International Telecommunication Union

ITU-T Technical Paper

TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU

(30 November 2016)

Analysis of Digital Data Technologies Toward Future Data Eco-Society

Forward

This Technical Paper has been developed by Mr. Jun Kyun Choi and Hwa Jong Kim.

Contents

Page

List of Figures

Page

List of Tables

Page

ITU-T Technical Paper

Analysis of Digital Data Technologies Toward Future Data Eco-Society

1 Scope

This document focuses on key issues and requirements of future data technology. First, the existing digital data technology and standards are reviewed and analysed. Toward future data eco-society, the key features of existing data formats and standards are reviewed. The data formats of telecommunication, broadcast, and Internet applications have been mainly analysed. Also, the other types of data applications like publishing, 3D, and IoT applications, are analysed to check whether these formats may be merged or integrated in a same way of future data format or not. Second, technical issues and key strategies relevant to future data standards are investigated. Key requirements for future data formats to enable efficient data sharing and aggregation are identified. These requirements can be used to initiate new ITU-T data technology standardization activities. Third, business opportunities and other similar standardization activities related to the open data eco-system are investigated. Finally, strategies the ITU-T could use to standardize open data formats and enable a data-driven world are proposed.

(Note) This report focuses on open data formats and technologies that were developed by and can be used by anyone for any purpose. Business-related or domain-specific issues like security, privacy, and digital right management are not investigated in depth.

2 Definitions

Several terms used in this document to describe the future data eco-society are defined here.

2.1 data intelligence: The use of data is to analyse human life and business, and system operations to make better decisions. (note) Data intelligence may be mistakenly referred to as business intelligence. Data intelligence focuses on data used for future.

2.2 data science: Data science focus on data quality to make a decision successfully as real and actionable data. It is inherently cross-functional to categorize data work flows such as getting data and analysing data. The high quality of data makes sure of applications to do predictive analysis. There are a lot of technologies on data science such as storing and query of structured/unstructured database, sorting and filtering of data with flexibility, scalability, and speed, etc. For social applications, tangible impact and sentiment analysis on data is used to demonstrate the value of data.

2.3 data index or data tag: Data index is to improve the speed of data retrieval operations. It is used to locate data for easy access and efficient search. The indexing or tagging methods are critical on performance which is depending on size and order of data structure.

2.4 data schema: Data schema describes the organization of data on how data is constructed. A data schema specifies knowledge representing element formats and constraints on data.

2.5 digital identity: Digital identity is an entity's online presence, encompassing personal identifying information. It can be interpreted as the codification of identity names and attributes of a physical instance. The use of digital identities is now widespread as the entire collection of information generated by a person’s online activity.

3 Abbreviations

This document uses the following abbreviations:

3D	3 Dimension
3DMLW	3D Markup Language for Web
3DXML	3D eXtensible Markup Language
6LoWPAN	IPv6 over Low-power WPAN
6Lo	IPv6 over Low-power
A/V	Audio/Video
ACC	Advanced Audio Coding
AI	Artificial Intelligence
AIDC	Automatic Identification and Data Capture
AIFF	Audio Interchange File Format
ALAC	Apple Lossless Audio Codec
ALE	Application Level Event
ANN	Artificial Neural Network
API	Application Program Interface
AR	Augmented Reality
ASCII	American Standard Code for Information Interchange
ASF	Advanced Systems Format
AVCHD	Advanced Video Coding High Definition
AVI	Audio Video Interleave
BIL	Band Interleaved by Line
BIP	Band Interleaved by Pixel
BSQ	Band Sequential
CAD	Computer Aided Design
CAGR	Compound Annual Growth Rate
CCTV	Closed-Circuit Television
CDF	Common Data Format
CGI	Commission for the management and application of Geoscience Information
CGM	Computer Graphics Metafile
CKAN	Comprehensive Knowledge Archive Network
CNN	Convolutional Neural Network
CoAP	Constrained Application Protocol
CoRE	Constrained RESTful Environments
CSDGM	Content Standard for Digital Geospatial Metadata
CSS	Cascading Style Sheets

CSV	Comma-Separated Value
DAAS	Data-as-a-Service
DAS	Direct Attached Storage
DCT	Discrete Cosine Transformation
DCW	Digital Chart of the World
DEM	Digital Elevation Model
DGN	DesiGN CAD file format
DLG	Digital Line Graphs
DOM	Document Object Model
DTD	Document Type Definition
DWF	Design Web Format
DWG	Drawing file format
EAV	Entity-Attribute-Value
ECW	Enhanced Compressed Wavelet
EPC	Electronic Product Code
EPUB	Electronic Publication
ESRI	Environmental Systems Research Institute
EUI	European University Institute
FGDC	Federal Geographic Data Committee
FITS	Flexible Image Transport System
FLAC	Free Lossless Audio Codec
FLV	Flash Video
GeoSciML	Geoscience Markup Language
GIF	Graphics Interchange Format
GIS	Geographic Information System
GISPopSci	GIS and Population Science
GML	Geography Markup Language
GPS	Global Positioning System
HDFS	Hadoop Distributed File System
HEVC	High Efficiency Video Coding
HF	High frequency
HPGL	Hewlett-Packard Graphic Language
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
IaaS	Infrastructure as a Service
ICA	International Cartographic Association
ICT	Information and Communication Technology

IDC	International Data Corporation
IDPF	International Digital Publishing Forum
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
IoT	Internet of Things
IP	Internet Protocol
IPv6	Internet Protocol version 6
IRFT	Internet Research Task Force
IRI	Internationalized Resource Identifier
ISBN	International Standard Book Number
ISM	Industry-Science-Medical
ISO	International Organization for Standardization
ISO/TS	ISO Technical Specification
ISSN	International Standard Serial Number
ITF	Interrogator-Talks-First
ITU-T	International Telecommunication Union Telecommunication Standardization Sector
IUGS	International Union of Geological Sciences
JFIF	JPEG File Interchange Format
JPEG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
JSON-LD	JavaScript Object Notation – Linked Data
KF8	Kindle Format 8
LF	Low Frequency
LLRP	Low Level Reader Protocol
LoD	Linked Open Data
LYR	LaYeR files
LZW	Lempel-Ziv-Welch
MCMC	Markov Chain Monte Carlo
MIB	Management Information Bases
MIF/MID	MapInfo Interchange File/Map Interchange Data
MIME	Multipurpose Internet Mail Extensions
MP3	MPEG-1/2 Audio Layer 3
MP4	MPEG-4 Part14
MPEG	Moving Picture Experts Group
MrSID	Multi-resolution Seamless Image Database
NaaS	Network as a Service
NAS	Network Attached Storage

NCSU	North Carolina State University
netCDF	network Common Data Format
NoSQL	Non Structured Query Language
O&M	Observations & Measurements
OGC	Open Geospatial Consortium
OS	Operating System
PaaS	Platform as a Service
PCM	Pulse-Code Modulation
PCX	PiCture eXchange file format
PDF	Portable Document Reader
PHP	Hypertext PreProcessor
PJM	Phase Jitter Modulation
PNG	Portable Network Graphics
PS	Adobe PostScript
QGIS	Quantum GIS
QoD	Quality of Data
QoS	Quality of Service
QT	QuickTime
RDF	Resource Description Framework
RDFa	Resource Description Framework in Attributes
RDFS	RED Schema
RFID	Radio Frequency Identification
ROLL	Routing Over Low-power and Lossy networks
RS-274X	extended Gerber file format
RSS	Really Simple Syndication
SaaS	Software as a Service
SDC	Smart Data Compression
SDTS	Spatial Data Transfer System
SMS	Short Message Service
SNMP	Simple Network Management Protocol
SOAP	Simple Object Access Protocol
SPARQL	SPARQL protocol and RDF query language
SSO	Single Sign-On
SVG	Scalable Vector Graphics
SWE	Sensor Web Enablement
T2TRG	Thing-to-Thing Research Group
TC	Technical Committee

TDF	Trusted Data Format
TDS	Tag Data Standard
TIFF	Tagged Image File Format
TIGER	Topologically Integrated Geographic Encoding and Referencing
TIN	Triangular Irregular Network
TR	Transcription Rules
UDDI	Universal Description, Discovery and Integration
UHF	Ultra High frequency
UML	Unified Modelling Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
USB	Universal Serial Bus
USGS	United States Geological Survey
USN	Ubiquitous Sensor Network
VCEG	Video Coding Experts Group
VPF	Vector Product Format
VR	Virtual Reality
VRML	Virtual Reality Markup Language
W3C	World Wide Web Consortium
WAV	Waveform Audio File
WebRTC	Web Real-Time Communication
WFS	Web Feature Service
WLAN	Wireless Local Area Network
WMA	Windows Media Audio
WMS	Web Map Services
WMV	Windows Media Video
WoT	Web of Thing
WoT IG	Web of Thing Interest Group
WSDL	Web Service Definition Language
WWW	World Wide Web
X3D	Web 3D Technology
XHTML	Extensible Hypertext Markup Language
XML	eXtensible Markup Language
XSD	XML Schema Definition

4 Overview and Introduction of Digital Data

Data standards are agreements on representations, formats, and definitions of data. Without data standards, the future of data interoperability is questionable. Data fields and the content of those fields need to be standardized. To drive data interoperability, the key standardization activities are primarily driven by regulation. But, since there are also many stakeholders, data standardized processes are hard on harmonization to manage data and enable data sharing. A consensus on data standardized process gathers interested individuals from industry and consumer groups, specialty domains, agencies, professional organizations, and vendors to develop a concept and express it in a standard.

The future digital data standards may allow various digital data types and formats including audio/video as well as files. In the first invent of digital data format, there were some arguments on character sets according to language. The electric documents had also suffered a lot of forms that intended to be used as printed output. In the development of computer network, the documents had been distributed in electric forms than printed materials. After improving electric display technologies, it is possible to view documents on screen instead of printing them. However, using electric document for final presentation instead of distribution and printing on the paper has some problems of multiple incompatible file format. Even more problems are connected with complex file formats of various word processor, spreadsheets and graphic software.

In telecommunication and broadcast applications, the electric document formats are used in recording and transmission, which include analogue and digitized contents. The contents can be delivered via transmission channels, encrypted in digital forms, recorded in storage and processing methods, and displayed on the screen. The metadata provides the descriptive information of the data like means, purpose, time and date, creator or author, and location.

For Internet applications, most file formats are designed to send, download, and process at the local computer system with digital data storage. Most internet files can be useful in web applications, but some electric files are not easily accessed directly through web browser. Depending on type of web pages and web applications, the document formats can be modified or processed to be visible on the screen. There are many file types and formats that are encoded for Internet. Some file formats like hypertext markup language (HTML), scalable vector graphics, and source codes are used with defined syntaxes. The human-readable parts of the data like "surname", "address", "rectangle", and "font" are also displayed in visible form. The chunk-based file format is used for play-out of audio/video file to screen. The descriptive information that identifies a particular "chunk" is called by "field name", "identifier", "label", or "tag". The data format with multipurpose Internet mail extensions (MIME) header, comma-separated value (CSV), extensible markup language (XML), and JavaScript object notation (JSON) are used on the web. Recently, unstructured file formats of raw sensing data are widely used by dumping memory or collecting sensing data of Internet of Things (IoT) devices. The unstructured data is difficult for reading and writing without conversion to a structured format. To identify, classify, and interpret a file, the internal descriptive metadata is stored inside the file itself. Typical file header contains metadata about content format, size, resolution, colour, and optional authoring information. Such metadata is used by reading, interpreting, and displaying the file. By relying on web pages, a lot of descriptive or useful information are shared with public people. The uniform resource locator (URL) is used to identify the accessible location of web page.

For the identification-related applications by using Internet of Thing (IoT), the identifier can be used to process of recognizing or identifying sensors, device/system, and persons as well as objects like file and application software. A lot of identification code including barcode which is mostly standardised by GS1 is increasingly being used in the industry [1], and the radio frequency identification (RFID) is being used as an alternative. In these applications, the identification is used to reduce running out of stock or wasted products. Credit cards and passports in the wallet are to

prove who you are. Recently, biometrics, iris recognition, and voice recognition technologies are used for identification. Theft and counterfeiting of critical or costly items like drugs, food, repair parts, or electronic components will be reduced because manufacturers will know where their products are at all times. Product wastage or spoilage will be reduced because environmental sensors will alert suppliers or consumers when sensitive products are exposed to excessive heat, cold, vibration, or other risks. Supply chains will operate far more efficiently because suppliers will ship only the products needed when and where they are needed. Consumer and supplier prices should also drop accordingly.

For the location-based applications, the geographic data format is used to capture, store, edit, analyse, share, and display spatial or geographical information. The geographical data are used for location-enabled services like transport/logistics, real estate, public safety, crime mapping, national defence, and climatology. Currently, Open Geospatial Consortium (OGC) is to make open standards for the global geospatial data [2]. The global positioning system (GPS) provides location and time information at certain application like weather forecast, e-commerce, electric map, and e-navigation [3]. The GPS-enabled mobile devices are used to display their location in relation to fixed objects (nearest restaurant, gas station, and fire hydrant) or mobile objects (friends, children, and police cars). The geographical data represent real objects like roads, lands, trees, houses, buildings, and waterways Moreover, abstraction references like images, vectors, points, lines, and polygons are mapped to location attributes. A new hybrid method of data is identifying the physical location which combines three-dimensional vector points of physical space. This location information is becoming more realistically visually descriptive. Recently, the web page with huge amounts of geographic data enables users to create customer applications and make complex spatial information, which is called mashup application of the web. An editable map of the geographical data is used to offer street maps, aerial/satellite imageries, geocoding, search, and car navigation

For data intensive applications, a large volumes of data typically terabytes in size and referred to as big data are processed. A database file format refers to how data is stored on a hard drive. Some database file is used for human readable or searchable. Cloud computing applications requiring large volumes of data and their processing times to I/O are deemed data intensive. The rapid growth of the Internet led to vast amounts of information available online. Parallel processing can typically involve partitioning or subdividing the data into multiple segments which can be processed independently using the same executable application program in parallel on an appropriate computing platform. The data-intensive computing is managing and processing exponentially growing data volumes, significantly reducing associated data analysis cycles to support practical and timely applications. Information extraction and indexing of web documents can derive significant performance benefits on data parallel executions since the web can be processed in parallel. The semantic query language like SPARQL protocol and RDF query language (SPARQL) is enabled to retrieve and manipulate data stored in RDF format of the web. Massive data from millions of IoT sensors needs the non-structured query language (NoSQL) database for storage and retrieval of data, making some operations faster than the relational database. The high-speed ICT infrastructure allows the data to be partitioned among the available computing resources and processed independently to achieve performance and scalability based on the amount of data. The cloud computing system controls the scheduling, execution, load balancing, communications, and movement of programs and data across the distributed computing clusters.

For data analytics, the first process is collecting data from a variety of sources. The data can be collected from sensors, instruments, and devices like remote surveillance camera, motion detecting device, recording system, and satellite. It can be also obtained through human interviews, Email, web page, social networking services, and on-line documents After cleaning incomplete, incorrect, and duplicated data, it is arranged for process of data exploration. Second, data is processed and organized so that a variety of techniques for data analytics are applied to obtain descriptive statistics in graphic form and additional insight for decision making. The results of data analytics are shown in human readable screen and communicate the message to remote audiences. Data visualization

uses information displays like tables or charts to explain the quantitative messages contained in the data.

For science and engineering applications, various types of signals or information like electromagnetic signals or biometric information are converted to digital forms. The weather conditions and chemical formula are represented by digital data. The conversion of analogue symbols or signals to digital is needed to relevant mapping methods or converting rules. The urgent problem is backward data compatibility since most scientific data has their own old data formats like common data format (CDF), network CDF (netCDF) and Flexible Image Transport System (FITS) in NASA [4]. Recently, Hadoop Distributed File System (HDFS) is widely used for big data analytics since it is a Java-based file system that provides scalable and reliable data storage and is designed to large group of servers with up to several hundred Petabytes of storages [5].

4.1 Data models

A data model is an abstract model that organizes elements of data. The generic data model is similar to that of a natural language. It defines the relation types like classification relation and part-whole relation. A data model enables the expression of binary relation between an individual thing and group of things, which are predefined in the fixed and limited domain. The conventional data models have some shortcomings for data exchange and data integration. Recently, a semantic data model in software engineering is a technique to define the meaning of data within the context of its interrelationship with other data. It defines how the objects or symbols are related to the real world. Therefore, a semantic data model is sometimes called a conceptual data model.

For current computer science and mathematics, the entity-attribute-value (EAV) model is a data model to describe entities where the numbers of attributes (properties, parameters) that can be used to describe them are potentially vast, but the number that actually applies to a given entity is relatively modest. This model is known as object-attribute-value model, vertical database model, and open schema. This data representation is analogous to space-efficient methods of storing a sparse matrix, where only non-empty values are stored. The data type of EAV offers a limited set of data types: byte, Boolean, DateTime, double, and string, in addition to dividing numeric data into int, long, or float. It also defines custom data types like a phone number, an e-mail address, geocode, and a medical record The cloud computing system offers data stores based on the EAV model, where an arbitrary number of attributes can be associated with a given entity. XML provides a framework on top of an EAV design and builds an application that has to manage data sets extremely complicated when using EAV models.

For audio/video and communication applications, the data serialization model is used. The context of data serialization is the process of translating data structures or objects into a format that can be stored in a file or memory buffer, or transmitted across the network. For communication network, this model is not straightforward since data serialization is formatted by their associated protocols. In addition, a communication network running on a different hardware and software architecture should be able to reliably reconstruct a serialized data stream. Serializing the data structure prevents the problems of byte ordering, memory layout.

For software engineering including data analytics, the metadata model is to analyse, construct, and develop the frame, rules, constraints, models, and theories, which are applicable and useful for the predefined data applications. It describes the contents and contexts of data or data files. Metadata is similar to the card catalogues of libraries. As information has become increasingly in digital form, metadata is used to describe digital data. For example, most files and documents include metadata specifying what language the page is written in, what format was used to create it, and where to find more information about the subject. There are two types of metadata: structural metadata and descriptive metadata. Structural metadata is the data about the containers of data. Descriptive metadata uses to describe individual instances of application data or the data contents. The main

purpose of metadata is to facilitate in the discovery of relevant information, more often classified as resource discovery. Metadata also helps organize electronic resources and provide digital identification.

For web application, the eXtensible Markup Language/Resource Description Framework (XML/ RDF) format can be used to the processing of XML/RDF document which is published at W3C [6]. Technically, a XML schema is an abstract collection of metadata, consisting of a set of schema components, mainly elements, attribute declarations, and complex and simple type definitions. These components are usually created by processing a collection of schema documents, which contain the source language definitions of the components. Schema documents are organized by web namespace including syntax and grammar. All the named schema components belong to a target namespace which includes types, values, properties, and attributes of the schema document. A schema document includes other schema documents by using the same namespace, which can import schema documents for a different namespace.

With the advent of web browser, there are many markup languages, especially the hypertext markup language (HTML), which is the standard markup language used in web pages [7]. It is a markup language that web browsers use to interpret and compose texts, images and other materials into web pages. Web browsers can read HTML files and render them into visible or audible web pages. HTML describes the structure of a website semantically for presentation, making it a markup language, rather than a programming language. The HTML elements form the building blocks of all websites. HTML allows images and objects to be hyperlinked, which can be used to create interactive forms. It provides a means to create the structured documents by denoting structural semantics for texts like headings, paragraphs, lists, links, quotes, and other items. The language is written in the form of HTML elements consisting of tags enclosed in angle brackets like <html>. Browsers do not display the HTML tags and scripts, but use them to interpret the contents of the web page. HTML can include scripts languages like JavaScript which affect the behaviours of HTML web pages. Web browsers can also refer to cascading style sheets (CSS) to define the look and layout of texts and other materials.

If the metadata is inserted in HTML format, it is easy for peoples to share web pages through network. The files representing metadata can be grouped into three parts: structured texts from reference points to data, how the files can be accessed, and location information of files. HTML prescribes how the texts will be formatted visually, which fonts will be used and on which place, where the image will be situated, and where the heading of the chapter is located However, it is typical for the descriptions of the documents to be classified into various categories. These categories form a certain hierarchy depending on their significance of contents. The differences of contents are not always represented visually in formatted documents, but they are important for the mass processing of metadata.

4.2 Data storage

There was a long history of writing, recording, and storing information. Recording can be done using virtually any form of energy, spanning from manual muscle power in handwriting, to acoustic vibrations in phonographic recording, to electromagnetic energy modulating magnetic tape and optical discs. Electronic data storage requires electrical power to store and retrieve data. Electromagnetic data is stored in either an analogue data or digital data on a variety of media. This type of data is considered to be electronically encoded data, whether it is electronically stored in a semiconductor device. Most electronically processed data storage media (including forms of digital data) are considered permanent (non-volatile) storage, that is, the data will remain stored when power is removed from the device.

Except for printed data, electronic data storage is easier to revise and is more cost effective than alternative methods due to smaller physical space requirements and the ease of replacing (rewriting)

data on the same medium. However, the durability of printed data is still superior to that of most electronic storage media. The durability limitations can overcome by the ease of duplicating (back-up) electronic data.

In the digital age, the long-term durability is more significant since valuable portions from more than several zeta-bytes of the data should be permanently stored for history. With advances of large scale cloud computing system, access speed and bandwidth of stored data are greatly improved, not only for achieving the storage capacity with more than several zeta-bytes. The data files stored on millions of cloud servers constitute educational, cultural, and scientific resources. Recently, "web culture" is characterized by the extreme rapidity of data-flows and rapid obsolescence. The average lifespan of an Internet page is less than one month. To archive data in the cloud computing system, the successive version of the same documents should be lined up with its date of release. Digital storage at the cloud computing are designed to provide acceptable performance in terms of access speed, life duration, and release time. In near future, the new storage technologies permit important advances regarding the accessibility and manageability of data.

4.3 Data classification and filtering

In the era of zeta-bytes, data searching and filtering technologies are important. The data creation is useless if the data could not be searched or indexed by search engine. Therefore, data would be well arranged, sorted, and prepared for searching, filtering, grouping and classification. Data with searchable index or tag is the process of organizing data into categories for effective and efficient use. A well-planned data classification form makes essential data easy to find and retrieve. This can be of particular importance for access and search. The relevant procedures for data classification should define what categories and criteria people will use to classify data. If a data-classification scheme has been decided, the appropriate index or classification procedures for each category should be addressed with data's lifecycle requirements. It is essential that data classification is closely linked with data categories. Data classification is clustering the data sets by an iterative process of data category. New data sets can be categorized by new classification rules of intelligence. The effectiveness of data classification is measured by searchability, speed of sorting and clustering, scalability on large amounts of data, and robustness of data quality.

In scientific and engineering applications handling massive amounts of data like satellite or geographical investigation, data classification raises the issues of identifying new observations. It can merge investigating, analysing, and learning processes at an instance. Total volumes of data can be minimized by grouping data into categories based on the measure of inherent similarity. Data clustering for pattern recognition from a large amount of image data is used to identify a member of possible classes with highest probability. Probabilistic algorithm with statistical inference is to find a best instance. In experimental and statistical analysis, data classification is done with logistic regression or a similar procedure. New observations on experimental results are referred to create new categories of possible values or outcomes.

4.4 Meaning of hyperlink at web page

The hyperlink over electric document is used to link any other information through the web page. In the web pages written in the hypertext markup language (HTML), the hyperlink is a reference to data that the reader can directly follow by clicking. Users navigate or browse the web page following the hyperlinks. Most hyperlinks cause the target document to replace the document being displayed. A link information from one web page to another is a uniform resource locator (URL). It is achieved by means of an element with a "name" or "id" attribute at the HTML document. The document containing a hyperlink is known as its source code document. For example, in a web site, many words and terms are hyperlinked to definitions of those terms. Hyperlinks are often used to

implement reference mechanisms, like tables of contents, footnotes, bibliographies, indexes, letters, and glossaries. The linked data describes a method of publishing structured data and enables data from different sources to be connected and queried. The linked open data (LoD) is the linked data that is open content. The URIs can be used to look up the things which name is identified by using open standards like RDF, SPARQL.

4.5 Digital publication in aspects of data format

Until now, books have been recognized as the useful material to write, print, and illustrate works of literature or human history. The electronic books are widely distributed for educational, living and business purposes. Digital printing has opened up the possibility of print-on-demand with relevant updates. It should be noted that digital books or magazines should not be modified or changed after their electronic publication. To face an ever-increasing rate of publishing, sometimes called data explosion, new contents and information are readily updated during the electronic printing of books. If digital technology is used in book design, there will be a new art of incorporating content, style, format, design, and sequence of a book. New digital books or magazines will be hyperlinked with interdisciplinary knowledge, ready for academic discussion, and collecting various opinions from social networking services. Digital books are constantly updated and hyperlinked with the advances of contents, which is similar to publication on websites.

Online newspapers can present breaking news in a timelier manner. The credibility and strong brand recognition of well-established newspapers are also seen by the newspaper industry as strengthening their chances of survival. Online newspapers are more or less like hard-copy newspapers and have the same legal boundaries, like laws regarding libel, privacy and copyright. A blog or a wiki is nevertheless not clear to the public. News reporters are being taught to shoot videos and to write the news pages.

In academic publishing, a scientific journal is a periodical publication intended to further the progress of science, usually by reporting new researches. Most journals have been peer reviewed to ensure that articles meet the journal's standards of quality and scientific validity across a wide range of scientific fields. The publications of scientific research are an essential part of the scientific method. If they describe experiments or calculations, they should supply enough details that an independent researcher could repeat the experiments or calculations to verify their results. Such an article in a journal becomes part of the permanent scientific records. Articles in scientific journals can be used in research and higher education. Scientific articles allow researchers to keep up to date with the developments of their research field. An essential part of a scientific article is the citation of earlier works. The impact of articles and journals is often assessed by counting citations.

Electronic publishing is a new area of information dissemination. In an electronic (non-paper) form, scholarly scientific results are written or created for publication or dissemination. The electronic journal is specifically designed to be presented on the website. The electronic journal will exist alongside the paper version because the latter is not expected to disappear in the future. The output on a screen is important for browsing and searching. Many journals are electronically available in formats readable on the screen via the web browsers. Electronic publishing of scientific journals is not costly, is accessible to many people, and is doable due to the availability of supplementary materials (data, graphics, and video).

In many fields in which even greater speed is wanted, the role of the journal in disseminating the latest researches has largely been replaced by electronic databases. An increasing number of electronic journals are available as open access. Individual articles from electronic journals can be found online and stored either in personal or community archives, or posted at websites as blogs.

4.6 Media in aspects of data format

Media is to transport information like newspapers, radio, and television. To indicate the means of human communication like language, reading, writing or audio/video/music, there are technologies and methods that support communication over distances in time and space. Media is physically stored content (in the case of files) or transferred content (in the case of messages), audio/video/music, film, photos or more generally of data. The media technology provides cost-effective solutions to help human intelligence by applying various scientific knowledge like electronics, telecommunication, computer science, mathematics, physics, material science, human-machine interaction, cognitive science, perception psychology, sociology, and economics However, today's media technologies are mainly built on electronic and computer systems, which are called digital media or social media.

A wiki is a website which allows collaborative modification of its content and structure directly from the web browser. In a typical wiki, texts are written using a simplified markup language, running on wiki software. There are at least tens of thousands of other wikis in use, both public and private, including wiki functions, notetaking tools, community websites and intranets. Some wiki engines permit control over different functions (levels of access). For example, editing rights may permit changing, adding or removing materials. Others may permit access without enforcing access control. A wiki allows evolving, complex, and networked texts with argument and interaction. A characteristic of wiki technology is the ease to find which pages can be created and updated. Many wikis are open to alteration by the public without requiring registration of user accounts. Many edits can be made in real-time and appear almost instantly online. Private wiki servers require user authentication to edit pages, and sometimes even to read them.

A blog is a personal online journal that is frequently updated and intended to the open public. A key characteristic of blogs is interactive, allowing visitors to leave comments and even messages to each other on the blogs. The interactivity of blogs distinguishes them from other static websites. "Bloggers do not only produce contents to post on their blogs, but also build social relations with their readers and other bloggers." The one is more personal online diaries and the other is more of an online brand advertising of a particular individual or company. A typical blog combines texts, images, and links to other blogs, web pages, and other media related to its topic. The ability of readers to leave comments in an interactive form is an important contribution to the popularity of many blogs. There are many different types of blogs: personal blogs, collaborative blogs, group blogs, microblogging (the practice of posting small pieces of digital content which could be texts, pictures, links, short videos, or other media on the Internet), corporate and organizational blogs.

As a popular free social networking service, Facebook allows registered users to create profiles, upload photos and video, send messages, and keep in touch with friends, family, and colleagues. Within each member's personal profile, there are several key networking components. The most popular feature is a virtual bulletin board in which messages left on a member's Wall can be texts, videos or photos. Another popular component is the virtual photo album in which photos can be uploaded from a desktop or directly from a smartphone camera. An interactive album allows the member's contacts to comment on each other's photos and identify (tag) people in the photos. A microblogging feature allows members to broadcast short announcements to their friends. All interactions are published in a news feed, which is distributed in real time to the member's friends. Recently, personal live video broadcasting service is available. Facebook offers a range of privacy options to its members. He can block specific connections and keep all his communications private. For members who wish to use Facebook to communicate privately, the messages closely resemble e-mails. Facebook allows for both asynchronous and synchronous dialogues and supports the integration of multimodal contents like user-created photographs, video, and URLs to other texts.

Twitter is a free social networking microblogging service that allows the registered members to broadcast short posts called tweets. Twitter members can broadcast tweets and follow other users'

tweets by using multiple platforms and devices. Tweets can be sent by cell phone text messages. To weave tweets into a conversation thread or connect them to a general topic, members can add hashtags to a keyword in their post. Tweets can be delivered to followers in real time, they might seem like instant messages to the novice user.

YouTube is a video-sharing website. Available contents include video clips, TV clips, music videos, and other contents like video blogging, short videos, and educational videos. Most of the contents on YouTube have been uploaded by individuals. The unregistered users can watch videos and the registered users can upload videos to their channels. However, at the time of uploading a video on YouTube, the copyright issues are controversial since there are still many unauthorized clips of copyrighted materials.

In comparison with other media, social media are a blending of technology and social interaction for the co-creation of values. People obtain information, news, and other data from electronic media. They enable anyone to publish information as a type of user-generated contents. Social media have provided an open environment where people are free to exchange ideas on technologies, applications, brands, and products. One characteristic of social media is the capability to reach small or large audiences, for example, either a blog post or a television show reaches some people or millions of people.

4.7 Web technologies for future data formats

Many people recognize that web technology is a web page by using the web browser. A web browser displays a web page on a display monitor or mobile device. With graphic user interface, the web page is what is displayed, usually written in HTML. Web browsers coordinate the various web resource elements for the web page like style sheets, scripts, and images. Typical web pages provide hypertexts which include the navigation menu to other web pages via hyperlinks. A web browser can retrieve a web page from a remote web server. The web browser uses the hypertext transfer protocol (HTTP) to make requests to the web server. The web server may restrict access to only a corporate network. A static web page is delivered exactly as stored in the web servers, while a dynamic web page is generated by a web application that is driven by server-side software or client-side scripting. A dynamic web page is created at the server side when it is requested and served to the end users. These types of web pages typically do not have a static URL. The design of a web page is personal according to one's own preferences. Many people edit the contents of a web page by using web templates.

A web document is similar in concept to a web page, but a web document has its own uniform resource identifiers (URIs). It should be noted that a web document is not the same as a file. A single web document can be available in many different formats and languages. A single file, for example a hypertext preprocessor (PHP) script, is responsible for generating a large number of web documents with different URIs. A web document is defined as HTML, Joint Photographic Experts Group (JPEG), or resource description framework (RDF) in response to HTTP requests. As for the resources identified by uniform resource identifier (URI), the user gets a readable representation of the web, in which the resources are not only web documents, but also real world objects like cars, buildings, sensors, and non-existing things. A web service is a software system designed to support interoperable machine-to-machine interaction over a network. Technically, a web service describes a standardized way of integrating the web-based applications using the XML, simple object access protocol (SOAP), web service definition language (WSDL) and universal description, discovery and integration (UDDI) open standards.

4.8 Big Data

At 21st century, so called "era of big data," the amount of data being generated is increasing dramatically. Big data is initiated be user generated video, picture upload services, social network services. Moreover, the ways that data is being used impact daily life, business, economy, industries, and education. With the wide use of smartphones and computers, new businesses have emerged that enable people all over the world to use their electronic devices and make use of countless data, resulting in innovative services and more personalized services. The data generated from various smart devices has evolved into the era of big data and the data business industry utilizing data is growing at a rapid pace.

Many people call these impact as a "big data". Even though there is no clear definition of big data, many companies and institutes are using big data, and related solutions are called as big data technologies. Examples of big data technologies include Hadoop, machine learning, NoSQL database, data visualization, natural language processing, and image analysis. The big data can be shortly stated that "process data in order to get some insight from the data", and this process is called academically as "data science". Data science is a new terminology that came with the big data, and its analytic processes, representing the whole process of data gathering, analysing, and applying the results to real world applications. Big data is becoming the key component of modern ICT applications. Big data is changing many business paradigms (data driven marketing), scientific research, social analysis, safety and security check, smart factories. Big data has initiated many changes in industries, ICT technologies and services.

4.9 Data Analytics

The amount of data being generated is increasing exponentially. IBM estimates that people create 2.5 quintillion bytes (i.e., 2,500,000,000 gigabytes) of data every day, so much that 90% of the data in the world today was created in the past two years alone. At 2014 study by global market intelligence firm IDC suggests that the 4.4 zettabytes (i.e., 4,400,000,000,000 gigabytes) of data that existed in 2013 will grow tenfold, to 44 zettabytes, by 2020.

The data is transforming the world. Advances in hardware, software, and algorithms used to analyze the data are revolutionizing domains from business to education to manufacturing. Existing technologies like urban planning tools and medical diagnosis systems are being reinvented by data. Emerging technologies like autonomous vehicles and augmented reality are made possible by data. Data will continue to influence, accelerate, and drive almost all human progress [8].

A data arms race has been started. Individuals, organizations, companies, and governments are rushing to develop increasingly sophisticated technologies that generate, manage, and analyze ever-increasing amounts of data [9]. Not all of these developments are good. As the world becomes more data-driven, data-related issues will become more prominent.

4.10 Machine Learning

Machine learning generally refers to the development of computer algorithms that identify patterns in data without being explicitly programed to do so by being shown many examples. These algorithms typically require a human to manually translate the raw data into a set of features the human thinks might capture the key aspects of the raw data.

There are three types of machine learning:

- (**Supervised learning**): Where the algorithm learns to map an input X to an output Y.
- (**Unsupervised learning**): Where the algorithm learns the characteristics of an input X.

- (**Reinforcement learning**): Where the algorithm learns what actions to perform to optimize target score S.

Some well-known examples of machine learning applications include recommendation systems on e-commerce sites and on-demand video platforms; spam detection in email; self-driving cars; credit card fraud detection systems; automatic speech to text transcription; and intelligent personal assistants. Most of these examples started being useful in day-to-day life over the past decade thanks to massive amounts of data, cheap access to good computing power, and advancements in algorithms like artificial neural networks.

4.11 Deep Learning

Deep Learning is a subset of Machine Learning. Artificial neural networks (ANN) are a type of machine learning algorithm that automatically identify patterns in data by mimicking the structure of the human brain. In ANNs layers of neurons are connected together by nonlinear functions to transform inputs, like the intensities in an image, into outputs, such as the number of the numeral in the image. Deep learning specifically refers to ANNs with several layers of neurons.

The most famous deep learning architecture is convolutional neural networks (CNNs). CNNs are a subset of ANNs that are modelled of the human visual cortex and are particularly good at image recognition tasks. In contrast to traditional neural networks, where every neuron one layer is connected to every neuron in the next layer, the connections in this type of neural network correspond to overlapping regions of the input field. These networks can have billions of connections between neurons and are trained on millions of images on GPU clusters. Once trained, networks are in the small hundreds of megabytes and can be easily deployed anywhere. Then, new images can be analysed in microseconds.

4.12 Emerging technologies for data intelligence

The recent new technologies like cloud computing, web/application software as well as 5G communication technologies are just the beginning of a wide variety of technological developments for the future. The future society is ready to invite new technologies like big data analytics, deep learning, augmented reality/virtual reality (AR/VR) as well as Internet of things (IoT). In the near future, network access speeds will be exceeding more than 1 terabit per second. The processing power of cloud computing will be more than several hundred petaflops at the surprising costs. The storage capability of individual smartphones or personal computers will be more than 1 terabyte. Many people feel that networking and computing resources are unlimited and plentiful like fresh air and fresh water.

Recent advances of artificial technology and machine learning give some feeling to many people that the thinking capability of computer is sometimes superior to that of human. Therefore, a lot of people think about how to utilize computer and network with their interest and business world. For analysing new information and fining new knowledge, humans focus on how to think rather than on how to remember. Humans welcome to utilize the storage and processing capability of the cloud computing system. The computer with artificial intelligence help human with how to think and remember. To overcome the language barrier, real-time language translation will be available. Additionally, the searching machine displays in advance the relevant information on the screen from the websites if people are discussing some outstanding issues.

