National Statistics Center
NSTAC

New data sources of Japanese official statistics in Big data era

8th Dec. 2017

Hiroe Tsubaki

CONTENTS

I Big Data Era in Official Statistics

1 Role of Statistician for Data Industrial Revolution

2 New Way: The Expansion of Data Sources for the Official Statistics

3 New Mission: The Social Needs of new data sources from Official Statistics

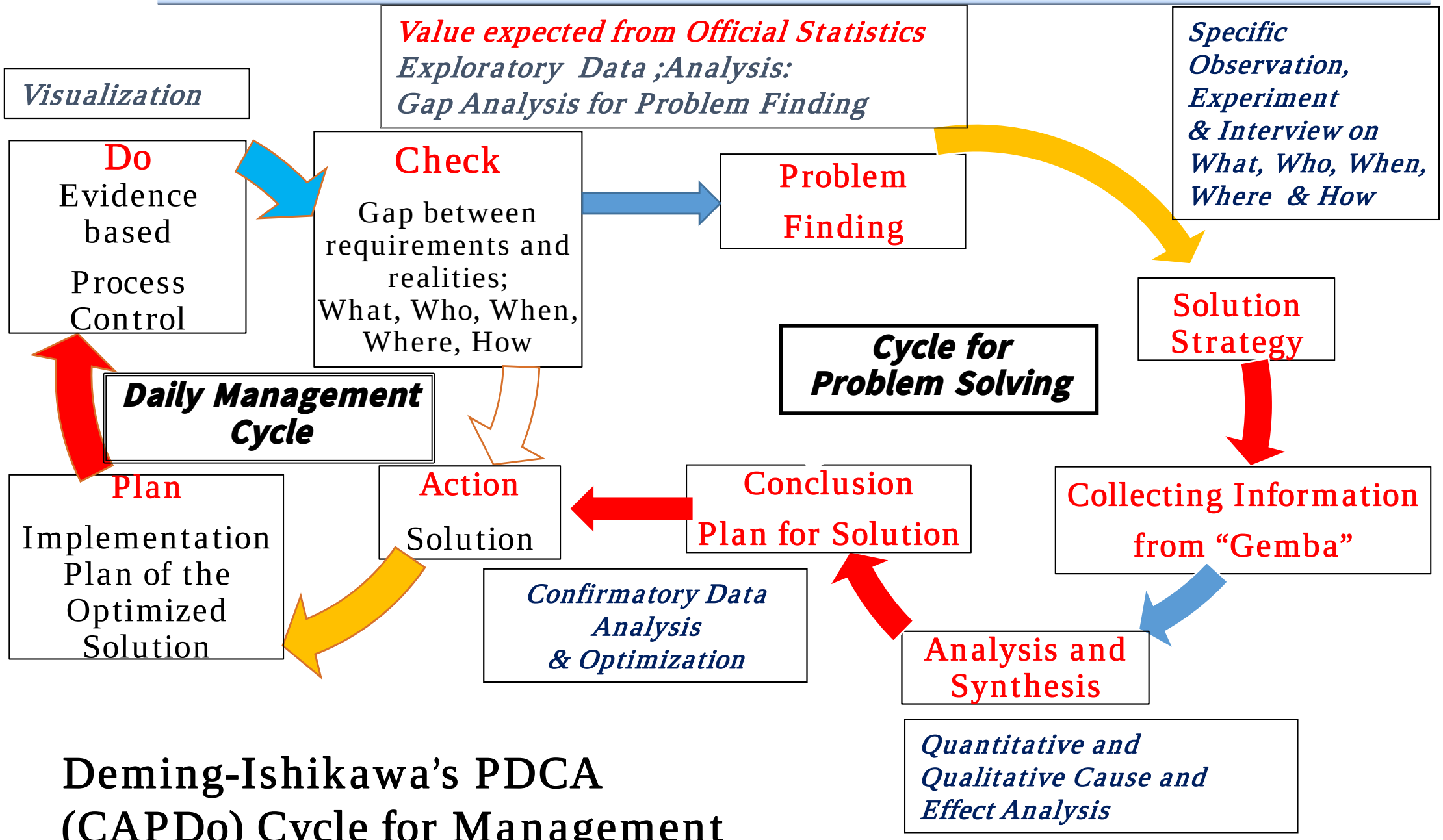
1 Role of Statistician for Data Industrial Revolution

Prof. Hal Varian

Commentary 2009/01 McKinsey & Company

- "The ability to take data – to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it – that's going to be a hugely important skill in the next decades."
- "Because now **we really do have essentially free and ubiquitous data.** So the complimentary scarce factor is the ability to understand that data and extract value from it"

