|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **International Telecommunication Union** | | |
|  | |  | | |
| **ITU-T** | **FG-SSC** | |
| TELECOMMUNICATION STANDARDIZATION SECTOR OF ITU | | (05/2015) |
|  | ITU-T Focus Group on Smart Sustainable Cities | | | |
|  | **Anonymization infrastructure and open data in smart sustainable cities** | | | |
|  | Focus Group Technical Report | | | |



FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of tele­com­mu­ni­ca­tions, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The procedures for establishment of focus groups are defined in Recommendation ITU-T A.7. ITU-T Study Group 5 set up the ITU-T Focus Group on Smart Sustainable Cities (FG-SSC) at its meeting in February 2013. ITU-T Study Group 5 is the parent group of FG-SSC.

Deliverables of focus groups can take the form of technical reports, specifications, etc., and aim to provide material for consideration by the parent group in its standardization activities. Deliverables of focus groups are not ITU-T Recommendations.

|  |
| --- |
| **SERIES OF FG-SSC TECHNICAL REPORTS/SPECIFICATIONS**  Technical Report on "Smart sustainable cities: a guide for city leaders"  Technical Report on "Master plan for smart sustainable cities"  Technical Report on "An overview of smart sustainable cities and the role of information and communication technologies"  Technical Report on "Smart sustainable cities: an analysis of definitions"  Technical Report on "Smart water management in cities"  Technical Report on "Electromagnetic field (EMF) considerations in smart sustainable cities"  Technical Specifications on "Overview of key performance indicators in smart sustainable cities"  Technical Report on "Information and communication technologies for climate change adaptation in cities"  Technical Report on "Cybersecurity, data protection and cyber resilience in smart sustainable cities"  Technical Report on "Integrated management for smart sustainable cities"  Technical Report on "Key performance indicators definitions for smart sustainable cities"  Technical Specifications on "Key performance indicators related to the use of information and communication technology in smart sustainable cities"  Technical Specifications on "Key performance indicators related to the sustainability impacts of information and communication technology in smart sustainable cities"  Technical Report on "Standardization roadmap for smart sustainable cities"  Technical Report on "Setting the stage for stakeholders’ engagement in smart sustainable cities"  Technical Report on "Overview of smart sustainable cities infrastructure"  Technical Specifications on "Setting the framework for an ICT architecture of a smart sustainable city"  Technical Specifications on "Multi-service infrastructure for smart sustainable cities in new-development areas"  Technical Report on "Intelligent sustainable buildings for smart sustainable cities"  Technical Report on "Anonymization infrastructure and open data in smart sustainable cities"  Technical Report on "Standardization activities for smart sustainable cities" |

© ITU 2015

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

**Anonymization infrastructure and   
open data in smart sustainable cities**

Acknowledgements

This Technical Report was researched and authored by Jie Chang (China Telecom), Lei Gu (China Telecom), Wei Liu (Shanghai Library), Kanae Matsui (Keio University), Hiroaki Nishi (Keio University) and Cuijuan Xia (Shanghai Library).

The authors wish to extend their gratitude to Ramy Ahmed (NTRA Egypt), Takafumi Hashitani (Fujitsu), Michael Mulquin (IS Communications), Ziqin Sang (Fiberhome Technologies Group), Mythili Menon (University of Geneva) and Junnosuke Sugiura (Fujitsu) for their valuable inputs and support.

Authors and contributors are listed in alphabetical order.

The authors acknowledge the guidance and support from all the members of the ITU-T Focus Group on Smart Sustainable Cities (FG-SSC) Management Team.

Additional information and materials relating to this Technical Report can be found at: [www.itu.int/itu-t/climatechange](http://www.itu.int/itu-t/climatechange). If you would like to provide any additional information, please contact Cristina Bueti (ITU) at [tsbsg5@itu.int](mailto:tsbsg5@itu.int).

Anonymization infrastructure and open data in smart sustainable cities

**Table of Contents**

Page

1 Introduction 2

1.1 Scope 2

1.2 Background 2

2 The demand of open data in SSC 5

2.1 The Contents of Open Data 5

2.2 The Objective of Open Data 6

2.3 Relationship between Open Data and Smart Sustainable Cities 7

3 The framework of open data 8

3.1 The Principles of Overall Design 8

3.2 The Framework of Overall Design 9

4 Open data handling in SSC 13

4.1 Barriers and Constraints in Open data 13

4.2 Security Protection and Privacy Preservation of Open Data 14

4.3 Policy and Regulation 15

5 Technologies related to open data in SSC 16

5.1 Metadata Management 16

5.2 Linked Data 17

5.3 Data Visualization 20

5.4 Social Networks 22

5.5 Anonymization technology 23

6 The management of open data in SSC 24

6.1 Executive Institution Management 24

6.2 User Management 25

6.3 Quality Management 25

6.4 Security Management 26

6.5 Operation Management 29

7 Use cases 30

7.1 Case study 1. Data.gov 31

7.2 Case study 2. Data.gov.uk 33

7.3 IT dashboard 33

7.4 Open Data Movement in China 34

7.5 Open Data Services with Privacy Data and Anonymization 36

8 Expected (Anonymous) applications of open data in smart sustainable cities 40

8.1 Application of Open Data on Smart Forecasts 40

Page

8.2 Data Anonymization for Smart Sustainable City 41

8.3 Application of Anonymization Method for a Smart Metering 45

8.4 Infrastructure of Secondary Use of Data 46

9 Milestone 52

Annex A – Application of anonymization for disaster recovery 54

Annex B – Abbreviations 56

Appendix I – Bibliography 58

Anonymization infrastructure and open data in smart sustainable cities

Overview

ITU-T Study Group 5 has established a new Focus Group on Smart Sustainable Cities (FG-SSC) to assess the standardization requirements of cities aiming to boost their social, economic and environmental sustainability through the integration of information and communication technologies (ICTs) in their infrastructures and operations.

This technical report focus on anonymization infrastructure and open data in smart sustainable cities (SSC), including the demand of open data in SSC, the framework of open data, the handling of open data in SSC, the technology of open data in SSC and the management of open data in SSC. Nevertheless, fruitful application examples are provided as use cases and expected (anonymous) applications of open data in SSC as well as anonymous application for disaster recovery.

Executive Summary

Focus Group on Smart Sustainable Cities (FG-SSC) has embarked on a journey for integration of open data and adoption of anonymization techniques for future smart sustainable cities.

This technical report has been specifically written for city officials, town planners, enterprise managers, developers, infrastructure providers, service providers, network operators, and citizens with the ultimate aim of ensuring data protection and improving data availability within their city of interest.

Open data is a current movement which requires machine-readable data to be freely available to everyone to use and republish as they wish, without restrictions of copyright, patents or other mechanisms of control. Currently, the open data system is utilized and is in high demand for creating new services from the level of local community to government. However, in order to beneficially utilize the open data, excessive exposure of private information which maybe included in the open data should be mitigated. Despite this need for protection of certain types of data, there are only few studies on open data utilization for smart sustainable cities (SSC), thereby making it difficult to get a comprehensive understanding of how these types of data can be both utilized and as well as protected.

To remedy the lack of credible research on this topic, this document provides a comprehensive study on open data in smart sustainable cities based on six aspects: (i) the demand of open data, (ii) the framework of open data, (iii) open data handling, (iv) the technology of open data, (v) the management of open data and (vi) application examples.

In addition, (a) practical methods of open data anonymization, (b) related information for the readers’ understanding that highlight the relationship between open data, anonymization, and (iv) methods of how to utilize the information practically are also provided.

The technical report further examines the following; (1) definition of the open data and its characteristic features and objectives to understand what types of dangers are predicted, (2) open data framework that provides a general structure to produce and apply open data, (3) related technologies of open data for maintenance and application, (4) management methods of the open data from the aspects of usage, quality, secturity, and operation, (5) use cases of open data driven by governments for smart forecast, city management, metering many types of data with an anonymization technique using information comunication technologies (ICT) infrastructure.

Further work will aim to analyse the impacts of open data anonymization and its effectiveness.

1 Introduction

1.1 Scope

This technical report proposes a summarized introduction on open data in smart sustainable cities and characterizes open data in smart sustainable cities based on the following six aspects (i) the demand of open data; (ii) the framework of open data; (iii) the constraint of open data; (iv) the technology of open data;(v) the management of open data; (vi) application examples of open data..

The target audience for this report include:

* City Officials
* Town planners
* Enterprise Manager
* Developers
* Infrastructure providers
* Service providers
* Network operators
* Citizens

This report provides guidance on implementation and promotes efficient deployment of open data in smart sustainable cities.

1.2 Background

Data is an extremely broad term, only slightly less vague than the nearly all-encompassing term information. Broadly speaking, data is structured information with potential for meaning.

Open data is data that can be freely used, re-used and redistributed by anyone – subject only, at most, to the requirement to attribute and share alike.

The term "open data" was initially used in the natural science field. The term was used to indicate basic and untreated scientific data. The first recognition of the use of "open data" to refer to a policy context, defining a scientific policy for a research project, was in the 1970s, during an international collaboration project in NASA[[1]](#footnote-1). The term “open data” becomes popular with the launch of open-data government initiatives such as Data.gov and Data.gov.uk [[2]](#footnote-2) [[3]](#footnote-3).

Currently, Open data is an international movement that certain data should be freely available to everyone to use and republish as they wish, without restrictions of copyright, patents or other mechanisms of control. “Open” in this context usually refers to machine processed online resources that are easy to access and are put under free licenses. These free licenses enable the re-use of data by anyone for any purpose at no charge, requiring at most attribution.

A summary of the main criteria for open data include:

* **Availability and Access:** The data must be available as a whole and at no more than a reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
* **Re-use and Redistribution:** The data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
* **Universal Participation:** Everyone must be able to use, re-use and redistribute . There should be no discrimination against fields of endeavor or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use or restrictions of use for certain purposes (e.g., only in education), are not allowed.

The idea of open government has an older origin, dated back to the early years following World War II at the time of the Great Depression in the U.S. There were increasing efforts to improve openness and the application of the right to know principle in the government during that period[[4]](#footnote-4) . The various components of open data as applied by National Government and the private sector are discussed in this Section. Though the term "open data" is not as old as open government. Open government data (OGD) is an important part of open data, which can be traced back to the 2003 Public Sector Information Directive by The European Commission[[5]](#footnote-5). As of January 2014, more than 700,000 OGD datasets have been put online by national and local governments from more than 50 countries. The first one-stop-shop open data portal Data.gov (www.data.gov) was launched in May 2009, as part of the open government directive of the Obama Administration. In the Memorandum on Transparency and Open Government, issued on January 21, 2009, the President instructed the Director of the Office of Management and Budget (OMB) to issue an Open Government Directive. Responding to this memorandum is intended to direct executive departments and agencies to take specific actions to implement the principles of transparency, participation, and collaboration set forth in the President's Memorandum[[6]](#footnote-6). By June 2014, the US open government data portal has already published 104779 datasets in a machine-readable format with 80486 geographical sets, 24293 non-geographical sets and additional 341 tools. In January 2010, the British government launched Data.gov.uk. The British open data portal applied Comprehensive Knowledge Archive Network (CKAN), which is an open‑source data management system (DMS) for powering data hubs and data portals. Thus, both open government and open data initiatives are not entirely a new concept.

At the conceptual level, the idea of openness has not significantly changed. For instance, the definition of open data in the current context still includes several important elements characterizing the preceding definition such as accessibility, availability, re-usability, re-distribution and participation[[7]](#footnote-7) in combination with the emphasis on information technology factors such as machine‑processable, non-proprietary and license-free[[8]](#footnote-8). On the other hand, openness as a concept is wider due to the emphasis on information technology usage and innovation.

Open Scientific Data (OSD) is the other important part of open data that came much before the inclusion of than OGD. The concept of open access to scientific data was institutionally established with the formation of the World Data Center system[[9]](#footnote-9) [[10]](#footnote-10), which was established by the International Council of Scientific Unions (now the International Council for Science) to minimize the risk of data loss as well as to maximize data accessibility, and further recommending that data be made available in machine‑readable format[[11]](#footnote-11) [[12]](#footnote-12). Additionally, in 2004, the Science Ministers of nations of the Organization for Economic Co-operation and Development (OECD), signed a declaration which essentially states that all publicly funded archive data should be made publicly available[[13]](#footnote-13) [[14]](#footnote-14). In 2007, the OECD Principles and Guidelines for Access to Research Data from Public Funding was also published as a soft-law recommendation[[15]](#footnote-15).

Open Industrial Data (OID) resulted from recent initiatives, including:

1. Vision 2020[[16]](#footnote-16): This is New York City's Comprehensive Waterfront Plan, which is led by the Department of City Planning of New York City. It sets the stage for expanded use of the waterfront for parks, housing and economic development, as well as waterways for transportation, recreation and natural habitats.
2. North Shore 2030: This initiative involves improving and reconnecting the North Shore's Unique and Historic Assets, and reviewing the City's industrial policies, including those relevant for Industrial Business Zones (IBZs) and Significant Maritime and Industrial Areas (SMIAs).

Another component of open data is open enterprise data (OED) which is best described as a way of doing business, forging strong relationships with the network of other organizations, customers, and potential customers. Alibaba, the China largest online B2C and B2B company, initiated Open Data Partnership Projects is an OED. In this project, ten open data partners are selected and desensitized online transaction data are open to these partners to innovate new applications.

Open data may need to be in a linked format or in another format that is easily readable by acomputer (for example comma-separated values (.csv),Excel spread sheet (.xls), or even PC‑axis(.px)formats. In addition, all websites and text documents are included. However, scanned paper documents(.pdf), or other image files are not considered machine-readable.

With the rise of smart phones and their built-in sensors as well as web-apps, an increasing amount of personal data is being silently collected. Personal data–digital information about users' location, calls, web-searches, and preferences–is undoubtedly the “oil” of the new digital economy. However, the lack of access to the data makes it very hard if not impossible for an individual to understand and manage the risks associated with the collected data. Therefore, advancements in using and mining this data have to evolve in parallel with considerations about ownership and privacy.

2 The demand of open data in SSC

2.1 The Contents of Open Data

Open data include various sources, including the Open Government Data (OGD), Open Industrial Data(OID), Open Enterprise Data(OED) and Open Personal Data(OPD). The OGD is the major part of open data because governments have accumulated a large amounts of data, have become the largest owner of data in terms of volume. Currently governments are assumed to publish open data to maximize public reuse, not only strengthen the transparency and promote efficiency and effectiveness in administration, but also to create economic opportunity and improve citizens' quality of life(QoL). The OGD includes geographical, environmental, weather, education, agriculture, and occupational safety as well as economic data, which help citizens to be more informed, and makes the government more efficient.

Open scientific data is another important source of open data, including experimental data, [genomes](http://en.wikipedia.org/wiki/Genome),[chemical compounds](http://en.wikipedia.org/wiki/Chemical_compound), mathematical and scientific formulae, medical data practice, bioscience biodiversity. Most of these fundamental researches are financed by governments and are funded for the purpose of disclosure of their works and face little limit for openness. Problems often arise in open industrial and enterprise data because these data are commercially valuable orcan be aggregated into works of value. In these cases, access to, or reuse of the data is controlled by organizations, including access restrictions, licenses, copyright, patents and charges for access or reuse. It is important that the data are re-usable without requiring further permission though the types of reuse (such as the creation of derivative works) may be controlled by a license. Open personal data is also used in research projects. Companies like Microsoft and Yahoo investigate their consumer internet behavior in accordance with their respective user approval policies.

It is important to note that data management from new aspects, especially, anonymization is an essential from a viewpoint of achieving open data management in smart sustainable city.

Various institutions such as medical facilities, transportation facilities, and government agencies must manage large amounts of data, which may include private customer information, medical records, and transaction information. This data, commonly stored in electronic form, often contains sensitive personal information. These types of data are useful in smart sustainable city establishments, and are frequently necessary, to facilitate the provision of advanced services. However, stored data may contain a considerable amount of personal and sensitive information about individuals. This information may include age, addresses as well as more sensitive items such as financial data, medical records, personal preferences and history of behavior. In the interest of the individuals, it is essential that the data containing sensitive information should be protected from unauthorized use.

Contrastingly, the organizations should provide the privacy data by transferring the data to useful services in smart sustainable cities. In providing the data, preserving privacy at a required level should be prioritized. It has the possibility that the two different data, generated by suitable anonymization process to the required anonymization level, reveal the original plain data. This situation can be harmful in publishing data even if the data is appropriately anonymized. Moreover, the following information is needed from the viewpoint of both original data provider and application servicer using the data. In providing data as open data, this anonymization process is indispensable to maximize the use of private data as rich services.**Contents of data**

Services or applications of smart sustainable cities mainly focuses on contents of data. The data can be separated into two parts: header as an index or tags from the viewpoint of data management and contents. Contents are the main part of data. For the entity providing an application and its user, the type of information included is considered important. The valuable application of the data strongly depends on this information.

* **Ownership of data**

For the application service provider of the data and its user, the data has to be reliable and used to address pertinent concerns. Data providers have to give the name of the organization in order to add value to the data. In short, the way to authorize the owner of the data is another issue for data services in a smart sustainable city.

* **Generation date/time and expiration of data**

For the application servicer provider of the data and its user, the date and time for when the data was generated is valuable information to determine the relevance and freshness of the data. In some cases, data analysis in its historical trend or changes is achieved. The date when the data was generated is fundamental information for this analysis. In addition to the information, the expiration date and time is required. Traffic information and market information can be the examples of data that the expiration date and time is indispensable.

* **Update of the contents**

Some data of smart sustainable city requires continuous updating. Instead of static information, historical information requires to be updated to keep the data fresh. The frequency or interval of this update may influence valuability of the data.

* **Anonymizer of the data**

From the application service provider of the data, and its user, the information regarding who or which organization anonymized the data is required to know whether the data is trustworthy and can be used An authorized anonymization servicer provider should anonymize the data for maximizing the value of the anonymized data. Namely, the way to authorize the anonymization servicer is another concerning in data services in smart sustainable cities.

* **Anonymized date/time of data**

The information of anonymized date/time may not be directly concerned with the applications. However, it is needed from the viewpoint of traceability of data processing. When some privacy pirating attacks occurs, the information becomes a significant source of its detection and prevention.

* **Anonymized method and level**

In data anonymization, the method of anonymization and the level of the anonymization is matter of concern to the data application providers and users. The level of anonymization could impact on the information loss of data[[17]](#footnote-17). The information loss is an index of similarity with the original plain data. The information could be changed or lost in anonymizing data, and information loss gives the level of the change or loss. The existing anonymization method and its level will be described in section 8.

2.2 The Objective of Open Data

In the open government point of view, the objective is to build more transparent, participatory and collaborative administration. However, in SSC, three other objectives should also be emphasized, one is to inspire innovation in the data industry and to create new businesses and building industry chain. The second is to discover deeply hidden relationships from multisource data. The third is to facilitate the citizen living with innovative applications of open data. Open data reduces barrier and cost for entrepreneurs and a new industrial chain of data consumption is built by combining the efforts of data creators, data collectors, data exchangers and data consumers in more efficient way. Innovation takes place in the new industry driven by open data and original applications are created that change the living. Open data and the open data driven-industry creates tremendous economic value. A report by Lateral Economics show that open data will lead to a Gross Domestic Product(GDP) increase of 13 trillion USD for G20 countries between 2015-2020[[18]](#footnote-18). The McKinsey Global Institute (MGI) predict that open data will increase direct and indirect benefits up to 3-5 trillion USD per year[[19]](#footnote-19) in seven areas including education, health, and transportation. Open Gov Data in European Union (EU) would increase business activity by up to €40 billion/year

Open data can help to make better decisions in ones life. It may also enable individuals to be more active in society. A woman in Denmark built findtoilet.dk, which showed all the Danish public toilets, so that people she knew with bladder problems could now trust themselves to go out more. In the Netherlands a service, vervuilingsalarm.nl, is available which warns you with a message if the air-quality in your vicinity is going to reach a self-defined threshold tomorrow. In New York, one may easily find out where to walk onesdog, as well as find other people who use the same parks. Services like 'mapumental' in the UK and ' mapnificent' in Germany allow you to find places to live, taking into account the duration of the individual’s commute to work, housing prices, and attractiveness of the residential compound. GoodGuide[[20]](#footnote-20)[4-4] take advantages of open data provides ratings for over 210,000 products. WeatherBill, the climate corporation based in San Francisco, use open data help farmers around the world adapt to climate change and increase their crop yields(see also Section 8.1).All the forementioned examples use open data. The NYU's GovLab launches Open Data 500[[21]](#footnote-21), which is first comprehensive study of U.S. companies including examples of 500 companies that use open government data to generate new business and develop new products and services.

