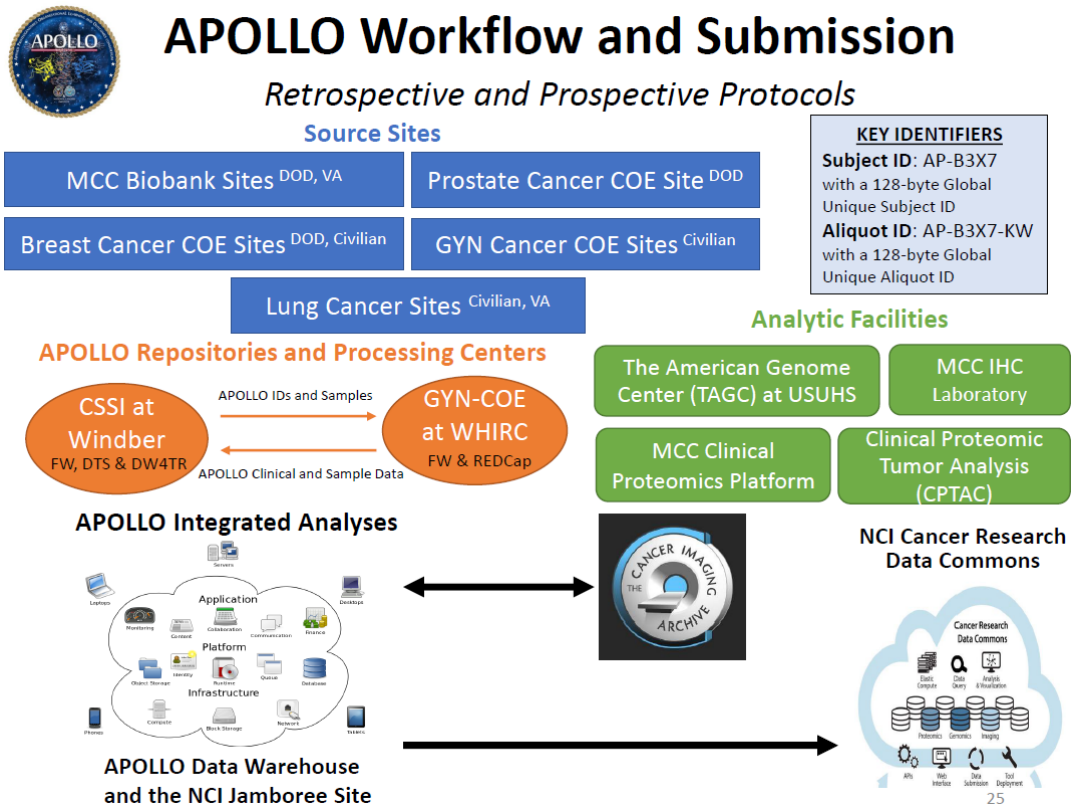
FIGURE 12: APOLLO IMAGING WORKFLOW



6.2.3 APOLLO WORKFLOW AND SUBMISSION

APOLLO Workflow and Submission provides retrospective and prospective Apollo protocols linked to Apollo IDs. Cancer patients are enrolled by VA, DOD, and civilian source sites. Samples are submitted to and distributed by Kappa accredited repositories. Free analytic processing, sequencing and proteomics are performed in state-of-the-art facilities. Data integration and analyses are accomplished using the Apollo data warehouse and the NCI provided Jamboree site. Feature annotation for clinical and tissue images are integrated into primary Apollo analyses, and then for public use in partnership with the NCI cancer imaging archive and finally Apollo data are submitted to the NCI data commons for public use. Figure 13 shows the APOLLO workflow and submission between different sites and their activities.

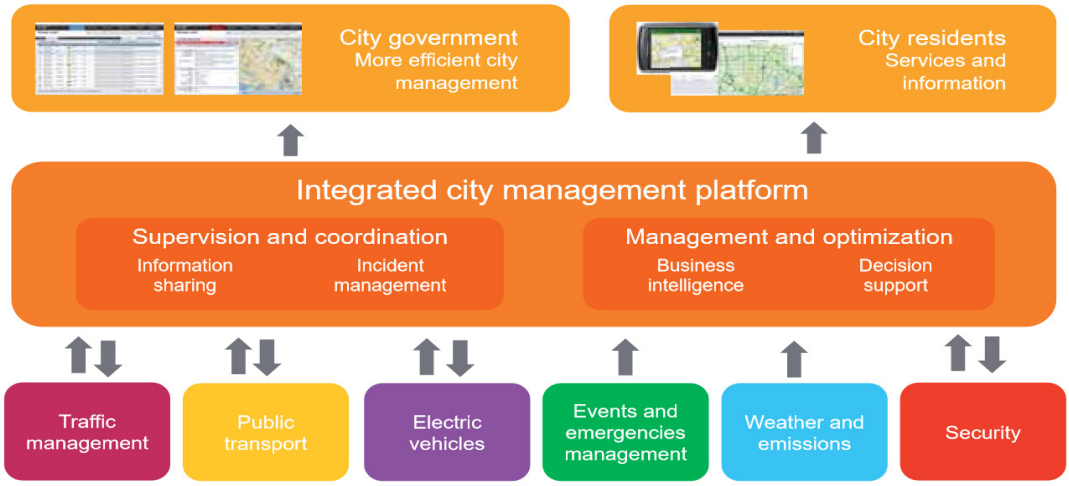
FIGURE 13: APOLLO WORKFLOW AND SUBMISSION APPROACH



6.2.4 SMART CITIES

Smart Cities provides a rich environment for data mashup when dealing with heterogeneous data from many diverse IoT sensors. The complexity of such data collection includes different realtime communication protocols, data formats, data stores, and data processing methods either at edge or at central office. The combined data enables decision making from city residents to city government. Section 3.6 case study on Smart Cities provides further descriptions and their challenges. Figure 14 shows the typical Smart Cities architecture.