1 Role of Statistician for Data Industrial Revolution

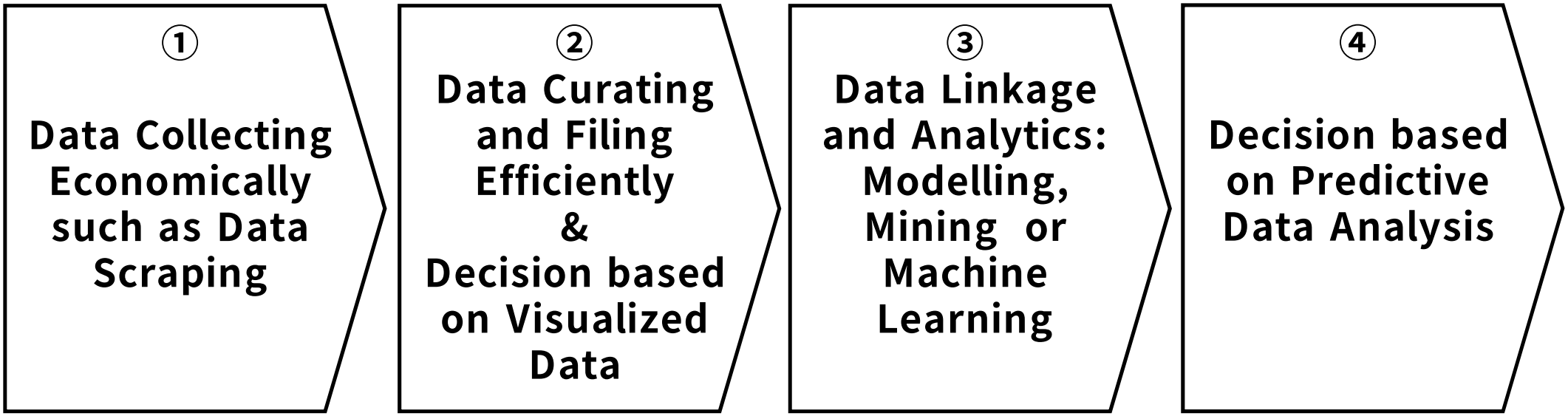


Deming-Ishikawa's PDCA (CAPDo) Cycle for Management and Improvement

1 Role of Statistician for Data Industrial Revolution

- Data Driven Consultation as a New Industry
  - Value Chain along Data Processing
- Data Driven (Evidence Based) Policy Making
  - New Way and Mission of Official Statistics Division

Conceptual Diagram of the Value Chain



Official Statistics as Social Foundation of Data Industry

1 Role of Statistician for Data Industrial Revolution

Age	Output: Main media of data provision	Input: Questionnaire
Before the Mainframe Computer Era	Paper media	Handwritten entry
Mainframe Computer Era	Paper and photographic media	Handwritten entry
After PCs appear	Paper media Text dataized documents	Handwritten OCR
Predawn of the Internet	Macro data provided on the web (PDF, Excell)	Handwritten OCR
Diffusion of the Internet	Data Provided on the web Publicly Accessible Database	Handwritten Online format

The statistical environment was changed where everybody could analyze data

The era has come when users want data to be provided in a user friendly format.

2 New Way: The Expansion of Data Sources for the Official Statistics

- Resource of data (Statistical data)
  Past: Official Survey results only
  Present: Existence of Big Data
  (incl. various private sector data)

Official statistics requires the utilization of Big Data
Because of :
- Deterioration of the official survey environment
- Timeliness of private sector data

3 New Mission: The Social Needs of new data sources from Official Statistics

- Provision of New Statistical results
  - Statistical software (easy use) & Many kinds of Data
  - Users want to analyze data by themselves

Official statistical agencies are expected to provide
- Micro Data (Raw Data)
- Easier use of Survey results
- Statistical training

II Utilization of New Input

1 Utilization of POS data

2 New challenge of utilization of Big Data

※ POS data

The data shows Point of Sales which include the information such as “what” , ”when” , ”where”, “how match” and ”how many” commodities or service were bought.

1 Utilization of POS data

- The objects of POS data are limited. The close inspection about the characteristics of the data and about the utilization are required in the official statistics.

Official Statistic

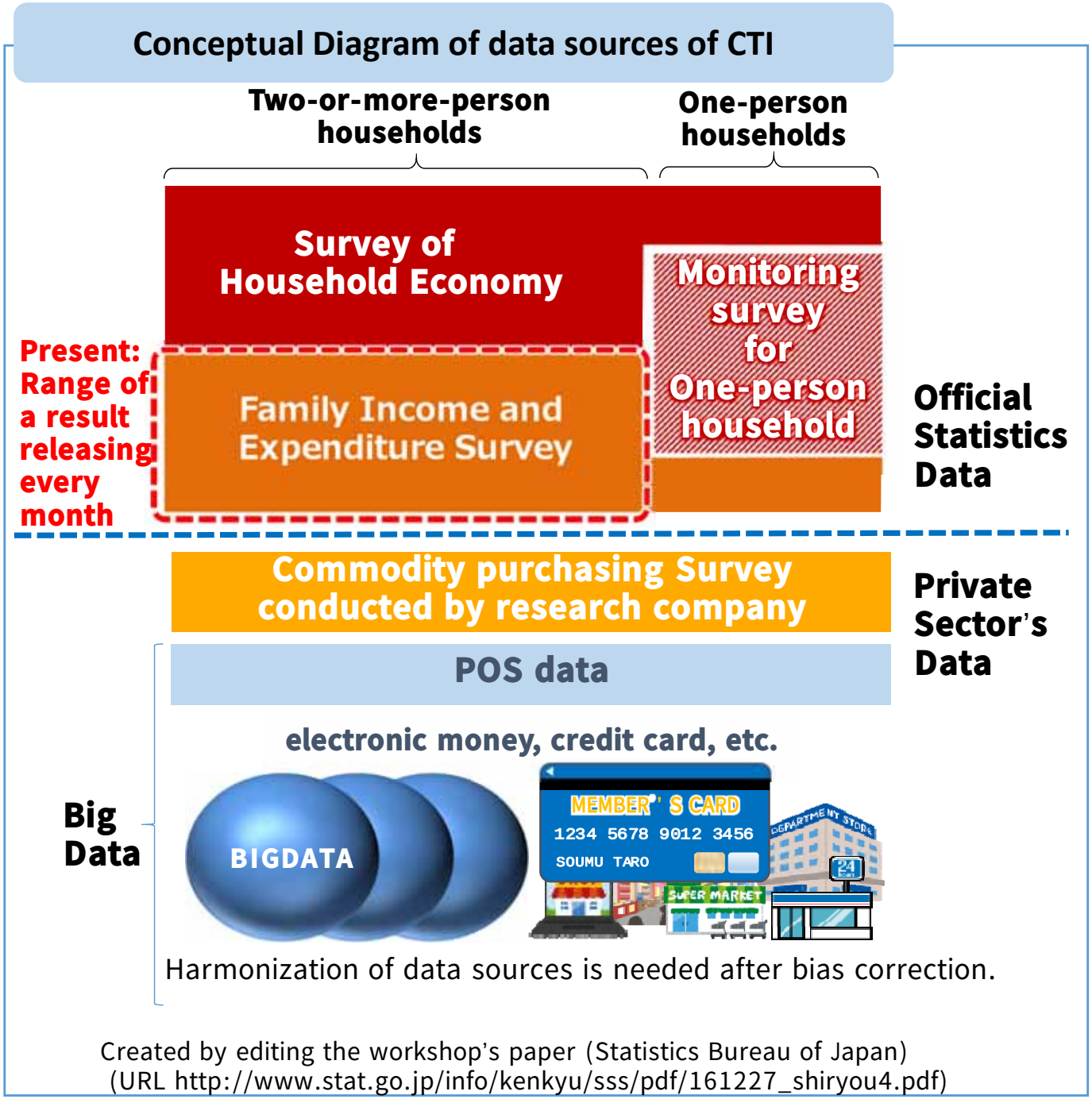
Monthly Consumer Price Index (Base 2015, 3commodities)

