

BIG DATA GOVERNANCE AND METADATA MANAGEMENT: STANDARDS ROADMAP

Authored by the

IEEE Big Data Governance and Metadata Management
Industry Connections Activity

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 <i>National Institute of Informatics, Japan</i>	Kathleen M. Darcy <i>Uniformed Services University of the Health Sciences, U.S.</i>	Yan Lu <i>National Institute of Standards and Technology, U.S.</i>
Paventhan Arumugam <i>National Research and Education Network of India, India</i>	Santegeeta Dhamdhare <i>Modern College of Arts, Science and Commerce, India</i>	Hiroshi Mano <i>Data Trading Alliance, Japan</i>
Claire Austin <i>Environment & Climate Change, Canada</i>	Mohsen Farid <i>University of Derby, UK</i>	Dhaivat Parikh <i>IBM, U.S.</i>
David Belanger <i>Stevens Institute of Technology, U.S.</i>	Jane Greenberg <i>Drexel University, U.S.</i>	Allison L. Powell <i>Corporation for National Research Initiatives (at time of contribution) MITRE (current), U.S.</i>
Elizabeth Chang <i>University of Maryland, U.S.</i>	Kathy Grise <i>IEEE Technical Activities</i>	Ann Racuya-Robbins <i>World Knowledge Bank, U.S.</i>
Wo Chang <i>National Institute of Standards and Technology, U.S.</i>	Keith Jeffery <i>Consultant</i>	Luis E. Selva <i>U.S. Department of Veterans Affairs</i>
Paolo Cervolo <i>Università degli Studi di Milano, Italy</i>	Rebecca Koskela <i>University of New Mexico, U.S.</i>	Cherry Tom <i>IEEE Standards Association</i>
Marl Conrad <i>National Archives and Records Administration, U.S.</i>	Larry Lannom <i>Corporation for National Research Initiatives, U.S.</i>	Denise Warzel <i>National Institutes of Health, U.S.</i>
Mahmoud Daneshmand <i>Stevens Institute of Technology, U.S.</i>		Joan Woolery <i>IEEE Standards Association</i>

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

Copyright © 2020 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved. July 2020. Printed in the United States of America.

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

IEEE is a registered trademark in the U. S. Patent & Trademark Office, owned by The Institute of Electrical and Electronics Engineers, Incorporated. All other trademarks are the property of the respective trademark owners.

IEEE prohibits discrimination, harassment, and bullying. For more information, visit <http://www.ieee.org/web/aboutus/whatis/policies/p9-26.html>.

No part of this publication may be reproduced in any form, in an electronic retrieval system, or otherwise, without the prior written permission of the publisher.

Find IEEE standards and standards-related product listings at: <http://standards.ieee.org>.

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 ICom 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

	Executive Summary.....	5
1.0	Introduction	6
2.0	Data Explosion	7
3.0	BDGMM Case Study	10
4.0	BDGMM Technical Requirements	26
5.0	Relevant Standardization Activities	28
6.0	Standard Technology Gap Analysis	35
7.0	Recommendation Standardization Areas and Issues to IEEE SA	44
8.0	References	46
	Appendix A BDGMM Cast Study with Extracted Requirements	50
	Appendix B BDGMM Workshops and Hackathons	61

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.



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

Dataset-1: Physician's Billing

	NPI	Entity	Provider Last Name	Provider First Name
2	1079576722	1	WIEBE	DAVID
3	1580667638	1	PILCHER	WILLIAM
4	1497758544	2	CUMBERLAND COUNTY HOSPITAL SYS	
5	1306849450	1	SMITSON	HAROLD
6	1215930367	1	GRESSOT	LAURENT
7	1023011178	2	COLLABRIA CARE	
8	1932102084	1	ADUSUMILLI	RAVI
9	1841293990	1	WORTSMAN	SUSAN
10	1750384806	1	BISBEE	ROBERT
11	1669475711	1	SUNG	BIN
12	1578566626	1	KUIPERS	WARF
13	1487657532	1	HUEBERT	ALLISON
14	1396748448	1	YOUNGMAYKA	CYNTHIA

Dataset-2: Physician Identifier

	npi	nppes	ncpes	description	average_Medica	average_subn	average_Medica
3	1003000126	M.D.	I Hospital observ		72.68	328	54.5026087
4	1003000126	M.D.	I Hospital observ		135.85	614	100.9588889
5	1003000126	M.D.	I Initial hospital i		101.3650848	333.2881356	79.46966102
6	1003000126	M.D.	I Initial hospital i		139.0104546	356.4924242	107.9334091
7	1003000126	M.D.	I Initial hospital i		205.1859546	631.1863636	157.9089546

Dataset-3: Physician Specialization

	Code	Grouping	Specialization	Definition	No
2	101Y00000X	Behavioral Health & Social Service Prov	A provider who is trained and educ		
3	101YA0400X	Behavioral Health (Addiction (Substan	Definition to come...		
4	101YM0800X	Behavioral Health (Mental Health	Definition to come...		
5	101YP1600X	Behavioral Health (Pastoral	Definition to come...		
6	101YP2500X	Behavioral Health (Professional	Definition to come...		
7	101YS0200X	Behavioral Health (School	Definition to come...		

