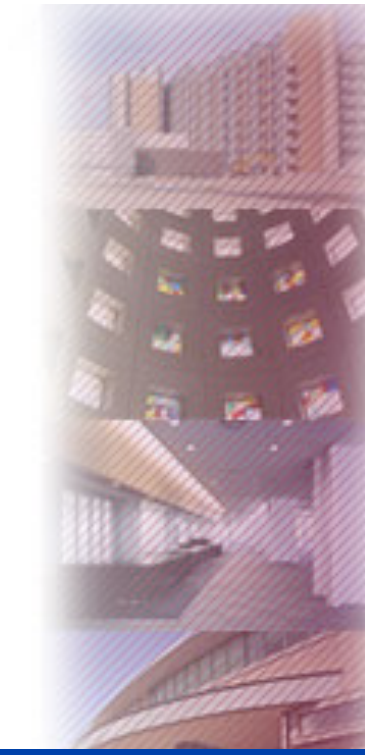
九州大学
KYUSHU UNIVERSITY



FREEDxDOM:

Toward Developing Foundation for Research Data on Cross Domains

Goal of This Talk: Introduction of Our New Project

Department of Informatics
Kyushu University, Japan
Daisuke Ikeda
daisuke@inf.kyushu-u.ac.jp

Workshop for promotion of"Open Science Data"
@Kyoto Univ. (2015.09.17--18)

My Research Interests: Data

Mining from Data

- Web mining, text mining, data mining
- Bioinformatics
- Mining from Research Data, e-Science

Data Infrastructure

- Database
- Information retrieval
- Automatic identification (e.g., bar codes, RFID, ICcards), security, ACL
- *Scholarly communication*, e.g, institutional repositories

Outline of the Project

KAKENHI (Grants-in-Aid for Scientific Research)

- Category: Scientific Research (B)
- Field: Library and information science
- Year: April 1, 2015〜March 31, 2019 (estimated)
- Project Number: 15H02787
- URL (KAKENHI): `https://kaken.nii.ac.jp/d/p/15H02787.en.html`

Members

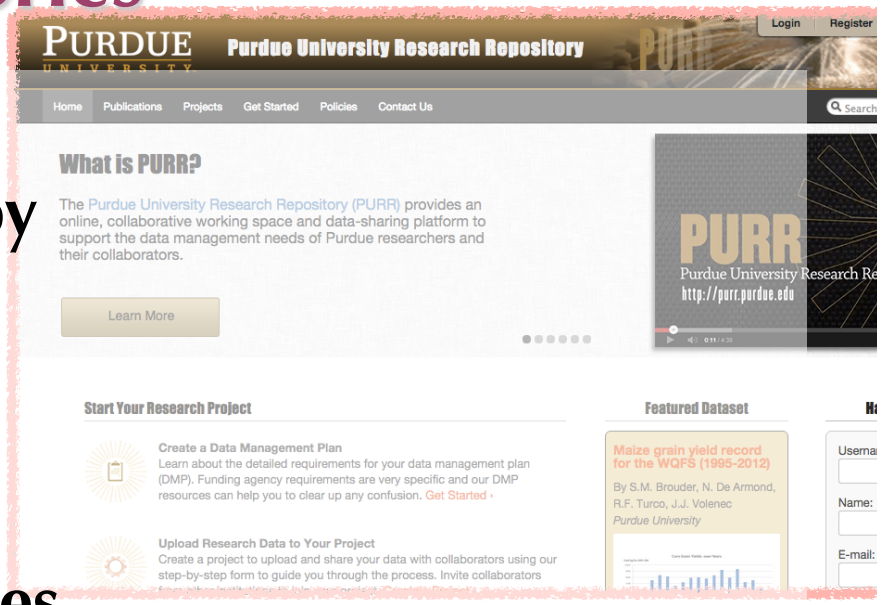
- PI: Ikeda, Daisuke (Kyushu Univ.)
- Three Core Co-Investigators
  Araki, Hiroyuki (Tokushima Univ.)
  Imai, Koji (JAXA)
  Koyama, Yukinobu (Research Organization of Information and Systems)
- Fields of four other co-investigators include
  Library and information science, Astronomy, Earth and planetary science, Bioinformatics.