Private sector's Statistics

Weekly: 2014/09~
SRI-Hitotsubashi Consumer Purchase Indices

Daily : 2013/05~
Nikkei CPINow

2 New challenge of utilization of Big Data



The new challenge of creating the "Consumption Trend Index (CTI)" is being considered by the Statistics Bureau of Japan.

Regarding data sources of the index, it is planned that not only official statistics but also private sector's data including Big Data will be used.

2 New challenge of utilization of Big Data

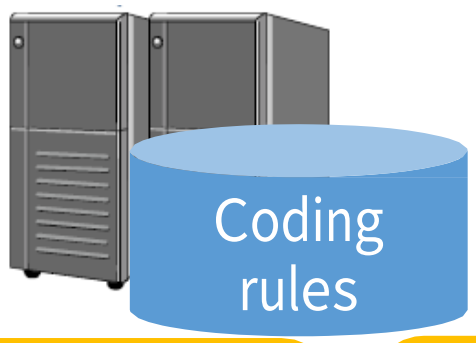
Various kinds of expressions

Character recognition

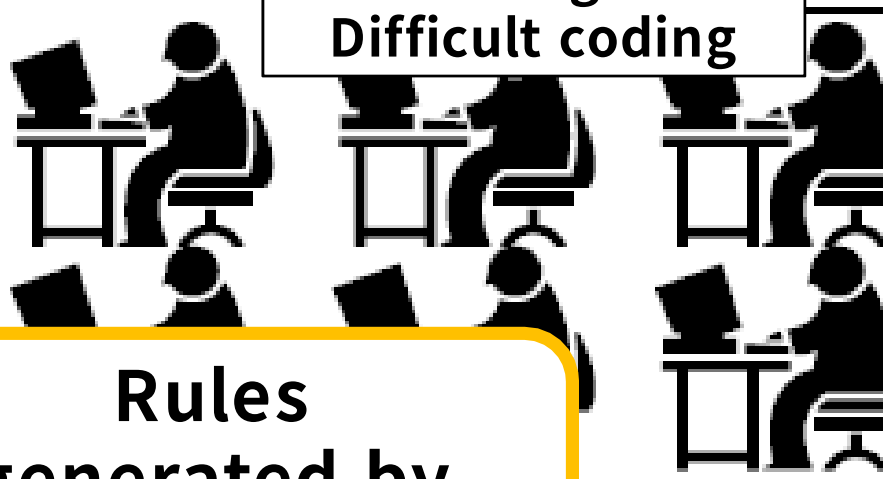
J （FINANCE AND INSURANC)

D （Sales workers)