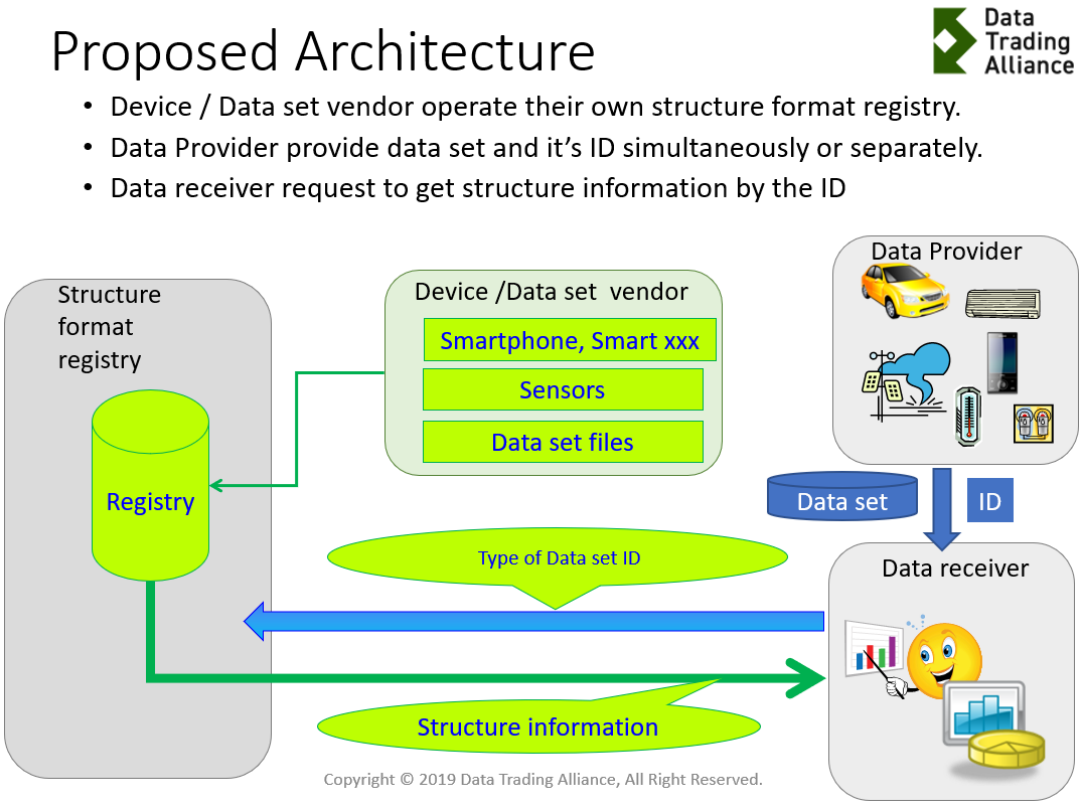
FIGURE 14: SMART CITIES ARCHITECTURE



6.2.5 DATA TRADING ALLIANCE

The purpose of the Alliance is to set up technical and systematic environments, so that users can easily judge, collect, and utilize data. There are major data trading players. First, the Data Trading Market Service Provider provides mediation between the data provider and the data user, and functions in exchange and settlement. Data Trading Market Service Provider should not perform any collection, holding, processing, and trading of data by themselves. Second, Data Provider Creator Broker to Data creates and acquires data by their own business or observation activities, organizes/processes, stores, and then deploys these data. Third, the Data User receives data from the data provider and utilizes it for services/products etc. or for their own business. Figure 15 shows the proposed architecture that would enable such data trading.

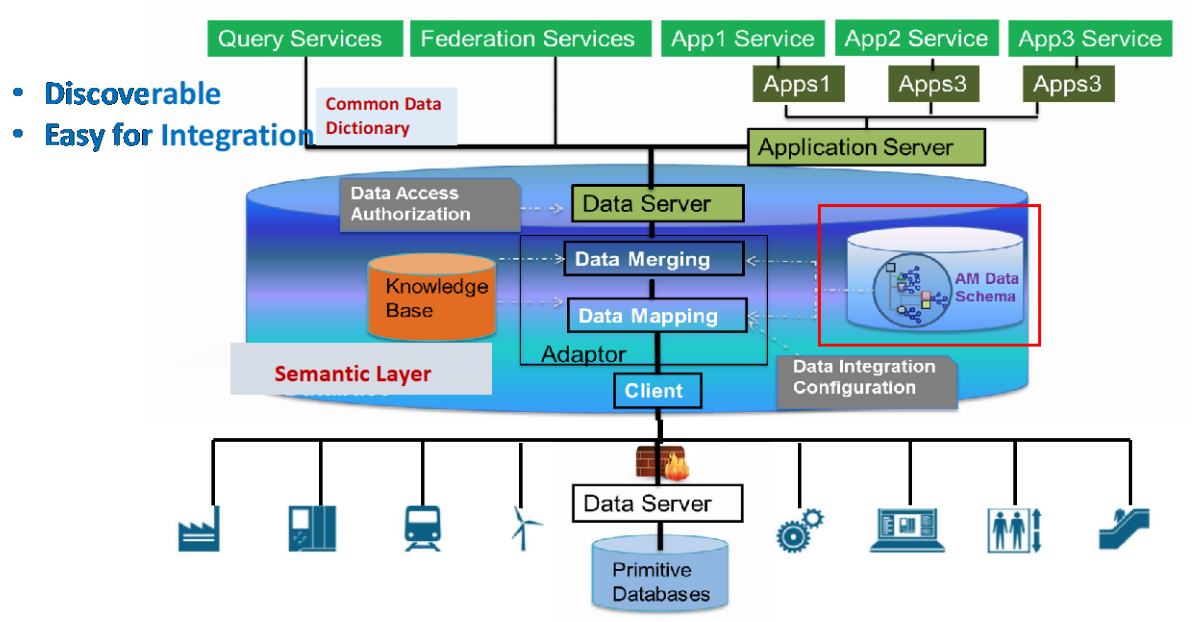
FIGURE 15: DATA TRADING ALLIANCE PROPOSED ARCHITECTURE



6.2.6 DATA INTEGRATION AND MANAGEMENT FOR ADDITIVE MANUFACTURING

Smart manufacturing plays a central role in data integration: from diverse supply chains of raw materials on product specifications to realtime quality monitoring throughout the production lifecycle. Additional data and metadata are generated from many different supporting sensors and machineries for realtime analysis and decision making in order to provide safe and healthy environments, bring precise and quality processes, and deliver reliable and superior products. Section 3.9 case study on Additive Manufacturing provides more description on processes and their challenges. Figure 16 shows the Additive Manufacturing architecture.

FIGURE 16: ADDITIVE MANUFACTURING ARCHITECTURE



6.3 IDENTIFYING GENERAL FRAMEWORK ADDRESSING BDGMM CURRENT AND FUTURE NEEDS

Data integration approaches from diverse application domains provides much insight into potential common architecture component needs across different domains. Table 2 shows the extraction and comparison between different data integration architectures from Section 6.2.

TABLE 2: COMMON FEATURE COMPONENTS FOR DATA INTEGRATION

Feature Components	CAP [54]	APOLLO Imaging [55]	APOLLO Workflow [56]	Smart Cities [57]	DTA [58]	Additive Manufact. [59]
Goals/Purposes	Share, Discover, Map, Access	Share, Discover	Share, Discover, Matching	Share, Discover, Access, Use	Share, Discover, Query, Answer	Share, Discover, Easy for integration
Metadata Registry (core, domain, relationship)	X			X	X	
Type Registry (types and properties)	X				X	
Data Catalog	X				X	
Data Dictionary	X		X	X		X
Data Model, Schema, Structure Info,	X	X	X	X	X	X
Applications/Tools/Services (query, visual, analytics, etc.)		X		X		X
Data Merging/Data Mapping			X	X		X

Table continues

Feature Components	CAP [54]	APOLLO Imaging [55]	APOLLO Workflow [56]	Smart Cities [57]	DTA [58]	Additive Manufact. [59]
Data Sources/Data Repositories	X	X	X	X	X	X
Persistent Identifier (PID), Participant ID, Global ID,	X	X	X		X	
API	X					X
Metadata Validation Tools, Verification,	X	X	X		X	X
Authentication, Data Encryption	X		X	X		X

