RCOS

国立情報学研究所
National Institute of Informatics

Overview of a Repository System with a Focus on the WEKO3 (Invenio) System

Masaharu Hayashi

Research Center for Open Science and Data Platform (RCOS),

National Institute of Informatics(NII), JAPAN

Eko-Konnect Repository Workshop and eduID Policy meeting

27 Jan 2020

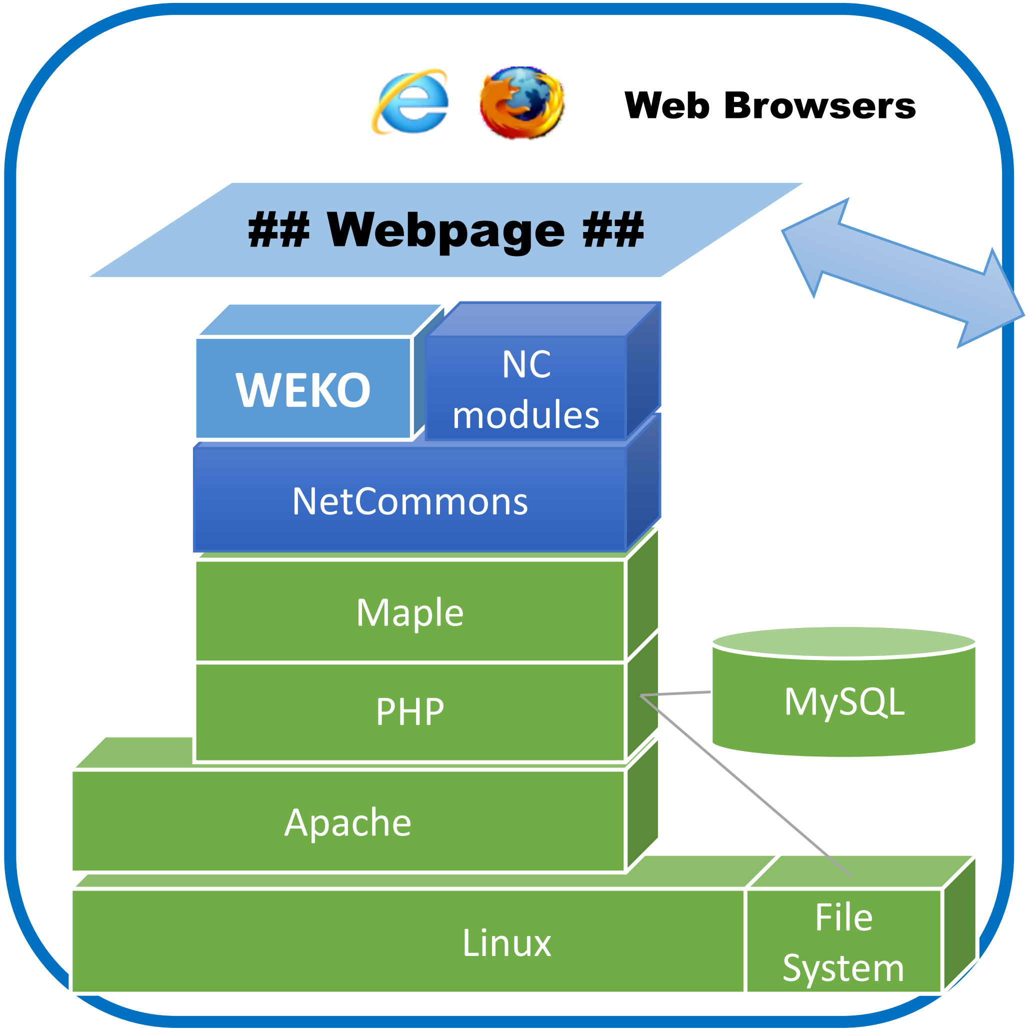
Table of Contents

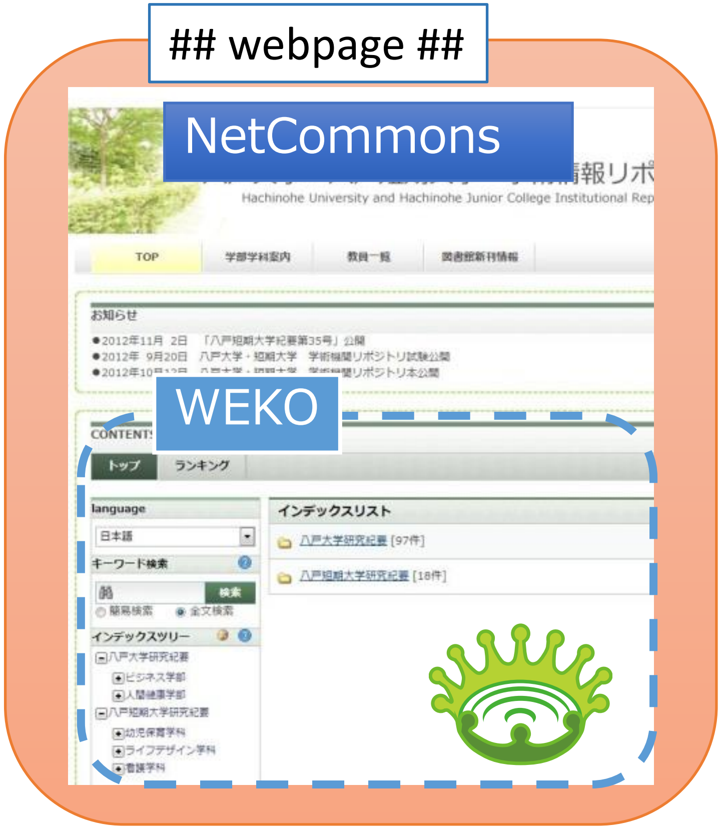
- **Introduction of WEKO2**
- JAIRO Cloud and WEKO2
- Issues of JAIRO Cloud/WEKO2
- Comparison of Opensource Repository Software
- Introduction of WEKO3
- Update of NGR Implementation

What is WEKO for Subtext message in Logo

WEKO means repository in Swahili. Once a researcher deposit his/her contents in WEKO, a circle of researcher have been made around the content.

System Architecture





…情報リポ
Hachinohe University and Hachinohe Junior College Institutional Rep

TOP | 学部学科案内 | 教員一覧 | 図書館新刊情報

お知らせ

- 2012年11月 2日 「八戸短期大学紀要第35号」公開
- 2012年 9月20日 八戸大学・短期大学 学術機関リポジトリ試験公開
- 2012年10月12日 八戸大学・短期大学 学術機関リポジトリ本公開

CONTENTS

トップ | ランキング

language

日本語

キーワード検索

検索

○ 簡易検索 ● 全文検索

インデックスツリー

- ⊟八戸大学研究紀要
  - ⊞ビジネス学部
  - ⊞人間健康学部
- ⊟八戸短期大学研究紀要
  - ⊞幼児保育学科
  - ⊞ライフデザイン学科
  - ⊞看護学科

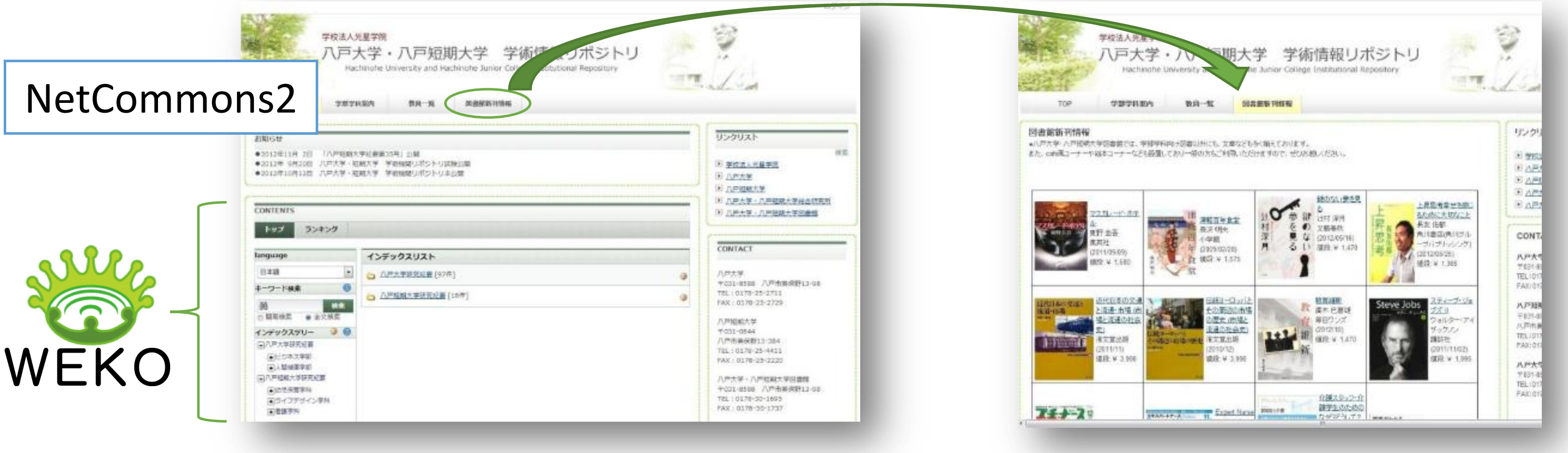
インデックスリスト

- 八戸大学研究紀要 [97件]
- 八戸短期大学研究紀要 [18件]

Linux, **A**pache, **M**ySQL, **P**HP = LAMP

Feature of WEKO

- High Functionality
  - It has almost all functionalities you need as a repository system.
- Easy
  - All functions can be customized and operated by browser.
- As you like
  - Not only the repository function but also variety of add-on can be utilized for designing your own web page.



NetCommons2

WEKO

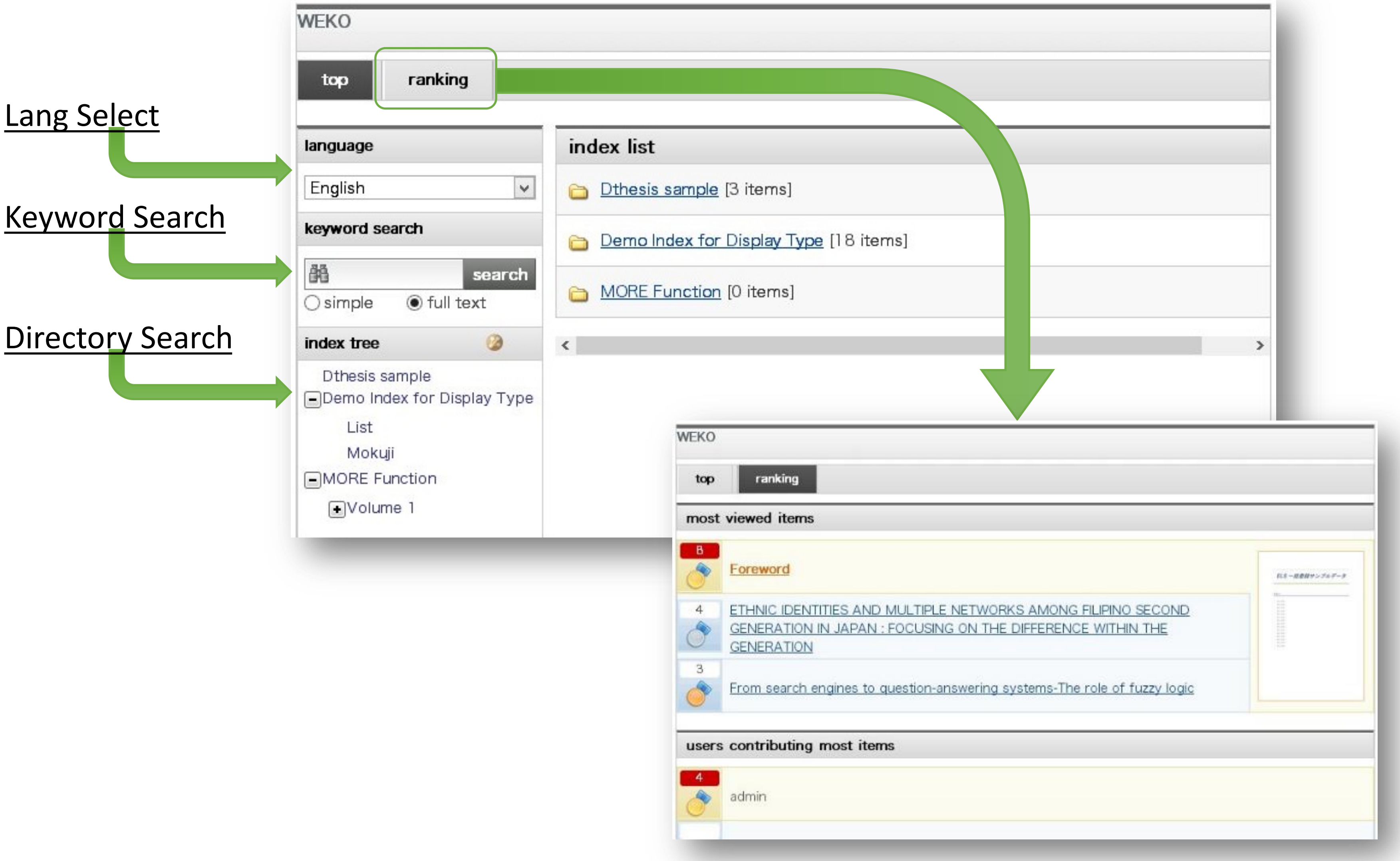
学校法人光星学院
八戸大学・八戸短期大学　学術情報リポジトリ
Hachinohe University and Hachinohe Junior College Institutional Repository
リンクリスト
インデックスツリー
図書館新刊情報
Steve Jobs

Summary of Functions

- Guest
  - Directory Search
  - Keyword Search
    - Metadata
    - Full Text
  - Ranking
- Registered User
  - Item Registration
  - Workflow

- Moderator
  - Item Type Mgmt
  - Item Mgmt
  - Index Tree Mgmt
  - Review System
  - Import
  - Log Analysis
  - WEKO Custom

WEKO: Guest User Functions



Lang Select

Keyword Search

Directory Search

WEKO

language

English

keyword search

○ simple ● full text

index tree

Dthesis sample
⊟Demo Index for Display Type
List
Mokuji
⊟MORE Function
⊞Volume 1

index list

Dthesis sample [3 items]

Demo Index for Display Type [18 items]

MORE Function [0 items]

WEKO

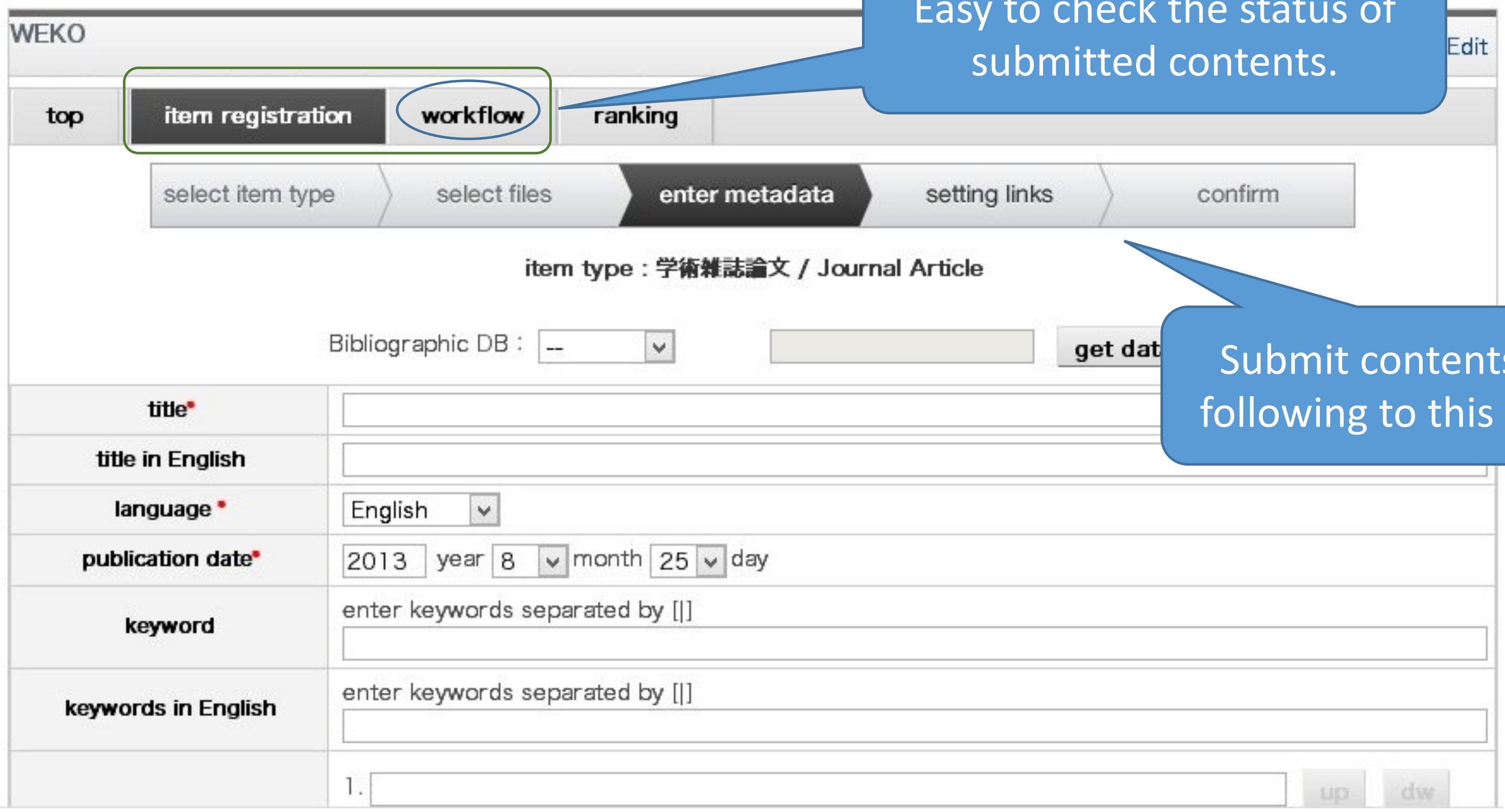
most viewed items

8	Foreword
4	ETHNIC IDENTITIES AND MULTIPLE NETWORKS AMONG FILIPINO SECOND GENERATION IN JAPAN : FOCUSING ON THE DIFFERENCE WITHIN THE GENERATION
3	From search engines to question-answering systems-The role of fuzzy logic