Support by ICT



Focusing on Difficult coding



Useful rules generated by staff exploiting their expertise

Rules generated by machine learning

III Production of New Output

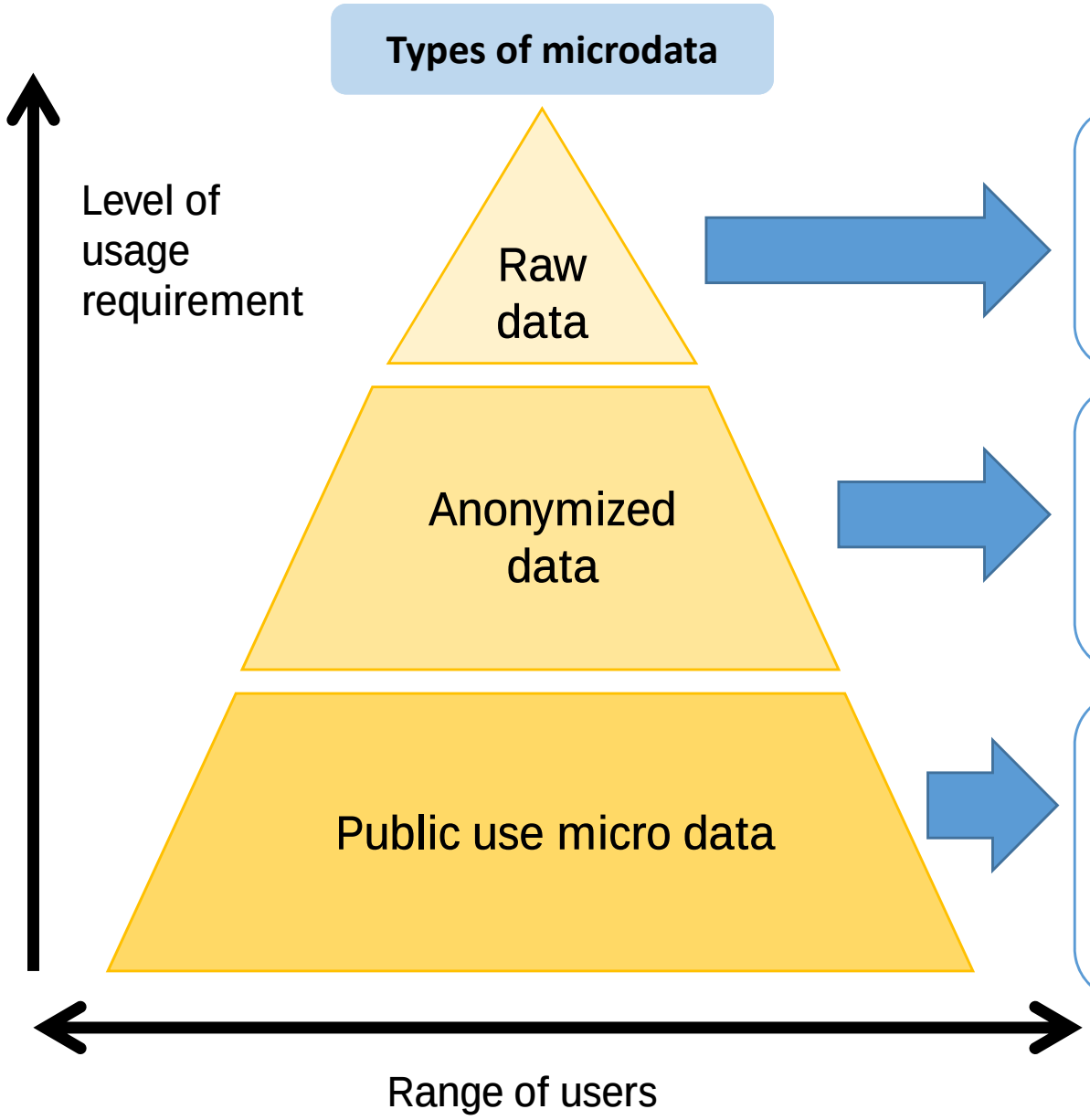
1 Different Micro Data Provision

2 Macro Data Provision in Big Data Era

3 Statistical Training

1 Different Micro Data Provision

Provision of micro data by demand level of users



Type of users & services

③On-site use of statistical microdata

Researchers who want to analyze micro data deeper even with strict restriction.

②Provision of Anonymized data

Researchers who want to analyze micro data. Because anonymized micro data is provided with lower restrictions.

①Release of microdata files for public use

Beginners, who want to know how to use micro data and want to analyze it simply and easily without any restriction.

① Release of microdata files for public use

Public use microdata files

- **PUMF can be used freely as data for preliminary study as part of data science in statistics education or business without entailing difficulties related to the confidentiality of individual household information.**
- **PUMF are not questionnaire information (microdata) itself, but simulative microdata which were randomly generated from the estimated correlation structure of the tabulated results.**

Type of household	Record count	Contents of household attributes, etc.	Classification of income and expenditure number
All households	45,811	7 items (3 major metropolitan areas or not, number of household members, number of earners, type of tenure of dwelling, characteristic of household head such as age(2 items), employment status)	12 items (Yearly income, consumption expenditures, 10 major groups)
			422 items (Yearly income, consumption expenditures, 10 major groups, 410 items)
Worker's households	26,239	4 items (characteristic of household head such as age, industry, occupation, size of enterprise)	Same as the above

② Provision of Anonymized data

Provision of Anonymized data

The anonymized data provision service aims to provide (lend) applicants who have applied for use of data with questionnaire data obtained from statistical surveys as anonymized data, which was processed so that no survey objects can be identified (anonymization processing: not only deleting information that allows for direct identification of individuals, such as name, but also categorizing information more broadly by integrating various detailed categories for regions and attributes, and deleting distinguishing data).

Survey Name	Time period
Population Census	2000, 2005
Labour Force Survey	1989～2012
Housing and Land Survey	1993, 1998, 2003
National Survey of Family Income and Expenditure	1989, 1994, 1999, 2004
Employment Status Survey	1992, 1997, 2002, 2007
Survey on Time Use and Leisure Activities	Questionnaire A 1991, 1996, 2001, 2006 Questionnaire B 2001, 2006
Comprehensive Survey of Living Conditions	1998, 2001, 2004, 2007, 2010

③ On-site use of statistical microdata

Conceptual Diagram of on-site use utilizing remote access

National government offices and ministries
(survey conductors, such as SBJ)

Decides on permission for application for use and taking data

- Registration for the questionnaire information.
- Entrusts necessary related business, such as service counter, to NSTAC.

- Service counter
- Formality check of application for use, and formality examination of taking data
- Management of the data and system, etc.