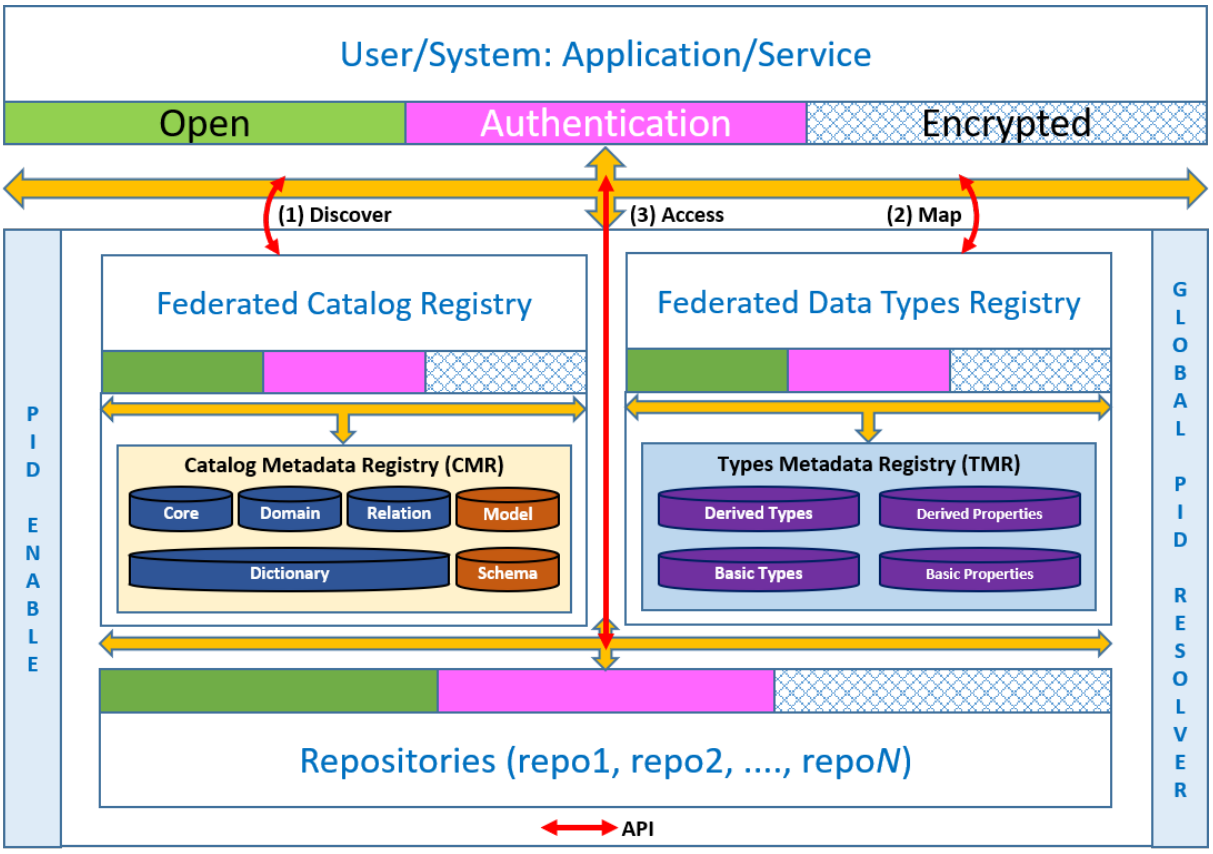
Feature components identified in Table 2 can be grouped into three main operations for the potential BDGMM Reference Architecture [60] as shown in Figure 17. The operations are as follows:

a) Discover: User/System could use either open, authentication, or encrypted interface into the Federated Catalog Registry to retrieve desirable datasets from a given set of repositories. The query operation will search datasets according to their PID-based registered data model, schema, definition, and metadata at their core, domain, and relationship with other similar datasets and data fields.

b) Map: from the Discover search result sets, retrieve other datasets that have the same or similar data models from the PID-based Federated Data Types Registry and return aggregated datasets to the User/System.

c) Access: User/System has the option to retrieve all or "top N" datasets from the repositories.

FIGURE 17: POTENTIAL BDGMM REFERENCE ARCHITECTURE



6.4 IDENTIFYING EXISTING STANDARDS SUPPORTING BDGMM GENERAL FRAMEWORK

BDGMM reference architecture functionalities can be grouped into the following four categories:

a) Big Data Governance Management for dealing with access control to query/retrieve data

b) Big Data Metadata Management for categorizing datasets with proper metadata, data model/schema with data definition

c) Big Data Integration Framework for enabling data mashup by discovering what collection of datasets would be consumed, performing mapping between datasets, and accessing all relevant datasets

d) Persistent Identifier Framework for tagging all datasets, catalog metadata, and type metadata with persistent identifiers (PIDs) so that via PIDs datasets from different repositories can be retrieved and consumed.

Table 3 shows a list of potential existing standards that can enable such operations.

TABLE 3: MAPPING EXISTING STANDARDS TO BDGMM NEEDS

Big Data Governance Management	Big Data Metadata Management	Big Data Integration Framework	Persistent Identifier Framework
ISO/IEC 38500 Governance of IT & Data Access Control • Authentication—Secure Internet Protocols ○ Web-based: Hyper Text Transfer Protocol Secure (HTTPS) ○ File-based: Secure File Transfer Protocol uses Secure Shell (SSH) encryption to transfer files over the Internet. ○ Login-based: Secures Shell (SSH) is a protocol used for securely logging in to a remote computer. ○ IP-based: Internet Protocol Security (IPsec) provides encrypted connections between computers on the Internet. ○ Socket-based: Secure Sockets Layer (SSL) encrypts data with two keys: a public key and a private key. • Encryption [61] ○ Homomorphic ○ Functional ○ Access Control Policy-based ○ Secure Multi-Party Computation ○ Blockchain	Federated Catalog Registry/ Catalog Metadata Registry (ISO 11179-2/-3/-7, ISO 19583-1/-2) • Core • Domain • Relation • Model • Schema • Dictionary Federated Types Registry/ Types Metadata Registry (types: ISO/IEC 11404—General Purpose Datatypes (GPD) • Basic Types • Derived Types • Basic Properties • Derived Properties	Discover: TBD Map: TBD Access: TBD	PID Enabler (yes and no: depends which PID you want to resolve) • Archival Resource Keys (ARKs) • Electronic Identifier Serial Publications (EISPs) • Digital Object Identifiers (DOIs), the Handle System • International eBook Identifier Numbers (IEINs) • Persistent Uniform Resource Locators (PURLs) • Uniform Resource Names (URNs) • Extensible Resource Identifiers (XRIs) Global PID Resolver: PID Resolution Services Best Practices —European [62] • Resolver services ○ Type of resolver services ○ Domain name services (DNS) ○ Local resolvers ○ Full resolvers ○ Meta-resolvers ○ Single-service resolver • Best practices for basic PID resolver functionality ○ Machine identifiable ○ PID functionality ○ PID resolution ○ Global scale • Best practices for advanced

Table continues

Big Data Governance Management	Big Data Metadata Management	Big Data Integration Framework	Persistent Identifier Framework
○ Hardware Primitives for Secure Computations • Open Access ○ Web-based: Simple Object Access Protocol (SOAP) is a messaging protocol specification for exchanging structured information in the implementation of web services in computer networks ○ Web-based: HyperText Transfer Protocol (HTTP) ○ File-based: File Transfer Protocol (FTP) ○ RESTful API: using HTTP methods (GET, POST, PUT, PATCH, DELETE)			PID resolver functionality ○ Advanced services ○ Content negotiation ○ Support for multiple locations ○ Support for direct access to content ○ Link Checking

7 RECOMMENDATION STANDARDIZATION AREAS AND ISSUES TO IEEE SA

In addition to the requirements extracted from case studies in Section 4 and gap analysis identified in Section 6, IEEE SA may want to consider standard development in Big Data Governance and Metadata Management framework that is scalable and can enable the Findability, Accessibility, Interoperability, and Reusability between corporate repositories for their heterogeneous datasets in order to empower data scientists and analytics systems to perform analytics processing without worrying about data source and structure. Specifically, the IEEE SA may want to focus on the standard development opportunities discussed in the subsections that follow.

7.1 BIG DATA GOVERNANCE MANAGEMENT

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.

7.2 BIG DATA METADATA MANAGEMENT

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.

7.3 BIG DATA INTEGRATION FRAMEWORK

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.

7.4 PERSISTENT IDENTIFIER FRAMEWORK

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.

8 REFERENCES

[1] Iskander, M. R. and Chamlou, N., "Corporate Governance: A Framework for Implementation Public," *World Bank Gr.*, pp. 636–644, 2000, doi: 10.1016/b978-0-12-373932-2.00098-3.