users contributing most items

4	admin

WEKO: Login User Functions



item type : 学術雑誌論文 / Journal Article

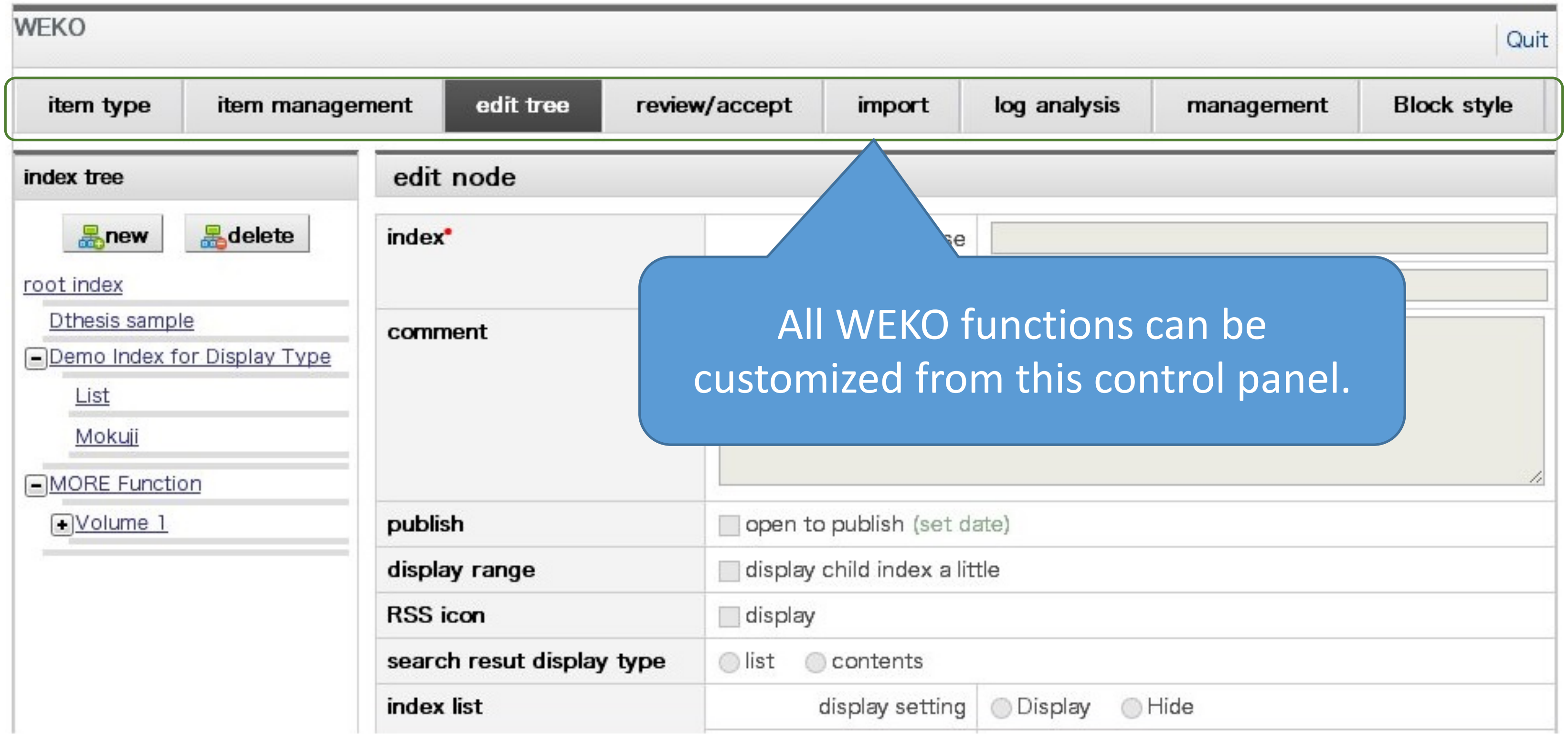
Bibliographic DB : --

title*	
title in English	
language *	English
publication date*	2013 year 8 month 25 day
keyword	enter keywords separated by [|]
keywords in English	enter keywords separated by [|]
	1.

Easy to check the status of submitted contents.

Submit contents by following to this flow.

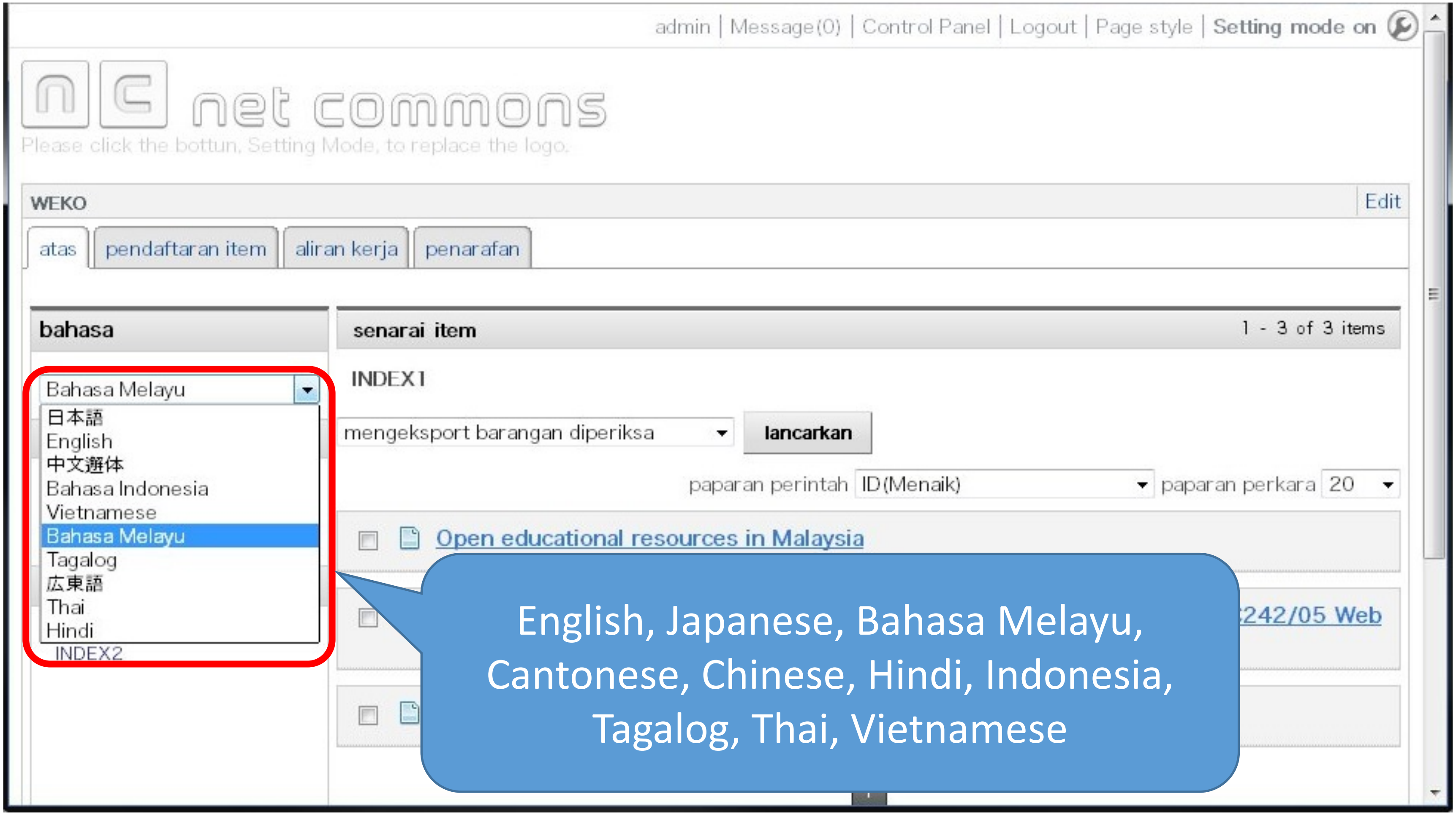
WEKO: Admin Functions



index*	
comment	
publish	□ open to publish (set date)
display range	□ display child index a little
RSS icon	□ display
search resut display type	○ list ○ contents
index list	display setting ○ Display ○ Hide

All WEKO functions can be customized from this control panel.

WEKO: Multilingual Functionality



admin | Message(0) | Control Panel | Logout | Page style | Setting mode on

net commons

Please click the bottun, Setting Mode, to replace the logo.

WEKO

Edit

atas | pendaftaran item | aliran kerja | penarafan

bahasa

Bahasa Melayu

- 日本語
- English
- 中文鐗体
- Bahasa Indonesia
- Vietnamese
- Bahasa Melayu
- Tagalog
- 広東語
- Thai
- Hindi

INDEX2

senarai item

1 - 3 of 3 items

INDEX1

mengeksport barangan diperiksa | lancarkan

paparan perintah ID(Menaik) | paparan perkara 20

□ Open educational resources in Malaysia

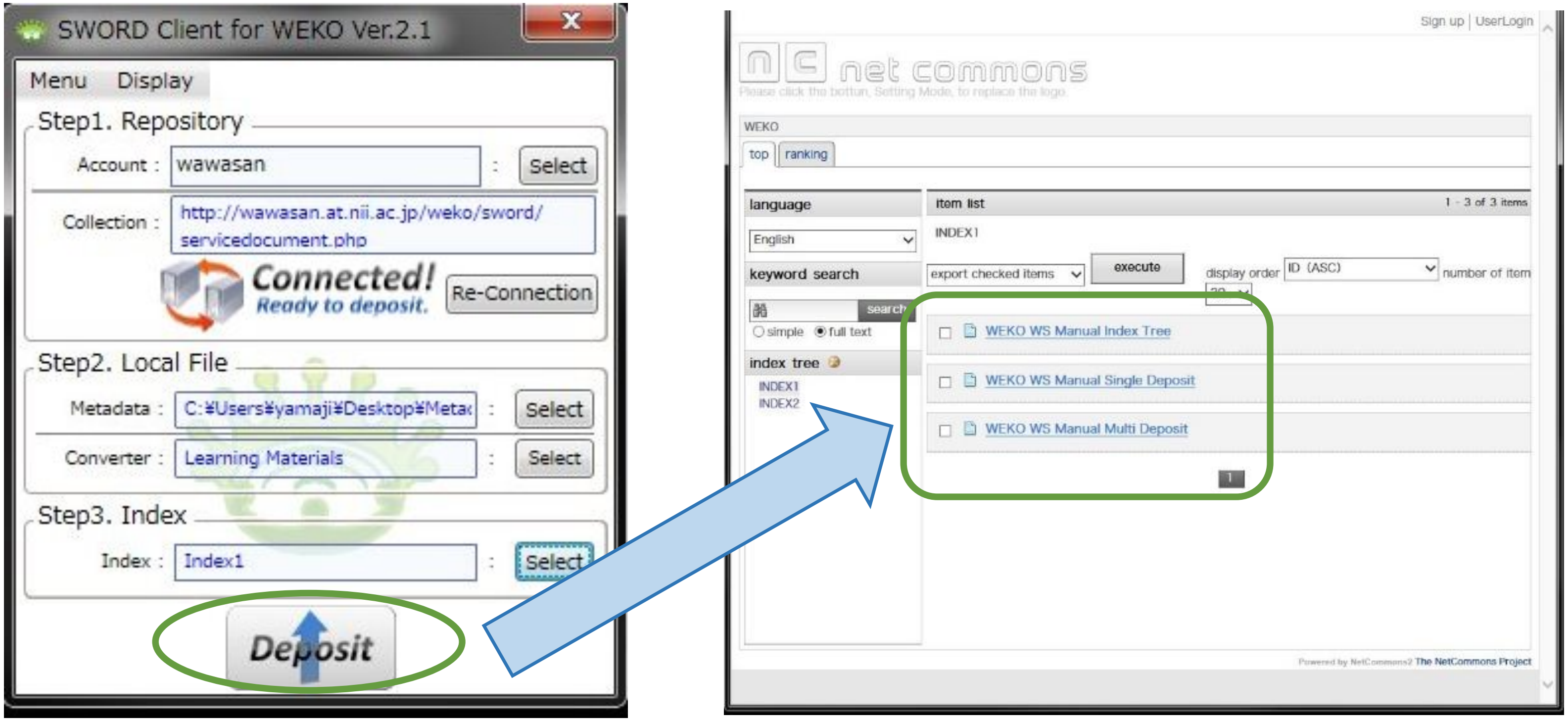
□ :242/05 Web

□

English, Japanese, Bahasa Melayu, Cantonese, Chinese, Hindi, Indonesia, Tagalog, Thai, Vietnamese

WEKO: Easy to Multiple Deposit

SWORD Client for WEKO



SWORD Client for WEKO Ver.2.1

Step1. Repository

Account : wawasan

Collection : http://wawasan.at.nii.ac.jp/weko/sword/servicedocument.php

Connected!
Ready to deposit.

Step2. Local File

Metadata : C:¥Users¥yamaji¥Desktop¥Meta(

Converter : Learning Materials

Step3. Index

Index : Index1

WEKO

language: English

keyword search: ○ simple ● full text

index tree
INDEX1
INDEX2

item list — 1 - 3 of 3 items

INDEX1

export checked items — execute — display order: ID (ASC) — number of item

□ WEKO WS Manual Index Tree

□ WEKO WS Manual Single Deposit

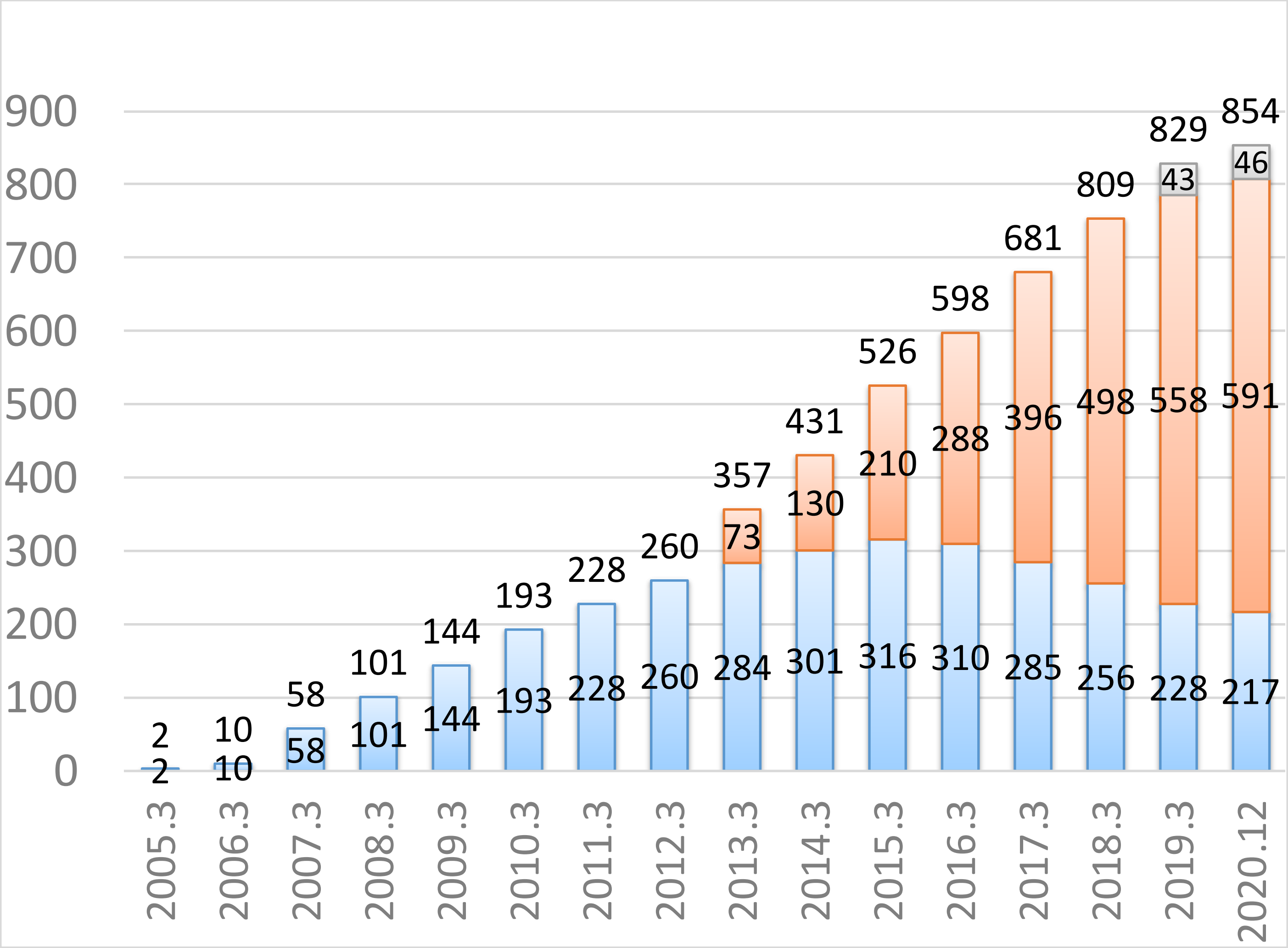
□ WEKO WS Manual Multi Deposit

Powered by NetCommons2 The NetCommons Project

Table of Contents

- Introduction of WEKO2
- **JAIRO Cloud and WEKO2**
- Issues of JAIRO Cloud/WEKO2
- Comparison of opensource repository software
- Introduction of WEKO3
- Update of NGR implementation