On-site facilities

Dedicated servers

On-site facilities

SINET
(Uses the VPN service)

Administrator

National Statistics Center
(Central-data-management facilities)

Virtual PCs

Operates virtual PCs by remote control

Users

Monitoring camera

On-site facilities

Displays the tabulated/analysis results.
(Using memory equipment, such as a USB memory stick, is prohibited and it cannot be used.)

↓

Formality examination is conducted when taking data

Users

On-site facilities

③ On-site use of statistical microdata

Merits of on-site use

Present (provide with DVD)

- **Use condition**: It must be the use of microdata in research deemed to provide a public-benefit.
- **Security**: Researchers are responsible for ensuring security at large.
- **Appli-cation**: User needs to obtain permission by submitting an application for use including the detailed design of tabulation and analysis.
- **Micro-data**: Only the minimum information required for the designed analysis is provided.

Exploratory and creative research is **difficult.**

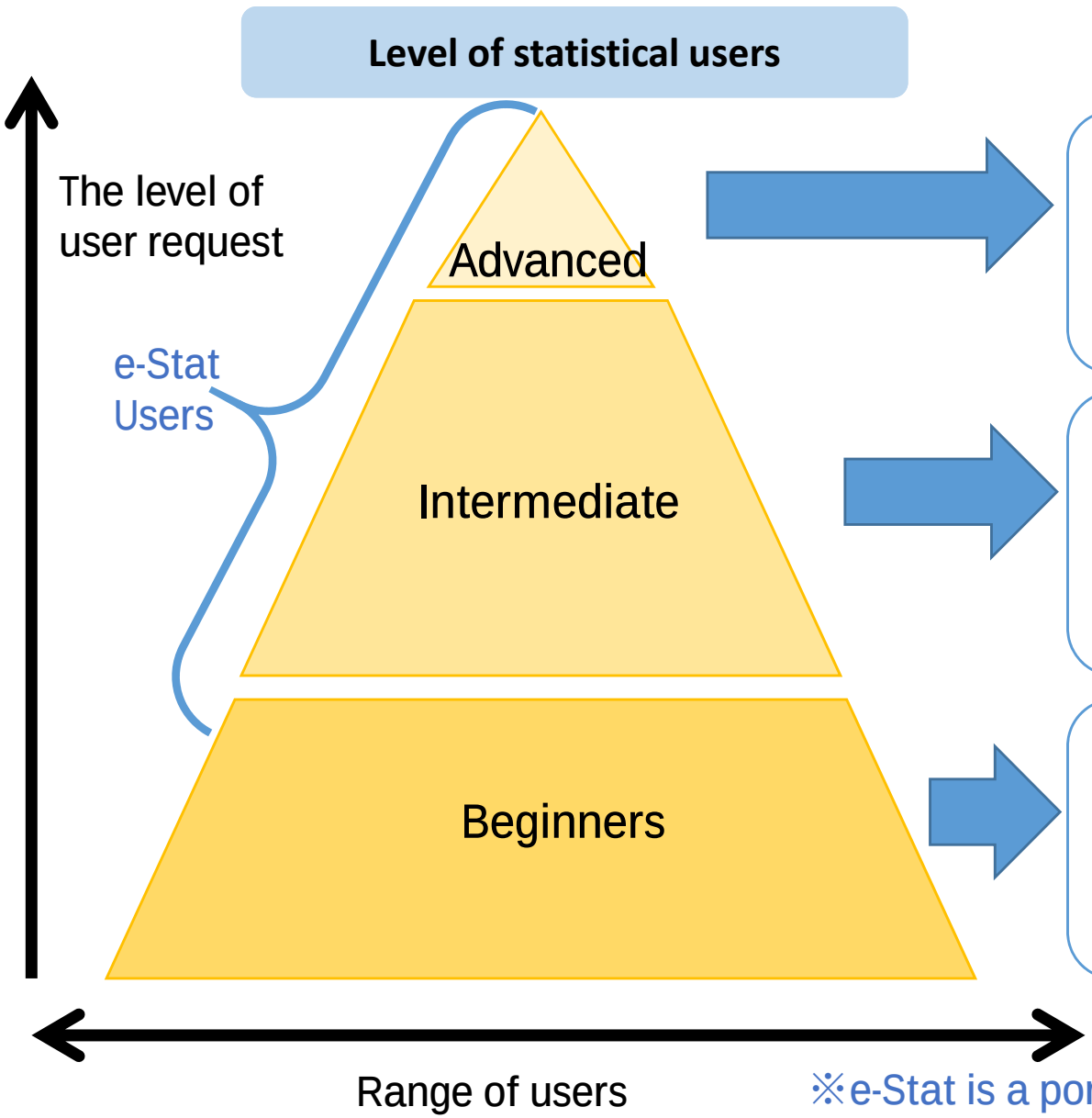
Future (on-site use)

- **Use condition**: It must be the use of microdata in research deemed to provide a public-benefit.
- **Security**: Facility installation personnel are responsible for ensuring a secure environment.
- **Appli-cation**: User burden is reduced by simplifying the application for use.
- **Micro-data**: All the information is available for use.

Exploratory and creative research is **possible.**

2 Macro Data Provision in Big Data Era

Provision of Macro Data for Potential User



Type of provision services

③Statistical LOD

This data is provided for e-Stat users who need next-generarion data provision service.

②API function & jSTAT MAP

These functions are provided for mainly e-Stat users who use official statistics regularly but want to use them more conveniently .

①Statistics Dashboard

This system is provided mainly for beginners who want to use official statistics but never or hardly used them before.