For future of media, the web technologies are important to enable content sharing, creating ideas collectively, and accumulating business intelligence. A cloud-based platform supports that the web technologies can enable for people to build up new ecosystem of life and business.

4.13 Emergence of Internet of Thing (IoT)

Advancements in electronics, software, and manufacturing in the form of Internet of Things (IoT) devices will generate vast amounts of data in many domains. Data from sources as diverse as weather sensors, smart electric power meters, and connected shipping containers will be fused to power more advanced, more personalized, and more widespread services that will in turn create even more demand for devices that generate yet more data. People may say that data would be analysed "later", because people do not know how much value the data might have when the data is generated (or gathered). Along with technology improvements, or due to new needs, people would try later to get more (hidden) value form the data.

Currently, fusing data from multiple sources involves gathering that data in one central location and analyzing all of the data at once. However, IoT devices are expected to generate so much data so quickly and in so many locations that the devices themselves will likely have to perform real-time analysis to decide things like what to do immediately, what data to send to the central servers for more processing, and what data to discard. Facilitating these types of systems will require developing new data analysis paradigms.

For remote surveillance or monitoring environments, the location of IoT devices should be remotely identified. Also, the detail profile and operational rules of IoT sensors should be well specified regardless of their locations. It means that overall application scenario should be defined to get a value when IoT sensors are setting up at the certain location. But, it notes that the identification mechanism of IoT devices have been standardized in alignment of services/applications, objects, and platform. But, it notes that simple identification code is not enough to configure IoT services. Moreover, the interoperability problems relating to IoT technologies and protocols as well as identifying IoT objects will be significant in near future.

To solve the interoperability problems, the concept of Web of Thing (WoT) is one of the suitable solution. The web technologies will be used to help IoT services. The web site or web page provides the overall information for IoT services. In a case, the relevant mechanism is needed to activate IoT devices and collect the sensing data. One solution is that the special Javascript code on web browser activates such IoT applications, which is mainly based on HTML. Otherwise, the SOAP over HTTP or web sockets including Web Real-Time Communication (WebRTC) is applicable to such real time IoT applications. If an IoT device has a role like web client, the data created or collected by IoT devices should be sent to web servers. In this case, the web server has a role of IoT service platform. The web server collects the status of IoT devices and the web browsers at the client side display their presence information to user screen. In case of emergency, an IoT sensor detects and sends its urgent notification information to web server. When the web server receives the urgent information, human identify the urgent status and quickly decide the responsible actions, in which all the operations will be shown at the web page. For the WoT scenarios, the users access the specific IoT services via the specific web platform which is identified by URL. It means that the web platform is used to solve all the interoperability problems nested in the IoT services.

In the WoT scenario, the web page has a kind of bridge role for information sharing between IoT sensors and human. The question is that "the current web architecture is suitable for handling IoT sensors?" If a web application has a direct interface to IoT sensor which is similar to the external input/output, the web page is a kind of the overall management platform including configuration, monitor, change, update, and download, etc. In this case, the web screen is the management screen for human decision. For WoT applications, it should be analysed whether the web architecture based on HTML is useful or effective or not. If some people does not agree to use URLs as management screen for WoT applications, new platform can be invented with acceptable level of consensus. The extension of the current social networking platform can be a candidate as future IoT service platform.

4.14 Emergence of new media

There are some evidences to suddenly appear new media with advances of information and communication technology. A lot of activities embrace new media formats and experiments in producing content in as many different ways as possible. The creative tools of new media are generally inexpensive, at least compared with the existing media tools. In social media environment, contents can begin on fully personal own platform and migrate into social distribution channels.

The rise of new media has increased communication between people all over the world. It has allowed people to express themselves through blogs, websites, videos, pictures, and other user-generated media. New media most commonly refers to contents available on-demand through the Internet, accessible on any digital device, usually containing interactive user feedbacks and creative participation. New media includes the existing social media like online newspapers, blogs, wikis, video and games. New media enables people around the world to share, comment on, and discuss a wide variety of topics. One of the key features of new media is denoted as interactivity among communities.

New media has the integrated forms of digital data and software that are manipulated, networkable, dense, compressible, and interactive. It combines Internet accessible digital texts, images, and video with web-links, creative participation of contributors, interactive feedback of users, and formation of a participant community of editors and donors for the benefit of non-community readers. People in virtual community's exchange information for life and business. New media has the ability to connect like-minded people worldwide and feeds into the process of guiding their future development.

By using smartphone and smart pad, mobile social media makes use of the location- and time-sensitive marketing and communication aspects like sales promotions/discounts and relationship development. Mobile social media offers that offline consumer moves to online world. Mobile social media applications are influencing an upward trend in the popularity and accessibility of e-commerce or online purchases.

4.15 Advent of Artificial Intelligence

Advancements in the algorithms used to analyze data, particularly in fields like machine learning and artificial intelligence (AI), are fundamentally changing the way that societies, economies, industries, and jobs work. In January 2016, the World Economy Forum reported that roughly 50% of current jobs will be fundamentally changed thanks to artificial intelligence, robotics, and other forms of automation. These machine learning and AI algorithms are powered by massive amounts of data. The better the data, and the more data there is, the better the algorithms perform. This means the quantity and quality of data that exists will continue to grow in importance.

There are two types of artificial intelligence (AI):

- (**Specialized AI**): which are better than the best humans at achieving a specific goal?
- (**Generalized AI**): which are as good as or better than the best humans at achieving any arbitrary goal?

Specialized AIs are good at singular tasks like "tell me what objects are in this image", "translate this English sentence in Chinese", "figure out what movies this Netflix user wants to watch next". Recent advancements in artificial neural networks have helped people figure out how to do better than humans in a lot of specific tasks like these. The rapid progress in building specialized AI, from

cute theoretical models to fully-deployed, human-level systems in less than five years, is what's caused the explosion of interest by technology visionaries, futurists, and the media.

Good language translators are not what is causing people to call AI "summoning the demon" (Elon Musk) or something that could "spell the end of the human race" (Stephen Hawking), nor what is making thousands of the world's foremost AI experts sign petitions calling for more research into proper AI safeguards, nor what made Silicon Valley tech luminaries donate $1 billion US Dollar in December 2015 to foster AI research "free from financial obligations, so we can better focus on a positive human impact", nor what is making Hollywood haemorrhaging movies exploring AI like Her (2013), Chappie (2015), Ex Machina (2015), Transcendence (2014), Avengers: Age of Ultron (2015), Big Hero 6 (2014), and Terminator Genisys (2015). All of this come a fear that generalized artificial intelligence would, either accidentally or purposefully, exterminate the human race. Fortunately, most AI experts think generalized AI won't come until 2050, so we have time to figure out how to teach AIs morality.

In the short term, most AI research is devoted towards improving specialized AIs, typically by training more and more powerful machine learning models on specific tasks.

4.16 Toward the data-driven world

A data-driven world is a world driven by data: everything is being improved, replaced, and redesigned by data, and anything not continually reinvented to incorporate new data quickly becomes obsolete. In a data-driven world, being competitive requires being able to get, analyze, and use data well. Currently, generating data is a resource-intensive process; analyzing data is a specialized field; and using data requires a lot of experience making data-driven decisions. Big companies who have the resources to pursue all three of these components are growing increasingly strong. Many small companies and individuals, however, do not have the resources to pursue any of those three components, much less all three. There is the increasing gap between those that have enough resources to pursue the value of data and those who do not continue to widen as more types of data. More complicated methods of analysis are developed. For everyone to benefit from the coming data-driven world, then, people need to explore new ways to generate, analyze, and use data efficiently enough to share the value of data fairly.

The ITU-T will play an important role in this exploration. Traditionally, the telecommunications sector focused on transmitting data from place to place efficiently. However, as the amount of data increases significantly, how a user uses data becomes much more important than how a user transmits data. Instead of developing entirely new solutions to develop guidelines for making using data more efficient, it makes sense to extend the work of the existing solutions with a lot of experience in developing guidelines for making communicating data more efficient.

5 Review of existing digital data format and standards

Since the beginning of the human history, the means of recording data and information have been developed and continuously evolved over time. To record and preserve information, mankind started to draw images and write letters on not only cave wall but also various materials, including wood, metal, clothes and etc. As one of the most revolutionary events in the mankind history, the invention of the paper and printing technologies provided humankind important means to record, preserve and distribute information. Furthermore, humans developed methods to record sounds, images, and videos in analog formats: film, tape, vinyl and etc. By the advent of computer and electronic memory disks, the history of recording data and information met another turning point from analog to digital. Including hard disk, floppy disk, optical disk, flash memory sticks and etc.,

numerous types of physical memory disks have been developed. Evolved from the physically restricted storage, distribution, and exchange forms, data and information are recently exchanged, shared, searched, discovered, accessed, linked, and connected through a new medium, so called the World Wide Web or simply known as the Web.

In these days, data has become ubiquitous and datasets are continuously getting larger and more complex. The purpose of this document is to propose the needs of standardized data formats for this newly emerging data and media era and the data format that will be discussed in the document denotes the method for representing data in a digital store. To exchange, perceive, manage, process and reconstruct data, in short, to prevent data format obsolescence and isolation, its stored format is one of the most important aspects of data. In advance of discussing the direction how new media is changing and what types of formats and standards are needed in details, existing formats of digital files —including publication, audio/video, and web— and some features of newly emerging areas with related standards are briefly discussed in this chapter. Since there exist an enormous number of data formats and standards, only most pervasively used ones in these days among the common data formats and standards will be discussed rather than enumerating the entire list.

5.1 Digital File

5.1.1 Introduction

According to Hilbert, the leading data stored type is gradually and rapidly changed from analog to digital between 1986 and 2007, and people have entered into the "digital age" in 2002 as the global information storage capacity of digital files exceeded 50% of the entire data files, and at the point of 2007, digital file took 94% of the total data store type [10]. Data in various types—mostly text, image, audio and video—are stored, processed, transported, and shared in digital formats in these days. As the purposes, types, characteristics, and etc. of data are getting more diverse, data formats and standards are adapting to the new data ecosystem and evolving as well.

In this section, existing digital file formats and standards of text, image, audio, and video will be discussed briefly in advance to propose an overall picture of how emerging data and media era is approaching.

5.1.2 Text

There exist a couple standards for encoding and representing texts in digital which embrace letters and numbers. The most well-known and widely used standard is American Standard Code for Information Interchange (ASCII). ASCII character set uses 7 bits per character—or 8th bit is used as a parity bit. That is, ASCII can generate total 128 number of codes. 95 are graphic codes which are displayed on a console and 33 are control codes which control communication or console features. The extended version of ASCII was developed to cover more number of characters by using the full 8-bit code to represent a character. However, 256 characters are not enough for international use. Unicode, another standard for texts, is developed as a superset of ASCII. Instead of using the 7-bit code, Unicode uses a 16-bit code, thus is able to represent 65536 characters. The first 256 characters in the Unicode character set are exactly the same as extended ASCII character set.

While ASCII is limited to represent some symbols and alphabet-based characters, Unicode is able to embrace different base characters, like Arabic, Korean, and etc. To improve the efficiency in storing and transferring digital data, many encoding methods were introduced, including keyword encoding, run-length encoding, and Huffman encoding. These encoding methods focus on the reduction of bit rate by eliminating redundancy or cleverly assigning shorter bit codes to frequently appearing letters or numbers.

As a basis of data store and transmit forms, text-based data formats will never wither even though new types of digital data and format will keep emerged. Due to its compactness and simplicity, it requires less processing overhead to store, transform, or transport data. Therefore, from simple data storage to big data analysis, they have been and will be widely used in various fields. Other than simple plain text, Comma-Separated Values (CSV) is one of the most frequently used text file format. CSV uses comma as a delimiter to separate fields—sometimes other delimiters are also used—and stores tabular data in plain texts. Because its file size is small and it is easy to process, CSV is often used in storing an enormous amount of data for big data analysis. However, CSV file format is not standardized, and because of its loose terminology, it is not possible to perceive the meaning of individual field and the value from each column without any additional descriptive information. Therefore, it is often accompanied by another text file, README for example, which contains descriptive information about the CSV file.

Although some limitations of text file formats exist, they still have outstanding advantages over other data type file formats: simplicity, compactness, and processability. In these reasons, text files are well used for information storage, transformation, and transport widely in a diverse range of fields. Even in newly emerging data and media era, text file formats will not vanish due to its irreplaceable characteristics and distinct advantages.

5.1.3 Image

Images can be digitalized by discretizing continuous analog imagery into digital space. Digital image files can be divided into three classes according to compression methods: uncompressed raw, lossless compressed, and lossy compressed. Uncompressed raw image files are digitalized image files without any compression process afterward. The quality of an original image can be retained, but the file size is large. Therefore, uncompressed raw images are not frequently used on media other than to preserve the original quality of the image for some specific purposes—for example in photography, medical analysis, and other fields where untouched details in images are needed. Lossless compression is often used to reduce file size without losing any data from the original resource. On the other hand, lossy compression tradeoffs the file size and the quality of the image. Therefore, a lossy compressed image has a smaller file size than lossless compressed one, but the lossless compressed image has better quality compared to the lossy compressed one.

However, digital graphic files can be separated into two main types according to composition and representation methods as well: raster image format and vector image format. Often called bitmap images, raster images are made of squared pixels just like a mosaic. The most common raster image formats are Joint Photographic Experts Group (JPEG), JPEG File Interchange Format (JFIF), Graphics Interchange Format (GIF), Portable Network Graphics (PNG), and Tagged Image File Format (TIFF). On the other hand, vector images are composed of geometric descriptions, like thin lines and curves. The most common vector image formats are Adobe PostScript (PS), Computer Graphics Metafile (CGM), Gerber format (RS-274X), and Scalable Vector Graphics (SVG). As shown in the example in Figure 1, when images are zoomed in, the square pixels of the raster image become visible while the lines of the vector image stay smooth and keep their geometric shapes. Since raster images are more commonly used in general, in this section popular raster image data formats will be shortly discussed.

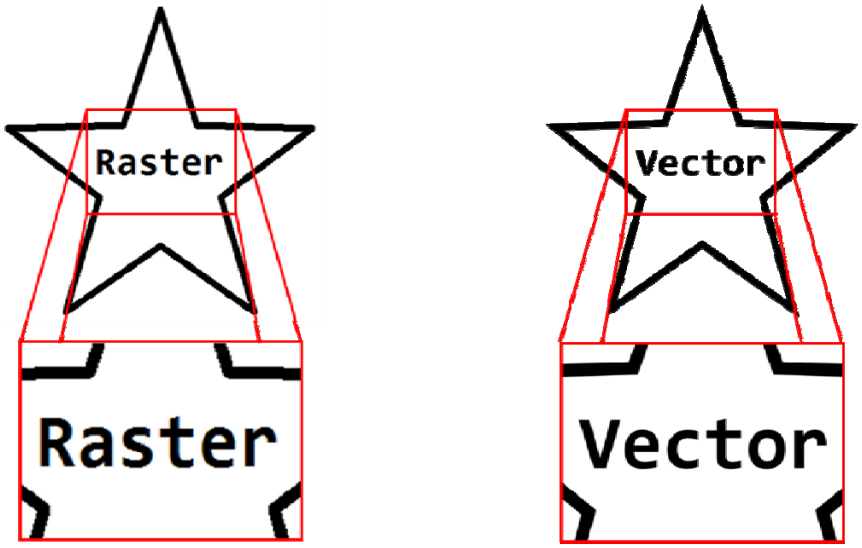


Figure 1. Raster and vector image comparison example

JPEG is a lossy compression method for still digital images and JFIF is a file format that commonly holds JPEG compressed data. However, regardless the format variations, image files compressed by JPEG are often called JPEG. JPEG is the most common file format for storing and transferring photographic digital image data on Web environment and many electronic devices. JPEG uses discrete cosine transform (DCT) and typically achieves 10:1 compression rate without perceptible degrade in image quality [11]. JPEG was designed for detail photographic images and during compression, it removes information unnoticeable by humans. Since it is a lossy compression that uses block DCT, JPEG is suitable for photographs or realistic images rather than created images with lines, texts, drawings, or elements with distinct geometric properties. When an image is compressed too much, JPEG compression results in some blocky noises in the image. The blocky noises degrade the overall quality of the image especially in distinct lines and geometric shapes, like letters, symbols, logos, animation, and etc. To complement the weakness of JPEG, JPEG 2000 was developed as a successor. Instead of DCT, JPEG 2000 uses wavelet transformation method. Although it is more efficient in compressing an image in the range of 20%, it has not been adopted by many fields since the implementation is too complex and complicated compared to its achievable improvement rate [12]. Therefore, JPEG is still the leading data format on the Web.

GIF uses lossless compression and is suitable for storing relatively simple images with few colors and shapes, like diagrams or cartoons. Although its compression ratio is low, it is widely used in image animation effects due to its animation capabilities. PNG was developed as an alternative to GIF by supporting more color depth and additional alpha channel for transparency. However, PNG is often not supported by many software and Web, thus not considered a good file format for widespread use. TIFF is a file format for digital images that uses Lempel-Ziv-Welch (LZW) compression method, and as a container, it can hold image files that are uncompressed or compressed by JPEG. Because TIFF's tagged structure is designed to be flexible and easily extendible, private tags can be applied inside a TIFF file, and it is often used as proprietary formats. It also supports layers and transparency. Since TIFF offer uncompressed file storing, it is suitable to edit and store files without degrading the original quality. However, TIFF is somewhat inappropriate on Web environment because of the large file size when it is not compressed.

Short summary on the digital image file formats and standards discussed in this subsection are described in Table 1.

Table 1. Summary of popular digital image file formats and standards

	Features	Recommended	Not Recommended
JPEG	• Lossy Compression • Detailed photographic images	• For photographic images • On Web	• To edit • For simple graphics
GIF	• Lossless Compression • Animated graphic images with small size	• For simple graphics and animations with limited color	• For photographic images
PNG	• Lossless Compression • Good replacement of GIF: Supports millions of colors and alpha channel for transparency	• For simple graphics and animations	• For photographic images • For widespread use: since not well-supported on Web and many software
TIFF	• Offers lossy (JPEG) or lossless (LZW) compressed and uncompressed files • Supports tags, layers, and transparency	• To edit and store: in case that the file is uncompressed	• To reduce file size • On Web

Each format has their own advantages over others. However, JPEG seems to be the leading image data file format on Web environment because of its compactness. Since JPEG reduces the file size in a great matter while minimizing the quality degrades, it is suitable for storing and transferring on Web environment.

5.1.4 Audio

Analog sound data can be digitalized by sampling, periodically measuring the voltage in the sound signal. The digital audio formats can be classified into three parts: lossy compressed audio formats, lossless compressed audio formats, and uncompressed raw audio formats. Lossy compressed audio is compressed with some loss of the sound data during the compression process to diminish the file size. The compression is done in the way not to degrade the sound quality as much as possible so that people cannot notice the degradation in most cases. The most common lossy compressed audio formats are MPEG-1 and/or MPEG2 Audio Layer 3 (MP3), Advanced Audio Coding (AAC), and Windows Media Audio (WMA). Lossless compressed audio is the compressed digital audio to reduce the file size without losing any sound data of the original complete source through post-processing. A lossless compressed audio file size is much larger compared to a lossy compressed audio file, but it has much higher sound quality. The most common lossless compressed audio formats are Free Lossless Audio Codec (FLAC) and Apple Lossless Audio Codec (ALAC). Uncompressed raw audio is an intact digital audio file converted from an analog form without any post-processing. The most common uncompressed audio formats are Pulse-Code Modulation (PCM), Waveform Audio File Format (WAV), and Audio Interchange File Format (AIFF). However, uncompressed audio file formats are hardly used.

The total amount of data created worldwide has exponentially increased, thus efficiently storing the data to save memory space and transfer the data fast with low cost have been becoming important. With these reasons, lossy compressed audio files are widely used in general in media and in ordinary life due to their advantage of the reduction in file size with small degrade in overall quality. Of course, uncompressed raw files and lossless compressed files are still used in special fields which require high-quality unrefined audio files or consumed by some people who have an exceptional hearing sense to feel the difference between lossy and lossless audio files. Details in

few widely used lossy compressed audio file formats —MP3, ACC, WMA— will be discussed separately in section 5.2 along with video formats.

5.1.5 Video

To reduce the file size for storing and transferring, video files are often compressed as well. Video file formats can be separated into codec and container formats. Codec encodes and decodes a digital data stream or signal to compress and decompress a video file. It interprets a video file and determines how the media file will be played. Container can be considered as a case that holds a set of data files, including the video and audio stream files, codecs for video and audio, metadata, and etc. as described in Figure 2. Container gives users some controls over how the media will be created or consumed by letting users choose which codec to be used each for video and audio. Common container formats are Audio Video Interleave (AVI), Advanced Systems Format (ASF), QuickTime (QT), MPEG-4 Part14 (MP4), and Flash Video (FLV). Widely used video codec formats are H.264 and Window Media Video (WMV). Details in these video formats will be discussed separately in section 5.2 along with audio formats.

Figure 2. Visual description for container format structure of audio video file

Data is incessantly evolving as adapting to new environments and ecosystems. Therefore, with the advance in technologies and the emergence of a new ecosystem, new types of data will appear and data formats and standards will keep evolving as well.

Short summary on the overall digital file formats and standards discussed in this section are described in Table 2.

Table 2. Summary of common digital data formats and standards

Data Type	Common Formats and Standards	
Text	Encoding Character Set	ASCII, Unicode
	Organizing format	Plain text, CSV (or with other diameters)
Image	Raster	JPEG/JFIF, GIF, PNG, TIFF

	Vector	PS, CGS, RS-274X, SVG
Audio	Uncompressed	PCM, WAV, AIFF
	Lossless Compressed	FLAC, ALAC
	Lossy Compressed	MP3, ACC, WMA
Video	Codec	H.264, MP4, WMV
	Container	WMV, AVI, ASF, QT, MP4, FLV

5.2 Audio / Video

5.2.1 Introduction

Among all the five senses that humans have, the sense of sight is frequently considered as the particularly crucial one for humans to perceive data and information. As the sense of sight can help to compensate a loss of the other senses, seeing contains much more information than we usually think. People obtain a lot of information by seeing. They perceive and understand faster and more easily what can be seen than what cannot be. Sound as well is an element that is hard to be described in words and takes a huge part in communicating information for humans. One can easily find that he or she obtains much more information—and much more precisely—from just one image or some sound than hundreds of lines of word descriptions. Therefore, audio and video files have been the most important and frequently used form to express, store or share information for humans. As popularly used mediums for data exchange, they have been continuously evolved over time along with the advance of technologies and there exist a countless number of data formats and standards for audio and video data. Therefore, in this section, the most common data formats and standards for audio/video will be discussed rather than the entire list.

5.2.2 Codec and Container

As the total amount of data created worldwide has exponentially increased, efficiently storing the data to save memory space and to transfer the data fast with low cost have been becoming important. With these reasons, lossy compressed audio files are widely used in general because of the reduction in file size with almost unnoticeable degrade in overall quality. Of course, uncompressed raw files and lossless compressed files are still used in special cases, but they are hardly used and often not well supported on many environments. Therefore, in this section, only few pervasively used ones from lossy compressed audio file formats—MP3, ACC, WMA— are discussed in detail.

To reduce the file size for storing and transferring, video files are often compressed as well. Video file formats can be separated into codec and container formats: compressing and holding the compressed data streams with the codec. Codec encodes and decodes a digital data stream or signal to compress and decompress a video file. It interprets a video file and determines how the media file will be played on a screen. Container is a solution to hold a set of data files, including the video and audio stream files, video and audio codecs, metadata, and etc. Common container formats can be listed as Audio Video Interleave, Advanced Systems Format, QuickTime, Advanced Video Coding High Definition, Flash Video, and etc. The widely used video codec formats are H.264 and Window Media Video.

In advance to discuss the details in popular audio and video file formats, many of which are developed by Moving Picture Experts Group (MPEG) and Video Coding Experts Group (VCEG), their works and progress are briefly described in Table 3 [13]. MPEG and VCEG are the two working groups that work side-by-side to lead setting standards for audio and video. MPEG is a working group formed by International Standard Organization (ISO) and International Electrotechnical Commission (IEC) and VCEG is a working group formed by International

Telecommunication Union Telecommunication Standardization Sector (ITU-T). Especially, MPEG plays an important key role in both audio and video formats and standards as they have developed many well-used techniques, formats, and standards on moving pictures with audio for decades.

Table 3. Summary of the works by MPEG

Release Year	Standard	MPEG	VCEG	Part	Type / Layer	Usage
1993	ISO/IEC 11172	MPEG-1		**Video Home System (VHS) and Television Recording**		
				Part 1	Systems	
			H.261	Part 2	Video	
				Part 3	Audio	
					Layer I	
					Layer II	
					Layer III	MP3
1995	ISO/IEC 13818	MPEG-2		**Broadcast, Distribution, DVD**		
				Part 1	Systems	
					Program Stream	
					Transport Stream	
			H.262	Part 2	Video	
				Part 3	Audio	
					Layer I	
					Layer II	
					Layer III	MP3
				Part 7	Advanced Audio Coding (AAC)	AAC
1999	ISO/IEC 15938	MPEG-4		**Broadcast, Internet, Blu-ray**		
				Part 1	Systems	
			H.263	Part 2	Video	
				Part 3	Audio	AAC
			H.264	Part 10	Advanced Video Coding (AVC)	MPEG-4 AVC
				Part 14	MP4 Container	MP4
2013	ISO/IEC 23008	MPEG-H	H.265	Part 2	Video	High Efficiency Video Coding (HEVC)

MPEG-1 Part 3 Layer III, well known as MP3, is the most pervasive digital audio coding format which uses a form of lossy data compression. It obtains a good compression rate by cutting off all the data beyond the common human hearing range and compressing the remaining data. Its good compression rate is one of the reasons for its popularity. ACC was also developed by MPEG as the successor to MP3. AAC uses more advanced compression technics than MP3; therefore, AAC

generally has a better sound quality at the same bitrate with MP3. Despite its advancing quality, AAC is not popular as much as MP3, but it is still widely used today. It is one of the standard audio compression method used by a lot of services including YouTube, iTunes, Nintendo, PlayStations and more [14]. WMA was designed by Microsoft to address some of the flaws in MP3, but it does not have a standing-out-benefit over MP3 or other audio formats; therefore, not many devices and platforms support WMA.

MPEG-4 absorbs and improves some features from MPEG-1 and 2, and extends new features like support for 3D rendering, object-oriented composite files, external digital rights management, and etc. [15]. MPEG-4 is evolving standard and adapting to the newly coming media area with the new features. H.264, also known as Advanced Video Coding (AVC), is developed by MPEG as part 10 of MPEG-4. It is developed to extend the variety of applications and maintain the quality of the video even at lower bit rates compared to previous standards. However, its processing power requirements can be high. Developed by Microsoft, WMV was originally developed for internet streaming applications. WMV codec is able to play the video even when only small part of a video is downloaded and the remaining part of the file is still being downloaded. However, other than this feature, it does not seem that WMV has more outstanding advantages over MPEG formats.

AVI and ASF are container formats developed by Microsoft and mostly holds data compressed with WMA or WMV codec. They had been prevalent container formats for a long term due in no small part to the stability of Microsoft, but recently its popularity has been dipped. QT is a container format used natively by the QuickTime framework. It can contain abstract data reference, separate the media data from the media offsets, and tracks edit lists; therefore, it is suited for editing. AVCHD often used to hold data compressed with H.264 codec and it's the 2.0 version supports 3D. These container formats tend to be highly dependent to codec formats. For example, MP4 is the most popular container format for MPEG-4 and H.264 formats. WMV and ASF are the most popular container format for WMV formats since they are developed together by the same group. Therefore, the leading container formats seem to be determined by which codec formats will lead and what types of contents will be consumed.

As ISO standards, MPEG audio and video codecs and container formats have been widely used with fewer restrictions on hardware and native environment since their standards were designed for universal usage, not for specific environments nor by certain companies. On the other hand, the performance and stability of the container formats designed by some specific companies, that are QT and WMA, tend to depend on their own native environments. Additionally, the popularities of all Window Media formats were affected by the fact that they were bundled with and preinstalled in Windows. Therefore, without outstanding efforts of other formats to outstrip MPEG, MPEG tends to lead the audio and video file formats currently. FLV has distinct advantages over other formats on the Web environment specifically, thus it is widely used for streaming videos across the internet. Although its file size is big because of its byte-aligned format, premising that a working flash browser plugin is installed, FLV is easy to use and computationally cheaper to play, thus suitable for the Web environment. However, Flash players may have problems on security and processing power. Some websites are currently moving away from Flash. Therefore, the usage of Flash for Web is declining.

MPEG has researched on compression and code expression methods to efficiently compress and transfer the data of moving pictures. Although some properties of the content will be continuously changed over time, effectively reducing audio/video file size and transferring data fast with low bit rate will be the main technological issues to solve even in the future. These properties are the key advantages that MPEG holds. For example, in these days the most image files we handle are in two dimensions, but in the future, it seems that three-dimensional images will emerge soon enough. Even when handling three-dimensional video contents, reserving and managing the memory storage well by appropriately compressing the file and transferring data efficiently with low cost will still

play the key roles in the newly emerging video ecosystem. Therefore, it seems that MPEG's influence will not be easily ceased but continued in the approaching future.

5.2.3 Metadata

Metadata is data describing information about other data. The ability of machines to directly perceive the content of audio and video by analysing the data itself is relatively low and imprecise at a moment. Therefore, when metadata associates with data in video forms, it plays a critical role for machines to understand the content and to parse or find wanted information. The metadata accompanying video is often inserted in the container formats in front of or at the end of the data stream. Currently, the metadata often contains the title, publisher, keywords, description, codec format, and other overall generalized information about the entire video contents or formats, but not specific information about certain events or each key-frame. Therefore, the current method to structure and insert metadata possibly has eased the needs for parsing and searching a video but does not have any advantage for parsing and finding certain sections with specific features inside a video stream. The current method of structuring audio video content related metadata is focused on describing what the complete video is about rather than what is happening in the video and what types of events are occurring. However, metadata can provide much deeper information beyond simple information about the general facts of the video. If metadata can be inserted in-between frames to describe the detail information in the video, as shown in Figure 3 and Figure 4, parsing and finding information can be much efficiently done. For example, in Romeo and Juliet movie, if a consumer wants to find the scene when Romeo and Juliet first meet, the consumer has to go over and find the scene manually. However, with event triggered metadata tags inserted in-between frames, machines can easily perceive events or key features based on a timeline, thus details in the content as well. Therefore, to be prepared for the emerging semantic data era, more works on metadata associating with audio and video formats seem to be needed.

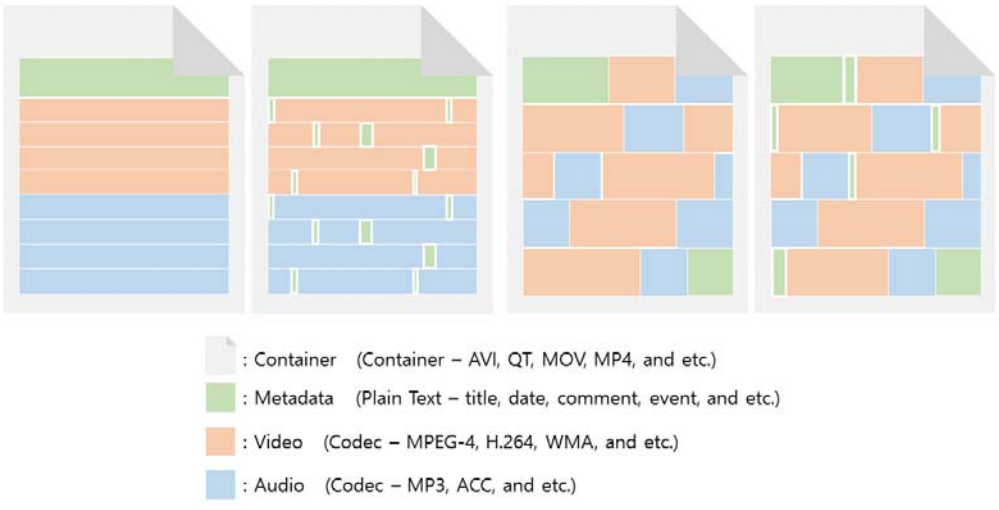


Figure 3. Metadata inserted in-between frames or data stream chunks

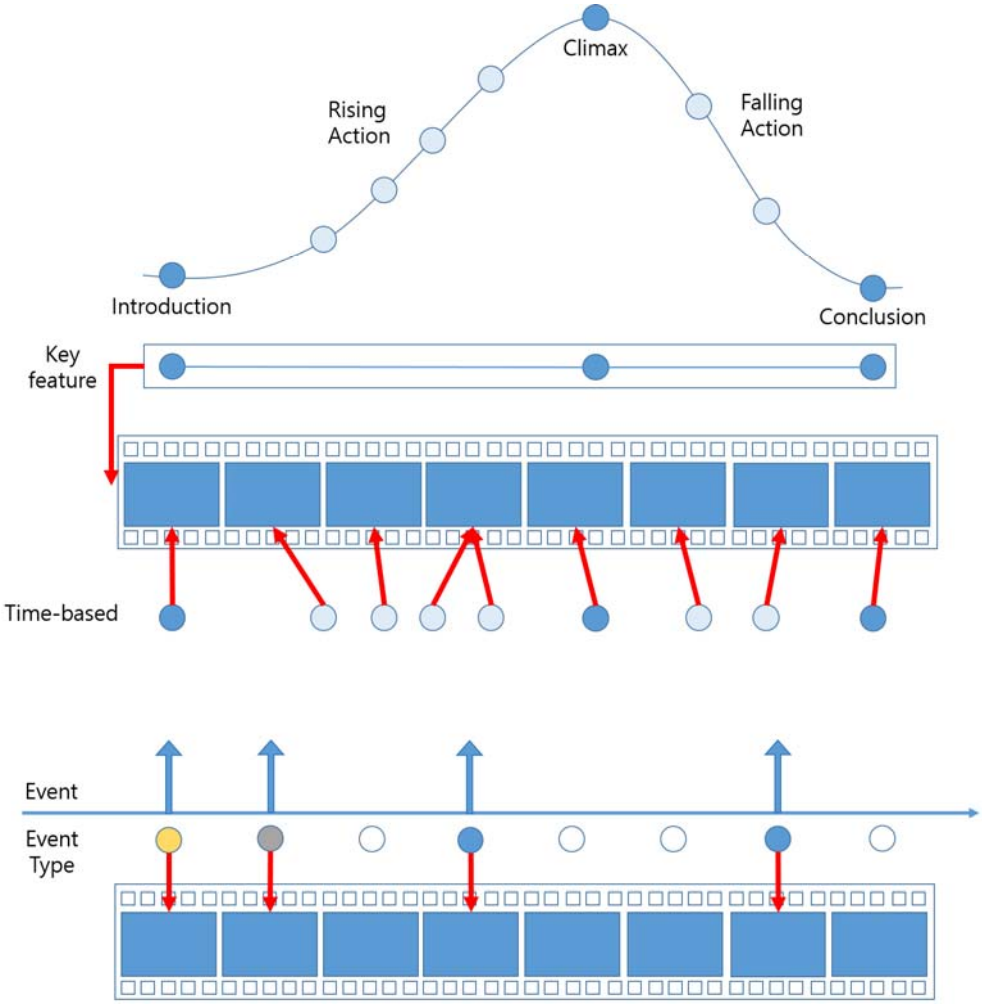


Figure 4. Time-based and event-type-based metadata

5.3 World Wide Web

5.3.1 Introduction

The World Wide Web, the so-called Web, was invented by Tim Berners-Lee in 1989, and first web server and client were deployed in 1990 [16]. Web pages are primarily text documents encoded with Hypertext Markup Language (HTML) and linked to one another through hyperlinks of their unique identifiers which are often referred as addresses of the website. Each page has a unique Uniform Resource Identifiers (URI) and web pages interlinked by hypertext links and accessed via the Internet and Web browsers. Web resources may contain not only formatted text but also image, audio, video and software components rendered in web browsers. Started from 1991 when only one website existed, the number of websites has been dramatically increased until now as its growth is shown in the Figure 5 [17].

