NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

IDR Informatics Research Data Repository

Sharing Real Business Purpose Datasets for Academic Research

Center for Dataset Sharing and Collaborative Research (DSC)

Keizo OYAMA, Tomoko OHSUGA

National Institute of Informatics, ROIS

Datasets provided via NII-IDR

16 datasets by 8 companies, and more…

- Yahoo! Dataset
- Rakuten Dataset
- Niconico Dataset
- Recruit Dataset
- Cookpad Dataset
- Lifull HOME's Dataset
- Fuman Dataset
- Sansan Dataset
- NTCIR Test Collections
- Speech Corpora

Simple Statistics of IDR

- 16 datasets by 8 private companies (as of Oct. 2017)
  - Users:
    - 692 labs. in total (482 distinct)
    - 2,356 individuals
  - Publications using the datasets:
    - 568 papers
- NTCIR Test Collection
  - Users: ＞4,000 labs; ～600 individuals
- Speech Corpus
  - Users: ＞3,600 labs. in total (1,200 distinct)
  - Publications: ～750

Why to share Real Business Purpose Datasets?

- Researchers' needs:

  Research and real application are getting closer in IT, AI, etc.

  → **Real and Large Scale Data generated in Real Business**

- Private Companies' incentives to provide data:
  - Social Contribution
  - Business/Technical Seeds Seeking
  - Future Collaboration
  - Recruitment …