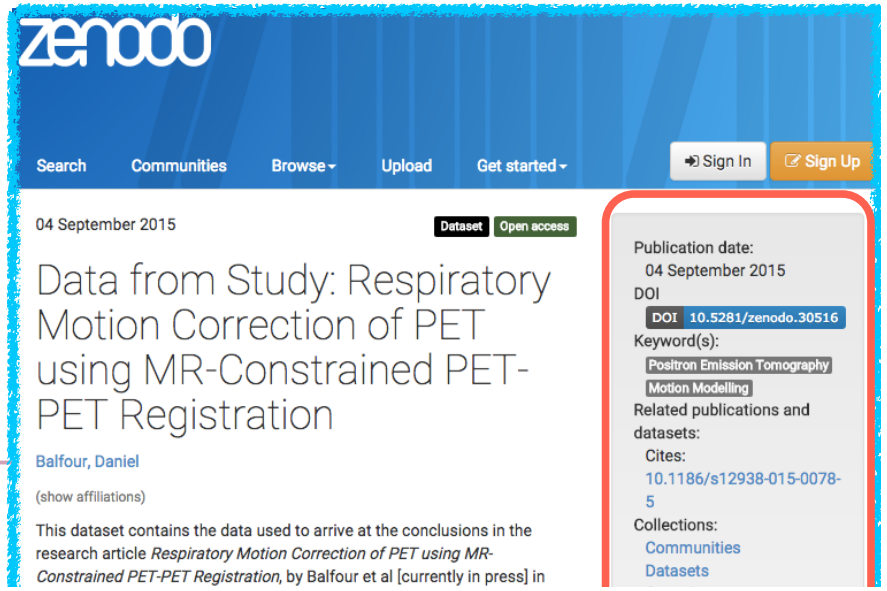
Background: Data Repositories

Two types:

1. **General type, usually hosted by libraries**
   - **PURR (Purdue univ.), zendoo (CERN)**

2. **Discipline Specific, usually hosted by research communities**
   - **detailed metadata**





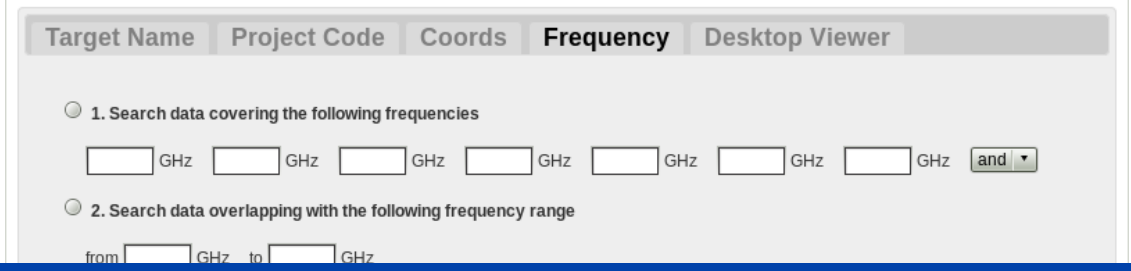
Motivation: Users in Other Disciplines

e-Science researches

- I have used data in different disciplines:
  - e.g., sequence data of bioinformatics, geomagnetic data