2.3 Relationship between Open Data and Smart Sustainable Cities

Smart sustainable cities are innovative efficient, gree, safe, and livable place for citizens. Besides the importance of ICT infrastructure and newly emerging innovative technologies, typically including Internet of Things (IoT), cloud computing and so on, soft skills including multi-source information integration and coordination, real-time response is even more critical. In the traditional cities, before the informative age, information and data are dispersive and not open to the public. Even where the information belongs to the same owner like government, when they are collected by different agencies there could be problems in coordination and contributing to real-time response. In SSC, smart application based on open data enable car drivers find the optimal way to avoid traffic jams. City planning with open data in SSC is also an important direction. Cities, especially the metropolis, are divided in regions with different functions. Central Business District (CBD) is located in the center of cities surrounded by residential living areas.. In some cities, sub-centers of cities emerge to ease the population and traffic pressure in the center cities. Open data from operation business companies can help to understand the population density and dynamic behaviors in the cities, which can assist city management planning for new area development and district function positioning.

3 The framework of open data

3.1 The Principles of Overall Design

Overall design of open data framework follows the following basic principles:

* **Openness**

Openness is a necessary process for system interconnection and data interaction; that is an important prerequisite to improve the maintenance of the system. From many aspects, openness can be reflected from the design of system architecture, database systems, operating systems, network, etc. It is convenient to adopt modular framework for a third party supplying a cross platform development tool. Alsoit is a continuous process to improve API and services relying on the feedback of the public and government departments.

* **Extensibility**

One important aspect of designing a framework is making sure the extensibility of the framework has been carefully considered. There isno dispute that data volumes are growing exponentially. Huge amounts of data are generated every hour of every day, and this data comes from an ever-increasing variety of sources. This requires system design to be considered from two aspects: one is high scalability of hardware; the other is component-based design of software.

* **Accessibility**

Accessibility is the degree to which a product, device, service, or the environment is available to as many people as possible. Accessibility can be viewed as the "ability to access" and benefit from some system or entity. The design objectives of open data are to strengthen the accessibility, utilization and availability of data. In data publishing and permission management, it is effective to ensure that published data can meet the needs of users, and this requires it based on the maximum degree of data sharing. On the other hand, providing high-value data sets released, without involving privacy, confidentiality, safety issues, should be based on the needs of each government department and the public. In data presentation, the public can easily get what they need through open data portal.

* **Simplicity**

There is no requirement that every dataset must be madeopen right now. Starting out by opening up just one dataset, or even one part of a large dataset, is fine – of course, the more datasets one can open up the better. Remember, moving as rapidly as possible is good because it means one can build momentum and learn from experience – innovation is as much about failure as success and not every dataset will be useful.

* **Maintainability**

Maintainability involves a system of continuous improvement – learning from the past in order to improve the ability to maintain systems, or improve reliability of systems based on maintenance experience. In other words, the maintainability of the system expressed as the probability that a system will be retained in or restored to a specified condition within a given period or expressed as the actual operating costs of the system. The availability of the system can be reflected from two aspects: hardware and software. Hardware equipment should be convenient for installation and upgrade; it should also have sufficient spare parts. Software should provide flexible and user-friendly admin interface for operators.

* **Security**

Security provides "a form of protection where a separation is created between the assets and the threat." These separations are generically called "controls," and sometimes include changes to the asset or mitigrate the threat. Generally, it is useful to ensure absolute security through permissions setting, security certifications, anonymization technique and other means.

* **Advancement**

Advancement is responsible for maintaining the data in Advance Client/Server(C/S), and relations database. This responsibility involves processing all data and keeping the data in database current. The system will use object-oriented design; use the web access technology in user interface design and then fully guarantee the requirements of application systems design and development.

3.2 The Framework of Overall Design

The framework of open data is presented in the following Figure 5.2.1. The top layer consists of applications of data products to meets public need. The second layer is a one-shop portal providing services including data catalog, data services, app services, developer center, service center and map service. The third layer, which is the technology layer, consists of anonymization technique, meta data management, linked data technique, data visualization and social network technique. The fourth layer is open data management, with user management, quality management, security management and operation management. The original data comes from government, social groups, enterprise, and individuals. On the two sides of the framework, open data regularities and open data security system protect the security and privacy of data.

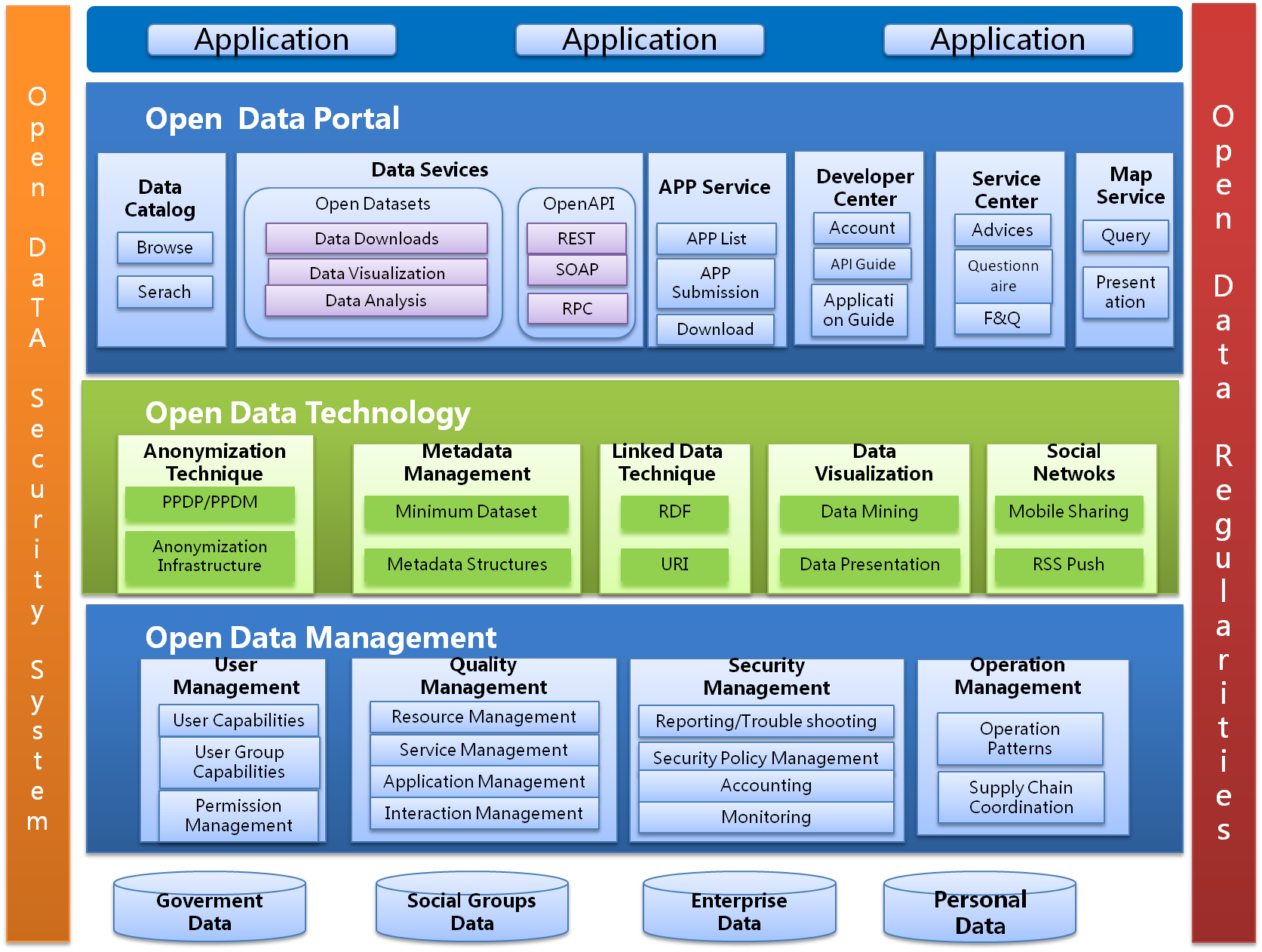


Figure 3.2.1 – The Framework of Open Data

The open data portal is the key part of open data which meets several goals:

* Offer a "one-stop-shop" experience to data consumers, saving the trouble of collecting data from various portals, authorities or offices with different controls and settings;
* Make it as easy as possible for government data stewards to make their datasets available for public consumption.

The open data portal provide four type of services, the data catalog for data sets browsing and searching, the open data service including data downloading, visualization and analysis as well as API access, the APP services and service center. Application developers access data through APIs like Representational State Transfer(REST), Simple Object Access Protocol(SOAP) or Remote Procedure Call Protocol(RPC). These developers create new web/mobile applications to facilitate service consumer and constitute and important part of the data industrial chain.

The use of data sets and App services can help users to obtain top recommended resources and applications that they may have a great of interest. Also, users can click on "More" buttons to thedemo page.

(1) Open Data Portal

Data Catalog

* Provide some levels of presentation, including unified list and isolated resource;
* Provide some levels of the classification system;
* Provide keyword and resource type to search;
* Provide topic, department, location, label or map to search.

Data Services

* Provide a variety of machine-readable data format, including XML, RSS, CSV/Text, KML/KMZ (Geospatial Data Format), ESRIShapefile (Geospatial Data Format);
* Provide the statistics of data download and data query;
* Provide the name, source institution, links, format, size, release data, etc.;
* Provide data related maps and visual display.

APP Service

* Provide some types of applications, including website and mobile app;
* APP List: include name, description, operating platform, download links, developer information and so on;
* APP Submission: include name, label, classification, description, version, charge mode, updated time, size, operating platform, download links and so on;
* APP Download: include name, label, classification, version, charge mode, rating, guidelines, and so on.

Developer Center

* Provide My Applications account;
* Provide related learning, exchanging, authorization, development and management of application;
* Provide different types of API for registered users to downloaddata or to call data: API download, API introduction, guide book, sample code, and other common problems.

Service Center

* Provide advices and Q&A through communication and interaction,includerecommendations for improvement, score and rank data resources and so on;
* Intercommunion and Interaction among users, including data resource evaluation;
* Intercommunion and Interaction between users and portal: include e-voting, online questionnaires, online communication;
* Provide personalized participation mechanisms through social network site (SNS) technology;
* Provide customized data based on ahot topic and user interest.

Map Service

* Provide data query services through binding data resources and geographic information, and then directly present on the map.

(2) Open Data Technology

The open data portal is supported by four main technologies, the metadata management, the linked data technique, the data visualization and social networks.

Metadata Management

* Provide minimum data set;
* Provide metadata structures.

Linked Data Technique

* The first step is ontology design, including vocabulary selection and vocabulary design;
* The second step is RDF structuration, including the type of data, data preparation, data storage, data publication and the type of publication;
* The third step is relationship establishment, including relationship links, links identification, vocabulary links and the type of relationship;
* The last step is data fusion, including crawling pattern, on-the-fly dereferencing pattern, query federation pattern and fusion mode.

Data Visualization

* Data mining;
* Data presentation.

Social Networks

* Mobile Sharing;
* RSS push technique.

Anonymization Technique

* Suppression;
* Pseudonymization (Hash function);
* Generalization

– Top/Bottom coding

– k-anonymity

– l-diversity

– t-closeness

– Non-intrusive Load Monitoring (NILM)

– Homomorphic encryption

– Randomization (Pk-anonymization)

– Perturbation

– Micro-aggregation

– Noise injection

– Data swapping

– Synthetic Microdata.

(3) Open Data Management

User Management

* Provide User Management Capabilities, User Group Management Capabilities, and Role-based (RBAC) Permission Management Capabilities;
* User management capabilities include query, new-built, account management, and so on;
* User group management capabilities include query, new-built, permission modification, and so on;
* Permission lists include function module uses-permission, data access permission, and so on.

Quality Management

* Provide Resource Management Capabilities, Service Management Capabilities, Application Management Capabilities and Interaction Management Capabilities;
* Resource management capabilities include resource exchange, resource synchronization, resource auditing, resource publishing, resource status monitor, and so on;
* Service management capabilities provide API interfaces for data resources so that developers can easily implement or attempt to access an external resource without download to local host. Under the premise of data update, it is possible to obtain the latest data resources;
* Application management capabilities mainly used to audit applications provided by developer, in addition to uniformly manage the on-line application of unified management, to monitor the entire application environment, and to timely investigate the abnormal operation of APP application;
* Interaction management capabilities mainly used to review applications of users, to respond to comments, to create questionnaires, to investigate statistical data, to update common questions and so on.

Operation Management

* Provide several operation patterns: include independent operation, cooperative operation, outsourcing operation, and so on;
* Provide supply chain coordination as well data flows both in and between links in the chain, which include government, enterprise, social groups and individuals.

Security Management

* Anonymization Technology is a part of security management in some studies;
* Anonymization technology is used for protecting original data. The main purpose of anonymization technology is to open data for everyone. The main point of anonymization for management is the level of anonymization. Although a strong anonymization makes it difficult to identify, it becomes difficult to use the data for services because it contains less information, vice versa.It is required to select best anonymization levels according to the target services.

(4) Open Data Security System

* Security system provides operation and system level security protection and privacy preservation in the open data system.

(5) Open Data Regularities

* Regularities provide top-down institution level security protection and privacy preservation in the open data system.

4 Open data handling in SSC

4.1 Barriers and Constraints in Open data

Data management is to do with managing the information life cycle which should include policies and processes for acquiring, validating, storing, protecting and processing data. The data infrastructure of smart sustainable city has to define the information and its expression as a clear format considering the interoperability with other standards. It also has to support the way to express the information as application programmable interface (API) for application programmers as a programming standard. As associated with these information, the information of management status is indispensable, such as duplication or backup layer of the data, storage of the data (datacenter or local server), latency and throughput to get the information, regulations in user, group, throughput, date or time, amount of fetching, count of accessing, ID or address in network and application layer, and accounting. Moreover, if API, security software/application, and other middlewares such as database, machine learning is used, the name and version of this software are required. This management can be the basement of privacy preserving. If the data management is not well organized, the following problems and risks will appear.

* **Falsifying data or illegal overwriting/deletion/wiping of data**

If appropriate security is not given, falsifying attacks damage data by illegal overwriting, deletion, and wiping.

* **Slow or imperfect recovery from attacks**

Data provider should provide the information on its data protection and security level before making it available.. In some cases, they will provide additional backups or options of security to meet their requirements. If it is not considered, slow recovery or imperfect data recovery may occur more serious than they expected.

* **Privacy invading**

Those who provide privacy data require a clear expression of the data, such as management status, number or type of applications, and number or type of users or the providing data.

* **Increasing the cost of data guarantee services and insurance services**

Data guarantee services and insurance services providers require to know the status of data management of smart sustainable city. If the status is not clear or missing, the cost of providing services will be increased. From the application viewpoints, these problems and risks may bring serious compromises by paralyzing infrastructures in smart sustainable city.

As described above, the infrastructure of SSC has to give a way to express the status of data management, the format of data contents, management information, authorization, and certification.

On that basis, the way to distribute data and to keep consistency of distributed data is required as a part of data infrastructure of SSC. According to the data distribution, hierarchical multi-grain network architecture for smart community is proposed by IEEE Standards Association[[22]](#footnote-22). Every service on a smart sustainable city should select the layer of the hierarchical network where a service is provided and processed. The consistency of the data is also managed by a distributed database in a hierarchical network architecture. This means discussion of locality and latency is the key to data consistency in the SSC. CAP theorem in theoretical computer science states consistency, availability, and partition tolerance cannot be guaranteed simultaneously, and one of them has to be omitted in designing systems. A model, which guarantees consistency and availability, supports hard real-time services with data consistency. In applications of SSC, the model is applicable to traffic signal control and power grid management of power stabilization service in a local area. However, single point of failure lurks in a system using the model because it does not guarantee partition-tolerance. A model, which guarantees availability and partition-tolerance, supports wide-area low-latency services such as naming rule service, sensor node management, and location services. In this model, consistency process becomes slower than other systems. A model, which guarantees consistency and partition-tolerance, supports wide-area low‑latency services such as trading, data broker service, and timing-critical data mining service. In this model, failure may degrade the functionality of separated subsystems. Some application will require the combination systems of two or more models.

These modes can be used at the same time. For example, a system locally guarantees consistency and availability and widely guarantees availability and partition-tolerance. Each service and application of SSC defines the marginal type of different models. Moreover, the service and application defines the service providing points in the hierarchical network structure, namely where the service and application are provided. The information infrastructure of SSC should have enough flexibility to manage all these model combinations and service providing points.

4.2 Security Protection and Privacy Preservation of Open Data

In open data issue, security protection and privacy preservation are crucial. In this globalized information society, it is impossible to attain security only by one enterprise of government because its complexity and meshed connections. To address this situation, new information infrastructure, and data processing rules are required. In some cases, the meaning of security protection in open data is equal to that of general security protection. However, open data is open to everyone, and technically it should be allowed to be accessed from anyone and anywhere. This means security issue is not as serious as generally discussed security. However, the access to the open data is regulated because of its license or its charge of usage. Accounting, usageconfirmation and illegal usage protection, could be the main issue of the security in using open data. Another security issue is original and unfalsified authenticity. This authenticity will be given by the technique of digital watermark, digital fingerprint with hash codes or the use of certificate authority system. Digital watermark and digital fingerprint are well-used technology to prevent falsifying. To use them, a common rule as standards is required to achieve an environment that everyone can check the original and unfalsified status of the data in the same way. For the use of certificate authority system, it requires a special organization like a certificate authority (CA) of public key infrastructure (PKI). The difference between the CA of open data and CA of PKI is that CA of open data focuses the point of data integrity in addition to the functions of CA of PKI, such as preventing spoofing, falsification, eavesdropping, and degeneration. CA of open data has to certify the data integrity whenever it is requested, and this means CA of open data has to manage all published open data and its fingerprints to verify the integrity.

According to the preservation of privacy, Privacy-Preserving Data Mining (PPDM)[[23]](#footnote-23)[[24]](#footnote-24) and Privacy-Preserving Data Publishing (PPDP)[[25]](#footnote-25)[[26]](#footnote-26) are well-known techniques. These techniques can mine or publish the data without personally identifiable information, thereby protecting the privacy. Anonymization is a practical technology that supports privacy protection. Anonymization technology can adjust to different privacy protection levels, thus providing flexible privacy protection. A considerable variety of studies on this technique have been performed owing to its high versatility. It is one of the most preeminent privacy protection technologies in current use. Generalization and deletion of the data are necessary to prevent privacy infringements. However, they reduce the value of the data. As a result, there is a trade-off relationship between privacy protection and the utilization of the data.

Although techniques such as PPDM and PPDP have been investigated in numerous studies, a method of securely publishing the data to enable secondary use has not been definitively established. This secondary use is the essential way of data to make interaction between different infrastructures. As shown in IEEE SMART GIRD VISION FOR VEHICULAR TECHNOLOGY: 2030 AND BEYOND, the future infrastructure exchanges data to use for inter-infrastructure smart services. PPDM and PPDP are an indispensable technique to maximize the distribution range of data. The anonymization method is the key of PPDM and PPDP. Anonymization enables the publication of private data by changing public data by omitting sensitive information.

However, from the viewpoint of security of anonymization, there is a possibility of privacy leak. After calculating and publishing anonymized data from a data source, another anonymized data set, calculated and published from the same source may cause a privacy information leak if an unauthorized person can access both sets of anonymized data. When calculating and publishing anonymized data, it is necessary to consider all of the previously published data from the same source. This leak will become larger when the data transaction in smart sustainable city becomes more active. To provide a way to protect the leak, new architecture of data management is required.