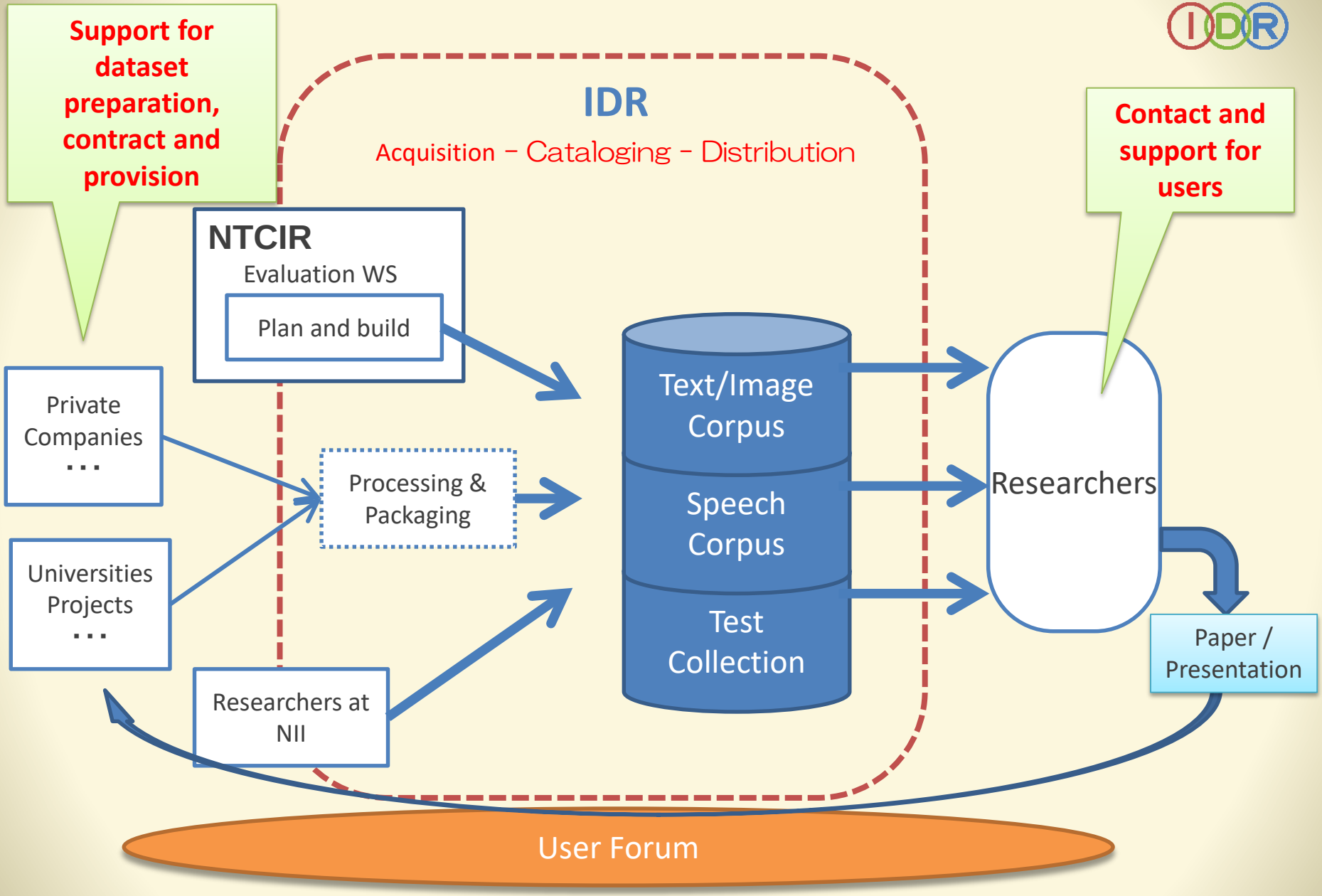
Problems

- Request-based approach:
  - Researchers:
    - difficult to know contact; no guarantee for identity …
  - Companies:
    - load of user-by-user dealing (data preparation, contract,etc.) …
- Do-It-Yourself approach (when data are accessible on the Web):
  - Researchers:
    - huge cost of crawling; risk of infringing others' right …
  - Companies:
    - load on their service system; risk of damaging business; unable to grasp/assess each user and usage …
- Open Access approach:
  - Companies:
    - difficulty of controlling risks; hard to grasp/assess each user and usage (even though they can claim for credit) …
  - Researchers:
    - real data are rarely provided …

Sharing Datasets via NII-IDR

IDR: dataset sharing activity at DSC

- Functions as a Hub:
  - Providing Common Datasets (that cannot be made open)
    - Collecting, Accepting, and Distributing Datasets
    - Mediating Researchers and Companies
  - Sharing Know-hows from Creation to Distribution
    - Various Know-hows are required for building/providing datasets:
      licensing; user agreement; risk management; data specifications; collecting/annotating/distributing method …
  - Activating Research by Creating/Connecting Communities across research fields, data owners/creators/users, etc.
    - → Promoting Collaborative Research
    - → Hosting Evaluation Workshops
    - → Holding Users' Forum


Support for dataset preparation, contract and provision
IDR
Acquisition – Cataloging – Distribution
Contact and support for users
NTCIR
Evaluation WS
Plan and build
Private Companies …
Universities Projects …
Processing & Packaging
Researchers at NII
Text/Image Corpus
Speech Corpus
Test Collection
Researchers
Paper / Presentation
User Forum

Origins of Datasets

- Real Business-purpose Data by Commercial Internet Services
  - Yahoo! Dataset
  - Rakuten Dataset
  - Niconico Dataset
  - Recruit Dataset
  - Cookpad Dataset
  - Lifull Home's Dataset
  - Fuman Dataset
- Research-purpose Data by Researchers/Research Organizations
  - Speech Corpus
- Research-purpose Data created via Evaluation Workshop
  - Sansan Dataset
  - NTCIR Test Collection (NII) *(some are using Real Business Data)*

Merits of Common Datasets

- for each Researcher:
  - Can ensure **reproducibility and transparency**
  - Easier to **compare** results with other research
  - Easy to **appeal** the research results
- for Research Community:
  - **Platform for Comparative Evaluation** of Techniques:
    setting common tasks, defining evaluation methods, accumulating research results, …
  - Enhance **Community** and open up **Cross-Disciplinary Collaboration**
- for Data Provider:
  - Reduce the load of user-by-user dealing
  - Make the social contribution known to the public
  - Can appeal openness and fairness

Risky Issues and Measure

- Risky issues (companies may not be aware of …)
  - Personal Information
    - Consent to provide to third party researchers for research purpose?
    - Possibility of being included in users' posts? System to eliminate them?
    - To what extent the related law allows to provide without the person's consent?
  - Copyright
    - Are posts accepted under proper terms and conditions?
    - Possibility of including posts infringing others' copyright? System to eliminate them?
  - Infringement of Privacy or other Human Rights
    - Possibility of including posts infringing others' privacy, slander or defamation? System to eliminate them?
  - Other Contents rousing Social Criticism, especially on the Net
- Measure
  - Advise the company to revise their terms of services, to process or remove some part of data, and so on
  - Prohibit researchers to disclose problematic content by user agreement
  - Set up a contact network with the users and the company to report problems

