Open Data Platform for DPM Standardization

Jun Kyun Choi
jkchoi59@kaist.ac.kr

Korea Advanced Institute of Science and Technology (KAIST)
Contents

1. Platform Revolution for Future Data Society

2. Data Model for Knowledge Society

3. Technical Issues for Data Standardization

4. Toward Open Data Platform
Platform Revolution for Future Data Society
Future Eco-Society

Human World

Physical World

Future Social Infrastructure
Knowledge Sharing Platform

- Sharing Information and Knowledge by using Cloud Platform
  - Public information of Government
  - Information of Science and Mathematics
  - Society and Life related Information
  - Bio and Medical Information
  - Energy, Automobile related Information
  - Nano, Semiconductor, and Component Information
  - Education, Culture and Art related Information
Smart City Eco-System

DIGITAL CITY ECOSYSTEM COMPONENTS

LOCAL CITIZENS
CIVIL SERVANTS
BUSINESSES
RESIDENTS
STUDENTS
BUSINESS TRAVELERS
TOURISTS

SMART CARD
SMART TV
WEARABLE DEVICES
SMART PHONE
TABLET
DIGITAL SIGNAGE
PC
SMART VEHICLES

EDUCATION
SAFETY SECURITY
ENVIRONMENT
ENERGY AND UTILITIES
ENTERTAINMENT AND RETAIL
HEALTH
TRANSPORTATION
GOVERNMENT

NEXT-GEN INFRASTRUCTURE
CITY CONNECTIVITY INFRASTRUCTURE
CITY DIGITAL INFRASTRUCTURE
CITY DIGITAL SECURITY INFRASTRUCTURE

DIGITAL ENABLING LEVERS
GOVERNANCE
OPEN GOVERNMENT AND ANALYTICS
INNOVATION AND ENTREPRENEURSHIP
HUMAN CAPITAL
CYBERSECURITY
POLICIES AND STANDARDS
STRATEGIC COMMUNICATION
PUBLIC-PRIVATE PARTNERSHIPS

http://www.urenio.org/2015/12/20/booz-allen-hamilton-on-smart-cities/
jkchoi59@kaist.ac.kr
Future Knowledge Convergence Infrastructure

1st Stage
ICT Convergence
- Vehicle
- Energy
- Construction
- Robot
- Health
- Machine
- Military
- Ship
- Textile
- Air

2nd Stage
Creative Convergence
- Human & Health care
- Nano devices, computers & Robotics
- Huge Complex Industrial systems
- Cyber Reality & Future Human Interface
- Security & Trust

3rd Stage
Creative Eco-Convergence
- Autono System
  - Intelligent Car, Robot
- uEco-City
  - U-Eco building & city
- Self-organized Services
- Knowledge Media
  - Lifelong education
- Governance
  - Personal & social security

Future Network and Computing Infrastructure

ICT Concepts
- (Conformity)
  - All Connected Things
- (Intelligence)
  - All Computerized Things
- (Resonance)
  - All Conscious Things

jkchoi59@kaist.ac.kr
Future Hardware and Software Ecosystems

How to handle billions of hardware components and devices?
How to arrange billions of software and business platforms?

(Key Strategies) Heterogeneous and Evergrowing Eco-environments

M2M Module = 6* x
Wearable Device = 7* x
Smartphone = 49* x
Tablet = 127* x
Laptop = 227* x

* Monthly basic mobile phone data traffic
Source: Cisco VNI Mobile, 2014

SMART SERVICES
Real-time | Big Data Analytics | Continuous Integration

Source: Ericsson

jkchoi59@kaist.ac.kr
What is the Language of Cyber World?