Considering these issues, it is crucial to establish a clear suggestion of technological guidance, an infrastructure, and a technical standard of protocols for the secondary use of data. The development of the protocol and infrastructure is especially important for the data infrastructure of smart sustainable city. It will facilitate collaboration between organizations that produce the data and the companies that require the data for secondary use, and thus increase their data publishing activity. It will develop new service and market for secondary uses of data in conjunction with advanced services such as market research, estimation of a route of infection, and traffic pattern analysis. Moreover, it will reduce the utilization costs for both providers and consumers of secondary use data, owing to the unification of data processing procedures.

4.3 Policy and Regulation

The Open Data policies are driven by a push for economic growth and job creation. President Obama made this clear when he announced his administration's new Open Data Policy in May 2013. This policy, which will make unprecedented amounts of federal data available in highly usable forms, has a business agenda first and foremost. Significantly, the President didn't make his announcement at a Washington press conference or in the Rose Garden but on a visit to a technology center in Austin, Texas. There he promised that governmental Open Data is going to help launch new businesses of all kinds in ways "that we haven't even imagined yet".

The Open Data Policy includes a detailed description of the criteria for government data to be released as Open Data, drawing on work done by the Open Knowledge Foundation in the United Kingdom, the Washington-basedSunlight Foundation, and others.

Technology of security protection and privacy preservation is the one wheel of a vehicle and the other wheel is political support by government. Open data and privacy issue conflict witheach other. To publish open data, private information has to be removed from the open data. Meanwhile, the value of open data will be degraded when private information is removed significantly. Although this trace-off problem can be ease by using technology, it cannot be perfectly solved only by using technology.

Political support is indispensable to give a guideline and guarantee the safe transaction of open data, namely publish and use of open data in smart sustainable city. Here, another trade-off comes. It is natural for data provider as the source of open data to minimize the leak of the information for preserving privacy. Political support gives the clear burden for fulfilling or negotiating both requirements of data provider and data services. This support should be varied according to the services and applications.

5 Technologies related to open data in SSC

It is found that Open Data is comprisedof a great diversity of research streams and related topics in SSC. However, most connected and influencing Open Data are the following technology streams.

5.1 Metadata Management

Metadata is "data about data", of any sort in any media[[27]](#footnote-27). In other words, metadata is data that describes other data, which facilitates the understanding, usage, and management of data, both by human and computers. Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier[[28]](#footnote-28). This commonly defines the structure or schema of the primary data. International standards apply to metadata. Much work is being accomplished in the national and international standards communities, especially ANSI (American National Standards Institute) and ISO (International Organization for Standardization) to reach consensus on standardizing metadata[[29]](#footnote-29).The core standard is ISO/IEC 11179-1:2004 and subsequent standards (see ISO/IEC 11179).Allt published registrations according to this standard cover just the definition of metadata and do not serve the structuring of metadata storage or retrieval neither any administrative standardization. It is important to note that this standard refers to metadata as the data about containers of the data and not to metadata as the data about the data contents. An important reason for creating descriptive metadata is to facilitate discovery of relevant information. In addition to resource discovery, metadata can help to organize electronic resources, facilitate interoperability and legacy resource integration, to provide digital identification, and to support archiving and preservation. According to NISO's definitions, there are three main types of metadata: descriptive metadata, structural metadata and administrative metadata[[30]](#footnote-30). Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords. There are many kinds of examples of descriptive metadata, such as CDWA (Categories for the Description of Works of Art), VRA (Visual Resources Association), DC (Dublin Core), FGDC (Content Standard for Digital Geospatial Metadata, for Federal Geospatial Data Committee), GILS (Government Information Locator Service), EAD (Encoded Archival Description), TEI (Text Encoding Initiative) and so on.Structural metadata indicates how compound objects are put together, for example, how pages are ordered to form chapters. Administrative metadata provides information to help manage a resource, such as when and how it was created, file type and other technical information, and who can access it.

Table 5.1.1 – Comparison and analysis of common metadata standards

|  | Applicable Type | User | Purpose |
| --- | --- | --- | --- |
| CDWA | Works of art, architecture, other material culture, groups and collections of works, and related images | Art historians, art information professionals, and information providers | Provide art categorization, make information of diverse systems both more compatible and more accessible |
| VRA | Works of visual culture as well as the images | Art Collection Organization | Description of works of visual culture as well as the images |
| DC | Online resources | Anyone, including experts, academics, students and library staff | Resource discovery |
| FGDC | Digital geospatial data | Government, research institute, and company | Share of geographic data, maps, and online services through an online portal |
| GILS | Federal information resources | Government | Identify, locate, and describe publicly available Federal information resources, including electronic information resources |
| EAD | Archival and manuscript collections at Harvard | Archives and manuscripts libraries | Materials, including letters, diaries, photographs, drawings, printed material, and objects |
| TEI | Electronic text | Libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation | A set of Guidelines that specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences, and linguistics |

Within open data initiatives/communities, metadata is used to support the description of data sets (including data services), as well as documents and applications. Only if metadata structure and meaning are sufficiently uniform or self-explanatory, a central portal can be realized, to consolidate various data offers and the contents of existing external metadata catalogs. The implementation of consistent metadata in SSC is often driven by public decision-makers, data providers, developers and other open data initiatives, or application requirements. Metadata can be the foundation of resource description that can facilitate a shared understanding across business and technical domains. Metadata focuses on the essentials along with great flexibility without wasting time to process and understand the described data. For that reason, making metadata machine readable greatly increases its utility, but requires more detailed open standardization.

5.2 Linked Data

Linked Data primarily describes the result of consistently applying semantic web principles and technologies when publishing structured data that allows metadata to be connected and enriched, so that different representations of the same content can be found, and links made between related resources. It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages for human readers, which extends them to share information in a way that can be read automatically by computers[[31]](#footnote-31). This enables data from different sources to be connected and queried. The exponential growth of subject-predicate-object expressions creating links between formerly disparate resources leads to what has been called the Linked Data cloud. Relentlessly, public and private organizations as well as individuals contribute their data following Semantic Web standards[[32]](#footnote-32). In 2006, Tim Berners Lee stipulated that interlinking all this data makes it more useful if 5 simple principles are followed:available, machine-readable, non‑proprietarydata formats, RDF data formatandinterlinked to other data by pointing at it[[33]](#footnote-33). Besides the large, global vision of linked data, its use in an organization to expose its public information, or even to manage internal data, brings new possibilities that traditional data management models have been notoriously bad at handling: It provides a model for naturally accessible and integrated data. In addition, the graph model it uses offers a level of flexibility that makes it possible to extend and enrich linked data incrementally, without having to reconsider the entire system: there is no system, only individual contributions.

As the SSC are a "system of systems", different systems give vast amount of information[[34]](#footnote-34). By using model smart city technologies, data amount increase more and more rapidly. This makes it possible to do many things that previously could not be done: spot business trends, prevent diseases, combat crime and so on. Managed well, the data can be used to unlock new sources of economicvalue, provide fresh insights into science and hold governments to account. However, the traditional data processing approaches cannot process such a vast amount of information. Big data is developed to deal this issue and make the city smarter than before. Linked Data makes the World Wide Web into a global database that we call the Web of Data. Developers can query Linked Data from multiple sources at once and combine it on the fly, something difficult or impossible to do with traditional data management technologies[[35]](#footnote-35).Many individuals and organizations collect a broad range of different types of data in order to perform their tasks. The Government is particularly significant in this respect, both because ofthe quantity and centrality of the data it collects, but also because most of that government data is public data by law, and, therefore, could be made open and made available for others to use. Linked data plays an important role in the construction and operation of the smart cities. When the smart city is constructed, open data can provide a large amount of data to assist the city planners and constructors. The citizens and city managers can make right decisions for city lives and managements.

Defining Standard Data layers and tools implemented for Open Data Portal can provide semantic agreement between heterogeneous data sources. These sources are mainly websites of different institutions and agencies, which offer data online in unstructured or semi-structured formats such as text documents, excel files or XML files. There are very few sources that can provide data structured in Entity-Relationship model. The importance of Standard Data layers to minimize the conflict of data generated by several Open Data Portals to publish their data using different models. Standard Data layers for Open Data Portal are divided into four layers as shown in Figure 5.2.1.



Figure 5.2.1 – Overall Standard Data layers for Open Data Portal

The first layer is raw data layer. According to Elmasru and Navathe andKent, data can be further classified into three components: structured data, semi-structured data, and unstructured data. Structured data are those organized in according toa rigid and pre-defined criteria such as respecting various fields (or attributes) of data, delimiting thes cope, domain (possible values) of data, data type, etc. This is the case, for example, with data involved in tables of Relational Databases (RDB) used by most institutions. The structured data extractor is responsible for data extraction of data stored on the RDB. An important aspect of data publication on the traditional Web is related to the loss of structure while transforming this data from RDB to Open Data Portals. These data are converted into current Web formats, making them unstructured. Semi-structured data are given in a way one cannot always predict all aspects of a given piece of data. Some of its general attributes can be known and required in advance, others added later depending on circumstances. References are an example of semi-structured dataset, containing fairly similar items. The semi-structured data extractor is responsible for data extraction of semi-structured data on the Open Data Portals. Unstructured data are those for which no scheme isspecified, containing only the content and a means of presenting it. An example is the text on a HTML page. The unstructured data extractor is responsible for data extraction of unstructured data on the Open Data Portals[[36]](#footnote-36).

The second layer is linked data layer. After gathering the raw data independently from the different sources, is to perform the converting from structured, unstructured and semi-structured data to semantic data. This converting is made by means of an ontology (e.g., vocabularies, taxonomies) that describes these data. To perform this converting, our approach makes a priori converting based on standards. This process is made by converting from the OWL ontology to RDF triples. We observed existing approaches to perform converting only from structured data to RDF.The RDF datasets are stored in a CKAN repository, which is made public and can be accessed via the CKAN web interface and CKAN API[[37]](#footnote-37)[7-10]. From a technical perspective, the objective is to use common standards and techniques to extend the Web by publishing data as RDF, creating well-formatted RDF links between the data items, and performing search on the data via standardized languages such as SPARQL query language for RDF, performing search on the data via standardized languages such as SPARQL query language for RDF. Query Interface, which enables the user community as well as the source institutions that offer these statistical data to pose queries upon it. This component consists of an online graphical interface as well as a SPARQL Endpoint. The results of a query may be displayed as structured Excel and RDF files tothe users. Query Interface layer is the sub-layer providing the open data, consisting of two components: SparQL endpoint and Query Processor. SparQL endpoint is the query interface of submission and retrieval results in open dataset submitted by Interconnect the dataset with other datasets. Query Processor analyze the SparQL queries to verify which artifacts stored in the semantic database will be used. There are two components: Query Analyzer and Semantic Reasoner. Query Analyzer analyzes SparQL query features to verification of the necessary elements to be used to return query results. Moreover, to improve the response time of a query uses the indices and metadata. Semantic Reasoner is responsible for generating knowledge derived from inference about the immediate knowledge. One consideration is that this mechanism degrades the performance of a query. So this mechanism will be activated dynamically according to the complexity of the query submitted. Interconnect the dataset with other datasets is the sub-layer that allows data fusion of semantic data[[38]](#footnote-38).The Open Data API is a RESTful, service-oriented platform that allows developers to easily access datasets and create independent services through these calls. REST uses the HTTP protocol and, as such, requests use the common URL format. The API provides simple methods that developers can use to tap into the functionality and rich datasets, and gather information, in JSON or XML format, related to different indicators and topics[[39]](#footnote-39).

A visualization service is delivered to the site and could include analytics, graphics, charting, and other ways of using the data. The enhanced visualization is built on top ofpublished APIs in collaboration with third party open source applications.

5.3 Data Visualization

Data visualization is a modern branch of descriptive statistics meant to allow people to both understand and communicate data clearly and efficiently via the data graphics selected, such as tables, maps, charts and so on[[40]](#footnote-40). In the context of SSC for visualization, part of thedata is stored in a digital file, typically in either text or binary form. Of course, potentially every piece of digital ephemera may be considered "data"—not just text, but bits and bytes representing images, audio, video, databases, streams, models, archives, and anything else. Nowadays, excess amounts of open data are overwhelming; raw data becomes useful only when we apply methods of deriving insight from it. Fortunately, we humans are intensely visual creatures. Only trained professionals can detect patterns among rows of numbers, but even young children can interpret bar charts, extracting meaning from those numbers' visual representations. For that reason, data visualization is a powerful exercise[[41]](#footnote-41). Visualizing data is the fastest way to communicate it to others. The definition of data visualization references its two objectives: understanding and communication[[42]](#footnote-42). This can also be referred to as exploratory and explanatory the visual representation of data respectively. This distinction is often overlooked, but it is extremely critical in the process of creating successful data visualization. If the motivation is to make sense of data, then the data visualization should be exploratory in nature. The process of visualization can help to see the world in a new way, revealing unexpected patterns and trends in the otherwise hidden data around us. At its best, data visualization is expert storytelling. However, if the analysis of the data is complete, then infographic should be used. More literally, visualization is a process of mapping complete data to visuals[[43]](#footnote-43). It is possible to craft rules that interpret data and express its values as visual properties. Without being conscious of one's data visualization motivations and goals, the process can be inefficient, misguided, or altogether unsuccessful.

Table 5.3.1 – Open Publishing of Ooof Data Visualization

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Revised Date | Production Name | Tag | Visualization Tools | Designer | First Time |
| 29-Sep-14 | World's Religious Numbers | World and Religion | yEd | Ooof | 2011-7-23 |
| 21-Feb-14 | India's religious development process | India, Religion, and Process | yEd | Ooof | 2011-7-23 |
| 21-Feb-14 | Journey to the West figure relationships | Journey to the west and figure | Gephi | Ooof | 2011-7-23 |
| 21-Feb-14 | 50 twitter's relationship Diagram | Twitter | Gephi | Ooof | 2011-7-23 |
| 21-Feb-14 | Pushing around 5000 years | History and 4000 | BB FlashBack | Ooof | 2011-7-23 |
| 21-Feb-14 | 107t twitter's relationship diagram | Twitter | Gephi | Ooof | 2011-7-23 |
| 1-Feb-13 | Wuxi training camp's NGO relationship | Wuxi and NGO | Gephi | Ooof | 2012-4-23 |
| 1-Feb-13 | Linfen's government budget information | Linfen, budget, and Government | Gephi | Ooof | 2012-4-23 |
| 1-Feb-13 | Relationship Visualization among linfen's 29 Weibo friends | Linfen, Weibo, and Realtionship | Gephi | Ooof | 2012-4-23 |
| 1-Feb-13 | Human emotion's spectrum | Emotion | Gephi | Ooof | 2012-4-23 |

Data visualization is increasingly becoming a popular way to portray and distribute complicated information in a consumable way to increase efficiency and work smarter[[44]](#footnote-44). By linking city's performance management systems to tools that provide instant access to current and historic records, more and more governments are allowing almost anyone to view and manipulate open data via vivid pie charts and line graphs. Users can even export raw numbers or high-impact graphics for use in meetings and communications materials. On one hand, the open platform is saving us hundreds of work hours by reducing the time that staff spends sorting through files of various data. On the other hand, as more citizens are using this technology we are starting to see many new inquiries come in from people interested in learning more about how smart city operates. By providing residents with the tools to visualize and work with open data to meet their individual needs, data visualization is not only increasing city management transparency and accountability but also enhancing the ability of city officials to be more responsive to citizens' needs.From other aspects, data always changes over time, only used static datasets barely reflect those little changes, the visual adjustments are made pretty with transitions, which can employ motion for perceptual benefit crosscutting all SSC domains.

5.4 Social Networks

Social networks are typically defined as a social structure made up of a set of social entities (such as individuals or organizations) and a set of the dyadic ties between these entities[[45]](#footnote-45)[7-18]. In other words, social networks are based on a certain structure that allow people to both express their individuality and meet people with similar interests. The social network perspective provides a set of methods for analyzing the structure of whole social entities, as well as a variety of theories explaining the patterns observed in these structures. The study of these structures uses social network analysis to identify local and global patterns, locate influential entities, and examine network dynamics. Social networking services (or social network sites) are websites or web-based services that allow people to build online communities where they can connect and interact with other people who have similar interests, identical backgrounds, or real-life connections[[46]](#footnote-46).The key components of a social network service are the ability for people to create a profile about themselves, the ability for people to create a list of other people who have a shared connection or similar interests, and the ability to view the connections made by other people in the network. Most social network services are web-based and provide means for users to interact over the Internet, such as e-mail and instant messaging. Examples of popular social network services include: Facebook, Twitter, LinkedIn, Pinterest, GovLoop, MySpace, and Friendster. Related to social networks, media-sharing services (video, photo/image, audio) are web-based platforms that allow people the ability to view, discuss, upload, distribute, notify and store digital content in a social environment. The power of media-sharing services is that they provide people or organizations a platform to discuss and disseminate information using multimedia content. Examples of various media-sharing services include: photo/image (Flickr, Photobucket, Picasa, SmugMug), video (YouTube, Vimeo, Veoh). Photo/image provides a visual representation to help people to understand easily and communicate with other people.

Without social networks, the ability for user-generated content to propagate and penetrate the public sphere would be seriously hindered. Therefore, citizen journalism or citizen reporting is heavily tied to user generated content and media-sharing services. Many governments have seen social networks and media-sharing as ways to disseminate the same information to different individuals and social groups. For example, Cisco's infographic on the Internet of Things explains the connotation and also forecasts that by 2020, there would be 50 Billion 'things' connected to the Internet. These things are interconnection of objects ranging from PCs, mobile, TVs, cars, vending-machines, cameras, alarm clocks, to even cattle and many more. The visual displays how these connected things could make a difference to daily lives. Along with multiple social networks presence: a Facebook page, a YouTube channel and so on. Many different governments and government agencies are now taking similar approaches to incorporation of social networks approaches like Facebook into their data and communication activities to promote access to and usage of open data.

5.5 Anonymization technology

Known anonymization technologies are as listed:

(1) Deletion of attributes

(1.1) Attributes suppression

To remove sensitive identifiers for protecting identification of personality.

(1.2) Pseudonymization

To replace sensitive identifiers or combinations of identifiers, such as name, date or birth to a code or number, etc.Hash function can be the candidate to calculate the code.

(2) Change of attributes

(2.1) Generalization

To replace an attribute to a generalized value or higher word in concepts. For example, 10-year steps, change cucumber to vegetables, etc. Rounding is a way of generalization.

(2.2) Top/Bottom coding

To put together small or large values into one attributes. Forexample, those who are older than 100 years is changed to ">100".

(3) Perturbation

(3.1) Micro-aggregation

After grouping the original data, each attribute of records in a same group is replaced by a representative value of the group.

(3.2) Noise injection

To add random noise into numeric attributes probabilistically.

(3.3) Data swapping

Stochastically swapping the values of the attribute between records.

(3.4) Synthetic Microdata

To create artificial synthetic data to be statistically similar to the original data.

(4) Other techniques

(4.1) Suppression or records

To delete records with a special attribute or value. For example, a record which has a value of more than 120 is deleted.

(4.2) Supression of cells

To delete sensitive attributes such as the attributes which should not be used for analysis.

(4.3) Sampling

To extract a value randomly from the entire original data at a constant rate or number.

(5) Advanced anonymization methods

(5.1) k-anonymization, l-diversity, l-closeness

These PPDP method is described in the section of Use Cases.

(5.2) Pk-anonymization

A method of anonymization that guarantee the probability to point out a personal record is less than 1/k in probability.

(6) Application specific anonymization methods

(6.1) Battery-load hiding

Battery-Load Hiding (BLH) techniques were proposed to ensure household privacy[[47]](#footnote-47)[[48]](#footnote-48). BLH hides or obscures real electric consumption by charging and discharging batteries, which are set with each household. This approach tries to keep electric power consumption data constant value and make it impossible to infer real usage of home appliances by NILM.

(6.2) Others

The anonymization methods explained in section 11 belongs this category.

