Japanese government project on innovation database platform
- As an infrastructure for improving quality of science, technology and innovation policy -

PICMET 2014 Conference
"Infrastructure and Service Integration"
July 27-31, 2014, ANA Crowne Plaza Hotel, Kanazawa, Japan

Hiroyuki Tomizawa[1], Natsuo Onodera[1], Yasuo Nakayama[1] and Kenta Nakamura[2]

[1] National Institute of Science and Technology Policy (NISTEP),
Ministry of Education, Culture, Sports, Science & Technology (MEXT), Japan

[2] Economic Department, Kobe University, Japan

Abstract

The Japanese government has been implementing a research program aimed at rationalizing the process of science and technology innovation policy-making. Part of this is a project to build an innovation database platform, which has been underway since 2011. The primary goal of this project is making available systematic data for promoting empirical quantitative research on science, technology and innovation policy. This paper describes the goal of this project and what the resultant data infrastructure could bring. Insights gained through past activities are reported as well.

What is valued in the public and higher education sectors is creating databases for bibliometrics, a powerful methodology for the quantitative analysis of output from scientific research. The project aims to eliminate the division between each levels of data (macro, mezzo and micro) arising from differences in the ways to create data, and to establish linkage between them, which we call a "vertical linkage" for its ability to connect upper and lower levels of data. The project also emphasizes "horizontal linkage," or links between bibliometric and non-bibliometric data such as data on R&D and research funding.

In the business sector, which is the other key area, the project focuses on establishing mutual connections among different data sets, primarily including patent and corporate data, as well as data on innovation activities, intellectual property right (IPR) activities, and investments in R&D and intangible assets. The biggest hurdle to overcome in forming such mutual links between various data held by different companies lies in the organizing of patent data, a task that would call for a process known as "data cleansing," as in the case of bibliometric data.

Almost three years have passed since the start of these efforts to develop databases. The databases are already available and have been used by policy researchers. The next challenge is how to exploit the outcome of such research for policy-making.

2

Outline of the presentation

- Basic concept of the project
  - Attempts towards "evidence-based policy making"
  - Increasing possibilities to use and analyze data
- Contents of Database platform and methodology
  - Project 1: Database platform on public R&D system
    - Horizontal data linking (between macro-level and micro-level data)
    - Vertical data linking (among multiple kinds of data)
    - "Name harmonization" of bibliometric data
  - Project 2: Database platform on R&D and innovation in industry
    - "Name harmonization" of patent data
    - Data representation on historical changes of companies
- Discussion:
  - Effects of name harmonization
  - Applications of the database platform

3

Part 1

Basic concept of the project

4

Constructing data infrastructure: basic concept

- Resources for policy studies on S&T and innovation
  - Improve scientific methodology of policy studies
  - Stimulate researchers to join in policy studies from various disciplines

- Foundation for more rational policy-making process
  - Data as "evidence" for policy making
  - Tool to improve quality of policy discussion
  - Base of policy evaluation and monitoring

- Public accountability
  - Data as a tool for explanation
  - Openness of policy formation process

5

Direction of initial stage of project

- Linking *input*, *output* and *outcome*
  - For analyzing productivity or spillover, it is needed not only for developing data of them but also for linking them to each other

- Active use of **micro data**
  - It enhances possibilities of analysis and to get high quality evidence
  - It is essential to link input, output and outcome
  - Micro data of government statistics recently became available for research purposes in Japan

- Not "**analysis for analysis**" but "**analysis for assessment of policy**"

- Develop the data demanded for research purposes
  not reasonable to be prepared by individual researchers
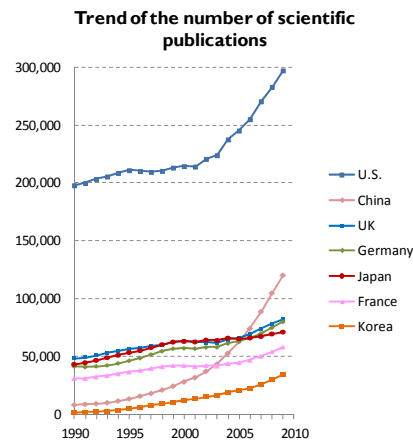
6

Part 3

Contents of Database platform

7

Project 1:
Database platform on Public R&D system

- Public R&D system:
  - Government organizations or R&D performers funded by the government
    - Higher education institutions (universities)
    - Public research institutes
    - Public R&D funding
  - Conventional S&T policy targets
  - Producer of seeds of innovation, not main performers of innovation
  - Producer of knowledge that is not provided by the market mechanism
    - S&T knowledge as public goods = Not clearly visible

8

A concern about performance of the public R&D system of Japan: the need for micro data

Trend of the number of scientific publications

- U.S.
- China
- UK
- Germany
- Japan
- France
- Korea

Source: *Benchmarking Scientific Research 2011*, NISTEP, 2011.
Data: Based on whole count of publications in *Web of Science*, Thomson.