Researches for global issues

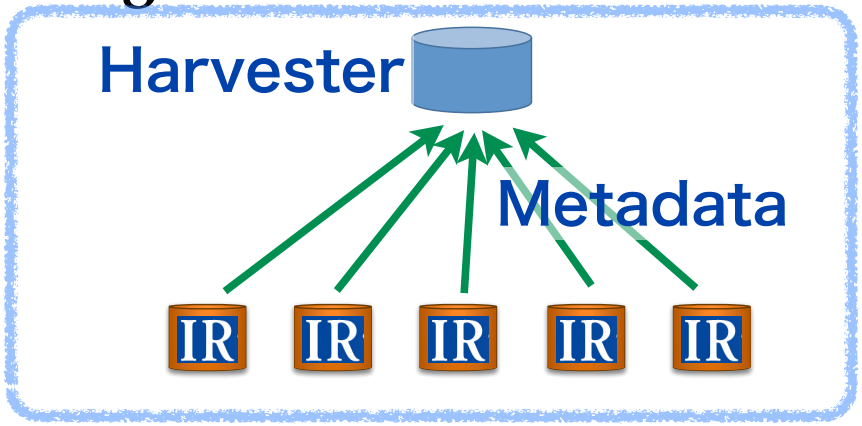
- e.g., GEOSS (Global Earth Observation System of Systems) and Future Earth
  - In these initiatives, inter- & trans-disciplinary are principle keywords: co-designing and co-producing of policy-makers, funders, academics, business and industry, and other sectors of civil society



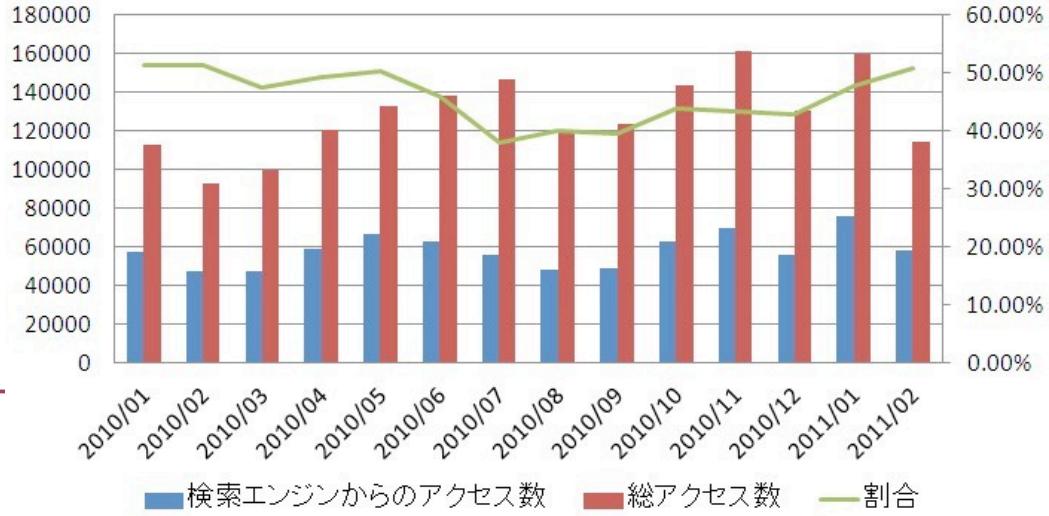
It is difficult for users in other disciplines to find appropriate data.

Access from General Public to IR (1/2)