6 The management of open data in SSC

6.1 Executive Institution Management

Open data is a valuable resource and a strategic asset to the government, its partners, and the public. Due to the existence of institutional mechanisms there are more barriers between information systems, data showing disorderly scattered islands situation, the actual utilization of data is very low. Scattered data means that the data from various agencies and institutions unable to form an open source browser-based map that enable citizens and enterprise hardly to visualize open data from agencies, institutions and other people. Also, a thorough analysis helps agencies and institutions pinpoint needs, priorities, and existing capabilities for open data. Before the analysis begins, however, it is wise to establish Executive Departments and Agencies for open data that we name it as Open Data and Manpower Bureau (ODMB) consisting of all important stakeholders for open data and data management processes. Many of the responsibilities of ODMB are the same, regardless of where the person falls within the organization. So the Heads of ODMB is a mandated role, and individuals appointed are responsible for ensuring that data is handled and managed appropriately. This means making sure that data is properly protected and that their value to the organization is fully exploited. Memorandum establishes a framework to help institutionalize the principles of effective information management at each stage of the information's life cycle to promote interoperability and openness. Whether or not particular information can be made public, agencies can apply this framework to all information resources to promote efficiency and produce value.

6.2 User Management

Broadly speaking, user management[[49]](#footnote-49) encompasses the processes and technologies that allow an organization to more securely and efficiently.

* Add, create, and delete users from its systems;
* Provision all the applications and resources a user needs;
* Enable users to manage their own profiles using self-service techniques.

6.3 Quality Management

* Provide Resource Management Capabilities, Service Management Capabilities, Application Management Capabilities and Interaction Management Capabilities;
* Resource management capabilities include resource exchange, resource synchronization, resource auditing, resource publishing, resource status monitor, and so on;
* Service management capabilities provide API interfaces for data resources so that developers can easily implement or attempt to access an external resource without download to the localhost. Under the premise of data update, it is possible to obtain the latest data resources;
* Applicationmanagement capabilities mainly used to audit applications provided by developer, in addition to uniformly manage the on-line application of unified management, to monitor the entire application environment, and to timely investigate the abnormal operation of APP application;
* Interaction management capabilities mainly used to review applications for users, to respond to comments, to create questionnaires, to investigate statistical data, to update common questions and so on.

(1) Data Acquisition Mechanism

Establishment ofbiodiversity data acquisition mechanisms of multi-side participation, division of labor, coordination of work and high effective operation, which is the effective method in the case of data comes from different sectors, areas, and geographies.

VerticalData Link, primarily consistof all levels of the data set, covering the areas of public safety, public services, transportation services, education technology, financial services, energy and environment, health and hygiene, culture, entertainment and other fields;

Horizontal Data Link, primarily by the way links can point to other open platforms, and then portals integrate applications and information and present them to the end users as one unified view.

Open data portal mainly collect and manage sources of data that is produced by various government departments, enterprises, social groups and individuals.

Government departments, enterprises, social groups should regularly publish their open-access data sets through open data portal, but these data sets should not appear before, or as non-downloaded format appeared on the network.

In order to make all government departments, enterprises, social groups are willing to and can add their data sets throughopen data portal,it should provide a variety of mechanisms to greatly ease publication process.

(2) Catalog System

Through a comprehensive Classification System of Data, Open data should provide integrated management features for data resources on the platform, and then according to different classification multiple formats of data resources should be compiled into the same level of catalog, such as Raw Data Catalog, Tools Catalog, Geographic Data Catalog, and so on.Data resources can also be classified based on a different topic and different institutions. By two-level catalog system, the classification of data resources should be set up macroscopically and microscopically, not only make clear of data resources appearance, but also facilitate discovery and use of data resources. This can easily improve the quality of open data portal and promote high efficiency of data resources.

(3) Data Availability Management

Establishing a strict review mechanism to avoid published data malicious use due to legal rules, such as privacy concerns, confidentiality, security,and so on. Government departments should monitor the quality and quantity of the published data so as to ensure full implementation of data publishing and to ensure public needs for relevant data.

Establishing a set of effective approachesfor government departments involved in open data, this can provide available indicators to measure the availability of published data, such as web site availability, expected response time, quantity of published data sets, APIs runtime, and so on.

Government departments should publish data as a priority, according to main responsibilities, institutional strategy, and public needs, in order to speed up the release of high-value data sets.

(4) Data Usage

Achieving open data platform statistical functions for key indicators to reflect the situation of data publishing and data using, this can provide the basis for feedback and then promote the quality of data platform construction. The key indicators of statistical functions includes page views, data downloads, API calls number, user rating statistics, top ranking data resources, government departments published data statistics, and so on.

Usage metrics should include daily visits of monthly statistics, user visits of monthly statistics, data sets downloads of monthly category statistics, the trend of monthly downloads, monthly views of web site, top 10 downloaded data sets, and so on.

(5) Data Accessibility

Data accessibility can be measured by the feedback of public information. Public users can comment and report issues based on the clarity and completeness of existing published structured data. Other metrical indicators should include setting mode of search keywords, use the degree of semantic web technology, detailed rating of datasets.

6.4 Security Management

Data security management is a way to maintain the integrity of data and to make sure that the data is not accessible to unauthorized parties or susceptible to corruption of data[[50]](#footnote-50). Data security is put in place to ensure privacy in addition or protecting this data. Data itself is stored on network servers, possible personal computers and in the form of columns and rows. This data can be anything from personal files to intellectual property and even top-secret information. Data can be considered as anything that can be understood and interpreted by humans. Because the internet is a growing phenomenon, there was and always will be an emphasis on protecting personal or company data. Computer users, as time goes on, tend to be slightly more aware of their files, but are still encouraged to use some sort of data security. Data security methods can be acquired by using specific software solutions or hardware mechanisms[[51]](#footnote-51).

Data can be encrypted or unreadable to a person with no access. When encrypting this data, mathematical sequences and algorithms are used to scramble data. Encryption allows only an approved party to decode this unreadable text with a key. Only those that have this key can access any information. Authentication is another form of data security to be used for more daily access[[52]](#footnote-52). A sign-on to an email account, bank account etc., only allows the user with the proper key or password. The most commonly used method of keeping data protected is with data security software. This software keeps unauthorized parties from accessing private data and offers a variety of different options. Some of these options include requiring a sign-on to email accounts, rewriting of software, and being able to control security options remotely. Data can also be protected with IP security. This means that data can be protected from a hacker while in transit.

One of the biggest reasons to keep data protected is because there are many corporations that hacker want to target and breach. Data security tends to be necessary for large businesses, but the small ones usually have fewer infrastructures in place, making the information, not a great loss if breached. Depending on the services and content that is to be protected, there can be preventative measures to protect the information further. For example Windows Rights Management Services (RMS) can be set to control whether or not the recipient of an email can be read and viewed, edited, copied or saved; these setting can also set an expiration date of a specific document[[53]](#footnote-53).

By keeping data secured, it is possible to give different access to different people. For instance, sales associates can have access to their sales databases but are unable to access another sales associate information or business information (e.g., accounts payable, accounts receivable). Creating a single storage location (or server) for the data, and assigning individuals with different access, keeping up with data is a breeze. It makes it easier to maintain the data and permits a quick transfer to another storage location if needed. Data security software can also serve as a source to make secure sites (that give access to data files) can only be accessed by authorized personnel.

Data Security and Privacy

Open data should focus on the application of information security standards and legal institutions of network security. The establishment of confidentiality rules and regulations for information resources, not only to strengthen the protection of information security, but also to strengthen the protection of user's personal information[[54]](#footnote-54). At the same time, the improvement of legal institutions can provide institutional safeguards.

Government departments, enterprises, social groups should protectpersonal privacy when they collect personal information, each department and agencies shall not publish personally identifiable information on the platform, also shall not violate national laws and legal rules by any means. All published data should comply with all related security and privacy requirements.

Privacy protection that is not relevant to data should guarantee the feedback provided by the platform is anonymous without recording trace information or identification information.

Data Auditing Mechanism

According to legal requirements, government departments, enterprises, social groups and individuals should review procedures involving data sets of national security and privacy[[55]](#footnote-55).

The establishment of data anonymization mechanisms to prevent linking attacks. Even if any single data set may not pose a threat to national security, or it may not cause a risk of leak privacy, it may increase this risk through a lot of published data sets. Therefore rigorous review procedures can help to reduce the risk, preventing sensitive information, personally identifiable information and national security-sensitive information being leaked intentionally or unintentionally.

The definition of secondary use of data is to permit the use the data for purposes not limited to the primary and original use. Secondary use of data by interacting data generated by different infrastructures is expected to create new services and businesses in smart sustainable city. Secondary uses of data, including location information recorded by mobile phones and data from electricity smart meters, are under consideration for new services. The location data of mobile phones will reveal the daily travels of their users. For example, some car navigation systems utilize mobile phones to connect to datacenters, and, therefore, it can obtain the car's location and other relevant data. The primary purposes of these data are to track the requirements of car's maintenance and to facilitate road services for drivers. By analyzing the data, it is possible to obtain the driving speed and location of the car. In addition, analysis of this data can identify intersections where drivers frequently brake in a sudden manner. a road maintenance squad can check the intersection by utilizing this information, where they may identify problems such as hidden or missing signs. Data from a smart meter can provide information about the daily activities of the household. Remote observation services that monitor elderly parents attract significant attention in an aging society.

Moreover, by analyzing household electric power consumption data, security companies may provide a service that alerts by e-mail when there is no consumption or when consumption is higher than usual when residents are not present at home. In addition, cleaning service of heating, ventilation, and air conditioning appliances can be provided. Such a service can use the air conditioning electric power consumption data to determine clogged filters. Eco-point services, such as discount coupons for various services, can use the data to determine incentives for households that avoid peak use of electricity. Various service providers, such as food service outlets, can also cooperate and share data with electric power companies. These examples demonstrate that the secondary use of data can potentially create new services while enhancing the data's value. From numerous viewpoints, the secondary use of data is under consideration, and its demand is increasing.

However, it is possible to know what kinds of home appliances are used in the house. Moreover, the family configuration and estimation of income could be analyzed from such data. In a smart grid and clean power conference in Britain, an executive of Siemens Energy said "We, Siemens, have the technology to record energy consumption every minute, second, microsecond, more or less live.From that, we can infer how many people are in the house, what they do, whether they're upstairs, downstairs, do you have a dog, when do you habitually get up, when did you get up this morning, when do you have a shower: masses of private data.". If such information is revealed, it may become a threat; e.g., a thief may enter the house when the residents are regularly absent.

In equal measure, this secondary use of data can result in privacy problems. In the previous examples, the location data produced by a smart phone reveals the user's location at a given time. The amount of electricity usage recorded by smart meters may reveal excessive power consumption by the household, potentially revealing their high-income status. Moreover, it is simple to publish sensitive data utilizing the Internet without proper regard to the privacy. If access to this information is not adequately restricted, it may promptly result in its unauthorized use. Aside from its usefulness, publishing the data may result in the infringement of privacy rights. Therefore, techniques for publishing the data while simultaneously protecting the privacy are required for the safe secondary use of the data.

6.5 Operation Management

Data operation management focuses on the delicate data management of internal business processes to produce and distribute products and services. Some of activities that are covered by data operation management include data creation, development, production and distribution. Other data operation management activities include managing purchases and evaluations. A great deal of the focus of data operation management is on the efficiency and effectiveness of data's processes. Therefore, data operation management often includes substantial measurement and analysis of internal processes. Ultimately, the nature of how data operation management is carried out in a city depends much on the nature of products or services. As with all forms of management, data operation management needs to be tailored to meet the specific needs and requirements of a city. Rather, it is gained through the utilization of thoroughly developed methods and processes, and shared with all members. Many factors need considering when planning, implementing and continually developing operational processes.

Supply chain management is defined as the management of data as well data flows both in and between links in the chain, which include government, enterprise, social groups and individuals. The key issue for successful supply chain management is the effective full-scale coordination between these different partners. Such relationships are dependent on the data sharing. Issues such as purchasing prices and the levels to be purchased, as well as, storage of raw data, and other product components are to be overseen. From an operations viewpoint, all of these various processes must be reviewed frequently and improved constantly in order to ensure 'smooth', efficient operations within the city.

This report gives a specific description to three-level system of nationality, state/province, and city. The government is in the core status, enterprise and social groups are the subjects of data resource exploitation and utilization. Data resource exploitation and utilization should increase public interest as a prerequisite and should not harm the privacy of owner's data source and national security.

Open data is a kind of service system which has universal social significance. It is used not only to increase public welfare of social members, but also to enhance the competitiveness of a specialist area, or the competitiveness of the region. At the background of opennessgovernment and third-party organizations should develop a win-win cooperate mode, andthen jointly maintain and developopen data.Through the cooperation establishing a service network covering the whole of society to benefit the public, this can provide the basis for social development and transformation.

Principles of business model for open data are government dominant, broad participation, and interactive sharing.

There are three specific patterns:

1. Independent Operation

Independent operation pattern totally relieson the strength of government departmentsto operate web site. At present independent operation, pattern is often used in thegovernment website, which requires sufficient human resources, financial guarantee, and technical support. However, this can give increasing fiscal burden to the government, and also is likely to cause technical problems and security technology due to lack of human resources.

2. Cooperative Operation

Cooperative operation pattern transfer some sections of the web site to third parties, not only ensures that government seize the initiative inthe central part, but alsoensure that other organizations work on what they love and master by a third party.Thus it will be conducive to daily operations and update maintenance of open data, and to reduce the pressure of government staffs. In addition, to promote the development of value-added utilization of data resources, it should encourage the participation of the private sector and non-governmental organizations, simultaneously promote fair competition in the public and private sectors. Due to remarkable gap between governmental data resources and diverse demand of the general public, some of the data should be professional analysis and research because most userscan not directly use the original data resources provided by Government. Through filtering, organizing, and then processing the original data resources, third-party organizations can provide secondary use of data or services for individuals or other organizations to use, this can bridge the gap between demand and supply.

3. Outsourcing Operation

Outsourcing operation can improve the efficiency of maintenance management through guaranteed services of professional organizations, this is a good way to cost savings and boost industrial development. Meanwhile government has a distinct advantage and effects in efficiency improvement, structures implifies, costs reduction and professional services development.

From now on, more and more open data projects gradually from independent operation pattern to cooperative operation pattern.

7 Use cases

Some use cases of open data are introduced in this Section. As mentioned previously,. open data are defined as the idea that the data are available for anyone can use for any purpose at no cost. Additionally, open data do not request users any management protocols such as copyrights, patents for reusing and republishing. "Open Knowledge Definition" is well-known definitions how the data can be said as open data. Table 7.1 shows 10 definitions of open data[[56]](#footnote-56). These definitions were created for smooth data use in this information society.

Table 7.1 – Open Knowledge Definition

|  |  |  |
| --- | --- | --- |
| 1 | Access | "Access" indicates that available as a whole as and at no more than a reasonable reproduction cost, preferably downloading via the Internet without charge. The work must also be available in a convenient and modifiable form. |
| 2 | Redistribution | "Redistribution" indicates that open access do not restrict any party from selling or giving away the work either on its own or as part of a package made from works from many different sources. The license shall not require a royalty or other fee for such sale or distribution. |
| 3 | Reuse | "Reuse" indicates that open data allow for modifications and derivative works and must allow them to be distributed under the terms of the original work. |
| 4 | Absence of Technological Restriction | "Absence of Technological Restriction" indicates that open access providers in such a form that there are no technological obstacles to the performance of the above activities. This can be achieved by the provision of the work in an open data format, i.e., one whose specification is publicly and freely available and which places no restrictions monetary or otherwise upon its use. |

Table 7.1 – Open Knowledge Definition

|  |  |  |
| --- | --- | --- |
| 5 | Attribution | "Attribution" requires as a condition for redistribution and reuse the attribution of the contributors and creators to the work. If this condition is imposed, it must not be onerous. For example, if attribution is required a list of those requiring attribution should accompany the work. |
| 6 | Integrity | The license requires as a condition for the work being distributed in modified form that the resulting work carry a different name or version number from the original work. |
| 7 | No Discrimination Against Persons or Groups | The license does not discriminate against any person or group of persons. |
| 8 | No Discrimination Against Fields of Endeavor | The license does not restrict anyone from making use of the work in a specific field of endeavor. For example, it may not restrict the work from being used in a business, or from being used for genetic research. |
| 9 | Distribution of License | The rights should be attached to the work must apply to all to whom it is redistributed without the need for execution of an additional license by those parties. |
| 10 | License Must Not Be Specific to a Package | The rights should be attached to the work must not depend on the work being part of a particular package. If the work is extracted from that package and used or distributed within the terms of the work's license, all parties to whom the work is redistributed should have the same rights as those that are granted in conjunction with the original package. |
| 11 | License Must Not Restrict the Distribution of Other Works | The license does not place restrictions on other works that are distributed along with the licensed work. For example, the license must not insist that all other works distributed on the same medium are open. |

One of brilliant achievement of arising open data movement was "Data.gov" that is a portal site established in the United States in 2009. The movement was provoked by "Open Government Initiative" that President Obama declared for smooth data use. Examples of the data utilization are (1) making web applications for business, (2) utilization of research, for instance, predictions of climate change.

The data include high publicness data, for instance, government data. More detailed information will be described in 7.1 that describes Data.gov in the United States. After this movement, the other countries have started to develop portal sites to provide the government data. The detailed information will be introduced among 7.2 that describes the U.K case, 7.3 that describes Japanese case, and 7.4 describes Chinese case.

7.1 Case study 1. Data.gov

Data.gov is a portal site, which can download many types of government data for the utilization for businesses, researches, and the other purposes. A background of arising open data movement was "Open Government Initiative" that President Obama declared on May 2009. This initiative has three main policies (1) transparency, (2) participation and (3) collaboration[[57]](#footnote-57). After the initiative, "Data.gov," which is a portal site that people can download the government data sets. The government data indicate that the federal offices have collected, for instance, data fromnational census, environment, and economic conditions. Figure 7.1.1 is a capture of the website.

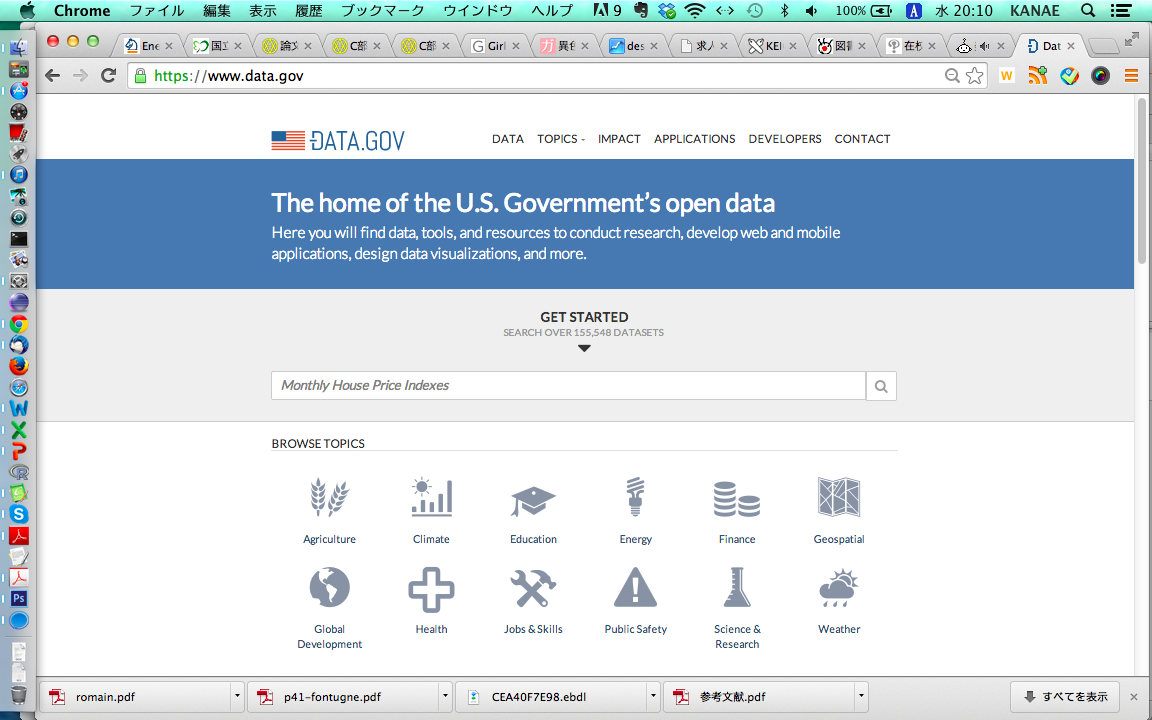


Figure 7.1.1 – Capture of the Postal Site, Data.gov

The government presented "Open Government Directive" to each the federal offices for sharing the data among any people, researchers, and companies via the Internet on December 2009. After the directive, Data.gov has published the government data, which categorized four data of raw, geo, data tool, and interactive. Users can select the type of data, public institution, data categories, topics, and web domains. Table 7.1.1 shows examples of data categories.