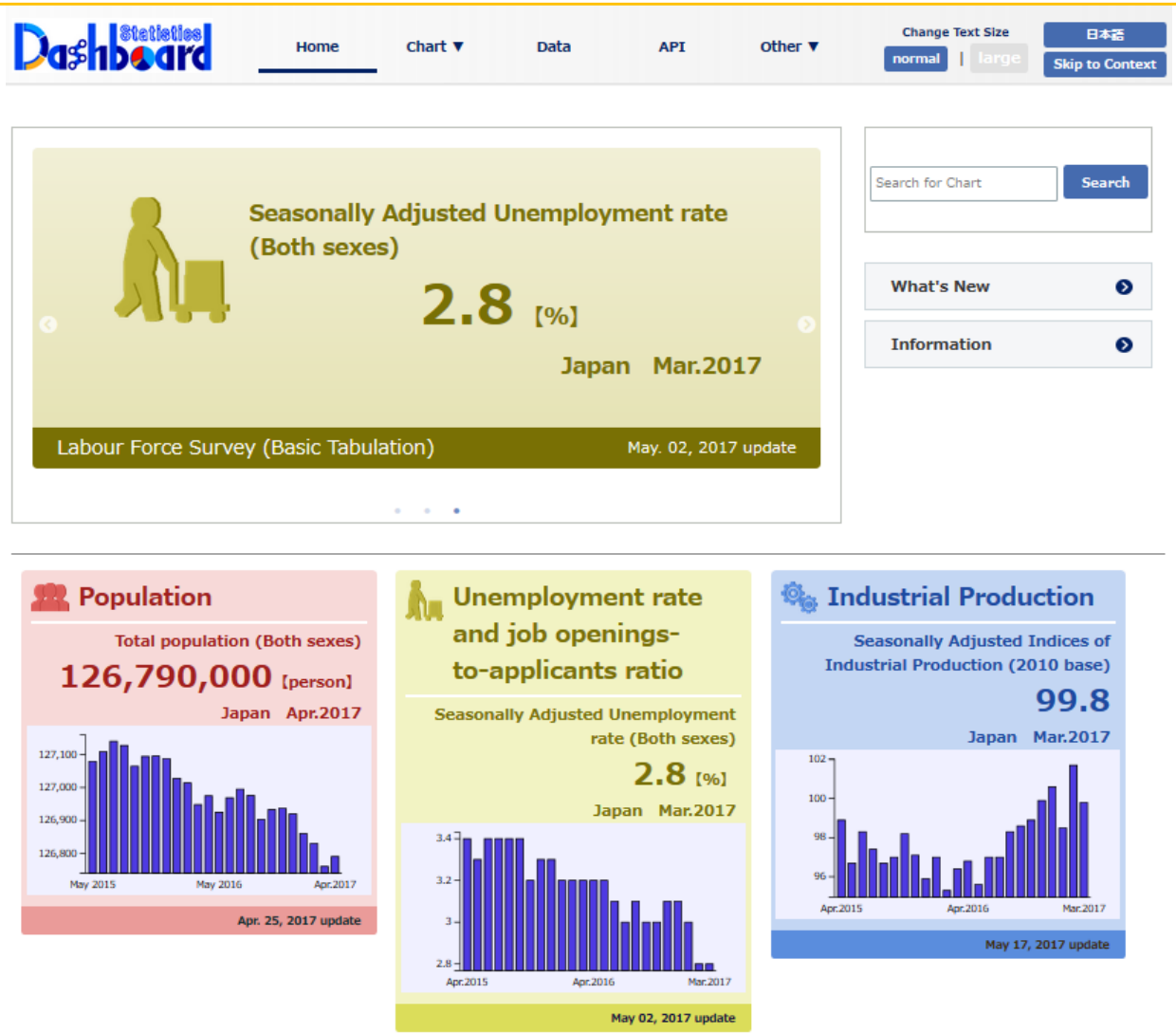
※e-Stat is a portal site for Japanese Government Statistics. Users can search and download statistical table freely.

① Statistics Dashboard

Statistics Dashboard

Statistics Dashboard is a system that summarizes various statistical data and displays a set of graphs and charts based on the processed data

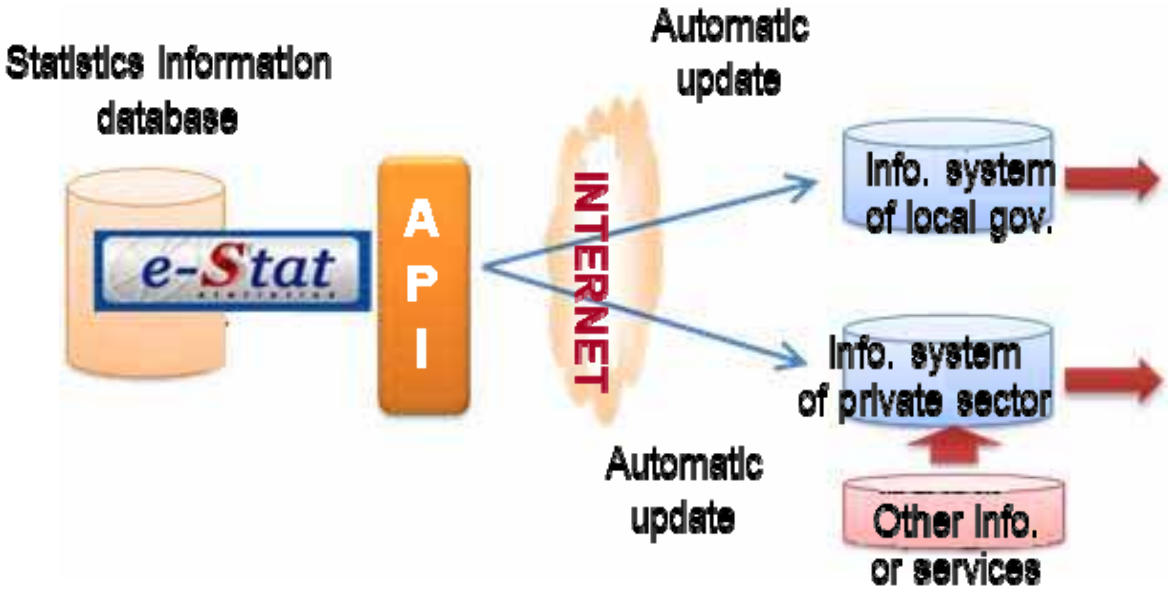
Example:
Top page of
the statistics dashboard
of one day