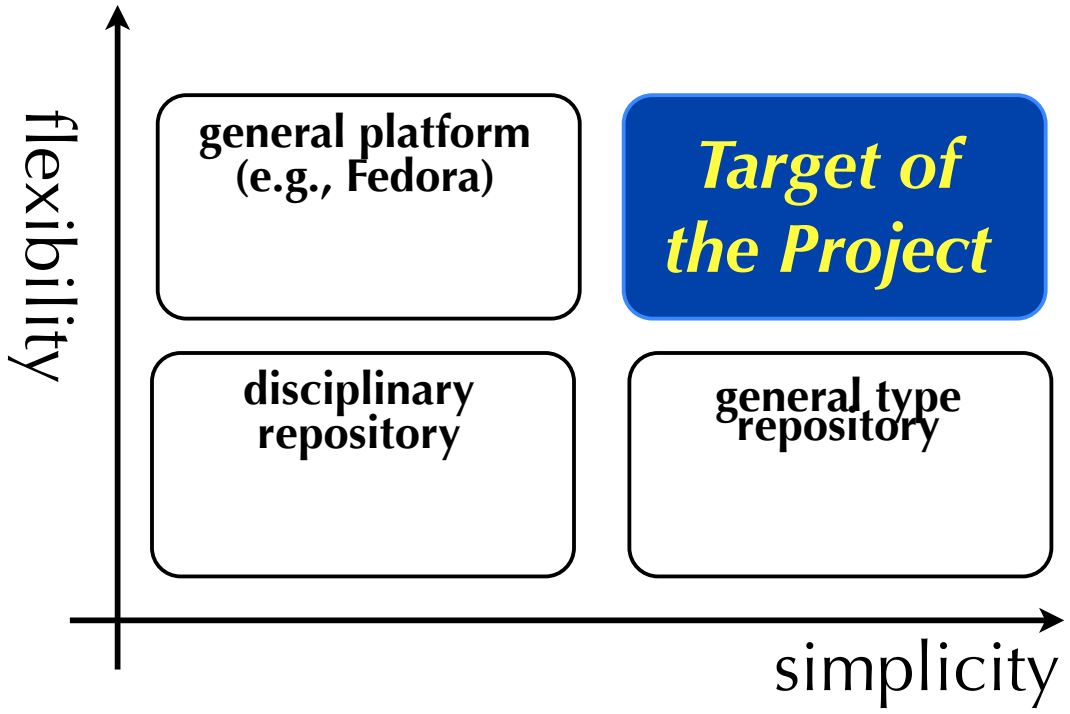
Original search model: Harvester + Metadata



Practical model: Search Engine + Index from Contents



Goal: Infrastructure for Scholarly Communication



To develop a general infrastructure of data repositories

Lesson Learned from Institutional Repositories

Original search model: Harvester + Metadata

- search terms are restricted to metadata, such as titles and author names.

→Practical model: Search Engine + Index from Contents

- users can search any terms included in academic thesis at popular search engines.

Rich "Metadata" and Simple Interface

Hypothesis: Vector Representation of Data

item (data) = vector of words

- **data1 = [“aurora”, “substorm”, “geomagnetic”, ...]**

Direct consequence:

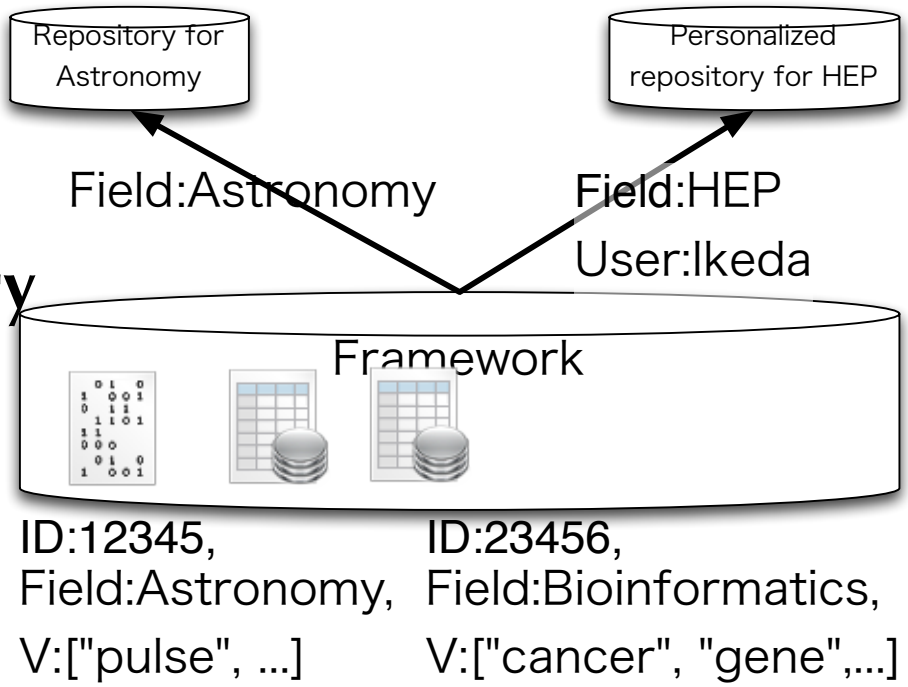
1. users can *search* data.