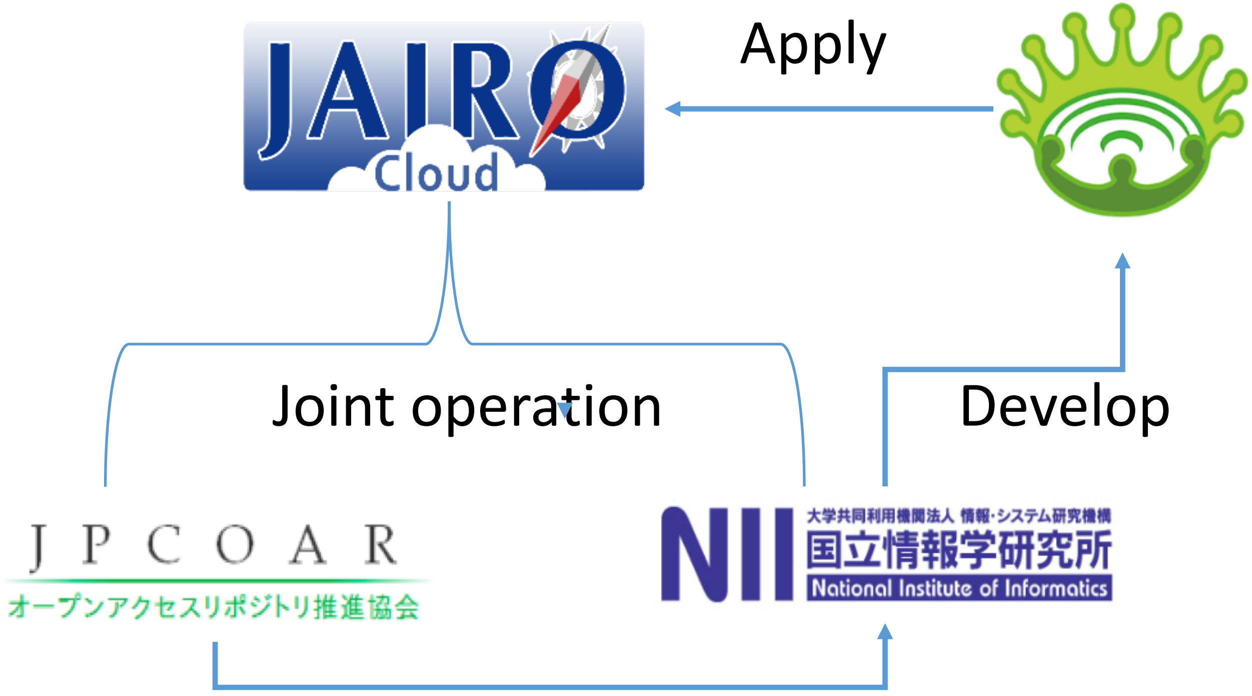
JAIRO Cloud: Institutional Repository Cloud Service



JC hosted IRs of 637 institutions as of Dec. 2019

JAIRO Cloud Operation

Joint Operation of NII and JPCOAR



JAIRO Cloud
JPCOAR
オープンアクセスリポジトリ推進協会
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

Apply

Joint operation

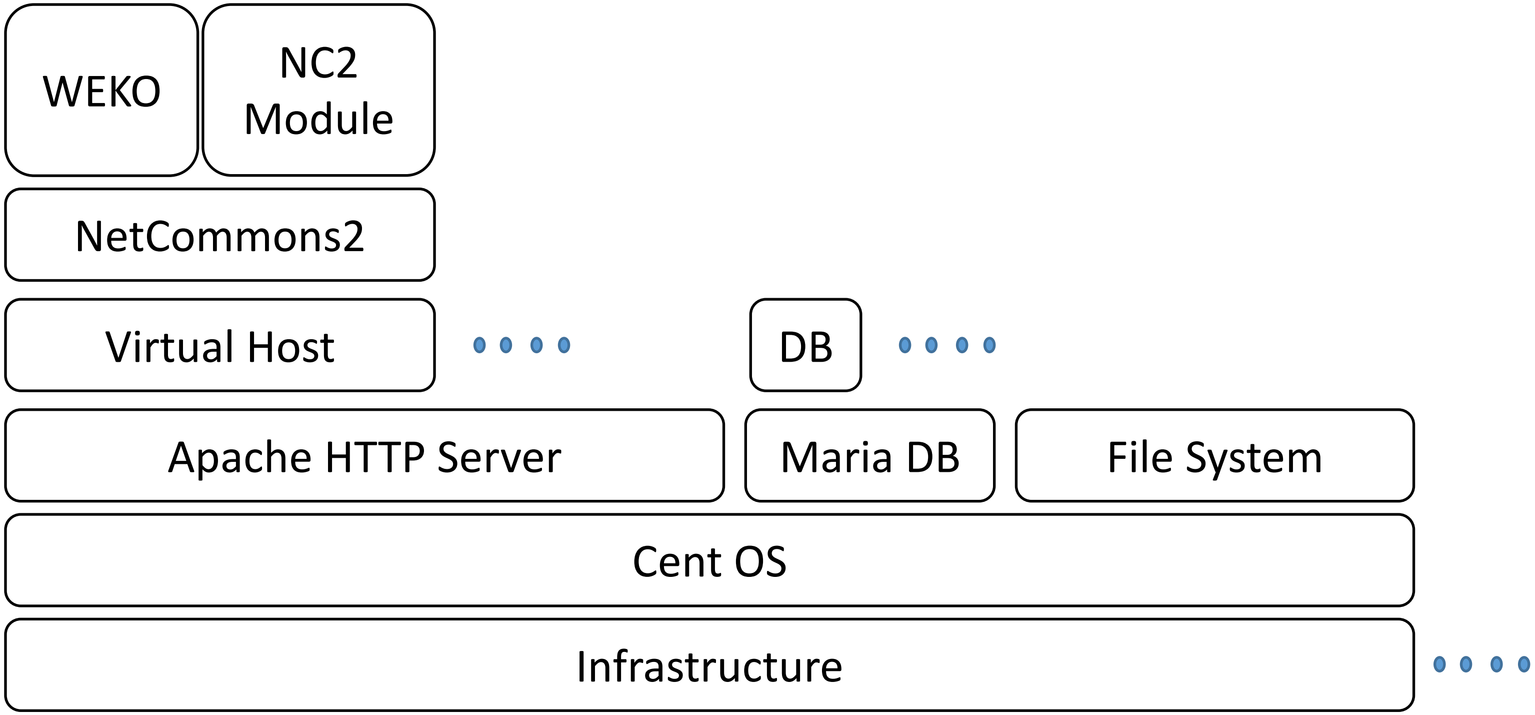
Develop

Assist to develop as user community

JC is operated with service fee of users

JAIRO Cloud System Architecture

JAIRO Cloud is shared repository service



WEKO

NC2 Module

NetCommons2

Virtual Host ・・・・

DB ・・・・

Apache HTTP Server

Maria DB

File System

Cent OS

Infrastructure ・・・・

Components of repository are logically separated for each repository to ensure security

Number of Repositories Per a Server Instance

JAIRO Cloud will save resources by a sharing server

Repository Size Type	Average	Max	Min
	(number of repositories)		
small	18.4	27	1
large	1.8	2	1

In case of large repository type
it occupy almost single instance whereas
small type lives with repositories
(average 18.4 repositories).

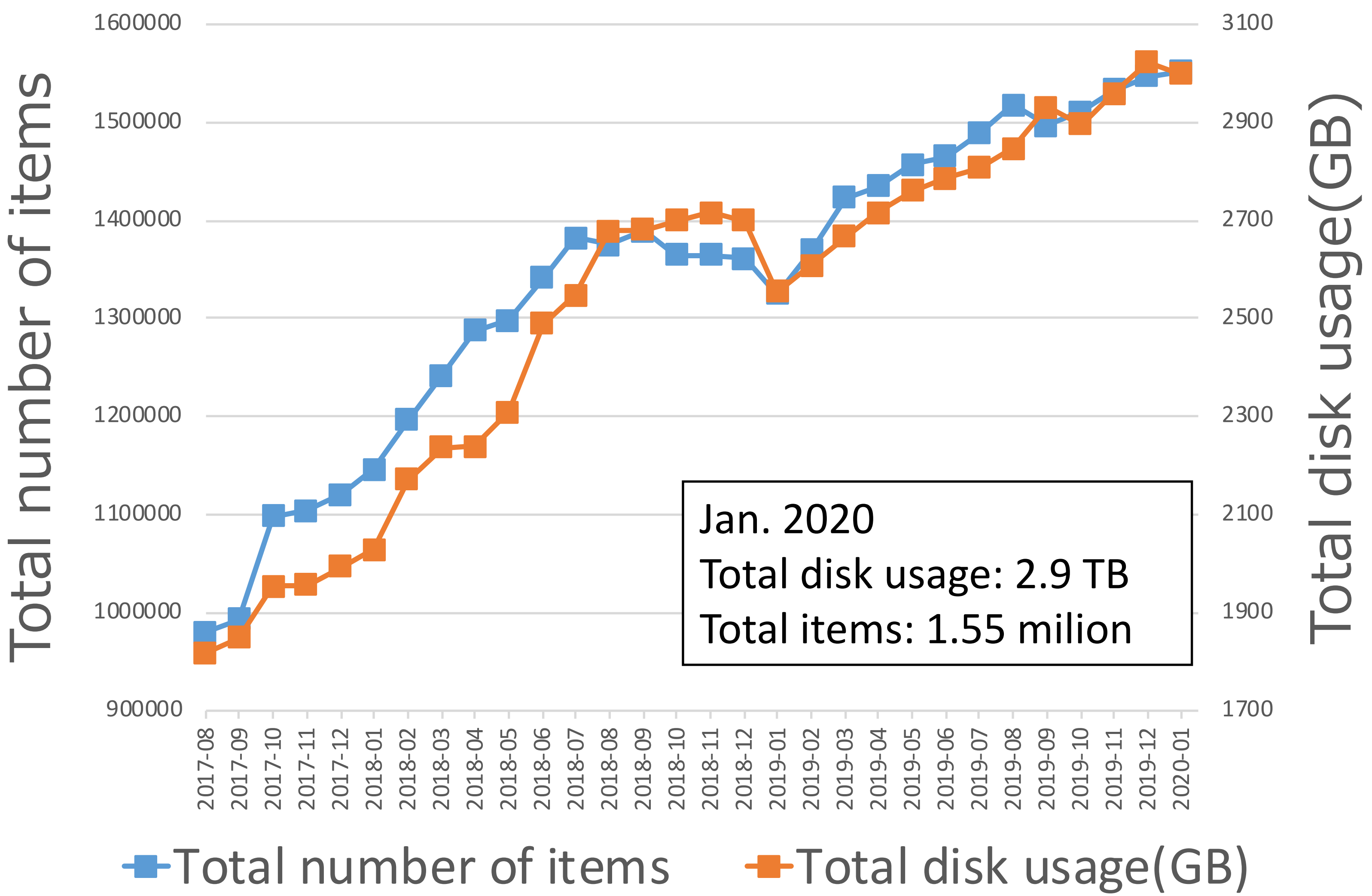
Specification of Server instances

JAIRO Cloud is composed of the following servers.

Role of Instances	CPU	Mem	SDD	HDD	Number of Instances
Shibboleth IdP	1	4GB	20GB		2
Monitoring	4	4GB	30GB		2
WEKO(Small)	6	16GB	20GB	250GB	31
WEKO(Large)	6	16GB	20GB	500GB	5
PID server	8	8GB	20GB		2

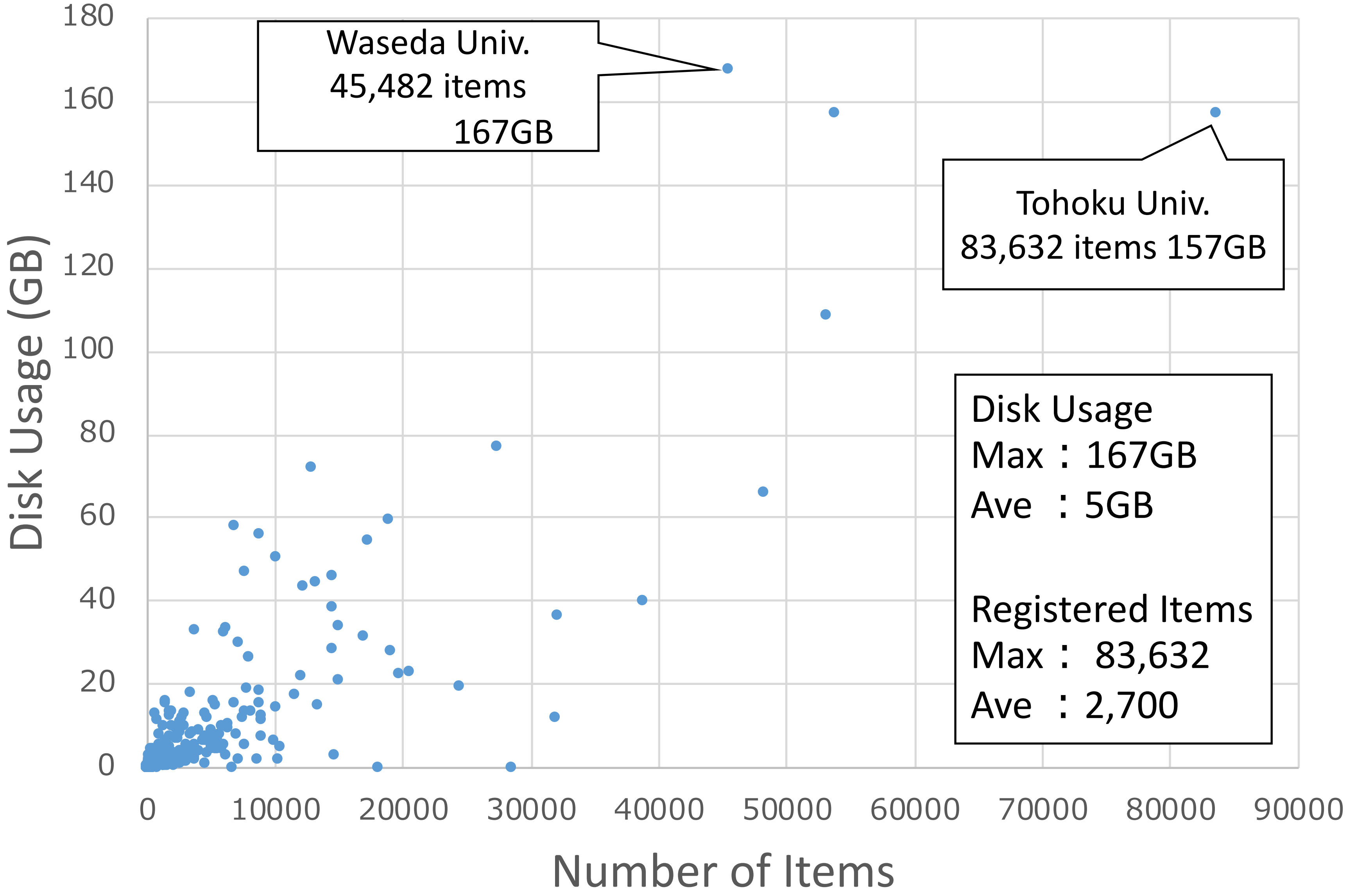
JAIRO Cloud is operated on 42 servers
for providing the service

Growth registered items and usage disk size



Jan. 2020
Total disk usage: 2.9 TB
Total items: 1.55 milion

Usage of JAIRO Cloud per IR



Waseda Univ.
45,482 items
167GB

Tohoku Univ.
83,632 items 157GB

Disk Usage
Max：167GB
Ave ：5GB

Registered Items
Max： 83,632
Ave ：2,700

How have functions of WEKO2 been developed?

NII and JPCOAR cooperate to develop functions.

1. JPCOAR makes a request list from users needs
2. JPCOAR set priority the request list
3. JPCOAR submits the request list to NII
4. NII decides functions to develop based on the JPCOAR's request list with NII’s strategy
5. NII provides feedback on the functions that have developed to JPCOAR
6. JPCOAR tests the functions and feedback the result to NII

JPCOAR オープンアクセスリポジトリ推進協会 ⟺ 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

JPCOAR Strategy from fiscal 2019 to 2021

1. To contribute to the promotion of **open science,** engage in leading initiatives on the public access to and circulation of **research data**
2. Establish an infrastructure for the circulation of scientific information that advances open access, and promote the distribution and use of contents
3. Strengthen functions as a community supporting open access repositories
4. Foster human resources able to advance open access and **open science**
5. Strengthen JPCOAR’s operational infrastructure and enhance its brand equity

JPCOAR's strategy moves from OA to OS.
JAIRO Cloud needs to respond to the strategy.

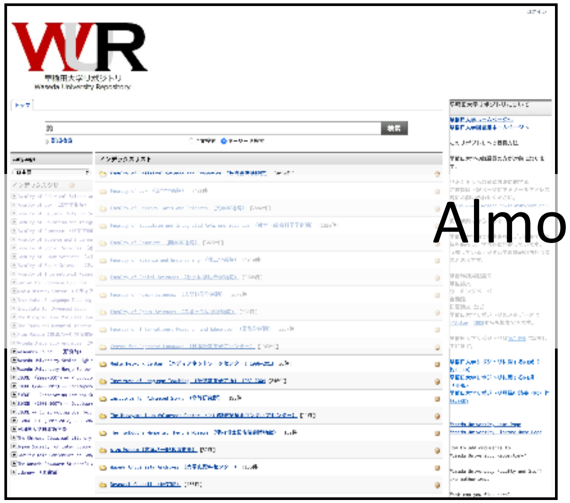
Contents

- Introduction of WEKO2
- JAIRO Cloud and WEKO2
- **Issues of JAIRO Cloud/WEKO2**
- Comparison of opensource repository software
- Introduction of WEKO3
- Update of NGR implementation

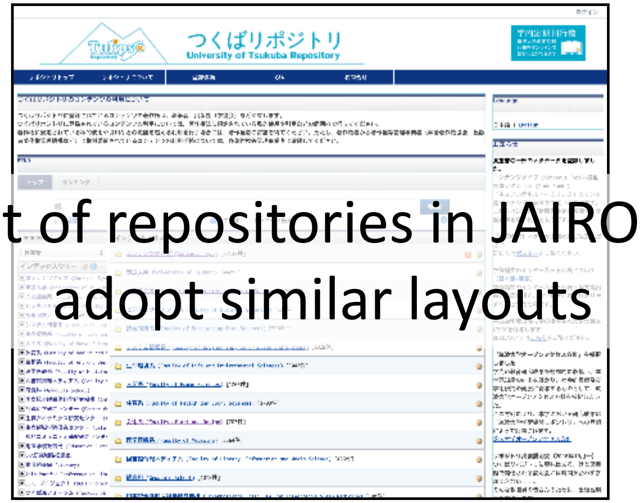
Necessity of Rich CMS Function

CMS integration was expected to promote community activities by using forum modules, FAQ modules, messaging modules etc.