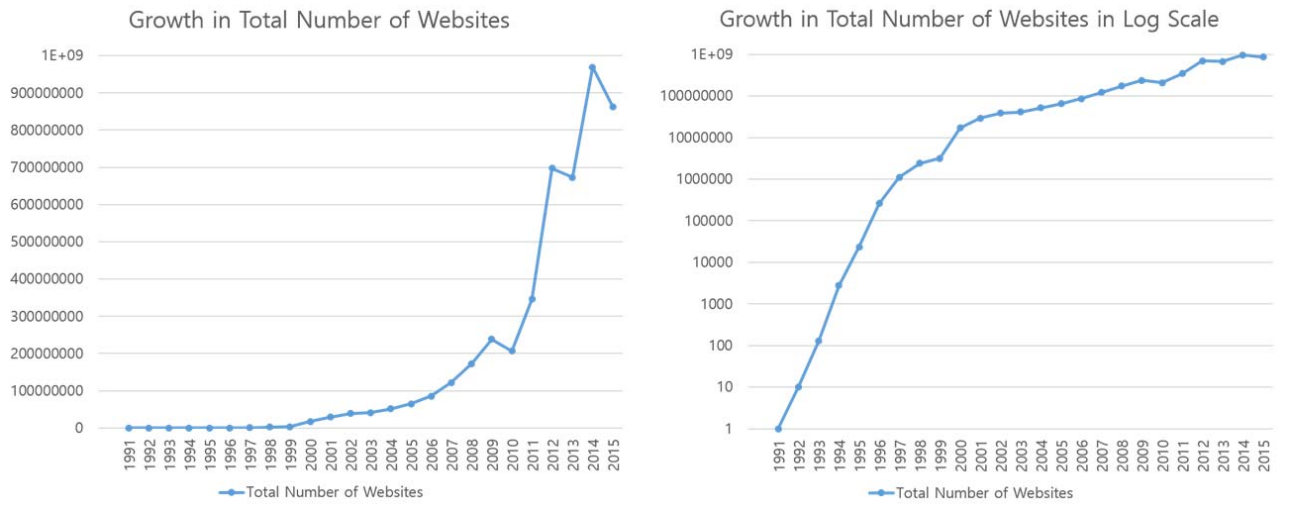


Figure 5. Growth in the number of websites

The Web has been the biggest medium for data exchange during a significant amount of time, and even at this moment it is continuously enlarging its power and influence. Although its growth speed is slowing down as it is approaching the saturation point, as the purposes and key features of the Web comply with the newly emerging fields and data properties, for example 3D and IoT, the new data era will accelerate the growth of the Web again rather than slow it down. Moreover, even if the growth in the number of existing websites has become on the wane, the growth in its functions, power, and importance are keep increasing more and more.

As the types of data are getting various and the amount of data is also wildly surging, the needs are rising to intelligently manage and control data, not only the ones newly being created but also the ones already generated. Therefore, machine-readable languages and data formats are getting spotlights to organize the resources on the Web processable by machines. However, the initially designed structure of the Web was focused more on the representation of the exchanged data but not on machines' capabilities to perceive or interpret the meanings or properties of the data. Although the recent trend of the Web is leaning toward semantic components from mere data sharing concepts, the current structure of the Web is yet more adequate for the aspect of data exchanges, and there still are some restrictions and shortages of existing ways to semantically manage the data on the Web.

As a virtual space for integrated data and information share, the Web is often used to search data or information. The keyword matching is the current major search method on the Web. This method is able to find web pages including the specifically searched keywords, but cannot take the contexts of the found web pages into consideration to fit the users' detailed intentions. Therefore, without technological help for machines to understand the contexts of data by semantically linking them together to represent their contexts, in relations, properties, constraints, and etc., users have to manually go through another post-processing of information which is given from the keyword matching search. To compensate this weak point of the structure of the Web, markup languages are on the rise, which describe the contexts of data and link them together according to their relations, properties, and etc.

In this section 5.4, markup languages are reviewed from the ones that have been the most pervasively used since the creation of the Web to the ones that are the recent trends. For the better understanding of the overall structure of the web, basic concepts of structural and operational components of the Web are briefly described at the following subsection. Then, semantic components of the Web will be introduced in the next subsection

5.3.2 Structural / Operational Components

In advance of semantic components of the Web, basic concepts of structural and operational components of the Web are briefly described to help the understanding of the overall structure of the Web.

Uniform Resource Identifiers (URI)

URI is a unique identifier of each resource on the Internet. Uniform Resource Locator (URL), also informally known as a web address, is the most popular form of a URI. Uniform Resource Name (URN) is also a type of URI. While URL resembles an apartment street address, URN resembles an owner's name of each unit. For instance, as the example in Figure 6, URL indicates the actual web address that a resource located and URN indicates the name of the resource.

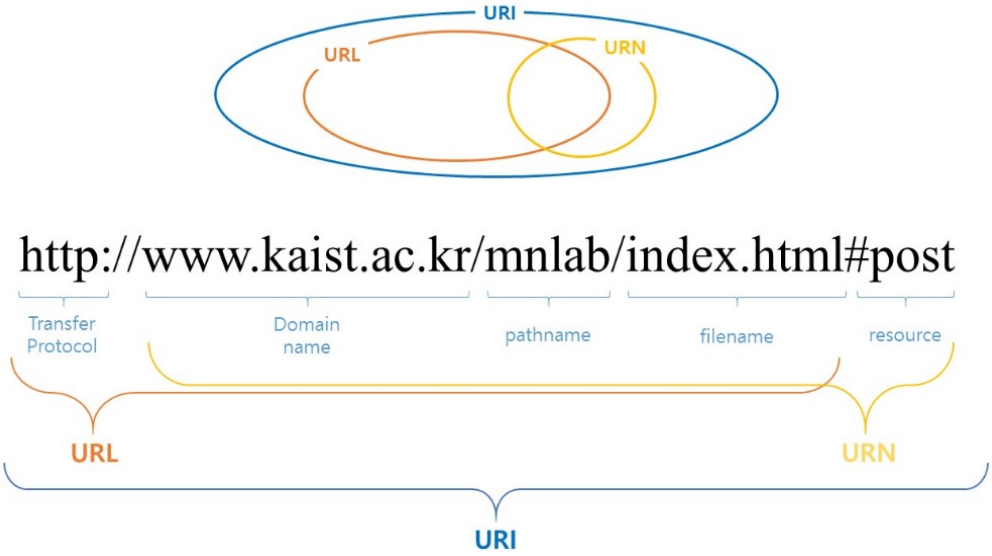


Figure 6. URI, URL, and URN example

Hypertext Transfer Protocol (HTTP)

The communication for the Web, HTTP is a request-response protocol between client and server. Relied on URI naming mechanism, HTTP is a simple request-response protocol as a client sends a request message and server replies with a response message. In these days, almost every data communication on the Web are made through HTTP.

Internet Protocol (IP) Address

A unique identification number, so called IP address, is assigned to every device connected to a network that uses the internet protocol for communication. Although it is assigned to every device, it is not a permanent address for the specific device. It is rather a network address for internet connection.

Communication on the Web

On the Web, hypertext links are linked and connected together according to their unique URIs. Clients use browsers and request contents of a web page with corresponding URI via HTTP.

Servers response with the requested Web contents or an error message. As client sides obtain the Web contents as a response, browsers render the contents. When a device is connected to the Internet for communication, a unique IP address will be assigned to each device so that servers can successfully receive a request from and send a response to the particular clients that asked for specific resource.

5.3.3 Semantic Components

This subsection focuses on semantic components of the Web. The current structures of data resources on the Web including their functional abilities are described. Only the most pervasively used ones are discussed: HyperText Markup Language (HTML), Cascading Style Sheets(CSS), JavaScript, Extensible Markup Language (XML), JavaScript Object Notation (JSON), and Resource Descriptive Framework (RDF).

The term 'Semantic Web' was first introduced by World Wide Web Consortium (W3C) referring its vision for the Web linked data [18]. W3C has devoted efforts to facilitate the Web of data environment in standard formats, and under the auspices of W3C, XML and RDF had been developed. A simple basic architectural layer stack of Semantic Web with RDF, XML, and their schemas are described in Figure 7 [19].

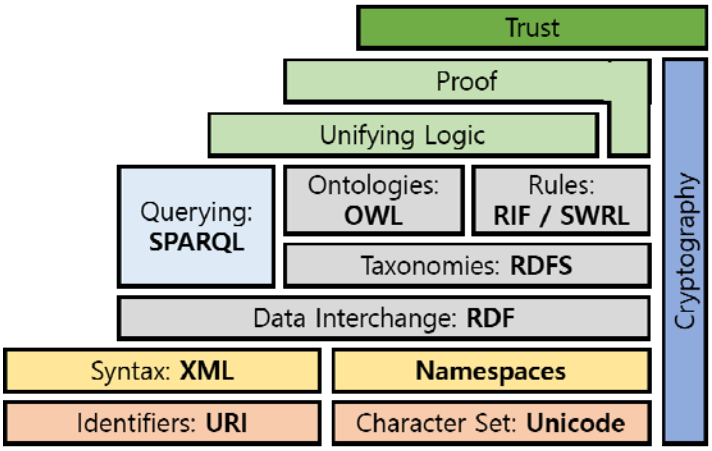


Figure 7. Semantic Web architectural layer stack

HyperText Markup Language (HTML)

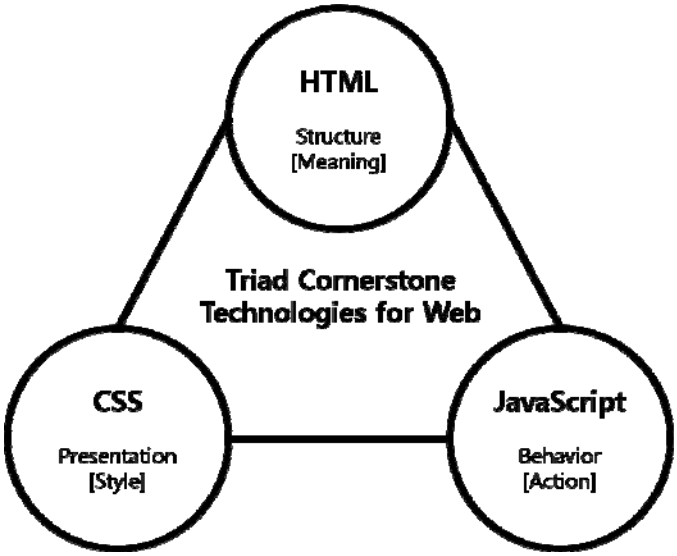


Figure 8. Triad cornerstone technologies for the Web: HTML, CSS, and JavaScript

HTML is the core markup language used to create web contents and structures of data documents which can be rendered on web pages by web browsers. It is one of the triad cornerstone technologies for the Web with Cascading Style Sheets(CSS) and JavaScript. HTML describes content and structure of a document with two-sided tags in text forms and facilitates a hypermedia environment on the semantic Web. A two-sided tag is composed of an opening tag and a closing tag as a pair, and document contents are contained in between the two-sided tags. Inside the opening tag, attributes to each element can be added to provide additional information about the element. HTML attributes come in name and value pairs. Attributes can be considered as the type of the tags assigned to an element—like 'html', 'title', 'body', 'img', 'table' and much more. HTML was initially designed to facilitate the Web environment by semantically describing presentation to organize and display content on The Web as it was initially designed. Since HTML focuses on content visual representation in the Web environment, it has limitations on semantically linking resource and data. Only the restricted range of predefined tags can be used, and there are some restrictions to express metadata; therefore, HTML often meets limitations on semantically describing structures or properties of data accurately or in details.

A brief example of HTML document to show the basic structure of an HTML page is described in Figure 9. As shown in the first line of Figure 9, HTML manages how the content looks like on a web browser environment. A document's background color, text color, font style, font size, text alignment and so many more styles can be added inline by using a style attribute in HTML elements or internally by inserting a <style> attribute in the head section. However, they also can be applied externally through using one or more external style sheets. That is where CSS comes in to separate from the data content by containing their visual representation style.

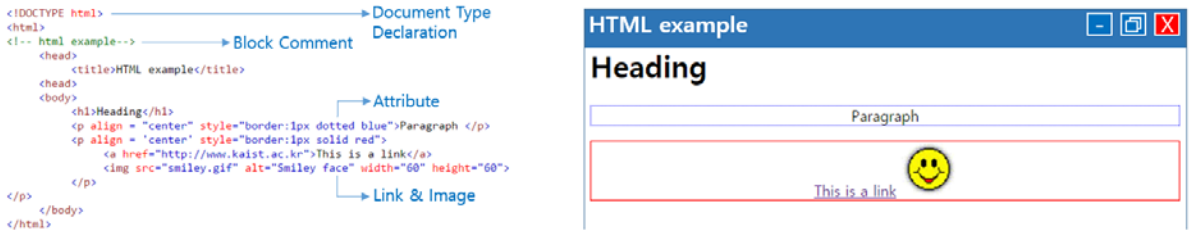


Figure 9. Simple example of HTML document

Cascading Style Sheet (CSS)

As one of the triad cornerstone technologies for the Web along with HTML and JavaScript, CSS is used to manage the visual style of documents written in markup languages, often as in HTML but still applicable to XML as well. The style definitions mostly saved in external files so that the entire presentation of a content can be transformed instantly by changing just one file. It describes how elements are to be displayed in various media, including screen and paper, and defines styles for web pages, including the design, layout, and variation in the display for different environments, like device types and screen sizes [20]. Since CSS removes the style formatting from documents written in a markup language by separating style presentation from contents, it saves a lot of work and time on generating web contents. It enables different markup pages to share style formats or augments the compatibility of a markup document at multiple devices. IT can easily render the same markup page to different styles or sizes of layouts. As shown in Figure 10, the same content can be displayed in multiple designs and styles easily by using different CSSs without changing the contents itself. Therefore, it improves accessibility and flexibility of the content and reduces complexity and unnecessary repetitions in formatting the visual style of markup documents [21].

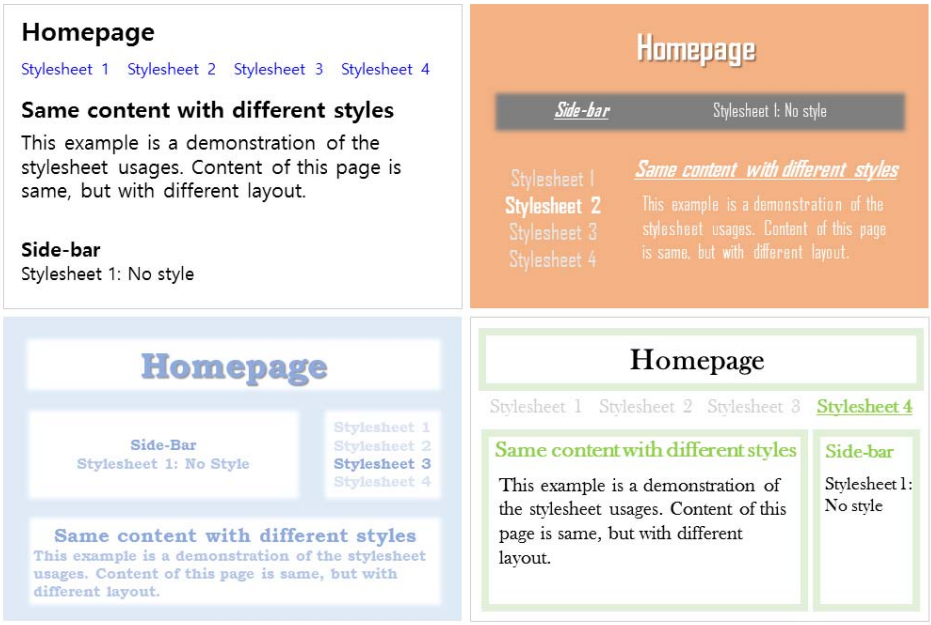


Figure 10. CSS examples

To represent a web page shown in Figure 11 as an example, three methods to manage style of a document can be used: inline, internal, and external. Without CSS or other style sheets, the presentation of a document written in a markup language can be described inline as shown in Figure 12. However, in this case, it is not simple to change the visual structural format since the content and style of the document are intertwined. As in Figure 13, style sheets can be included in the document but in between the head tags. It is easier to make some changes in the visual representation of the document, but it is ineffective to manage the document since the content and visual style are separated in some ways but still contained in the same document. Finally, style sheets can be saved as external files and linked as shown in Figure 14. The last method is effective to manage style sheets since the style components are completely separated from the content. Therefore, it is much simpler and takes fewer efforts to make any changes in styles.

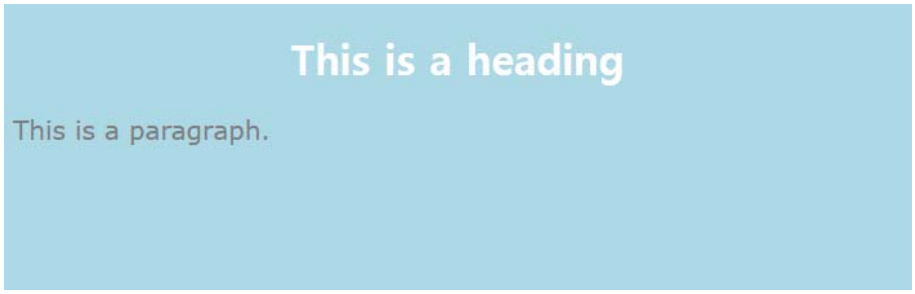


Figure 11. Style sheet example

```
<!-- inline -->

<!DOCTYPE html>
<html>
<body style="background-color:lightblue;">

<h1 style="color:white;text-align:center;">This is a heading</h1>
<p style="color:grey;font-family:verdana;font-size:20px;">This is a paragraph.
</p>

</body>
</html>
```

Figure 12. Inline style example

```
<!-- internal -->

<!DOCTYPE html>
<html>
<head>
<style>
body {
    background-color: lightblue;
}

h1 {
    color: white;
    text-align: center;
}

p {
    color: gray;
    font-family: verdana;
    font-size: 20px;
}
</style>
</head>
<body>

<h1>This is a heading</h1>
<p>This is a paragraph.</p>

</body>
</html>
```

Figure 13. Internal style sheet example

```
<!-- external -->

<!DOCTYPE html>
<html>
<head>
<link rel="StyleSheet" type="text/css" href="CSSexample.css">
</head>
<body>

<h1>This is a heading</h1>
<p>This is a paragraph.</p>

</body>
</html>
```

```
<!-- CSS file -->

body {
    background-color: lightblue;
}

h1 {
    color: white;
    text-align: center;
}

p {
    color: gray;
    font-family: verdana;
    font-size: 20px;
}
```

Figure 14. External style sheet example

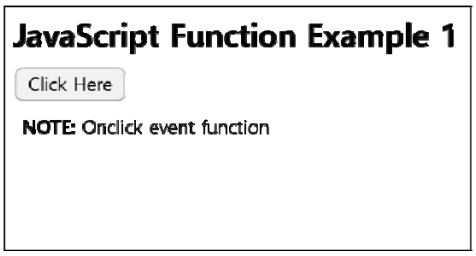
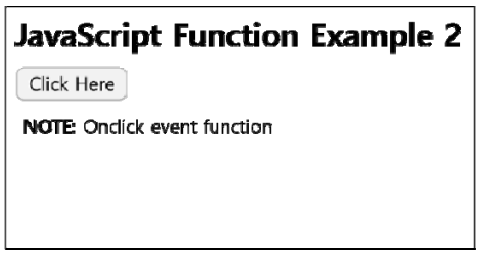
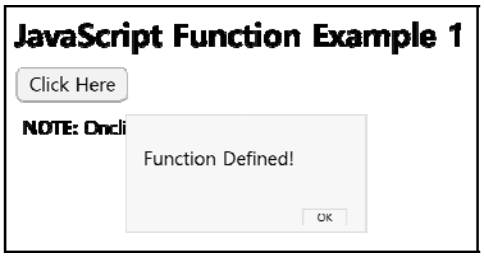
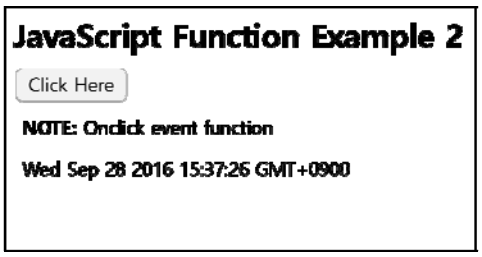
As shown in the example described in the previous paragraph, with CSS, the documents represented in the markup language gets much easier and more effective. To modify visual structure or representation of a document, only the CSS file has to be changed and the main document containing the content does not need to be touched. To share the same style for different contents or on different environment, there are no needs to repetitively and manually define the style again with CSS. These outstanding advantages of CSS, separation of style from content, is one of the necessary aspects of emerging media and data formats. Data itself has been getting bigger and much more complex exponentially, and methods to keep data simple and efficient are needed more than ever so that it is manageable and processable. Therefore, CSS may play a crucial role in the coming more complex and enlarged data era. Both data representation and content are getting intricate, thus to manage them well, the needs of their separation will be magnified and the role of CSS will be also emphasized.

JavaScript

JavaScript is a client-side scripting language designed to adds dynamic and interactive elements on the web pages. JavaScript code is written inside and sent within HTML documents to the browser to execute certain works. It is supported by all modern Web browsers without plug-ins. Instead of reloading a new version of the entire web page, it allows making modifications to the certain content of a page as executed by the browser automatically for a run-time environment or in response to one-by-one event triggered by users. JavaScript facilitates the Web to incorporate information from the user's environment as well, like current location, current time, and etc. As shown in Table 4, with JavaScript the web page modifies its content more interactively without reloading the entire page in response to an event triggered by users, clicking the button in this case.

Table 4. JavaScript examples

	Example 1	Example 2
Script	<!DOCTYPE html> <html lang="en"> <head> <meta charset="utf-8"> <title>JavaScript Function Example 1</title> <script type="text/javascript"> function myFunction(){alert('Function Defined!');} </script> </head> <body> JavaScript Function Example 1 <button type="button" onclick="myFunction();"> Click Here</button> NOTE:Onclick event function </body> </html>	<!DOCTYPE html> <html lang="en"> <body> JavaScript Function Example 2 <button type="button" onclick="document.getElementById('de,p').innerHTML = Date()"> Click Here</button> NOTE:Onclick event function </body> </html>

Before		
After		

Security vulnerability issues of JavaScript, like Cross-Site Scripting or Cross-Site Request Forgery, have been continuously arisen because there is always a possibility that JavaScript with malicious contents may run automatically on client-side applications to take certain unintended actions. However, efforts to prevent these issues have been made by restricting the actions that JavaScript can execute, like reading or writing files, executing other programs, or making any connection to the computer other than for requesting HTML pages and sending emails. There are some people who disable JavaScript for the security reasons, but JavaScript is used in the Web environment so prevalently that disabling it may cause problems to post comments or writings, log in, or etc. Although it was initially used for simple basic task like hovering the mouse over icons or showing alert boxes, in these days JavaScript is implemented as the most fundamental parts of the modern Web to do more complicated actions, like dynamically loading images as scroll or showing popups, previews, or thumbnails, thus without JavaScript there are too many restrictions for many websites to work properly.

JavaScript enables interactive and dynamic Web environment while saving bandwidth usage and server strength usage but enhancing client user experiences. Client-side scripting can cause security issues or high development costs, but it allows interactive immediate responses to users' actions and fast executions. The most prevalently used client-side scripting language, JavaScript is so integrated into the modern Web environment that it has become one of the most fundamental components of the Web. Therefore, even in the newly approaching media era, JavaScript seems to keep its firm position as an inextricable element of the Web.

Extensible Markup Language (XML)

XML is a markup language to set rules for encoding documents [22]. XML is much like HTML and emerged as a front runner format when XML was first introduced. However, different from HTML, XML focuses on the visual structure of contents, XML is designed to carry data and be self-descriptive. Therefore, XML is independent of both software and hardware to store and transport data [23]. Unlike HTML, XML tags are not predefined, but the author is able to freely define tags as needed; therefore, the authors are capable to well describe the structure of data by themselves. In this point of view, XML is more appropriate to facilitate and draw more synergy of the semantic Web. However, because of freely definable tags, a proper schema is required to semantically perceive the structure and successfully render the document on the Web. As the example document written in XML shown in Figure 15, it is self-descriptive and clear enough to see that it is a note from Vanessa to Laura created on December 10th, 2016 at 2:30 p.m. to remind the meeting on weekends. The tags are defined in a human-readable way with natural language, but machines need

proper schemas or cannot perceive the tags freely defined by the authors as humans do. XML is suitable for both human-authored documents and all kinds of machine-to-machine data transfer, with proper schema, since the content is separated from visualizing structure information and user definable tags allow to describe detail relations among elements.

Figure 15. XML example

Because XML is software and hardware independent, it eases the cost of storing, transporting, exchanging, and sharing data even between incompatible systems. Moreover, interoperability between the old version and a new version of the XML-based applications is supported.

XML Schema

Also referred to as XML Schema Definition (XSD), XML schema is a language for expressing constraints about and structure of XML documents [24]. XSD is used to define the legal building blocks and validate the data structure of an XML document [25]. It defines the elements and attributes types and their constraints like amounts, orders, relationships, and etc. A schema is needed for interchanging data between independent groups of people and when there have to be mutual understandings and agreements with formal descriptions of documents. With a schema defining the structure of data, XML documents become both human-readable and machine-processable. As the example in Figure 16, ‘2016-10-12' can be interpreted as October 12, 2016, or December 10, 2016. However, XML element with a data type, ‘<date type="date">2016-12-10</date>', refers December 10, 2016 without any ambiguity since the format of 'date' is defined as 'YYYY-MM-DD'. Therefore, independent groups of people can interpret the given data in the same way without misunderstanding.

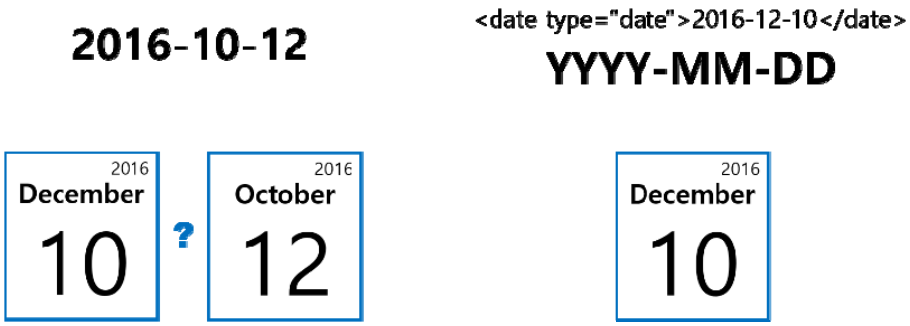


Figure 16. XML element with date data type

Every XML Schema has the <schema> element as the root element and the <schema> element contains some attributes. In Figure 17, the line, 'xmlns:xs="http://www.w3.org/2001/XMLSchema"', indicates the elements and data types used in the given schema are from the specified namespace and prefixed with 'xs'. The next line, 'targetNamespace="http://www.w3schools.com"', also indicates the elements defined in the schema come from the specified URI. The line, 'xmlns="http://www.w3schools.com"', indicates the default namespace URI. Figure 18 describes how to reference 'note.xsd' file in 'note.xml' file. The line 'xmlns="http://www.w3schools.com"' indicates declaring that all the elements used in the XML document are declared in the given namespace. XML Schema Instance namespace and schema location attribute are declared in this example as well. By defining namespaces, XML prevent conflictions between elements with the same name.

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xml:base="http://www.animals.fake/animals#">

<rdf:Description rdf:ID="animal">
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
</rdf:Description>

<rdf:Description rdf:ID="dog">
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  <rdfs:subClassOf rdf:resource="#animal"/>
</rdf:Description>

</rdf:RDF>
```

Figure 17. XML Schema example Note.xsd

```
<?xml version="1.0"?>

<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xml:base="http://www.animals.fake/animals#">

<rdfs:Class rdf:ID="animal" />

<rdfs:Class rdf:ID="dog">
  <rdfs:subClassOf rdf:resource="#animal"/>
</rdfs:Class>

</rdf:RDF>
```

Figure 18. Example of referencing XSD file in an XML document

XSD supports data types, hierarchies, properties and more information about data itself which allow machines to interpret the meanings and contexts of the data. As shown Figure 17 and Figure 18, by defining that 'dog' is a subclass of 'animal', now machines can understand that 'dog' is included in 'animal'. Otherwise, machines would not be able to interpret the meaning of 'dog' and what the word 'dog' would be. Humans acquire information as growing up and naturally interpret languages, images, audios, and etc. Therefore, people understand the semantic meanings of data so handily that people are unsurprisingly able to perceive 'love' is a type of 'feeling', and 'feeling' falls into 'abstract concept'. However, without relation constraints or expressions, machines cannot know what 'love' is and what kinds of properties that the word 'love' have.

Defining rules for encoding documents, XML carries self-descriptive data. It improves interoperability of data since it is independent of software and hardware. By setting formal descriptions of data to establish mutual understandings and agreements between independent groups of people, XML Schema facilitates the machine-readability and exchangeability of data. As XML and XSD support data types, hierarchies, relationships, and more structural information about data itself which allow machines to interpret the meaning and contexts of the data. However, since XML and XSD use non-predefined tags and focuses on the structural aspects of data, they still leave some ambiguities in properties, hierarchies and other relations between data. RDF is often used to prevent the ambiguities caused by the usage of free tags of XML and clarify the meaning of the data.

JavaScript Object Notation (JSON)

JSON is a lightweight data-interchange format designed to transmit data objects in a form of attribute-value pair or ordered list of values as shown in Figure 19 and Figure 20. It is not a language, but a type of data format. It is completely language independent as it is transmitted through HTTP requests in a text format, which makes it an ideal data-interchange language [26]. As its structure is light and easy to parse, JSON is becoming one of the most frequently used data format for asynchronous browser or server communication [27].

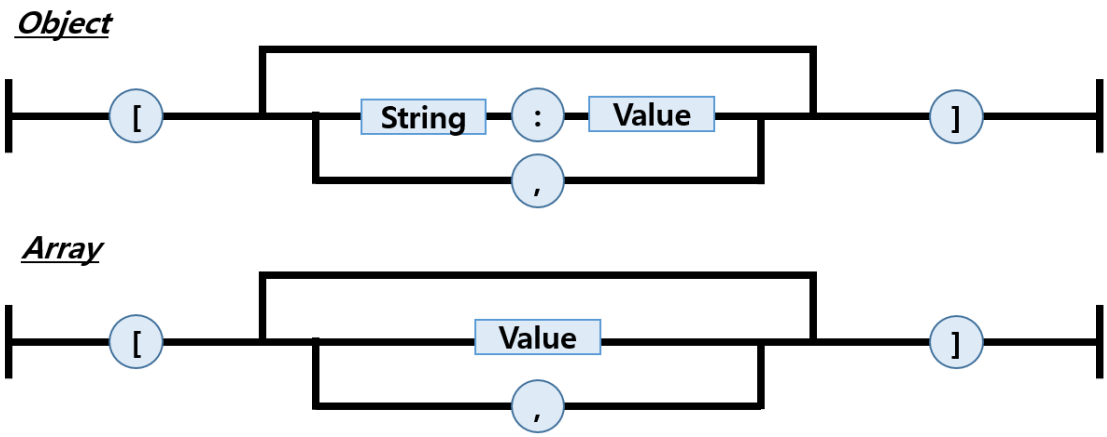


Figure 19. JSON basic forms of transmitted data

```
{
    "firstName": "Laura",
    "lastName": "Smith",
    "age": 25,
    "address":{
        "street": "855 Peachtree Street",
        "city": "Atlanta",
        "state": "GA",
        "postalCode": "30308"
    },
    "phoneNumber":[
        {
        "type": "mobile",
        "number": "678-515-6789"
        },
        {
        "type": "home",
        "number": "801-833-3852"
        }
    ]
}
```

Figure 20. JSON example script

JSON was not designed to do what XML or other markup languages do. It is designed to structure the data in object oriented way, so JSON is more compact compared to XML, but it is designed to link data. Therefore, JSON is often used for transmitting data between websites and in browsers rather than for semantically linking and referencing resources on the Web.

JSON Schema

JSON schema defines and describes the structure of JSON data. The Figure 21 is a JSON schema document that describes the structural constraints of the JSON data described in the Figure 20. It is used to restrict the data type, set the minimum or the maximum, insert descriptions, define required properties or etc. However, JSON Schema cannot define any functional or relational limits, but it can only constrain the structure of JSON data as shown.

```
{
    "title": "Personal Information",
    "type": "object",
    "properties":{
        "firstName":{"type": "string"},
        "lastName": {     "type": "string"},
        "age": {
                "description": "in years",
                "type": "integer",
                "minimum": 0,
        },
        "address":{
                "street":{"type": "string"},
                "city":{"type": "string"},
                "state":{"type": "string"},
                "postalCode":{"type": "integer"}
        },
        "phoneNumber":{
                "type": "array",
                "items":{
                        "properties":{
                                "type":{"type": "string"},
                                "number":{"type": "string"},
                        }
                }
        }
    },
    "required": ["firstName", "lastName", "age"]
}
```

Figure 21. JSON Schema example script

Resource Descriptive Framework (RDF)

RDF is a standard model designed as a metadata model for data interchange on the Web and is based on the idea of making statements about resources in the form of triples: subject-predicate-object or entity-attribute-value structured basic building blocks [28]. The entity denotes a resource and the attribute denotes relationships between the entity and the value. As described in Figure 22, the subject is an address of a web page, and the object is the title of the given web address. The predicate represents the relations between the subject and the object as a resource and its title. Therefore, the subject is a web page, the object is a string variable, and their relationship indicates that the string variable is the title of the page. RDF identifies resources with URIs, thus URIs are the basis of the RDF graphs. Semantically intertwined triples compose the labelled directed graphs as the bases of the Semantic Web structure. Resources are connected and mapped on the graphs via their URIs and properties.

Subject	Predicate	Object
http://www.w3.org/	http://purl.org/dc/elements/1.1/title	"World Wide Web Consortium"

```
<?xml version="1.0"?>
<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:dc="http://purl.org/dc/elements/1.1/">

        <rdf:Description rdf:about="http://www.w3.org/">
                <dc:title>World Wide Web Consortium</dc:title>
        </rdf:Description>