<table>
<thead>
<tr>
<th>Cyber Language</th>
<th>Physical Language</th>
<th>Natural Language</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Language for Semantic Knowledge</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Language for Application Platform</strong></td>
<td>Human Language</td>
<td>Human Language</td>
</tr>
<tr>
<td><strong>Language for System and Application</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Operating System (OS)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Machine Language (e.g., device driver)</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Technical Challenges for Future Social Infrastructure

- **Internet of Things**
  (The ongoing convergence of evolution of devices)

- **Cloud Computing**
  (Deployment of large shared infrastructure)

- **Big Data**
  (Accumulation of data from sensors and social networks)
Cost Declining of Cloud Computing

Compute Costs Declining = 33% Annually, 1990-2013

Decreasing cost / performance curve enables computational power @ core of digital infrastructure

Global Compute Cost Trends

$ per 1 MM transistors

$0.01 $0.10 $1.00 $10.00 $100.00 $1,000.00

$527


$0.05

Note: Y-axis on graph is logarithmic scale.
12 TeraFlops GPU Processing Power

Table overlay: Theoretical SP GFLOPS/W of NVIDIA Tesla cards. Grey *italic* text represents my guesses.
Cost of IoT Modules

Average Cost of IoT Sensors

- 2004 average cost: $1.30
- 2020 average cost forecast: $0.38

Data: Goldman Sachs, BI Intelligence Estimates
IoT Sensor Modules

- Multi-sensor module
- WiFi module
- Arduino compatible LoRa Module
- Bluetooth modules
- Raspberry PI module
- Intel Edison module
- Intel Curie module
Wireless IoT Modules

M32W + ZigBee + Bluetooth = Wireless IoT Gateway

433/868/915MHz RF

Radio Module
Performance Trends of Cloud Computing

**STORAGE**
1000X IOPS and 50X+ lower latency, Transition from HDD to Flash

**COMPUTE**
200X per Moore’s Law, 2X every 1.5 - 2 years

**NETWORKING**
10X (1GbE to 10GbE) with ongoing transition to 25/50GbE

*Mainstream server deployed in a data-center*
Cloud Computing for IoT Applications

- Surveillance
- Critical infrastructure monitoring
- Environment Monitoring
- Health Monitoring
- Smart Transportation

- Visualization
- SaaS
- PaaS
- IaaS
- Analytics
- Computation
- Storage

- Security, Re-configurability, Quality of Service, Communication protocols, Location Awareness, Compressive Sensing

- Wireless Sensor Networks
- “Network of Things”
Concept of Cyber Physical System (CPS)
CPS: Intelligent System for Future Eco-Society

“Intelligent Systems will exist in environments they sense and perceive, and from which they learn and continually act to achieve their objectives.”

Intelligent and Self-Learning Computing System

1. **sense** the real-world environments,
2. **perceive** the world using world models,
3. **adapt** to different environments and changes,
4. **learn** and build knowledge, and
5. **act** to control their environments.
Data Model for Knowledge Society
Data Model of IoT Data Analytics

Data Processing for Knowledge Engineering

A

Raw data → Pre-processing → Clean data

ACGTC
GCGTA
GTCCG
TTAGT
CGTAG
AGAAA

Features

Model

Results

B

Supervised

Unsupervised

\( x \rightarrow y \)

• Linear regression
• Logistic regression
• Random Forest
• SVM
• ...

\( x \)

• PCA
• Factor analysis
• Clustering
• Outlier detection
• ...

C

Raw data → Feature extraction → Discriminative features

D

Label

Layer 2

TSS
Intron
Exon

Layer 1

Raw data

jkchoi59@kaist.ac.kr
Data Model for IoT Applications

- **IoT data model as open standards**
  - RAML, JSON-LD or CBOR for IoT sensors (e.g., mode, property, action, value, etc.)

- **How to describe semantics for IoT applications**
  - **Who** (Id): device name, product code, and identification no.
  - **Where** (Location): city, car, smart grid, home, farm, etc.
  - **What** (Purpose): safety, surveillance, monitoring, protection
  - **When & How** (Events): value, condition, and to activate or trigger

  → context-aware and environmental data model

New form of Linked Data?
- Lookup table or directory for device id

jkchoi59@kaist.ac.kr
Data → Information → Knowledge → Wisdom Framework

- Synthesis of multiple sources of information over time,
- Organization and processing to convey understanding, experience and accumulated learning,
- Mix of contextual information, values, experience and rules.