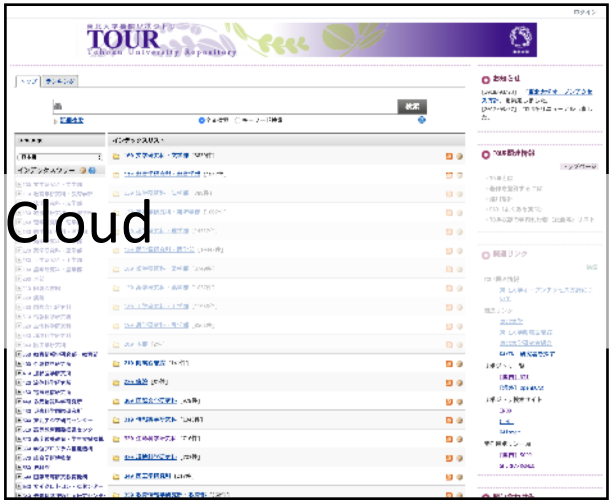
However, frequently used function was the site design funcion.



Waseda Univ.



Tsukuba Univ.



Tohoku Univ.

Almost of repositories in JAIRO Cloud adopt similar layouts

Site design function is important than other functions of supporting community activity on CMS

Issues of repository function extensibility

- **Expressiveness of metadata**
  - WEKO2 has a flexible metadata editor, but it does not support hierarchical metadata.
    → A flexible metadata editor that supports hierarchical metadata is required.
- **Function extendibility**
  - WEKO2 supports multimedia file and multiple OAI-PMH metadata schema. However, they are not extensible.
    → For supporting various use cases, pluggable file previewer, extendibility of OAI-PMH provider function and customizable workflow function are required
- **Function scalability**
  - WEKO2 is not designed for large-scale data processing. As a result, there are problems with faster search functions and batch processing speed.
    →A function design that realizes scalability is required.

Issues of JAIRO Cloud Operation

- **Effective use of resources**
  - JAIRO Cloud is a shared repository service. Multiple repository share CPU, memory and disk in a server. However, it cannot be shared between servers.
    →It is necessary to design a system that shares computer resources as much as possible and uses them effectively.
- **Scalability**
  - A repository is a kind of archive system. The metadata and files stored there continue to increase.
    →System design and operation that can cope with the increase in data volume and metadata is required.
- **Safe Deploy**
  - Function development can be centralized by sharing the repository as infrastructure. However, developments can affect the entire cloud server.
    →System design and operation for realizing function development and secure deployment while continuing operation are required.

Minimum Requirement for Supporting Research Data

- **Versioning**
  - Research data change dynamically during research cycle.

  →Versioning management function is required to track the changes of research data.
- **Metadata customization**
  - Each domain has a different set of metadata.
    → flexible metadata editor is required.
- **Flexible workflow**
  - Different laboratory have different research data archiving workflow .
    →Customizable workflow function is required for supporting laboratory operation.

How to solve the problems and to develop new functions

	Pros	Cons
Homegrown software	• Can respond freely to user needs • Can control everything in development of software	• have to develop everything • Difficult to catch up with the new technology
Development using existing software	• Easy to develop than scratch development • Joint development of common functions is possible	• Difficult to develop functions freely • Requires communication with developers

In order to consider the implementation policy of next version WEKO, we compared existing open source repository software.

Table of Contents

- Introduction of WEKO2
- JAIRO Cloud and WEKO2
- Issues of JAIRO Cloud/WEKO2
- **Comparison of opensource repository software**
- Introduction of WEKO3
- Update of NGR implementation

History of OSS Repository Development

Dawn of repository software

- 2000 eprints repository software
- 2002 DSPACE
- CDSware(INVENIO)
- 2003 Fedora™

OAI-PMH 1.0 (2001)

OAI-PMH 2.0 (2002)

Growth of repository software

- 2006 Islandora
- dvn(The Dataverse Project)
- 2008 Repository Module on NC2 WEKO
- 2009 hydra (samvera)
- 2010 hubzero

SWORD 1.0 (2007)

SWORD 2.0 (2012)

ResourceSync (2014)

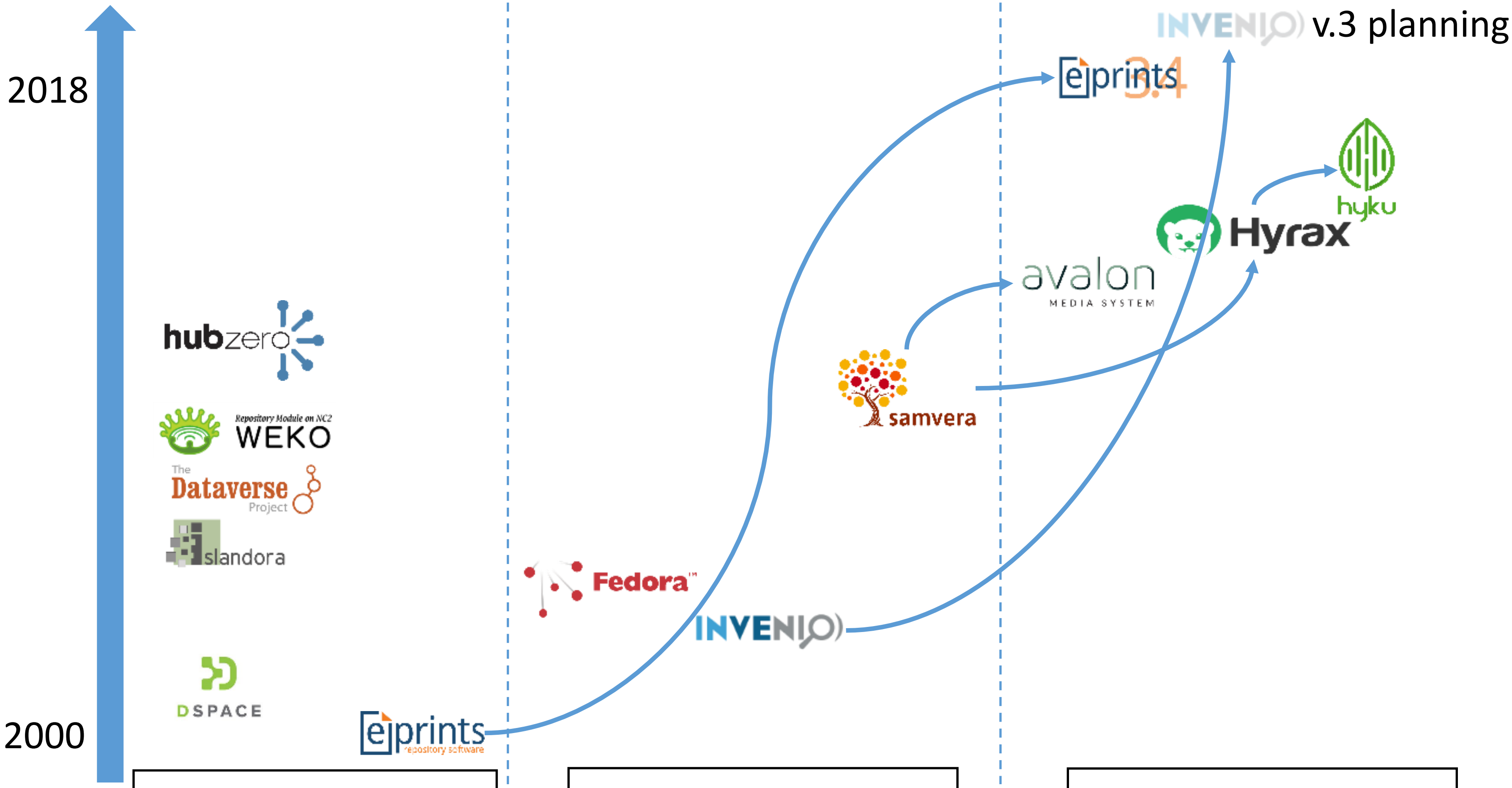
New age of repository software

- 2016 Hyrax
- hyku

ResourceSync revised (2017)

COAR NGR (2017)

Trend: Package, Modules, Flavors



PACKAGE

All functions are provided as one package

MODULES

Part of the functions that compose the system

FLAVOR

A set of modules defined for a specific purpose

Trend: Content Management System (CMS) based Repository

CMS based Repository

Islandora	HUBzero	WEKO
v.1.0 (2009/4/3)	v.0.8.0(2010/9/15)	v.1.0 (2008/3/28)
Drupal 5.x - 8.x	Joomla!	NetCommons2

Pros and cons of CMS based Repository

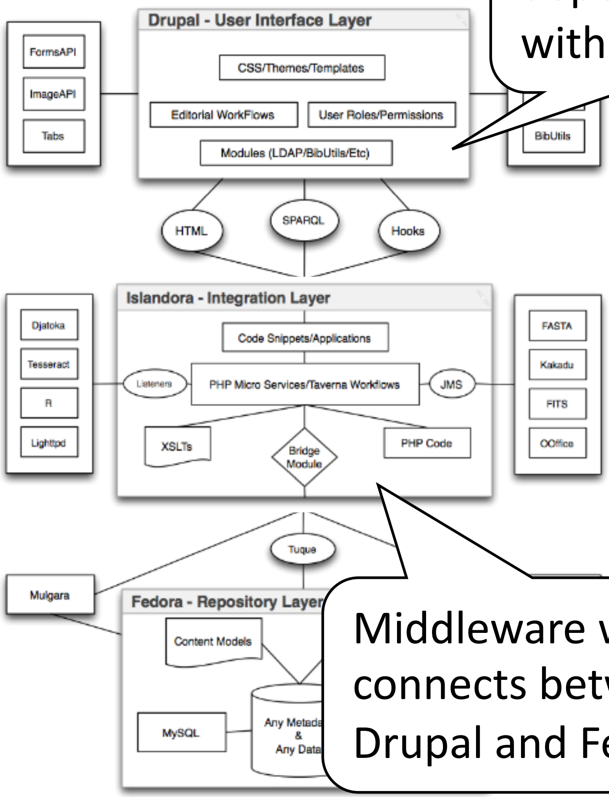
Pros.	Cons.
Make it to manage web site and users easily	It's functions are restricted by CMS
Make it to extend functionality using modules in CMS	it has a strong dependency with CMS
Make it to support scholarly and social communication using modules in CMS	

Islandora CLAW gets independence and horizontal scalability by organizing the dependency (microservice architecture)

Islandora CLAW adopts microservice architecture

Microservice architecture is an architectural style that structures an application as a collection of loosely coupled services.

Islandora 7.x

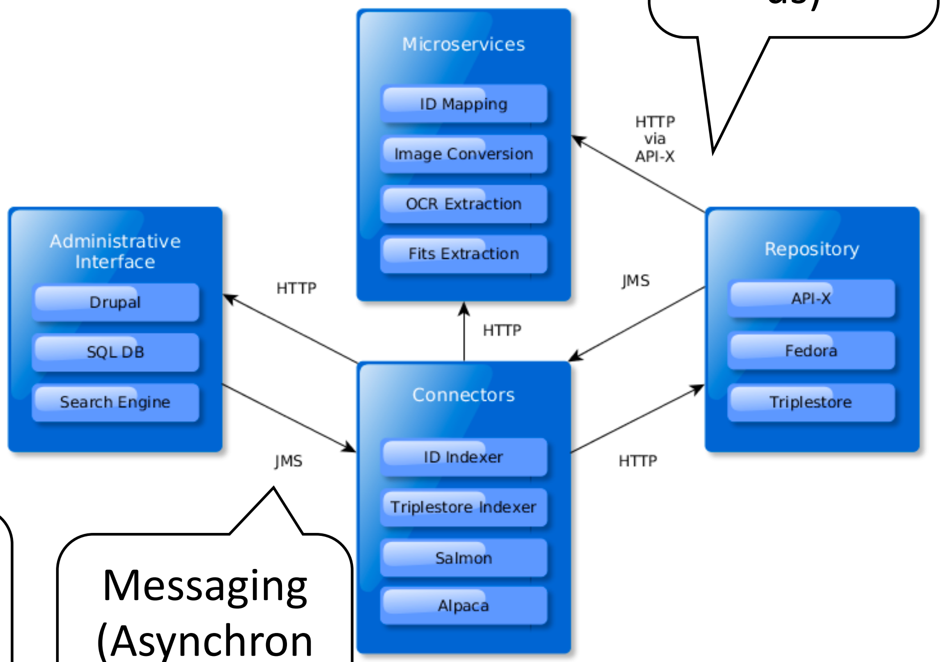


strong dependency with Drupal

Middleware which connects between Drupal and Fedora

Image by https://wiki.duraspace.org/display/ISLANDORA/About+Islandora

Islandora CLAW



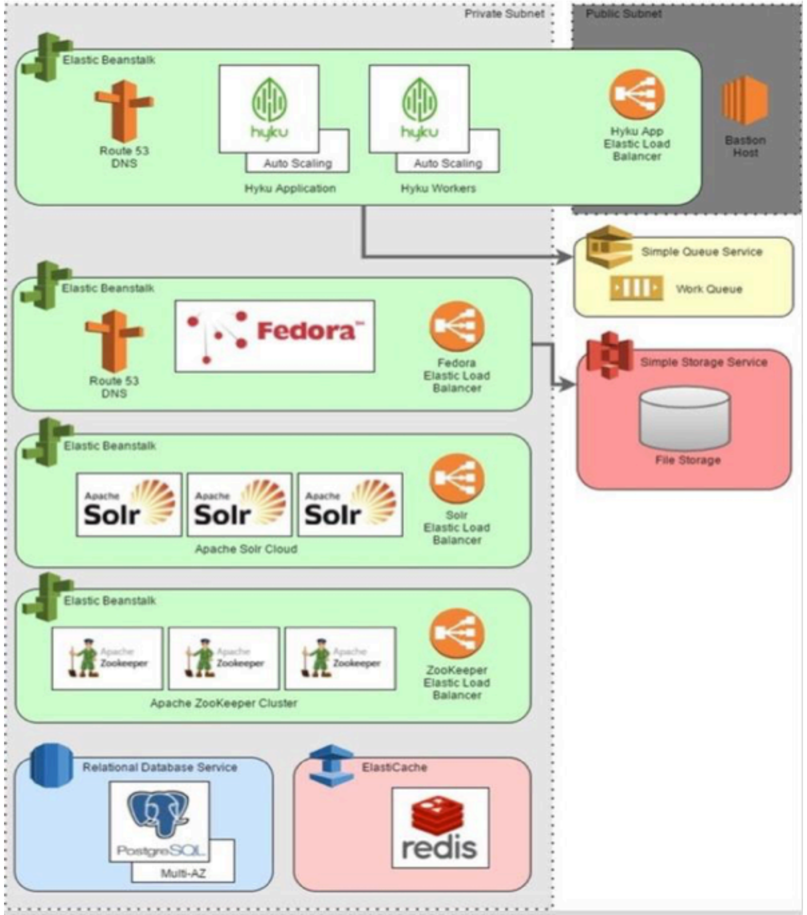
REST (Synchronous)

Messaging (Asynchronous)

Image by https://islandora-claw.github.io/CLAW/mvp/mvp_doc/

Trend: Standalone to Cloud

- Japan national repository service(WEKO)
  - Start : 2012/4
  - Organization: JPCOAR & NII
- DSpaceDirect (DSpace)
  - Start: 2014/03
  - Organization: DuraSpace
- HykuDirect (Hyku)
  - Start: 2017/2
  - Organization: DuraSpace
- No Hassle Hub (HUBzero)
  - Start: Unknown(could not find)
  - Organization: Purdue University
- EPrints Services (EPrints)
  - Start: Unknown(could not find)
  - Organization: University of Southampton



Hyku system architecture

Ref. Tom Cramer, Hannah Frost, Michele Kimpton and Debra Hanken Kurtz: Hydra-in-a-Box Final Project Report, CNI Forum 2017

Trend: Software Development

Name	Version	License	Version control	Issue tracker	CI	Mainling list	Chat	Vagrant	Docker	cloud deploy
DSpace6	6.2	BSD 3-Clause	GitHub	JIRA	Travis CI	Yes	IRC, Slack	Yes	-	-
DSpace7	-	BSD 3-Clause	GitHub	JIRA	Travis CI	Yes	IRC, Slack	-	-	-
EPrints3.3	3.3.16	GPL v.3	GitHub	GitHub	-	Yes	Gitter	Yes	-	-
Eprints3.4	3.4.0	GPL v.3	GitHub	GitHub	-	Yes	Gitter	-	-	-
Fedora 4	4.7.5	Apache 2.0	GitHub	JIRA, GitHub	Travis CI	Yes	IRC, Slack	Yes	-	-
HUBzero	2.1.15	GPL v.2	GitHub	Hubzero	Travis CI	-	-	Yes	-	-
Invenio3	3.0.0rc1	MIT	GitHub	GitHub	Travis CI	-	Gitter	Yes	Yes	OpenShift template
Zenodo	-	GPL v.2	GitHub	GitHub	Travis CI	-	Gitter	-	Yes	-
Hyrax	2.1.0.rc1	Apache 2.0	GitHub	GitHub	Travis CI	Yes	Slack	Yes	-	-
Hyku	v1.0.0.beta2	Apache 2.0	GitHub	GitHub	Travis CI	Yes	Slack	Yes	Yes	AWS template
Samvera	-	Apache 2.0	GitHub	GitHub	Travis CI	Yes	Slack	Yes	-	-
Islandra	7.x-1.10	GPL v.3	GitHub	JIRA	Travis CI	Yes	IRC	Yes	-	-
Islandora CLAW	-	GPL v.2	GitHub	GitHub	Travis CI	Yes	IRC	Yes	-	-
Dataverse4	4.8.6	Apache 2.0	GitHub	GitHub	Travis CI	Yes	IRC	Yes	Yes	OpenShift template
WEKO2	2.4.2	BSD 2-Clause	Bitbuket	-	-	Yes	-	Yes	Yes	-