Licensing to User

Depending on provider's choice …

- Direct licensing by provider
  - Concluding a contract
    - Rakuten
  - Agreement to provider's user policy & approval by provider
    - Cookpad
  - Agreement to provider's user policy and online registration
    - Niconico, Sansan, NTCIR (some need contract)
- Sublicensing by NII
  - Concluding a contract
    - Yahoo!, Recruit, Lifull, Fuman
  - Agreement to provider's user policy & approval by NII
    - Speech Corpus
- Open Access
  - Datasets provided by some researcher/research organization

Restrictions on Data Usage

- Worries of data providers (especially for private companies)
  - Copyright
  - Privacy and Personal Information of its Service Users
  - Flaming caused by Abuse
  - Damage to the Business

  → Managing users and restricting usage are necessary in most cases.
- Restrictions depend on the nature of data and company's policy
  - Prohibited by all companies:
    - providing data to third party; commercial use
    - disclosure of identified person/organization even in academic publication
  - Prohibited by some companies:
    - matching data with other data (especially data from the originating Internet service)
  - Required by some companies:
    - checking content before publishing research result

  … still difficult to provide data sensitive even if only a little

Activating Research

- Sharing Research Results
  - Supporting Research Meeting focusing on Data Set
    e.g. Rakuten R&D Symposium
- Sharing Problems and Ideas
  - Planning Meeting gathering Data Owner and Researcher
    - Ideathon held in advance of releasing Recruit Data
    - Session in 2015 HCG Symposium “Forefront of research using large scale cooking recipe data” Dec. 18, 2015 in Toyama
    - Ideathon: “Workshop on Open Data of Japanese Classical Documents” Dec. 18, 2015 in Kyoto
- Creating Communities
  - Evaluation Forum using Data Sets (e.g., NTCIR)
    - Community QA Pilot Task using Yahoo! Chiebukuro Data
    - Cooking Recipe Search Pilot Task using Rakuten Recipe Data
  - Accumulating and Sharing Know-hows for Competitions
    - BIGCHA – Big Data Programming Challenge using Common Data Sets
      e.g. Yahoo!, Rakuten, Niconico, Recruit, Cookpad, …

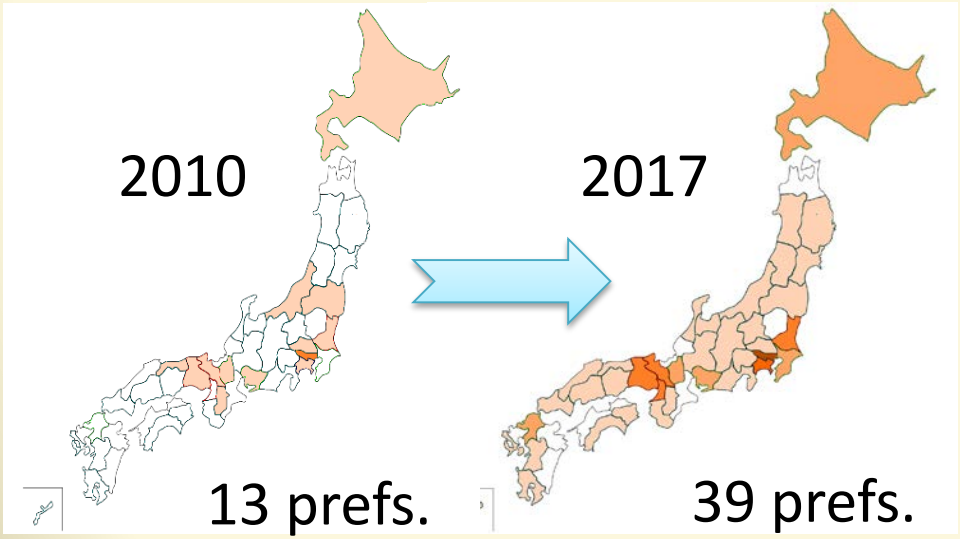
Simple Statistics of IDR

- 16 datasets by 8 private companies (as of Oct. 2017)
  - Users:
    - 692 labs. in total (482 distinct)
    - 2,356 individuals
  - Publications using the datasets:
    - 568 papers
- NTCIR Test Collection
  - Users: ＞4,000 labs; ～600 individuals
- Speech Corpus
  - Users: 3,600 labs. in total (1,200 distinct)
  - Publications: ～750