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

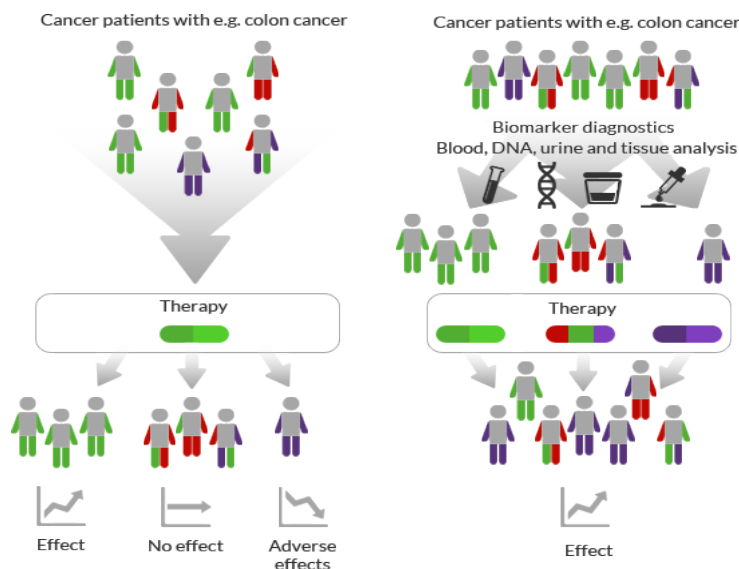
ASCII, CSV files, file-based

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.

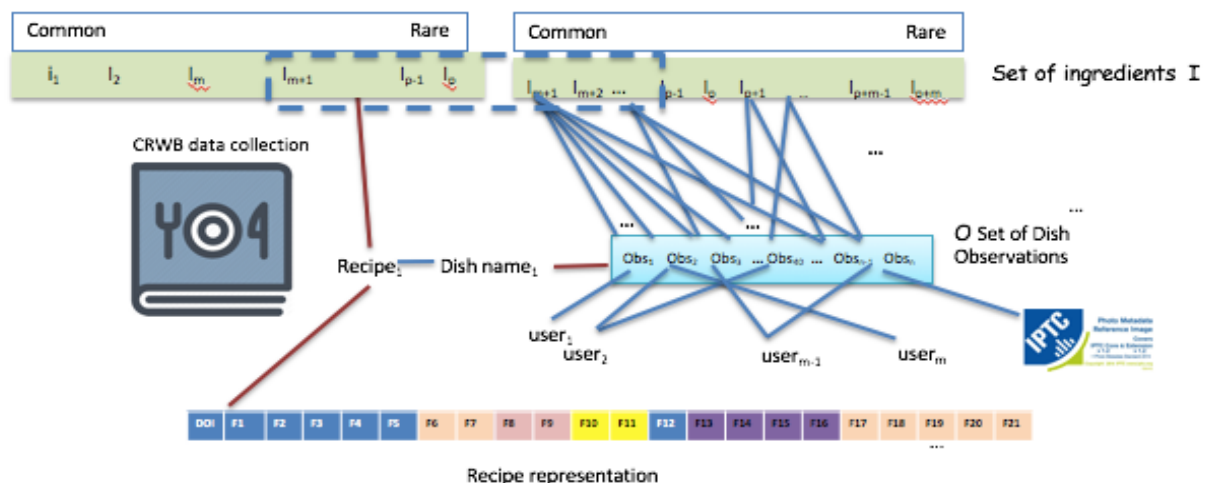
3.3 CASE STUDY #3: INTELLIGENT FOOD AND COOKING RECIPE