Use of GitHub,Travis CI and Vagrant are de facto standards

Use of Slack, Gitter are spreading

Invenio, Hyku and Dataverse are providing cloud deploy template using AWS or OpenShift

Trend: System Architecture

Name	Lang	Framework	Base	DataBase	O/Rmapping	Search	Full Text	REST API	Messaging	Hosting
DSpace6	Java	Spring	-	PostgreSQL, Oracle	Hibernate	Solr	Yes	Yes	-	Yes
DSpace7	Java	Spring	-	PostgreSQL, Oracle	Hibernate	Solr		Yes	-	-
EPrints3.3	Perl	-	-	MySQL		Internal	Yes	Yes	-	Yes
Eprints3.4	Perl	-	-	MySQL		Internal	Yes	Yes	-	Yes
Fedora 4	Java	Spring	ModeShape	ModeShape		Internal	-	Yes	-	-
HUBzero	PHP	Joomla!	-	MySQL		Internal	-	-	-	Yes
Zenodo	Python	Flask	Invenio3	SQLite, PostgreSQL, MySQL	SQLAlchemy	Elasticsearch	-	Yes	Celery/Rabbitmq	-
Invenio3	Python	Flask	-	SQLite, PostgreSQL, MySQL	SQLAlchemy	Elasticsearch	-	Yes	Celery/Rabbitmq	-
Hyrax	Ruby	Rails	Samvera	SQLite, PostgreSQL, MySQL	ActiveFedora	Solr	Yes	Yes	Sidekiq/Redis	-
Hyku	Ruby	Rails	Hyrax	SQLite, PostgreSQL, MySQL	ActiveFedora	Solr	Yes	Yes	Sidekiq/Redis	Yes
Samvera	Ruby	Rails	Fedora4	-	ActiveRecord	Solr	-	Yes	-	-
Islandra	PHP	Drupal7	Fedora3	MySQL	Doctrine	Solr	Yes	Yes	-	-
Islandora CLAW	PHP	Drupal8	Fedora4	MySQL	Doctrine	Solr	-	Yes	API-X(Karaf), ActiveMQ	-
Dataverse4	Java	primefaces	-	PostgreSQL	Hibernate	Solr	Yes	Yes	-	-
WEKO2	PHP	NetCommons2	-	MariaDB		Mroonga	Yes	-	-	Yes

Software development has become complicated
However, automating build/test/deployment with CI/CD tool can help developers
Container orchestration software will improve maintenance and operability

What can be seen from History

Trends

- Package, Modules, Flavors
- Content Management System based Repository
- Standalone to Cloud
- Software Development
- System Architecture

Provide as **Cloud Service**
Developed by **Agile**
Composed by **Micro Services**

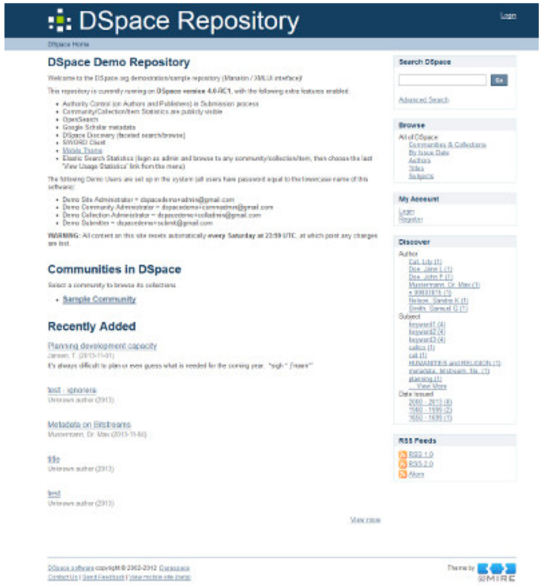
Repository System for Research Data

Flexible and Acceptable to Various Requirements in Developmental and Operational Aspects

Especially Focusing on DSpace and Invenio

DSpace

- DSpace is the most popular repository software in the world.
- It is java based "out-of-the-box" repository software developed by community based opensource project led by DuraSpace.



DSpace Repository

Login

DSpace Home

DSpace Demo Repository

Welcome to the DSpace.org demonstration/sample repository (Mirage / XMLUI interface)!

This repository is currently running on **DSpace version 4.0-RC1**, with the following extra features enabled:

- Authority Control (on Authors and Publishers) in Submission process
- Community/Collection/Item Statistics are publicly visible
- OpenSearch
- Google Scholar metadata
- DSpace Discovery (faceted search/browse)
- SWORD Client
- Mobile Theme
- Elastic Search Statistics (login as admin and browse to any community/collection/item, then choose the last "View Usage Statistics" link from the menu)

The following Demo Users are set up in the system (all users have password equal to the lowercase name of this software):

- Demo Site Administrator = dspacedemo+admin@gmail.com
- Demo Community Administrator = dspacedemo+commadmin@gmail.com
- Demo Collection Administrator = dspacedemo+colladmin@gmail.com
- Demo Submitter = dspacedemo+submit@gmail.com

WARNING: All content on this site resets automatically **every Saturday at 23:59 UTC**, at which point any changes are lost.

Communities in DSpace

Select a community to browse its collections.

- Sample Community

Recently Added

Planning development capacity
Jansen, T. (2013-11-01)
It's always difficult to plan or even guess what is needed for the coming year. *sigh* [more]

test : ignorera
Unknown author (2013)

Metadata on Bitstreams
Mustermann, Dr. Max (2013-11-06)

title
Unknown author (2013)

test
Unknown author (2013)

View more

Search DSpace

Go

Advanced Search

Browse

All of DSpace
Communities & Collections
By Issue Date
Authors
Titles
Subjects

My Account

Login
Register

Discover

Author
Call, Lily (1)
Doe, Jane L (1)
Doe, John F (1)
Mustermann, Dr. Max (1)
[illegible] (1)
Nelson, Sandra K (1)
Smith, Samuel G (1)

Subject
keyword1 (4)
keyword2 (4)
keyword3 (4)
calico (1)
cat (1)
HUMANITIES and RELIGION (1)
[illegible] (1)
planning (1)
... View More

Date Issued
2000 - 2013 (8)
1900 - 1999 (2)
1800 - 1899 (1)

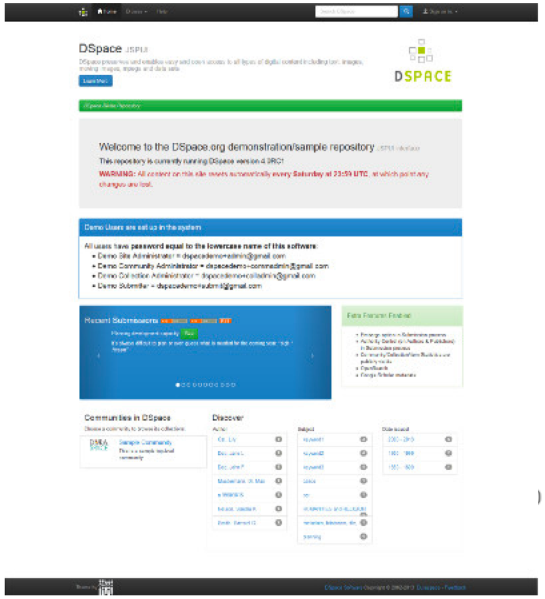
RSS Feeds

RSS 1.0
RSS 2.0
Atom

DSpace software copyright © 2002-2012 Duraspace
Contact Us | Send Feedback | View mobile site (beta)

Theme by @MIRE

XMLUI



Home DSpace Help Search DSpace Sign on to:

DSpace JSPUI

DSpace preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets

Learn More

DSPACE

DSpace Demo Repository

Welcome to the DSpace.org demonstration/sample repository JSPUI interface

This repository is currently running DSpace version 4.0RC1

WARNING: All content on this site resets automatically every Saturday at 23:59 UTC, at which point any changes are lost.

Demo Users are set up in the system

All users have **password equal to the lowercase name of this software**:

- Demo Site Administrator = dspacedemo+admin@gmail.com
- Demo Community Administrator = dspacedemo+commadmin@gmail.com
- Demo Collection Administrator = dspacedemo+colladmin@gmail.com
- Demo Submitter = dspacedemo+submit@gmail.com

Recent Submissions

Planning development capacity

It's always difficult to plan or even guess what is needed for the coming year. *sigh* [more]

Extra Features Enabled

- [illegible]
- Authority Control (on Authors & Publishers) in Submission process
- Community/Collection/Item Statistics are publicly visible
- OpenSearch
- Google Scholar metadata

Communities in DSpace

Choose a community to browse its collections.

Sample Community
This is a sample top-level community

Discover

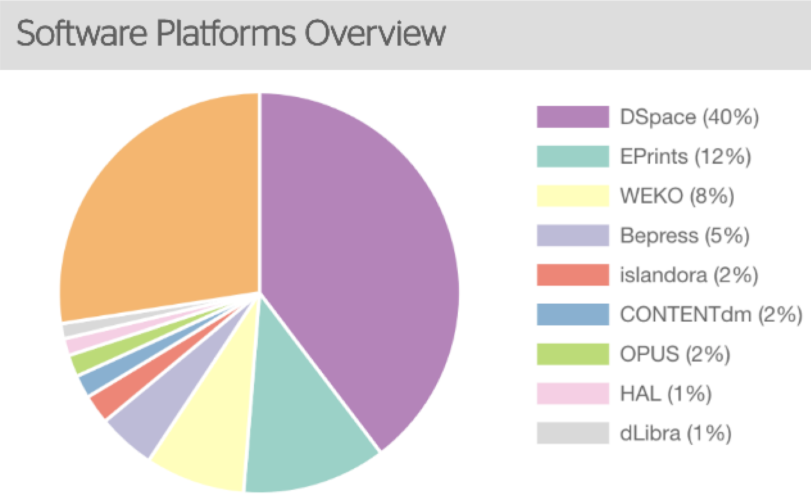
Author
Call, Lily
Doe, Jane L
Doe, John F
Mustermann, Dr. Max
[illegible]
Nelson, Sandra K
Smith, Samuel G

Subject
keyword1
keyword2
keyword3
calico
cat
HUMANITIES and RELIGION
[illegible]
planning

Date issued
2000 - 2013
1900 - 1999
1800 - 1899

DSpace Software Copyright © 2002-2013 Duraspace - Feedback

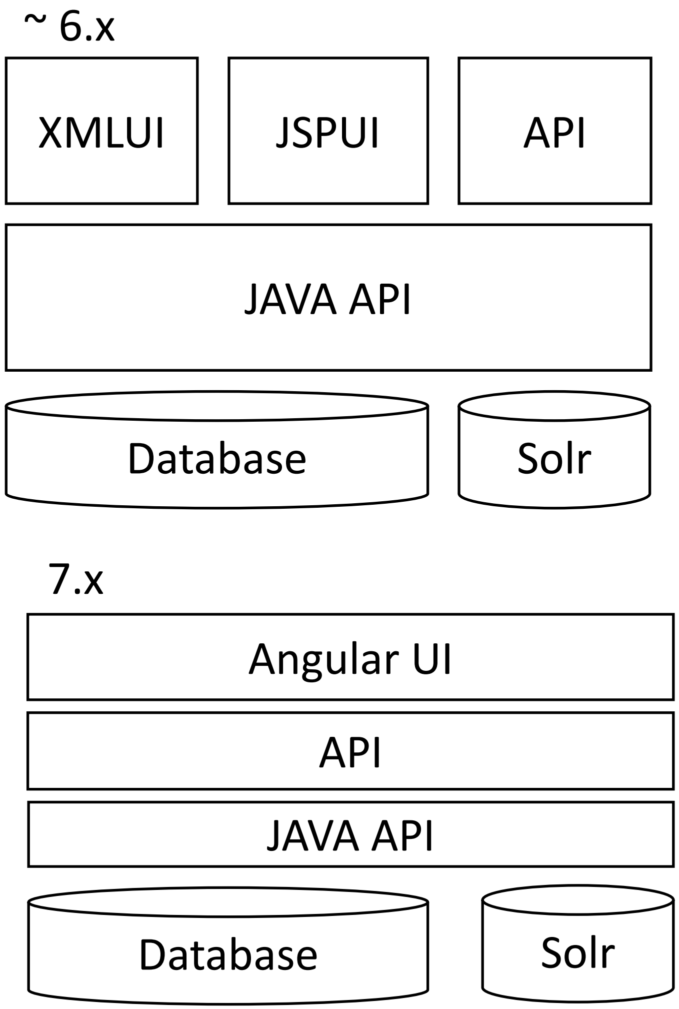
JSPUI



By OpenDOAR Statistics(2020/1/20)

DSpace architecture

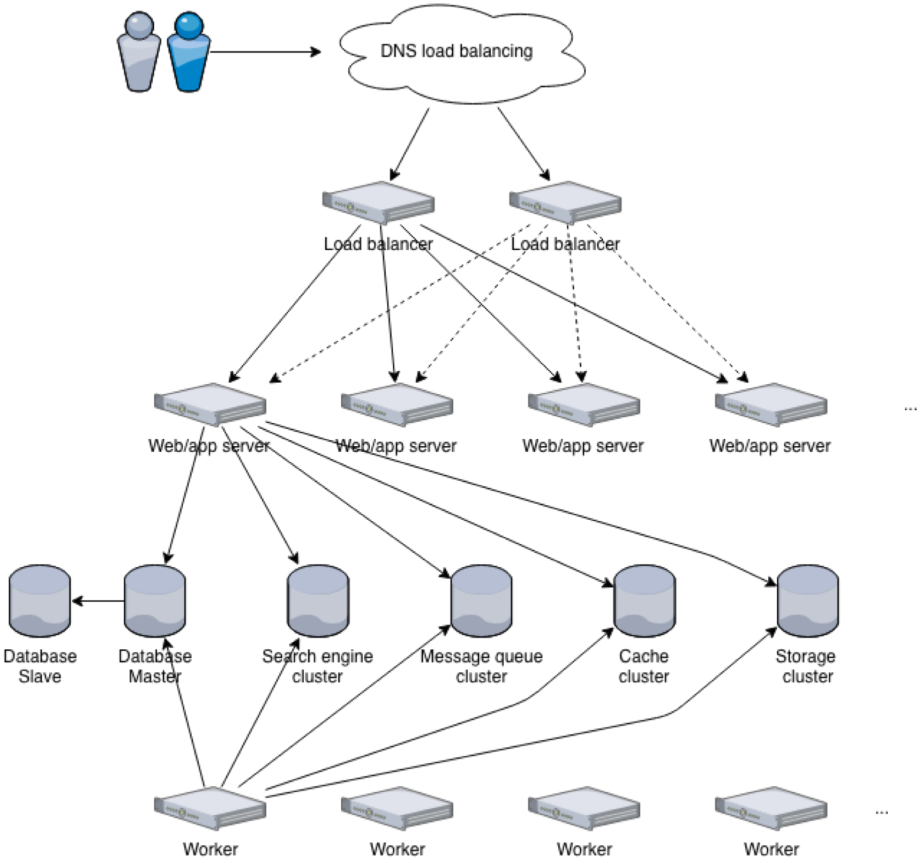
- DSpace 6.x adopts monolithic architecture based on Java servlet container.
  - It is difficult to support research data metadata because It does not support hierarchical metadata(community level implement is exist).
- DSpace 7.x is will start to move microservice architecture.
  - It will support configurable entities function.