Table 7.1.1 – Example of Data Categories

|  |  |
| --- | --- |
|  | Categories |
| 1 | Transportation |
| 2 | Environment |
| 3 | Crimes |
| 4 | Medicine |
| 5 | Education |
| 6 | Nutrition |
| 7 | Security |

After sharing the data, new businesses have arisen, which were similar movements like that many weather news businesses started after National Oceanic and Atmospheric Administration opened their data. As the recent case, many new businesses have started up with data of "Health Data Initiative" shared[[58]](#footnote-58).

One of the examples to utilize the open data of energy usage is "Opower"[[59]](#footnote-59). They have used energy consumption data, which are collecting by smart meters, and made reports how to conserve the energy for companies and households. Their analyzed data are almost free. Therefore, their running costs are dramatically low compared with before. They have expanded their business to seven countries and had 16 million customers in 2013.

7.2 Case study 2. Data.gov.uk

"Data.gov.uk" is a portal site that the government has shared public data. However, their purpose includes enhancing the people's understanding how government works and how policies are made. Data.gov.uk is placed as a key part of the government's work on "transparency" between the government and people. Open Data team in the Cabinet Office has led Data.gov.uk. Figure 7.2.1 shows a capture of top page, Data.gov.uk.

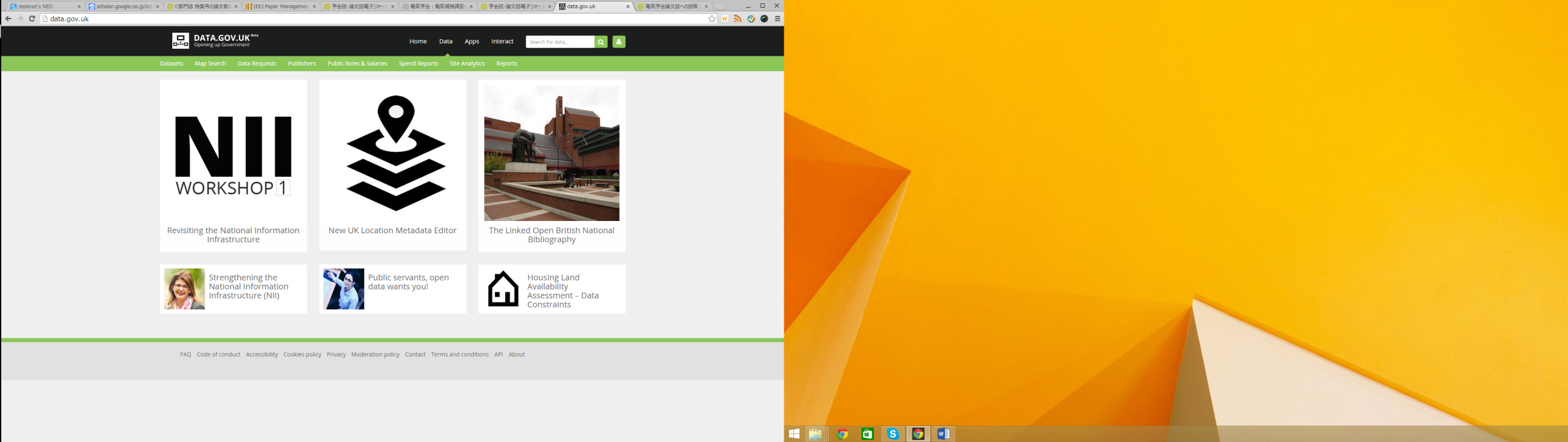


Figure 7.2.1 – Capture of Postal Site, Data.gov.uk

Hear, more about the Government's Transparency agenda from the Prime Minister in the UK[[60]](#footnote-60). Data.gov.uk provides datasets available from all central government departments and a number of other public sector bodies and local authorities. From the portal side of data.gov.uk, users can access the raw data driving government forward like Data.gov provided by the United States. Examples of general public information are tax credits or tax your car. So, the portal site enhances the utilization data not only for

These data sets are allowed to use to analyze trends over time from one policy area or to compare how different parts of government go about their work. Technical users will be able to create useful applications out of the raw data files.

7.3 IT dashboard

"IT dashboard" is a portal site that National Strategy Office of Information and Communications Technology in Japanese government created. This movement arose by "Declaration to be the World's Most Advanced IT Nation" on June 2013[[61]](#footnote-61). Figure 7.3.1 shows a top page of IT dashboard.

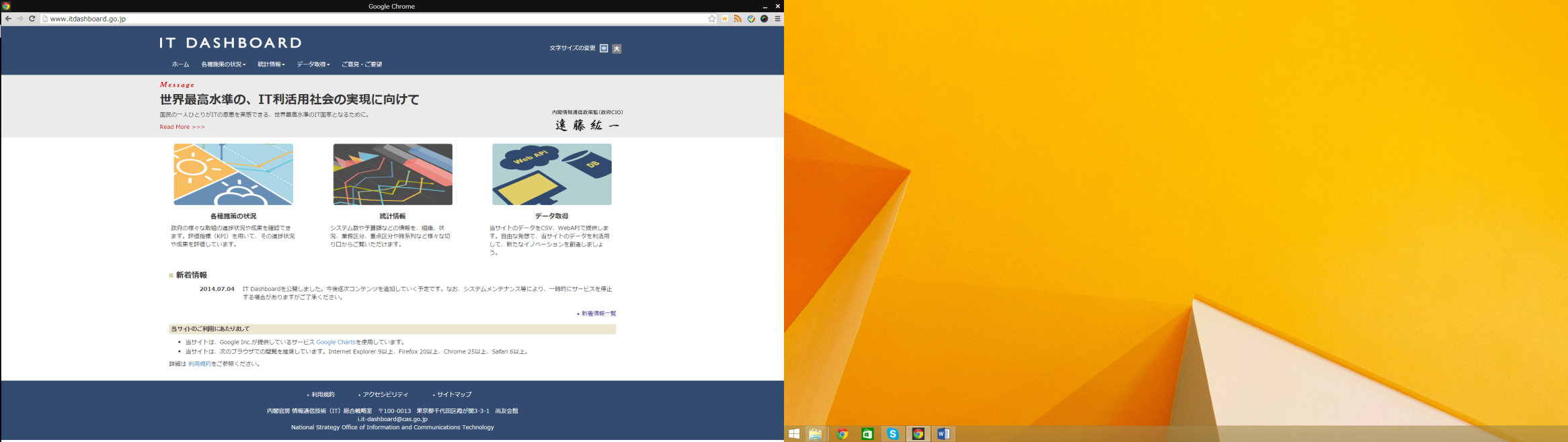


Figure 7.3.1 – Capture of Postal Site, IT Dashboard

This declaration indicated that people living in Japan would be contributed by ICT; therefore quality of ICT should be higher. The purposes of this declaration are introduced below:

(1) Contributing to the creation of new industry and growth in all industry fields through the creation of innovative technologies and integrated services that will enable the public to experience personally the recovery of the Japanese economy through the use of IT and data;

(2) Contributing to the improvement of the world's safest and most disaster-resilient society where people can live safely, with peace of mind, and comfortably; and

(3) Contributing to the provision of electronic government services and government reform from the perspective of user public to enable one-stop public services that anyone can access and use from anywhere.

In order to provide the data sets, the government developed a website "IT dashboard"[[62]](#footnote-62). 15 federal offices shared the data sets via the website. Categories and percentages of the open data are geographical data (33.0%), static data (28.1%), white paper (14.1%), protection against disasters (14.7%), technical data of using IT dashboard (7.0%), and the other (2.2%). Used data formats are pdf, html, xls, gif, csv, pdtxt, zip, epub, jpg, xlsx, ppt, docx,pdf", and mp3.

7.4 Open Data Movement in China

There were only three non-user friendly government open data sites and a smattering of open data enthusiasts who often had to find their own data sources and even create hardware to generate their own data in China. They were not a formally connected group but rather, individuals who created open data apps out of personal interest. Now, the recently launched Open Data Community is trying to create a multi-disciplinary network of businesses, research institutes, and NGOs interested in open data.

The Open Data Community is currently working on three projects, one of which is a comprehensive timeline of open data in China where OFKN China has potentially traced the open data "movement" to its beginnings.According to the timeline, the Chinese government's first open data website was Shanghai's Internal Data Directory launched sometime around September 2011, though the date is not clear. The government does little to publicize the launch of these sites. The current data list includes 425 data sets. The Shanghai government later released some data on the Shanghai data portal (see Figure 7.4.1), launched in December 2012. Beijing's open data site went online in October 2012 with 4,000 datasets to date, followed by that of the National Bureau of Statistics site in September 2013.

Several months later, China held its first ever hackathon using public data called Code for Climate Change. The creator of a hackerspace called Xindanwei, meaning 'new work unit,' which is a play on the government work units. This is the first time that the government is providing all this data to the start-up and creative community and is working together with them by providing data sets. Also, top researchers from all over China are providing insights and knowledge.

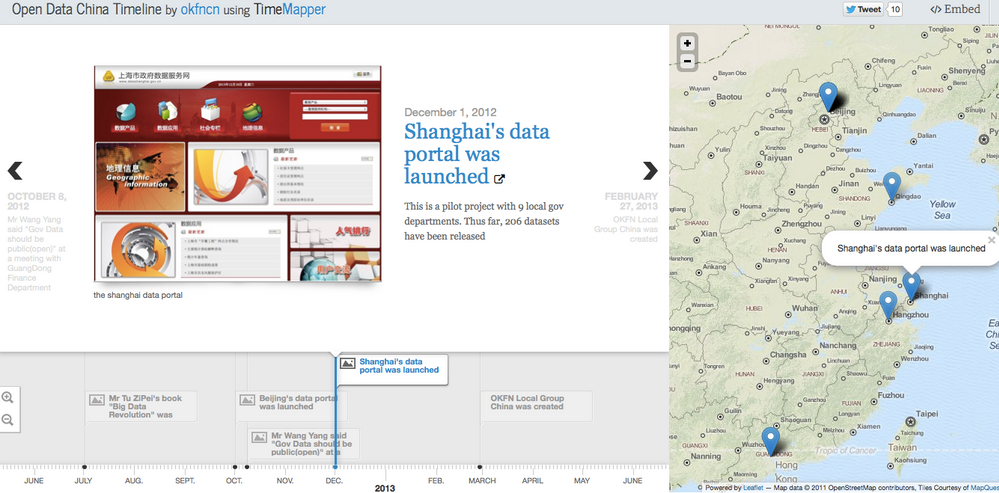


Figure 7.4.1 – The timeline provides the history of open data in China

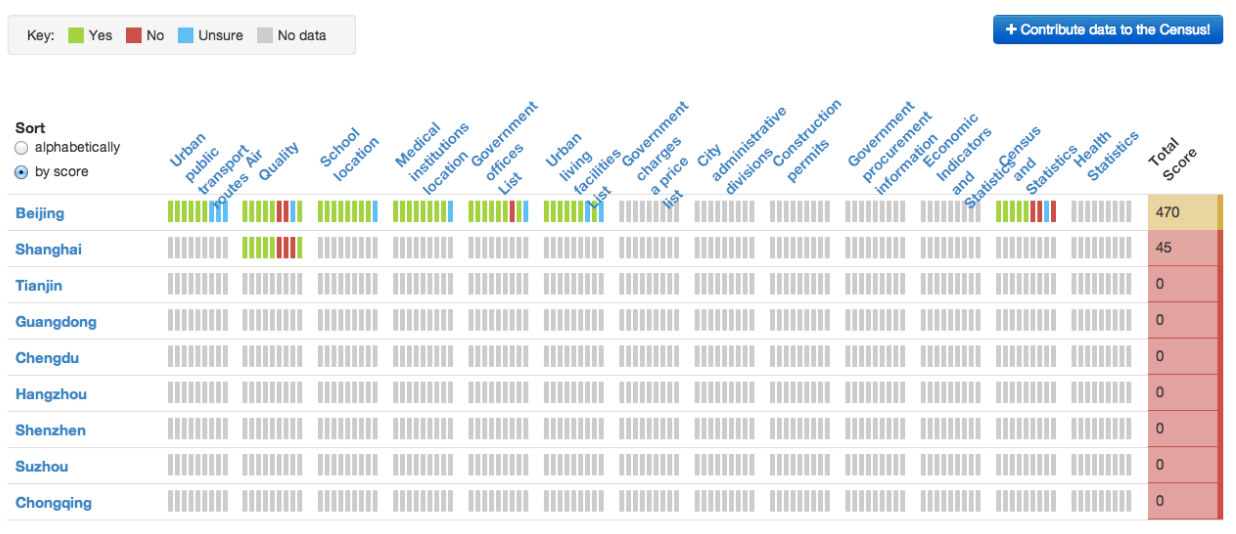


Figure 7.4.2 – The color-coded to show open data's availability in China

Before you can begin to use data, you have to know what's available. So another ongoing project of the Open Data Community is a graded survey of all available data in major Chinese citiesbased on a number of categories such as urban public transportation, air quality, census, health, and economic indicators, among others. As of now, of the two cities that have been graded, about 50 percent of their datasets are open (see Figure 7.4.2), meaning they are machine-readable. Still, the Open Data Community hopes that after making a coordinated effort to clean up the data, it will encourage civil society to use it and that the Open Data Community can serve as an incubator for data apps and projects[[63]](#footnote-63).

In the last decade, a few pioneers in library and information institution have introduced the idea and principles of Semantic Web and Linked Data technology into practice. Some researchers in National Library of China, Institute of Scientific and Technological Information of China and Shanghai Library have done several experimental projects to implement Linked Data. However, there is no practical system that can provide online services by the end of 2014.

Shanghai Library has initiated a project that aimed to establish a genealogy system based on Linked data architecture and the related semantic technologies. There are substantial genealogy materials collected by Shanghai Library. Genealogy contains rich information about families and local history, such as the ancestors, notable people, events, family members migration information,etc. Shanghai Library has the biggest Genealogy collection in China. However, the database is based on MARC format, which is the standard bibliography format for Library resources in the last six decades. It should be transformed into a RDF-enabled format so as to applied to the up-to-date semantic technology.

With the development of Digital Humanity and Semantic Web, the old database based on MARC related technologies are outdated for its obsolete data model. It cannot deeply reveal the properties of the content of genealogy resources, and also cannot encode the relationship among the entities (People, Places, Temporal, and Events etc.). The Linked Data technology is the feasible solution to resolve the problems. As the Lightweight solution of Semantic Web, Linked Data[[64]](#footnote-64) is rooted in the exist basic technologies of Web. In Linked Data architecture, all resources are identified by URI, which act as not only the unique name of the resource on the web but the locator to access the resource as well. The information about the resource is encoded with RDF data model (sometimes can be normalized with OWL encoded ontologies to represent domain knowledge). Its serialization can be understood by machine and indicate the relationship among resources.

Shanghai Library has published a genealogy ontology based on BIBFRAME model and vocabulary. BIBFRAME was initiated by LOC and other Libraries as the next generation bibliography framework intended to replace MARC. It provides the possibility to bring the old-fashioned bibliographic data on the Semantic Web, as a Linked Data service. Currently, Shanghai Library has developed a practical data model with RDF, and start to map the genealogy data into RDF. Now a demo system is under construction. It aims to publish the authority controlled data and bibliography data as Linked Data, and to establish an exhibit layer to access and search the data by multiple facet such as people, place, time, etc. It will provide a SPARQL endpoint to facilitate querying triples. There is also provided a data visualization tool to display the relationship of the data Shanghai Library will keep on applying Semantic Web Technologies to build its digital humanity platform. The Genealogy Linked Data System acts as a demonstrative solution to recognize the traditional library can fit into Semantic Web with its rich, treasured and sophisticated resources.

7.5 Open Data Services with Privacy Data and Anonymization

In this section, we explained how open data are protected. Basic ways of data and information anonymization are shown in Table 7.5.1.

Table 7.5.1 – Ways of Information Anonymization

|  |  |  |
| --- | --- | --- |
| 1 | Replacement | Substitute identifying numbers |
| 22 | Suppression | Omit from the released data |
| 33 | Generalization | Replace birth date with something less specific, like a year of birth |
| 44 | Perturbation | Make random changes to the data |

These technologies are necessary to utilize the open data with privacy because the government data includes the people's privacy. In order to create "Smart Sustainable City," this point should have careful consideration. For a referece, we introduce a coding event of "Open Data for Development Challenge" which was held in Montreal, Canada on January 2014[[65]](#footnote-65).

The Foreign Affairs, Trade and Development Canada (DFATD) was hosting this 36 hours "codathon" bringing together Canadian and international technical experts and policy makers. The event enhanced to generate new tools and ideas with the open data and aid transparency among the users. This kind of events has held in the countries, which have promoted the movement of the open data.

In Japan, several smart cities project including the services using open or private data are conducted. Here two projects are shown as examples. Kawasaki city, Kanagawa prefecture conducts Kawasaki Smart City Project. This project picks up three areas in this city and drives different and localized smart city projects. One is the area of the bay of chemical and oil industries. Hydrogen pipelines and delivery system interconnect different factories and shares the energy of hydrogen and electric power generation between them. One is the area of center station of Kawasaki, and it drives cluster energy management system with aggregation service of energy management data. The aim of this area is to achieve effective use of electricity and reduce the amount of the use. The last area around Musashi‑Kosugi uses data anonymization of private data. Kawasaki city has shared their data via the website "CityData"[[66]](#footnote-66). They shared the data sets of protection against disasters included places of shelters, temporary evacuation facilities), meteorology information, disaster prevention information, such as water stage in a river, rainfall, the place of firehouses, fire prevention water tanks, population statistics, industry statistics. Moreover, information directly connects to usual life, such as receptions, garbage separation and collection, and governmental information. Especially in Musashi-Kosugi area, Home energy management system (HEMS) is introduced into the houses in this area and gathers the data from its smart meter to use as localized services by anonymizing with the data sets. HEMS is a core device to recommend the effective use of electricity to households. However, the households become tired with the information and come to ignore the information. Frequent check of the information of HEMS is indispensable to reduce the use of electricity as self‑action. The specialized HEMS in this area show the information of the data sets and life information as summarized and customized information to change it into fascinating information. Moreover, the data captured by a smart meter is gathered and anonymized to use as the watching services of elder families and persons, door-to-door delivery services to optimize their delivery route by checking the houses' stay, ads and recommendation services, and insurance services for households.

Kurihara City, Miyagi Prefecture conducts Green Society ICT Life Infrastructure Project.This project aims the integration and sophistication of medical, energy, agriculture, climate, local government, disaster prevention infrastructure, and designs information management system as a cloud service are designed as the core of this infrastructure. This system utilizes standards and supports several new smart sustainable city services by providing open interfaces for data analysis and anonymization. Here, several examples of its application are provided.

This system enables to predict the number of patient of thermic fever in the area by learning the trend of a number of the patient and weather information and assembling the climate prediction data. This result enables the city government to predict and prepare required emergency medical service and number of hospital beds in the future. This result also helps the design of future medical services. As the application of this data, elder people are likely to catch a thermic fever easily because they do not feel and think heat. It causes frequent return to the hospitals. Energy management system and smart agricultural system monitors the temperature, humidity and other environmental information by embedded sensor networks. In the future, doctors will use these environmental information to use as a lifestyle recognition and guidance for the elder peoples to　the prevent frequent return to the hospitals.

Another application is energy area. The system gets the information that the power consumption of this area increases by 4% 50 years later.

Hence, applied systems are also proposed. For example, a prediction of data of methodology prediction and energy demand showed 4% higher after 50 years. Additionally, a prediction of increasing hot summer days by eight times showed 4% higher demand converges on summer, it means that electrical power problems would be increased. For the reason of that, demand response services, which have methods of peak cuts and so on, would be enriched. Sensing for cultivated lands would lead the problems of variety selection and relationships between of people's health and farm work in the agriculture field.