</rdf:RDF>
```

Figure 22. RDF triples and graph data model example written in XML

The goal of RDF is to define a mechanism for describing resources that does not make assumptions about a particular application domain [29]. As it is developed to be interpreted by machines rather than humans, it focuses more on linking and connecting data semantically in machine-interpretable ways than displaying or managing visual structure of data. Since it is designed for the expression of arbitrary information about arbitrary things, it supports the evolution of data schemas over time to facilitates data interoperability even with different underlying schemas [30]. Each triple composes a labeled directed graph and they are semantically intertwined to build connected graph maps over the semantic web as shown in Figure 23, a simple example representing the basic connection structure.

There always have existed and will be multiple leading data formats, instead of simply just one. Hence RDF's ability to semantically merge data with different underlying schemas is essential. In this complex data era, new types of data and more data formats will be created. Therefore, importance in converging data with different data formats will rise in the future as well to bring a significant synergy and magnify the potential in data.

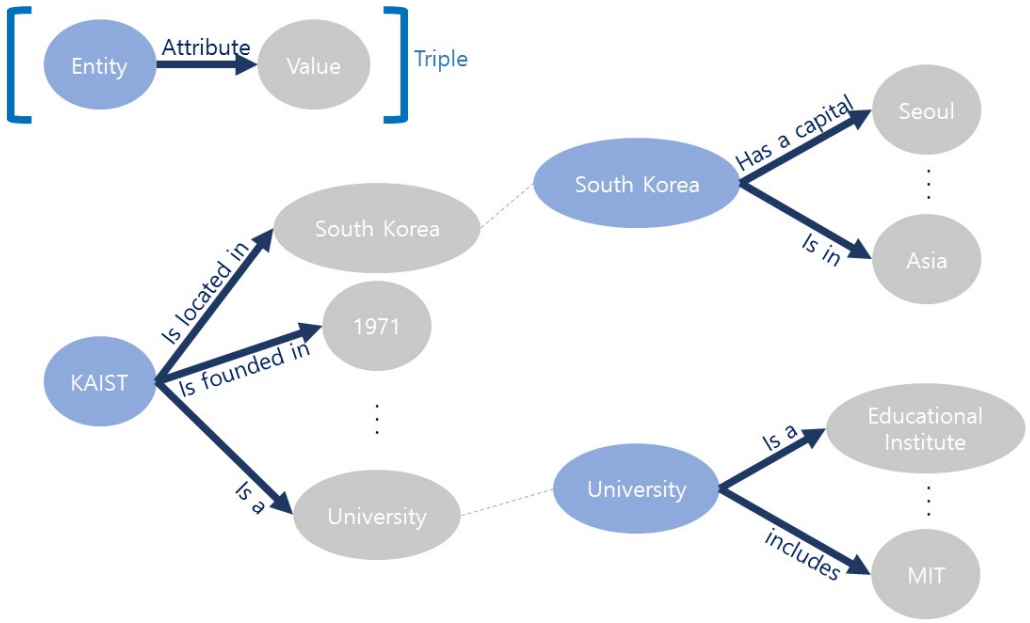


Figure 23. RDF in the semantic web structure

RDF Schema (RDFS)

Contrast to XML Schema that limits the structural constraints prescriptively, RDFS descriptively provides facts and base inferences. RDFS provides the framework for specific classes and properties. Moreover, it defines resources as instances of classes or subclasses of the classes, and by defining vocabulary shared between different applications, it eases the needs of data interchangeability and interoperability between them [31]. As an example shown in Figure 24, if it is stated that South Korea has a capital Seoul, it can be inferred that Seoul is not only a capital but also a city because capital is a subclass of city. By following the labeled directed graph, it can also be inferred that South Korea is a country as a geographic entity. Moreover, the property itself can be an instance and its properties as well can be described. The domain of having a capital should be restricted by country and the range should only contain all the existing capitals.

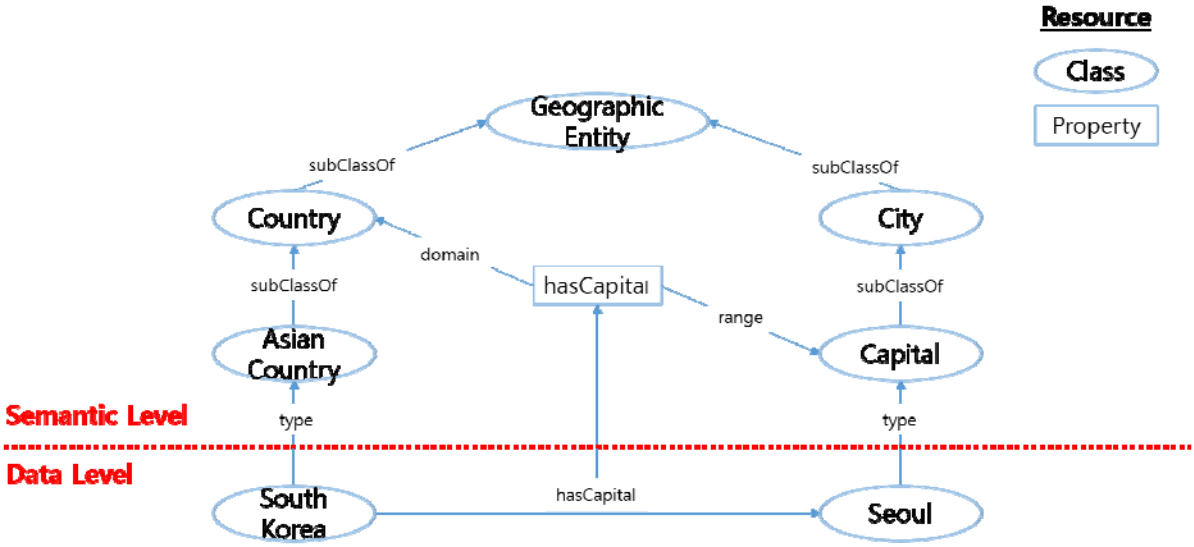


Figure 24. RDFS semantic structure

In another example shown in Figure 25 and Figure 26, if it is stated that Anna has a mother Lauren and mother is a subclass of female, then it can be automatically inferred that Lauren is female. By following the labeled directed graph, it can also be inferred that Lauren is in the classes of parent and person as well.

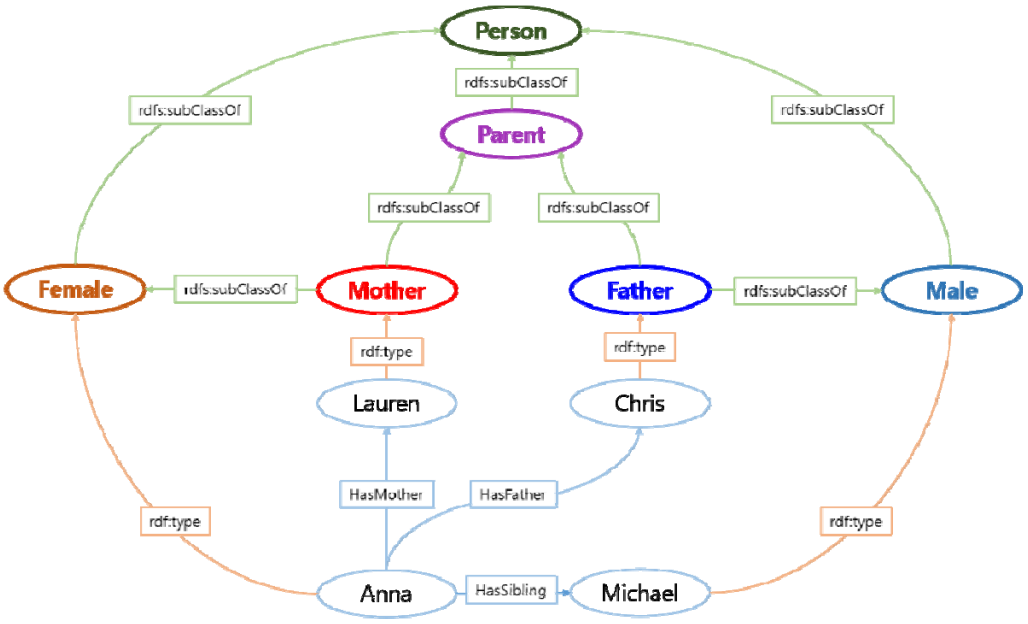


Figure 25. RDFS example class labelled directed graph

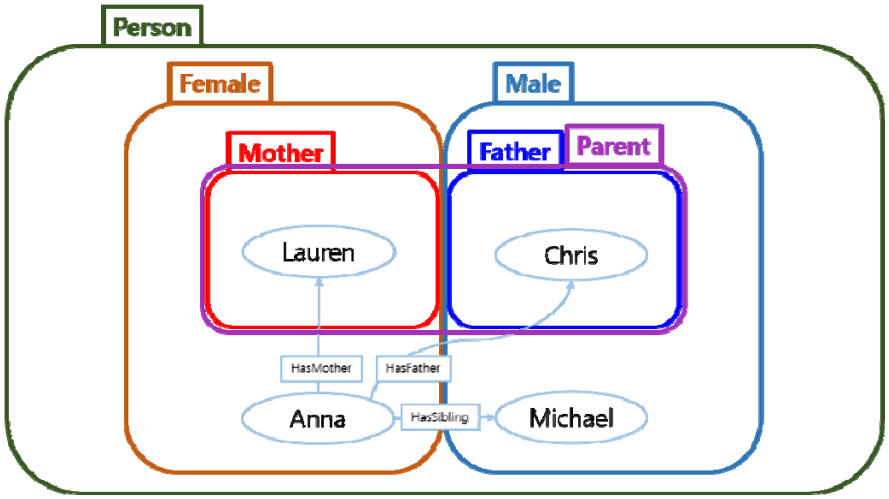


Figure 26. RDFS example class map

RDF and RDFS grammatically so liberal that it allows the data be described for arbitrary purpose but it often causes some ambiguities. They do not have a clear distinction between instance and class. As an example shown in Table 5, in one document, elements can be used both as instances and classes. Moreover, not only instances and classes but also properties can have properties of themselves as shown in the previous examples. Therefore, RDFS does not have clear boundaries between instances, classes, and properties.

Table 5. RDFS instance and class

Example	Instance	Class
<AsianCountry, type, Country>	AsianCountry	Country
<EastAsianCountry, type, AsianCountry>	EastAsianCountry	AsianCountry
<SouthKorea, type, EastAsianCounty>	SouthKorea	EastAsianCountry

Another weakness of RDFS is its limitations in expressions. Using triple structure forms, RDFS can represent only one directional binary relationship between a subject and an object, but not any multi-directional properties, like complimentary, transitive, inverse, symmetric or etc. For example, by stating that Anna has a mother Lauren, humans can be easily deduced that it expresses the same meaning that Lauren has a child Anna, but RDFS cannot express the property that having a mother is the inverse of having a child or having a sibling is a complimentary of having a brother. The statement Anna has a sibling Michael also means the same thing as Michael has a sibling Anna, but RDFS again cannot express these properties are symmetric of each other. Moreover, other than stating all of the relationships one by one, it is not possible to state that every instance of person must have one biological mother who is also a person nor that the range of person having a child must be person as well.

RDF and RDFS are indispensable elements of the Web to merge and semantically link the diversely dispersed data and resources on the Web environment. They facilitate linked data of the Web by converging resources with different underlying schemas. Supporting semantically linked data, they aim not just simply to merge the data, but to define the meanings of resources and connect them. However, even though RDF and RDFS designed for semantically linked Web data environment,

there still are some ambiguities that the current structure of them could not resolve. In this ubiquitous data era, managing data in machine-readable ways is becoming inevitable and crucial more than ever. Accordingly, RDF and RDFS seem to play important roles in emerging data era, but they still have some space to be improved for them to do more and maximize their potential.

Summary and Conclusion

The Web is initially developed for data exchanges and focused on their visual representations, but how the data is represented and structured also have been evolved as technologies are advanced and the needs are risen for more complicated and advanced methods to manage data—store, exchange, search, parse, and more.

As the root of the Web, HTML was designed for visual structural representation of data in the beginning. With HTML, CSS facilitates the separation of the style from the content, and JavaScript adds dynamics and interactivity, but they are not enough to complement the limitation of HTML, lacking the abilities representing the meanings, properties, hierarchies of data beyond simply linking resources. HTML can only use predefined tags and cannot represent the structure of data in meaningful ways. Therefore, HTML has no capability to assert the meanings or relations of the information or the context of data which it contains.

As entering to the Web of linked data or Semantic Web era, XML was then developed to compensate this type of capability shortages of existing languages. XML constraints structural formats or types of data. It is independent of both software and hardware to store and transport data so that interoperability and exchangeability of data are improved, separates content from styles so that the reusability of the data increases, and uses non-predefined tags so that the structure of data can be freely and well described. However, as XML uses non-predefined tags, it still leaves some ambiguities in describing properties, hierarchies and other relations between data for machines to fully interpret meanings. RDF is often used to clarify the ambiguities caused by the usage of free tags of XML, but inversely RDF itself does not provide any constraint on structures or data types as XML does. RDF rather focuses on representing the meaning of the data in relations. RDF expresses resources as metadata and provides standard methods to represent them. With RDF, machines can perceive, process, and manage the data by merging and representing dispersed data on the Web together by connecting them according to their relationships. Often considered as a possible alternative or a competitor of XML, JSON is a compact object-oriented data structure frequently used for transmitting data back and forth between websites. However, while XML and XSD are a type of markup languages and schemas focused on expressing the properties of data resources, JSON is a data format type to efficiently store and transport data with low overheads. Therefore, their purposes emphasize different points of view on data storing and transporting structure, and it is difficult to make a direct comparison between them.

As data has become ubiquitous, data is ceaselessly getting more complicated and wildly overwhelming. Hence the issues to intelligently manage the data on the Web is coming into the spotlight. Moreover, the contents of the Web resources of the same URI may differ over time while the defined properties stay as permanent concepts, thus to be precise and consistent, the information about the resources and data itself also have to be updated accordingly. However, the current structure of the Web is yet well adequate for the newly emerging data era, and the needs of better semantic components and structures for context-aware data managing by machines are growing.

To compensate the structural weak points of the current Web, markup languages and data formats are on the rise of the trend to describe the context of data and information and link them together according to their current relations, properties, and etc. For better context-aware data managing in the future, the roles of markup languages, data formats, and their metadata discussed in this section seem to be crucial for expressing meanings and contexts of data in machine-processable and

integrable ways. However, beyond describing relationships or properties among resources, managing detail-content-based information seems critical as well.

The structure of the modern bible has some desirable features for inserting metadata to data. As the modern bible is divided into book names, chapters, and verses, one can easily use the form of "Book Chapter:Verses" to represent specific phrases, sentences or paragraphs from the bible. For example, John 3:16 represents the exact sentence 'For God so loved the world that he gave his one and only Son, that whoever believes in him shall not perish but have eternal life'. It is similar how URI works on the Web. The book name and Chapter 'John 3' can be considered as a URL and a URN of the described sentence, and people use this form of identifier to annotates, cites, references, shares, add descriptions, and do more. In this way, people can access the same contents of the specific parts of the bible without confusion. The difference between the bible and the Web, in this case, is that the resources in the bible are divided into many small pieces, verses, which allow to point the exactly wanted data out. As the contents of the data are also divided and described, it is efficient to search, parse, manage the content information of the Bible in detail. For example, if one wants to make an annotation saying 'I think it is the central theme of the Bible!' to the sentence, 'For God so loved the world that he gave his one and only Son, that whoever believes in him shall not perish but have eternal life', he or she can make that note and reference 'John 3:16'. In this way, the person can not only easily perceive that the annotation is made for the exact sentence whenever he or she revisited the note but also precisely share the content and the descriptive information together with other people.

However, how metadata are defined and used in the Web environment in these days is somewhat different from the structure of the bible as described in the previous paragraph. The metadata used in the Web focuses more on the just linking and connecting the resources based on their relationships and properties. Moreover, unlike the previous bible example, the resources on the Web are grouped as big chunks rather than divided small pieces is, and their descriptions or metadata are often generalized for overall contents. For example, if a user uploads a recorded surveillance video, the structure of the Web focuses more on what this video is mainly about, who uploaded it, where it is located, and other general information rather than what events are occurring in this video, when exactly they happen, and other detailed content based information. If an analogy is made about how the content of the bible can be described in the structure used for the Web environment, it would look like the following example. The bible has a book of John which is a part of New Testament and is written by John who was one of 12 apostles of Jesus, and so on. In this book of John, the author stated that one can earn an eternal life by believing in God. It would be extremely difficult to find the specific sentences that correspond the description given above.

Linking is important but context aware is also vital in the same manner. To facilitate context-aware data managing by machines for approaching data era, the needs of improved metadata managing methods for detail context-awareness seems to be inevitable. However, the influence and power of the mainstream formats for the linked data on the Web, especially XML and RDF, will not be easily ceased but continued in the future due in no small part of their interoperability and compatibility.

5.4 Geospatial Data

5.4.1 Introduction

Geospatial data, also known as spatial data or geolocation data, is data that can be mapped to a particular location in geographic coordinate systems. Examples include satellite imagery, latitude and longitude coordinates, and addresses.

Geospatial data is typically divided into two types: raster data and vector data. Esri, the company behind the popular geospatial information system ArcGIS (which captures, stores, verifies, visualizes, and analyzes geospatial data), gives the following definitions: [32].

- Raster data models: A representation of the world as a surface divided into a regular grid of cells. Raster models are useful for storing data that varies continuously, as in an aerial photograph, a satellite image, a surface of chemical concentrations, or an elevation surface [33].
- Vector data models: A representation of the world using points, lines, and polygons. Vector models are useful for storing data that has discrete boundaries, like country borders, land parcels, and streets [34].

In The GIS Primer: An Introduction to Geographic Information Systems, David J. Buckley summarizes the advantages and disadvantages of the two types of formats as follows: [35]

Vector Data Advantages and Disadvantages

Advantages:

- Data can be represented at its original resolution and form without generalization.
- Graphic output is usually more aesthetically pleasing (traditional cartographic representation).
- Since most data, e.g. hard copy maps, is in vector form no data conversion is required.
- Accurate geographic location of data is maintained.
- Allows for efficient encoding of topology, and as a result more efficient operations that require topological information, e.g. proximity, network analysis.

Disadvantages:

- The location of each vertex needs to be stored explicitly.
- For effective analysis, vector data must be converted into a topological structure. This is often processing intensive and usually requires extensive data cleaning. As well, topology is static, and any updating or editing of the vector data requires re-building of the topology.
- Algorithms for manipulative and analysis functions are complex and may be processing intensive. Often, this inherently limits the functionality for large data sets, e.g. a large number of features.
- Continuous data, like elevation data, is not effectively represented in vector form. Usually substantial data generalization or interpolation is required for these data layers.
- Spatial analysis and filtering within polygons is impossible.

Raster Data Advantages and Disadvantages

Advantages:

- The geographic location of each cell is implied by its position in the cell matrix. Accordingly, other than an origin point, e.g. bottom left corner, no geographic coordinates are stored.
- Due to the nature of the data storage technique data analysis is usually easy to program and quick to perform.
- The inherent nature of raster maps, e.g. one attribute maps, is ideally suited for mathematical modeling and quantitative analysis.
- Discrete data, e.g. forestry stands, is accommodated equally well as continuous data, e.g. elevation data, and facilitates the integration of the two data types.
- Grid-cell systems are compatible with raster-based output devices, e.g., electrostatic plotters, graphic terminals.

Disadvantages:

- The cell size determines the resolution at which the data is represented.
- It is especially difficult to adequately represent linear features depending on the cell resolution. Accordingly, network linkages are difficult to establish.
- Processing of associated attribute data may be cumbersome if large amounts of data exists. Raster maps inherently reflect only one attribute or characteristic for an area.
- Since most input data is in vector form, data must undergo vector-to-raster conversion. Besides increased processing requirements this may introduce data integrity concerns due to generalization and choice of inappropriate cell size.
- Most output maps from grid-cell systems do not conform to high-quality cartographic needs.

5.4.2 Geolocation Data

Geolocation data is an important subset of geospatial data. Geolocation data is information about the location of a particular device like a smartphone. This data is often combined with other types of data to help understanding what is happening at that location at that particular time. As the number of smart devices increases to an estimated 21 billion devices in 2020, the volume and importance of geolocation data will only continue to increase [36].

Data Formats

Geospatial data is created, shared, and stored in many different formats. The following list names and briefly describes many common formats in alphabetic order. The descriptions are adapted from lists released by NCSU Libraries [37] and Geo-community [38], the premier portal for geospatial technology professionals.

Vector Data File Formats

- (**Arc Export**): Arc Export is a transfer format, either ASCII or compressed into binary used to transfer files between different versions of ARC/INFO. It is undocumented and will work only with ESRI products.
- (**ARC/INFO Coverages**): An ARC/INFO "coverage" is a set of internal binary files used by ARC/INFO, a GIS program. This file format is proprietary and not readily usable by other programs. An ArcInfo coverage does not have an individual file extension. Instead it is composed of two folders within a "workspace" which each contain multiple files. One of the two folders carries the name of the coverage, and contains a number of various .adf files. The other folder is an "info" folder, which typically contains .dat and .nit files for all the coverage and grids in the workspace.
- (**AutoCAD" Drawing Files**): DWG is the internal, proprietary format used in AutoCAD® software, which is a computer-aided design/drafting (CAD) program. Despite its proprietary nature, AutoCAD can convert any DWG file to a DXF file (described below) without loss of graphic information. As with DXF files, there are a number of ways to store attribute information in DWG files. The emerging standard is one that uses Extended Entity Data (EED) to link attributes, but many others are possible. However, the lack of one standard for linking attributes can cause problems when data is transferred between systems.
- (**Autodesk's Data Interchange File**): DXF is probably the most widely used vector data transfer format, and a file in DXF format offers some strong advantages. It contains complete

display information, and almost every graphics program can read it. However, there are several different ways to store attribute information in DXF and to link DXF entities to external attributes. Because there are no attribute standards, many programs that claim to read DXF files still do not import attribute information properly.

- (**Digital Line Graphs**): DLG, a transfer format used by the US Geological Survey (USGS), depicts vector information portrayed on printed paper maps. It carries accurate coordinate information and sophisticated feature-classification information but no other attribute data. DLG does not include any display information. The DLG standard is significant because the USGS and other US government agencies have used it to publish large numbers of digital maps.
- (**E00: Arc Export or Interchange Format**): .e00 files are ArcInfo Interchange or export files, used to conveniently copy and move ArcInfo GIS coverages and grids.
- (**Hewlett-Packard Graphic Language**): HPGL is a language that controls computer plotters; it contains display information but no geographic coordinates or attribute data. It is usually not appropriate for the storage or transfer of GIS data.
- (**LYR: Layer File**): A .lyr file is directly readable only by ArcGIS software and other newer software applications. This file does not contain actual geographic data, but rather contains specifications for the presentation of other datasets. Such specifications include color shading, naming, label properties (font, color, placements). Such presentation properties are usually time consuming to create, so a .lyr file allows these settings to be saved and shared. In order to use a .lyr file, you must also have a separate data file with the same prefix name saved in the same file space.
- (**MapInfo" Data Transfer Files**): MIF/MID is a transfer standard used by MapInfo, a desktop mapping system. It carries all three types of GIS information: geographic, attribute, and display. Attribute links are implicit in the file format.
- (**Micro Station Design Files**): DGN is the internal format used by Bentley Systems Inc.'s Micro Station, a CAD program. It is well documented and standardized, so it may also be used as a transfer standard. DGN files contain detailed display information. The most common way to store attributes is to place them in an external database file and record links in the MSLINK field-a data item carried for each element in the DGN file.
- (**SHP: Shape file**): The ESRI Shape file has become an industry standard geospatial data format, and is compatible to some extent with practically all recently released GIS software. To have a complete shape file, you must have at least 3 files with the same prefix name and with the following extensions: .shp = shapefile, .shx = header and .dbf = associated database file.
- (**Smart Data Compression**): SDC is ESRI's highly compressed format, which is directly readable by ArcGIS software, but not by ArcView 3.x. Many ESRI Data and Maps datasets are natively in SDC format.
- (**Spatial Data Transfer System**): SDTS, a new transfer format developed by the US government, was designed to handle all types of geographic data. Virtually all geographic concepts can be encoded in SDTS, including coordinate information, complex attribute information, and display information. This versatility causes a corresponding increase in complexity. To simplify things, several standard subsets of SDTS have been adopted.
- (**Topologically Integrated Geographic Encoding and Referencing Files**): TIGER is an ASCII transfer format used by the US Census Bureau to store the street maps constructed for the 1990 census. It contains complete geographic coordinates and is line, not polygon, based (although polygons can be constructed from its attribute information). The most important

attributes include street name and address information. TIGER does not contain display information. Maps of the entire US are available in TIGER format.

- (**Triangular Irregular Network**): A TIN is a vector-based model which represents geographic surfaces as contiguous non-overlapping triangles. The vertices of each triangle are known data points (x, y) with values in the third dimension (z) taken from surveys, topographic maps, or digital elevations models (DEMs). The surface of each triangle has a slope, aspect, surface area, and continuous, interpolated elevation values. The selective inclusion of points within a TIN gives the triangles their irregular pattern and reduces the amount of data storage required relative to the regularly distributed points in a DEM.
- (**Vector Product Format**): VPF is a binary format used by the US Defense Mapping Agency. It is well documented and can be used as an internal format and as a transfer format. It carries geographic and attribute information but no display data. VPF files are sometimes referred to as VMAP products. The Digital Chart of the World (DCW) is published in this format.

Raster Data File Formats

- (**ArcInfo Grid**): An ArcInfo Grid does not have an individual file extension. Instead it is composed of two folders within a "workspace" which each contain multiple files. One of the two folders carries the name of the grid, and contains a number of various .adf files. The other folder is an "info" folder, which typically contains .dat and .nit files for all the coverage and grids in the workspace. The best way to manage (copy, move, delete, rename) ArcInfo Grids is with ArcCatalog or ArcInfo Workstation.
- (**Band Interleaved by Line**): Band Interleaved by Pixel (BIP), and Band Sequential (BSQ). BIL, BIP, and BSQ are formats produced by remote-sensing systems. The primary difference among them is the technique used to store brightness values captured simultaneously in each of several colors or spectral bands.
- (**Digital Elevation Model**): DEM is a raster format used by the USGS to record elevation information. Unlike other raster file formats, DEM cells do not represent color brightness values, but rather the elevations of points on the earth’s surface.
- (**ECW**): ECW is a proprietary format of ERMapper for imagery compression. It is a more recent format than MrSID, but is gaining popularity because of free compression utilities available from ER Mapper's website [39].
- (**JPEG 2000**): JPEG 2000 is a non-proprietary image compression format based on ISO standards, and typically uses .jp2 as the file extension. It's advantages are that it offers loss and lossless compression, and world files (.j2w) can be used to geo-reference an image in GIS software. Compression ratios are similar to MrSID and ECW formats.
- (**MrSID**): MrSID is a proprietary format of LizardTech's GeoExpress software for imagery compression, and is commonly used on orthoimages. The MrSID file extension is .sid. A companion file with a .sdw extension and the same prefix name as the .sid is used as a world file for georeferencing a MrSID image. Most greyscale TIFF images are compressed with MrSID to 10:1 or 15:1. Color images are usually compressed to 30:1 or 40:1. GeoExpress is also commonly used to create image mosaics. Most recent GIS software, including ArcGIS, are able to read MrSID compressed images without any additional extensions. ArcView 3.x, however, requires a MrSID Extension for image access. Plugins for other software, like AutoCAD and Photoshop, may or may not be required.
- (**Spatial Data Transfer Standard**): As was indicated under vector formats above, SDTS is a general-purpose format designed to transfer geographic information. One SDTS variant is the

raster profile, designed as a standard format for transferring raster data. However, this protocol has not as yet been finalized.

- (**Tagged Image File Format**): Like PCX, TIFF is a common raster format produced by PC drawing programs and scanners.

5.4.3 Standards

There are many efforts to develop standards that promote interoperability amongst the many geospatial data formats. Below is a list of the names and brief descriptions of some of the more well-known examples. The descriptions are adapted from the organizations that developed the respective standards.

For a more comprehensive list, see the Spatial Data Standards list curated by the Spatial Data Standards Commission of the International Cartographic Association (ICA) [ref-9, ref-10].

IUGS Commission for the Management and Application of Geoscience Information (IUGS-CGI) Standards

CGI's mission is to foster the interoperability and exchange of geoscience information, by active community leadership, collaboration, education, and the development and promotion of geoscience information standards and best practice [42].

- **GeoSciML**
  - GeoSciML is an XML–based data transfer standard for the exchange of digital geoscientific information. It accommodates the representation and description of features typically found on geological maps, as well as being extensible to other geoscience data like drilling, sampling, and analytical data.
  - GeoSciML provides a standard data structure for a suite of common geologic features (e.g., geologic units, structures, earth materials) and artefacts of geological investigations (e.g., boreholes, specimens, measurements). Supporting objects like the geologic timescale and vocabularies are also provided as linked resources, so that they can be used as classifiers for the primary objects in the GeoSciML standard.
  - The GeoSciML data standard is underpinned by several established OGC and ISO standards, including Web Feature Service (WFS – ISO 19142), Geography Markup Language (GML – ISO 19136), Observations & Measurements (O&M – ISO 19156), and SWE Common.
  - A parallel data standard for simple map visualisation, GeoSciML-Portrayal, has also been developed. It enables portrayal of a small simplified subset of the GeoSciML data model using Web Map Services (WMS) or simple Web Feature Services (WFS) [43].

- **EarthResourceML**
  - EarthResourceML is an XML–based data transfer standard for the exchange of digital information for mineral occurrences, mines and mining activity. EarthResourceML describes the geological characteristics and setting of mineral occurrences, their contained commodities, and their mineral resource and reserve endowment. It is also able to describe mines and mining activities, and production of concentrates, refined product, and waste materials.

- EarthResourceML makes use of the existing GeoSciML data standard for describing geological materials associated with mineral deposits. It is also underpinned by established OGC and ISO standards, including Web Feature Service (WFS – ISO 19142), Geography Mark-up Language (GML – ISO 19136), and SWE Common [44].

Federal Geographic Data Committee (FGDC) Standards

The United States Federal Geographic Data Committee (FGDC) is an organized structure of Federal geospatial professionals and constituents that provide executive, managerial, and advisory direction and oversight for geospatial decisions and initiatives across the Federal government.

The FGDC is responsible for developing the National Spatial Data Infrastructure. The tools, policies, standards, and communities that compose the NSDI can be accessed online on the FGDC GeoPlatform [46]. For a full list of all standards developed by or endorsed by the FGDC, see FGDC.gov Geospatial Standards [47].

- **Metadata Standards**
  - The FGDC currently supports multiple metadata standards. Historically, the FGDC recommended using the Content Standard for Digital Geospatial Metadata (CSDGM), a consensus-based standard developed by U.S. Federal Agencies. In 2010, however, the FGDC began recommending the International Standards Organization (ISO) geographic metadata standard (19115), a consensus-based standard developed by the international community, because the ISO metadata standard better supported data sharing across national and cultural boundaries [48].
  - According to the FGDC, these are the core components of geospatial metadata:
    - (**Metadata Record Information**): Information about the metadata record including the language in which the record is written, a unique file identifier for the metadata record, the metadata standard used to organize the record, a point of contact for the metadata record, and the date that the metadata record written.
    - (**Identification Information**): Citation-level information about the data including the title, abstract, purpose for creation, status, keywords (theme and place), and extent (temporal, vertical and horizontal). Constraints information about legal and security limitations to data access and use.
    - (**Data Quality Information**): Information about the processes and sources used to develop the data and positional and/or accuracy assessments performed.
    - (**Maintenance Information**): Information about the scope and frequency of data updates. Spatial representation information about the mechanism used to represent spatial data (grid, point, and vector).
    - (**Reference System Information**): Information about the reference systems used to represent geographic position and time.
    - (**Content Information**): Information about the data set entities and attributes.
    - (**Symbology Information**): Information about the symbols used to represent spatial features.
    - (**Distribution Information**): Information about the data distributors and methods for obtaining the data.
    - (**Metadata Extension Information**): Information about custom, user-based, changes to the elements, domains or conditionality of the standard.
    - (**Application Schema Information**): Information about the schema or data models used to structure the data [49].

ISO Geospatial Metadata Standard

In 1999 the International Organization for Standardization (ISO) Technical Committee (TC) 211 Geographic Information / Geomatics was tasked to harmonize the FGDC Content Standard for Digital Geospatial Metadata (CSDGM) with geospatial metadata standards. The resultant ISO 19115: Geographic information - Metadata standard was finalized in 2003 and endorsed by the FGDC in 2010. A series of additional ISO 191** standards have been developed to update, extend, and supplement the 19115 standard [50].

- ISO 19115-1:2014

- ISO 19115-1:2014 defines the schema required for describing geographic information and services by means of metadata. It provides information about the identification, the extent, the quality, the spatial and temporal aspects, the content, the spatial reference, the portrayal, distribution, and other properties of digital geographic data and services.
- ISO 19115-1:2014 is applicable to:
  - The cataloguing of all types of resources, clearinghouse activities, and the full description of datasets and services.
  - Geographic services, geographic datasets, dataset series, and individual geographic features and feature properties
- ISO 19115-1:2014 defines:
  - Mandatory and conditional metadata sections, metadata entities, and metadata elements.
  - The minimum set of metadata required to serve most metadata applications (data discovery, determining data fitness for use, data access, data transfer, and use of digital data and services).
  - Optional metadata elements to allow for a more extensive standard description of resources, if required.
  - A method for extending metadata to fit specialized needs.
- Though ISO 19115-1:2014 is applicable to digital data and services, its principles can be extended to many other types of resources like maps, charts, and textual documents as well as non-geographic data. Certain conditional metadata elements might not apply to these other forms of data [51].

- ISO 19115-2:2009

- ISO 19115-2:2009 extends the existing geographic metadata standard by defining the schema required for describing imagery and gridded data. It provides information about the properties of the measuring equipment used to acquire the data, the geometry of the measuring process employed by the equipment, and the production process used to digitize the raw data. This extension deals with metadata needed to describe the derivation of geographic information from raw data, including the properties of the measuring system, and the numerical methods and computational procedures used in the derivation. The metadata required to address coverage data in general is addressed sufficiently in the general part of ISO 19115 [52].

- ISO 19115-3:2016

- ISO/TS 19115-3:2016 defines an integrated XML implementation of ISO 19115-1, ISO 19115-2, and concepts from ISO/TS 19139 by defining the following artefacts:

- A set of XML schema required to validate metadata instance documents conforming to conceptual model elements defined in ISO 19115-1, ISO 19115-2, and ISO/TS 19139.
- A set of ISO/IEC 19757-3 (Schematron) rules that implement validation constraints in the ISO 19115-1 and ISO 19115-2 UML models that are not validated by the XML schema.
- An Extensible Stylesheet Language Transformation (XSLT) for transforming ISO 19115-1 metadata encoded using the ISO/TS 19139 XML schema and ISO 19115-2 metadata encoded using the ISO/TS 19139-2 XML schema into an equivalent document that is valid against the XML schema defined in this document.

- ISO/TS 19115-3:2016 describes the procedure used to generate XML schema from ISO geographic information conceptual models related to metadata. The procedure includes creation of an UML model for XML implementation derived from the conceptual UML model.
- This implementation model does not alter the semantics of the target conceptual model, but adds abstract classes that remove dependencies between model packages, tagged values and stereotypes required by the UML to XML transformation software, and refactors the packaging of a few elements into XML namespaces. The XML schema has been generated systematically from the UML model for XML implementation according to the rules defined in ISO/TS 19139 or ISO 19118.

5.4.4 Geospatial Applications

Introduction

Geospatial analysis involves analyzing data that contains geospatial information, typically using geographic information systems and geomatics. This analysis was originally developed to solve problems in the environmental and life science, particularly ecology, geology, and epidemiology, but has since extended to almost all industries including defense, intelligence, utilities, natural resources, social sciences, medicine, public safety, and disaster management.

Typical examples might be combining two or more maps or more layers; identifying regions of a map within a specified distance of one or more features, like towns, roads or rivers; understanding how characteristics like temperature vary across a surface; selecting routes, pipeline, and facility locations; assessing the flow of water or traffic over time; and more.

Geospatial Analysis Tools

Below are names and brief descriptions of some of the most common geospatial analysis software packages. Names and descriptions are adapted from a list [54] curated by the Geospatial Information System Population Sciences Center [55], a collaboration between the Population Research Institute (The Pennsylvania State University) and the Center for Spatially Integrated Social Science (University of California, Santa Barbara). For more comprehensive crowd sourced lists include in Wikipedia [56].

- (**ArcGIS**): An integrated collection of GIS software products that offers a platform for spatial analysis, data management, and mapping. The most widely used GIS software in the world, offers a broad set of tools and applications.

- (**BioMedware**): Provides software for the visualization, analysis, modeling and interactive exploration of spatiotemporal data, detection and analysis of event clusters and detection and analysis of geographic boundaries.
- (**GeoDa**): The GeoDa center offers a number of free and paid software for the purpose of conducting geospatial analysis, geo-visualization, geo-simulation, spatial econometrics, crime analysis, and more. Developed by GISPopSci collaborator Luc Anselin and colleagues.
- (**GRASS**): A free Geographic Information System (GIS) software used for geospatial data management and analysis, image processing, graphics/maps production, spatial modeling, and visualization.
- (**QGIS**): A free, open source GIS software. Similar to other software GIS systems, QGIS allows users to create maps with many layers using different map projections. Maps can be assembled in different formats and for different uses. QGIS allows maps to be composed of raster or vector layers. Typical for this kind of software, the vector data is stored as either point, line, or polygon-feature. Different kinds of raster images are supported, and the software can georeference images.
- (**R**): A free, widely used, open source statistical software that is commonly applied used in spatial analysis.
- (**WinBUGS/ GeoBUGS**): A set of free and open source software that allows for the estimation of models (including spatial models) by means of the Gibbs sampler/Markov Chain Monte Carlo (MCMC) methods.
-

5.5 E-Book Data

5.5.1 Introduction

Merriam-Webster defines e-books as "books composed in or converted to digital format for display on a computer screen or handheld device". These e-books have become more and more popular in recent years [57]. In 2012, the net sales revenue from e-books surpassed the net sales revenue from hardcover books. In 2013, 23% of American adults had read an e-book that year. That number increased to 28% in 2014 [58].

The percentage of American adults who owned tablets or e-readers (devices dedicated to reading e-books) saw an even larger increase, from 30% in 2013 to 50% in 2014. At the same time, the 28% of American adults reading e-books on their phones in 2013 grew to 32% in 2014.

It's worth noting that those who read e-books also read print books: only 4% of readers are "e-book only" [59].

5.5.2 Data Formats

E-books are created, shared, and stored in many different formats. The following list names and briefly describes many common formats in alphabetic order. The descriptions are adapted from a list released by eBook Architects, which provides professional consulting from one of the eBook industry's most recognized experts, as well as quality assurance and diagnostic testing on your eBook files. For more comprehensive crowd sourced lists is mobileread.com [ref-29, ref-30].

EPUB2

The EPUB format was developed as an industry-wide standard for eBooks. It is based on a variety of other technologies and standards, like Open eBook and XHTML 1.1, but its uniqueness is in how

it combines these standards to provide a solid formatting foundation for eBooks of just about every shape and size.

The EPUB standard is maintained by the International Digital Publishing Forum (IDPF), a non-profit organization made up of technology and publishing companies. EBook Architects is an active member in the IDPF.

EPUB2 was first introduced in 2007, and had a minor maintenance update in 2009. It is widely used by a large number of retailers, and is the most common eBook format used on the market. However, EPUB3, the latest version of the EPUB standard, is the format that most retailers are moving toward. It contains more enhanced functionality than EPUB2, as well as a large number of important core updates.

EPUB3

EPUB3 is the latest version of the EPUB format. The EPUB standard is maintained by the International Digital Publishing Forum (IDPF), a non-profit organization made up of technology and publishing companies.

EPUB3 was released in 2011, and has been gaining traction in the marketplace since that time. It is currently officially supported on only a few reading systems, but it can be ingested by most of the eBook retailers.

EPUB3 was updated to include better support for foreign languages, embedded media, and other core features and enhancements. It also has a fixed layout formatting option that is beginning to be adopted for children's eBooks and even for non-fiction titles.

iBooks Author

iBooks Author is a proprietary eBook format created by Apple and intended for complex non-fiction eBooks like textbooks, cookbooks. The iBooks Author format is only able to be read in the iBooks ecosystem, and it can only be created in the iBooks Author program on a Macintosh computer.

iBooks Author files are inherently fixed layout, meaning that the design is static and does not allow the reader to change the font size or other visual settings. However, iBooks Author also has an option for creating a reflowable version of the content that can be accessed by changing the orientation of the device. Normally the fixed layout design is implemented in landscape while the reflowable design is implemented in portrait, but this can be determined by the designer creating the file.

iBooks Author files should not be used for children's eBooks because it does not have support for narration overlays (the Read Aloud function that is available in standard children's fixed layout eBooks for Apple).

Apple calls these files "multi-touch" because they are designed to include interactive features and content, like video, audio, and widgets. The format is also structured in a consistent way for every eBook, with the chapters and each page within those chapters easily accessible.

While iBooks Author files are similar to EPUB files in structure, they are not the same format and are not interchangeable with EPUB files. In addition, iBooks Author widgets are built in Dashcode, which is mainly used to make widgets for on Macintosh computers. While Dashcode widgets can be built in a variety of ways, they will not always be easily usable in other eBook formats or devices.

KF8

Kindle Format 8 (KF8) was released by Amazon in late 2011. It is the successor to the old Mobipocket format, and has been updated to include a variety of new features and functionality. KF8 has support for HTML5 and CSS3, and it also has a built-in fixed layout format that is especially well-suited for children's eBooks.

Mobipocket (MOBI)

Mobipocket is a French company that developed its eBook creation and reading software when eBooks were still in their infancy, and managed to flourish in the nascent market, eventually being purchased by Amazon in 2005. When Amazon decided to develop the Kindle, it was a logical step for the company to use its own proprietary format for the new device's eBooks. The Mobipocket format is based loosely on HTML 3.2 and includes some unique formatting requirements.

NOOK Kids

"ePib" is the unofficial name of the NOOK Kids format. This is a fixed layout format that is only used for children's eBooks in the Barnes & Noble NOOK platform. It is not supported on other devices or platforms, and it is not possible to sell it from your own website.

While the file structure behind an "ePib" file is similar to that of an EPUB file, the NOOK Kids format is dramatically different in many ways. These files are created from PDF files, not HTML and CSS, and the resulting code is not editable or able to be enhanced. NOOK Kids files are built with a proprietary tool that is only available to publishers with a direct account with Barnes & Noble; these files cannot be sold through the NOOK Press self-publishing portal, so you must have a publisher account or use a distributor to sell them.

The NOOK Kids format has a few useful features like audio narration and region magnification, but it does not allow embedded video, page zooming, or other features that are available in other children's eBook formats. For more information, please see our Children's eBooks page.

PDF

PDF is common document format. While most people will use Adobe products like InDesign or Acrobat to generate PDF files, there are a variety of other programs that can create PDF files, including some that install on your computer like a printer allowing you to "print" a PDF from almost any application.

The main problem PDF files have in the modern eBook world is that the text in PDF files cannot reflow to fit small screens. eBook devices like the Kindle and Nook, as well as phones and tablets like the iPhone, iPad, and Kindle Fire, all have screens that are smaller than a typical computer screen or even than a standard print book. As a result, when a PDF file is loaded on such a device it usually needs to be zoomed in to be readable, forcing the user to scroll left and right to be able to read all of the text on a line. This is why most eBooks formats like EPUB and Kindle are reflowable.

In addition to this, PDF files are not sold by any of the standard retailers, so it is difficult to distribute them to consumers unless you are selling them on your own website.

5.5.3 Standards

There are not many efforts to develop standards that promote e-book interoperability. Many of the aforementioned e-book formats are proprietary and thus their details unknowable. The EPUB format, however, has developed as the de facto industry standard, and most major e-book retailers and e-readers (with the notable exception of Amazon and its e-readers) now accept it.