Ref. Tim Donohue, Art Lowel, Andrea Bollini, On the Road to DSpace 7: Angular UI + REST,OR2017, https://www.slideshare.net/tdonohue/on-the-road-to-dspace-7-angular-ui-rest

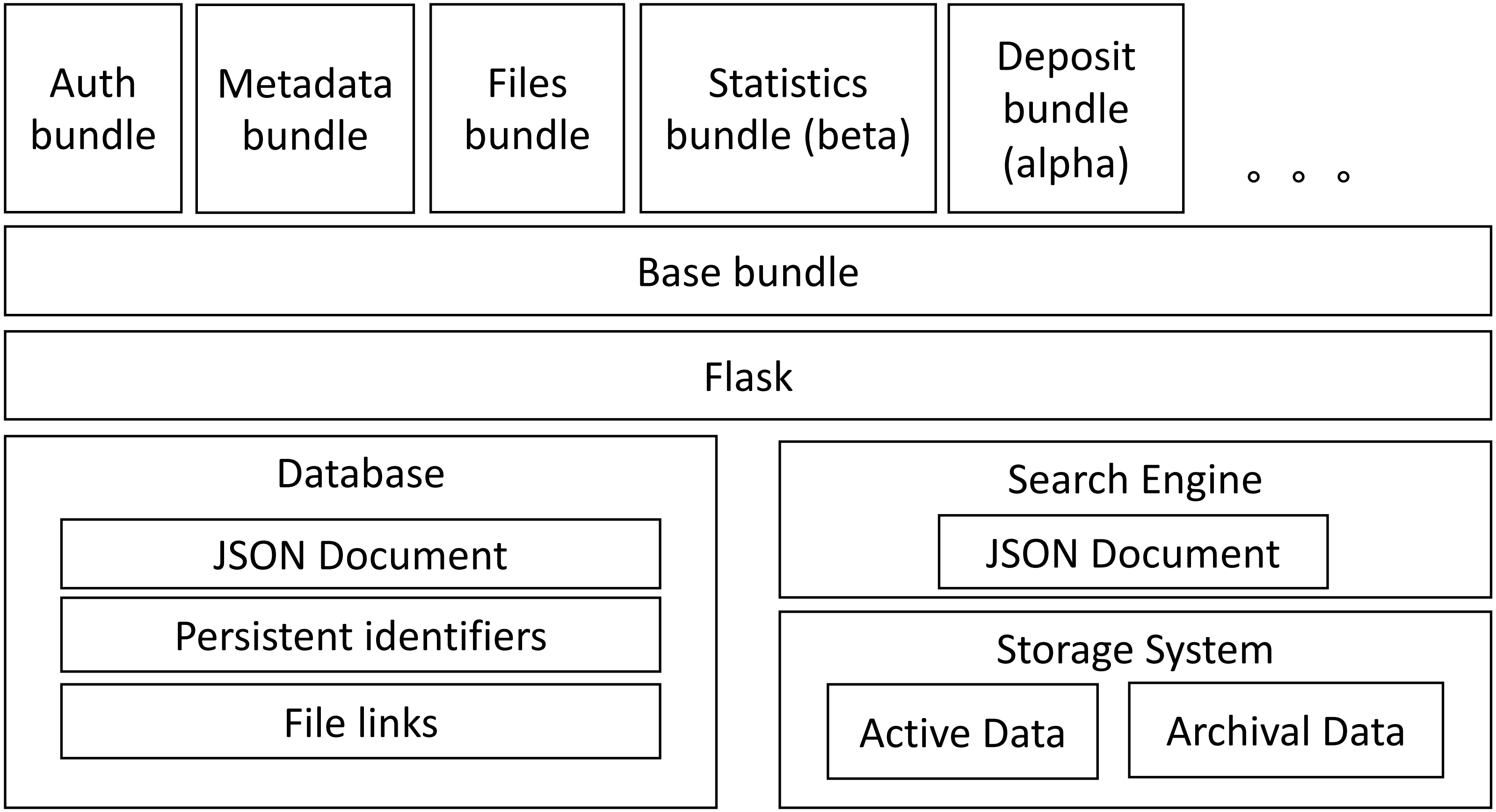
Invenio

- open source framework for large-scale digital repositories led by CERN
- developed based on zenodo code and related projects
- Store metadata as JSON Document
- adopts message-oriented middleware to assist scalability.



Ref. https://invenio.readthedocs.io/en/latest/architecture/infrastructure.html

Invenio3 Architecture



Auth bundle

Metadata bundle

Files bundle

Statistics bundle (beta)

Deposit bundle (alpha)

。。。

Base bundle

Flask

Database

JSON Document

Persistent identifiers

File links

Search Engine

JSON Document

Storage System

Active Data

Archival Data

https://invenio.readthedocs.io/en/latest/general/bundles.html

DSpace vs. Invenio

	Pros	Cons
DSPACE	• Out of the box repository software • Large user community	• Difficult to reflect opinions in development • Low Metadata extensibility, But 7.x will fix this issues
INVENIO	• System is designed for supporting scalability • flexible metadata model is supported using JSON document	• No out of the box repository, require to develop. But recently InvenioRDM project is started. • Small user community

Invenio v3 is best open source repository

- Because…
  - Invenio v3 has **a flexible data model** using JSON document.
    → It can realize to develop flexibility of metadata like WEKO level.
  - Invenio v3 has **excellent extensibility and scalability**.
    →Scalability requirements were essential when assuming JAIRO Cloud operation. Invenio v3 is loosely coupled with the middleware was also attractive.
  - Invenio v3 is a **simple architecture**.
    →We thought that this would help the developer's understanding and donate to future maintenance.

Development of homegrown software Vs. development using existing software

- Repository functionality is getting larger and more complex, scratch development becoming difficult.
- Invenio v3 is a module-based repository framework, it allows almost managed of development, such as developing new functions while combining existing functions.
  - homegrown software ✖
  - ○ development using existing software

As a result, we decided to develop the next version of WEKO using Invenio3.

Table of Contents

- Introduction of WEKO2
- JAIRO Cloud and WEKO2
- Issues of JAIRO Cloud/WEKO2
- Comparison of opensource repository software
- **Introduction of WEKO3**
- Update of NGR implementation

Issues of repository function extensibility

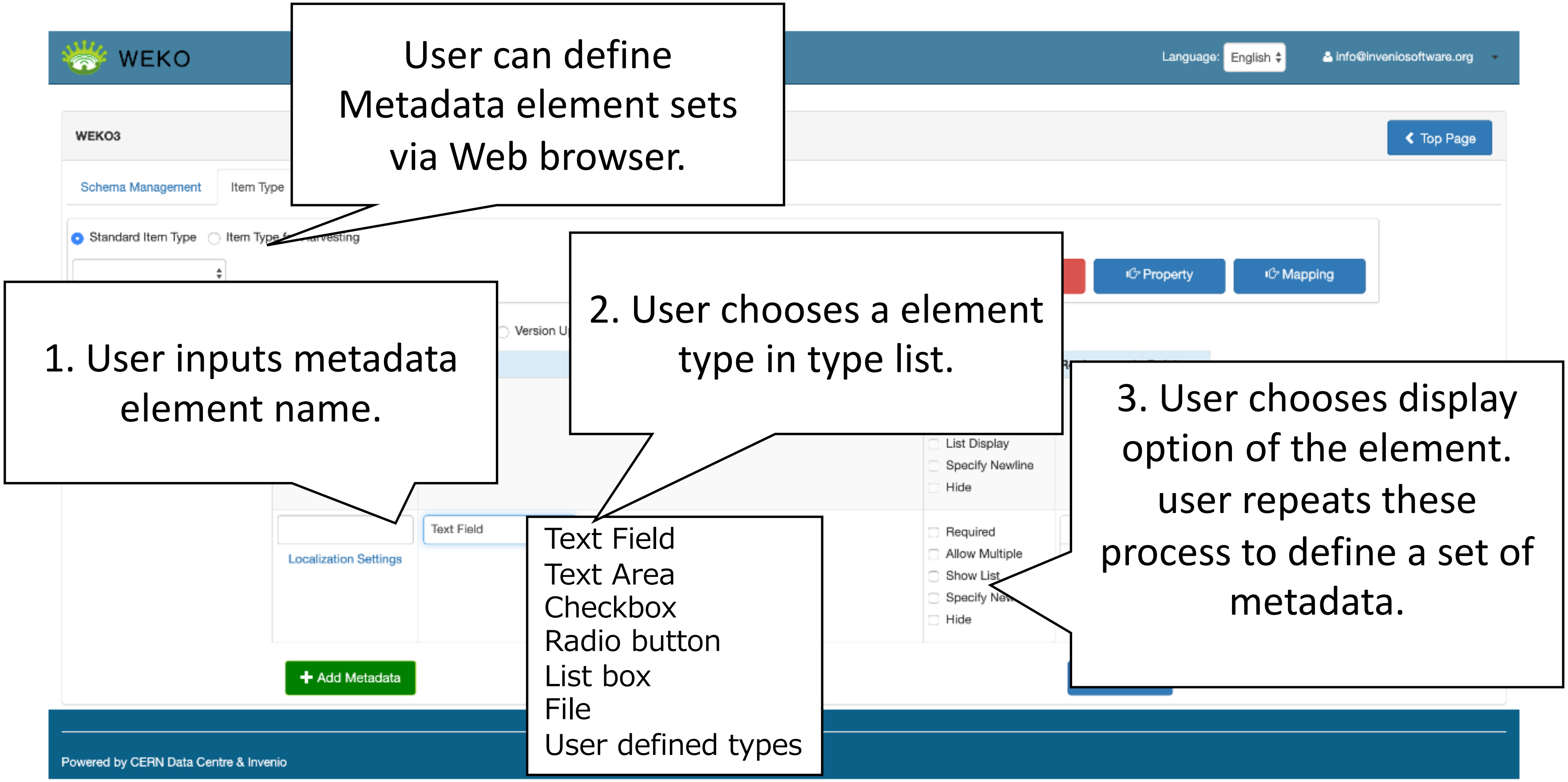
- **Expressiveness of metadata**
  - WEKO2 has a flexible metadata editor, but it does not support hierarchical metadata.
    → A flexible metadata editor that supports hierarchical metadata is required.
- **Function extendibility**
  - WEKO2 supports multimedia file previewer, OAI-PMH provider with multiple metadata schemas, workflow function but they are not extensible.
    → pluggable file previewer, extendibility of OAI-PMH provider function, customizable workflow function are required for supporting various use cases.
- **Function scalability**
  - WEKO2 is not designed for large-scale data processing. As a result, there are problems with faster search functions and batch processing speed.
    →A function design that realizes scalability is required.

WEKO3

- A flavor of Invenio v3 for Institutional Repository
- Main features of WEKO3
  - Integrate WEKO2 functions for IR based on Practical Requests from JAIRO Cloud Users
  - Flexible **metadata editor** and OAI-PMH provider feature
  - Flexible **workflow** feature for ingest workflow
  - **Cloud storage integration** for storing research data
  - **Multitenancy** features for providing perspective on needs of institution and the JAIRO Cloud service

Metadata management function(1/3)

WEKO3 provides flexible metadata management function



User can define Metadata element sets via Web browser.

1. User inputs metadata element name.

2. User chooses a element type in type list.

Text Field
Text Area
Checkbox
Radio button
List box
File
User defined types

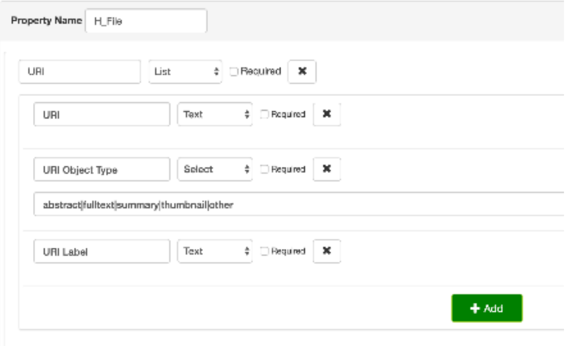
3. User chooses display option of the element. user repeats these process to define a set of metadata.

All operation related metadata management can be done on the web browser

Metadata management function(2/3)

WEKO3 has JSON schema + JSON form editor as "metadata element sets editor”

1. User defines an original element type in “element type editor”.



Property Name: H_File

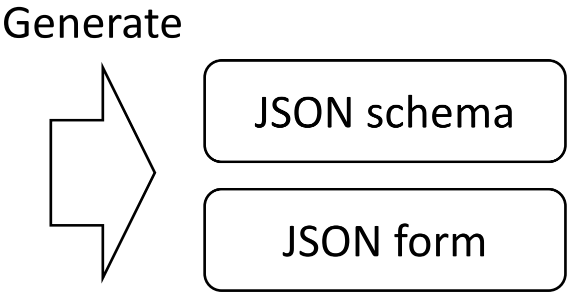
URI | List | ☐ Required | ✖

URI | Text | ☐ Required | ✖

URI Object Type | Select | ☐ Required | ✖

abstract|fulltext|summary|thumbnail|other

URI Label | Text | ☐ Required | ✖

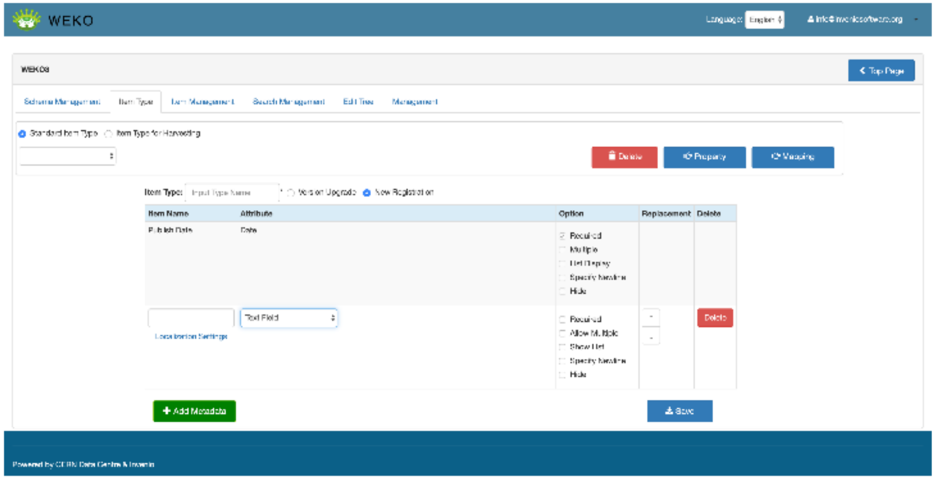


Generate

JSON schema

JSON form

2. User combines built-in and user-defined element types to create a metadata set.

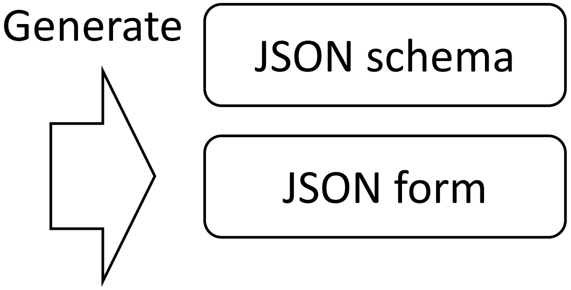


● Standard Item Type ○ Item Type for Harvesting

Item Type: Input Type Name * ○ Version Upgrade ● New Registration

Item Name	Attribute	Option	Replacement	Delete
Publish Date	Date	☑ Required ☐ Multiple ☐ List Display ☐ Specify Newline ☐ Hide		
	Text Field	☐ Required ☐ Allow Multiple ☐ Show List ☐ Specify Newline ☐ Hide		Delete

Powered by CERN Data Centre & Invenio



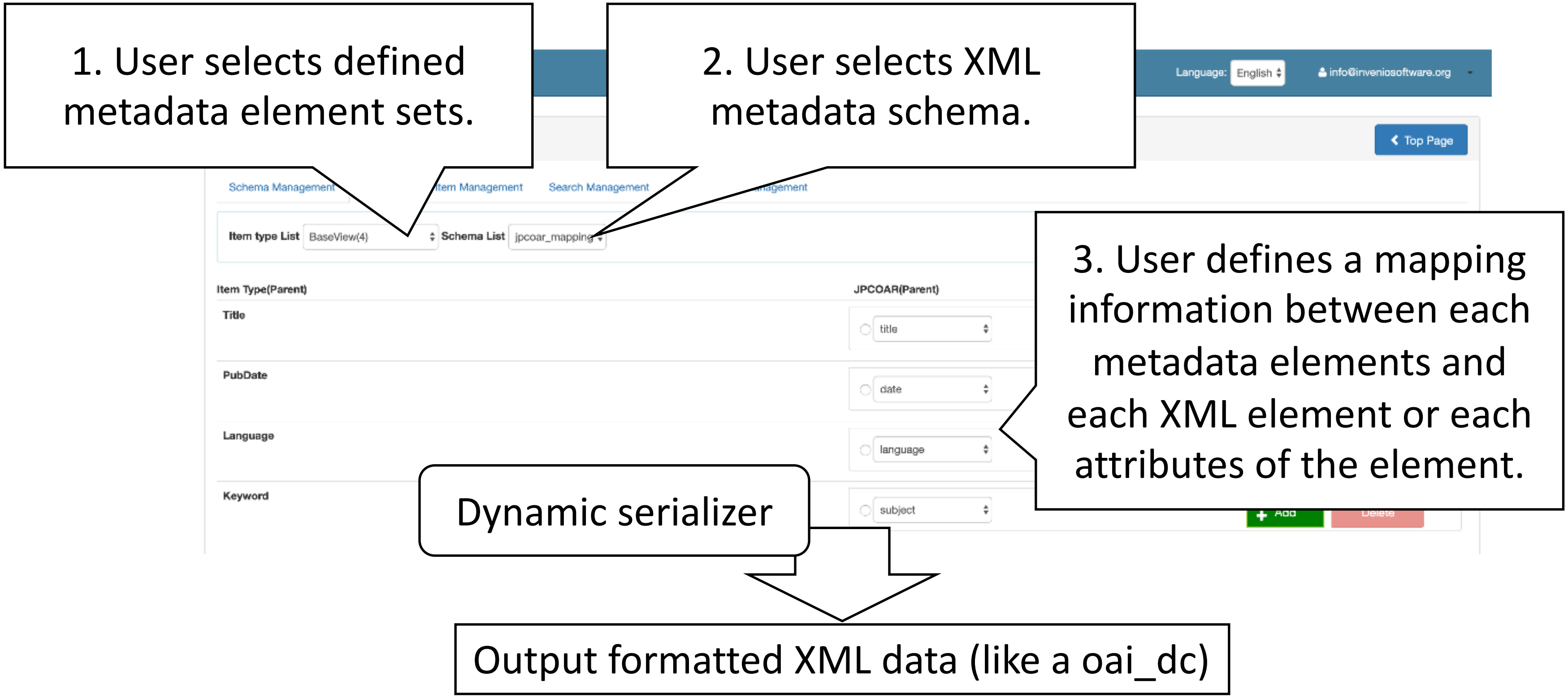
Generate

JSON schema

JSON form

Metadata management function(3/3)

WEKO3 has a sort of dynamic serializer with XML Schema mapping



1. User selects defined metadata element sets.

2. User selects XML metadata schema.

3. User defines a mapping information between each metadata elements and each XML element or each attributes of the element.