2. FW for different disciplinary repositories

- **like hashtags in microblogs**

3. Access Control

4. Version Control



Comparison with Data Identity

Each DR is expressed by a search query, such as, “Field: xyz”.

- using a query to specify a subset of data sets, we can dynamically cite the data.
  c.f., Dynamic Citation WG of RDA

ID for data

- extensional approach, that is, we can enumerate each items.
  - e.g., the set of odd numbers are extensionally given as $\{1, 3, 5, \dots\}$
- our approach is intensional.
  - e.g., the set of odd numbers are intensionally given as $\{2n + 1\}$

Scope of the Project: Search and QuickLook

Before: to check a data, we need to know details of the data.

After: easy-to-check

- **Still, we need to know details of the data and related tools if you want to use it.**

Verify the Hypothesis: Search and QucikLook

Technical Background: Vector Space Model

Salton '70

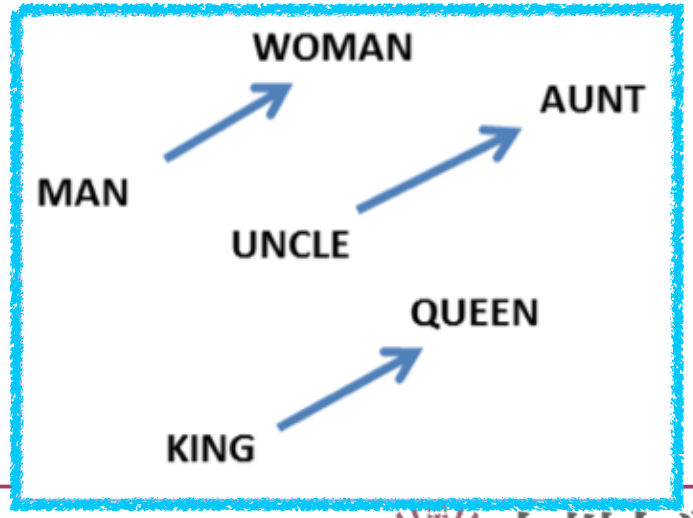
Basic idea: a dimension <=> a (indexed) word

A document: a vector of weights, such as TF/IDF, of words

- **TF/IDF gives a large weight to a word if it appears frequently in the document and it does not prevail in the whole data set.**

Similarly, we can consider a word vector and a context vector [Hosoi et al. '14].

- **we can calculate words or contexts.**
  - **e.g., 'king' - 'man' + 'woman' = 'queen' (from word2vec)**



Technical Background: Vectors for Data

Idea: use data journals to assign words to a data set

- in a data journal, a data set has an identification, such as DOI.
- we can obtain related words to the data from corresponding description.

Future work:

- expand words list for general public.