- The Japanese government has given preference to the R&D budget for the past 15 years
- But the number of scientific publications has fallen into stagnation
- This should be investigated:
  - **Micro data** is essentially important ⋯
  - If we use macro data only, the object of analysis cannot be treated but as a black box

Research input → Public research system → Research output

9

Multi-level data linking

- **Horizontal link:** various data is linked to each other
  - Especially, link between input data and output data
- **Vertical link:** macro, meso and micro data are linked
  - Consistency among levels: Macro and meso data can be acquired by adding up micro-level data
- Developing multi-level data that enables both horizontal and vertical linking

- Example:
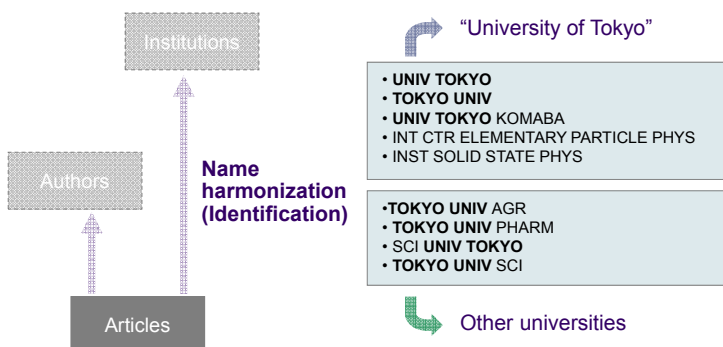 Link individual data of R&D statistics with the bibliometric data
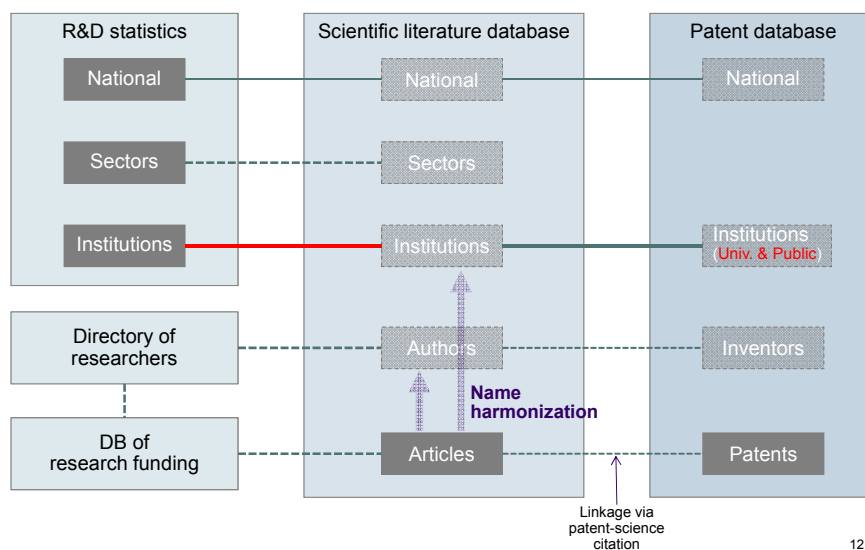 - R&D input and output are linked with each other at the multi level

Macro ⟷ Macro

Meso ⟷ Meso

Micro ⟷ Micro

R&D statistics data   **Bibliometric data**
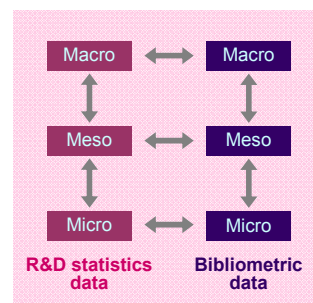
10

A methodological challenge : "Name harmonization" of bibliometric data

■ Micro data in scientific publication database

- Individual article ↔ each record of database
- Individual author ↔ recorded in db but problems of fluctuations in writing, same names, etc.
- Individual institute ↔ recorded in db but problems of fluctuating names, etc.

Institutions

"University of Tokyo"

• **UNIV TOKYO**
• **TOKYO UNIV**
• **UNIV TOKYO** KOMABA
• INT CTR ELEMENTARY PARTICLE PHYS
• INST SOLID STATE PHYS

Name harmonization (Identification)

Authors

•**TOKYO UNIV** AGR
• **TOKYO UNIV** PHARM
• SCI **UNIV TOKYO**
• **TOKYO UNIV** SCI

Articles

Other universities

11

Multi-level data link on public R&D system

R&D statistics	Scientific literature database	Patent database
National	National	National
Sectors	Sectors	
Institutions	Institutions	Institutions (Univ. & Public)
Directory of researchers	Authors	Inventors
DB of research funding	Articles	Patents