In February 2013, a workshop hosted in New York by the Book Industry Study Group, the International Digital Publishing Forum (the group behind the EPUB format), and the World Wide Web Consortium hosted a workshop called "eBooks: Great Expectations for Web Standards" that discussed the future of e-book standards with a focus on EPUB 3.0 as the standard currently leading the way. This is the standard we will focus on in this section of the report.

International Digital Publishing Forum

The International Digital Publishing Forum (IDPF) is the global trade and standards organization dedicated to the development and promotion of electronic publishing and content consumption.

The work of the IDPF promotes the development of electronic publishing applications and products that will benefit creators of content, makers of reading systems, and consumers. The IDPF develops and maintains the EPUB content publication standard that enables the creation and transport of reflowable digital books and other types of content as digital publications that are interoperable between disparate EPUB-compliant reading devices and applications.

EPUB3.0.1

EPUB 3.0.1 is the current version of the EPUB standard. It is a minor maintenance update to EPUB 3.0. On June 26, 2014 the IDPF announced that the IDPF membership had overwhelmingly approved elevation of EPUB 3.0.1 to final Recommended Specification status.

The full, detailed specifications of the EPUB format can be found idpf.org [62]. Here are the main features of EPUB3.0:

Table 6. EPUB 3.0 main features

Main Features & Functions	Description
Multimedia representation	Rich media representation, like Audio & Video, using HTML5
Script support	Implement dynamic events (graph, input) using JavaScript (EcmaScript)
Metadata	Provides Dublin Core-based eBook metadata (identifiers, titles, language elements)
MathML acceptation	Expressions in text form, not graphic or image form
CSS3 support	More finer adjustments for row assignments, hyphenation
Multi style sheet	Dynamic horizontal writing and vertical writing available
OTF&WOFF	Allows fonts that are not installed on the user's system to be displayed inside the EPUB file
SVG support	Not only SVG file itself, but inline vector graphic representation inside content

Here are the main components of the EPUB3.0 standard:

Table 7. EPUB 3.0 components

Standard	Descriptions
EPUB3 Overview	Description of EPUB 3.0 Overview
EPUB Publications3.0	Define the semantics and hierarchical conformity requirements of the publication level
EPUB Content Documents 3.0	Define profiles like XHTML (HTML5), SVG, and CSS using for publishing content
EPUB Open Container Format 3.0	Processing format and file format definition for encapsulating data in the form of a single file (ZIP) for publication
EPUB Media Overlays 3.0	Define processing method and formats for text and audio synchronization
EPUB Canonical Fragment Identifier (epubcfi)	Definition of how to refer to various materials within EPUB publications through the use of object identifiers

EPUB Package Document

The Package Document is an XML document that consists of a set of elements that each encapsulate information about a particular aspect of the EPUB Package.

These elements serve to centralize metadata, detail the individual resources that compose the Package and provide the reading order and other information necessary to render the Rendition.

- (**Metadata**): Mechanisms to include and/or reference metadata applicable to the given Rendition of the EPUB Publication.
- (**Manifest**): Identifies (via IRI) and describes (via MIME media type) the set of resources that collectively compose the given Rendition.
- (**A spine**): An ordered sequence of id references to top-level resources in the manifest from which all other resources in the set can be reached or utilized.
- (**Collections**): A method of encapsulating and identifying subcomponents within the Package.
- (**Manifest fallback chains**): A mechanism that defines an ordered list of top-level resources as content equivalents.

When an element defined in this section has mandatory text content, content is referred to as the value of the element in the explanatory descriptions [63].

5.5.4 E-readers

There are many devices that exist to make reading e-books easier. Here is a non-exhaustive list of popular hardware and software products.

For a more comprehensive, crowd sourced list, see Wikipedia.org e-book_readers [64].

Hardware Products

- Amazon Kindle
- Android (WordPlayer, FBReader, Aldiko)
- Barnes & Noble Nook
- iRiver Story & iRiver Cover Story
- Sharp Zaurus, Nokia 770, n800, n 810
- Sony Reader
- iPad, iPhone, iPod Touch (Lexcycle Stanza, Glider, iFlow Reader, iBooks)

Software Products

Table 8. E-reader Software Products

e-reader	Operating System
Adobe Digital Editions	OS
Aldiko	Android
BookGlutton	Web
e-book Reader	Opera widget
EPUBReader	Firefox add-on
FBReader	Windows, Linux, PDAs
Google Books	Web
iBooks	iOS
Okular	Linux
WordPlayer	Android
Talking Clipboard	Windows
URead	Windows

5.6 Logistics RFID/USN Data

5.6.1 Introduction

The RFID journal defines radio frequency identification, or RFID, as a generic term for technologies that use radio waves to automatically identify people or objects. The most common identification method involves storing a serial number on a microchip attached to an antenna (the chip and the antenna together are called an RFID transponder or an RFID tag) which lets the chip transmit the identification information to a reader which converts the radio waves into digital information that is then passed to computers that can make use of it [65].

The ITU defines Ubiquitous Sensor Networks (USN) as networks of intelligent sensors that could, one day, become ubiquitous: available "anywhere" (i.e., anywhere that it is useful and economically viable to expect to find a sensor) [66].

RFID USNs could be used to solve many problems. For example, they might help supply chain managers track goods. This application of radio frequency identification (including RFID tags with sensors) corresponds to the lower layers in the schematic model for USN as follows: [67].

- (**RFID Tags**): An RFID processor that may be either passive or active (with potentially read/write functions, wider communication ranges and independent power supplies). An active RFID chip is capable of two-way communication whereas a passive tag is read-only.
- (**RFID Reader**): The reader senses and "reads" the information on the tag and passes it on for analysis.
- (**RFID Middleware**): Like the USN, the RFID network may have its own software for the collection and processing of data.

5.6.2 Data Formats

Electronic Product Codes

The Electronic Product Code™ (EPC) is syntax for unique identifiers assigned to physical objects, unit loads, locations, or other identifiable entity playing a role in business operations.

EPCs have multiple representations, including binary forms suitable for use on Radio Frequency Identification (RFID) tags, and text forms suitable for data sharing among enterprise information systems.

GS1's EPC Tag Data Standard (TDS) specifies the data format of the EPC, and provides encodings for numbering schemes -- including the GS1 Keys -- within an EPC.

When unique EPCs are encoded onto individual RFID tags, radio waves can be used to capture the unique identifiers at extremely high rates and at distances well in excess of 10 metrics, without line-of-sight contact.

Radio Frequency Bands

RFID uses the following radio frequency bands:

Table 9. RFID radio frequency bands

Band	Regulations	Range	Data speed	ISO/IEC 18000 Section	Remarks	Approximate tag cost In volume (2006) US$
120–150 kHz (LF)	Unregulated	10 cm	Low	Part 2	Animal identification, factory data collection	$1
13.56 MHz (HF)	ISM band worldwide	10 cm–1 m	Low to moderate	Part 3	Smart cards (ISO/IEC 15693, ISO/IEC 14443 A,B). Non fully ISO compatible memory cards (Mifare Classic, iCLASS, Legic, Felica ...).	$0.50 to $5

					Microprocessor ISO compatible cards (Desfire EV1, Seos)	
433 MHz (UHF)	Short Range Devices	1–100 m	Moderate	Part 7	Defense applications, with active tags	$5
865-868 MHz (Europe) 902-928 MHz (North America) UHF	ISM band	1–12 m	Moderate to high	Part 6	EAN, various standards	$0.15 (passive tags)
2450-5800 MHz (microwave)	ISM band	1–2 m	High	Part 4	802.11 WLAN, Bluetooth standards	$25 (active tags)
3.1–10 GHz (microwave)	Ultra wide band	up to 200 m	High	Not Defined	requires semi-active or active tags	$5 projected

5.6.3 Standards

There are many efforts to develop standards that promote interoperability amongst the many RFID/USN technologies. Below is a list of the names and brief descriptions of some of the more well-known examples. The descriptions are adapted from the organizations that developed the respective standards.

For a more comprehensive list, see the crowd sourced list RFID journal [69]. For more detail on many of the standards described www.gs1.org [70].

5.6.3.1 GS1 Standards

GS1 is an international non-profit organization that manages international supply chain standards [71].

Tag Data Standard

GS1's EPC Tag Data Standard (TDS) defines the Electronic Product Code (EPC), including its correspondence to GS1 keys and other existing codes. TDS also specifies data that is carried on Gen 2 RFID tags, including the EPC, User Memory data, control information, and tag manufacture information [72].

Tag Data Translation Standard

This EPC Tag Data Translation standard is concerned with a machine-readable version of the EPC Tag Data Standards specification. The machine-readable version can be readily used for validating EPC formats as well as translating between the different levels of representation in a consistent way. This specification describes how to interpret the machine-readable version. It contains details of the structure and elements of the machine-readable markup files and provides guidance on how it might be used in automatic translation or validation software, whether standalone or embedded in other systems [73].

UHF Gen2 Air Interface Protocol

GS1's EPC "Gen2" air interface protocol, first published by EPCglobal in 2004, defines the physical and logical requirements for an RFID system of interrogators and passive tags, operating in the 860 MHz - 960 MHz UHF range. Over the past decade, EPC Gen2 has established itself as the standard for UHF implementations across multiple sectors, and is at the heart of more and more RFID implementations.

2008 saw the publication of Gen 2 Version 1.2.0 which incorporated a number of enhancements requested by the retail community to support their RFID rollouts at item level.

The most recent update, Gen2v2, was developed in response to the requirements of the EPCglobal user community, and features a number of backwards-compatible, optional features [74].

HF RFID Air Interface Protocol

This standard defines the physical and logical requirements for a passive-backscatter, Interrogator-talks-first (ITF), radio-frequency identification (RFID) system operating in at 13.56 MHz frequency. The system comprises Interrogators (also known as Readers), and Tags (also known as Labels). The EPC HF air interface protocol V2.0.3 provides item-level tagging capabilities for HF at speeds faster than current HF protocols for RFID. This standard uses signaling (ASK) that is backwards compatible to ISO 15693. Also included is an optional signaling method (PJM) [75].

Low Level Reader Protocol

LLRP is a "low level" protocol between software and a reader. LLRP provides fine control over the operation of a single reader. It composed of almost 100 standard commands and provides an interface to low level functionality that is uniform across different reader vendors. Reader vendors don’t have to throw away their existing vendor-specific command language. Instead, many reader vendors support LLRP commands in parallel with their vendor-specific interface. If middleware or application software uses the LLRP interface, portability will be increased [76].

Discovery Configuration and Initialization

This GS1 EPCglobal standard specifies an interface between RFID Readers and Access Controllers and the network on which they operate. The purpose of this document is to specify the necessary and optional operations of a Reader and Client that allow them to utilize the network to which they are connected to communicate with other devices, exchange configuration information, and initialize the operation of each Reader, so that the Reader Operations Protocols can be used to control the operation of the Readers to provide tag and other information to the Client. To facilitate these operations by the Reader, an Access Controller provides several functions, described [77].

Reader Management

The current RM Standard Version 1.0.1 of the wire protocol used by management software to monitor the operating status and health of EPCglobal compliant RFID Readers. This document complements the EPCglobal Reader Protocol Version 1.1 specification [RP1]. In addition, this document defines Version 1.0 of the EPCglobal SNMP RFID MIB. Version 1.0.1 corrects errata discovered replaces Version 1.0. See additional PowerPoint that explains the errata corrected [78].

Application Level Events (ALE) Standard

The Application Level Events standard specifies an interface through which clients may obtain filtered, consolidated consolidated data capture information for physical events and related data from a variety of sources.

ALE provides a starting point for writing business logic, because it hides a lot of low level details. In particular, ALE clients do not need to know which make or model of reader is being used or even how many readers or antennas are in use. ALE delivers decoded data (e.g., an EPC URI) rather than the raw binary contents of tags. Additionally, an ALE client only needs to specify its information requirements (e.g., "give me a report once a minute of all new item-level tags that pass through loading door #5"), allowing the ALE implementation to figure out the best way to fulfill that request using the capabilities of the underlying readers. This function is implemented by running an "ALE filtering and collection engine", which can be obtained from certain reader vendors and also software vendors who specialize in RFID middleware. There is also an open-source ALE engine which can be downloaded from the Fosstrak ALE site. The ALE engine works as background process and provides a web-based Application Programming Interface (API) to user applications. ALE includes features both for reading and writing RFID tags [79].

ISO/IEC Standards

Both the International Standards Organization (ISO) and the International Electrotechnical Commission (IEC) have developed many international standards related to RFID/USNs. Below is a list of the relevant standards with some descriptive text adapted from a body of work published by the RFID in Europe group, an extension of a European Commission FP7 Thematic Network called RACE network RFID that promotes the adoption of Radio Frequency Identification and related technology solutions in Europe [80].

International standards have evolved for various sectors of RFID usage, notably in the areas of animal identification (ISO 11784 and 11785 with further development through ISO 14223/1) and contactless smart cards (ISO 10536, ISO 14443 and ISO 15693).

Other standards, having a specific application focus, can also be recognized including identification for freight containers using 2.45GHz transponders (ISO 10374) and data carriers for tools and clamping devices (ISO 69873).

The need to produce broader based standards to accommodate supply chain item management requirements has resulted in significant standardization activity being pursued through ISO/IEC JTC1 SC31 WG4 – RFID Item Management (ISO 18000 series - Information Technology – Automatic Identification and data capture techniques - Radio frequency identification for item management) with the following air-interface and data structure standards now available:

- ISO/IEC 18000-1 Part 1 – Reference architecture and definition of parameters.
- ISO/IEC 18000-2 Part 2 - Parameters for Air Interface Communications below 135 kHz
- ISO/IEC 18000-3 Part 3 - Parameters for Air Interface Communications at 13.56 MHz
- ISO/IEC 18000-4 Part 4 - Parameters for Air Interface Communications at 2.45 GHz
- ISO/IEC 18000-5 Part 5 - Parameters for Air Interface Communications at 5.8 GHz - abandoned project.
- ISO/IEC 18000-6 Part 6 - Parameters for Air Interface Communications at 860 to 930 MHz
- ISO/IEC 18000-7 Part 7 - Parameters for Air Interface Communications at 433 MHz
- ISO/IEC 15961 RFID for Item Management - Data protocol: Application interface.
- ISO/IEC 15962 RFID for Item Management - Protocol: Data encoding rules and logical memory functions.

- ISO/IEC 15963 RFID for Item Management – Unique Identification of RF Tag.
- ISO/IEC 15459-4 - System of Unique Item Identification Codes.

Work on further ISO/IEC standards is in progress for conformance, applications profiling and interfacing:

- ISO/IEC 18001 RFID for Item Management – Application requirements profiles.
- ISO/IEC TR18047 Technical Report - RFID Conformance Test Methods.
- ISO/IEC 18047-2 Part 2: Parameters for air interface communications below 135 kHz
- ISO/IEC 18047-3 Part 3: Parameters for air interface communications at 13.56 MHz
- ISO/IEC 18047-4 Part 4: Parameters for air interface communications at 2.45 MHz
- ISO/IEC 18047-6 Part 6: Parameters for air interface communications at 860-960 MHz
- ISO/IEC 18047-7 Part 7: Parameters for air interface communications at 433 MHz
- ISO/IEC 19789 RFID for Item Management – Application Programmer Interface (API).
- ISO/IEC TR 24710 Information Technology AIDC Techniques – RFID for Item Management-ISO 18000 Air Interface.

Communications Elementary Tag license plate functionality for ISO 18000 air interface definitions.

6 Trends toward future digital data format

6.1 Technical issues of future data formats

Paradox between open data and data business

Most people may confront basic questions while they produce valuable contents and materials: how to distribute and share their contents and materials with others. If people create some documents by using desktop computer, they may send files to others via Email or file transfer protocol. If people want to post their documents to public or anonymous, the web portal sites or personal live TV programs are applicable. If some people want to use their own social networking services like blog and Facebook, they can post their created contents at that site.

Here, the first question is how to distribute their created contents widely if possible. The probability how many users visit at those web sites and TV channels is important to evaluate their media impacts. Most power bloggers or smart marketers want to reach their contents to target audiences and potential customers as much as possible. Most contents may be easy to get lost in the crowd since the incredible number of contents and materials are created and uploaded every day. Without filtering mechanism of spam or unwanted advertisements, the push messages like Short Message Service (SMS) or Really Simple Syndication (RSS) are not recommended. In the zeta-bytes era, the existing document formats with short description may be lost or unnoticed by audience. Most users feel some difficulty to find valuable information among millions of documents through the web sites. They also suffer to find correct document at the same web site since a web site may have a lot of hierarchically linked pages or sub-categories. Otherwise, the simple delivery of each document like email is more acceptable for novice customers. When new version of software and application packages are released, most people may not recognize the fact of new release at acceptable duration of time. Without help of active update tools or notification messages, it is not easy for individuals to upgrade the software on their own desktop.

The second question is how to distribute the user owned contents only for the registered customers or the negotiated correspondents correctly and confidentially. When the users send the documents to the target business customer, they may be worry about illegal distribution of the contents. The users can post some payable contents at the private web sites which assume to be only allowed for the subscribed customers. If users may suspect target customers, they do not want to send or post the documents by on-line manner. Even though the original contents are encrypted with proper security code, the destination customers may re-distribute the received documents to others after decoding the documents with relevant key. They may illegally distribute the original documents and key information together. It means that the security protection on document is not effective if the receivers intentionally and illegally distribute the original document to others. The relevant tracking or certification mechanisms from original owners are needed. The delivery of raw data file with/without security protection are not recommended to cope with this types of behaviors.

Both questions mentioned above contain paradox. There is paradox between open data for public and data security for business. The public available documents may be used sometimes for business. Sometimes, the securely protected documents may be open to public. It means that the existing data formats are not designed to solve these issues at same time. The contents or document originally designed to be open to public are not easy to extend for business. Otherwise, the unacceptable situation may happen during conversion of document. If some people try to find business opportunities from open data, they can add on the irreducible values over original open data with security protection. If the documents created for private purpose are unacceptably open to public, privacy problems may happen. If some search engines find the relevant document among millions of open data repository at an instance, the users can save the searching time to find the document. On the other hand, customer feels some difficulty to find out private documents with security protection. Without help of advertisements, the owner of private and secured documents has some difficulty to search for target customers. To make a business from public or anonymous, the title of documents with short descriptions should be open and searchable.

Web as a Platform

Internet is originally designed for global networking of computing resources. Basic service of Internets is web services/applications as well as email, telnet, and ftp. For world wide web (WWW), all the resources on the Internet are communicated by using hypertext transfer protocol (HTTP). The WWW is a way of exchanging information between computers on the Internet and sharing a vast collection of interactive multimedia resources. The URL is used to specify addresses in the Web. The URL is the fundamental network identification scheme for any resource connected to the Web. The web site is a collection of various pages written in HTML. This is a location where people can find hypertext pages, images, and sound file. Currently, there are billions of web sites available on the Internet. All the web site is installed at computer called by web server. The web server is connected by a unique web address known as domain name like www.itu.int. To access web sites, the web browser software is needed.

In the concept of client/server platform, the web platform is building resources for a better web regardless of manufacturer, browser or platform venders. In the client aspect, the web platform is not run by a single corporation or a single person. Anyone is welcome to write and edit the documents, share and comment web blog posts, and communicate each other. The web platform provides links to various helpful resources which is intended for a general audience. The web platform includes a list of browser technologies being currently developed, implemented, and tested.

A computing platform provides low level functionality for application process which include hardware, an operating system, and runtime libraries. A web browser itself runs on a hardware platform with relevant operating system. An application like spreadsheet or word processor is a fully-fledged application as a platform. The cloud computing platform allows application developers to build software out of components that are hosted not only by the developer, but also

by provider. The web platform is compared with computing platform that includes any piece of software described above. If some virtualized hardware, operating system, software, and storage are mapped to the resources defined by web script language and web technologies, the web platform could be recognized as one of computing platform.

To support web platform, JavaScript as a high level, dynamic, interpreted programming language, is supported by most web browsers without plug-ins. It can support object-oriented, imperative, and functional programming styles. It has an API for working with texts, array, dates, and regular expressions. But, since most web browser does not include any external interface like networking, storage, external graphic interfaces and peripherals, JavaScript is used to support the real-time applications including IoT and WebRTC interface. Additionally, the server-side networking programming with run-time environment can use node.js which is an open-source, cross-platform JavaScript runtime environment for developing a diverse variety of tools and applications.

Impacts of web site or web page

The web pages in on-line electric form replace traditional books to share valuable information of documents. The HTML format of web page includes CSS for data representation, XML/RDF for explaining data, and JavaScript for data processing. JavaScript Object Notation (JSON) is also a typical data format widely used in the web services as well as web page. Recently, multimedia forms of audio/video, image, drawing and 2D/3D animation may be included at the web page. The concept of hyperlink is useful to link the related documents or the other web pages. It notes that all the on-line documents and materials should be well described through the web page.

Recently, the web page can include real-time live channels like personal broadcast service, remote surveillance camera, and motion detection sensors. The multimodal interfaces by using external I/O (Input/Output) devices (like mouse, speaker, touch pad, and other peripheral devices) and plentiful Internet of Thing (IoT) sensors can be also utilized through the web page. Additionally, devices and systems which are mainly used for other industries such as manufacturing, remote surveillance, and vehicles, etc. are merged and integrated into a web-based application platform.

These trends mean that the web page is evolved as a human readable application platform. In an application, the URL provides the entrance gate of specific application platform. Recently, major companies may advertise for the users to remember their URLs. Otherwise, they should rely on the relevant searching machines to click their URL as preferable bookmark. In this case, the URL is mainly guiding the users to enter specific application platform by replacing the download of specific application software.

Here, traditional media like television, news, magazine, and animation/game has some difficulty to advertise their new surprising published contents and materials. They utilize web pages (i.e., URL information) to announce new information to public. Some users may feel strange to click the URLs in order to enjoy the interesting television programs. But, young people easily enjoy the menu offered by specific web sites which are arranged to lot of media contents and materials. Therefore, many media service providers utilize the web technologies on their own media platform. The web technologies can provide the useful capabilities for such integrated or mashup service environments.

Some solution providers with a lot of utility software packages want to advertise their solutions through web sites. But, to run specific utility software, target service environment may be well configured with proper data structure. Target platform is not easy to operate the application software since the complicate back-end processing and background tasks should be aligned. Since the concept of web browser is not designed for that purpose, the web platform may have some limitations to support real-time application including data collection, processing, and storage. In a case, the web platform should invoke or activate the executable file and software. JavaScript provides a unique solution to support the executable web applications. In case that some special software packages or application platforms are linked through web sites, a web browser could not

interactively communicate with other web browser. For support of future mashup application, the existing design concepts of web sites or web pages should be reviewed and enhanced. As a solution, the current concepts of web application program interfaces (APIs) may give a way to invite millions of application software within the web site or web page. If the web page provides a way for executable files through open API, the web technologies will be common basis of future convergence applications. This issue is summarized as "Are web sites or web pages recognized as open application platform?"

Effects of web screen as human interface

The original design concept of web is simple. Through web site, any people can see the web contents which are open to public. The web page assumes to layout the contents at window screen, which is encoded by HTML. The dynamic frame format of HTML5 are tuned to different size of screens like smartphone or smart pad. With web-based applications, users access the contents via the web browser. The web browser displays the web contents to screen regardless of the operating systems. Since web browser is widely applicable to smart phone and smart pad as well as desktop computer, many researchers and software developers are interesting on how to utilize web over their application environments. Therefore, peoples utilize the web technologies for a variety of complex and sophisticated applications domains. In addition, the multifaceted functionality of the web applications provides unique features on usability, performance, security, and ability.

For rendering data at web page, there are the standard Document Object Model (DOM) encoded by HTML, JavaScript for execution, and CSS for style. When the volume of data is increasing so that it is impossible to display the content at single web screen, simple rendering the data is not effective. More efficient rendering skill is needed to layout the number of web pages and volume of data file to screen. If a large volume of data is collected and accumulated by measuring devices and on-line channels, direct rendering their contents to web-enabled screen is not efficient. Data visualization tools or other rendering machines help people analyse and perceive data patterns. The cloud computing platform can collect, store, and process the data efficiently, in which data visualization tool as a concept of SaaS (Software as a Service) provides a way for rendering the data to web screen. It notes that the direct rendering at web page may have some limitation since human perception is a key of intelligence extracted from data. Then, people utilize data visualization tools over cloud computing platform to interpret the meaning of data.

In advance of augment reality/virtual reality (AR/VR), a number of web screens are used to render 2D/3D information. The rendering technologies over multiple screens are important especially for synchronization of contents. Web screens designed by HTML should be upgraded to display audio/visual information to screens simultaneously and synchronously. The web browser contains the location information of multiple screens and handle time information of each individual contents as well as powerful decoding software of audio/video contents. Sometimes, high performance hardware solutions may assist web browser to arrange multiple screens at the same time. When a lot of animation images or videos for augmented reality are overlapped over background images of screen, multiple rendering engines should coordinate multiple screens simultaneously with accurate timing control.

How to activate application software at the web environment?

There are a lot of web services written by WSDL (Web Service Description Language). The SOAP (Simple Object Access Protocol) over HTTP is defined to activate web services. Otherwise, the external inputs (e.g., clicking mouse or touching screen) can activate some tasks which are enabled by JavaScript. Some extension of HTTP (e.g., XMLHttpRrequest) can invoke a web services alerted by IoT devices. It provides the communication between web elements and external components. The HTTP which is originally designed for human-to-machine communication, is

utilized for machine-to-machine communication as a part of IoT services. The machine readable web data formats like XML and JSON are applicable to IoT services. The in-line JavaScript code offered by web browser can be used to change, modify, and process the web resources which are directly linked to IoT devices. But, there are some constraints since all the web resources should be specified in a form of XML/RDF and their schema. It is not same with the existing application software. Various application software is originally designed to consist of specific hardware and software environments including operating systems (e.g., Android, IOS, and Window). The equivalent capability with the existing application software is not easily realized at the web environment. The web API solutions provide the hardware- and OS-independent application software. But, there are uncountable number of issues to be solved. If some web resources should be handled with accurate timing and performances, simple JavaScript codes may be not relevant. Moreover, if some metadata including tag information or the secured protection field of data are included at web service, the relevant background tasks are ready to extract values from raw data. If users click specific words or fields on web screen, a lot of application tasks like email, short message, video chatting, and voice recording have to be waked up. Also, the external inputs from IoT sensors invoke some tasks and represent their current status on web site.

In a conclusion, the simple or basic rules to activate tasks should be clearly specified or standardized to keep that most web services and web data formats are interoperable. The step-wise deployment strategies for activating software or tasks within the web environment should be established. The standards of web-based data format will be well developed for future flexibility, scalability, and adaptability of the corresponding application software. It means that the data format gives a clear guidance how to run the related application software. The existing billions of software and application programs can be well arranged according to data format.

Open application program interfaces (APIs) at web environment

A lot of software solution providers investigate the web-enabled versions of their own software since the web originally intends to open to public. The web service provides an object-oriented interface to application servers including database. It also has a user interface by using smartphone as well as desktop personal computer. Another common application of web is a mashup capability where the web services are interoperable at difference machines and various operating systems. Heterogeneous web applications from different locations can be mashed up by using web APIs. The web service provides a standardized way of integrating web-based application using open standards over Internet like XML, SOAP, WSDL, and UDDI. The web technologies including HTTP and HTML are designed for human-to-machine communication as well as machine-to-machine communication. The cloud computing systems are used to exchange data with each other and a web service provides a methods of communication over the cloud. With web concepts, the methods of data exchange should not depend on particular platforms and operating systems. XML tag and XML schema can be interpreted for data exchange. The communication rules between web resources are defined in WSDL which describes more flexible web services.

Most cloud computing systems assume to support web services, which are called by SaaS (Software as a Service), PaaS (Platform as a Service), IaaS (Infrastructure as a Service), and NaaS (Network as a Service). It means that their software, platform, infrastructure, and even network environments will be merged into the concepts of cloud computing. Here, most web services rely on the cloud computing for data storage and processing.

There are more than billion software in the ICT world until now. The current activities of open-source software (e.g., GitHub) are re-issued and combined with web concepts. Currently, any device, system, and platform should be equipped with the relevant software to operate and activate. The software is a tool to activate device and platform. The users want to minimize a learning or studying activities in using, modifying, and sharing the software while they operate the devices and platforms. A similar software program is distributed to the neighbours so that they can use it for the

expected operation of devices. The users want to enhance a program and the modified source codes are also available to others. The concepts of free and open software will be well aligned with web culture. In the web service or web architecture, the key issues are how to share and re-use the existing software at the other applications. At the web environment, the relevant web APIs including data format and access technology should be defined. The JavaScript codes can be extended to free and open software environments.

How to make links among data to extract values?

In the zeta byte era, simple data delivery is not significant issues. More than million video files and game applications per day are uploaded and downloaded. Traffic volumes of virtual reality and video surveillance sharply increases every year. The data analytics are examining large volumes of data to uncover hidden patterns, unknown correlations, market trends, and user preferences. The real-time data and mobile applications requires high processing capabilities to manipulate large and diverse data sets across heterogeneous systems. The data analytic technologies including Hadoop and NoSQL databases forms the core of an open source software framework to handle incoming stream of raw data. The large amount of data may cause some headaches on data management system. The linked information of data is useful through semantic queries. While creating, sending, and storing data, a linked information among structured data become more useful for data analytics.

The current linked data format of W3C has four principles as follows [81]

- Use URIs to name (identify) things.
- Use HTTP URIs so that these things can be looked up (interpreted).
- Provide useful information about what a name identifies when it's looked up
- Refer to other things using their HTTP URI-based names when publishing data

Here, these principles are mainly applicable to XML/RDF files. The different data sources like audio/video, surveillance, energy metering, and status monitoring are not relevant to directly satisfy these rules. Data linking tools for more than millions of data package provide complicate semantic process. The collaboratively-created linked datasets are useful for data analytics. Individual data sets without describing properties and relationships are not easy to extract a significant value. It concludes that all data formats are equipped with linked data information when they are created, processed, updated, and distributed. The linked data information is essential to provide add-on values which is called by “Data is new Oil”. A structured data like spreadsheet contains the formula which deliver linked information of multiple sheets and eliminate identical sets of data. The structuring process of linked data saves time, reduces errors, and improve data integrity. The well-structured spreadsheet with link information provides better understanding and creates more values.

The hyperlink concept of web page is used to point a whole document or a specific element as a reference. Hyperlinks are used to implement reference mechanisms like image, thumbnail, cropped section, tables of contents, footnotes, bibliographies, indexes, and glossaries. For extension, hyperlink is a link bound to a portion of document or a hot area in an image. A boundary on geographical map is hyperlinked to further information about the area. A separate invisible hot area allows to connect the related information. Simple image, thumbnail, low resolution preview, and cropped section of a web page contain an external link to indicate full contents or categories of the related page layouts.

In XML/RDF documents, the multidirectional links describes a greater degree of functionality, which describe data classification, linked relationship, and sorting, etc. As a part of data management, data classification is particular important for some applications like risk management,

legal discovery, and government regulation. The structured form of data including relational or tabular data and audio/video data is relatively simple to make a process. Effective data classification can significantly improve their performance and utilization according to specific applications.

Data classification and data aggregation

Data classification is a part of data management. To categorize data, the following question is identified which is similar to description of metadata.

- What data types are available?
- Where are certain data located?
- What access levels are defined?
- What protection level is needed and does it adhere to regulations?

The data classification is depending on applications. Some private data may be accessible through proprietary application program. An application produces the structured data stored at database. Some IoT applications produces the un-structured data which is not easy for context-aware intelligent decision making. Audio/video data is relatively simple process of data classification since creation, delivery and distribution of A/V data are well designed and formatted until now. The complex procedures of data classification ensure adequate quality provisioning of data analytics. Benefits of well classified data can significantly improve data handling performance and save the resources including storage. Therefore, efficient data classification can reduce costs and administration overheads. Data indexing and tagging also improve user access speed.

For business, data classification is to cluster the data set used for category. The algorithm used on the categories is creating a descriptive model for individual applications. The effectiveness of data classification is measured by [82]

- **Predictive accuracy**: How well does it predict the categories for new observations?
- **Speed**: What is the computational cost of using the classifier?
- **Robustness**: How well do the models created perform if data quality is low?
- **Scalability**: Does the classifier function efficiently with large amounts of data?
- **Interpretability**: Are the results understandable to users?

Data tag and data index

For web search engine, the index is used to collect, parse, and store data to facilitate fast and accurate information retrieval. The creation of index incorporates knowledge accumulation with interdisciplinary concepts from linguistics, cognitive psychology, mathematics, and informatics. The index process is closely related to search engines which is designed to find web page on the Internet. Popular engines focus on the full-text indexing of online, natural language documents. The real-time media like audio, video, and image are also searchable. For the files with metadata description, search engines can use the index of audio/video applications. The recorded media can be searched by indexing at a predetermined time interval due to the required processing time.

The purpose of inserting or storing an index is to optimize speed and performance in finding relevant documents for a search query. Without an index, the search engine would scan all

document in the database which require considerable time and computing power. The computation time to store index is traded off for the time saved during information retrieval. Indexes are useful for many applications, but come with some limitations. With an index, the database operation has much less computationally burdens than a full database scan. For sequential search is performed, index is applicable for high performance of lookup. In this case, the index is used to locate the data record from which the required data is read.

In database, the index is used to quickly locate data without having to search and improves the speed of data retrieval operation. It can be created for both rapid random lookups and efficient access of ordered records. It can include direct link to the whole database to be searched efficiently. Since most databases contains millions of objects or documents and the lookup is a common operation, the index technology can improve the performance of lookup regardless of database structure. An index also supports fast searching for the recorded structure of database. If the data are contained in arbitrary order, the index can specify the logical ordering. The index can be expressed in the sorted order regardless of physical data structure. If the distributed data is clustering into a certain distinct order, the clustered index can greatly increase overall speed of retrieval. Since some data is physically separated and indexed some blocks into separate files, the clustered index can put in order for different data blocks within the same physical file.

On the other hand, tag is keyword or term assigned to a piece of information like bookmark on the web page. Tag helps describe an item and allows it to be found by browsing or searching. Tagging are carried out to Tags are created informally and personally depending on the system and applications. Recently, tagging is an important feature of many web 2.0 services and it is available at most browser.

Data schema

In general, schema can be well described a pattern of thought or behavior that organizes categories of information and the relationship among them. The primary purpose of schema is to specify what the structure of a document can be. A schema is analogous to a grammar for a language. It also describes the structure of data format and a framework to represent a system of organizing and perceiving information. Schema contains some attention of new knowledge. The complex thought is quickly perceived by using schema since schema organizes current knowledge and provides a framework for future understanding. Example of schema include types, role, script, and viewpoint. By using schema, a technique to encode and retrieve data is not much required for strenuous processing. Schema is generally thought to have a level of activation which is known as a generic knowledge of sequences of actions. The schema is determined by user experience and expertise. It allows the common explanation to be chosen from the given data format.

In XML/RDF document, the schema describes a type of XML/RDF document, which is expressed by some combination of grammatical rules of data elements. It includes data types governing the content of elements and attributes, and more specialized rules. The XML document may not be usable in the absence of its schema. The XML document should be produced or parsed according to XML schema specification. The document type definition (DTD) is widely used for XML specification for the expression of schema. The DTD can be defined for the textual content of data elements and attributes, for example, by specifying ranges of values, regular expressions, or by enumerating the permitted values. Some schema is written for editing and transforming the document, in which XML schema is designed to make manipulation of the XML instance in application programs. Recently, some XML schema for convergence applications supports the unordered and non-deterministic contents, sometimes, with irregular expression and unstructured form. It would be unusual to create schema for the unstructured form of contents. However, some intensive discussion on tags or attribute names should be needed for schema design. The naming conventions of natural language can be extended for unstructured form of data.