[2] Patrizio, Andy, "IDC: Expect 175 zettabytes of data worldwide by 2025."*Networkworld.* December 3, 2018. Available: https://www.networkworld.com/article/3325397/idc-expect-175-zettabytes-of-data-worldwide-by-2025.html.

[3] Vision Critical, "13 Stunning Stats on the Internet of Things."28 April 2017. Available: https://www.visioncritical.com/blog/internet-of-things-stats.

[4] Kosowatz, John. "Top 10 Growing Smart Cities." The American Society of Mechanical Engineers, 3 Feb 2020. Available: https://www.asme.org/topics-resources/content/top-10-growing-smart-cities.

[5] Q. Yan *et al.*, "A Review of 3D Printing Technology for Medical Applications," *Engineering*, vol. 4, no. 5, pp. 729–742, 2018, doi: 10.1016/j.eng.2018.07.021.

[6] Saltonstall, Polly, "University of Maine: 3Dirigo—World's largest 3D-printed boat." [Online]. Available: https://maineboats.com/print/issue-162/university-maine-3dirigo.

[7] Shkabura, Oleksandr, "The Main Benefits and Challenges of Industry 4.0 Adoption in Manufacturing." *Infopulse*, 12 Feb 2019. Available: https://www.infopulse.com/blog/the-main-benefits-and-challenges-of-industry-4-0-adoption-in-manufacturing/.

[8] Seal, Alan, "Five 5G Statistics You Need to Know." *vxChange*, 10 April 2020. Available: https://www.vxchnge.com/blog/5g-statistics.

[9] "IoT finds its way into hospitals." [Online]. Available: https://www.paessler.com/iot/healthcare?utm_source=google&utm_medium=cpc&utm_campaign=GBR_EN_DSA_website_Categories&utm_adgroup=prtg network monitor&utm_adnum=dsa_en_04&utm_campaignid=362067254&utm_adgroupid=35330199082&utm_targetid=dsa-162655701618&utm_customerid=136-338-8887&utm_location=9046417&gclid=CjwKCAjwkun1BRAIEiwA2mJRWSjLGJXlU4Vao4aHn7Lh3SJgOADJcpwyGg4hlQRSuoqOfb6SrItUWhoC6GoQAvD_BwE.

[10] HIT Infrastructure, "Organizations See 878% Health Data Growth Rate Since 2016." [Online]. Available: https://hitinfrastructure.com/news/organizations-see-878-health-data-growth-rate-since-2016.

[11] "Top health industry issues of 2020: Will digital start to show an ROI?" [Online]. Available: ghttps://www.pwc.com/us/en/industries/health-industries/top-health-industry-issues.html.

[12] White, SE. Predictive modeling 101. How CMS's newest fraud prevention tool works and what it means for providers. *Journal of AHIMA*. 2011;82(9):46-47.

[13] Frederic, A., "The CRWB RSbench: Towards a cooking recipe benchmark initiative," in *Proceedings—IEEE 34th International Conference on Data Engineering Workshops, ICDEW 2018*, 2018, doi: 10.1109/ICDEW.2018.00032.

[14] Frederic, A., "The Cooking Recipes without Border Data set: FAIR challenges," in International Workshop on Data Science—Present & Future of Open Data & Open Science, Mishima, Shizuoka, Japan, 2018.

[15] IEC, “IoT 2020: Smart and secure IoT platform,” *IEC*, vol. 47, no. SUPPL. 3, pp. 1–194, 2016, doi: 10.1053/j.ajkd.2006.03.010.

[16] *IEC White Paper—Internet of Things: Sensor Networks.* [Online]. Available: https://www.iec.ch/whitepaper/internetofthings/.

[17] Gibson, David V.,et al., The Technopolis Phenomenon: Smart Cities, Fast Systems, Global Networks. 1992.

[18] Caragliu, Andrea, “Smart Cities in Europe.” *Journal of Urban Technology,* Vol. 18, Issue 2. pp. 65−82. Available: https://www.tandfonline.com/doi/full/10.1080/10630732.2011.601117?scroll=top&needAccess=true.

[19] *IEC White Paper—Orchestrating infrastructure for sustainable on Smart Cities.* [Online]. Available: https://www.iec.ch/whitepaper/smartcities.

[20] Kalule, M. P., “The Legal Challenges of Social Media,” *SCRIPT-ed*, vol. 15, no. 1, pp. 141–148, 2018, doi: 10.2966/scrip.150118.141.

[21] Abir Troudi, Corinne Amel Zayani, Salma Jamoussi, *A New Social Media Mashup Approach*. Springer, 2017.

[22] A. Troudi, C. A. Zayani, S. Jamoussi, and I. A. Ben Amor, “A New Mashup Based Method for Event Detection from Social Media,” *Inf. Syst. Front.*, vol. 20, no. 5, pp. 981–992, 2018, doi: 10.1007/s10796-018-9828-9.

[23] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, “Social media analytics—Challenges in topic discovery, data collection, and data preparation,” *Int. J. Inf. Manage.*, vol. 39, no. October 2017, pp. 156–168, 2018, doi: 10.1016/j.ijinfomgt.2017.12.002.

[24] “Toreador Project.” [Online]. Available: http://www.toreador-project.eu.

[25] “Gartner Glossary on Information Governance.” [Online]. Available: https://www.gartner.com/en/information-technology/glossary/information-governance.

[26] ISO/IEC JTC 1/SC 32. Data management and interchange. [Online]. Available: https://www.iso.org/committee/45342.html.

[27] ISO/IEC JTC 1/SC 40. IT Service Management and IT Governance. [Online]. Available: https://www.iso.org/committee/5013818.html.

[28] ISO TC 20/SC 13. Space data and information transfer systems. [Online]. Available: https://www.iso.org/committee/46612.html.

[29] “Dublin Core Metadata Initiative.” [Online]. Available: http://dublincore.org/documents/dces/.

[30] “W3C Data Catalog Vocabulary.” [Online]. Available: https://www.w3.org/TR/vocab-dcat-2/.

[31] “Common European Research Information Format.” [Online]. Available: http://www.dcc.ac.uk/resources/metadata-standards/cerif-common-european-research-information-format.

[32] ISO/IEC 11179-2:2019, Information technology—Metadata registries (MDR)—Part 2: Classification. Available: https://www.iso.org/standard/74570.html.

[33] ISO/IEC 11179-3:2013, Information technology—Metadata registries (MDR)—Part 3:Registry metamodel and basic attributes. Available: https://www.iso.org/standard/50340.html.

[34] ISO/IEC 19763-1:2015, Information technology—Metamodel framework for interoperability (MFI)—Part 1: Framework. Available: https://www.iso.org/standard/64637.html.

[35] ISO/IEC 19763-3:2010, Information technology—Metamodel framework for interoperability (MFI)—Part 3: Metamodel for ontology registration. Available: https://www.iso.org/standard/52069.html.

[36] ISO/IEC 11404:2007, Information technology—General-Purpose Datatypes (GPD). Available: https://www.iso.org/standard/39479.html.

[37] ISO/IEC 38505-1:2017, Information technology—Governance of IT—Governance of data—Part 1: Application of ISO/IEC 38500 to the governance of data. Available: https://www.iso.org/standard/56639.html.

[38] ISO/IEC TR 38505-2:2018.Information technology—Governance of IT—Governance of data—Part 2: Implications of ISO/IEC 38505-1 for data management. Available: https://www.iso.org/standard/70911.html.

[39] "Consultative Committee for Space Data Systems (CCSDS)." [Online]. Available: https://public.ccsds.org/about/default.aspx.

