SHARING DATA SETS AS RESEARCH RESOURCES

Center for Dataset Sharing and Collaborative Research (DSC)

National Institute of Informatics (NII) / SOKENDAI
Keizo Oyama

2015.10.21 第2回 SPARC Japan セミナー2015
Research and Real Application are getting closer in ICT

→ Needs as an Indispensable Research Resources for:

**Real and Large Scale Data generated by Real Services**

- Incentives for Private Companies to provide data for Academia
  - Social Contribution, Future Collaboration, Recruitment ...
- Open Access cannot be a solution for most of such data

*If each researcher tries to obtain such a data set ...*

**Users**: difficult to know contact, no guarantee for identity, huge cost of crawling, risk of infringement upon others’ right ...

**Owners**: made busy dealing with, hard to grasp/evaluate users and usage, suffer damage to business ...
Merits of Common Data Sets

- For each researcher:
  - Can ensure **reproducibility and transparency**
  - Easier to **compare** results with other research
  - Easy to **appeal** the research
- For research community:
  - Platform for Comparative Evaluation of Techniques
    - Setting common tasks, defining evaluation methods, accumulating research results, ...
  - Enhance **Community** and open up **Cross-Discipline Collaboration**
- For data provider:
  - Make the social contribution known to the public
  - Can appeal openness and fairness
For Promoting Research in Informatics ...

- Providing Data Sets
  - Collecting, Accepting, and Distributing
- Sharing Know-hows from Creation to Distribution
  - Various Know-hows are required for building / providing data sets specs, collecting, annotating, distribution method, licensing, user agreement ...
- Activating Research by Creating / Connecting Communities (data owners, creators, users)
  - Hosting Ideathons and Evaluation Workshops
  - Promoting Collaborative Research
Data Sets provided by IDR

- Yahoo! Data Set
- Rakuten Data Set
- Niconico Data Set
- Recruit Data Set
- Cookpad Data Set
- NTCIR Test Collection
- Speech Corpus
- Image Database of Japanese Classical Documents (planned)
- Two more Data Sets in preparation
Origins of Data Sets

- Real Data generated by Commercial Internet Services
  - Yahoo! Data Set
  - Rakuten Data Set
  - Niconico Data Set
  - Recruit Data Set
  - Cookpad Data Set
  - Two more Data Sets in preparation

- Research-purpose Data created by Researchers and Research Organizations
  - Speech Corpus
  - Image Database of Japanese Classical Documents (planned)

- Research-purpose Data created via Evaluation Workshop organized by NII
  - NTCIR Test Collection
Distribution Procedure

- **Contract-based (based on provider’s choice)**
  - Yahoo! Data Set Contract with NII
  - Rakuten Data Set Contract with Provider
  - Recruit Data Set Contract with NII
  - Cookpad Data Set Agreement & Approval by Provider
  - Speech Corpus Agreement & Approval by NII
  - NTCIR Test Collection Contract with NII
    (Web Archive Data, Task Data derived from Yahoo! Data Set)
  - *Two more Data Sets in preparation*

- **Registration Only**
  - Niconico Data Set Name (of any type) & optional e-mail, etc.
  - NTCIR Test Collection Name, Affiliation, e-mail, etc. → move to OA (?)

- **Open Access**
  - Image Database of Japanese Classical Documents (planned)
    CC BY-SA
Worries of data providers (especially for private companies)
- Copyright
- Privacy and Personal Information of its Service Users
- Flaming caused by Abuse
- Damage to Property Value
→ Controlling users and restricting usage are necessary in most cases.

Restrictions depend on the nature of data and company
- All prohibit:
  - providing data to third party; commercial use
  - disclosure of identified person/organization even in academic publication
- Some prohibit:
  - match data with information on the Internet
- Some require:
  - check the content of publication in advance
Tasks of data set sharing

Support for contract on build and provision

IDR
Acquisition - Cataloging - Distribution

NTCIR
Evaluation WS
Plan and build

IDR
Packaging

Universities
Projects

SRC
Processing

Private Companies

Researchers at NII

Researchers

Text Corpus
Speech Corpus
Video Corpus

Contact and support for users

Paper / Presentation

Plan and build

IDR

Paper / Presentation
Approaching to Data Providers

- Offering **Know-hows of Data Provision** to Data Owners and Producers (i.e. potential providers)
  - Reconsideration of Service Users’ Policy
    - copyright issues, ...
  - Procedure and Content of Contract with Data Users
- Proposing **Ideathon**, etc. for Preparation of Data Set Provision
  - Recruit
  - National Institute of Japanese Literature (国文研)
- Feed-back of Research Results
  - Grasping research and technology trends
  - Knowing active researchers
    → Collaborative Research, recruit of students
Sharing Research Results

- Research Meeting focusing on Data Set

  e.g., Rakuten R&D Symposium

Plan meeting gathering data owner and researcher

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

Ideathon held in advance of releasing Recruit Data

https://twitter.com/arg/status/440822789646217216
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


https://sites.google.com/site/ntcir11recipesearch/
Creating Communities

- Accumulating and Sharing Know-hows for Competitions

**BIGCHA** –
Big Data Programming Challenge using Common Data Sets
  - Yahoo!
  - Rakuten
  - Niconico
  - Recruit
  - Cookpad
  - ...

http://bigcha.net/
Connecting Communities

- Cross-border Collaboration — Informatics and ...
  - Cookpad Recipe Data
    Nutritional Science, Economics, Environmental Studies
  - Rakuten Travel Data
    Tourism Studies
  - Niconico Comment Data
    Musicology, Entertainment studies; Gap of Age

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?
     - Expectation for Data Journal

(3) How can we measure the value of each data set?

(4) How can we evaluate the effectiveness of our activity?
   We want to shift to OA, but ...
   
   Without evaluation, we cannot get funds for our activity.
Future Direction

(1) Sharing Data and Tools based on Cloud-style Environment
   - Data sets unable to distribute due to:
     - Huge data size
     - Personal information protection
     - High commercial value
   - Cloud-type Data Sharing Research Platform
   - Evaluation as a Service (EaaS)

(2) Clarifying Social and Technical Problems on Research Use of Privacy Sensitive Data

(3) Research on System Environment for Research using Deep Data based on Customized Licensing Scheme