Dynamic serializer

Output formatted XML data (like a oai_dc)

Language: English

info@inveniosoftware.org

Top Page

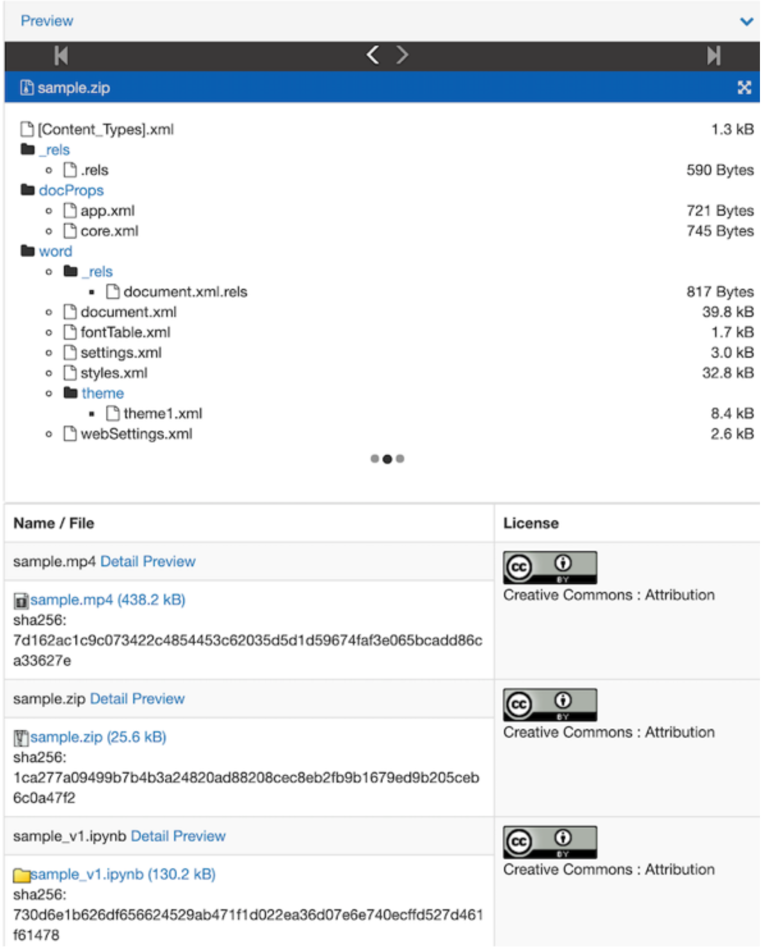
Schema Management | Item Management | Search Management | …nagement

Item type List: BaseView(4) Schema List: jpcoar_mapping

Item Type(Parent)	JPCOAR(Parent)
Title	○ title
PubDate	○ date
Language	○ language
Keyword	○ subject

WEKO3 possible to support a wide variety of metadata formats

Pluggable File Previewer



Preview

sample.zip

```
[Content_Types].xml                1.3 kB
_rels
  ○ .rels                          590 Bytes
docProps
  ○ app.xml                        721 Bytes
  ○ core.xml                       745 Bytes
word
  ○ _rels
      ▪ document.xml.rels          817 Bytes
  ○ document.xml                   39.8 kB
  ○ fontTable.xml                  1.7 kB
  ○ settings.xml                   3.0 kB
  ○ styles.xml                     32.8 kB
  ○ theme
      ▪ theme1.xml                 8.4 kB
  ○ webSettings.xml                2.6 kB
```

Name / File	License
sample.mp4 Detail Preview	Creative Commons : Attribution
sample.mp4 (438.2 kB) sha256: 7d162ac1c9c073422c4854453c62035d5d1d59674faf3e065bcadd86ca33627e	
sample.zip Detail Preview	Creative Commons : Attribution
sample.zip (25.6 kB) sha256: 1ca277a09499b7b4b3a24820ad88208cec8eb2fb9b1679ed9b205ceb6c0a47f2	
sample_v1.ipynb Detail Preview	Creative Commons : Attribution
sample_v1.ipynb (130.2 kB) sha256: 730d6e1b626df656624529ab471f1d022ea36d07e6e740ecffd527d461f61478	

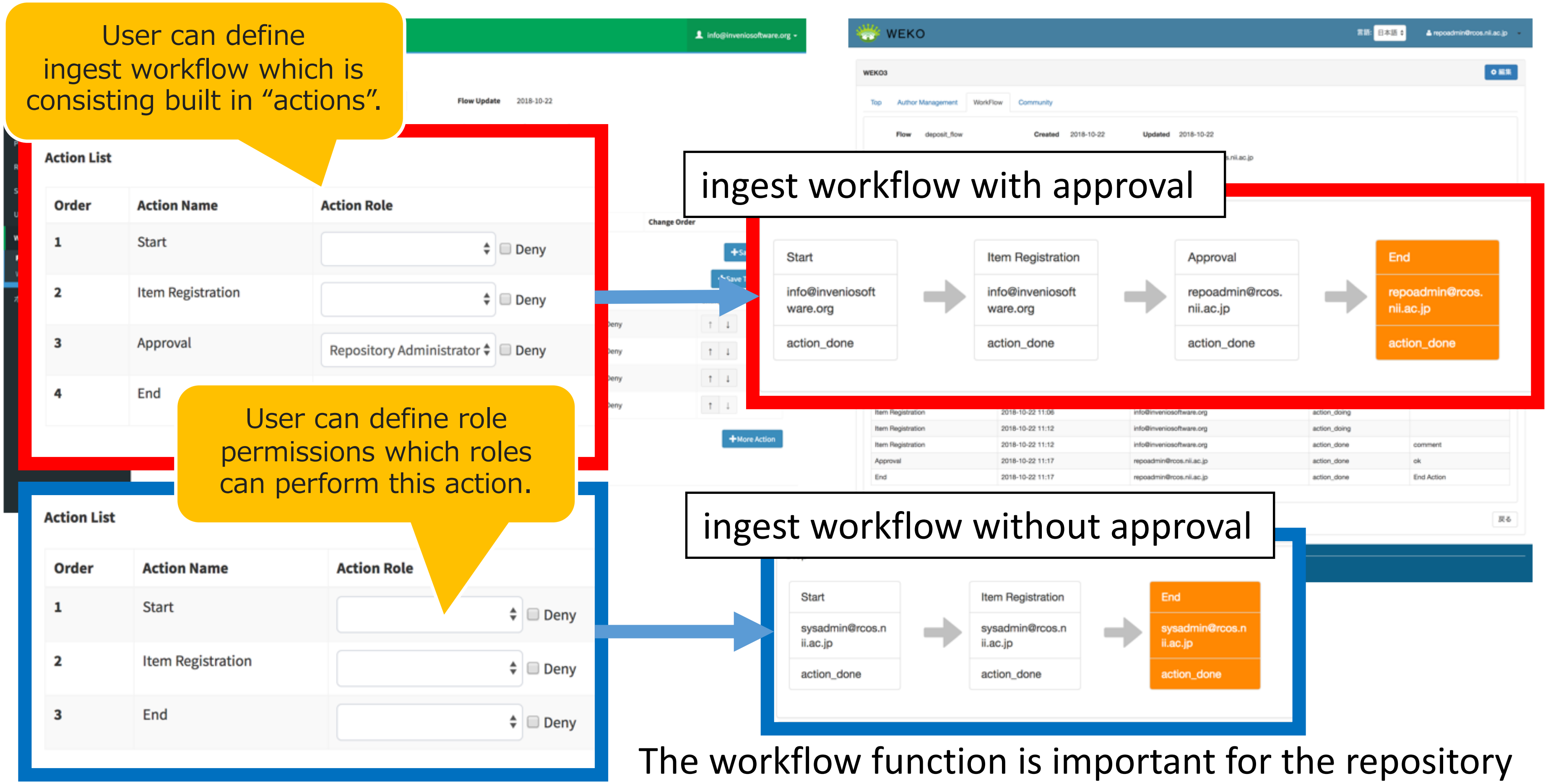
Previewable file types

- Document
  - pdf, doc, docx, xls, xlsx, ppt, pptx
- Movie
  - mp3
- Audio
  - mp3, ogg, wav
- Image
  - jpg, png
- Archive
  - zip
- Others
  - ipynb, wavefront obj(alpha), dicom(alpha)

WEKO3 uses Invenio-previewer
that is pluggable file previewer module of invenio3

Customizable workflow function

In WEKO3, user can customize workflow according to the repository policy



User can define ingest workflow which is consisting built in “actions”.

Flow Update 2018-10-22

Action List

Order	Action Name	Action Role	
1	Start		□ Deny
2	Item Registration		□ Deny
3	Approval	Repository Administrator	□ Deny
4	End		

User can define role permissions which roles can perform this action.

Action List

Order	Action Name	Action Role	
1	Start		□ Deny
2	Item Registration		□ Deny
3	End		□ Deny

ingest workflow with approval

Flow deposit_flow Created 2018-10-22 Updated 2018-10-22

Start	Item Registration	Approval	End
info@inveniosoftware.org	info@inveniosoftware.org	repoadmin@rcos.nii.ac.jp	repoadmin@rcos.nii.ac.jp
action_done	action_done	action_done	action_done

Item Registration	2018-10-22 11:06	info@inveniosoftware.org	action_doing	
Item Registration	2018-10-22 11:12	info@inveniosoftware.org	action_doing	
Item Registration	2018-10-22 11:12	info@inveniosoftware.org	action_done	comment
Approval	2018-10-22 11:17	repoadmin@rcos.nii.ac.jp	action_done	ok
End	2018-10-22 11:17	repoadmin@rcos.nii.ac.jp	action_done	End Action

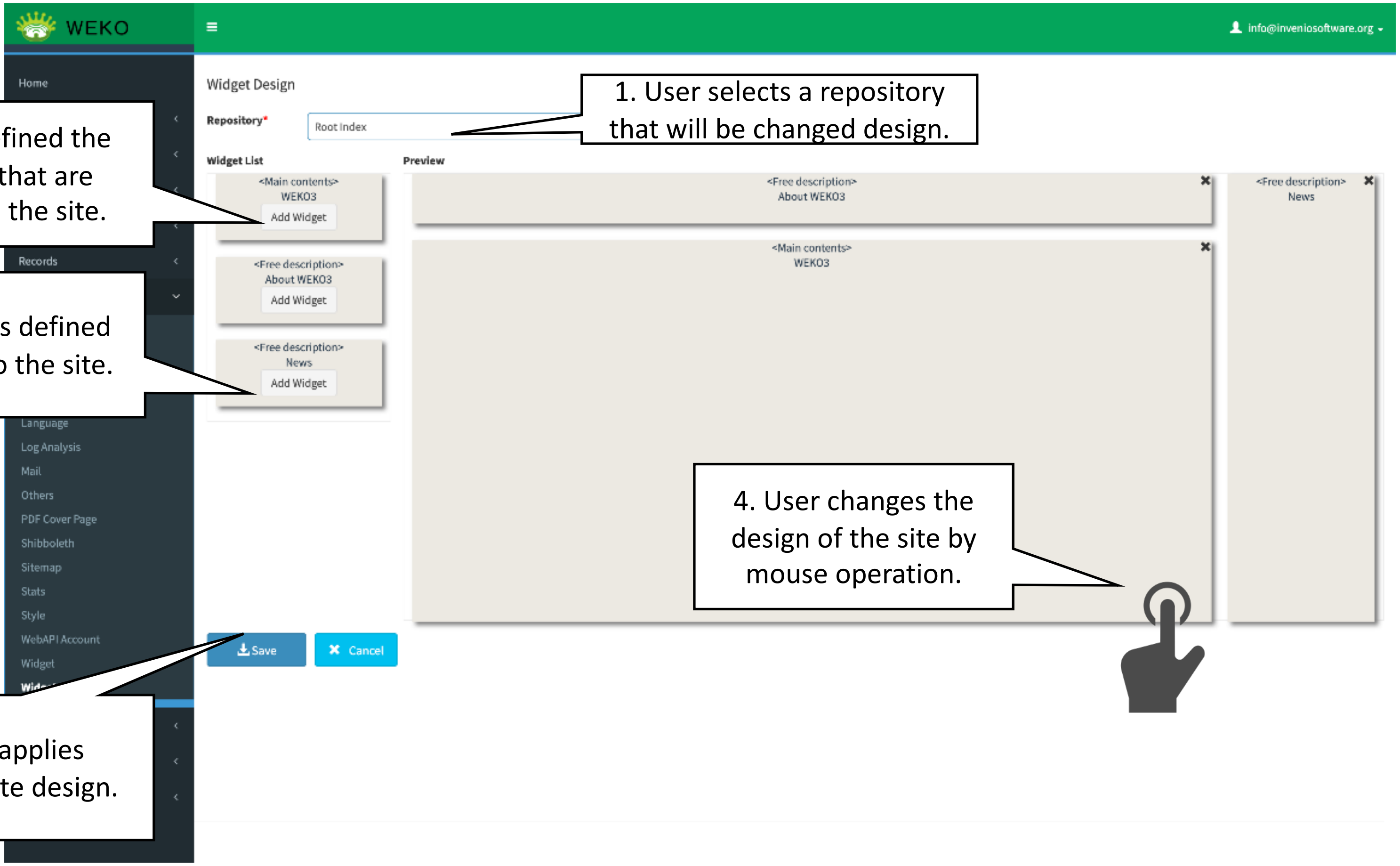
ingest workflow without approval

Start	Item Registration	End
sysadmin@rcos.nii.ac.jp	sysadmin@rcos.nii.ac.jp	sysadmin@rcos.nii.ac.jp
action_done	action_done	action_done

The workflow function is important for the repository to handle research data

Repository Site Designing function(1/2)

WEKO3 can customize site design like a previous version of WEKO



Repository*: Root Index

1. User selects a repository that will be changed design.

2. User defined the widgets that are placed on the site.

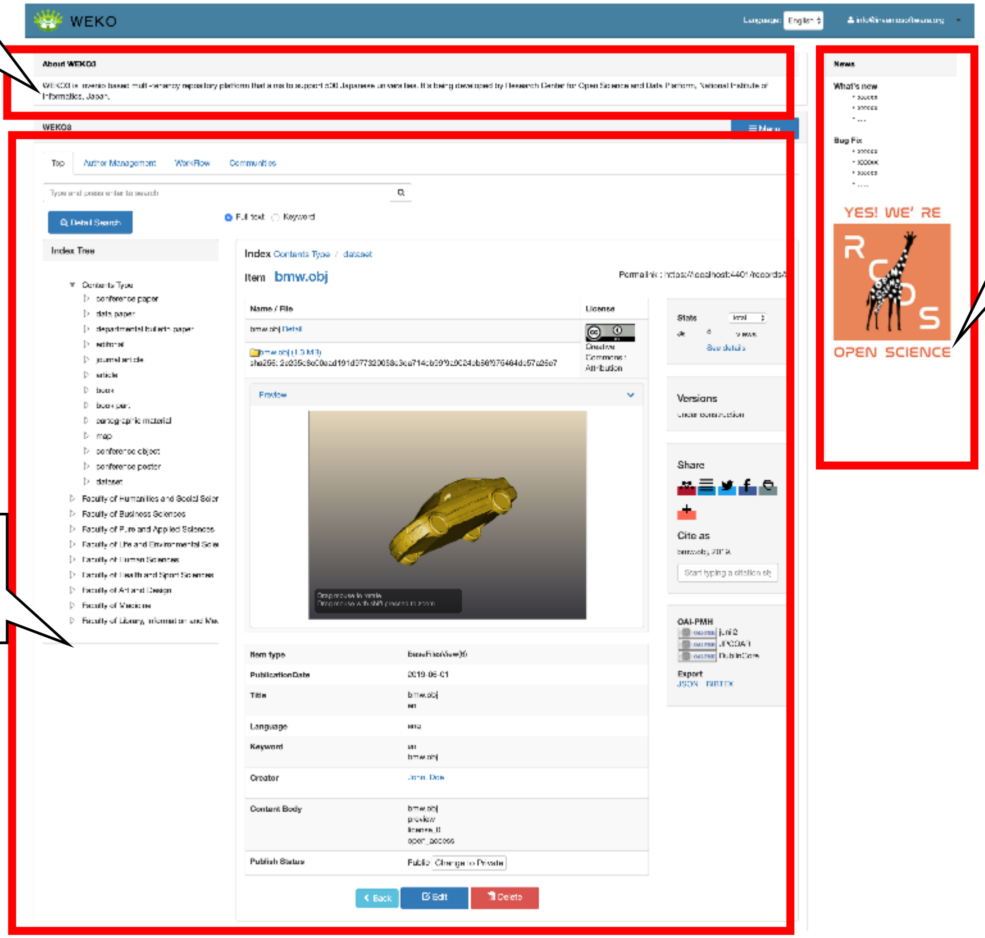
3. User adds defined widgets into the site.

4. User changes the design of the site by mouse operation.

5. User applies changed site design.

Repository Site Designing function(2/2)

WEKO3 displays widgets, repositories
in a user’s design layout



WEKO   Language: English   info@inveniosoftware.org

About WEKO3

WEKO3 is invenio based multi-tenancy repository platform that aims to support 800 Japanese universities. It's being developed by Research Center for Open Science and Data Platform, National Institute of Informatics, Japan.

WEKO3   ☰ Menu

Top | Author Management | WorkFlow | Communities

Type and press enter to search

Detail Search   ● Full text ○ Keyword

Index Tree

- Contents Type
  - conference paper
  - data paper
  - departmental bulletin paper
  - editorial
  - journal article
  - article
  - book
  - book part
  - cartographic material
  - map
  - conference object
  - conference poster
  - dataset
- Faculty of Humanities and Social Scier
- Faculty of Business Sciences
- Faculty of Pure and Applied Sciences
- Faculty of Life and Environmental Scie
- Faculty of Human Sciences
- Faculty of Health and Sport Sciences
- Faculty of Art and Design
- Faculty of Medicine
- Faculty of Library, Information and Med

Index Contents Type / dataset

Item bmw.obj   Permalink : https://localhost:4401/records/

Name / File	License
bmw.obj Detail	Creative Commons : Attribution
bmw.obj (1.3 MB) sha256: 2c035c8c00acd191d377320058c3ea714cb39f0c9024cb5619764 84dc57a26c7	

Preview

Drag mouse to rotate
Drag mouse with shift pressed to zoom

Item type	BaseFileView(6)
PublicationDate	2019-06-01
Title	bmw.obj en
Language	eng
Keyword	en bmw.obj
Creator	John Doe
Content Body	bmw.obj preview license_0 open_access
Publish Status	Public Change to Private

Back | Edit | Delete

Stats   total
See details

Versions
Under construction

Share

Cite as
bmw.obj, 2019.
Start typing a citation style

OAI-PMH
jpcoar
JPCOAR
DublinCore

Export
JSON BIBTEX

News

What's new
- xxxxxx
- xxxxxx
- ....