[40] ISO 14721:2012, Space data and information transfer systems—Open archival information system (OAIS)—Reference mode.

[41] ISO 16363:2012, Space data and information transfer systems—Audit and certification of trustworthy digital repositories. (See also: Z. Xiaolin and N. Petri, "Audit and Certification of Trustworthy Digital Repositories Certification of Trustworthy Digital," *Practice*, no. September, pp. 59–68, 2011.)

[42] ISO 20652:2006, Space data and information transfer systems—Producer-archive interface—Methodology abstract standard.

[43] ISO 20104:2015, Space data and information transfer systems—Producer-Archive Interface Specification (PAIS). (See also: Consultative Committee for Space Data Systems, "Producer-archive interface specification (PAIS): Blue book," no. February, 2014.)

[44] "MOIMS Publications." [Online]. Available: https://public.ccsds.org/publications/MOIMS.aspx.

[45] K. Jeffery and R. Koskela, "M69 IEEE BDGMM Invited Speakers: Metadata for Big Data and DCAT." 2018.

[46] "W3C eGov Interest Group." [Online]. Available: http://www.w3.org/egov/.

[47] "W3C Government Linked Data." [Online]. Available: http://www.w3.org/2011/gld/.

[48] R. Kahn and R. Wilensky, "A framework for distributed digital object services," *Int. J. Digit. Libr.*, vol. 6, no. 2, pp. 115–123, 2006, doi: 10.1007/s00799-005-0128-x.

[49] Corporation for National Research Initiatives, "Overview of the digital object architecture," pp. 1–3, 2012.

[50] P. J. Denning and B. Rous, "The ACM Electronic Publishing Plan," *Commun. ACM*, vol. 38, no. 4, pp. 97–109, 1995, doi: 10.1145/205323.205348.

[51] C. for N. R. I. Robert E. Kahn, "The Architectural Evolution of the Internet."

[52] ITU-T X.1255, Framework for discovery of identity management information, 2013.

[53] "175 Zettabytes By 2025." [Online]. Available: https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/#587370055459.

[54] Chang, W., "M14 IEEE BDGMM Invited Speaker: NIST Common Access Platform (CAP) Architecture," 2017.

[55] Selva, Luis E., “Aggregating and Sharing De-Identified Clinical, Imaging, and Genomic Data from the VA to External Repositories for the APOLLO Network,” in *4th IEEE Big Data Governance and Metadata Management*, 2018.

[56] Darcy, K.M., “M104 IEEE BDGMM Inivtied Speaker: APOLLO a High Profile Use Case with Unique Challenges for the Cancer Research Data Commons,” no. September, 2018.

[57] Parikh, D. “M117 IEEE BDGMM Use Case—Smart Cities.” 2018.

[58] Mano, H., “M144 IEEE BDGMM Invited Speaker: Data Trading Alliance Challenges.” 2019.

[59] Lu, Y., “M148 IEEE BDGMM Invited Speaker on Data Integration and Management for ADDITIVE MANUFACTURING (AM),” 2019.

[60] W. Chang, “M215 BDGMM White Paper BDGMM RA Diagram.” 2020.

[61] NBDPWG—deSecurity and Privacy Subgroup, “NIST Special Publication 1500-4 NIST Big Data Interoperability Framework : Volume 4 , Security and Privacy,” vol. 4, 2019, doi: 10.6028/NIST.SP.1500-5.

[62] Wimalaratne, S. and Fenner, Martin, “D2.1 PID Resolution Services Best Practices,” no. 777523, 2018. FREYA Consortium. https://www.project-freya.eu/en/deliverables/freya_d2-1.pdf.

[63] Chang, W. “M25 IEEE BDGMM Hackathon Topic Proposal—Big Data Analytics on Healthcare Fraud Detection.” 2017.

[64] Chang, E., “M72 IEEE BDGMM Hackathon Topic Proposal—Personalized Medicine for Drug Targeting in Prostate Cancer Patients.” 2019.

[65] Arumugam, P. “M112 IEEE BDGMM Application Use Case on IoT.” 2018.

[66] Parikh, D., “M154 IEEE BDGMM Case Study Requirements.” 2019.

[67] Ceravolo, Paolo (Università degli Studi di Milano, “M134 IEEE BDGMM Application User Case—Log Analysis In Outsourcing.” 2019.

[68] Andres, F. and Dhamdhere, S.,“M135 IEEE BDGMM Use Case—Social Media,” 2019.

[69] Andres, F., “M107 IEEE BDGMM Application User Case—Intelligent Food and Cooking Recipes.” 2018.

APPENDIX A

BDGMM CAST STUDY WITH EXTRACTED REQUIREMENTS

Note—Black Text: case study challenges; Blue Text: case study specific requirements; Red Text: general requirements.

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
1	Healthcare Fraud Detection [63]	Different organizations have different governance management practices. Mechanisms and rights for accessing healthcare data from such diversified organizations can be challenging. 1. Users need to sign Data Release Agreement in order to access datasets. 2. Datasets are available via RESTful API. 3. Use "date" as the revision control method for releasing delta update from the earlier version. GR1: Access Rights—needs to support signing Data Release Agreement before access datasets. GR2: Access Method—needs to support RESTful API to retrieve datasets. GR3: Revision Control—needs to support revision control on datasets.	Healthcare data most likely coming from various organizations like healthcare providers and insurance and therefore their metadata tagging (naming, structuring, meaning, etc.) are different. Furthermore, the unit of measurement on data fields can also vary, for example, patient temperature, U.S. uses Fahrenheit where UK uses Celsius. 1. Each healthcare dataset comes with community-based data dictionary. 2. Each data dictionary is available online per each dataset. 3. Each data field provides description along with unit measurement. MR1: Metadata Schema—Needs to support any schema. MR2: Schema Data Elements—Needs to support online data	Working with multiple datasets with different file formats (csv, etc.) from different repositories (file-based, RESTful API, etc.) across different organizations and/or national boundaries; be able to align or link data fields to expand/discover new knowledge at the machine level without human in the loop. 1. Each healthcare dataset comes with data dictionary. 2. Each dataset provide a unique ID. 3. Using unique ID can correlate with another dataset. DMR1: Needs to support data dictionary for any given data field. DMR2: Needs to support unique ID for each unique data field. DMR3: Needs to support	Identify and applying appropriate tools for statistical analysis, machine learning, and visualization for predictive modelling to detect irregularities and prevent healthcare payment fraud. 1: Specific challenging analytics questions are given. 2. Some recommended tools are given. AR1: Needs to support deployable analytics tools AR2: Needs to support integration of 3rd parties of system utilities and tools.	ASCII, CSV files, file-based 1. Data are in ASCII format. 2. Data are in CSV structure. 3. Data are file-based. DCR1: File Format—needs to support ASCII file format DCR2: Data Structure—Needs to support CSV file structure. DCR3: Data Location—Needs to support local file system.