"Social capital" is an important factor in local communities in sociology. The concept of the word indicates importance that relationships of mutual trust and social structure in local areas and communities. In order to calculate indexes of social capital, holding questionnaire is one of the ways to make the indexes. Former studies showed a relationship between medical treatment and social capital has significantly interrelated. If one area or community has high index of social capital, aged people are relatively healthy, and if the hospital actions are taken there, the ratio of improvement is higher compared with areas and communities where low ratio of social capital index has. A map layered indexes of social capital in Kurihara-city was created and used for long-distance health guidance using ICT. These results indicate that installing HEMS and community grids to the areas and communities having the indexes of social capital would have the possibility to be a success more than the areas and communities having the lower indexes. Moreover, social capital indexes can utilize not only electricity conservation and a predication of sunstroke prevention using comfort indexes that are different depending on the people.

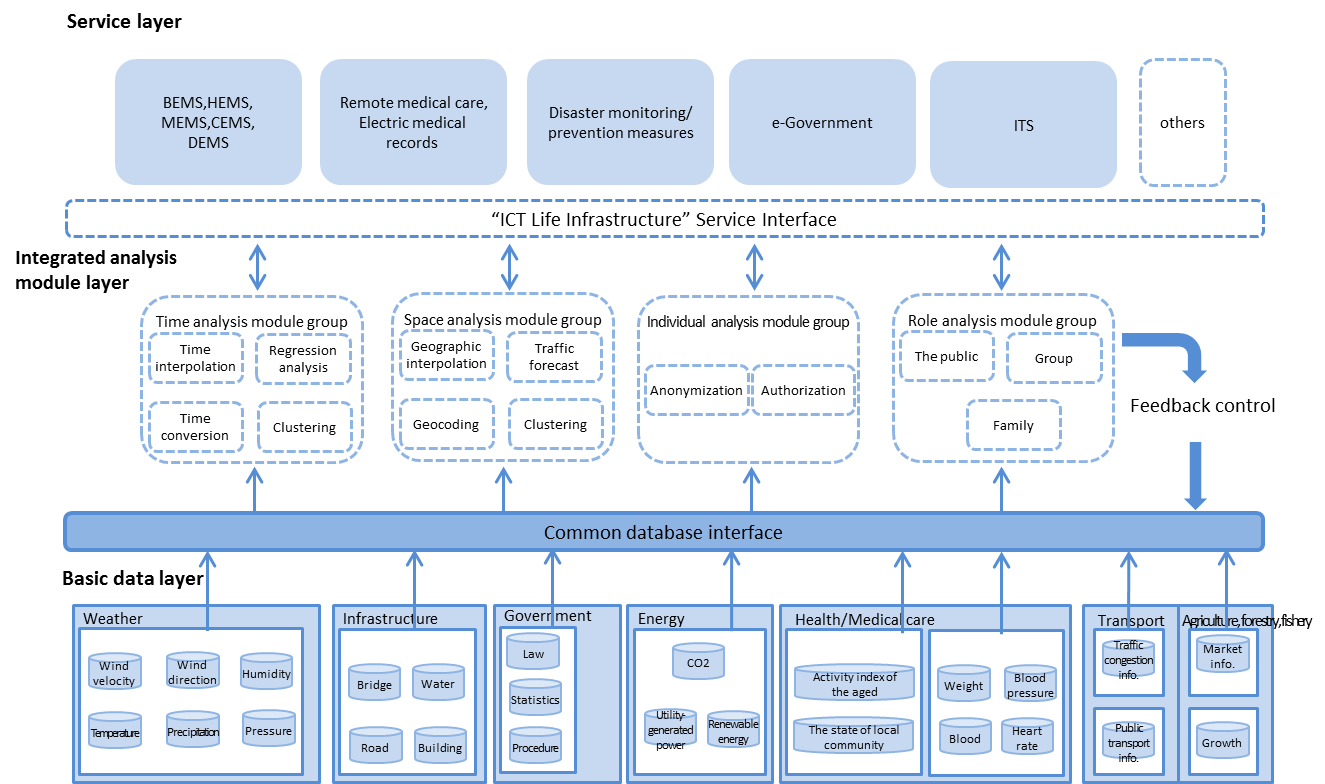


Figure 7.5.1 – Management Infrastructure Model for Data of Smart Sustainable City Originally created by Prof. Uehara, Keio University, Japan

Here, the secondary use services in the world are given. Canada has two laws for privacy preservation: Privacy Act for government institutes and Personal Information Protection and Electronic Document Act (PIPEDA) for private companies. Montreal city in Quebec introduced Opus, an IC card for domestic transportation services in 2008. Then provided a mobile application "STM Merci". This application offers personalized information like shops, concerts, other events. These services use Opus card user information, such as using lines, age, and sex to personalize the information. The user can select the "offers" of the application, and they can get a recommendation and discount tickets of the shop. The major of partner companies as a service provider is bicycle/car sharing services, events, and shops.

These services are well-known in smart phone application by using the personal information. In some cases, private information is used to provide rich services. The utilization rate of the offer is about 4%, and it is twice as large as other general one-to-one marketing. Concordia University in Montreal also leads several approaches using private data with data mining in healthcare field.

Ontario State sells private information of healthcare to private companies and provides it to public institute or organization for free of charge. Personal health Information Protection Act (PHIPA) permits the secondary use of health data by anonymizing it. This data is provided by using k‑anonymity, and Privacy Analytics Risk Assessment Tool (PARAT) was used to confirm the risk of data. One of the schemes is the use in The Children's Hospital of Eastern Ontario (CHEO). The Canadian Institutes of Health Research (CICH), the National Science and Engineering Research Council (NSERC), and the Social Sciences and Humanities Research Council (SSHRC) issues the guidance of research ethics for the use of secondary medical data.

In US, Federal Trade Commissioner is developing laws prevents the illegal use of private data according FTC Act. Health Insurance Portability and Accountability Act (HIPAA) of US takes similar approaches with Canada

In UK, Information Commissioner Office of UK Ministry of Justice supports the penetration, justice, regulation of data privacy. In medical area, UK National the Health and Social Care Information Centre (NHS IC) supports Secondary Uses Service (SUS) and Pseudonymization Implementation Project (PIP). These laws support the secondary use of medical information under the supervision of UK government.

8 Expected (Anonymous) applications of open data in smart sustainable cities

8.1 Application of Open Data on Smart Forecasts

TheClimate Corporation in San Francisco originally called WeatherBill was started to sell powerful software and weather insurance, but it's grown into a company that could helpfarmers around the world adapt to climate change, increase their crop yields.The company's proprietary technology platform combines hyper-local weather monitoring, agronomic data modeling, and high-resolution weather simulations to deliver climate.com, a solution that helps farmers improve their profits by making better informed operating and financing decisions, and Total Weather Insurance, an insurance offering that pays farmers automatically for bad weather that may impact their profits.It is a perfect example of open data application that shows what government data can tell them.The Climate Corporation can display hundreds of different data-driven views of the planet through some projectors, showing changing wind patterns, temperature, ocean currents, or whatever you'd like to look at. In the face of increasingly volatile weather, the companyprovides famers with the industry's most powerful full-stack risk management solutiondepends on the company's unique technologies to help stabilize and improve profits and, ultimately, help feed the world.With a few exceptions, data that fuel the company are freely available to anyone.

The company started by working with data from 200 weather stations across the country. As a prospective policyholder, a business would go to the company's website, pick a nearby weather station, and buy insurance against bad weather that the station would measure. The company would analyze historical weather data for that station, predict the likely weather mathematically, and write an appropriate policy. Farmers in the United States generate about $500 billion a year in revenue, and they make about $100 billion a year in operating profits. So farming is about a 20-percent-margin business on average. The one source of variability for revenue nowadays is the weather, because all the other risks of farming have largely been eliminated through herbicide, fungicide, and insecticide technologies. Weather can be a very big driver for outcomes: farmers can end up losing everything. Slight variations in weather can cause significant losses in profit. Moreover, farmers were significantly underinsured under the federal crop program.

As The Climate Corporation began to turn its attention to farmers, the company found that data from 200 weather stations across the United States imply wasn't precise enough to model the weather at local farms. They expanded to get data from 2,000 stations, but that was still not enough. So they used what is called Common Land Unit data that shows the location, shape, and size of all the farmed fields in the country. Even though this is free, public data, it took many Freedom of Information Act requests and collaboration with Stanford University and other research institutions to get the U.S. Department of Agriculture to release it. Next, The Climate Corporation used government data to assess the weather atall those fields more precisely. Using Doppler radar, it is now possible to measure how much rain falls on a given farmer's field in a day, to an accuracy of almost 1/100th of an inch. The company also got maps of terrain and soil type from the U.S. Geological Survey, built from on‑the-ground soil surveys and satellite images, which give accurate pictures of squares of land 10 meters ona side. Farmers don't necessarily care about how much rain fell. "What they really need to care about is how much water is in their ground," which is determined by both rainfall and the soil. Their goal is to be able to increase a farmer's profitability by 20 or 30 percent – a huge increase in this vulnerable industry.

In the end, it can seem like a conundrum: the U.S. government has invested huge amounts to generate data, but it is taken a private company to put the data to use. In fact, though, this is exactly how many advocates for Open Data think it should be. You have to go outside the government to use the capitalist economic model that says, Take a risk and make more return. However, without government support, none of that innovation could happen. In the government provide infrastructure services. That final point is a critical one. Through an Open Data infrastructure, government can spur innovation by providing the foundation for data-driven businesses. It is been true for GPS and weather data, and it is starting to be true for health data as well.

8.2 Data Anonymization for Smart Sustainable City

Anonymization is one of the methods included in PPDM and PPDP. This method protects sensitive information by masking or generalizing the sensitive data. In addition, it allows the adjustment of the privacy protection level. There are several generalization methods available for anonymization. In the following paragraphs, two relatively basic and frequently referenced generalization methods, - anonymity and -diversity are explained.

anonymity

Anonymity is one of the methods utilized for generalization[[67]](#footnote-67), and it is the base of l-diversity. Further explanation of this method will incorporate the various definitions listed below.

(i) Data table:

A data list similar to a database table is termed a "data table." Its column is termed an "attribute." Address, birth, and gender are examples of attributes. One group of data corresponding to the person or group of people is termed a "data set" and one data set is termed a "tuple".

(ii) Attribute:

An attribute among a group of related attributes that can identify a corresponding person by itself, such as name or unique ID, is termed an "identifier," and others that cannot identify a group on their own, however, it can provide identification when combined with other attributes, such as illness, birth, gender, is termed a "quasi-identifier".

(iii) Sensitive attribute:

A significant attribute for secondary use is termed a "sensitive attribute," which can be selected from attributes that are not identifiers. The method will exclude this attribute from masking or generalization by anonymization. Furthermore, tuple groups that have the same quasi-identifier values are termed "q\*-block".

The definition of k-anonymity is as follows: "In each q\*-block in the data table, at least k tuples are included".

Table 8.2.1 represents an example of a medical records data table. In this table, the sensitive attribute is "Problem" and the quasi-identifiers are "Birth,""Gender," and "ID." The data consists of a t1~t3 q\*-block, a t4, t5 q\*-block, and a t6, t7 q\*-block. It represents k=2. Even if an attacker attempts to ascertain a specific individual's problem and has already obtained the individual's quasi-identifier, the attacker can narrow the results down to only two tuples. Table 8.2.2 indicates that the anonymization results from Table 8.2.1 are k=3. The results displayed in this table demonstrate that anonymization methods provide the required privacy protection level, utilizing masking or generalization.

Table 8.2.1 – Medical record

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Birth | Gender | ID | Problem |
|  | 1970 | male | 121 | cold |
|  | 1970 | male | 121 | obesity |
|  | 1970 | male | 121 | diabetes |
|  | 1980 | female | 121 | diabetes |
|  | 1980 | female | 121 | obesity |
|  | 1981 | male | 125 | diabetes |
|  | 1981 | male | 125 | cold |

Table 8.2.2 – Anonymized medical record

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Birth | Gender | ID | Problem |
|  | 1970 | male | 121 | cold |
|  | 1970 | male | 121 | obesity |
|  | 1970 | male | 121 | diabetes |
|  | 198\* | human | 12\* | diabetes |
|  | 198\* | human | 12\* | obesity |
|  | 198\* | human | 12\* | diabetes |
|  | 198\* | human | 12\* | cold |

As displayed in these tables, the masking or generalization processes prevent an attacker from identifying a specific person. There are several algorithms for calculating masking or generalization. The most popular algorithm is the heuristic searching method, utilizing double-nested loops.

diversity

Diversity is a method designed to protect the privacy of data[[68]](#footnote-68). This method considers the diversity of sensitive attributes, and it is, therefore, different from -anonymity.

The definition of -diversity is as follows: "In all q\*-blocks in a data table, there are at least l different sensitive attributes."

Researchers designed this method to provide protection from the following attacks.

(i) Homogeneity attack:

Table 8.2.3 is an additional example of a medical record data table. In this case, if an attacker has acquired Alice's quasi-identifier, the attacker can read Alice's problem from this table because no diversity exists for the sensitive attributes in the q\*-block.

(ii) Background knowledge attack:

Although theq\*-block in the table has a diversity of sensitive attributes, if the probability of poor circulation is very low for males and an attacker is aware of that, the attacker can read Bob's problem from the table.

*l*-diversity provides more security than -anonymity for preserving privacy. However, the calculation cost of *l*-diversity is higher than anonymity.

Table 8.2.3 – Anonymized medical record

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Birth | Gender | ID | Problem |  |
| 1970 | female | 121 | cold | Alice |
| 1970 | female | 121 | cold |  |
| 1970 | female | 121 | cold |  |
| 198\* | human | 12\* | poor circulation |  |
| 198\* | human | 12\* | poor circulation |  |
| 198\* | human | 12\* | Headache | Bob |
| 198\* | human | 12\* | Headache |  |

The demand for the secondary use of the data such as medical records is increasing, because it may enable the estimation of infection routes. However, medical data frequently includes sensitive and private information. The medical data providers should define the anonymization methods and the related privacy protection levels when publishing the data. In addition, when the data provider permits several methods of anonymization, the consumers of the data must select a method that matches their requirements. Moreover, consumers of the anonymized data should avoid obtaining private data that exceeds their requirements, including situations where the data provider permits the lower protection level and thus provides the private data. Therefore, the anonymization data infrastructure should provide a method to define anonymization methods and protection levels that fulfill the requirements for both data providers and data consumers.

In order to meet these requirements, data publishing with anonymization is required. However, PPDP utilizing anonymization has numerous problems. One of the problems is that no protocols and formats currently exist to enable secure data publishing, as described in the introduction. The other is loss of anonymity by publishing the same data multiple times. Table 8.2.1 is an example of a medical record data table. Table 8.2.2 is an anonymized data table with data from Table 8.2.1, and Table 8.2.3 is another anonymized data table with data from Table 8.2.1. In this case, those who can obtain both the anonymized data ofand k=3can obtain thedata, including situations where the data provider did not permit the publishing of k=1 data.This results in the leak of privacy information. One cause of this problem is that previously published data is not referenced in the anonymization process; as a result the coherence between the and k=3 data was severed. Table 8.2.4 is another example of a *k* = 3 data table. Utilizing Table 8.2.4 instead of Table 8.2.3 avoids the problem described above. Table 8.2.4 was generated by anonymizing Table 8.2.2 instead of anonymizing Table 8.2.1, to maintain coherency in masking and generalization. This anonymizing process can prevent further leaks of privacy information.

To address these problems, a data-publishing infrastructure is shown as a solution. It manages the previously published data for the anonymization without the loss of anonymity and provides safe secondary use and anonymization. For encryption technology, it utilizes Public Key Infrastructure (PKI). Certificate Authority serves a function as an authorized organization for certifying the public key of servers on the Internet. For this discussion, the anonymization technology and this infrastructure can be associated with the encryption technology and PKI, respectively.

Table 8.2.1 – Medical record

|  |  |  |
| --- | --- | --- |
| Birth | Gender | Problem |
| 1970 | male | cold |
| 1970 | male | obesity |
| 1970 | male | diabetes |
| 1981 | male | diabetes |
| 1981 | female | obesity |
| 1982 | female | diabetes |
| 1982 | female | cold |

Table 8.2.2 – Anonymized medical record ()

|  |  |  |
| --- | --- | --- |
| Birth | Gender | Problem |
| 1970 | male | cold |
| 1970 | male | obesity |
| 1970 | male | diabetes |
| 1981 | human | diabetes |
| 1981 | human | obesity |
| 1982 | female | diabetes |
| 1982 | female | cold |

Table 8.2.3 – Anonymized medical record (1)

|  |  |  |
| --- | --- | --- |
| Birth | Gender | Problem |
| 19\* | male | cold |
| 19\* | male | obesity |
| 19\* | male | diabetes |
| 19\* | male | diabetes |
| 198\* | female | obesity |
| 198\* | female | diabetes |
| 198\* | female | cold |

Table 8.2.4 – Anonymized medical record (2)

|  |  |  |
| --- | --- | --- |
| Birth | Gender | Problem |
| 1970 | male | cold |
| 1970 | male | obesity |
| 1970 | male | diabetes |
| 198\* | human | diabetes |
| 198\* | human | obesity |
| 198\* | human | diabetes |
| 198\* | human | cold |

8.3 Application of Anonymization Method for a Smart Metering

Recently, owing to the evolution of cloud services, discussion of secondary uses of data has attracted attention, especially for big data. However, preserving privacy is a significant problem. As a typical application, here we have focused on demand response services in a smart grid as a promising application of smart sustainable city.

The introduction of a smart meter has been considered and achieved around the world. It manages the energy use of a home to achieve a balance between energy saving and a comfortable lifestyle. Smart meters have a communication function to transmit the electric power consumption of a household at regular intervals. The primary use of smart meters is for fare correction. In this case, power consumption data should be collected without loss of data. This means that private information is included in the data.

Demand response (DR) is a typical application of the secondary use of electric power consumption data measured by a smart meter. DR can achieve peak-cut and peak-shift of electricity use by changing the price of electricity or by providing incentives to encourage customers to change their normal consumption pattern when demand for electric power is high. Variable pricing may encourage consumers to reduce electricity consumption and will provide an opportunity to think more about how and when we use electricity. Electric companies or aggregators create DR messages to households according to their current electric power demand. An operation test conducted in the United States demonstrated electric power demand reduction of 10-20%[[69]](#footnote-69). DR can be achieved using electric power consumption data transmitted from residential smart meters.

As an example of DR, electric power consumption will be used widely for secondary use due to its flexibility in applications. The cost of introducing smart meter and devices to control home electric appliances is comparatively higher than the reduction cost of electricity by itself. DR services use a smart meter, which is an electric power meter with a communication function to transmit the electric power consumption of a household to a datacenter. This power consumption data can be used to develop and exploit new services. Recently, secondary use of such data has been considered for such services. DR services are not only for electric companies that collect raw data from smart meters. For other companies, collecting private information with such meters is prohibited. To avoid disclosing private information, it is sufficient for such companies to use generalized or anonymized data if the quality of their services can be guaranteed. However, as described in section 7, electric power consumption data must be treated with significant care. An anonymizing method for electric power consumption data that preserves personal information can be given as an example. This method converts data to distribution data by considering anonymity[[70]](#footnote-70).

Here, as an example, a new anonymizing method for electric power consumption data. This method anonymizes data using the following steps. First, this method generates clusters using -member clustering[[71]](#footnote-71). After -member clustering, the average and width of each cluster can be extracted. By using this parameter, existence probability can be generated from the average and width of each cluster. At this time, the width is modified to control the anonymization level. After creating the existence probability for all clusters, a convolution is given to all clusters. All existence probabilities created from each cluster is summed up and transformed the area generated in this summation process into 1.

In order to achieve DR, one major solution is to change the price of electricity or to provide incentives to encourage customers to change their typical consumption pattern when electricity demand is high. Using this anonymized electric power consumption distribution, a DR service can be provided without obtaining raw data. From historical power trends of anonymized data, it is still possible to predict electric power demand for the next 30-min interval. When the predicted value exceeds a threshold, the system sends a reduction message as a DR message. Figure 8.3.1 shows the image of the anonymized data of electricity consumption distribution of all houses. In this graph, both numbers of houses and power consumption of each house are hidden. In this figure, the DR control group is also given, and different DR signal is issued by these four groups independently. Namely, Group 4 will receive DR message with higher reduction than other groups to observe graduated DR for maintaining fairness. This method also reduces the total calculating cost of DR and number of messages and occupation throughput of network.