Microdata format without learning

A microdata format uses HTML and/or XML tags to convey additional metadata and other attributes in web pages. It allows software to process information intended for end-users automatically, like contact information, geographic coordinates, calendar events, and similar information. Although the content of web pages has been capable of some "automated processing", such processing is difficult because the markup tags do not describe what the information means. Microdata formats can bridge this gap by attaching semantics and automated processing of web page, like natural language processing or screen scraping. The microdata format enables XML/RDF data items to be indexed, searched for, saved or cross-referenced, so that information can be reused or combined. It allows the encoding and extraction of data details, contact information, social relationships and similar information. The microdata format like hCard are published on the web page more than alternatives like schema (microdata) and RDFa.

The microdata formats for HTML provide additional formatting and semantic data. For example, web crawlers can collect data about on-line resources. The use of microdata formats can also facilitate "mash ups", for example, exporting all of the geographical locations on a web page. Several browsers provide the microdata formats at an HTML page. Some browser extensions allow the microdata formats for hCard or hCalendar at the applications with contact information and calendar schedule. They allow the location data format to geographical applications. By utilizing microdata format, the web browser can get the following information.

- Knows what applications are accessible to the user and what the user's preferences are
- Easy to develop web site if web designer does not need to handle "appearance" or "action" issues
- Keep backwards compatibility with web browsers that don't support microdata formats
- Simplifies security issues since the web browser presents a single point of entry
-

The design principles and practical aspects of microdata formats have been compared to other approaches. The spread and use of microdata formats has been advocated to see a bunch of microdata formats being developed. It assume the way how the semantic web will be built. The design principles of microdata format can be summarized as follows:

- **Reduce**: favor the simplest solutions and focus attention on specific problems,
- **Reuse**: work from experience and favor examples of current practice,
- **Recycle**: encourage modularity and the ability to embed, reused in blog posts, RSS feeds.

Microdata formats are not the only solution for providing "more intelligent data" on the web: alternative approaches are used and are under development. For example, the use of XML markup and standards of the Semantic Web are cited as alternative approaches. One advantage of microdata format is to lower the barrier to learn the new data format. It is similar when the people want to study new language. For some applications, if the type of data to be described does not map to an existing microdata, new RDF schema can embed arbitrary vocabularies into HTML, like for example domain-specific scientific data such as zoological or chemical data for which there is no microdata format. The web standards allow microdata formats to be converted into data compatible

with the Semantic Web. The microdata formats provide an easy way for many people to understand semantic data to the web.

For the structured form of micro-data format, for example, the information in a relational database refers to the structured data with a high level of organization. When information is highly structured and predictable, search engines can more easily organize and display it in creative ways. The standard structured data in a variety of online applications is widely accepted at open community. The structured data markup is most easily represented in JSON-LD (JavaScript Object Notation – Linked Data) format. The structured data markup describes things or objects on the web, along with their properties. For example, at the web sites, people could use markup to describe properties for each recipe, like the summary, the URL to a photo for the dish, and its overall rating from users. When users provide structured data markup for their online contents, users make that content eligible to appear in two categories of Google Search features:

- **Rich results** - Structured data for things like recipes, articles, and videos can appear in Rich Cards, as either a single element or a list of items. Other kinds of structured data can enhance the appearance of personal web sites for easy search.
- **Knowledge Graph** - If some people are the authority for certain contents, the search engine can treat the structured data on their site as factual and import it into the Knowledge Graph [83], where it can power prominent answers in search engine. Knowledge Graph appears for authoritative data about organizations, and events based on reviews and ranking. It enhances the search engine's search results with semantic search information gathered from a wide variety of sources.

The following table lists the supporting formats and corresponding feature of microdata format [84].

Table 10. A list of microdata formats

Format	Descriptions
JSON-LD (JavaScript Object Notation – Linked Data)	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.
Microdata	It is the nested structured data within HTML contents. It uses HTML tag attributes to name the properties of the structured data.
RDFa (Resource Description Framework in Attribute)	The HTML5 supports the linked data by introducing HTML tag attributes that correspond to the user-visible contents

Data format for IoT applications

The Internet of Things (IoT) is designed to connect the physical devices, vehicles, buildings and other items to ICT world. It provides network connectivity on electronics, software, sensors, actuators that enable these physical objects to collect and exchange data. The potential IoT applications include smart city, smart car, smart grid, smart home, and smart industries. Public safety, remote surveillance, environment protection, smart agriculture/farm and smart tourism also acquire high attention from people as IoT applications.

In data aspects, IoT provide advanced connectivity of devices, system, and services within ICT infrastructure. IoT data are classified into sensing data from IoT devices, aggregation data from IoT gateway/brokers, and context-aware data from IoT servers. IoT devices collect useful data and send it to IoT gateway or other devices. For example, smart home devices control lighting, smart thermostat for heating, ventilation, air conditioning system, and a lot of appliances like washer/dryer, robotic vacuum cleaner, air purifier, oven, refrigerator/freezer by using WiFi or near field communication technology. Since IoT application is expected to generate large amounts of data from diverse devices, IoT data should be aggregated, stored, and processed in more effective manner. In this sense, the IoT platform looks like the management platform for smart city and smart grid. A lot of small packets of data from IoT devices are merged to the gateway node and automate the operation from home appliances to entire factories. If all the IoT devices are equipped with identifiers, the IoT platform analyses the context-ware data from sensors and manage the proper actions. To implement IoT applications, all the object data with identifying device or machine-readable identifiers is to be transformed to make an intelligent control. The objects in the IoT application is not only devices with sensory capability, but also provide actuation capabilities (e.g., bulb or switches controlled though Internet). IoT systems could also be responsible for performing actions, not just sensing things. Some IoT sensors could monitor user's purchasing behaviours in a store by tracking their smartphone. Based on analysis of monitoring results, the special offers on user's favourite products can be provided. Other IoT applications are extended to home security and home automation. The accumulated credit by using home IoT sensors make smart home being more secure and safe.

There are various types of IoT devices which measure and collect the sensing data stream. The sensing data consists of the unique identifier and context-aware information. It is transferred to the gateway node by using relevant transport protocol with IP address. The IoT data is accumulated or stored with the relevant structured form of database. IoT platform performs some actions by analysing all the sensing data and extracting context-aware information.

Metadata format

Title, name, and type of data should be well identified at instances that data sets are created, delivered, and consumed. With good description of data, users decide their usages and the application programs take relevant actions. Generally, data description would cover [85].

- What the data is
- Who can use it
- When it can be used
- How it can be used
- What it might be used for
- Where it can be found
- How long it will be available

Most of existing files and documents contains metadata information like library catalogues, museum collections, digital audio/video file, web sites, and on-line books. For example, metadata for digital audio/video/image describes how large the picture is, color depth, quality/resolution, frame speed, and other data. A text document's metadata contains information about how long the document is, who the author is, when the document was written, and a short summary of the document. Metadata within web pages can also contain descriptions of page content, as well as key

words linked to the content. These links are often called "Metatags", which were used as the primary factor in determining order for a web search.

For individual applications, some examples of metadata include

- Means of creation of the data
- Purpose of the data
- Time and date of creation
- Creator or author of the data
- Location on a network where the data was created
- Standards used
- File size

While the metadata application covers a large variety of fields, there are specialized and well-accepted models to specify types of metadata. Descriptive metadata is typically used for discovery and identification, as information to search and locate an object, such as title, author, subjects, keywords, and publisher. For guide information, metadata helps humans find specific items and are usually expressed as a set of keywords in a natural language. For technical matters, metadata corresponds to internal metadata, and business metadata corresponds to external metadata. Structural metadata describes how the components of an object are organized. It describes the structure of database objects such as tables, columns, keys and indexes. An example of structural metadata would be how pages are ordered to form chapters of a book. Finally, administrative metadata gives information to help manage the source. Administrative metadata refers to the technical information, including file type, or when and how the file was created. Two sub-types of administrative metadata are rights management metadata and preservation metadata. Rights management metadata explains intellectual property rights, while preservation metadata contains information to preserve and save a resource.

Metadata can be stored and managed in a database, often called a metadata registry or metadata repository. However, without context and a point of reference, it might be impossible to identify metadata just by looking at it. For example, by itself, a serial number at database could be the results of calculations or a list of numbers perceived as the data. But if given the context that the database is a log of a book collection, ISBNs, information that refers to the book, are not itself the information within the book. "Structural metadata" i.e. "data about the containers of data are usually found in library catalogues.

As for typical example of metadata description, the following gives some general guideline of document [86].

Table 11. An example of metadata description at document

Title	Name of the dataset or research project that produced it
Creator	Names and addresses of the organization or people who created the data

Identifier	Number used to identify the data, even if it is just an internal project reference number
Subject	Keywords or phrases describing the subject or content of the data
Funders	Organizations or agencies who funded the research
Rights	Any known intellectual property rights held for the data
Access information	Where and how your data can be accessed by other researchers
Language	Language(s) of the intellectual content of the resource, when applicable
Dates	Key dates associated with the data, including: project start and end date; release date; time period covered by the data; and other dates associated with the data lifespan, e.g., maintenance cycle, update schedule
Location	Where the data relates to a physical location, record information about its spatial coverage
Methodology	How the data was generated, including equipment or software used, experimental protocol, other things one might include in a lab notebook
Data processing	Along the way, record any information on how the data has been altered or processed
Sources	Citations to material for data derived from other sources, including details of where the source data is held and how it was accessed
List of file names	List of all data files associated with the project, with their names and file extensions
File Formats	Format(s) of the data, e.g. FITS, SPSS, HTML, JPEG, and any software required to read the data
File structure	Organization of the data file(s) and the layout of the variables, when applicable
Variable list	List of variables in the data files, when applicable

Code lists	Explanation of codes or abbreviations used in either the file names or the variables in the data files
Versions	Date/time stamp for each file, and use a separate ID for each version
Checksums	To test if your file has changed over time

For the web-based data format, the "Dublin Core" achieved wide dissemination of metadata standard [87]. It focuses on generic data model for metadata. Dublin core defines most popular metadata for use with the Resource Description Framework (RDF) and linked open data. The key idea is "simple metadata for resource discovery", which is based on current natural language.

Data format for accumulation of experience and insights

The knowledge-accumulation process can be conceptualized at a high level of abstraction as composed of two principal elements: mechanisms for search and means to select the contents. The term 'search' here defines the process by which people select among the documents to a defined problem. This characterization of innovation-oriented activities implies a four-phase problem-solving cycle: a gap between actual and desired performance on a given dimension of managers, designers, or engineers. Experience is essential to data process. To choose the texts to be searched, people employ previously acquired knowledge, distilled in experience. At the problem-solving cycle, decisions must be made about which potential solutions are to be tested. The more and better the knowledge users can bring to bear on the preliminary selection process, the higher will be the likelihood that the chosen set meets these criteria. Different forms of experience can yield at least three categories of useful knowledge: insight into which problems are most valuable to solve and information about where solutions are most likely to be found. Each of these can be considered separately as forms of learning at which the learning occurs and is stored.

Knowledge of which problems are most important comes especially from previous experiences in related fields. Experience with the application of new process, for example, leads to focus on problems more directly relevant to actual concerns. Learning occurs at the person who select and frame the portfolio of innovation. Experience also improves understanding of the search process itself. Such knowledge can manifest in changes to the relative allocation of resources to various search strategies, for example to shifts from basic to applied research. Similarly, knowledge of where solutions are most likely to be found comes especially from prior experience. Such experience can increase the confidence that it is aware of all potential solutions and that the option set within which it is searching includes a viable solution.

Tacit knowledge can be defined as skills, ideas and experiences that people have in their minds and are, therefore, difficult to access because it is often not codified and may not be easily expressed. With tacit knowledge, people are not often aware of the knowledge they possess or how it can be valuable to others. Effective transfer of tacit knowledge generally requires extensive personal contact, regular interaction and trust. This kind of knowledge can only be revealed through practice in a particular context and transmitted through social networks. Some examples of tacit knowledge are: riding a bike, playing the piano, driving a car, hitting a nail with a hammer, and putting together pieces of a complex puzzle, interpreting a complex statistical equation.

In the field of knowledge management, the concept of tacit knowledge refers to a knowledge which cannot be fully codified. Therefore, an individual can acquire tacit knowledge without language. Apprentices, for example, work with their mentors and learn craftsmanship not through language

but by observation, imitation, and practice. The key to acquiring tacit knowledge is experience. Without some form of shared experience, it is extremely difficult for people to share each other's thinking processes. Tacit knowledge has been described as "know-how" – as opposed to "know-that" (facts). Tacit knowledge involves learning and skill but not in a way that can be written down. Knowing-how or embodied knowledge is characteristic of the expert, who acts, makes judgments, and so forth without explicitly reflecting on the principles or rules involved. The expert works without having a theory of his or her work; he or she just performs skillfully without deliberation or focused attention. Although it is possible to distinguish conceptually between explicit and tacit knowledge, they are not separate and discrete in practice. The interaction between these two modes of knowing is vital for the creation of new knowledge.

To convert tacit knowledge into explicit form of data, the following issues should be investigated: [88].

- **Data Schema or mechanism of transferring knowledge**

  Even though tacit knowledge is intuitive and unarticulated, the codified schema or mechanism which can "write it down", "put it into words", or "draw a picture" would be communicated, understood, or used without "knowing in detail". This form of data format requires close interaction, shared understanding, and trust among people. For example, if people recognizes a person's face among a thousand, some mechanism puts into words as a schema of image and all other people recognize the face image as a whole.

- **Acquisition and accumulation of experience**

  Since tacit knowledge is acquired and accumulated through practical experience, logical deduction of relevant contexts of documents can be generated. Therefore, schema for explicit data contexts are not fixed, which can be interactively modified or enhanced by accumulation of experiences. For example, to learn a language by being taught the rules of grammar, a young age entirely unaware of the formal grammar can be interactively taught by a native speaker. Other examples are how to ride bike, how tight to make a bandage, or knowing whether a senior surgeon feels an intern is ready to learn the intricacies of surgery.

- **Appropriation and assessments of tacit contexts**

  Though tacit knowledge is personal contextual, the final form of knowledge can be aggregated and stored at a certain database, and appropriated without accumulation of individual insights. The realization of tacit knowledge requires the close involvement and cooperation of the knowing object.

The process of transforming tacit knowledge into explicit or specifiable knowledge is known as codification, articulation, or specification. Tacit knowledge should be codified and transmitted through accumulation of personal experience. When codifying or articulating tacit knowledge into explicit knowledge, the novice becomes an expert by acquiring skills or declarative knowledge. All the propositional knowledge (knowledge that) is ultimately reducible to practical knowledge (knowledge how).

The descriptive knowledge is the type of knowledge, by its nature, expressed in declarative sentences or indicative propositions. It distinguishes descriptive knowledge from wat is commonly known as "know-how", or procedural knowledge (the knowledge of how, and especially how best, to perform some task). For knowledge in science and engineering, a scientist picks a question of interest based on previous knowledge and hypothesis. The scientist designs the controlled experiment which allow him to test the hypotheses against the previous experience. In this case, the experiment procedure is well described if the observations match the predictions based on the

hypotheses. A hypothesis has been shown to accurately and reliably predict and characterize some physical phenomenon. New knowledge generated by scientific experiments are well described for human reasoning as a piece of scientific work.

Data management and data governance

Data intelligence is extracting from data to information and knowledge. It can be used by companies to extend their services or business. For business intelligence, data should be managed as a valuable resource. Based on needs of an enterprise, data is properly managed by policies and procedures to control, protect, deliver and enhance the value of data and information assets. If the data is not well defined, the data would be mis-used in applications. If the data management process is not well defined, it is impossible to meet user needs.

Some data require to protect relevant rights and freedoms. In particular, privacy right is significant while processing personal data. Anyone holding personal data for other purposes is legally obliged. Technically, personal data should be handled by acceptable protection process. The Trusted Data Format (TDF) is used for the purposes of enabling file level tagging and security features [89]. It includes assertion of data properties or tags, cryptographic binding and data encryption. Both data and metadata objects can be associated with a block of encryption information which is used by any TDF consumer to decrypt the associated data or metadata if it had been encrypted.

Interesting topics in data management include [90].

- Data governance
- Data Architecture, Analysis and Design
- Database Management
  - Include data maintenance, database administration, database management system
- Data Security Management
  - Include data access, data erasure, data privacy
- Data Quality Management
  - Include data cleansing, data integrity, data enrichment, data quality assurance
- Reference and Master Data Management
  - Include data integration and reference data
- Data Warehousing and Business Intelligence Management
  - Include business intelligence and data mining
- Document, Record and Content Management
  - Include document management system, records management
- Meta Data Management
  - Include metadata discovery, metadata publishing, and metadata registry
- Contact Data Management
  - Include business continuity planning, marketing operations, customer data integration, and identity management

For future ICT world, a systematic approach including data management functions described above should be carefully selected. The associated technical, legal or economic attributes of data can

enhance the ability of digital assets or specific objectives of business. The data management process ensures that important data assets are managed through information value-chains. Data can be trusted and accountable for any negative events that happens because of low data quality. Data quality control is key discipline for assessing, managing, using, improving, monitoring, maintaining, and protecting data assets. Some goals for data governance include [91]

- Increasing consistency and confidence in decision making
- Decreasing the risk of regulatory fines
- Improving data security, also defining and verifying the requirements for data distribution policies
- Maximizing the income generation potential of data
- Designating accountability for information quality
- Enable better planning by supervisory staff
- Minimizing or eliminating re-work
- Optimize staff effectiveness
- Establish process performance baselines to enable improvement efforts
- Acknowledge and hold all gain

Data security, data protection, and privacy

Toward data era, most businesses are concerned about data protection. Network security including firewalled approach is not a total solution since there is a lot of vulnerability at business data level. The ability to control the information and who can access that information has become a growing concern through Internet. These concerns include whether email can be stored or read by third parties without consent, or whether third parties can continue to track the web sites someone has visited. Another concern is web sites which are visited to collect, store, and possibly share personally identifiable information about users.

The advent of various search engines and the data mining technologies can create a capability for personal data which is collected and combined from a wide variety of sources easily. In order not to give away too much personal information, e-mails should be encrypted and browsing of webpages as well as other online activities should be done trace-less. Since everything is accessible over the internet nowadays, a major issue with privacy relates to social networking. For example, there are millions of users on Facebook and regulations have changed. People may be tagged in photos or have valuable information exposed about themselves either by choice or most of the time unexpectedly by others. It is important to be cautious of what is being said over the internet and what information is being displayed as well as photos because this all can be searched across the web and used to access private databases making it easy for anyone to quickly go online and profile a person.

Information about a person's financial transactions, including stocks, funds, debts, and purchases can be sensitive. If criminals access to information like a person's accounts or credit card numbers, that person could become the victim of fraud or identity theft. Information about a person's purchases can reveal a great deal about that person's history, like places he/she has visited, whom he/she has contacted with, products he/she has used, his/her activities and habits, or medications he/she has used. In some cases a company might wish to use this information to target individuals with marketing customized towards those individual's personal preferences, something which that person may or may not approve.

The internet users always give away a lot of information about themselves: Unencrypted e-mails can be read by the administrators of the e-mail server, if the connection is not encrypted (no https), and also the internet service provider and other parties sniffing the traffic of that connection are able to know the contents. Furthermore, the same applies to any kind of traffic generated on the Internet (web-browsing, instant messaging, among others). There has been interest in improving computer privacy through individualization. Researchers further improve each person's compliance with computer security and privacy.

In computer security, identity management is to enable the right individuals to access the right resources at the right times and for the right reasons. Identity-management systems, products, applications and platforms manage identifying and ancillary data about entities that include individuals, computer-related hardware, and software applications. It covers issues like how users gain an identity, the protection of that identity and the technologies supporting that protection (e.g., network protocols, digital certificates, passwords). Identity management includes information that authenticates the identity of a user, and information that describes information and actions they are authorized to access and/or perform. It also includes the management of descriptive information about the user and how and by whom that information can be accessed and modified. Managed entities typically include users, hardware and network resources and even applications.

Digital identity and digital identifier

Digital identity is an entity's online presence, encompassing personal identifying information. It can be interpreted as the codification of identity names and attributes of a physical instance. The use of digital identities is now widespread as the entire collection of information generated by a person’s online activity. This includes usernames and passwords, online search activities, birth date, social security, and purchasing history. In this wider sense, a digital identity is a version, or facet, of a person's social identity. In general, an entity can have multiple identities and each identity can encompass multiple attributes. In a model of digital identity, a given identity object consists of a set of properties. These properties contain information about the object for classification and retrieval. For example, a digital signature or software token is used to verify the identity. For purposes of information security, digital signature is simply stored, maintained, and retrieved. Identity management can be defined as a set of operations on a given identity model. User access control is typically used for a specific digital identity across applications. It simplifies access monitoring and verification. User access can be tracked from initiation to termination of user access. For an identity management process, the motivation is normally not primarily to manage a set of identities, but rather to grant appropriate access rights to those entities via their identities. Increasingly, identity management is evolving to control access to all digital assets, including devices, network equipment, servers, portals, content, applications and/or products. Its service requires access to extensive information about a user, including address books, preferences, entitlements and contact information. Social web and online social networking services make heavy use of identity management. Users decide how to manage access to their personal information.

Digital identifiers are the key used by the parties to an identification relationship to agree on the entity being represented. There are many different schemes and formats for digital identifiers. The most widely used is Uniform Resource Identifier (URI) which is the standard for identifiers on the World Wide Web. In the ICT world, a digital object has a machine and platform independent structure that allows it to be identified, accessed and protected. A digital object incorporates not only informational elements, i.e., a digitized paper, movie or sound, but also the unique identifier of the digital object and other metadata about the digital object. The metadata may include restrictions on access to digital objects, notices of ownership, and identifiers for licensing agreements, if appropriate. A distributed information system provides efficient, extensible, and secure identifier and resolution services for use on networks like the internet. It includes protocols, a namespace, and a reference implementation of the protocols. The protocols resolve the information necessary to

locate, access, contact, authenticate, or otherwise make use of the resources. This information can be changed to reflect the current state of the identified resource without changing its identifier.

Technically, to represent a digital entity, the party must trust that the claim of an attribute (such as name, location, role as an employee, or age) is correct and associated with the person or thing. Conversely, the individual may only grant selective access to its information, e.g. when one proves identity by authentication for payment at a web site. In this way, digital identity is better understood as a particular viewpoint within a mutually-agreed relationship than as an objective property. This nature of digital identity is referred to as contextual identity. In establishing the contextual relationship of identity attributes to one another, digital identifiers are able to represent identity in terms of pre-defined structures. This in turn allows computer applications to process identity attributes in a reliable and useful manner. XML (eXtensible Markup Language) has become a de facto standard for the abstract description of structured data. The digital identity solutions can interoperate taxonomically-diverse representations of digital identity. Free-tagging has emerged recently as an effective way of the identity of digital entities like bookmarks and photos.

Authentication is a key aspect of trust-based identity, providing an assurance of the digital identity. Authentication methodologies include the presentation of a unique object like a bank credit card, the provision of confidential information such as a password or the answer to a pre-arranged question, the confirmation of ownership, and more robust costly solutions utilizing encryption methodologies. Physical authentication techniques like iris scanning, hand printing, and voice printing are currently being developed to protect against identity theft. The strong authentication for online payment transactions links a verified person to an account, where such person has been identified in prior to account being opened. While automated face recognition, tagging, location tracking, and digital authentication systems become easily associated with identity, privacy may be lost. An identity system should provide privacy and enhance security for digital services and transactions. Thus, authentication system allows individuals to verify their identities to others without revealing to them the digital representation of their identities.

Authorization is the determination of any entity that controls resources. Authorization depends on authentication, because authorization requires that the critical attribute of the authorizer must be verified. For example, a database management system might be designed to provide certain specific individuals who can retrieve information from a database, but cannot change data stored in the database, while allowing other individuals the ability to change data. When a person rents a car and checks into a hotel with a credit card, the car rental and hotel company may request authentication that there is credit enough for an accident or profligate spending on room service. Thus a card may be refused when trying to book the hotel, though there is adequate credit to pay for the rental and the hotel. Then when the person leaves the hotel and returns the car, the actual charges are authorized.

Three dimensional data format

3D data has a lot of applications like virtual game, interactive video conference, 3D movies, computer-aided design and augmented reality/virtual reality, etc. 3D animation and 3D navigation are also emerging areas. 3D applications are also used for medicine, structural engineering, automobile, construction, military, and cultural heritage, and so on.

The 3D data format consists of 3D contents, 3D file formats, and 3D viewers. 3D contents could be classified into three categories: geometry, image, and scene. The 3D geometry model is for graphic applications which render a set of 3D points and series of polygons. The 3D image model applies two dimensional images to mapping the corresponding 3D vertex. The scene model is regarded to the camera. The 3D vector of the camera indicates the position on spaces and the direction of light sources including colour/intensity. The 3D file format is used for storing 3D data and various

software packages for viewing 3D data. 3D software is designed for viewing and converting 3D data files.

There are a list of 3D file formats depending on application types:

- 3D animation, 3D image, 3D drawing, and 3D computer graphics
- Interactive 3D game applications
- VRML (Virtual Reality Markup Language) and X3D (Web 3D Technology) [92]
- 3DMLW (3D Markup Language for Web) and 3DXML (3D XML) for 3D web sites
- Database files including all scenes, objects, meshes, and textures
- Digital content creation for Autodesk and 3D studio
- 3D geometry for polygon
- 3D stereolithography for CAD and medical application
- 3D manufacturing including 3D printing and additive manufacturing

Many 3D file formats are currently available, which should be classified according to categories and applications. The relevant standards can reduce the conversion overheads.

How to store and search large volumes of data

In future zeta byte era, the large volumes of data are stored at widely distributed cloud system. The storage infrastructure is designed to store, manage, and retrieve massive amounts of data. Big data storage enables the storage and sorting of big data in such a way that it can easily be accessed, used, and processed by application and services. It supports storage and input/output operations on storage with a large number of data files and objects. A typical data storage architecture is consisted of a redundant and scalable configuration of direct attached storage (DAS) or clustered network attached storage (NAS). The storage infrastructure is connected to cloud servers that enable quick processing and retrieval of big quantities of data. Moreover, data sets grow rapidly because cheap and numerous mobile devices, remote sensing, software logs, cameras, microphones, RFID readers, and wireless sensors are increasingly gathering the data.

The flexible and scalable storage platform empowers people to access, query and manipulate the data to transform it into with metadata information. Metadata gives some emphasis on the content, meaning and value of information over the media, type and location of data. Data storage platform enables people to take a single, unified approach to managing data across large, distributed locations which includes the use of content and metadata indexing. Data storage platform virtualizes aggregate file systems into a single namespace. The file, full text index and custom metadata is collected and stored in a distributed metadata repository. This repository is leveraged to enable speed and accuracy of search and discovery, and to extract value leading to informed business decisions and analytics. Data storage platform can be globally distributed through the unique namespace, eliminating data silos and improving storage utilization.

The cloud computing provides shared computing resources including storage. The cloud system is enabling ubiquitous, on-demand access to a shared pool of configurable computing resources, which can be rapidly provisioned and released with minimal operating cost. Cloud storage platform provide users to store and share their data across the world.

Key issues on data analytics

Understanding the process of data analytics is the first step towards understanding the data-related challenges. Data analysis can be broken into six steps, each often performed multiple times when trying to solve a single problem:

1. Identifying Opportunities
2. Getting Data
3. Exploring Data
4. Preparing Data
5. Analyzing Data
6. Applying Results

The first and the sixth steps refer to understanding the contexts in which analysis is performed: discovering problems; setting goals; making decisions; and so on. These are context-dependent and business-oriented. The technical side of data analytics is described in steps 2 through 5.

- **Getting Data**

  Translating the problems and opportunities which are identified in Step 1 is currently more of an art than a science. Traditionally, data scientists might brainstorm what kind of data could solve the problem and then try to find it or look at the data they already have and brainstorm how it could be used to solve the problem. If the data scientists have a lot of time and money, they might dream up the perfect data that could solve their problem and set about trying to generate it.

  With the advent of big data, however, an increasingly wide variety of data types can be used to solve an increasingly wide variety of problems. Common data types include SMS data, web usage data, sensor data, weather data, traffic data, energy data, and so on. In the future, computers will likely take a much more active role in experimenting with different types of data that help solve the problem.

  In the meantime, here are some of the characteristics data scientists commonly look for in potentially useful datasets:

  - (**Quality**): The data should be high quality. Bad data gives bad results.
  - (**Cost**): Purchasing the data can't be too expensive or time-consuming.
  - (**Size**): Something too small might not be useful. Something too big might be too unmanageable.
  - (**Cleanliness**): Datasets that don't require a lot of cleaning are preferred.

  Once a good dataset is identified, it needs to be gathered and stored on local machines. This involves interacting with physical devices like hard drives, or web-based services like APIs, or through a particular communication channel on a specific sensor network. Which data storage system to choose depends on factors like access speed and cost.

- **Exploring Data**

  After getting data, people must examine the data to verify that it is what the users expect and start planning their own analysis strategy. Exploring data might include checking simple characteristics like the number of columns, and statistical characteristics like the distribution

and range of values in each column. This typically involves characterizing individual variables in the data and the relationships between multiple variables, often through some form of visualization.

Exploring data is actually analyzing data roughly. In many cases, valuable insights are obtained during the exploratory process.

- **Preparing Data**

  Preparing data involves using the knowledge gained during the previous steps to prepare the data for analysis. This typically involves converting the data to different formats, dealing with missing values, reducing the dimensionality of the data, normalizing the remaining data, extracting features, and changing the values of variables in particular examples.

- **Analyzing Data**

  This is where the real value is extracted from data. Traditionally, people manually involve this step by using statistical methods to assess the relationships between variables, often using spreadsheet tools like Microsoft Excel. Recently, however, machine learning has begun dominating this step in the analysis process. The widespread impact of machine learning algorithms has raised fears about the impact that artificial intelligence algorithms will have on society.

6.2 Key requirements of future digital data format

Currently of telecommunication, broadcast, and internet/web applications, large variety of new data formats may be emerging. Also, different and heterogeneous metadata concepts and formats from consortia/fora or individual organizations are investigated without consensus of future data eco-society. Data interoperability and data format conversion will be one of urgent issues to be solved at near future. Moreover, other industries like energy, transportation, health, biology, and geography are developing their own data format. For future convergence markets including IoT applications, the future data formats including metadata should be well defined and standardized both for ICT industries and other industries.

For future digital data format, the key assumption is easy to search or discover. The document should be produced to be easily searched or found without any additional skill. The metadata helps people find out the correct documents among large volumes of data files. The management data is also essential if the created documents are used for business. To get initial consensus, the following generic requirements of future digital data formats may be concerned.

(**Raw Data Format**) Until now, various data formats are already used in the real world. All the data formats are acceptably tuned with their own objectives for creation, delivery, processing, storing, sharing, and distribution. The people have to understand their specifications for reading or writing the raw data files since all the data formats have their own rules and syntax. Additionally, people also have to understand the corresponding protocols while transferring, processing, converting, sharing, and distributing the data file, which are depending on hardware types of devices, network, and storage as well as software environments such operating systems, database, and application platforms. The future data applications for convergence of energy, transportation, health industries as well as IoT applications need to create a lot of new data formats. Too many new specifications

may be published to handle data formats for future convergence applications. People feel some difficulty while they have to understand new data specifications and the relating protocols.

Here, the outstanding issues are how to minimize the understanding of data specifications and protocols while handling the data. The generic requirements of future data formats include as follows:

- Minimize the interpretation and understanding of data formats and syntax
- Minimize conversion and translation overheads among data formats
- Minimize the dependency of hardware types and software environments including encoding and decoding technologies

To meet these generic requirements, the well-known data formats data are recommended, which are easily understood or interpreted by both human and computing system. It means that future data formats should be well defined without complicate specifications or additional explanation. Currently, the XML/RDF data format is self-descriptive without any additional explanation. Additionally, the existing data format can be converted to XML/RDF forms. Otherwise, if future data formats have similar rules and syntax with natural human language, it is more easy for people to understand new data specifications.

(**Descriptive Data or Metadata Format**) For searching the data files or data sources, they should be well identified by their physical or virtual locations. To find out the right or correct contents, the data sources should be described what kinds of contents are included. Also, the data encoding format should be declared. To solve these issues, the data sources should be well described and the corresponding metadata contain the semantic information for identifying, processing, and managing the data files. Some data file or sources may include tag or index information which is linked by metadata information and/or detected by searching machine.

Currently, there are many different metadata standards for specific purpose, specific domain or particular types of data. The metadata specify the meaning of data sources, document format, and representation rules. Additionally, many different metadata schemes are developed according to applications such as e-commerce, education, science and engineering, etc. At future convergence applications, people with domain-specific or industrial-specific knowledge need new metadata format. Similar with raw data format, the outstanding issues are how to minimize the metadata format without any additional interpretation. Therefore, the generic requirements of metadata format include as follows:

- Well specified without any confusions and mis-interpretation
- Minimize the interpretation and understanding of metadata formats and syntax
- Minimize the searching, sorting, classification capabilities of raw data file or data sources