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
			elements definition MR3: Schema Data Values—Needs to support any data unit measurements.	cross reference between datasets using unique ID. DMR4: Needs to support any streaming data source. DMR5: Needs to support any heterogeneous datasets.		
2	Personalized Medicine [64]	Publicly available genomic data and their associated clinical information are generated from multiple organizations, each with their own methodology of data collection and dissemination. Mechanisms and rights for accessing genomic data from such diversified organizations can be challenging. 1. Data are publicly available. 2. Users need to navigate web portal to access appropriate datasets. GR1: Access Rights—Needs to support with or without signing Data Release Agreement before access data. GR2: Access Method—Needs to support web portal navigation to access appropriate datasets.	Genomic data from sequencing or analytical platforms have different metadata structures from different data hubs. Understanding semantic meaning and syntactic structures of these data fields can be challenging. 1. Each genome dataset comes with well-defined online data dictionary. 2. Each data field comes with a well-defined description as metadata. MR1: Metadata Schema—Needs to support any given schema MR2: Schema Data Elements—Needs to support online data element definition available for given datasets. MR3: Schema Data Values—Needs to support online data unit measurement for given data field.	While NCI GDC genomic data are well curated with different levels of description, applying appropriate linkage between data fields without human in loop would be challenging. 1. Each dataset comes with well-defined description. 2. Each dataset comes with unique ID (patient ID). 3. Each unique ID comes with well-defined description. DMR1: Needs to support data dictionary for any given genotype/ phenotype data field across all genomic cancer research data DMR2: Needs to support unique ID for each unique data field. DMR3: Needs to support cross reference between datasets using unique ID DMR4: Needs to support any streaming data	Aside from applying appropriate statistical analysis tools to find significant p-value, without specific parameters, combining datasets to do discovery will be challenging and many times a SME is necessary. 1. Specific challenging analytics questions are given. 2. Some recommended tools are given. AR1: Needs to support deployable analytics tools AR2: Needs to support integration of 3rd parties of system utilities and tools.	ASCII, TSV or JSON, file-based 1. Data are in ASCII format 2. Data are in TSV and JSON structure 3. Data are file-based. DCR1: File Format—Needs to support ASCII file format DCR2: Data Structure—Needs to support TSV and JSON file structures DCR3: Data Location—Needs to support local file system and streaming data.

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
				source. DMR5: Needs to support any heterogeneous datasets.		
3	Internet of Things (IoT) [65]	Management of data from heterogeneous data sources like various sensors, video cameras both in discrete and streaming mode, and providing a data access mechanism between diverse set of applications is challenge. Also, the data across various IoT devices and vendors needs to be normalized before they can be used for analysis. Some of the data access request from Industrial application will have to be met in realtime. 1. Be able to publish IoT sensor data and enable stakeholders to consume manage data securely from IoT ecosystem. 2. Be able to handle heterogeneous sensors (discrete sensor data and video camera streams) 3. Be able to work with static sensor sources to mobile video streams. 4. Be able to provide standardized APIs/protocols for external access models. 5. Be able to normalize the structured and unstructured sensor data compatible for	Towards providing effective data exchange between entities, IoT data needs to be catalogued having all the relevant information like owner of the data, its components, formats, authorized users, data source is stationary or mobile, anonymization requirement, etc. 1. IoT dataset can come from diverse field sensors, actuators, cameras. 2. There are also operational data that is specific to a particular IoT application like water supply, surveillance, street lights, traffic, etc. MR1: Metadata Schema—Information model, schema, ontologies. MR2: Derived data—Labeling of raw data for advanced analysis. MR3: Context data—Linking data with different contexts can enable discovery of new patterns and relationship.	Standardized interfaces for data sharing, discovering and interoperability challenges for information mashups. 1. IoT dataset comprise of static information like IoT device ID, location. 2. Meta information like message type, resources. 3. Actual message—sensed information. DMR1: data catalog, schema that is specific to the sensor device and resources being managed. DMR2: Enable search feature of the data catalog from the web. DMR3: Able to support mashup discrete and streaming data from heterogeneous sensor resources. DMR4: Vendor independent interfaces to access IoT devices.	Placement of the analytics functionality in the IoT network—in-device, edge or in the cloud—depending on both the historical and realtime data. 1. IoT analytics should support combining heterogeneous datasets. 2. Processing both at the edge and the cloud. AR1: Needs to support right context and semantics depending on the IoT application AR2: Needs to support development of smart applications leveraging heterogeneous dataset (static context, live sensor, video feeds etc.)	Audio, video, text, numeric, both discrete and streaming. 1. Data can be in diverse forms—numerical, video, audio 2. data catalog describes the structure/format 3. Majority of data are collected realtime either discrete or streaming. DCR1: Information model needs to support diverse data format. DCR2: Extensible data schema enabling easy development of enhanced services. DCR3: Data storage is in the cloud and it can be partly local (edge of the IoT network).

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
		application consumption. GR1 Access Method: Needs to support standard interfaces based on open networking protocols to access IoT sensors data.				
4	Smart Cities [66]	1. Management of data from heterogeneous data sources like various sensors, video cameras, social media and local/national agencies DB in both in discrete and streaming mode as well as historical data aggregations is challenging. 2. Since the various applications have different levels of rights/authorization to various type of data sources—managing the entitlement as well as data masking or aggregation required for different application is another challenge. 3. The data collected in such framework can potentially have characteristics that are covered by PII, HIPPA, or GDPR and other similar regulations. Even when not covered by such regulations the data can be very potent in identifying social, financial, political, health, etc. characteristics of individuals and if misused can cause great harm. 1. Data need to be masked for	The data sources for a Smarter Cities framework are varied and include multiple contributors and users, primarily government, or quasi-government agencies. The naming and definition of data entities are different for the same concept. Moreover, data representation of the same sounding entity might be different depending on purpose and technologies involved (e.g., vehicle identifier—is it license plate or VIN?). Also, the methodology to measure or capture the quantity might be different leading to different accuracy and tolerance for the measurement (e.g., laser vs. ultrasound to capture speed) and the unit of measure varies across boundaries (e.g., lb vs. kg of CO_2 emission equivalent) 1. All sources of data are not in uniform modality or source system, which is also	Working with multiple formats of input data, working with multiple representation of the same data, different technologies and communication protocols involved (REST APIs vs. CSV and mobile protocols vs. IoT devices communicating over proprietary protocols), inconsistent formatting and nuances of human generated data (e.g., voice). 1. Each dataset comes with unique sensor ID. 2. Each data comes with unique timestamp. 3. Each dataset comes with or can be correlated with centralized repository of unique sensor IDs and their characteristics. 4. PII and other regulation related elements in the data might be masked. DMR1: Needs to support data dictionary for any	Other than algorithmic complexity and issues related to metadata and normalization, following are key analytics challenges: 1. Making sure that judicious use of algorithms produces results that are unbiased and do not violate legal requirements of the various overlapping jurisdictions 2. Realtime decision making is separated and narrower in scope compared to long term analysis and its impact on policy making 3. The data samples are representative to generalize the use of results across the stakeholder base. 4. The objective functions represent the interest of all the stakeholders. 1. Some analytics are run at rest (after aggregating the data from sensor	Multiple formats—Structured (CSV, Relational Databases), Semi-Structured (JSON, XML) and Unstructured (Video, Voice, Image, TXT) 1. Need to support ASCII and File-based datasets. 2. Need to support JSON and XML schema-based datasets. 3. Need to support binary (or other encoding) dataset (either as part of objects in JSON/XML schema or as stand-alone datasets) for non-text data (e.g., image). DCR1: File Format—Needs to support ASCII file format DCR2: Data Structure—Needs to support TSV and JSON file structures

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
		most purposes not to reveal PII. 2. Data element level authorization and access control/rights needs to be established to understand who can access what data elements and under what condition. 3. Dataset access needs to be traceable and revocable to be compliant with various regulations (e.g., GDPR). GR1: Access Rights—Needs to support signing Data Release Agreement before access datasets. GR2: Access Method—Needs to support RESTful API to retrieve datasets. GR3: Revision Control—Needs to support revision control on datasets.	pervasive at metadata level. Ability to manage heterogeneous metadata. 2. Each sensor supplies its own well-defined data field description, content and unit of measurement 3. Central repository of the metadata for each sensor is available in shared environment (e.g., web-based). MR1: Metadata Schema—Needs to support any given schema MR2: Schema Data Elements—Needs to support online data element definition available for given datasets. MR3: Schema Data Values—Needs to support online data unit measurement for given data field.	given genotype/ phenotype data field across all genomic cancer research data DMR2: Needs to support unique ID for each unique data field. DMR3: Needs to support cross reference between datasets using unique ID.	network over time) and some are run in-stream, real time 2. Some analytics processing needs to happen at the edge (next to the sensor) without the data traversing the network 3. Environmental effects need to be accounted/normalized (e.g., ambient temperature values while analyzing the data from temp. sensors) 4. Analytical models (or pre-processing data normalization models) need to support multi-format, multi-modal data (e.g., temp reading and images and acoustic data) 5. Algorithms and analytics models need to be scrutinized by independent parties to make sure they are free from bias and are not harmful to vulnerable population. AR1: Needs to support deployable analytics tools. AR2: Needs to support integration of 3rd parties of system utilities and tools.	DCR3: Data Location—Needs to support local file system and streaming data.