Context

Data

Feedback

Understanding relations

Understanding patterns

Knowledge

Wisdom

-Who
-What
-When
-Where

-How

-Why

Understanding principles
Key Features of Knowledge Networking

Future Creative Industry from Innovation of Traditional Industries

Networking of Distributed and Scattered Knowledge

Innovation & Creation

Information Gathering & Analyzing

Gathering and Analysis of Heterogeneous Information
Knowledge Accumulation Process

Hypotheses

induction

observation
deduction

Test of Predictions

Predictions

jkchoi59@kaist.ac.kr
Value of Knowledge

- **Some statements on knowledge**
  - “All knowledge is of itself of some value” in Samuel Johnson in 1775
  - “Thought, not money is the real business capital” Firestone in relation to business
  - Knowledge is power in its own right

- **New business values from industrial products and goods**
  - Shifts from tangible knowledge to **intangible knowledge** will be revolutionized

- **In the information and communication world**
  - **Data transfer** can be viewed as a process whereby knowledge can be transferred
    - Used to all forms of recording and transferring knowledge
    - Millions of books in the libraries can be transferred in electrical manner
  - **Mobile wireless** offers individuals a means to connect anywhere and anytime
  - **New technologies** with reflection and critical thinking create intelligence and valuable meaning for knowledge creations
Cross-Disciplinary and Interdisciplinary of Knowledge

Key characteristics of knowledge

- Synthesize broad perspectives, skill, experiences, and know-hows by crossing boundaries
  - Interdisciplinary is applied not only with education and training, but also with research and development
- To solve the global common tasks on climate change and health, interdisciplinary studies are important to connect peoples with different knowledge disciplines
- Synergy effects among peoples with broad dimensions and different experiences are promising
  - IoT/M2M (Internet of Thing/Machine to Machine) technologies are needed to collaborate with cultural and social sciences as well as economics
- Knowledge eco-environments may consist of people, organization, and processes that work together
Creating New Data World

- **Need new data types for knowledge society**
  - (**Mapping**) Data types connecting physical world to cyber world
    - Identification/classification, Location, Status, Role/Function, etc.
    - (Example) GPS, address, serial no. etc.
  - (**Vitalization**) Data types for human understanding, meaning, interpretation, and translation, perception, recognition based on accumulated knowledge, reasoning, learning, action, behaviors, and experience
    - (Example) Pythagorean Theorem, know-how of medical treatments, etc.
  - (**Visualization**) Data types for communication, sharing, seeing, rendering, hearing, and writing, etc.
    - (Example) icon, logo, graphic image, character,
Approaches toward New Data Society

- Key Issues to realize New Data Society
  - How to utilize interactive power of computing and communications technology?

- General Approaches toward New Data Society
  - Behavior Cycle for future human life and business culture
  - Open Environments for Network, Software, and Device
  - New data format to create, deliver, and consume
  - Eco-systems between physical society and cyber society
  - Intelligent emerging devices including smart phone, smart TV, smart car, smart building, and smart things, etc.
How to accumulate Knowledge

- **How to connect the forms of knowledge**
  - A lot of forms or types to represent knowledge or intelligences
  - Tacit knowledge (or implicit knowledge) is difficult to transfer

- **New forms of development, acquisition, and spread of knowledge**
  - New forms of capturing, developing, sharing, and effectively spreading knowledge would be needed
  - Texts and audio-visual forms are extended to use five sense of human organs including see, hear, taste, smell, and touch

- **New media as an useful tools of Internet and web**
  - Digital technologies including file format is expanded to provide new means of communication and expression of knowledge
  - HTML is only designated to the text-based name of media types

- **Cultural and linguistic diversity of knowledge**
  - Linguistic diversity plays a vital role in knowledge creation and accumulation
Technical Issues on Data Standardization
Data Models depending on Applications

- **Telecommunication and Broadcast Industry**
  - Telephony, SMS, AV/Multimedia stream, and AR/VR, etc.