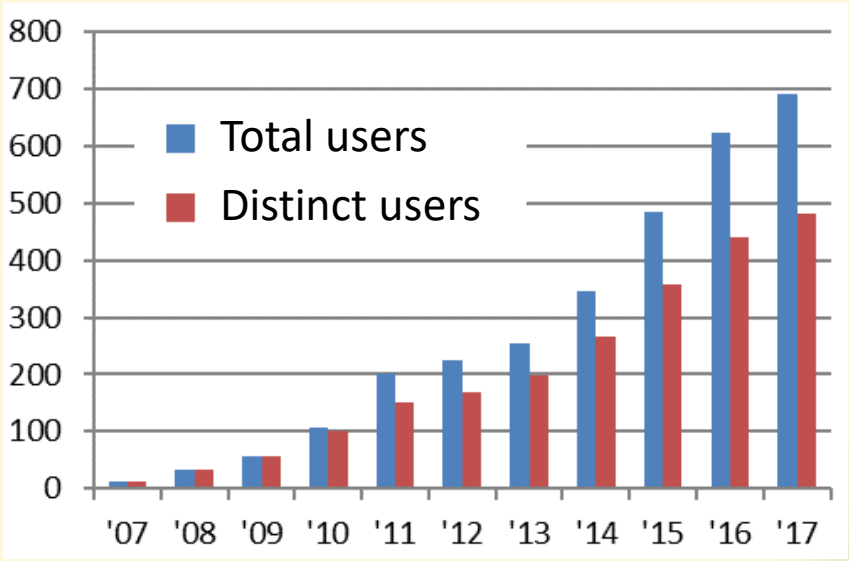
Trend of Dataset Users (～2017.11)

datasets by private companies only

User Loction
(Lab. Users only)

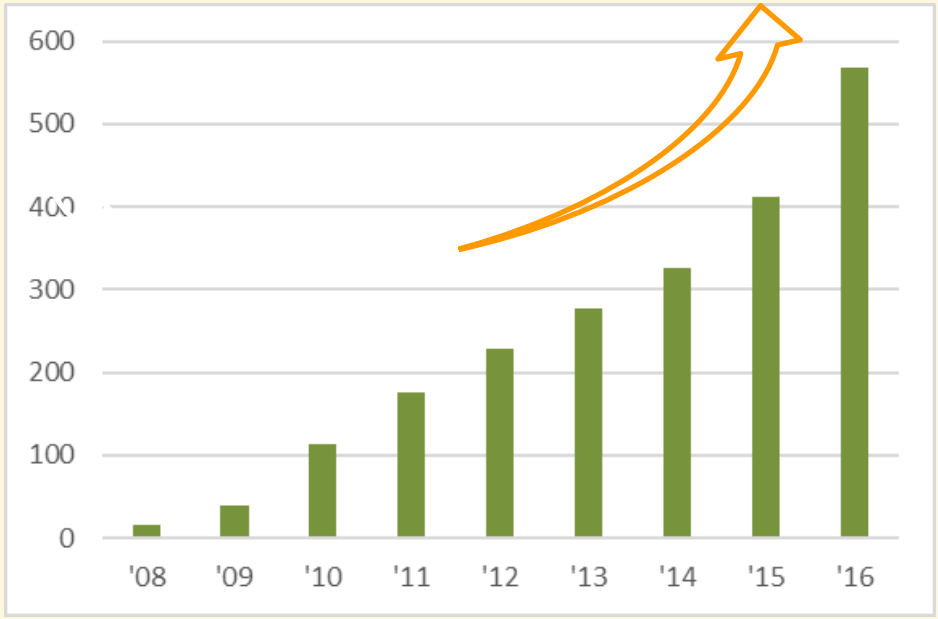


Number of User Labs.



Trend of Publications (～2016.3)

datasets by private companies only



(Cumulative)

Evaluation Issues

(1) How can we capture the users?
- Easy for contract-based distribution (users report once a year)
- **Difficult to trace the users for registration-only distribution**
- **How can we capture the users of OA data sets?**

(2) How can we capture the research results?
- Request users to report once a year
  (effective for contract-based distribution)
- Ask users to mention in acknowledgment. **But, how can we collect?**
- Better to be cited in the references. **But, what to cite?**
  - **Needs for Data DOI and Data Journal**

(3) How can we measure the value of each data set?
- Data citation may be an answer, but others?

(4) How can we evaluate the effectiveness of our activity?

Future Direction

(1) Sharing Data and Tools in Cloud-style Environment

- Datasets unable to distribute due to:
  - Huge data size
  - Personal information protection
  - High commercial value
  - Stream-type/Real-time data
- Conflict between protection and freedom
  - Data must be kept inside
  - Access to tools/resources outside (Internet or user site) is requested

→ Novel Cloud-type Data Sharing Research Platform

→ Evaluation as a Service (EaaS)

(2) Guidelines for Data Creators/Providers