② API function & jSTAT MAP

Introduce API functions

API (Application Programming Interface) functions provide statistical data converted to machine readable data



Example 1:
Update data of e-Stat automatically

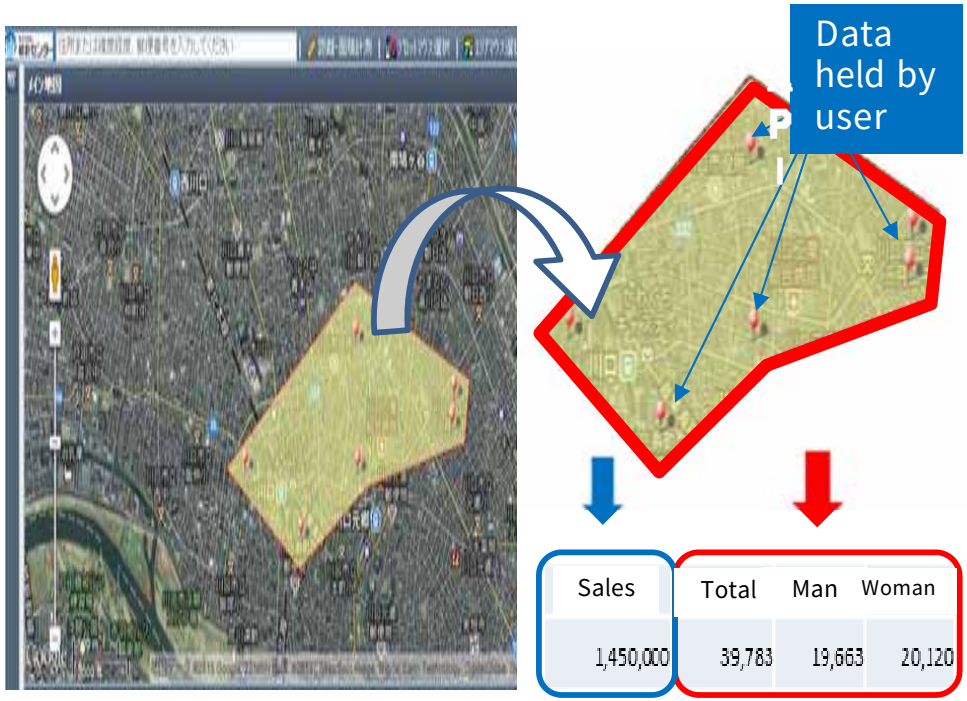
Example 2:
Mash-up with other data of user or data available from the Internet

Developer support information is also available

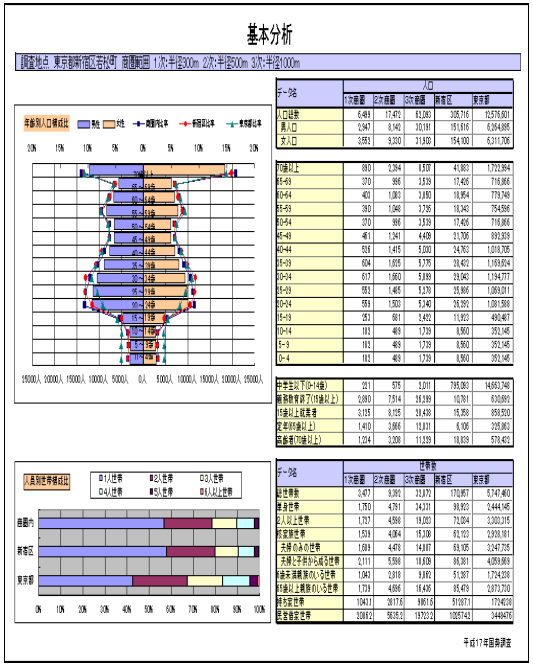
② API function & jSTAT MAP

Small area analytics on maps (jSTAT MAP)

Provide functions that enable users to tabulate statistics in any arbitrarily designated area and to import and tabulate any data owned by users



Example 1:
The function enables retrieving various data held by data or making use of statistics data in an arbitrarily designated area



Example 2:
Prepare a report on the results of basic analysis including the age structure in the selected area

An App for Tablet is also available

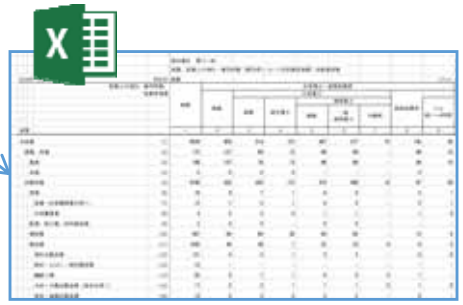
③ Statistical LOD

Statistical LOD - Five Levels of Open Data -

From link to files to link to data

Link to files

Addresses are given to each file
(http://www.e-stat.go.jp/xls0001.xls)