- **Internet and Web Industry**
  - File, image, documents, and social media, etc.
  - Virtual/Object Data, and web of data

- **Location related Industry**
  - GPS, CPS, and physical 2D/3D geographic location
  - Transport and Logistics (Geolocation Map)

- **Identification related Industry**
  - Sensor/RFID code, product code, bar code, and blockchain, etc.
  - Trade, copyright, and ownership (Shipping code, product code, watermark, etc.)

- **Data Intensive or Contexts related Industry**
  - Big data analytics, Healthcare and medical applications
Key Strategies for Data Model - 1

- **Data Model according to**
  - Data Creation and Collection (device type, protocol, etc.)
  - Data Storage and Processing (Database, R, Hadoop, Excel)
  - Data Delivery (protocol)
  - Data Perception (web, speaker, screen, touch, peripheral, etc.)

- **How to avoid dependency of platform, OS, and device**
  - Minimize the commonality → minimum interoperability
    - GPU, famous OS, and cloud platform as a common?
  - Review the existing and on-going data models for their own applications
    - First, at A/V and telecommunication application as well as on-line media
    - Secondly, Geolocation, Game, Transport, Energy, Health
  - Identify the benefits of web-based data model for applications
    - XML/RDF, and HTML
Key Strategies for Data Model - 2

- **Similar to Natural Language**
  - No rule, No grammar, just talk and use the word when people thinks!
  - But, human-understandable and machine-readable logical forms
  - Minimize the preliminary and/or background knowledge
  - Minimize data exchange or summary if both sides have background information
  - Audio/visual signals are also tools to deliver information and knowledge
    - But, decoding software and perception algorithm is prerequisite

- **Data Serialization Model**
  - Used for data delivery via communication network
  - In-band or out-of-band signaling/management protocol may be needed
  - Dynamic header with semantic metadata information can be designed → parsed by intelligent processors
    - Sorting, alignment, classification, and filtering, etc.
    - (note) Most existing protocols should be designed with fixed header format
Key Strategies for Data Model - 3

- **Entity-Attribute-Value (EAV) Data model**
  - **Well-known data type or model** should be defined
    - JSON/XML/RDF, Microdata, JSON-LD, etc.
    - Binary, String, Boolean, date, day, month, etc.
    - ID, phone, email, SNS, name, repository, product code, on-line sheet, address, etc.
  - Data model for search, classification and database
    - Catalog, repository, Big file, table, structure/unstructured file, etc.
    - Hierarchical model or hash algorithm
  - **Extension of dynamic or active data model**
    - Calculate algorithm, external I/O, statistics, query, events and triggering conditions according to situation and contexts,
    - Change or update microdata or JavaScript
  - **Future flexibility to invite new data model and format**
    - Ready for and abstract/clustering of virtual cyber space
Key Strategies for Data Model - 4

- **Metadata model** as a part of data model with intelligent linked information
  - Depending on usage and business applications
    - Same data contain different metadata for delivery, processing, search, visualization
  - Relationship between Data and Metadata
    - Same format, but, metadata is contextual data that describe data or its characteristics

- **Data model for identification, authentication and authorization**
  - Data encryption and dynamic access control by using login procedure, converting table (i.e., hash table) and priority, etc.

- **Web-based data model for visualization**
  - For rendering and screen display
    - Minimize transformation overheads of XML/RDF data to HTML
    - Limitation of HTML parser → Extension of CSS and JavaScript
    - Synchronize multiple screen at separate location!
  - Web UI/UX for touch screen and peripherals including mouse and joystick
Key Issues for Data Model and Format - 1

- **Data Interoperability → web-based data model?**
  - All the stakeholder already has their own data model and format
  - Web-based data interoperability at minimum
    - HTML/XML/RDF-based (e.g., DTD, Syntax, Schema, Semantics, etc.)