Figure 8.3.1 – Threshold value for clustering

8.4 Infrastructure of Secondary Use of Data

This infrastructure can be divided into four organizations as follows (Figure 8.4.1)[[72]](#footnote-72).

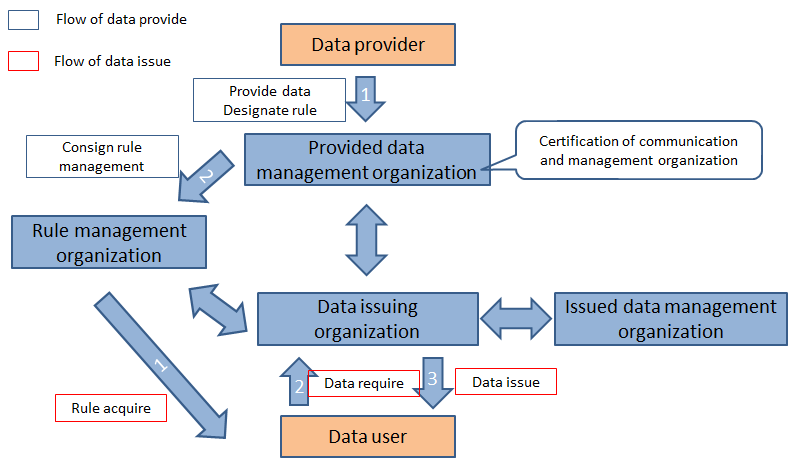


Figure 8.4.1 – Overview of data anonymization infrastructure

(i) Original data storeroom organization (ODS)

This organization manages data provided by the data folder. The data folder is considered the data provider when the data is managed by ODS. When providing data to ODS, the data folder prepares data for publishing and provides an allowance rule by utilizing a specially designed format. This format is termed XML-based Anonymization Sheets (XAS). The details of XAS are described in the following section. Publishing rule descriptions utilize a subset of XAS, termed XML-based Anonymization Rules (XAR). The data folder generates data as D-XAS, and the publishing rules (P-XAR) correspond to the D-XAS. D-XAS should include the link to the P-XAR. ODS should be responsible for maintaining the original data written as D-XAS in a secure manner. This data registration process is based on the PUT method.

(ii) Anonymizing rules storeroom organization (ARS)

This organization manages P-XAR. P-XAR will be openly published for users who need to access anonymized data based on the original data. P-XARs stored in the ARS can exhibit data when it is available for its secondary use. A P-XAR is stored by utilizing a PUT method issued by ODS.

(iii) Data anonymizing and publishing organization (DAP)

This organization anonymizes the original data (D-XAS) based on a publishing rule (P-XAR) and a request rule (R-XAR). A secondary use data consumer generates an R-XAR and provides it to the DAP. An R-XAR contains relevant information for D-XASs such as a URL, the requested anonymization method, its privacy level and anonymization range required to obtain the data for secondary use. The DAP receives the header of the requested D-XAS to access the link of the R-XAR. This header information does not include data. This header information is also described by using an XAR termed H-XAR; the DAP verifies its compliance by checking with the R-XAR and P-XAR requested from the ARS, according to the H-XAR. In this process, a user utilizes a GET method in conjunction with the R-XAR option. If it returns a compliance error, the user receives an appropriate error message. This message utilizes the HTTP error message protocol. If no error occurs, DAP issues a GET message to obtain the D-XAS from the ODS, and issues a subsequent GET message to receive the published XAS (P-XAS) from the PDS. The PDS is described in the following paragraph (iv). The DAP generates P-XASs as anonymized data and the response from the R-XAR of the user. The user receives the anonymized data resulting from the GET method. Finally, the DAP stores the generated P-XAS issues by utilizing the PUSH method. This P‑XAS is utilized to prevent further privacy leaks.

(iv) Published data storeroom organization (PDS)

This organization manages data previously published by the DAP as P-XASs. It may store all anonymized data generated by the DAP. However, to optimize data storage capacity, it is sufficient for the PDS to store only one P-XAS as anonymized data for each D-XAS, according to the one-direction anonymization policy. When generating P-XASs from D‑XASs according to the requested R-XAR, it is sufficient to generate P-XASs according to the R-XAR, and store the P-XAS to the PDS. However, when generating another P-XAS from the same D-XAS according to another R-XAR, the DAP should obtain all P-XASs related to the D-XAS from the PDS. The DAP should consider all of these P-XASs when generating new P-XASs to observe P-XARs. Therefore, we propose one-directional anonymization to avoid this process. The process is as follows:

(i) The DAP generates P-XASs according to P-XARs, instead of R-XARs, and stores it in the PDS. Therefore, the PDS stores the anonymized data, and it is anonymized according to the declared level in P-XAR. This P-XAS is not sent to the users if the requested level in the R-XAR is higher than the level in the P-XAR; this indicates thevalue is larger than that of the P-XAR in -anonymity.

(ii) DAP generates P-XASs according to the R-XARs. In this generation, the DAP only uses the first P-XAS generated from the P-XAR. DAP generalizes new P-XASs by adding "wild cards" as masking from the initial P-XAS. The DAP does not remove any of the "wild cards" provided as masking in the first P-XAS. Therefore, a one-directional anonymizing process should be considered.

(iii) The DAP can generate any type of P-XAS that satisfies both the R-XAR and the P-XAR by following the process described in (i) and (ii). In a scenario where -anonymity and ‑diversity are mixed, it is sufficient to generate a P-XAS that has a lower anonymization level than-anonymity and -diversity. For example, assume that 3‑anonymity and 3-diversity are permitted in P-XARs, and 4-diversity is requested by R‑XAR. In this case, DAP generates the initial P-XAR by utilizing 3-anonymity. The DAP can generate any type of P-XAR by utilizing the initial P-XAR, according to the one-directional anonymizing process.

In order to enable the data transfer between these organizations, data providers, and data consumers will utilize SSL and PKI if they transfer the data over the Internet. In the following discussions, four organizations are exhibited in order to clarify each role. It is possible to merge some of them into a single organization. Figure 8.4.2 represents an organizational structure and data connections between the organizations.

XML-based Anonymization Sheets (XAS) is a format to define the rules and data descriptions. To distinguish the rules from the data, XML-based Anonymization Rules (XAR) are also shown as a subset of XAS. XAS and XAR differ because XAR does not contain data as contents. All transactions in this infrastructure utilize the XAS and its subset, XAR. XAS is designed according to Extensible Markup Language (XML). Figure 8.4.2 lists an example of D-XAS. It includes the information to enable anonymization, including combinations of the sensitive attribute names and quasi-identifiers, permitted anonymization methods and levels, and data attributes such as created date, updated date and history, ownership, copyrights, comments, and others. Figure 8.4.3 lists an example of a P-XAR. It does not contain raw data; it only declares the required anonymization methods and levels. To enable masking or generalization processes, it can define the delimiter for distinguishing data sections. In this example, "BirthDay" is split utilizing the '-' character. During the anonymizing process, the character is used to define the generalization boundary. If the data employs a general and standardized format, for example, BirthDay should be separated by '-' it can generalize the data entry by referring to the default rule. As an additional feature, the data provider may publish data samples without data publishing limits to publicize the data's availability. This open information is termed "open attribute." This open attribute can be declared in the data entry.

1 <?xml version="1.0" encoding="utf-8"?>

2 <?xml-anonymize type="text/xas" href="p-xar.xas"?>

3 <list>

4 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:v="http://www.w3.org/2006/vcard/ns#">

5 <v:Kind rdf:about = "http://foo.com/me/hogehoge">

6 <v:fn>Hoge Foo</v:fn>

7 <v:bday>1980-01-01</v:bday>

8 <v:hasTelephone>

9 <rdf:Description>

10 <rdf:value>+81-45-566-1454</rdf:value>

11 <rdf:type rdf:resource="http://www.w3.org/2006/vcard/ns#Work"/>

12 <rdf:type rdf:resource="http://www.w3.org/2006/vcard/ns#Voice"/>

13 </rdf:Description>

14 </v:hasTelephone>

15 <v:hasAddress>

16 <rdf:Description>

17 <v:street-address>123-45 Hoge Village</v:street-address>

18 <v:locality>FooCity</v:locality>

19 <v:postal-code>5555</v:postal-code>

20 <v:country-name>Japan</v:country-name>

21 </rdf:Description>

22 </v:hasAddress>

23 </v:Kind>

24 </rdf:RDF>

25 <OfficeScale>100ha</OfficeScale>

26 <PowerConsumption>10kWh</PowerConsumption>

27

28 <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#" xmlns:v="http://www.w3.org/2006/vcard/ns#">

29 <v:Kind rdf:about = "http://foo.com/me/db">

Fig. 8.4.2 D-XAS Example (Extract)

1 <?xml version="1.0" encoding="utf-8"?>

2 <anonymize>

3 <head>

4 <publishacceptance sensitive="divisional" quasi="divisional" />

5 <firstdatasetposition>

6 <list>

7 <rdf:RDF />

8 </list>

9 </firstdatasetposition>

10 <sensitive type="k(>=3), l(>=2)">

11 <rdf:RDF>

12 <v:Kind>

13 <v:hasTelephone>

14 <rdf:Description>

15 <rdf:type number="2" />

16 </rdf:Description>

17 </v:hasTelephone>

18 </v:Kind>

19 </rdf:RDF>

20 <PowerConsumption />

21 </sensitive>

22 <sensitive type="k(>=3), l(>=2)">

23 <OfficeScale />

24 </sensitive>

25 <group name="addr" type="quasi" level="k(>=3), l(>=3)"/>

26 </head>

27 <rdf:RDF>

28 <v:Kind>

29 <v:fn note="Full Name" />

30 <v:bday note="BirthDay" type="quasi" level="k(>=2)" sprit="-" />

31 <v:hasTelephone>

32 <rdf:Description>

33 <rdf:value note="TelephoneNumber" type="open" sprit="\s" />

34 <rdf:type note="Number Type" attribute="rdf:resource" number="2" />

35 </rdf:Description>

36 </v:hasTelephone>

37 <v:hasAddress>

38 <rdf:Description note="Addresses">

39 <v:street-address group="addr" priority="4" />

40 <v:locality group="addr" priority="3" />

41 <v:postal-code group="addr" priority="2" />

42 <v:country-name group="addr" priority="1" />

43 </rdf:Description>

44 </v:hasAddress>

45 </v:Kind>

46 </rdf:RDF>

47 <OfficeScale note="OfficeScale" />

48 <PowerConsumption type="open" note="PowerConsumption" />

49 </anonymize>

Fig. 8.4.3 P-XAR Example

The secondary data user can request access to the open attributes by utilizing R-XAR. Figure 8.4.4 lists an example of an R-XAR. If the secondary data consumer requests attribute identified as quasi‑identifiers, DAP publishes anonymized data that contains attributes calculated as quasi‑identifiers. The user also declares the required anonymization method, privacy protection level, sensitive attributes combinations, open attributes, and quasi-identifiers utilizing the R-XAR.

1 <?xml version="1.0" encoding="utf-8"?>

2 <anonymize type="k(3)">

3 <head>

4 <sensitive>

5 <rdf:RDF>

6 <v:Kind>

7 <v:hasTelephone>

8 <rdf:Description>

9 <rdf:type number="2" />

10 </rdf:Description>

11 </v:hasTelephone>

12 </v:Kind>

13 </rdf:RDF>

14 <PowerConsumption />

15 </sensitive>

16 <group name="addr" type="quasi" />

17 </head>

18 <rdf:RDF>

19 <v:Kind>

20 <v:bday />

21 <v:hasTelephone>

22 <rdf:Description>

23 <rdf:value note="TelephoneNumber" type="quasi" />

24 </rdf:Description>

25 </v:hasTelephone>

26 </v:Kind>

27 </rdf:RDF>

28 <PowerConsumption note="PowerConsumption" />

29 </anonymize>

Fig. 8.4.4 R-XAR Example

The formats of XAS and its subset XAR utilize the Cascading Style Sheets (CSS) format and the Semantic Web standard. The XAS can be processed utilizing an XML schema, RDL schema, OWL method, and other related tools.

9 Milestone

Here, we give the milestone of open data innovation in smart sustainable city mainly from the viewpoint of data management, infrastructure. In this document, each chapter has their dedicated focuses. However, several common problems and solutions potentially lie over different discussions. Considering the overlap of each chapter, the roadmap model described here is categorized into the following three groups; 1st Open data issue, 2nd Application Services in Smart Sustainable City, 3rd Security and Anonymization. All items covered in this roadmap are given in this document and are arranged into the three groups above.

Timeline:

(today-2020, 2020-2040, 2040 and beyond)

+-------------------------------------------------------> (normal scale)

+------------------------------#------------------------> (with the point of technological accomplishment)

Open data issue

- Open data in smart sustainable city

Open data is the key of services in smart sustainable city.

(today-2020, 2020-2040, 2040 and beyond)

+---------------------------------------------->

- The use of smart energy data

Energy data as open data changes grid system to smart grid, which becomes a component of smart sustainable city.

(today-2020, 2020-2040, 2040 and beyond)

+----------------------------->

- Smart transportation data

ITS and automatic driving is one of the major component of smart sustainable city.

(today-2020, 2020-2040, 2040 and beyond)

+---------------------------#-------------------------->

- Location data

Location services is useful for every smart sustainable city services.

Application Services in Smart Sustainable City

- Recommendation service

Recommendation services, such as concierge service, will be penetrated using data in smart sustainable city.

(today-2020, 2020-2040, 2040 and beyond)

+-------------#------------------------------>

Security and Anonymization

- Security of smart sustainable city data

(today-2020, 2020-2040, 2040 and beyond)

+----------------#------------------------------------->

- Anonymization of smart sustainable city data

(today-2020, 2020-2040, 2040 and beyond)

+-----------------------#------------------------------>

Annex A  
  
Application of anonymization for disaster recovery

One important application of governmental open data is disaster minimization and recovery.Disaster is a social phenomenon, such as threaten human society or economic activity brought by physical hazard and a vulnerability in the society. Disaster management is the way to eliminate the vulnerability and is depicted as Figure 10.1.

Preparedness

Recovery &

Reconstruction

Response

Damage

Assessment

Mitigation

Prediction &

Early Warning

Disaster

Emergency Management

Risk Management

Figure 10.1 – Phase of disaster recovery

1) Pre-disaster: Risk Management

**Prevention of damage**

This action includes hardware approach, such as building a bank and aseismic reinforcement of building. In this case, integration with data generated by the use of structural health monitoring is indispensable. The data enables to provide applications, such as disaster prediction and early warning of disaster. The data should be published as open data to encourage these useful applications. However, this data is also a critical data from a viewpoint of privacy and security. In some cases, these data should be anonymized.

2) Post-disaster: emergency management

Damage evaluation is achieved by using data of global earthquake monitoring and earth scanning from satellite or airplane. The data is used for firefighting, rescue effort and medical activity, recovery of city functions and improvement of them. Damage evaluation also reveals the weak points of the structures. From a viewpoint of privacy and security, the data of evaluation should be anonymized.

3) Data Acquisition

The following systems generate data and is useful for governmental open data:

* Data generated by sensors in a building with anti-shaking system like active/passive dampers
* Transportation monitoring/management system for congestion or traffic accident regulation
* Electronic health records in hospitals
* Agricultures, especially state-of-art automated environment control system in a greenhouse
* e-government including residentiary and geographical data, administrative services and social services
* Smart infrastructures, such as smart water, smart grid, smart community

These data sometimes includes privacy information of personals. From a viewpoint of privacy and security. In some cases, these data should be anonymized.

Annex B  
  
Abbreviations

This Technical Report uses the following abbreviations:

ANSI American National Standards Institute

ARS Anonymizing Rules Storeroom organization

CA Certificate Authority

CICH Canadian Institutes of Health Research

CDWA Categories for the Description of Works of Art

CHEO Children’s Hospital of Eastern Ontario

CKAN Comprehensive Knowledge Archive Network

C/S Client/Server

CSS Cascading Style Sheets

DAP Data Anonymizing and Publishing organization

DC Dublin Core

DR Demand Response

EAD Encoded Archival Description

FGDC Federal Geospatial Data Committee

FTC Federal Trade Commissioner of United States of America

GDP Gross Domestic Product

GILS Government Information Locator Service

HIPAA Health Insurance Portability and Accountability Act

HTTP Hypertext Transfer Protocol

IoT Internet of Things

ISO International Organization for Standardization

ITU International Telecommunication Union

NASA National Aeronautics and Space Administration

NHS IC National Health and Social Care Information Centre of United Kingdom

NILM Non-intrusive Load Monitoring

NSERC National Science and Engineering Research Council

ODMB Open Data and Manpower Bureau

OECD Economic Co-operation and Development

OED Open Enterprise Data

ODS Original Data Storeroom organization

OGD Open Government Data

OID Open Industrial Data

OMB Office of Management and Budget

OSD Open Scientific Data

OWL Web Ontology Language

PARAT Privacy Analytics Risk Assessment Tool

PHIPA Personal Health Information Protection Act

PIP Pseudonymization Implementation Project

PIPEDA Personal Information Protection and Electronic Document Act

PKI Public Key Infrastructure

PPDM Privacy-Preserving Data Mining

PPDP Privacy-Preserving Data Publishing

QoL Quality of Life

RDB Relational Databases

RDF Resource Description Framework

RDL Report Definition Language

REST Representational State Transfer

RPC Remote Procedure Call Protocol

SOAP Simple Object Access Protocol

SSC Smart Sustainable Cities

SSHRC Social Sciences and Humanities Research Council

SUS Secondary Uses Service

TEI Text Encoding Initiative

URI Uniform Resource Identifier

VRA Visual Resources Association

XAS XML-based Anonymization Sheets

XAR XML-based Anonymization Rules

XML Extensible Markup Language

Appendix I  
  
Bibliography

[b-FG-SSC overview] FG-SSC deliverable (2014), Technical Report on an overview of smart sustainable cities and the role of information and communication technologies.

[b-FG-SSC KPIsICT] FG-SSC deliverable, Key performance indicators related to the use of information and communication technology in smart sustainable cities.

[b-FG-SSC infrastructure] FG-SSC deliverable, Smart Sustainable Cities Infrastructure.

[b-FG-SSC security] FG-SSC deliverable, Technical Report on cyber-security, data protection and cyber-resilience in smart sustainable Cities.

[b-FG-SSC management] FG-SSC deliverable, Technical Report on integrated management for smart sustainable cities.

[b-FG-SSC KPIs metrics] FG-SSC deliverable, Technical Report on metrics and evaluation of keyperformanceindicators for smart sustainable cities.

Accenture | Strategy. Understanding Data Visualization. 2013

A. Machanavajjhala, D.Kifer, J. Gehrke, et al., *L-diversity: Privacy beyond k-anonymity*, ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 1, No. 1, 2007.

B. P. Allen and J. T. Tennis, *Building metadata-based navigation using semantic Web standards: the Dublin Core 2003*, Conference Proceedings 2004 Proceedings of the 2004 Joint ACM/IEEE Conference on. 2004.

B. C. Chen, D. Kifer, K. LeFevre, A. Machanavajjhala, *Privacy-Preserving Data Publishing*, Foundations and Trends in Databases, Vol. 2, No. 1-2, pp. 1-167, 2009.

Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council (2008). Earth Observations from Space: The First 50 Years of Scientific Achievements.

The National Academies Press. p. 6. ISBN 0-309-11095-5. Retrieved 2010-11-24.

C. M. F. Benjamin, K. Wang, R. Chen, et al., *Privacy-preserving data publishing: A survey of recent developments*, ACM Computing Surveys (CSUR), Vol. 42, No. 4, 2010.

Data.gov.uk, URL: <http://data.gov.uk/about> (2014).

Declaration to be the World’s Most Advanced IT Nation, URL:

[http://japan.kantei.go.jp/policy/it/20140624\_decration.pdf (2014)](http://japan.kantei.go.jp/policy/it/20140624_decration.pdf%20(2014)).

Digital Government: Building a 21st Century Platform to Better Serve the American People, URL <http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html> (2014).

D. Li, L. Timothy, S. E. John, et al., *TWC LOGD: A portal for linked open government data ecosystems*, Web Semantics: Science, Services and Agents on the World Wide Web 9 (2011) 325–333.