- Be flexible or future safe on technology developments toward future knowledge society
- Optionally, the descriptive metadata is located at inside data file or linked to other separate forms

To meet these generic metadata requirements, XML/RDF schema for semantic web is safe as one of future metadata format. But, the complexity of current XML/RDF schema should be solved. The

metadata specification is tightly aligned with the usage of raw data format, especially for level of intelligence on target applications.

(**Management Data Format**) Most of data files may assume to be open to public. But, some data materials are applicable for specific business or markets. In this case, data may include privacy information as well as market sensitive information. Some data may need data security and protect copy right. But, many different security and copy right protection mechanisms are widely developed and used at current digital market. Some mechanisms are belonging to specific market and applications. For future convergence applications, multiple security and protection technologies should be developed. Therefore, the generic requirements of management data format include as follows:

- Minimize the data field format for management purposes both in raw data format and metadata format
- Be convertible or interchangeable between the original data sources and encrypted data sources for management if the data and metadata information are designed to be open to public.
- Minimize the interpretation and understanding of management formats and syntax

Technically, the outstanding technical issues for standardization of future data formats are described as follows.

High-level requirements for data classifications and data format

- Classification rules and principles: according to layer, application, property, and others
- Mapping of existing data format according to classification (backward compatibility)
- Define and classify raw data, descriptive data, and management data
    - Data with/without tag or index information
    - Index or summary metadata linked with raw data
- Classify data from single sources and multiple sources (by aggregation, multiplexing, package, summation)
- Classify homogeneous and heterogeneous forms of data
- Classify structured data and unstructured data
- Classify file type data and channel/stream type data
- Classify n^{th} tier data formats (i.e., data → information → knowledge)
- Classify according to environments of data creation, storing, processing, delivery, and consuming (to minimize format conversion)
    - Data creation devices (e.g., camera, microphone, sensor, office package, controller)
    - Data storage systems (e.g., memory, hard disk, magnetic tape, USB memory)
    - Data processing platforms (e.g., PC, smart device, or cloud system)
    - Data delivery protocols (e.g., Internet, wireless, MPEG, and USB)
    - Data output devices (type of layout, screen size)

High-level requirements on data applications

- Web and Internet applications (publication, news, entertainment, e-commerce, banking)
- Audio/video applications
- Geographical applications including 3D map and logistics
- Virtual space applications including virtual game, augmented reality and virtual reality
- IoT application (e.g., monitoring, sensing, and alarm)
- Data analytics applications (e.g., machine learning, big data processing, large cloud system)

High-level requirements of data identification

- Rules and Principles according to levels (e.g., component, system, aggregation, grouping/package, organization/company, and global association)
- Uniqueness and coverage (e.g., within area or domain, company, community, nation-wide or global)
    - Public or international authority operates public identity management system which is mandatorily registered by individuals or organizations (e.g., personal id, health insurance id, credit card no., or bank account no.)
    - Individual organization manages their own identification system (e.g., username/ password, employee id., group code no., system id, product code no, component id.)
    - Service provider manage customer information (e.g., user name, personal id, patient code, account no.)
- Usage of URL/URI/URN for data identification (that is, the related or additional information freely and arbitrarily contained in web sites)
    - Combination of URI/URL/URN with relevant search engine may not need any specific identification rule
- Dependency of search engine and discovery mechanism
    - data discovery, service discovery, application discovery
    - lookup table, database, distributed storage, yellow books, or whole web site
- Update and Life time of data identification
    - One-time usage like SSO (Single Sign-On)
    - Subscription and registration time with period
    - Temporary assignment during visited period
- Identification structure
    - Independent or individual id structure
    - Hierarchical identification structure (e.g., company code – department code – personal id)
    - Interlinked id structure (e.g., product/sale code – manufacture code – component code)
- Interoperability or backward compatibility of existing identification and service repository (e.g., GS1, RFID/USN, EPCglobal, ISBN/ISSN, EUI-64, UDDI)

High-level functional requirements for data processing and functional capabilities

- Minimize processing costs on creation, delivery, filtering, sorting, and processing
- Flexibility and scalability on data format conversion and header information
- Human readability by word, name, summary
- Machine readable and processable capabilities
- Binding mechanism of linked data: raw data + descriptive data (metadata) + management data (identifier, name, index)
- Overhead of tag insertion on raw data (what words, images, fields, and regions in a file)
- Relationship between tag/index information and search/discovery mechanisms
- Processing overheads of metadata and management data
- semantic ontology
- relationship of application software for triggering, decoding, understanding, interpretation, explanation

Requirements of metadata format

- Classification of metadata: procedural, descriptive, administrative, technical, or business transactional
- Schema including attribute and property in the data resources (e.g., date, origin, size, format, version, authors)
- Schema depending on applications (e.g., audio/video, photo, music, art, file, book, image, map, telecommunication, game, or healthcare)
- Direct or Indirect linkage or alignment between raw data and metadata
- Granularity or depth in description of metadata

Requirements for data storage and search

- How to collect and store documents from million sources (upload time for whole or prefix portions of sources), current SPARQL or NoSQL should be reviewed
- How to provide the linked or connected information (just URL only) of documents in case that owner does not to upload or send the original documents or live sources
- How to bind the output screens from multiple sources of documents or millions of live channels
- How to render output screen from multiple sources of IoT sensors and presence information
- How to search for the tightly or loosely related documents (with/without hyperlinked information)

Requirements for data sharing and fast analytics

- Usage of webs site or web pages for sharing
    - Open to public
    - Open within communities by access procedure like username and password

- Searchable or discoverable data by search engine and discovery mechanism on the web
- Standardized machine readable and sharing data format
- Platform- and application-dependent (e.g., only with relevant software or access procedure)
- Enabling convergence services and creating new businesses
- Minimizing resource usage (channel, memory and CPU)
- Metadata for sharing data: To help searching, gathering (combining), and analyzing

Requirement, mechanism, and technologies for 3D data format

- How to orchestrate multiple sources (input) and multiple screens (output) for rendering (like orchestration by using music sheets)
- Orchestration of various input/output devices in time and space
- Generation time of each input sources (with type of media, quality, time, location, synchronized other inputs, metadata, additional management information)
- Rendering of each document (file, AV, text, etc.) with type of output devices, resolution, size, time, location, angle, other types of synchronized devices)
- How to insert or provider timing information (like music sheet) to a lot of rendering devices with mediation (increase amplitude or remove some input sources)
- Location of input sources are widely distributed, recognized by URL/URI
- Real-time cloud platform arranges multiple live sources with intelligent cache or storage
- Each output devices can synchronously render the live materials or non-real time documents at certain point of window screen in alignment with other screen

Requirements for data analytics

- (**Tracking Data Value**) How to trace the data value chains from data sources to data users
    - Is it possible to trace the value chain of data from sources to analytics process?
    - Incentive mechanism for data donor
    - Tracking data value chain
    - Is it possible to support data market?
    - **Ownership**: how can people track who owns what data (and identify legal rights)?
    - Can people define the quality of data?
    - What would be common category for quality? (Define quality dimension?)
    - Timeliness(version), integrity, accuracy, purpose-dependency
- (**Data Fusion**) How to define a data interoperability based on unified metadata for data search and fusion
    - Is it possible to define a unified metadata format (or schema) for easy data search and fusion?
    - **Revision history**: how can we track who modified the data in what way and when?
    - Can users set any restriction to data usage or analytics?
- (**Data Reduction**) How to reduce data traffic at the source
    - Can network support “edge analytics” functionality generally to reduce traffic?

- Supporting distributed analytics
- Is it possible to reduce traffic due to redundant copy of similar data (99.9%) for analytics purpose?
- Can people prevent useless copy of data?

6.3 Key Strategies of future digital data format

Key features for future digital data

The future data format with open accessibility and website usability are important to enable content sharing, creating ideas, and accumulating collective intelligence of people. The web-based cloud platform provides the efficient way to share and store documents. The emerging features of future data formats will be based on web technologies as follows:

- (**Voice**) The voice data formats representing speech dialogue and speech recognition are defined in a form of XML. With voice interface at the web document, it is possible to create new interactive voice applications like a voice browser.
- (**Video**) The high quality video data format is mainly designed for efficient transport and layout to the screen. For searching large amounts of video streams, the metadata information is useful for query. The tagging technologies over video data format are needed to clip out a specific portion of a video stream. For augmented reality and virtual reality applications, the location information of video sources is delivered to align with screen positions and angles. Additionally, the timing information for creating and rendering video streams are also used to synchronize multiple video streams over target screens. The timed metadata information in alignment with video streams has to be concurrently delivered.
- (**Image**) The images and animation data format can be defined for interpretation and suitable for user interaction. The tagging and indexing technologies are needed to search for image data. The images or symbols with textual information are indexed by the search engines. In addition, the extraction of metadata information from images and symbols is used to represent color code and indicate real physical components connected to animated symbols. A portion of images may contain metadata tag information for a specific query.
- (**Table**) To lay out the data with table format, the techniques for indexing and analyzing tables are investigated at the web format. The sequence of strings of columns and rows in the table structure can appear in a graphical form.
- (**Graph**) The graph representation in the web data format is used to deliver the logical structure of tasks, algorithms, functionalities, or heuristics. The graph model for HTML documents includes the tree-structured hierarchy when parsing the tags. To connect nodes in the hierarchy of a graph data, there are the incoming/outgoing links for query and process. The hyperlink in a graph data is used to distinguish nodes of external references.
- (**Index**) The database or the directory would be well formatted and indexed. The string of texts in the database is tagged or hyperlinked to the specific URIs of the web. Some images in the 3D database are linked with real geographical locations.
- (**Semantics**) Technically, the semantic web with tags will be coming to mark up semantics on HTML and XML/RDF as well as traditional book-like documents. Most contents have various ontology/XML standard formats, which are stored in databases with index.
- (**Multimodal**) The multimodal interface of future web documents is one of the outstanding issues to be solved in the near future. The context-aware and presence information can be

extracted by the combination of IoT applications. The data streams from multimodal interface are mashed up with those of IoT applications.

- (**Language**) For exploiting knowledge from data on the web, the integration of XML technologies is used for natural language processing. With syntactic and semantic analysis of language, the self-explaining XML tags can be used to recognize concepts and extract knowledge from the data files.

To cultivate future digital data format, there are many ways for creating, delivering, and rendering data files. A metadata description is used for the composition of data sources (character, shot, scene) with other data objects (text, sound, image, etc.). The composite data formats from multiple sources have more complex and sophisticated presentation. For example, a character in a video stream is introduced by displaying a textual description when that character occurs. A word in a text sentence is highlighted when an audio plays out. A hyperlink or tag information is set on a video object or on a particular region of an image. A start time of the multiple video streams is used to synchronize the word in the texts and time location of word pronunciation among multiple audio streams. The timing information of the video objects is used to coordinate the related image regions especially for augmented reality applications. The structured data format with descriptive metadata will include the context-aware information available for composition process. A structured data format contains not only raw data, but also a hierarchical description of data. Many data formats have more complex presentation scenarios and require more flexible presentation services including interactions. For the future data formats with XML description, the web browsers can implement the temporal and spatial models to present the documents.

Key Strategies of data analytics

For the exploration of data analytics, three key issues have been identified: data value, data fusion, and data reduction.

Data analytics is the study and practice of extracting value from data. However, most of the value extracted from data is concentrated into the hands of a few. The tools to ensure the benefits of data analysis might be fairly distributed to all stakeholders. Extracting value from data almost always involves combining different types of data from multiple sources. However, current technologies and industry practices make this a complex and time-consuming process. It is hoped that people can easily share and combine different types of data from different sources. Additionally, the amount of data that people want to communicate, store, and analyze is continuing to grow exponentially. However, current technologies will probably not be able to meet the growing processing time, memory, and communication bandwidth requirements. The footprint of data can be reduced to a manageable size without losing too much value. Figure 27 shows the three key issues of data analytics.

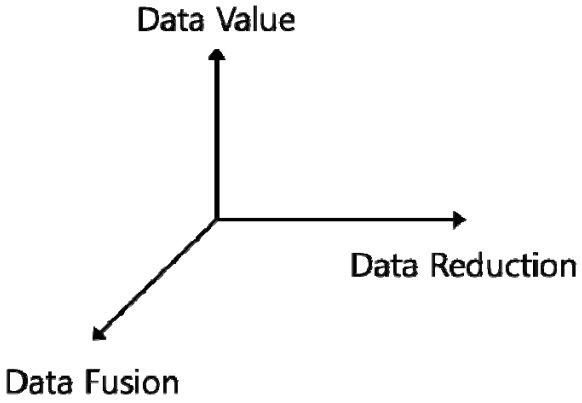


Figure 27. Three key issues facing data analytics

Figure 28 shows the relations between these three issues. Data fusion will deliver extra value from data. Data reduction will help to extract value from data efficiently. Additionally, data reduction will make finding and combining data from multiple sources easier.

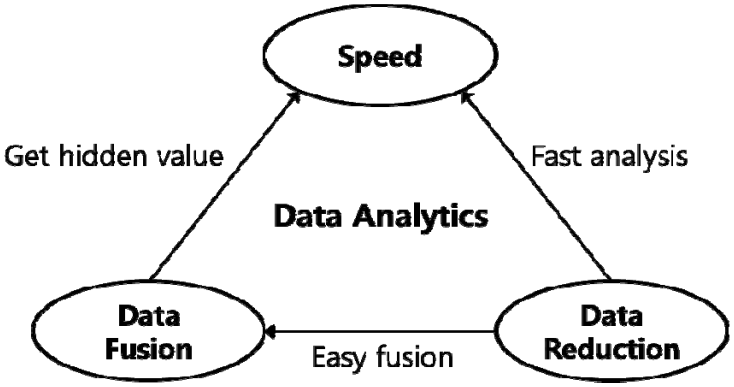


Figure 28. The relationships between three key issues of data analytics

Three key issues of data analytics can be translated into three key requirements that need to be satisfied to usher in a truly data-driven world that is fair and equitable to all:

- (**Data Value**): able to track the value extracted from data and devise mechanisms that support the fair sharing of the values amongst all stakeholders.
- (**Data Fusion**): make it easy for people to share many types of data with others for different purposes.
- (**Data Reduction**): make extracting values from data more efficient when it requires resources like processing time, communication bandwidth, and memory.

Data value as key requirements of data analytics

The key requirements of data value are able to track the value extracted from data and devise mechanisms that support the fair sharing of that value amongst all stakeholders.

- **Estimating Value**

  Value means many different things to many different people. A salesperson might care about increasing the number of refrigerators sold, increasing the satisfaction of the customers purchasing refrigerators, or decreasing the amount of effort it takes to sell each refrigerator. A doctor might care about increasing the lifespan of the patients, decreasing the number of occurrences of a particular disease, or decreasing the cost of each intervention. A newspaper editor might care about increasing the total number of readers, increasing the amount of each article the readers remember, or decreasing the amount of factual inaccuracies in each article.

  Data can help achieve all of those goals and more. Specifically, data could be analyzed to help achieve those goals.

  Data about what goods were sold to whom, what sales pitches were used, and how satisfied the customer was six months later can help the salesperson decide which products to recommend to which customers and which pitches to use during those recommendations.

  Data about causes of death or about the details of patients with particular diseases, or about the details of all instances of a particular intervention across multiple hospitals can help the doctors decide what lifestyle changes to recommend, understand what causes particular diseases, and understand what parts of the intervention need to be changed to drive costs down.

  Data about what articles are read by what people, what those people are interested in, how long each person spent on the article, and how well they performed on recall test months later would help the editors decide what articles to advertise to what people and what edits to recommend to the writers.

  Unfortunately, quantifying just how much analyzing the data helped achieve those goals is often difficult. Maybe the increase in sales made by the salesperson is because the demands of customers are changing, or because customers have more money to spend, or because a massage parlor opened nearby and relaxed customers are more likely to buy things. Maybe the doctor's patients live longer because the amount of pollution in the air decreased, not thanks to any lifestyle changes the doctor recommended. Maybe the newspaper readership increased because the newspaper was spontaneously endorsed by a celebrity, not because the newspaper got better at matching articles to potential readers.

  Additionally, the metrics measured might only be poor approximations of what people really care about. How do you measure the value a refrigerator adds to a customer's life? Or the health of a patient? Or how informed a population is? Or how much good you're causing? Or how much suffering you're preventing? Maybe we care less about the amount of money in the bank and more about the amount of happiness in our lives. But how do you measure happiness?

  Measuring how much achieving a goal is worth is extraordinarily difficult, as is measuring just how much a particular sequence of actions helped achieve that goal. This makes measuring the value of a particular instance of data analysis difficult.

  To make matters worse, quantifying the value of a particular dataset is even more difficult since the value of a dataset comes from what analysis is performed on it [93]. But one dataset might be analyzed many times, to help achieve many goals, over its effectively infinite

lifespan. Nowaday, the cost of data is now so cheap that anyone who wants to save a particular dataset can do so almost indefinitely for an almost negligible cost.

All of these factors mean that it is impossible to quantify, with 100% accuracy, exactly how valuable a particular dataset is at any point in time. However, decisions still need to be made about what data to generate, what price that data should be sold at, and how to compensate all stakeholders fairly and equitably. This means that the accuracy of the mechanisms might be continually improved to estimate the value of data.

- **Sharing Value**

Currently, the person who analyzes the data is the one who gets the value from the data. The analyst typically purchases (or gets for free) the data and the tools used to do the analysis for a fixed upfront cost. If the analyst spent hundreds of dollars on the tools, but did analysis worth millions of dollars, none of that extra benefit makes its way back to the people who generated the data or the people who developed the tools used in the analysis.

There are many online marketplaces like Axiom [94] and InfoChimps, as well as public data archives like data.org and archive.org [95], where people can obtain data. In most markets, the supplier would eventually realize the true value of the data and increase the price. However, because the value of the data is difficult to estimate and there are currently no mechanisms for the data owner to track the values extracted from the data over time, realizing this "true value" is almost impossible. This situation favors the data purchaser much more than the data owner.

Because the value extracted from the data that lies almost entirely with person who analyzes the data, many people are reluctant to share the data they've generated. After all, generating and releasing data is a lot of effort, and why should they give that efforts away for free or for what is, in the long term, a low price? They could just hold on to the data and eventually do the analysis themselves.

There are more stakeholders than just data owner and data purchaser. Depending on the situation, the stakeholders include the data generator, the sensor designers, the data manager, the data seller, the data buyer, the data analyzer, the government regulating everything, the citizens in the society in question, and all humankinds.

- **Tracking Value**

It is currently impossible to track the total amount of value that a particular dataset generates over its lifetime.

Data owners could hire data scientists to analyze the data and see exactly how much values can be extracted from their data. However, not every company employs data scientists or even knows how to hire good data scientists, and much of the values of the data depends on contexts in which the results of the analysis will be used, which only how the results of the analysis are used, which depends on the contexts, which only the end user knows, and thus the figure the data owner's data scientists calculate might be different from the figure the data buyer's data scientists calculate.

There needs to be some mechanisms. A data value chain model can track the values generated by specific datasets over time. This is the first step towards ensuring that those values can be fairly distributed to all stakeholders. This will help everyone involved understand why data can be valuable, what data is valuable, roughly how valuable that data is, and how to extract that values from data. This information is needed so people can guide conversations about how that values should be shared amongst different stakeholders. Overall, this should encourage more data owners to share their data for others to analyze, as some of the values extracted by others would make its way back to them.

Data Fusion as key requirements of data analytics

The key requirements of data fusion are to make it easy for people to share many types of data with other people for different purposes.

- **Common Difficulties**

  Extracting values from data almost always involves combining different types of data from multiple sources. However, because many people want to hoard their data in their respective organizational silos, it is difficult to discover, explore, and work with data that people does not already own [96]. this is one of the key bottlenecks to achieving a truly data-driven world [97,18].

  There are a number of reasons why this is the case:

  - (**Everyone comes up with their own custom data formats**): Often each data owner comes up with his or her own way to structure data, particularly metadata. The lack of commonly accepted industry standards makes ingesting someone else's data difficult.
  - (**Infrastructure is often incompatible**): The tools one person uses to manage data (e.g., databases, compression algorithms) may not be compatible with the tools another person uses. Conversion is often difficult to perform and often gives imperfect results.
  - (**Finding data takes a long time**): There is no commonly used search engine for data. It is impossible to know all of the data that is out there, and people might never know if someone in the building next door has the data users need to solve their problem.
  - (**Getting data takes a lot of effort**): Often the bureaucrats who make the decision to release the data does not understand all of the legal and ethical nuances of handling data. Privacy concerns, sometimes unjustified, often block any attempts to share any data.
  - (**Checking data takes a long time**): Often understanding if data might be worth fusing requires downloading the entire raw dataset and exploring it. Often this process ends with the realization that the dataset is not useful.
  - (**Tracking who has what data is difficult**): Data is property, though it's not always clear whose property it is.

  Many of these problems can be solved with good, consensus-based standards. Many of these problems need to be solved before people can truly usher in a data-driven world.

- **Use of Metadata**

  Sharing raw data consumes a lot of resources. Instead of sharing raw data, people often share metadata to people who may be interested in the full data. The metadata is meant to be enough for the users to understand what the raw dataset contains. The users can then decide if they want to access the raw data and begin the actual analysis. This supports the freemium model well: metadata is free to access but raw data requires payment.

  Metadata means many things to many people, from simple descriptions to detailed summarizations of the raw data. In many instances, the structure of metadata is based on the Dublin Core Metadata Initiative [99]. However, these standards are far from universally accepted, and few of the existing standards are designed for cross-domain applications that involve data fusion [100].

As an example, tags are a common feature of many metadata structures. Tags are user-defined descriptions that help users sort, search, and explore many files or datasets at once. Specialized, pre-defined tags can be invented for things like privacy control [101]. When used effectively, tags can help people easily analyze related data (e.g., analyzing multiple CCTV recordings simultaneously to find a person moving across the cameras) and cross-domain data sources (e.g. checking nearby public transportation and event schedules to guess why the person is moving in that direction). These tags, and other forms of metadata, help data users decide how to preprocess the raw data; data owners promote their data on the data market for free or for a price; and data analysts find partners to perform mutually beneficial data fusion.

However, there is no commonly accepted method of tagging: the tags are often invented on the spot by the persons responsible for managing the data. This means the tags of one dataset are often unrelated to the tags on another dataset, which makes using tags to identify related datasets much more difficult. This applies to many other forms of metadata. As a result, company and institutes often design their own metadata format, which makes sharing and understanding metadata difficult.

- **Features of Good Metadata**

Good metadata by itself should be enough to understand the main characteristics of the raw data. It should have similar structure across many different types of data so analysts can easily examine metadata of many different data formats to decide what data to combine. Good metadata might contain components like summaries, indexes, and tags that convey information suggested by standards like Dublin Core as well as additional information like how to access the data, how much of the data is available, and whatever else the data owner wishes to share.

The metadata might have the following items:

- Author/Contributor
- Date/Location
- Title/Description
- Format/Size/Data type
- Revision history
- How much of the data is available (just metadata, a sample or full data)
- Link to a representative sample of the raw data
- Link to the raw data
- Information on how the summary is structured
- The summary

Good metadata can also contain information specific to different data formats. For example:

- Metadata for numerical data could include max, min, average, quartile values; outliers and missing values; and graphical representations of the data.
- Metadata for text data could include the title of the document, the entities in the document, a text summary, and the captions of pictures.
- Metadata for video data could include entities in the video, object descriptions, object appearance statistics, and video summarizations.
- Metadata for sensor data could include contextual information, the purpose of sensing, and more.

Data Reduction as key requirements of data analytics

The key requirements of data reduction are to make extracting values from data more efficient when it comes to resources like processing time, communication bandwidth, and memory.

- **Backgrounds**

  The amount of data people wants to communicate, store, and analyze is continuing to grow exponentially. However, current technologies will probably not be able to meet the growing processing time, memory, and communication bandwidth requirements. The footprint of data should be reduced to a manageable size without losing too much value.

  To help reduce the footprint of data, the first option involves adding analysis capabilities to sensors to make decisions as the data is being generated. The second option involves performing preliminary analyses to summarize key characteristics of the data to help the central server prioritize what data to analyze.

- **Edge Analysis**

  Traditionally, sensors generate data and then send the entire data to a centralized server to be stored and analyzed. This can be visualized as a network of devices with many sensors on the edge of the network all connected to centralized servers in the middle of the network. Some analytics is performed at the edge of the network, either on the sensors that generate the data or in a local processing hub. The analysis result could be used to decide what, if any, commands should be sent to what other nodes in the network. For example, a motion sensor that notices a car pulling into the driveway might tell a smart garage door to open without first checking with a centralized server.

  Edge analytics has become a popular architecture for networks of connected devices. Recently IDC Future Scape for IoT report estimates that by 2018, 40% of IoT data will be stored, processed, analyzed, and acted upon at the edge of the network where it is created.

- **Preliminary Analysis**

  Sometimes the edge node might not be able to make a decision on its own. Sometimes it might make a decision but want the centralized server to tell it what it should have done based on the data it saw. Sometimes it might decide not to do anything but tell the centralized server anyway just in case. In each of these cases and many more, the sensor still wants to send its data to a centralized server.

  The centralized server, however, is receiving many different streams of data from many different sensors. It needs a way to quickly triage which data streams to pay attention to and analyze first. Analyzing the raw data is inefficient. Instead, the sensors can summarize the raw data and send that summary with suggested priority levels. The centralized server can then check the summary and adjust the priority level depending on the contexts.

  Imagine a security camera constantly generating audio/video data of a bank lobby. Instead of sending the raw data to the centralized server to be processed, the security camera could perform computer vision analysis and identify objects, faces, postures, and actions. It could then interpret these discoveries (“Nothing’s happening.” versus “A lot of action!”) and send those interpretations, along with metadata like timestamps and capture location, to the centralized server. If the message is “no entities discovered”, the message might be the lowest priority. If the message is “potential robbery in progress”, the message might be the highest priority. The centralized server would then look at the messages in order of priority and decide whether to ask for more data, whether to continue as otherwise planned, or some other actions.

There can be many different levels of summaries. For example, the "no entities discovered" message could just be an empty message. However, in the case of the "potential robbery in progress" message, the centralized server will probably want to see the raw data to do things like send it to a human to double check the alert. This message might thus contain both the "potential robbery in progress" message and the raw data.

Figure 29 shows some potential preliminary analysis disclosure levels. The raw data can disclose in three levels: (1) a summary of the data; (2) a sample of the data; and (3) the full raw dataset. Each level of disclosure could then be accompanied by descriptive statistics and visualizations. Table 12 summarizes the key aspects of these three disclosure levels as well as the visualizations and descriptive statistics that might accompany the disclosure.

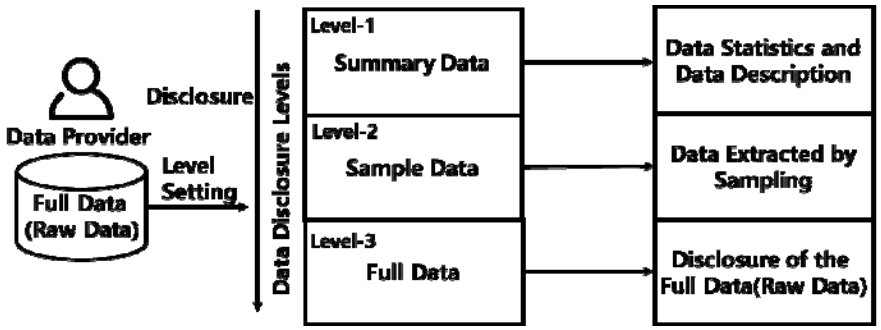


Figure 29. Three disclosure levels of a potential preliminary analysis system

Table 12. Descriptions of the three levels of disclosures as well as the type of a data visualizations

Disclosure Level	Data	Description
1	Summary Data	Information used in data mining tools to extract values from the data like the distribution the data follows most closely, minimum value, maximum value, intermediate value, average value, and range of data values
2	Sample Data	Sample extraction methods include random extraction, systematic extraction (collect every Kth data point), probability proportion extraction (extraction method reflecting the distribution of data), floor sampling, cluster extraction
3	Raw Data	The entire raw dataset

Quality of Service in the aspects of data analytics

Telecommunications services are traditionally composed of devices, transmission, switching, and application software that allow users to communicate data in the form of bit strings between different nodes on communication networks like the telephone network or Internet. Data users, however, are much more focused on doing something with the data once they have it than they are on simply having data. As the nature of how data is used is changing, so telecommunications services change to take into account new needs of data users.

Quality of Service (QoS) is the metric traditionally used to define the minimum bandwidth, delay, and latency requirements for each service. The QoS metric helped communication service providers make sure their services are performing adequately. For example, real-time voice conversations require fewer than 200ms delays. However, most telecommunication channels can satisfy this requirement easily, and additional improvements are not that noticeable to human ear. Additional features are needed to differentiate telecommunication services.

Users now care about things like:

- (**Data Value**): Is it possible to track how much value a data has generated?
- (**Data Quality**): Is the data likely to be valuable to the data users?
- (**Data Analytics**): Can the data be analyzed easily?
- (**Data Fusion**): Can the data be fused with other types of data easily?
- (**Data Reduction**): Does the data have a good utility to footprint ratio?

- **Quality of Data**

  The Quality of Service (QoS) metric assumes "the more data the better". As the amount of data continues increasing, however, it becomes more and more important to make sure that the most valuable data is analyzed first. We need a metric that assesses the value of data to help decide what data should be analyzed first.

  However, it is difficult to evaluate the value of data in advance because the value of the data depends on contextual information like who used the data, how they used the data, when they used the data, and what they used the data for that can't be known until after the analysis is done. However, it is still necessary to estimate, however roughly, how valuable data is expected to be, so that analysts know what data to analyze first.

  We call this estimate the Quality of Data (QoD). In general, quality data is expected to become valuable data. However, the quality metric for any particular dataset might change depending on the stated purpose of the analyst, the other data that can be combined with it, the tools available to do the analysis, and many other factors.

  The study of data quality has traditionally been the domain of data management specialists. Typical considerations include timeliness, accuracy, integrity, consistency, accessibility, and cost. Of these factors, accuracy is traditionally the most important, as bad data gives bad results.

  But in the era of big data, the sheer volume of the data makes up for slight inaccuracies, and now accuracy is not always the most desired attribute. This situation needs to expand the things considered when calculating the quality of a particular dataset.

  These considerations should include, but is not limited to, the following:

  - How easy the data is to fuse with other types of data
  - How easy the data is to analyze
  - How easy the data is to use for AI applications
  - Who owns the data
  - How we know we can trust the data
  - The history of the data
  - How quickly we can access the data

- How consistent the data is
- The integrity of the data
- The accuracy of the data
- The usefulness of the metadata
- The value of the data to specific stakeholders
- Why the data generators released the data (if they did)

Emergence of new media format

By reviewing the existing media format including audio/video as well as social media, there is some evidence of the emergence of new media with advances of information and communication technology. The rise of new media format has increased communication between people all over the world. It has allowed people to express themselves through blogs, websites, videos, images, and other user-created media. New media format most commonly refers to digital documents available on-demand through the Internet, accessible on any digital equipment, usually containing interactive user feedbacks and creative participation. New media format may include the existing social media like online newspapers, blogs, wikis, video and games. New media format can enable people around the world to share, comment on, and discuss a wide variety of topics. One of the key features of new media format is denoted as interactivity among communities.

New media format represents the digital forms with technologies that are manipulated, networkable, dense, compressible, and interactive. As a result of the evolution of new media technologies, virtual communities are being established online by eliminating geographical boundaries and social restrictions. People in virtual communities use words on screens to exchange information for life and business. New media has the ability to connect like-minded people worldwide and feeds into the process of guiding their future development.

Although there are several ways that new media may be described, Lev Manovich, in an introduction to "The New Media Reader", defines New Media by using eight propositions [102]:

- **New media versus cyberculture** – Cyberculture is the various social phenomena that are associated with the Internet and network communications (blogs, online multi-player gaming), whereas new media is concerned more with cultural objects and paradigms.
- **New media as computer technology** – New Media are the cultural objects which use digital computer technology for distribution and exhibition, e.g. websites, computer multimedia, and Blu-ray disks.
- **New media as digital data controlled by software** – New media is based on the assumption that all cultural objects rely on digital representation and computer-based delivery. New media is the digital data that can be manipulated by software. New media can create several versions of the same object. As an example, an image stored as matrix form can be manipulated and altered according to the additional algorithms implemented, such as color inversion, grey-scaling, sharpening, and rasterizing.
- **New media as the mix between existing cultural conventions and software** – New media can be understood as the mix between older cultural conventions and newer conventions for data representation, access, and manipulation. Software using computer animation can help representations of visual reality and human experience.
- **New media as the aesthetics** – If many aesthetic strategies may reappear, a much more comprehensive analysis on new media would correlate the history of technology with social, political, and economical histories.

- **New media as faster execution of algorithms** – High performance computers can make many new forms of media art such as interactive multimedia, 3D virtual reality, and video games.
- **New media as meta-media** – New media is about new ways of accessing and manipulating information (e.g., hypermedia, databases, search engines). Meta-media is an example of how quantity can change into quality as in new media technology. The manipulation techniques can recode modernist aesthetics into a different postmodern aesthetics.
- **New media as parallel articulation of post art and modern computing** – Post art or "combinatorics" involves creating images by systematically changing a single parameter. This leads to the creation of remarkably similar images and spatial structures. It means that algorithms as an essential part of new media do not depend on technology, but can be executed by humans.

7 Business opportunities in data eco-systems

7.1 Data Is Eating The World

In 2011, Silicon Valley venture capitalist Marc Andreessen penned his famous "Why Software Is Eating The World" article in the Wall Street Journal, where he argued that software companies were dominating or on their way to dominating almost every industry, "eating markets far larger than the technology industry has historically been able to pursue" [103]. He later argued that "Software will eat the world, in all sectors. Companies need to adapt or they will become extinct. In the future, every company will become a software company" [104].

Now it is data's turn. Data-driven companies are dominating or on their way to dominating almost every industry. Data companies are eating markets larger than software companies ever dreamt of. Data is eating the world, in all sectors, and now every company will become a data company.

The best companies in the world are already data companies. Six of the top ten most valuable publicly listed companies in the world by market capitalization are technology companies (Apple, Alphabet, Microsoft, Amazon, Facebook, and Tencent), all of which have earned their titles by building data-related products [105]. The remaining four (Exxon Mobil, Berkshire Hathaway, Johnson & Johnson, General Electric) all use data to turbocharge the effectiveness of their respective products.

It is not just the top companies that are seeing performance boosts. In 2012, the Harvard Business Review (HBR) wrote about a new style of data-driven management, where data guiding the decision-making process instead of fallible gut instinct, where every goal was linked to metrics, and where "analytics" was becoming de rigueur [106].

The more data-driven companies became, the better they performed. The companies that embraced data the most were 5% more productive and 6% more profitable than the companies that did not, and when your valuation is measuring in the tens or hundreds of billions, an extra 5% goes a long way. Data technology has come a long way in the four years since and the performance gap has likely continued to grow.

This massive impact is reflected in estimations for data-related markets. Below is a sample of some of the most recent estimates of the massive amount of value data technologies are generating for businesses.

7.2 Data-Related Market Estimates

There have been many attempts to quantify the market for data generation, management, analytics, and applications. Below are many of the most recent attempts sorted by area of interest.

7.2.1 Overall Market

IDC estimates that the big data and business analytics revenues will grow from nearly $122 billion in 2015 to more than $187 billion in 2019. The largest revenue opportunities are in Discrete Manufacturing (expected $22.8 billion in revenue in 2019), Banking ($22.1 billion), and Process Manufacturing ($16.4 billion) [107].

A June 2016 report by SNS Research showed that data-related investments are continuing to gain momentum around the world. They estimated that hardware, software, and professional services

sales related to technologies that capture, store, manage, and analyze data will net over $46 billion in revenue in 2016 alone. This number is expected to grow at a CAGR of 12% to over $72 billion in 2020.

Most of that market is dominated by hardware and professional services sales. By the end of 2020, however, SNS Research expects software revenue, led mainly by increase in sales of analytics software, to exceed hardware revenue by over $7 billion [108].

Wikibon estimates the global big data market will grow at a 14% CAGR from $18.3 billion in 2014 to $92.2 billion in 2026. The four fastest-growing sub segments are data management (14% CAGR), databases (18% CAGR), big data applications and analytics tools (23%), and core technologies like Hadoop (24% CAGR) [109].

451 Research estimates that the total data market is expected to nearly double in size over the next five years, growing from $69.6 billion in revenue in 2015 to $132.3 billion in revenue in 2020 [110].

Statistics estimates that the big data analytics and Hadoop market will grow at a CAGR of 42.1% from $8.48 billion in 2015 to $99.31 billion in 2022 [111].

7.2.2 Market Components

Companies

SNS Research found that nearly every large scale IT vendor maintains a Big Data portfolio and that competition between these vendors is heating up [108]. Gartner found that 10% of organizations have some form of prescriptive analytics. This number is expected to grow to 35% by 2020, largely due to adoption from large organizations in mature economies [112].