Link to data

	Total 総数（男女別）			Male 男				Fem 女	
	…	44歳【人】	45歳【人】	…	…	44歳【人】	45歳【人】	…	…
…	…	…	…	…	…	…	…	…	…
Saitama さいたま市	…	16,130	19,245	…	…	8,293	9,938	…	…
Kawaguchi 川口市	…	6,582	8,022	…	…	3,526	4,289	…	…
…	…	…	…	…	…	…	…	…	…

Addresses are given to each data item
(http://data.e-stat.go.jp/lod/…/obs00001)

DBpedia Japanese

ウィキペディア Wikipedia

③ Statistical LOD

Applications will be standardized

Data are standardized
(Use of international standards (RDF※1))

Access procedures are standardized.
(Use of international standards (SPARQL※2)

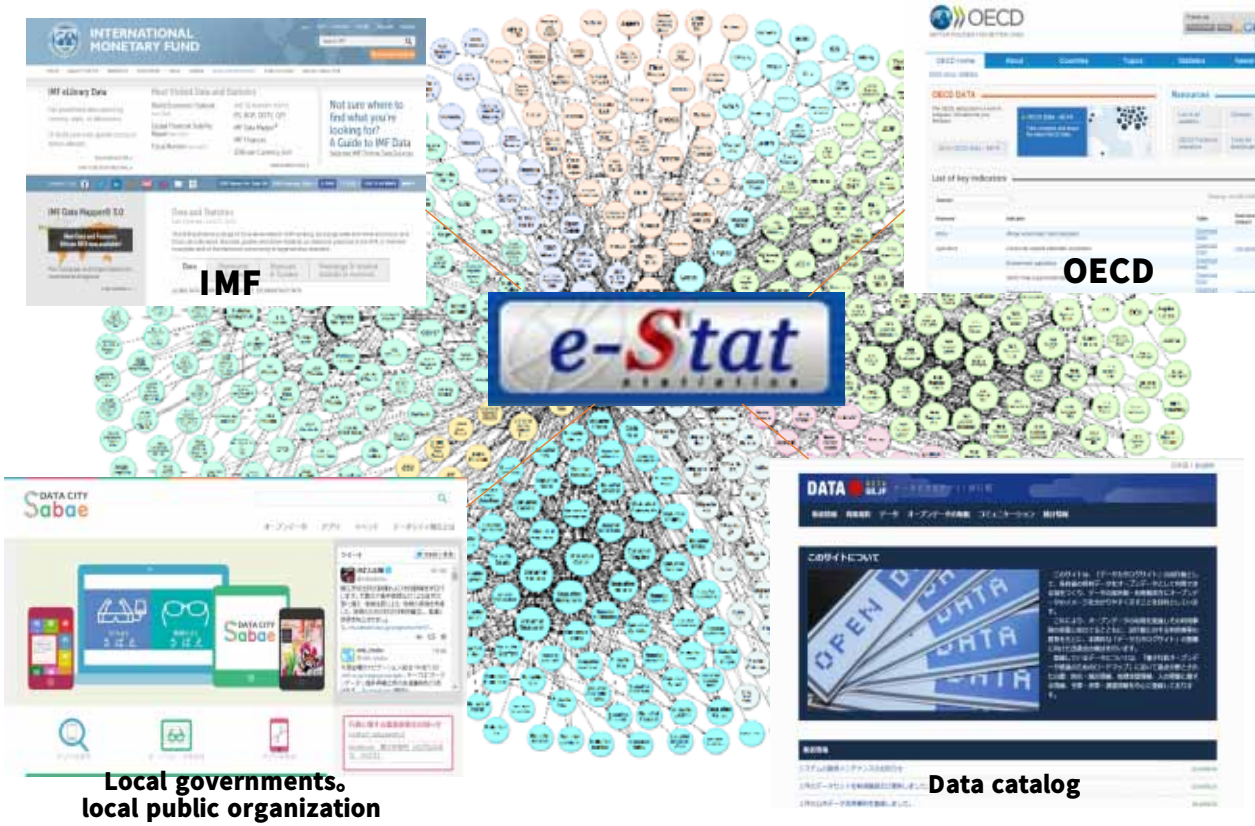
※1 RDF(Resource Description Framework): A unified framework recommended by W3C (1999/02) (an international body which promotes to standardize techniques used on web.

※2 SPARQL: A language recommended by W3C (2008/01) to search RDF.

Merits as LOD of statistical data

Defining data uniquely in the Internet by (URI), and expressing relationships by links.

Links with other data



Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/ CC BY-SA

3 Statistical Training

The first Massive Open Online Course(MOOC) provided by the Government of Japan

(1)Introduction course　(2015～)

about 10 minutes × 4～7videos × 4 weeks

(2)Pratical course for business　(2016～)

about 10 minutes × 5～6videos × 5 weeks

(3)Pratical course for official statistics　(June 2017～)

about 10 minutes × 5～7videos × 4 weeks

Not only for officials but also for general users

IV Conclusion

- Big Data Era is requiring from statistics agencies both new data and human resources for new value creation.

- Official statistics agencies should make their best efforts to realize the true value co-creation with their users

Thank you!

National Statistics Center (NSTAC)
http://www.nstac.go.jp/en/index.html