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
5	IoT Sensor Network [66]	1. Ownership of data—the various sensors participating in network might be owned by different entities and so the question of who owns the data aggregated over the network 2. Ownership of analytics—Moreover the result of the analytics on this aggregated data is shared among the network participants is going to be complicated 3. Adverse impact with aggregation—It is easy to control and confirm that any individual sensor is not collecting any information that can affect the individual adversely. But, if the analytics on the shared data created with Sensor Network results into PII or other results that can affect individuals adversely, it would be very hard to control the outcome. 1. Datasets aggregation needs to be studies to understand what aggregation/combination of data can reveal PII or run afoul of regulation—and such aggregations need to be prohibited. 2. Data element level authorization and access control/rights needs to be established to understand	The data sources for an IoT Sensor Network are varied and include multiple contributors and users, primarily sensors measuring physical quantities (e.g., temp., flow, vibrations, speed, etc.) as well as locations, images and video and audio signatures. The naming and definition of data entities are different for the same concept. Moreover, data representation of the same sounding entity might be different depending on purpose and technologies involved (e.g., vehicle identifier—is it license plate or VIN?). Also, the methodology to measure or capture the quantity might be different leading to different accuracy and tolerance for the measurement (e.g., laser vs. ultrasound to capture speed) and the unit of measure varies across boundaries (e.g., lb vs. kg of CO_2 emission equivalent). 1. All sources of data are not in uniform modality or source system, which is also pervasive at metadata level. Ability to manage heterogeneous metadata.	Data Mashup challenges for IoT network are very similar to the challenges of IoT sensor that are stand-alone. Working with multiple formats of input data, working with multiple representation of the same data, inconsistent formatting and nuances of human generated data (e.g., voice). On the top of these usual and similar challenges, the IoT sensor network poses two additional complications – 1) Due to the massive amount of data, a lot of processing and mashing has to happen at the edge (as opposed to central data store) and only a selected amount of data should be brought back in to the central data store 2) Beyond the data formats and representation, different technologies and communication protocols involved in sensor to sensor communication alter the raw data transmission rates and formats (REST APIs, mobile protocols and IoT devices communicating over	Other than algorithmic complexity and issues related to metadata and normalization, following are key analytics challenges: 1. The amount of data involved is massive and is growing at a fast rate (this is projected to be in ZB range soon). Without automation, it will not be humanly possible to analyze and create analytical models with this amount of data. 2. If the network of sensors and machines involved in the process start making autonomous decisions based on analytics, the liability and legality of such a situation and the role of the machine "owner/operator" vs. the "autonomous decision power of the machine" could pose thorny issues. 1. Some analytics are run at rest (after aggregating the data from sensor network over time) and some are run in-stream, real time 2. Some analytics processing needs to happen at the edge (next	Multiple formats—Structured (CSV), Semi-Structured (JSON, XML) and Unstructured (Video, Voice, Image, TXT) 1. Need to support ASCII and File-based datasets. 2. Need to support JSON and XML schema-based datasets. 3. Need to support binary (or other encoding) dataset (either as part of objects in JSON/XML schema or as stand-alone datasets) for non-text data (e.g., image). DCR1: File Format—Needs to support ASCII file format DCR2: Data Structure—Needs to support TSV and JSON file structures DCR3: Data Location—Needs to support local file system and streaming data.

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
		who can access what data elements and under what condition, especially since each dataset might be coming from different sets of sensors owned by different entities. 3. Dataset access needs to be traceable and revocable to be compliant with various regulations (e.g., GDPR). GR1: Access Rights—Needs to support signing Data Release Agreement before access datasets. GR2: Access Method—Needs to support RESTful API to retrieve datasets. GR3: Revision Control—Needs to support revision control on datasets.	2. Each sensor supplies its own well-defined data field description, content and unit of measurement 3. Central repository of the metadata for each sensor is available in shared environment (e.g., web-based). MR1: Metadata Schema—needs to support any given schema MR2: Schema Data Elements—needs to support online data element definition available for given datasets. MR3: Schema Data Values—needs to support online data unit measurement for given data field.	proprietary protocols), posing additional issues with data mashup. 1. Each dataset comes with unique sensor ID. 2. Each data comes with unique timestamp. 3. Each dataset comes with or can be correlated with centralized repository of unique sensor IDs and their characteristics. 4. PII and other regulation related elements in the data might be masked. DMR1: Needs to support data dictionary for any given genotype/ phenotype data field across all genomic cancer research data DMR2: Needs to support unique ID for each unique data field. DMR3: Needs to support cross reference between datasets using unique ID.	to the sensor) without the data traversing the network 3. Analytical models (or pre-processing data normalization models) need to support multi-format, multi-modal data (e.g., temp reading and images and acoustic data). AR1: Needs to support deployable analytics tools. AR2: Needs to support integration of 3rd parties of system utilities and tools.	
6	Log Analysis Outsourcing [67]	The challenges related to governance are mainly dependent to the ownership of data, as customers may be interested in outsourcing the audit of application logs but, at the same time, the risk of data leakage can limit their availability. Making	To address the proposed case study by a modular solution metadata are of paramount importance. It is actually necessary to disambiguate the different columns of a record, their confidence level, their format and eventually their	Data mashup is relevant to this case study as multiple log files may be integrated during the auditing. A domain specific language for addressing data integration by a fast configuration is for	The proposed case study underlines that metadata must support the reuse of data analytics and service. Consider for example that a new request could emerge for the organization concerned with auditing.	Multiple formats - Structured (CSV), Semi-Structured (JSON, XML) and Unstructured Grap-based additive data (JSON-LD, RDF) 1. Need to support