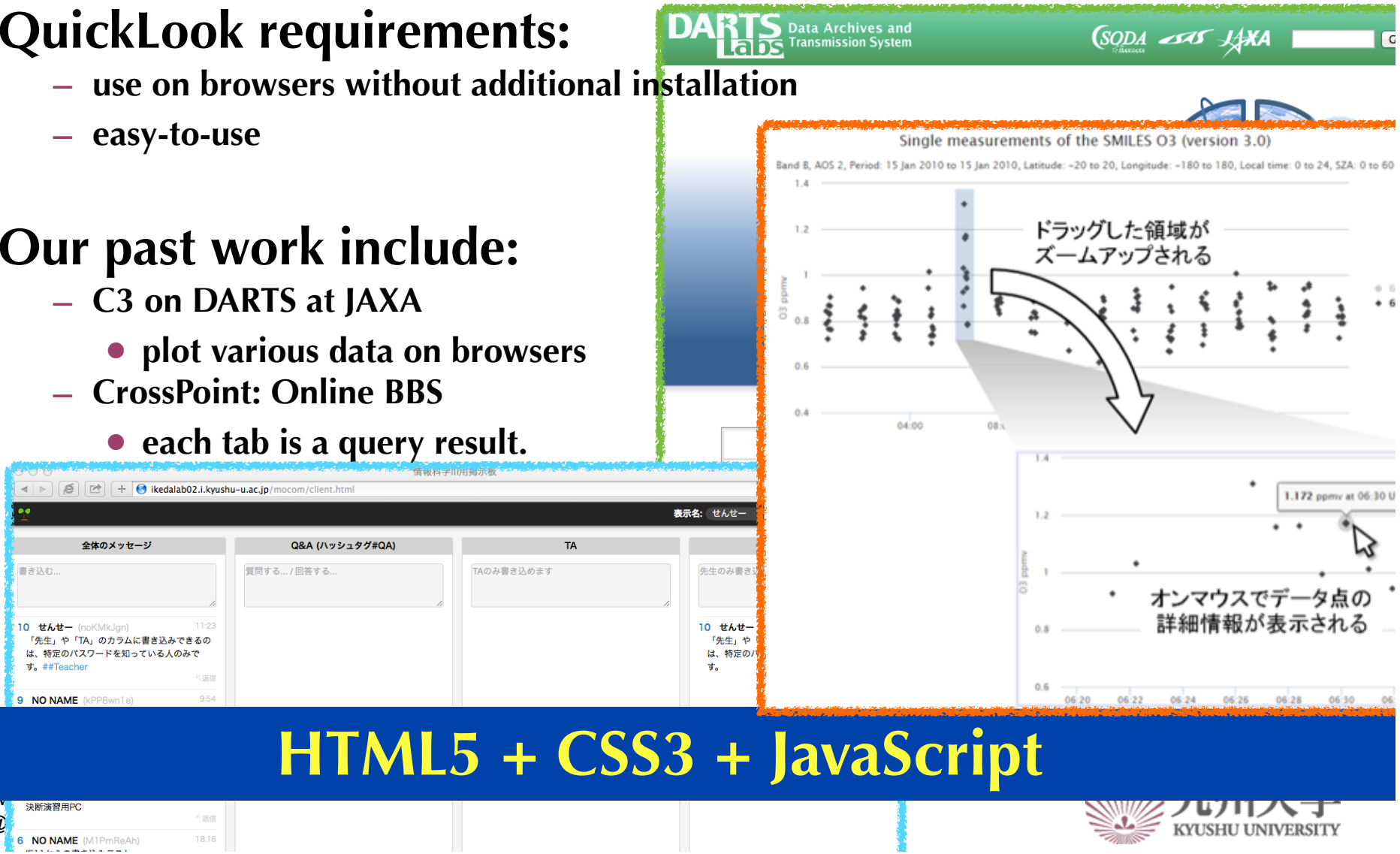
Technical Background: User Interface

QuickLook requirements:

- use on browsers without additional installation
- easy-to-use

Our past work include:

- C3 on DARTS at JAXA
  - plot various data on browsers
- CrossPoint: Online BBS
  - each tab is a query result.



HTML5 + CSS3 + JavaScript

KYUSHU UNIVERSITY

Conclusion

FREEDxDOM: four-year-project from this April.

Aim: General framework for data repositories

Ideas: Searchable data (via vectors) and QuickLook