D. S. Raghuwanshi, M. R. Rajagopalan, *MS2: Practical data privacy and security framework for data at rest in cloud*. Computer Applications and Information Systems (WCCAIS), 2014 World Congress on. 2014, Page(s): 1- 8.

GoodGuide,[www.goodguide.com](http://www.goodguide.com) (2014).

Health Data Initiative, HelthData.gov, URL: <http://healthdata.gov/> (2014).

H. M. Yu, Designing Software to Shape Open Government Policy, Princeton University, 2012.

H. Nishi, K. Inoue, et al., *IEEE SMART GIRD VISION FOR VEHICULAR TECHNOLOGY: 2030 AND BEYOND*, IEEE Standards Association, pp. 41-42, Jan., 2014.

<https://technet.microsoft.com/en-us/library/dd277323.aspx> .

<http://datasecuritymanagement.com/>.

<http://en.factolex.com/metadata> .

<http://en.wikipedia.org/wiki/Data_integrity>.

<http://en.wikipedia.org/wiki/Data_visualization>.

<http://en.wikipedia.org/wiki/Linked_data>.

<http://en.wikipedia.org/wiki/Metadata>.

<http://en.wikipedia.org/wiki/Open_data>.

<http://en.wikipedia.org/wiki/Representational_state_transfer>.

<http://en.wikipedia.org/wiki/Social_network>.

<http://en.wikipedia.org/wiki/SPARQL>.

<http://opendatachina.com/the-business-case-for-open-data/(2014)>.

<https://technet.microsoft.com/en-us/library/dd277323.aspx>.

http://techpresident.com/news/wegov/24940/China-Open-Data-Movement-Starting-Take-Off.

<http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Tech-Labs-Data-Visualization-Full-Paper.pdf>.

<http://www.emprata.com/what-we-do/data-visualization/>.

<http://www.nyc.gov/html/dcp/html/about/pr031411.shtml>.

<http://www.nyc.gov/html/dcp/html/cwp/index.shtml> .

<http://www.spamlaws.com/data-security.html> .

IT dash board, National Strategy Office of Information and Communications Technology, URL: http://www.itdashboard.go.jp/ (2014).

[J. Gurin](http://www.amazon.cn/s/ref=dp_byline_sr_book_1?ie=UTF8&field-author=Joel+Gurin&search-alias=books), *Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation*, McGraw-Hill Professional.

J. Mize, R.T. Habermann, *Automating metadata for dynamic dataset,* OCEANS 2010. 2010, Page(s): 1-6.

J. W. Byun, A. kamura, E. Bertino, et al., *Efficient k-anonymization using clustering techniques*, Advances in Databases: Concepts, Systems and Applications, pp. 188–200, 2007.

Kawasaki City Open Data, Kawasaki City Office, URL: http://citydata.jp/%E7%A5%9E%E5%A5%88%E5%B7%9D%E7%9C%8C/%E5%B7%9D%E5%B4%8E%E5%B8%82 (2014).

K. Okada and H. Nishi, *Big Data Anonymization Method for Demand Response Services*, ICOMP'14, The 2014 International Conference on Internet Computing and Big Data, 2014.

L. M. Prieto, A. C. Rodríguez, J. Pimiento, *Implementation Framework for Open Data in Colombia*, ICEGOV '12, October 22-25 2012.

L. N. Mutuku and J. Colaco, *Increasing Kenyan Open Data Consumption: A Design Thinking Approach*, ICEGOV '12, October 22 - 25 2012, Albany, NY, USA.

L. Seeney, *k-anonymity: A model for protecting privacy*, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

L. Yang, X. Chen, J. Zhang, et al., *Optimal Privacy-Preserving Energy Management for Smart Meters*, IEEE INFOCOM 2014, Page(s): 513-521.

L. Zhu, R.D. Kent, A. Aggarwal, et al., *Construction of a Webportal and User Management Framework for Grid*, High Performance Computing Systems and Applications, HPCS 2007, 21st International Symposium on. 2007, Page(s): 14.

M. Franck, M. Johan and F. Catherin, *A survey of RDB to RDF translation approaches and tools*.

NISO Press, National Information Standards Organization, Understanding Metadata.

N. Shadbolt, K. O’Hara, T. Berners-Lee, et al., *Linked Open Government Data: Lessons from Data.gov.uk*, IEEE INTELLIGENT SYSTEMS, May-June 2012, 16-24 (Volume: 27, Issue: 3).

OECD Declaration on Open Access to publicly funded data.

OECD Principles and Guidelines for Access to Research Data from Public Funding.

[Open Data 500](http://www.baidu.com/link?url=E0LtWD1fsrb_HVUbVeFEbZWo8xy58WdNgEnYXEo0rMl6UxfLhBevxqAa7tCGVEP2), [www.opendata500.com](http://www.opendata500.com) (2014).

Open Data for Development Challenge, Foreign Affairs, Trade and Development Canada,

URL: http://www.acdi-cida.gc.ca/acdi-cida/acdi-cida.nsf/eng/DEN-1223131242-PCZ (2014) .

Open Definition, URL: http://opendefinition.org/od/#sthash.PmqzjbQW.dpuf (2014).

Opower Official Website, URL: http://opower.com/ (2014).

Q. Xiao, W. Yan, H. Zhang, *Application of Visualization Technology in Spatial Data Mining*. *Computing*, Control and Industrial Engineering (CCIE), 2010 International Conference on. 2010, Page(s): 153-157.

R. Agrawal, R. Srikant, *Privacy-preserving data mining*, SIG-MOD, Vol. 29, pp. 439-450, 2000.

R. Irfan, G. Bickler, S. U. Khan, et al., *Survey on social networking services*, IET Networks, 2013, Page(s): 224-234.

S. McLaughlin, P. McDaniel, W. Aiello, *Protecting consumer privacy from electric load monitoring*, in Proceedings of the 18th ACM conference on Computer and communications security CCS ’11, 2011.

S. Turchi, F. Paganelli, L. Bianchi, et al., *A lightweight linked data implementation for modeling the Web of Things*, Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on 2014, Page(s): 123-128.

T. B. Lee, *Linked Data*, 2011-05-15, <http://www.w3.org/DesignIssues/LinkedData.html>.

T. Hattori, N. Toda, *Demand response programs for residential customers in the United States—Evaluation of the pilot programs and the issues in practice*, MARCH 2011.

T. Heath and C. Bizer, *Linked Data: Evolving the Web into a Global Data Space Heath*, 2011.

World Data Center System (2009-09-18). "About the World Data Center System". NOAA, National Geophysical Data Center. Retrieved 2010-11-24.

W. David, Z. Marsha, R. Luke, et al., *Linked Data: Structured Data on the Web*.

W. Parks, *Open Government Principle: Applying the Right to Know Under the Constitution*. The George Washington Law Review. 26, 1 (1957), 1-22.

Y. Lindell, B. Pinkas, *Privacy Preserving Data Mining*, Journal of Cryptology, Vol. 15, pp. 177-206, 2002.

Y. Nakamura, K. Matsui and H. Nishi, *Anonymization Infrastructure for Secondary Use of Data*, ICOMP'14 - The 2014 International Conference on Internet Computing and Big Data, 2014.

Y. Yu, J. Ni, M. H. Au, et al., *On the Security of a Public Auditing Mechanism for Shared Cloud Data Service*, Services Computing, IEEE Transactions on. 2014, Page(s): 1.

1. Yu, H.M.-T. 2012. Designing Software to Shape Open Government Policy. Princeton University. [↑](#footnote-ref-1)
2. Parks, W. 1957. Open Government Principle: Applying the Right to Know Under the Constitution. The George Washington Law Review. 26, 1 (1957), 1–22. [↑](#footnote-ref-2)
3. Nigel Shadbolt, Kieron O’Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall and m.c. schraefel; *Linked Open Government Data: Lessons from Data.gov.uk*; IEEE INTELLIGENT SYSTEMS, May-June 2012 16-24 (Volume:27, Issue: 3). [↑](#footnote-ref-3)
4. Li Ding , Timothy Lebo, John S. Erickson, Dominic DiFranzo, Gregory Todd Williams, Xian Li,James Michaelis, Alvaro Graves, Jin Guang Zheng, Zhenning Shangguan, Johanna Flores, Deborah L. McGuinness, James A. Hendler, *TWC LOGD: A portal for linked open government data ecosystems*; Web Semantics: Science, Services and Agents on the World Wide Web 9 (2011) 325–333. [↑](#footnote-ref-4)
5. Leonida N. Mutuku and Jessica Colaco; *Increasing Kenyan Open Data Consumption: A Design Thinking Approach*; ICEGOV '12, October 22 - 25 2012, Albany, NY, USA. [↑](#footnote-ref-5)
6. Lydia Marleny Prieto, Ana Carolina Rodríguez, Johanna Pimiento; *Implementation Framework for Open Data in Colombia*; ICEGOV '12, October 22 - 25 2012. [↑](#footnote-ref-6)
7. Committee on Scientific Accomplishments of Earth Observations from Space, National Research Council (2008). Earth Observations from Space: The First 50 Years of Scientific Achievements. The National Academies Press. p. 6. ISBN 0-309-11095-5. Retrieved 2010-11-24. [↑](#footnote-ref-7)
8. World Data Center System (2009-09-18). *About the World Data Center System*. NOAA, National Geophysical Data Center. Retrieved 2010-11-24. [↑](#footnote-ref-8)
9. OECD Declaration on Open Access to publicly funded data. [↑](#footnote-ref-9)
10. OECD Principles and Guidelines for Access to Research Data from Public Funding. [↑](#footnote-ref-10)
11. Please see: <http://www.nyc.gov/html/dcp/html/about/pr031411.shtml>. [↑](#footnote-ref-11)
12. Ibid. OECD Principles and Guidelines for Access to Research Data from Public Funding. [↑](#footnote-ref-12)
13. Please see: <http://en.wikipedia.org/wiki/Open_data>. [↑](#footnote-ref-13)
14. Ibid. OECD Principles and Guidelines for Access to Research Data from Public Funding. [↑](#footnote-ref-14)
15. Ibid. OECD Principles and Guidelines for Access to Research Data from Public Funding [↑](#footnote-ref-15)
16. Joel Gurin, Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation, McGraw-Hill Professional. [↑](#footnote-ref-16)
17. Yuichi Nakamura, Kanae Matsui and Hiroaki Nishi, *Anonymization Infrastructure for Secondary Use of Data*, ICOMP'14 - The 2014 International Conference on Internet Computing and Big Data, 2014. [↑](#footnote-ref-17)
18. Please see: <http://opendatachina.com/the-business-case-for-open-data/(2014)>. [↑](#footnote-ref-18)
19. Joel Gurin, Open Data Now: The Secret to Hot Startups, Smart Investing, Savvy Marketing, and Fast Innovation, McGraw-Hill Professional. [↑](#footnote-ref-19)
20. Good guide, [www.goodguide.com](http://www.goodguide.com) (2014). [↑](#footnote-ref-20)
21. Open Data 500, www.opendata500.com (2014). [↑](#footnote-ref-21)
22. Hiroaki Nishi, Koichi Inoue, et al.; *IEEE SMART GIRD VISION FOR VEHICULAR TECHNOLOGY: 2030 AND BEYOND*; IEEE Standards Association, pp. 41-42, Jan., 2014. [↑](#footnote-ref-22)
23. Rakesh Agrawal; Ramakrishnan Srikant; *Privacy-preserving data mining*; SIG-MOD, Vol. 29, pp. 439-450, 2000. [↑](#footnote-ref-23)
24. Yehuda Lindell; Benny Pinkas; “*Privacy Preserving Data Mining*; Journal of Cryptology, Vol. 15, pp. 177-206, 2002. [↑](#footnote-ref-24)
25. Bee-Chung Chen; Daniel Kifer; Kristen LeFevre; AshwinMachanavajjhala; *Privacy-Preserving Data Publishing*; Foundations and Trends in Databases, Vol. 2, No. 1-2, pp. 1-167, 2009. [↑](#footnote-ref-25)
26. Benjamin C. M. Fung; Ke Wang; Rui Chen; Philip S. Yu; *Privacy-preserving data publishing: A survey of recent developments*; ACM Computing Surveys (CSUR), Vol. 42, No. 4, 2010. [↑](#footnote-ref-26)
27. Please see: <http://en.factolex.com/metadata>. [↑](#footnote-ref-27)
28. Mize, J.; Habermann, *R.T. Automating metadata for dynamic datasets*. OCEANS 2010. 2010, Page(s): 1-6. [↑](#footnote-ref-28)
29. Please see: <http://en.wikipedia.org/wiki/Metadata>. [↑](#footnote-ref-29)
30. NISO Press, National Information Standards Organization. Understanding Metadata. [↑](#footnote-ref-30)
31. T.; Bizer, C.; Linked Data: Evolving the Web into a Global Data Space Heath. 2011. [↑](#footnote-ref-31)
32. Allen, B.P.; Tennis, J.T.; *Building metadata-based navigation using semantic Web standards: the Dublin Core 2003*; Conference Proceedings, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on. 200 [↑](#footnote-ref-32)
33. Please see: <http://en.wikipedia.org/wiki/Linked_data>. [↑](#footnote-ref-33)
34. Turchi, S.; Paganelli, F.; Bianchi, L.; Giuli, *D.A lightweight linked data implementation for modeling the Web of Things*. Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on 2014, Page(s): 123-128. [↑](#footnote-ref-34)
35. David W., Marsha Z. and Luke R. with Michael H.; *Linked Data: Structured Data on the Web*. [↑](#footnote-ref-35)
36. Please see: <http://en.wikipedia.org/wiki/Linked_data>. [↑](#footnote-ref-36)
37. Franck M., Johan M. and Catherine F.; A survey of RDB to RDF translation approaches and tools. [↑](#footnote-ref-37)
38. Please see: <http://en.wikipedia.org/wiki/SPARQL>. [↑](#footnote-ref-38)
39. Please see: <http://en.wikipedia.org/wiki/Representational_state_transfer>. [↑](#footnote-ref-39)
40. Please see: <http://en.wikipedia.org/wiki/Data_visualization>. [↑](#footnote-ref-40)
41. Xiao Q., Yan W., Zhang H.; *Application of Visualization Technology in Spatial Data Mining*; Computing, Control and Industrial Engineering (CCIE), 2010 International Conference on. 2010, Page(s): 153-157. [↑](#footnote-ref-41)
42. Accenture | Strategy. Understanding Data Visualization. 2013 [↑](#footnote-ref-42)
43. Please see: <http://www.accenture.com/SiteCollectionDocuments/PDF/Accenture-Tech-Labs-Data-Visualization-Full-Paper.pdf>. [↑](#footnote-ref-43)
44. Please see: <http://www.emprata.com/what-we-do/data-visualization/>. [↑](#footnote-ref-44)
45. Please see: <http://en.wikipedia.org/wiki/Social_network>. [↑](#footnote-ref-45)
46. Irfan, R., Bickler, G., Khan, S.U., Kolodziej, J., Hongxiang Li, Dan Chen, Lizhe Wang, Hayat, K., Madani, S.A., Nazir, B., Khan, I.A., Ranjan, R; *Survey on social networking services.* Networks, IET. 2013, Page(s): 224-234. [↑](#footnote-ref-46)
47. McLaughlin, S., McDaniel, P., & Aiello, W, *Protecting consumer privacy from electric load monitoring*; In Proceedings of the 18th ACM conference on Computer and communications security CCS ’11 ,2011. [↑](#footnote-ref-47)
48. Yang, L., Chen, X., Zhang, J., & Poor, H.; *Optimal Privacy-Preserving Energy Management for Smart Meters*, IEEE INFOCOM, 2014,Page(s): 513-521. [↑](#footnote-ref-48)
49. Lichun Zhu, Kent, R.D., Aggarwal, A., Viranthi, P., Rahman, Q., Elamsy, T., Ejelike, O; *Construction of a Webportal and User Management Framework for Grid*. High Performance Computing Systems and Applications, 2007. HPCS 2007. 21st International Symposium on. 2007, Page(s): 14. [↑](#footnote-ref-49)
50. Please see: <http://en.wikipedia.org/wiki/Data_integrity>. [↑](#footnote-ref-50)
51. Please see: <http://datasecuritymanagement.com/>. [↑](#footnote-ref-51)
52. Please see: <http://www.spamlaws.com/data-security.html>. [↑](#footnote-ref-52)
53. Please see:<https://technet.microsoft.com/en-us/library/dd277323.aspx>. [↑](#footnote-ref-53)
54. Raghuwanshi, D.S., Rajagopalan, M.R.MS2; *Practical data privacy and security framework for data at rest in cloud*. Computer Applications and Information Systems (WCCAIS), 2014 World Congress on. 2014, Page(s): 1- 8. [↑](#footnote-ref-54)
55. Yu, Y., Ni, J., Au, M.H., Mu, Y., Wang, B., Li, H.; *On the Security of a Public Auditing Mechanism for Shared Cloud Data Service*. Services Computing, IEEE Transactions on. 2014 , Page(s): 1. [↑](#footnote-ref-55)
56. Open Definition, URL: <http://opendefinition.org/od/#sthash.PmqzjbQW.dpuf> (2014). [↑](#footnote-ref-56)
57. Digital Government: Building a 21st Century Platform to Better Serve the American People, URL <http://www.whitehouse.gov/sites/default/files/omb/egov/digital-government/digital-government.html> (2014). [↑](#footnote-ref-57)
58. Health Data Initiative, HelthData.gov, URL: <http://healthdata.gov/> (2014). [↑](#footnote-ref-58)
59. Opower Official Website, URL: <http://opower.com/> (2014). [↑](#footnote-ref-59)
60. Data.gov.uk, URL: <http://data.gov.uk/about> (2014). [↑](#footnote-ref-60)
61. Declaration to be the World’s Most Advanced IT Nation, URL: <http://japan.kantei.go.jp/policy/it/20140624_decration.pdf> (2014). [↑](#footnote-ref-61)
62. IT dash board, National Strategy Office of Information and Communications Technology, URL: <http://www.itdashboard.go.jp/> (2014). [↑](#footnote-ref-62)
63. Please see: <http://techpresident.com/news/wegov/24940/China-Open-Data-Movement-Starting-Take-Off>. [↑](#footnote-ref-63)
64. Tim Berners-Lee. Linked Data [EB/OL]. 2011-05-15. <http://www.w3.org/DesignIssues/LinkedData.html>. [↑](#footnote-ref-64)
65. Open Data for Development Challenge, Foreign Affairs, Trade and Development Canada, <URL:http://www.acdi-cida.gc.ca/acdi-cida/acdi-cida.nsf/eng/DEN-1223131242-PCZ> (2014). [↑](#footnote-ref-65)
66. Kawasaki City Open Data, Kawasaki City Office, URL: <http://citydata.jp/%E7%A5%9E%E5%A5%88%E5%B7%9D%E7%9C%8C/%E5%B7%9D%E5%B4%8E%E5%B8%82> (2014). [↑](#footnote-ref-66)
67. L. Seeney; *K-anonymity: A model for protecting privacy*; International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002. [↑](#footnote-ref-67)
68. Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, MuthuramakrishnanVenkitasubramaniam; *L-diversity: Privacy beyond k-anonymity*; ACM Transactions on Knowledge Discovery from Data (TKDD), Vol. 1, No. 1, 2007. [↑](#footnote-ref-68)
69. Yuichi Nakamura, Kanae Matsui and Hiroaki Nishi; *Anonymization Infrastructure for Secondary Use of Data*; ICOMP'14 - The 2014 International Conference on Internet Computing and Big Data, 2014. [↑](#footnote-ref-69)
70. T. Hattori, N. Toda; Demand response programs for residential customers in the United States—Evaluation of the pilot programs and the issues in practice; March 2011. [↑](#footnote-ref-70)
71. Kengo Okada and Hiroaki Nishi; *Big Data Anonymization Method for Demand Response Services*; ICOMP'14, The 2014 International Conference on Internet Computing and Big Data, 2014. [↑](#footnote-ref-71)
72. J. -W. Byun, A. kamura, E. Bertino, N. Li; *Efficient k-anonymization using clustering techniques*; Advances in Databases: Concepts, Systems and Applications, pp. 188–200, 2007. [↑](#footnote-ref-72)