Big companies are not the only ones rushing to build data-related portfolios. Firstmark found that big data startups received $6.64 billion in venture capital investment in 2015, 11% of the total amount venture capitalists invested in technology [113]. Additionally, Deloitte found that over $1 billion in venture capital funding went towards cognitive technologies in 2014 and 2015 [114].

People

IDC predicts a need for 181,000 people with deep analytical skills in the US by 2018 and a requirement for five times that number of positions with data management and interpretation capabilities [115].

Gartner estimates that by 2020, only 50% of chief analytics officers will have successfully created a narrative that links financial objectives to business intelligence and analytics initiatives and investments [116].

Cloud-based Services

IDC estimates that spending on cloud-based big data and analytics technology will grow 4.5x faster than spending on on-premises technologies [117].

Business Intelligence

Gartner estimates that the global revenue for the business intelligence and analytics market will reach $16.9 billion in 2016, an increase of 5.2% from 2015 [118].

Gartner estimates that by 2020, 40% of enterprises' new investments in business intelligence and analytics will go towards predictive and prescriptive analytics [119].

Gartner estimates that the prescriptive analytics software market will grow at a 22% CAGR from $415 million in 2014 to $1.1 billion in 2019 [120].

IDC estimates that 50% of all business analytics software will incorporate prescriptive analytics built on cognitive computing functionality by 2020 [121].

Internet of Things

Cisco estimates that the number of machine-to-machine connections should grow from 4.9 billion in 2015 to 12.2 billion in 2020, almost half of all connected devices [122]. By 2022, Cisco expects this global IoT market to be worth $14.4 trillion [123].

Industry-specific use cases like smart grids and connected personal vehicles will be responsible for $9.5 trillion (66%) of this amount. Smart factories alone will contribute $1.95 trillion of value. Cross-industry use cases like telecommuting and collaboration technologies will be responsible for $4.9 trillion (34%). Four industries compose more than half of this amount: manufacturing (27%), retail trade (11%), information services (9%), and finance and insurance (9%). The remaining 14 industries range from 1 to 7 percent.

The largest area of investment will be related to improving customer experiences. Additional areas of investment include reducing time-to-market ($3 trillion), improving supply chain and logistics ($2.7 trillion), reducing cost ($2.5 trillion), and increasing employee productivity ($2.5 trillion) [124].

Cisco also found that 50% of IoT activity today is in manufacturing, transformation, smart cities, and consumer markets.

IDC estimates that the worldwide IoT market will grow from $655.8 billion in 2014 to $1.7 trillion in 2020. Devices, connectivity, and IT services will likely make up two-thirds of the IoT market in 2020, with devices (including modules and sensors) alone representing more than 30 percent of the total [125].

Woodside Capital Partners estimates that IoT-related services will grow at a CAGR of 15.71% from $50 billion in 2012 to $120 billion in 2018 [126].

Telecoms

The data-driven telecom analytics market is expected to have a CAGR of nearly 50 percent, with annual revenues expected to reach $5.4 billion at the end of 2019 [127].

Ericsson estimates that by 2020, the number of smartphone subscriptions will have increased from today's 2.7 billion to 6.1 billion, and the total amount of mobile traffic generated by smartphones will be five times that of today [128].

7.3 Example Applications

There are many examples of applications enabled by data technologies. Well-known examples of emerging products made possible by data technologies include self-driving cars, smart wearable, smart homes, augmented and virtual reality, and crypto currencies. However, it is not necessary to develop entirely new products to take advantage of data technologies.

This section will explore three ways companies can take advantage of data technologies:

- Personalizing services
- Improving operations

- Developing new business models

Each subsection will contain a brief overview of that category and a list of several common examples of how individuals, companies, and organizations are taking advantage of these opportunities. Example industries that will see benefits from these applications include:

- Automotive, Aerospace & Transportation
- Banking & Securities
- Defense & Intelligence
- Education
- Healthcare & Pharmaceutical
- Smart Cities & Intelligent Buildings
- Insurance
- Manufacturing & Natural Resources
- Web, Media & Entertainment
- Public Safety & Homeland Security
- Public Services
- Retail, Wholesale & Hospitality
- Telecommunications
- Utilities & Energy
- Others

7.3.1 Personalizing Services

Data technologies enables personalization at scale. Algorithms can now look at a user's activity logs and automatically make decisions based on the preferences that particular person has expressed in the past through their actions. Here are several examples of data-driven personalized services:

- Customizing treatment plans based on a patient's medical history
- Identifying the best postoperative treatment plan to minimize each patient's chance of readmission
- Writing a news article about a particular event, customized for a particular reader's level of interest and experience level
- Ordering search engine results based on the searcher's prior history
- Identifying the most relevant ad to show a particular person
- Customizing the homework assignment based on the student's current level of understanding
- Matching students with educational support like tutors on a topic-by-topic basis
- Creating customized insurance packages for individuals based on risk factors
- Automatically giving financial planning and investment advice
- Providing personalized mental health therapy for individuals in need

7.3.2 Improving Operations

Data technologies help make things more efficient. Algorithms can now weigh all of the possible options and make judgment calls at a scale and level of accuracy impossible for humans to ever achieve.

Here are several examples of using data to improve the day-to-day operations of a service, piece of infrastructure, or product:

- Identifying fake news stories circulating on social media
- Checking the facts mentioned on a website, or in scientific papers, as a first pass before peer review
- Flagging suspicious transactions that may be fraudulent
- Reading handwritten addresses on envelopes to automatically route mail
- Auditing business financials to prevent fraud and mismanagement
- Deciding where to acquire what goods to optimize the supply chain for a portfolio of products
- Identifying bottlenecks in particular steps on factory floors
- Choosing shipping routes that minimize cost of resources consumed
- Adjusting an investment portfolio to hedge against price changes and maximize profits
- Deciding where in the energy grid to store surplus energy to most effectively meet increased demands during peak hours

7.3.3 Developing New Business Models

The newfound importance of data is changing the way many businesses work. It's also generating a lot of new business opportunities that did not previously exist or existed but were much smaller. Here are several examples of business models that are now significantly more valuable thanks to data technologies:

- Offering a "free" service in exchange for ownership of the data generated while using the service (e.g. free navigation applications that generate data about where people go when)
- Helping others determine and control what data is being generated about their activity
- Buying data from data owners and selling it to potential data users
- Helping data owners clean and sell the data on their own
- Building and selling sensors that generate data
- Managing sensor networks that generate data that can be sold
- Helping data owners understand the potential value in the data they own
- Helping data owners manage their data
- Helping data owners analyze their data
- Helping data owners apply the results of their analyses

7.4 Public open data related activities

7.4.1 Comprehensive Knowledge Archive Network (CKAN) activities

7.4.1.1 Introduction to CKAN

The Comprehensive Knowledge Archive Network (CKAN) is an open source data platform software package developed by the UK-based Open Knowledge, a worldwide non-profit network of people passionate about openness, using advocacy, technology and training to unlock information and enable people to work with it to create and share knowledge [130].

CKAN is used to manage data repositories, which serve as central locations to find data, standard practices, and showpiece use cases. These repositories are places where people inside and often outside of the hosting organization in question can search for and download the data they want [iot-7].

7.4.1.2 Examples of using CKAN

CKAN is used by more than 20 national governments and many more local governments, research institutions, companies, and organizations around the world. Notable examples include the federal governments of the United States, the United Kingdom, Mexico, Japan, and many other institutions around the world [132]. These examples often have many tens of thousands or hundreds of thousands of freely available datasets. For example, Data.gov.uk has 40,243 published datasets, Data.gov has 192,897 published datasets, and Data.gov.au has 235,000 published datasets [133].

Currently, CKAN is mostly used to power open data websites that offer data for free of charge. However, the software is open source, and could be used to power data markets where people pay to obtain data. Additionally, various organizations and people around the world offer paid services to help install and maintain the CKAN software.

Notable features include configurable metadata, user-friendly web interfaces, fine-grained authorization levels, and APIs [134]. Here is a more detailed list of features according to the CKAN website: [135].

- Data entry via web UI, APIs or spreadsheet import
- versioned metadata
- configurable user roles and permissions
- data previewing/visualisation
- user extensible metadata fields
- a license picker
- quality assurance indicator
- organisations, tags, collections, groups
- unique IDs and cool URIs
- comprehensive search features
- geospacial features
- social: comments, feeds, notifications, sharing, following, activity streams
- data visualisation (tables, graphs, maps, images)
- datastore (‘dynamic data’) + file store + catalogue
- extensible through over 60 extensions and a rich API for all core features

- can harvest metadata and is harvestable, too

7.5 IoT data related activities

7.5.1 Introduction

There exist a huge number of data creating devices being connected in these days. As described in Table 13, Gartner forecasted that 6.4 billion connected things will be in use worldwide in 2016 and will reach 20.8 billion by 2020 [136]. Following this trend, many of groups have been created and activities are occurring. Many organizations and associations, such as International Electrical Electronics Engineering (IEEE), United States Department of Transportation, European Commission and etc., are focusing on and piloting IoT technologies as parts of their smart city projects. Internet Engineering Task Force (IETF), Internet Research Task Force (IRTF) and W3C as well have been working on key IoT related standards and guidance for a significant amount of time. In the following section, their works and activities to approach toward the seamless Web and IoT environments will be shortly introduced.

Table 13. IoT units installed base by category (Millions of Units)

Category	2014	2015	2016 (forecasted)	2020 (forecasted)
Consumer	2,227	3,023	4,024	13,509
Business: Cross-Industry	632	815	1,092	4,408
Business: Vertical-Specific	898	1,065	1,276	2,880
Grand Total	**3,807**	**4,902**	**6,392**	**20,797**

7.5.2 IoT related groups and their works

IETF has been researching in mature standards and guidance for IoT over the past decade and recently emphasizing the importance of Web and IoT technologies strongly more than ever. In a IETF journal published in April 2016, IETF states that beyond the IETF work specifically focusing on IoT scenarios, other IETF working groups will likely end up being use for also IoT [137].

IETF's affiliate organization, IRTF, as well is paying attention in the potential of IoT. IRTF chartered Thing-to-Thing Research Group (T2TRG) in 2015 to investigate open research issues in IoT focusing on the topics of semantic interoperability, security and lifecycle management, ways to use the REST paradigm in IoT scenarios, and etc. [138].

W3C set up Web of Thing Interest Group (WoT IG) as well to support interoperability and inter connectivity of IoT technologies and platforms in the Web-based level through the global reach of Web standards [139]. W3C takes an amount of efforts to provide web standards and data formats related to IoT technologies to build interoperable Web services, such as Microdata or JSON-LD which allow machine-readable data to be embedded in the Web documents.

Some of the IoT related groups affiliated to these organizations and their descriptions along with their works are briefly described in Table 14 [138, 139, 140].

Table 14. IoT related groups and their activities

Group	Initiated	Concluded	Descriptions / Works
IETF			
Internet Protocol version 6 (IPv6) over Low-power WPAN **(6LoWPAN)**	Mar 2005	Jan 2014	• Methods for adapting IPv6 to WPAN network
IPv6 over Low-power **(6Lo)**	Oct 2013	Active	• Methods for adapting IPv6 to a wider range of radio technologies, including Bluetooth Low Energy
Routing Over Low-power and Lossy networks **(ROLL)**	Feb 2008	Active	• Specifications for both the IPv6 Routing Protocol for Low-Power and Lossy Networks and a set of related extensions for various routing metrics, objective functions, and multicast.
Constrained RESTful Environments **(CoRE)**	Mar 2010	Active	• Constrained Application Protocol (CoAP), a radically simplified User Datagram Protocol based analog to HTTP. • Extensions to CoAP enable group communications and low-complexity server-push for the observation of resources • A discovery and self-description mechanism based on a web link format suitable for constrained devices • currently looking at a data format to represent sensor measurements
IRTF			
Things-to-Thing Research Group **(T2TRG)**	Dec 2015	Active	• Investigating open research issues in leading a true IoT into reality • Focusing on issues related to standardization in the IETF
W3C			
Web of Thing Interest Group **(WoT IG)**	Aug 2016	Active	• Countering the fragmentation of the IoT by introducing a Web-based abstraction layer capable of interconnecting existing IoT platforms and complementing available standards

8 Standardization strategies for future data format and standards

Standardizing open data formats, something the ITU-T is well-positioned to do, would maximize the collective intelligence of people in the future data eco-society.

(Note) In this section, the standardization strategies are focused on open data formats. Only the technical aspects of private and confidential data formats are taken into consideration.

Principles for Data Eco-society

The future data eco-society is based on the ability of everyone to share data, access data, and transform that data into knowledge. Here are the basic principles that are essential for the future data eco-society:

- (**Open**) Open data can be used in various parts of human life and business. It makes it easier for people to access valuable data. From the viewpoint of standardization, "open" data is available and within the reach of the public, without barriers for its reuse and consumption. The future data eco-society provides open, equal, inclusive, and universal access of data to anyone who wants it.
- (**Trust**) Data trust (including security and privacy) is a prerequisite for the development of the future data eco-society. Published documents should be digitally signed and include information like publication/creation date, authenticity, and integrity. Digital signatures help people trust that the data has not been modified since it was published. This trust is required to reduce all kinds of risk when it comes to using open data.
- (**Equal Opportunity**) The opportunity to benefit from open data should be equally available to everyone. This works when open data is not restricted by privacy, security, and privilege limitations as well as copyright, patent, and trade regulations. This requires some harmony between the private sector and the public/social/government organizations.
- (**Linked Open Data**) Data is composed of sets of data records linked together and organized by those links. These linked lists are useful for retrieving and identifying the properties of the data records. Metadata includes the descriptions and other related information of those links. These links and metadata information help define the relationship between data.
- (**Collaboration**) ICT technologies help open data freely benefit everyone. The ability to better share information encourages innovation by encouraging collaboration. It is important to think about how open data standards can technically and legally ensure users' access to relevant and reliable public data. To promote the spread and sharing of open data, step-wise standardization is essential.

Collaborations for Open Data Standards

There is a long history of data standards including formal international standards bodies like ITU and ISO/IEC and independent standards bodies like the World Wide Web (W3) consortium. The openness of the standards is essential. If some data standards are sometimes proprietary and only available under restrictive contract terms from the organization that own the relevant copyright, such specifications are not considered to be fully open and cannot be called open standards.

The development of open data standards is a highly complex socio-technical negotiation process. The understanding of the social, economic, political, and technical considerations of a future data eco-society is important for these negotiations. To achieve a clear consensus on future data formats, the conceptual framework of the future data eco-society should guide research about data applications. This research should then guide the discussion about what data format best enables current data applications and future data applications. This conceptual model of a future data eco-society is part of the knowledge information infrastructure and will help build an open, voluntary, and consensus-based process to standardize data formats.

To extend the open data standards to other industries, future data formats should take into account computing, storage, and networking resources needed for energy, transportation, health, education, and other environments. Because the other industries have traditionally used their own data formats, data sharing platforms that use open standards will be important for the adoption of those open standards in multiple environments. In the future data eco-society, the collective intelligence framework is required to understand data from various sensors, networking systems, and cloud servers.

Currently, URI/URL/URN schemas are the primary way to identify digital data resources. However, for IoT applications that rely heavily on fusing data from multiple sources, other identification,

numbering, and addressing schemas need to be developed that support better IoT/M2M devices and systems.

Here are the issues that need to be investigated to develop a common understanding of the future data eco-society and thus facilitate standards that enable that society:

- **How to share the data for knowledge accumulation**
  - Future data files include metadata like authors, date, genre, and short summaries for easier search and discovery. The documents in same area could be classified and sorted to share individual experiences and opinions. Future data formats should provide a way to share, aggregate, and classify large volumes of data.
- **Metadata and data schema are the key essences for future data eco-society**
  - Various types of metadata provide useful information about when data is created, delivered, processed, shared, and consumed by users.
  - Data by itself is not often self-descriptive. Metadata helps interpret the data structure. Depending on application, metadata is parsed to help better understand the actual content of the data, instead of just the structure of the data.
- **New forms of social media for development, acquisition, and spread of knowledge**
  - New forms of social media are used to create, collect, accumulate, share, and distribute data. This data can be summarized to invent new forms of knowledge. These summaries may naturally evolve from social media with the progress of user interface and human perception technologies.
- **New web as a useful tool for future data eco-society**
  - Existing web technology based on HTML has some limitations that prevent it from being entirely useful for IoT/M2M applications. New markup languages that can facilitate the accumulation of human experiences and opinions may be needed.
  - Web-based application programming interfaces (APIs) for sharing content, documents, and files should be similarly enhanced. Additionally, new mechanisms for sharing data between human-to-machine and machine-to-machine communication channels are needed.

ITU-T initiatives for public open data standardization

There exist many definitions on data and metadata formats for existing applications. From the perspective of information and communication technology, the knowledge information infrastructure is difficult to realize since any form of knowledge should be represented and shared in the form of data. Designing effective and efficient data formats including metadata is a key research challenge. Various types of new data models can be analysed and suggested before even starting the formal standardization process. The collective behaviour of people developing and using these data applications will generate new ideas and drive the discussion towards eventual global consensus. This global consensus will be the basis for developing the relevant format standards. The process of developing these standards is thus based on a fair and equitable way that typically ensures the high-quality output and market relevance.

Another important component of the standards is backwards compatibility with existing data standards. Standardization can be achieved on many different levels resulting in a uniform and integrated system with harmonized process flows. This harmonization will facilitate information flow between different organizations with their own accumulated experience. The right level of data

standardization depends on the individual member's conditions, working structure, management maturity, and objectives.

The current ITU-T standardization working methods could be supplemented by the different views, opinions, and technical experiences from other industries and other standard bodies. Brainstorming with these other organizations may be needed to get rough and common consensus. The initial consensus target could be the open data format standards.

Since all the data formats affect and are affected by hardware systems and software environments, the standardization strategies for open data formats may need clear guidelines and established principles.

Many experts on market-specific or business critical applications try to add custom components to the data formats that they use. This happens often to a large variety of markets, resulting in a large variety of complicated and incompatible data formats. For example, there are already a wide variety of metadata and 3D geolocation data formats depending on the specific application, and it is nearly impossible to reach global consensus about metadata and 3D data formats.

Security and protection have recently become important data-related issues. Most ICT people assume that the data formats for telecommunication, broadcast, and Internet/web applications are easily extended to other industries and form the basis for common data formats of the future data eco-society. However, experts of other industries like energy, health, transportation, logistics, and environment may want their data formats to serve as the foundation for the future data eco-society. The ITU-T can focus on open data formats to avoid conflicting with the actions of these other industries. This is also more feasible than standardizing domain-specific formats: web-based data formats are easier to standardize.

The following approaches are recommended for open data standardization under ITU-T, which may take the form of a joint research group organized in collaboration with other SDOs:

- Concept and basic principles for future data eco-society
    - Review the concepts and understanding of data format to develop a common language for the future cyber world,
    - Identify the definition, property, and functional capability of future data society,
    - Analyze the relationship between data, information, and knowledge, and
    - Investigate the use cases and examples of data, information, and knowledge.

- Open data formats for future data eco-society
    - Review existing data types and formats in both digital and analogue forms,
    - Investigate the definition, property, and relationship of open data formats including raw data, metadata (descriptive data), and management data,
    - Investigate the descriptive methods of metadata for common and specific applications,
    - Investigate the tagging and indexing models used to organize data,
    - Investigate data formats for specific applications (e.g. 3D geographical information, anatomy information of human body, and composition of texts, image, symbol, and audio/visual information),
    - Investigate the descriptive format and processing methods used in data analytics,
    - Analyse the backwards compatibility of open data formats with existing data formats.

- Public social media as typical use cases of public open data format
    - Review existing social media services and technologies,
    - Investigate concepts and principles of social media related to the future data eco-society,
    - Investigate how web technologies and web services relate to social media,
    - Investigate web-based public open software related to the future data eco-society,
    - Investigate the step-wise deployment scenario and roadmap of future data eco-society

Standardization process of open data under ITU-T

The efforts of the ITU-T have focused on helping people communicate data efficiently. However, having data is now much less important than being able to extract value from it, and the ITU should expand its focus to include helping people use data efficiently. This report explores the likely future of data analytics to identify key issues the ITU can consider while developing new analytics-related standards. This report proposes that the ITU-T follow these open data format standardization strategies:

- 1st stage: establish a strategic group for discussing open data formats and technology
- 2nd stage: identify action plans and terms of references for open data standards
- 3rd stage: establish the relevant working group to develop the standards
- 4th stage: collaborate with ITU-T study groups and organizations outside of the ITU-T

Section 6 describes the key requirements of terms of references of open data standards for future data formats. The technical issues described in section 6.1 are discussed in detail. The strategic priorities described in section 6.3 can modify influence the organization of the working items described in section 6.2.

The digital identity management and identification issues for future data formats can be explored by Study Group 2. The overall functional capabilities for future data platform can be explored by Study Group 13. The audio/video data formats are explored by Study Group 16. The data formats for IoT and convergence applications are explored by Study Group 20. The data format issues on security and copy protection are explored by Study Group 17.

Collaborations with organizations outside ITU-T

The other standards development organizations (SDOs) have their own working methods to produce documents, reports, and implementation agreements. Collaboration between the ITU-T and the other mission-oriented SDOs may be needed. This requires discussing priorities, action items, and collaborative working methods with other SDOs to build a global consensus about what the future data eco-society should look like and how to best achieve it.

Web-based data formats are developed by the W3C consortium. Generic metadata have been produced by ISO/IEC. Some identification standards for digital electric code have been organized by GS1. Data formats for health, geography, and transportation data are driven by their respective industries.

The following working methods are recommended to facilitate collaboration with organizations outside ITU-T:

- Harmony among ITU-T, other standard bodies, and the private sector (W3C)
    - (ITU-T) Open data formats for ICT environments with common and mutual benefits
    - (Other Standard Bodies) Open data formats for their specific domains that are aligned with ICT environment
    - (Private) Technical solutions for open data formats
- Collaborations with academia for innovation and technical breakthroughs
    - Building open common collaborative environments and interoperability tests
    - Developing a global consensus requires understanding the opinions and needs of members of academia
- Market feedback on technical specifications
    - Identify technical problems and difficulties during market adoption of new data formats
    - Establish processes to translate market feedback into technical changes

9 Conclusions

This document is aligned with the ITU-T Technical Report "Future Social Media and Knowledge Society" published on 30 November 2015. This report focuses on standardization strategies that will help open data formats enable the future data eco-society. Currently, many organizations develop their own custom data formats regardless of similar activities of other bodies. This applies to organizations in the telecommunication, broadcast and Internet industries as well as organizations in other industries like energy, transportation, health, environment, and public safety. Some technical experts believe that data interoperability is most significant near future issue and that a future data eco-society requires any data formats without interoperability be demolished. The data formats will be continued only if people including data platforms feel easy and comfortable to share their data with others.

IoT standards must involve knowledge from many domains. Currently, most IoT platforms use their own custom data formats to structure the data generated from their sensors. IoT services then need combine data from multiple sensors to understand the context of the specific situation. This often involves fusing the IoT sensor data other types of data like human behavioural data, weather, and event calendars. Additionally, other industries like energy, health, and transportation are starting to develop their own data and metadata formats to accommodate requirements specific to their domain. This means that IoT data cannot be developed solely by the ITU-T. Instead, ITU-T can use its experience as a respected international standards body to collaborate with other standard bodies and institutions in many domains to develop IoT standards.

Over the last several decades, information and communication technology (ICT) was one of the key drivers of innovation and technological breakthroughs throughout the world. ICT will continue to act as a driving force for all industries. ICT technologies can no longer focus solely on communicating a lot of data from place to place. Instead, they need to help people and computing platforms easily understand and interpret data stored in many formats. Because data formats are closely related to the corresponding hardware and software, the standardization of future data formats will help structure billions of software packages and hardware devices.

This report can be summarized as follows:

- Review the existing data formats in terms of sharing and searching

- Data formats for publications, files, and audio/video applications have been analysed. Most data formats assume a specific service environment for the creation, delivery and consumption of data.
- Metadata formats, which are essential for searching and sharing data in the era of big data, are not a significant component of most data formats.
- There is no concept of structured metadata in web-based data formats like XML/RDF. Current standardization efforts aim to make future data formats interpretable without prior knowledge.
- Geolocation data formats and 3D data formats will be based on web architectures.

- Analyze key trends and technical issues toward future data formats
  - There is some paradox between data openness and data businesses. Some people want to advertise their purchasable datasets on open data repositories. However, as they only want paying customers to access their data, they only post summaries of their data and links of how to pay for the full dataset. This is at odds with the stated purpose of most open repositories, which only want to publish freely available datasets.
  - The URI/URL/URN schema is analysed to see how it helps enable the sharing of data. This report assumes that web architectures will continue being the primary platform of a future open data eco-society.
  - Because data formats are tightly linked with corresponding software, future data format standards explain how billions of software packages can handle data without being dependent on specific hardware or service environments. The hardware solution providers can just use the predefined microdata format without understanding the actual content, just like they do with email template standards.
  - Data indices, data tags, and data schemas are important for creating, delivering, sharing, and consuming data.
  - Microdata formats are a great way to deliver IoT data because no explanation of data semantics is needed. Hardware solution providers will not need to know complicated contextual information about the IoT ecosystem.
  - Data formats that accumulate domain-specific experiences and insights are investigated. Future data schemas may explain how information and knowledge can be extracted from raw data. But, in-depth studies on data semantics still need to be performed.
  - Data management and governance issues like digital identity, data security, and data protection were only slightly investigated in this report. More in-depth studies need to be performed.
  - Standardizing 3D data formats will be difficult because there are many hardware and software platforms that are already in the marketplace, each with their own 3D data format. The critical factor is the web friendliness of the format.

- Identify key requirements to initiate standardization activities for future open data format
  - Open data is classified and applications are identified. The classification depends on factors like the source of the data, whether the data is structured or unstructured, and whether the data is a file or a stream of data.
  - Data identification is important for searching and discovering data.
  - Metadata formats are important for helping people search, store, share data.

- Analyze new market opportunities of data eco-society
  - New data practices and market trends are investigated. A lot of new data-related business opportunities can be found.

- New data products and new media markets rely on human intelligence.
- New media technologies that evolve from existing social networking services may initiate new markets.
- Virtual reality for practices and new experiences of tacit knowledge will be upcoming.
- New markets for the cyber physical system are combined with IoT/M2M technologies.

- ITU-T has a responsibility to get a consensus for future data eco-society
  - ITU can have a leadership role in introducing future data standards by achieving global consensus about the nature of the future knowledge society.
  - Standards for future data technology and data-centric industries are necessary to realize a knowledge society

Finally, ITU-T has the chance to lead efforts to standardize open data formats by trying to initiate new working methods in collaboration with organizations in the private sector and academia that help develop standards that enable the future data eco-society.

10 References and bibliography

Here are several of the many resources we are exploring to learn more about various topics within this document, particularly how data science is changing now / will change in the future to better understand requirements from the user side (this is just the first section of links we have stored; we keep hundreds more in various places but we don't want to add tens of pages of just links we'll remove later anyway).

[1] www.gs1.org, The global language of business

[2] www.opengeospatial.org

[3] https://en.wikipedia.org/wiki/Global_Positioning_System, last modified on 27 September 2016

[4] https://en.wikipedia.org/wiki/Common_Data_Format, last modified on 21 July 2016

[5] https://en.wikipedia.org/wiki/Apache_Hadoop, last modified on 29 September 2016

[6] https://www.w3.org/TR/REC-rdf-syntax/, last modified on 10 February September 2004

[7] http://www.w3schools.com/html/

[8] Chen Hsinchun, Roger HL Chiang, and Veda C. Storey.: Business Intelligence and Analytics: From Big Data to Big Impact, MIS quarterly, Vol. 36 No. 4, pp.1165-1188, (2012)

[9] Kim Gang-Hoon, Silvana Trimi, and Ji-Hyong Chung.: Big-data applications in the government sector, Communications of the ACM, Vol. 57 No. 3, pp. 78-85, (2014)

[10] http://www.martinhilbert.net/WorldInfoCapacity.html/, last modified on 11 Apr 2011

[11] https://en.wikipedia.org/wiki/JPEG, last modified on 18 Nov 2016

[12] https://en.wikipedia.org/wiki/JPEG_2000, last modified on 11 Apr 2011

[13] http://mpeg.chiariglione.org/standards

[14] https://en.wikipedia.org/wiki/Advanced_Audio_Coding, last modified on 19 Nov 2016

[15] https://en.wikipedia.org/wiki/MPEG-4, last modified on 23 Sept 2016

[16] https://en.wikipedia.org/wiki/World_Wide_Web, last modified on 19 Nov 2016

[17] http://www.internetlivestats.com/total-number-of-websites/, last modified on 11 Apr 2011

[18] https://www.w3.org/standards/semanticweb/

[19] https://en.wikipedia.org/wiki/Semantic_Web_Stack, last modified on 17 Sept 2015

[20] http://www.w3schools.com/html/html_css.asp

[21] https://en.wikipedia.org/wiki/Cascading_Style_Sheets, last modified on 21 Nov 2016

[22] https://en.wikipedia.org/wiki/XML, last modified on 24 Nov 2016

[23] http://www.w3schools.com/xml/xml_whatis.asp

[24] https://www.w3.org/standards/xml/schema

[25] http://www.w3schools.com/Xml/schema_intro.asp

[26] http://www.json.org/

[27] https://en.wikipedia.org/wiki/JSON, last modified 28 Nov, 2016

[28] https://en.wikipedia.org/wiki/Resource_Description_Framework, last modified on 23 Nov 2016

[29] https://www.w3.org/TR/NOTE-rdf-simple-intro, last modified on 13 Nov 1997

[30] https://www.w3.org/RDF/, last modified on 15 Mar 2014

[31] http://www.w3schools.com/xml/xml_rdf.asp

[32] http://nationalgeographic.org/encyclopedia/geographic-information-system-gis/

[33] http://support.esri.com/sitecore/content/support/Home/other-resources/gis-dictionary/term/raster%20data%20model

[34] http://support.esri.com/other-resources/gis-dictionary/search/vector%20data%20model

[35] http://planet.botany.uwc.ac.za/nisl/GIS/GIS_primer/page_19.htm

[36] http://www.gartner.com/newsroom/id/3165317, last modified on 10 Nov 2015

[37] https://www.lib.ncsu.edu/gis/formats.html

[38] http://data.geocomm.com/helpdesk/formats.html

[39] https://en.wikipedia.org/wiki/ECW_(file_format), last modified on 18 May 2016

[40] http://ncl.sbs.ohio-state.edu/ica/3_spatial.html

[41] http://ncl.sbs.ohio-state.edu/ica/1_about.html

[42] http://www.cgi-iugs.org/

[43] http://www.geosciml.org/

[44] http://www.cgi-iugs.org/tech_collaboration/earthResourceML.html

[45] https://www.fgdc.gov/organization

[46] https://www.geoplatform.gov/

[47] https://www.fgdc.gov/standards/list#under-development

[48] https://www.fgdc.gov/standards/fgdc-endorsed-external-standards/index_html

[49] https://www.fgdc.gov/resources/factsheets/documents/GeospatialMetadata-July2011.pdf

[50] https://www.fgdc.gov/metadata/iso-standards

[51] http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=53798, last modified on 1 Apr 2014

[52] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=39229, last modified on 14 July 2014

[53] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=39229, last modified on 15 Aug 2016

[54] http://gispopsci.org/software/

[55] http://gispopsci.org/

[56] https://en.wikipedia.org/wiki/List_of_spatial_analysis_software, last modified on 23 Nov 2016

[57] http://www.merriam-webster.com/dictionary/e-book

[58] http://www.adweek.com/galleycat/ebooks-top-hardcover-revenues-in-q1/54094?red=as, last modified on 15 Jun 2012

[59] http://www.digitalbookworld.com/2014/more-americans-now-reading-ebooks-new-pew-data-show/, last modified on 16 Jan 2014

[60] http://wiki.mobileread.com/wiki/E-book_formats, last modified on 24 Oct 2015

[61] https://en.wikipedia.org/wiki/Comparison_of_e-book_formats, last modified on 13 Nov 2016

[62], http://www.idpf.org/epub/31/spec/epub-packages.html#sec-package-def, last modified on 14 Oct 2016

[63] http://www.idpf.org/epub/31/spec/epub-packages.html#sec-package-doc

[64] https://en.wikipedia.org/wiki/Comparison_of_e-book_readers, last modified on 24 Nov 2016

[65] http://www.rfidjournal.com/site/faqs#Anchor-What-363

[66] http://www.itu.int/oth/T2301000021/en, last modified on 3 Oct 2013

[67] http://www.itu.int/dms_pub/itu-t/oth/23/01/T23010000040001PDFE.pdf

[68] Sen, Dipankar; Sen, Prosenjit; Das, Anand M.: RFID For Energy and Utility Industries, PennWell, ISBN 978-1-59370-105-5, pp. 1-48, (2009)

[69] http://www.rfidjournal.com/articles/view?1335/, last modified on 16 Jan 2005

[70] http://www.gs1.org/about

[71] https://en.wikipedia.org/wiki/GS1, last modified on 24 Nov, 2016

[72] http://www.gs1.org/epc-rfid

[73] http://www.gs1.org/epc/tag-data-translation-standard

[74] http://www.gs1.org/epcrfid/epc-rfid-uhf-air-interface-protocol/2-0-1

[75] http://www.gs1.org/epcrfid/epc-rfid-hf-air-interface-protocol/latest

[76] http://www.gs1.org/epcrfid/epc-rfid-llrp/latest

[77] http://www.gs1.org/epcrfid/epc-rfid-dci/latest

[78] http://www.gs1.org/epcrfid/epc-rfid-rm/latest

[79] http://www.gs1.org/ale

[80] http://www.rfidineurope.eu/sr, last modified on Sep 2016

[81] https://en.wikipedia.org/wiki/Linked_data, last modified on 1 October 2016

[82] https://en.wikipedia.org/wiki/Data_classification_(business_intelligence), last modified on 18 January 2015

[83] https://en.wikipedia.org/wiki/Knowledge_Graph, last modified on 7 November 2016

[84] https://developers.google.com/search/docs/guides/intro-structured-data?rd=1, last modified on 22 June 2016

[85] https://en.wikipedia.org/wiki/Metadata, last modified on 15 October 2016

[86] http://guides.library.ucla.edu/c.php?g=180539&p=1190454, Last Updated: Sep 21, 2016

[87] http://www.ietf.org/rfc/rfc5013.txt

[88] https://en.wikipedia.org/wiki/Tacit_knowledge, last modified on 2 October 2016

[89] https://en.wikipedia.org/wiki/Trusted_Data_Format, last modified on 14 October 2016

[90] https://en.wikipedia.org/wiki/Data_management, last modified on 16 June 2016

[91] https://en.wikipedia.org/wiki/Data_governance, last modified on 17 June 2016

[92] http://edutechwiki.unige.ch/en/X3DV#X3D_Definition, last modified on 22 August 2016

[93] Deelman Ewa, and Ann Chervenak.: Data management challenges of data-intensive scientific workflows, Cluster Computing and the Grid, 2008. CCGRID'08. 8th IEEE International Symposium on. IEEE, pp. 687-692, (2008)

[94] https://developer.myacxiom.com/code/api/data-bundles/main

[95] Schomm, Fabian, Florian Stahl, and Gottfried Vossen.: Marketplaces for data: an initial survey, ACM SIGMOD Record 42.1, Vol.1, pp. 15-26, (2013)

[96] Pingzhi Fan.: Coping with the big data: convergence of communications, computing and storage, China Communications, Vol. 13 Issue. 9, pp. 203-207, (2016)

[97] Bauer, Florian, and Martin Kaltenböck.: Linked open data: The essentials, Edition mono/monochrom, Vienna (2011)

[98] Steinke, Gerhard.: Data privacy approaches from US and EU perspectives, Telematics and Informatics, Vol. 19 No. 2, pp. 193-200, (2002)

[99] http://dublincore.org/

[100] Arash Shahi.: Activity-Based Data Fusion for the Automated Progress Tracking of Construction Projects, PhD Thesis, presented to University of Waterloo, Ontario, Canada (2012)

[101] http://datatags.org/

[102] Manovich, Lev, *New Media From Borges to HTML*, The New Media Reader, Ed. Noah Wardrip-Fruin & Nick Montfort, Cambridge, Massachusetts, 2003, 13-25. ISBN 0-262-23227-8.

[103] http://www.wsj.com/articles/SB10001424053111903480904576512250915629460, last modified on 20 Aug 2011

[104] http://www.hurriyetdailynews.com/every-company-will-become-a-software-company-.aspx?PageID=238&NID=96253&NewsCatID=407, last modified on 10 Mar 2016

[105] https://en.wikipedia.org/wiki/List_of_public_corporations_by_market_capitalization#2016, last modified on 21 Nov 2016

[106] https://hbr.org/2012/10/big-data-the-management-revolution, last modified on Oct 2012

[107] IDC Worldwide Big Data and Business Analytics Revenues Forecast, https://www.idc.com/getdoc.jsp?containerId=prUS41306516, last modified on 23 May 2016

[108] http://www.snstelecom.com/bigdata, last modified on Jun 2016

[109] http://siliconangle.com/blog/2016/03/30/wikibon-forecasts-big-data-market-to-hit-92-2bn-by-2026/, last modified on 30 Mar 2016

[110] https://451research.com/report-short?entityId=89339, last modified on 14 Jun 2016

[111] http://www.strategymrc.com/report/big-data-analytics-hadoop-market, last modified on July 2016

[112] http://www.gartner.com/document/3202617

[113] http://mattturck.com/2016/02/01/big-data-landscape/#more-917, last modified on 1 Fab 2016

[114] https://www2.deloitte.com/content/dam/html/us/analytics-trends/2016-analytics-trends/pdf/analytics-trends.pdf

[115] http://www.idc.com/research/viewtoc.jsp?containerId=253423

[116] http://www.gartner.com/document/3263218

[117] https://www.cloudera.com/content/dam/www/static/documents/analyst-reports/idc-futurescape.pdf

[118] http://www.gartner.com/newsroom/id/3198917, last modified on 3 Feb 2016

[119] http://www.gartner.com/document/3263218

[120] http://www.gartner.com/document/3202617

[121] https://www.cloudera.com/content/dam/www/static/documents/analyst-reports/idc-futurescape.pdf

[122] http://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html?CAMPAIGN=VNI+2016&COUNTRY_SITE=us&POSITION=Press+Release&REFERRING_SITE=Cisco+page&CREATIVE=PR+to+VNI+web+page

[123] http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoE_Economy.pdf

[124] http://tamarafranklin.com/wp-content/uploads/2015/09/Oracle-Internet-of-Things-Cloud-Service_RGB.pdf

[125] http://blogs.wsj.com/cio/2015/06/02/internet-of-things-market-to-reach-1-7-trillion-by-2020-idc/, last modified on 2 Jun 2015

[126] http://www.woodsidecap.com/wp-content/uploads/2015/03/WCP-IOT-M_and_A-REPORT-2015-3.pdf

[127] http://www.researchandmarkets.com/reports/3334586/carrier-b2b-data-revenue-big-data-analytics, last modified on July 2015

[128] http://www.ericsson.com/res/docs/2015/ericsson-mobility-report-feb-2015-interim.pdf, last modified on Feb 2015

[129] http://www.snstelecom.com/bigdata, last modified on Jun 2016

[130] http://docs.ckan.org/en/latest/user-guide.html#what-is-ckan

[131] http://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management/

[132] http://ckan.org/instances/#

[133] http://data.gov.au/

[134] http://opendatahandbook.org/glossary/en/terms/ckan/

[135] http://orbital.blogs.lincoln.ac.uk/2012/09/06/choosing-ckan-for-research-data-management/

[136] http://www.gartner.com/newsroom/id/3165317, last modified Nov 10, 2015.

[137] https://www.internetsociety.org/sites/default/files/IETF-Journal-apr2016.pdf

[138] https://irtf.org/t2trg, last modified Nov 15, 2016.

[139] https://www.w3.org/2016/07/wot-ig-charter.html

[140] https://tools.ietf.org, last modified Nov 6, 2016.