- **Audio/video and communication data**
  - Metadata including time, location, format is mandate
  - Additional information may be carried by linked information (LoD)

- **Unstructured data for extracting values**
  - Out-of-band or off-line delivery of metadata
  - Raw data is associated or linked with metadata
  - Classify time-critical or valuable extracted or filtered from ordinary data

- **Data access structure**
  - Based on real-time processing, filtering requirements, access or priority class
  - Directory access protocol and distributed storage structure
Key Issues for Data Model and Format - 2

- **Directory and Data Search**
  - Directory structure for Regular or Periodic data (AMI, GIS data, etc.)
  - Storage and search structure of Web and unstructured data
  - Public open and privately preference structure (protected by authentication)
  - Identifier-based, Affiliation/repository-based dynamic data structure
  - Dynamic indexing/tagging structure based on behaviors/history and intelligence

- **Metadata, Syntax, and Semantics**
  - Human readable, self-descriptive or well-known!
  - Accumulation by collective intelligence (recursively with domain knowledge!)
  - In-text metadata or hyperlinked metadata (depending on applications)

- **Data Analytics**
  - The existing data format based on web of data
  - Step-wise add-on values for future IoT or CPS environments
Data Syntax, Schema and Semantics

- **Data Model and Data Schema**
  - **Conceptual or semantic schema**: describes the semantics of a data domain
    - Classification rule of the interest area of an organization or an industry
    - Classes and relationships between pairs of entity classes
    - Semantic data model defines the meaning of data within the context of its interrelationships with other data
  - **Logical schema**: describes the structure of information
    - Tables, columns, object-oriented classes, and XML tags
  - **Physical schema**: describes the physical means used to store data
    - Memory segment, CPUs, array, and buffer, etc.

- **Combined with grammatical rule and types**
  - Document type definition (DTD), XML/RDF schema
  - Document template and sheets (with JSON and microdata, etc.)
  - Word, keyword, macro, file, hybrid documents, etc.
Data Tag and Data Index

- Knowledge accumulation ➔ recursive with AI and machine learning
  - Extracted by Experience, Information, Knowledge and AI, etc.

- Classification rules and directory structure
  - According to structure of data, document, and contents
    - How to attach tag and create the index table?
  - Tag or Index Types (in relation with recording and storage formats)
    - Applications and contents like keyword, name, area, Image, A/V, etc.
    - Types of data: string, identifier, structured form, etc. ➔ machine readable metadata

- Search engine in terms of speed and performance
  - Search speed depending on algorithm (hash table) and directory structure
  - Index table in relation with physical/logical structure of database and memory

- Relation with software and algorithm
  - May trigger or wake up the corresponding software
  - Activate I/O devices and access the linked contents
Metadata Format

- **Metadata Classifications**
  - Structural, procedural, and descriptive metadata and their combinations
  - Depending on business and application
  - Relationship between metadata by using AI or Machine Learning algorithm

- **Level of Metadata Information**
  - 1\(^{st}\) level (**classification and index**): table, category, type, element, cost, author, etc.
  - 2\(^{nd}\) level (**processing and filtering**): condition, person, version, time, status, etc.
  - 3\(^{rd}\) level (**context-aware**): procedure, environmental or triggering condition, access rule, digital right management, etc.
  - 3\(^{rd}\) level (**intelligent**): reasoning, interest, preference, prediction, external interface with domain knowledge, etc.

- **Instance of Metadata Application**
  - Data creation instance: content metadata
  - Data collection, delivery and processing: service metadata
  - Data Consumption: usage metadata
# Microdata Format

<table>
<thead>
<tr>
<th>Format</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>JSON-LD</td>
<td>JSON-LD data is dynamically injected into web page’s contents like JavaScript code or embedded widgets since JavaScript notation is separated from the body of HTML itself.</td>
</tr>
<tr>
<td>Microdata</td>
<td>It is the nested structured data within HTML contents. It uses HTML tag attributes to name the properties of the structured data.</td>
</tr>
<tr>
<td>RDFa</td>
<td>The HTML5 supports the linked data by introducing HTML tag attributes that correspond to the user-visible contents</td>
</tr>
</tbody>
</table>
Toward Web-based Open Data Platform

- **Web-based UI/UX to help human perception**
  - CSS+JavaScript, RESTful multimodal interface, etc.