Name harmonization
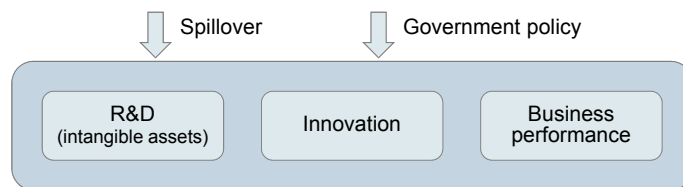
Linkage via patent-science citation
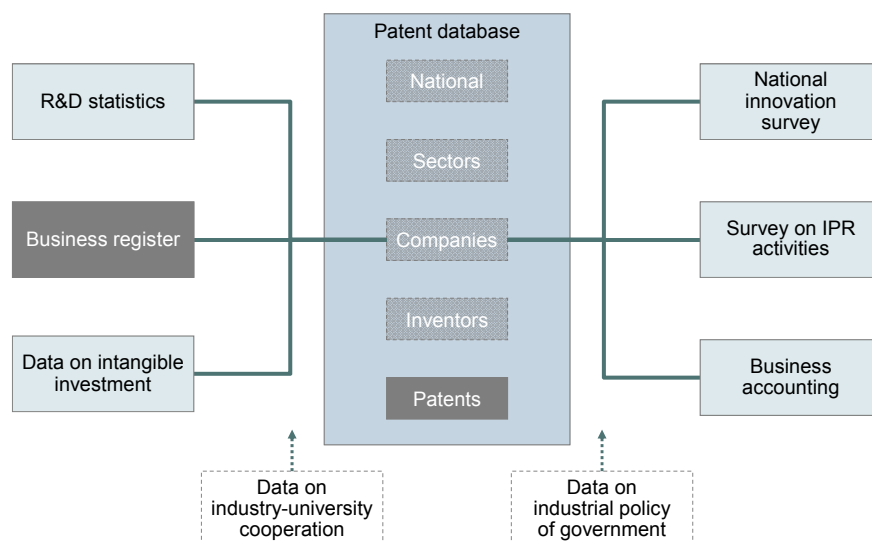
12

Project 2:
Database platform on R&D and innovation in industry

- Purpose of data use
  - Understanding innovation in industry
  - Grasping and analyzing spillover of public R&D to innovation
  - Analysis and evaluation of effects of policy for innovation
  - (contribution of innovation to economic growth)
- Approach to data organizing
  - At the level of individual companies, linking data on innovation activities, investment in R&D and intangible assets and IPR activities to each other
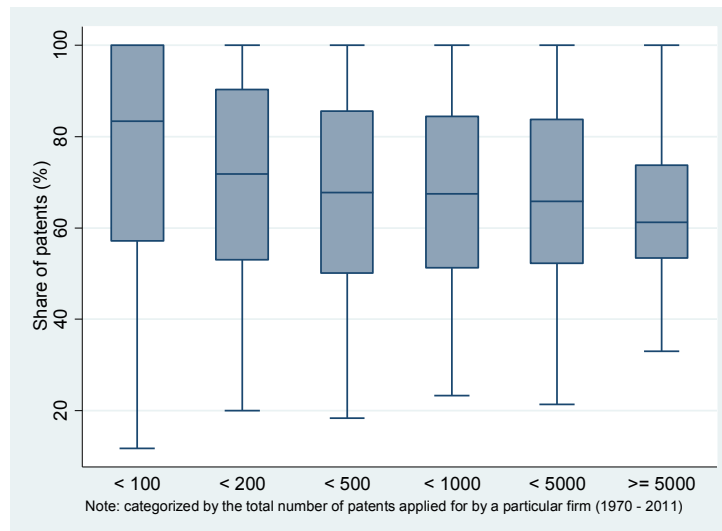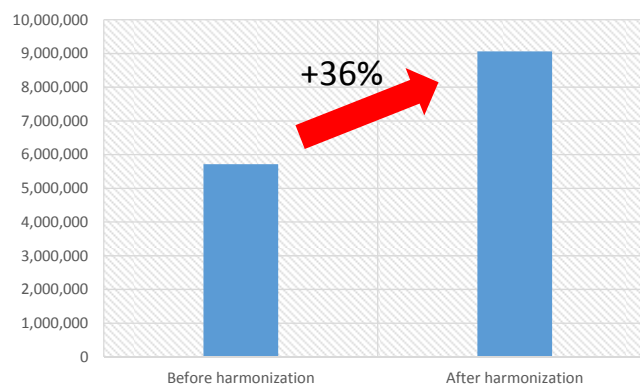
Spillover   Government policy

| R&D (intangible assets) | Innovation | Business performance |

13

Linking data on industrial innovation

Patent database

| R&D statistics | | National | | National innovation survey |

Business register — Companies — Survey on IPR activities

Sectors

Inventors

| Data on intangible investment | Patents | Business accounting |

Data on industry-university cooperation   Data on industrial policy of government

14

The share of patents of the original applicant name having the most patents

Note: categorized by the total number of patents applied for by a particular firm (1970 - 2011)

The number of patents before-and-after name harmonization

+36%

Summary of conclusions

- Technical achievement
  - "Name harmonization" of bibliometric and patent data were successful
  - The unsolved problem remains of how to represent historical changes of companies in the database
- What does using a database platform enable?
  - Expanding range of effective data
  - Reducing burden of data cleaning in policy research
  - Ensuring consistency among macro, meso and micro data
  - Analyzing multiple data sets (e.g., innovation input and output) at several levels becomes possible

17