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

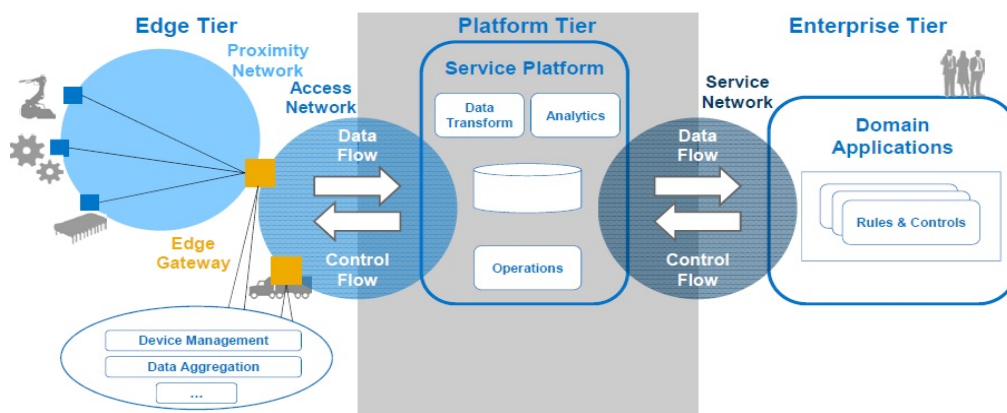
CSV or JSON .format

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.

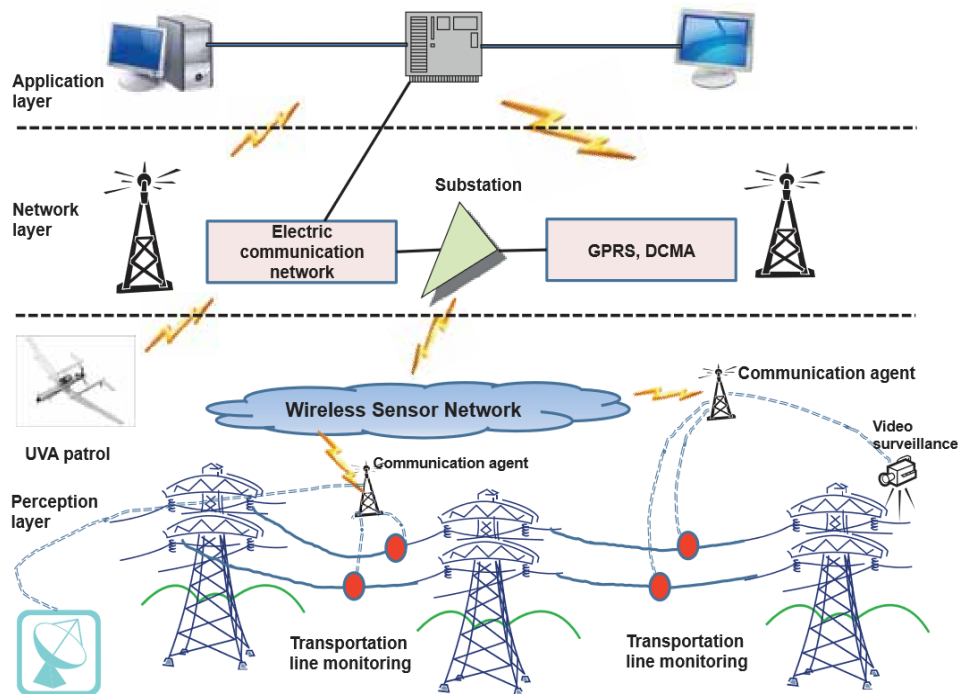
3.5 CASE STUDY #5: IOT SENSOR NETWORK

BACKGROUND

IoT Sensor Network can be defined as uniquely identifiable objects and their virtual representations in an “internet-like” structure (Ashon, 1999 - <http://www.rfidjournal.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 CO₂ 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

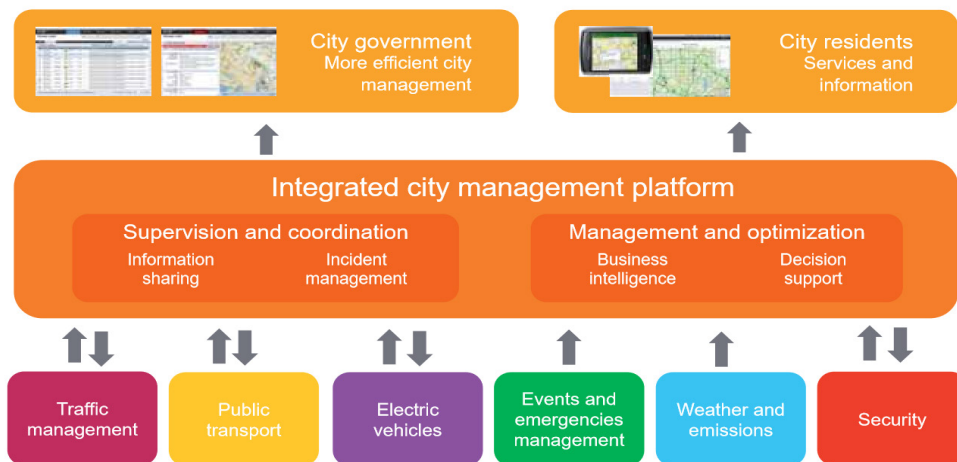
BACKGROUND

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 CO₂ 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).