- **Data and Metadata Together → New Media**
  - Media object and media resource model
  - Media Ontology, Media Annotation (AI-based semantics)
  - Microdata format for accumulation of collective intelligence including experience, opinion, and knowledge
  - Data Format for Mashup (e.g., Energy, Health, Transport)

- **Open Data Model similar to Human Genome Data**
  - Open, Auto-configurable, and future flexible
    - But, securable and manageable for good data governance is in question?
Web-based IoT Data Model

- **XML/RDF Schema**
  - Well-known data model?
    - RAML, JSON, Microdata (e.g., vCard, hCalendar), ATOM/RSS
  - DTD, syntax, schema, and semantics, etc.
    - Tag, Index, Summary, Thumbnail, Preview, etc.

- **IoT Data Model for future life and business**
  - Starting from OCF data model, but, …
  - Micro-service, microdata for semantics and AI processing
  - Continue to URI/URL/URN-based?
  - Mashup data model for heterogeneous applications like energy, transportation, health, etc.
Toward Open Data Platform
Strategies for Open Data Standards

☐ **How to share data for knowledge accumulation**
  - Include metadata like authors, date, genre, and short summaries for easier search and discovery
  - Share individual experiences and opinions

☐ **Metadata and data schema are the key essences**
  - Provide useful information according to 5W1H rules (when data is created, delivered, processed, shared, and consumed by users)
  - Easy to extract value and reduce time during search and processing

☐ **New forms of data**
  - Create, collect, accumulate, share, and distribute data,
  - Maximize human perception by excellent user interface,
  - Accumulation of human experience and knowledge
  - Easy to find new values by AI/ML technologies
(Metadata) Data could not stand alone without metadata or descriptors

- Re-define Data and Metadata → Metadata is not only descriptive information of data
  - used for handling, sharing, and processing data

- New data format including metadata is needed
  - Active Hyperlink or Active JavaScript running at intelligent web parser → like computer virus
  - New Metadata is different from existing metadata standards
  - Recursive data format according to levels of perception and intelligence

(ref) http://aspiresquared.co.uk/
Open Data Platform - 2

- (Platform) new model for data processing including database
  - “Data + Processing + Storage” in harmony → Web-based Cloud Platform
  - New version of web platform → Web-based IoT/WoT world?
    - How to contain location, status, behaviors information that is not descriptive from existing web standards (such as text script/binary-based web)
    - Recursive data format to support complex and iterative algorithm or logics
    - New Markup language to adopt new UI/UX tools (e.g., 2D/3D drawing, gesture, expression, etc.)
  - Data platform for IoT/WoT application!
    - Web platform to reflect physical world (e.g., new organic sensors, etc.)
    - Semantics for experience/knowledge accumulation from IoT devices
  - How to build Cyber Physical System for future flexibility?
    - 3D virtual space, location, and depth/granularity/tier/level/attribute, etc.
Open Data Platform - 3

☐ (Next Generation Web) good for future open data world?
- Common platform for data creation, delivery, share, and consumption
- Keep Simple User Interface and allow billions of software/applications
  - Utilize existing wireline/wireless network, computing/software, and database
- Support future flexible data platform for energy, transport, medical/health, education, and safety, etc.
  - Don’t steal data from owners and customers → good data governance!

☐ (Cloud) sharing data by using cloud platform
- Performance, security, availability, and manageability of cloud platform are suitable for future data eco-society
- Open ecosystem for data sharing and good data governance
Toward Knowledge Information Infrastructure

- **Forms of knowledge in relationship with data**
  - Recursive mechanism to accumulate individual knowledge and opinions including tacit knowledge

- **Metadata as glue to connect data, information, and knowledge**
  - Various types of metadata may be defined
  - extract the meanings from data with intelligent processing
  - pre-processing or post-processing of data with context-aware information such as condition, situation, and environment

- **New forms of development, acquisition, and spread of knowledge**
  - new tools from social media with advances of user interface and human perception technologies

- **New Web as an useful tools**
  - new markup languages may be needed (more than HTML)
  - new sharing and communication mechanism between human-to-machine and machine-to-machine are needed (new web API)
Thank you!