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
		application logs auditable even when anonymized or obfuscated is then a crucial challenge in order to support the proposed scenario. Dividing the auditing task in a training and an alerting (basically, a classification) stage, makes possible to deploy them on infrastructures subject to different controls, limiting the risk of leakage. However, the service provider may also be concerned by the ownership of analytics or even by the ownership of specific boosting strategies incarnated in a pipeline of services and analytics. The deployment procedure is then expected to be modular, to adapt to the specific needs of a customer, but not necessarily transparent to all the actors involved in. 1. Be able to represent a pipeline of computational tasks. 2. Be able to express rights and access control rules on specific components of the pipeline. 3. Be able to represent access control rules on the outputs of a specific component. GR1: Access Rights—needs to	linkage conditions. 1. Metadata can express data type at each column. 2. Metadata can express the confidence level of data at each column. 3. Metadata can express linkage conditions of data at column level. MR1: Schema Data Elements—needs to support online data elements definition. MR2: Context data—linking sensor data with different contexts can enable discovery of new patterns and relationship. MR3: Metadata schema must be additive (graph-based data model).	example an important challenge we would like to highlight. 1. Each dataset comes with unique sensor ID 2. Each data comes with unique timestamp 3. Data integration layer can support the data mashup by simple matching rules DMR1: needs to support unique ID for each unique data field. DMR2: needs to support cross reference between datasets using unique ID DMR3 a simple data integration language is required to aggregate data.	For example, Bob is now interested in dormant accounts, i.e., accounts which are not being used for a long duration and which suddenly become active. Detecting them may require deriving a new feature from the dataset, for example an "elapsed time" column, which will contain, for each entry, how many seconds passed since the same user performed an action. This implies a new data preparation must be included in the pipeline; however, the rest of the pipeline is not modified. A good set of metadata can allow deploying the new pipeline simply by reusing the previous set of specifications with the required addition or removal of services. This required the system to support a representation of Big Data pipelines to support their reproducibility and verifiability. 1. Some analytics are run at rest (after aggregating the data from sensor network over time) and some are run in-stream,	ASCII and File-based datasets 2. Need to support JSON and XML schema-based datasets 3. Need to support binary (or other encoding) dataset (either as part of objects in JSON/XML schema or as stand-alone datasets) for non-text data (e.g., image) 4. Needs to support the graph-based data model (JSON-LD, RDF). DCR1: File Format—Needs to support ASCII file format DCR2: Data Structure—Needs to support TSV and JSON file structures DCR3: Data Location—Needs to support local file system and streaming data. DCR4: Needs to support the graph-based data model (JSON-LD, RDF).

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
		support signing Data Release Agreement before access datasets. GR2: Revision Control: needs to support revision control on datasets.			realtime 2. Some analytics processing needs to happen at the edge (next to the sensor) without the data traversing the network 3. Analytical models (or pre-processing data normalization models) need to support multi-format, multi-modal data (e.g., temp reading and images and acoustic data) 4. Analytics are partially configured by reading metadata 5. A consistency check between metadata and analytics can be configured. AR1: Needs to support deployable analytics tools AR2: Needs to support right context and semantics AR3: Consistency between analytics and metadata can be checked.	
7	Social Media [68]	The primary challenge is the scope of the governance model. Target social channels along with policies and guidelines. For example, one organization could have either separate or common governance model for internal	The social media data sources are varied and include multiple contributors and users. It influences on various business aspects, including human resources. We can classify big data as personal	Social media gives opportunities to discover, report, and share different types of events. Social media can be considered as a dynamic source of information that enables individuals	The social media big data analytics is a highly complex process with different aspects. It involves four distinct steps: (1) data discovery, (2) collection, (3) preparation, and (4)	Social media is a source of realtime data that individuals create and voluntarily share on major social media generators such as Facebook, Twitter,

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
		and external communities. Three-dimension-based model 1) Branding dimension: Branding guidelines for external channels (e.g., branded templates for social channels like Twitter, social sharing and follow icons, as well as the use of company logo and related elements on external-facing channels). 2) Training & Education dimension: Plenty of educational resources for employees training (e.g. responding to customer feedback, both positive and negative). 3) Approval Processes & Continuity Planning: approval processes in place for employees to engage in social media. Challenges: Can everyone participate (highly recommended) or only members of certain external-facing groups can engage via the company's social channels? What is the process for getting approval for an official account?	and professional social media data.it will impact metadata and its management. In addition, there are several legal challenges impacting the metadata management: (1) the social media and law, (2) public order in a virtual space, and (3) private law responses to social media.	to stay informed of all real-world events. In this use case, the focus is the detection of events from social media: (1) the noise of data within the social media and the semantic detection of the main dimensions such as the topic, the time and the location. (2) mapping social media big data like lack of demographic profile, data integration problems, issues with user privacy and locational privacy, need of multidisciplinary collaborations, contextual analysis, filtering noises and difficulty of falsification of hypotheses and theories along with data inconsistency issues in case of updates and major revisions on service.	analysis. While there is a great deal of literature on the challenges and difficulties involving specific data analysis methods, there hardly exists research on the stages of data discovery, collection, and preparation. The volume scale of social media big data is the first challenge to impact the stages of data discovery, collection, and preparation We can point out three additional challenges—the velocity scale of social media big data, to extract valuable insights from the social media response in a realtime manner, the variety scale of social media big data analytics on spatial-temporal data and the veracity scale of unstructured and uncertain nation of social media big data in order to evaluate the quality data and manage the value of that data.	Google, Yahoo, and Instagram. Linked data in social media present both challenges and opportunities for feature selection. It includes user-user and user-post relationships manifested in linked data. Four types of relationships characterize linked data from Social Media [5]. We will detail these relationships in a next version. Coming actions will be to propose a data model for social media to scope the different challenges, including social media data quality.
8	Intelligent Food and Cooking Recipes	Open data cooking recipe dataset and associated nutritional and health-related information are generated	Cooking recipe collections as raw data or as linked open data have different metadata structures from	The CRWB dataset will facilitate the Big Data Mashup and H2M including IoT mashup	Aside from applying appropriate classifier tools to find similarity between recipes and	CSV or JSON format

Table continues

	Use Case	Governance	Metadata	Data Mashup	Analytics	Data Characteristics
	[69]	from multiple organizations, each with their own methodology of data collection and dissemination. Mechanisms and rights for accessing enhanced cooking recipe data from such diversified organizations can be challenging.	different data hubs. Promoting interoperable semantic meaning and syntactic structures of these data fields can be challenging.	tools, Cooker robots (Human-Machine interaction) as the integration of heterogeneous cooking recipes (including tasting expectation) and various applications from multiple sources including IoT for healthcare and other research purposes.	combination of ingredients, generating new recipes as healthy and gastronomical discovery will be challenging.	

APPENDIX B
BDGMM WORKSHOPS AND HACKATHONS

1. 1st BDGMM Workshop/Hackathon, Bostin, MA, US, Dec. 11–12, 2017, https://bigdatawg.nist.gov/bdgmm/event01.html

   8 participants (7 teams) in the hackathon, 30+ attendees for the workshop, one keynote speaker, 2 invited speakers, 3 peer-review papers accepted.

   IEEE SA and IEEE Future Direction donated $4000 for hackathon cash awards

2. 2nd BDGMM Workshop/Hackathon, Berlin, Germany, Mar. 19–20, 2018, https://bigdatawg.nist.gov/bdgmm/event02.html

   9 participants (5 teams) in the hackathon 20+ attendees for the workshop, one keynote speaker, 2 invited speakers, 4 papers submitted and 2 presented.

   FIESTA-IoT EU H2020 Project donated 3500 EURO (~$4300) for hackathon cash awards

3. 3rd BDGMM Hackathon, Tokyo, Japan, July 23–24, 2018, https://bigdatawg.nist.gov/bdgmm/event03.html

   30 attendees: 24 hackathon participants in 7 teams, 2 subject matter experts, 4 judges.

   IEEE Brain Initiative donated $900 for hackathon cash awards

4. 4th BDGMM Workshop/Hackathon, Seattle, WA, Dec. 10–11, 2018, https://bigdatawg.nist.gov/bdgmm/event04.html

   35+ participants (8 teams) in the hackathon, 30+ attendees for the workshop, one keynote speaker, 2 invited speakers, 2 peer-review papers accepted.

   IEEE SA, IEEE Brain, and IEEE Future Direction donated $6000 for hackathon cash awards

RAISING THE WORLD'S STANDARDS

IEEE

3 Park Avenue I New York, NY 10016-5997 USA standards.ieee.org

Tel.+1732-981-0060 Fax+1732-562-1571