Bug Fix
- xxxxxx
- xxxxxx
- xxxxxx
- .....

HTML Widget

HTML Widget

Repository Widget

Our development and forked/extended modules list

- Forked modules of Invenio v3
    - invenio-files-rest
    - invenio-s3
    - invenio-oaiharvester
    - invenio-previewer
    - invenio-stats
    - Invenio-records
    - invenio-records-rest
    - invenio-communities
    - invenio-mail
    - invenio-oaiserver
    - invenio-deposit
- Extension modules of Invenio v3
    - weko-theme
    - weko-accounts
    - weko-admin
    - weko-logging
    - weko-groups
    - weko-search-ui
    - weko-records
    - weko-records-ui
    - weko-user-profiles
    - weko-sitemap
    - weko-deposit
- Original modules
    - weko-schema-ui
    - weko-index-tree
    - weko-authors
    - weko-workflow
    - weko-indextree-journal
    - weko-bulkupdate
    - weko-gridlayout
    - weko-itemtypes-ui
    - weko-items-ui
    - weko-items-autofill
    - weko-handle
    - invenio-resourcesyncclient
    - invenio-resourcesyncserver

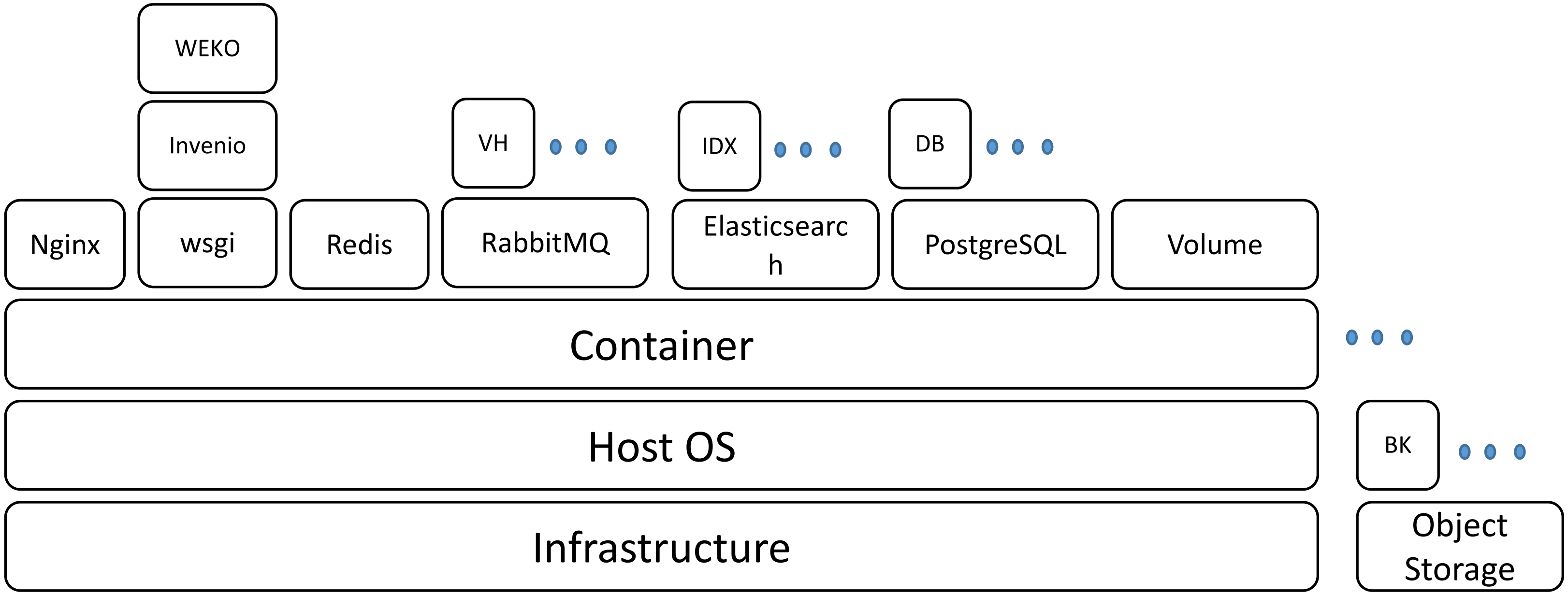
Improving coding skills and moving to test-driven development are the challenges for our team.

Issues of JAIRO Cloud Operation

- **Effective use of resources**
  - JAIRO Cloud is shared repository service, but the sharing scope is one server. Effective utilization of resources is limited because CPU, memory and disk cannot be shared between servers.
    →It is necessary to design a system that shares computer resources as much as possible and uses them effectively.
- **Scalability**
  - A repository is a kind of archive system. The metadata and files stored there continue to increase.
    →System design and operation that can cope with the increase in data volume and metadata is required.
- **Safe Deploy**
  - Function development can be centralized by sharing the repository as infrastructure. However, developments can affect the entire cloud server.
    →System design and operation for realizing function development and secure deployment while continuing operation are required.

Next version of JAIRO Cloud architecture

JAIRO Cloud using WEKO3 adopt Container based architecture



WEKO

Invenio

VH • • •

IDX • • •

DB • • •

Nginx

wsgi

Redis

RabbitMQ

Elasticsearc
h

PostgreSQL

Volume

Container • • •

Host OS

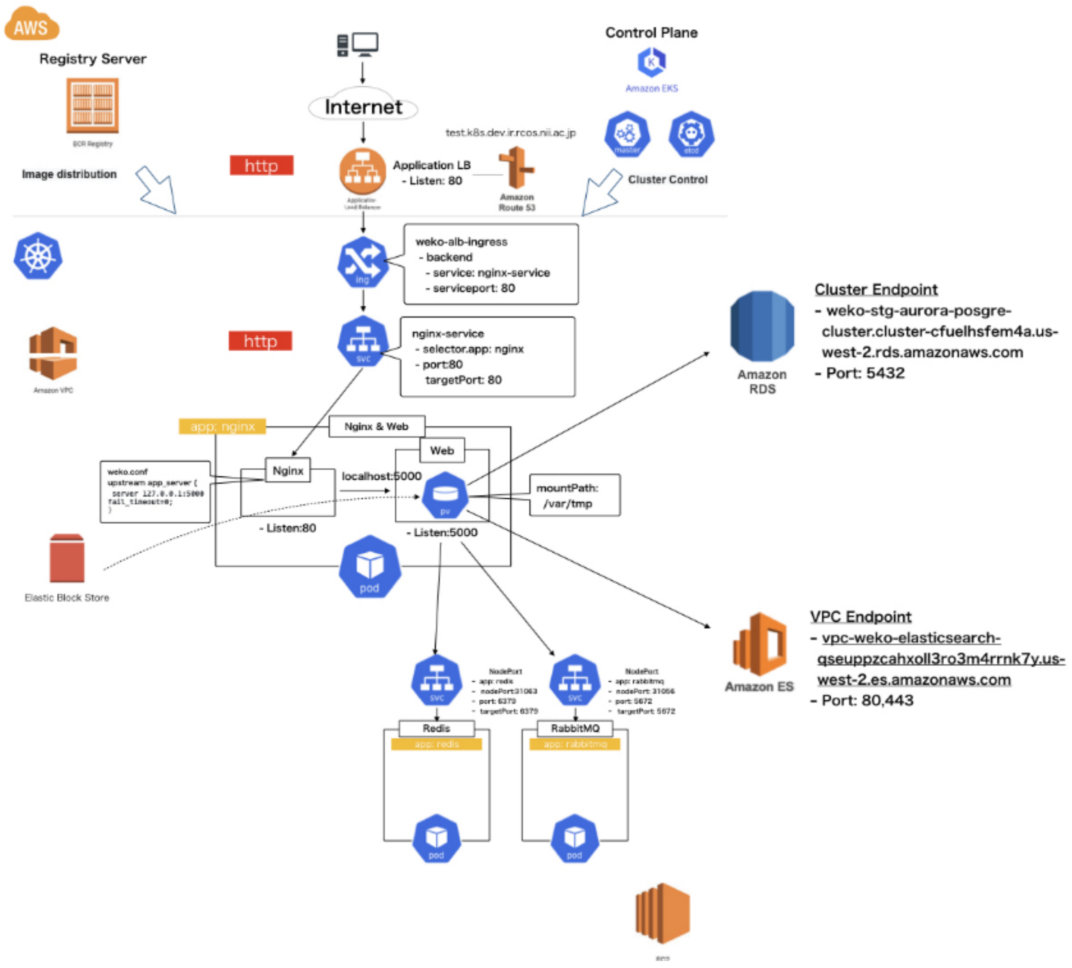
BK • • •

Infrastructure

Object Storage

Computer resource sharing can be realized even more than before

WEKO3 on Kubernetes prototype (AWS EKS)



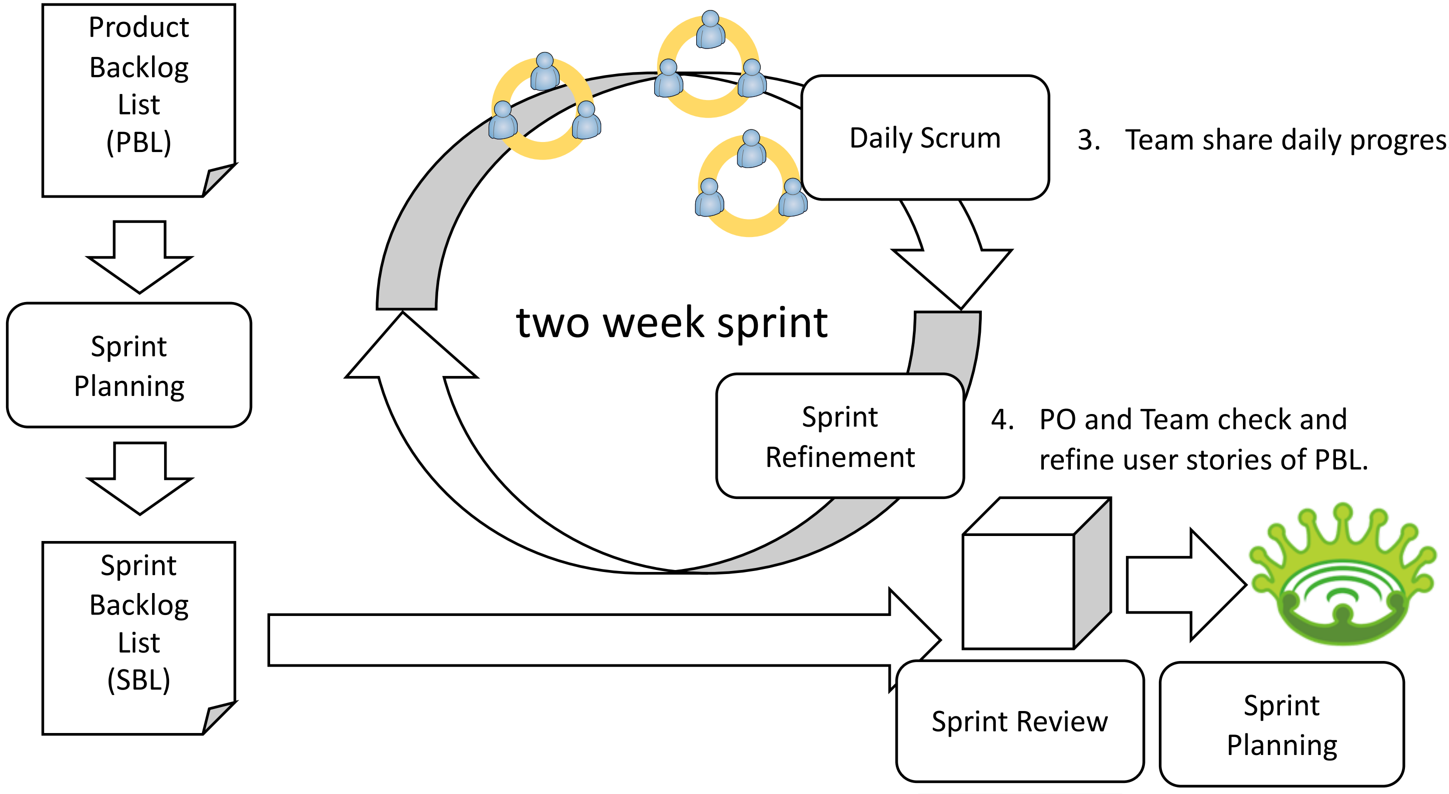
```
weko.conf
upstream app_server {
 server 127.0.0.1:5000
fail_timeout=0;
}
```

Reduce operator load by automation

WEKO3 Development Style

Adopting **Scrum**, an **Agile** software development style

NII: Product owner(PO) , 5 people
Software Development Company: Team,24 people
(SE x 6, PG x 13 including offshore members)



Product Backlog List (PBL)

Sprint Planning

Sprint Backlog List (SBL)

two week sprint

Daily Scrum

Sprint Refinement

Sprint Review

Sprint Planning

Sprint Retrospective

1. PO determine the user story priority
2. PO and Team put user stories of PBL into SBL in Sprint Planning Meeting.
3. Team share daily progress
4. PO and Team check and refine user stories of PBL.
5. PO and Team review the work in this sprint and discuss the output and reflect next sprint.

REDMINE
flexible project management

DevOps in WEKO3 development and deployment

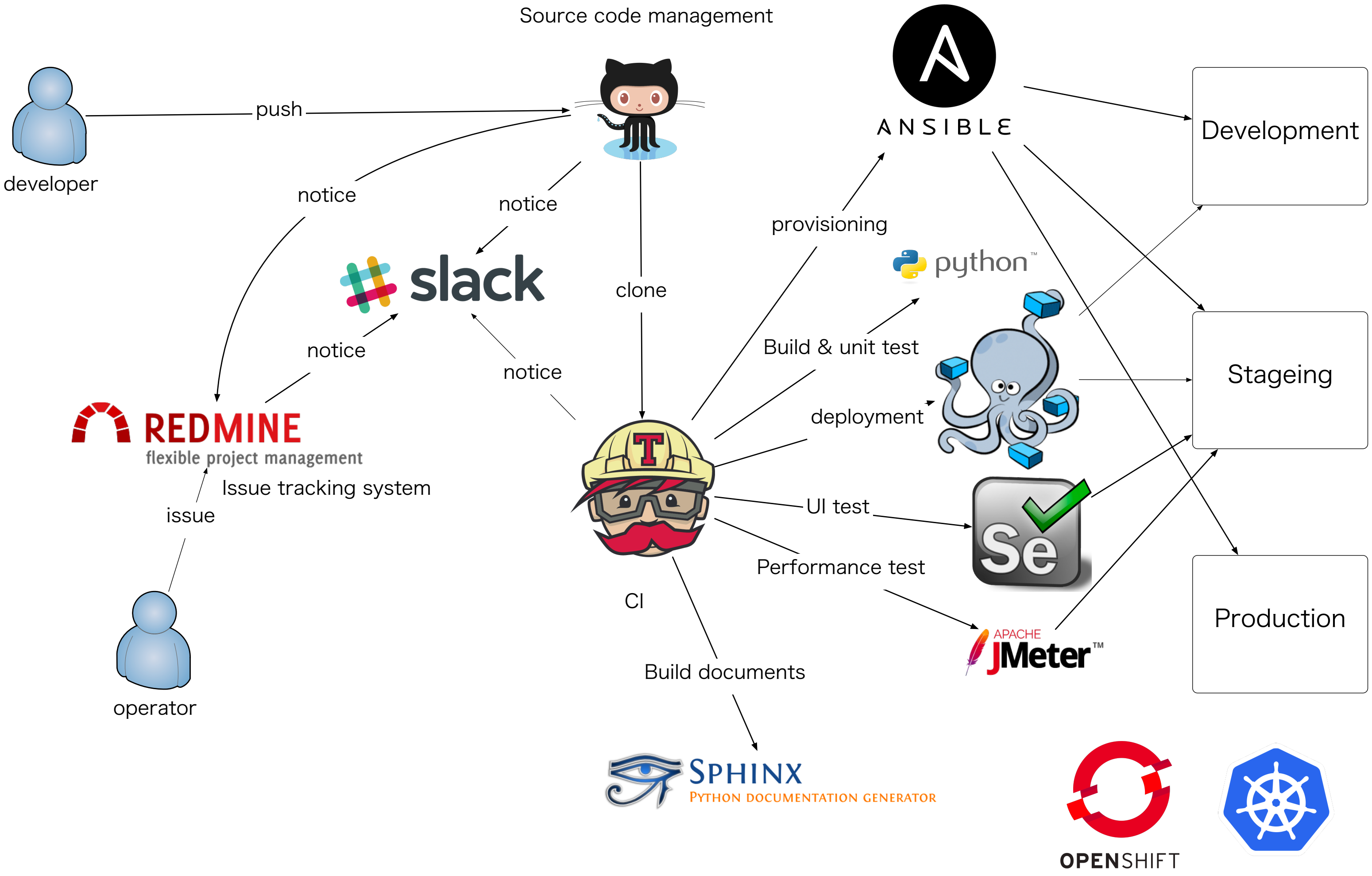


Table of Contents

- Introduction of WEKO2
- JAIRO Cloud and WEKO2
- Issues of JAIRO Cloud/WEKO2
- Comparison of opensource repository software
- Introduction of WEKO3
- **Update of NGR implementation**

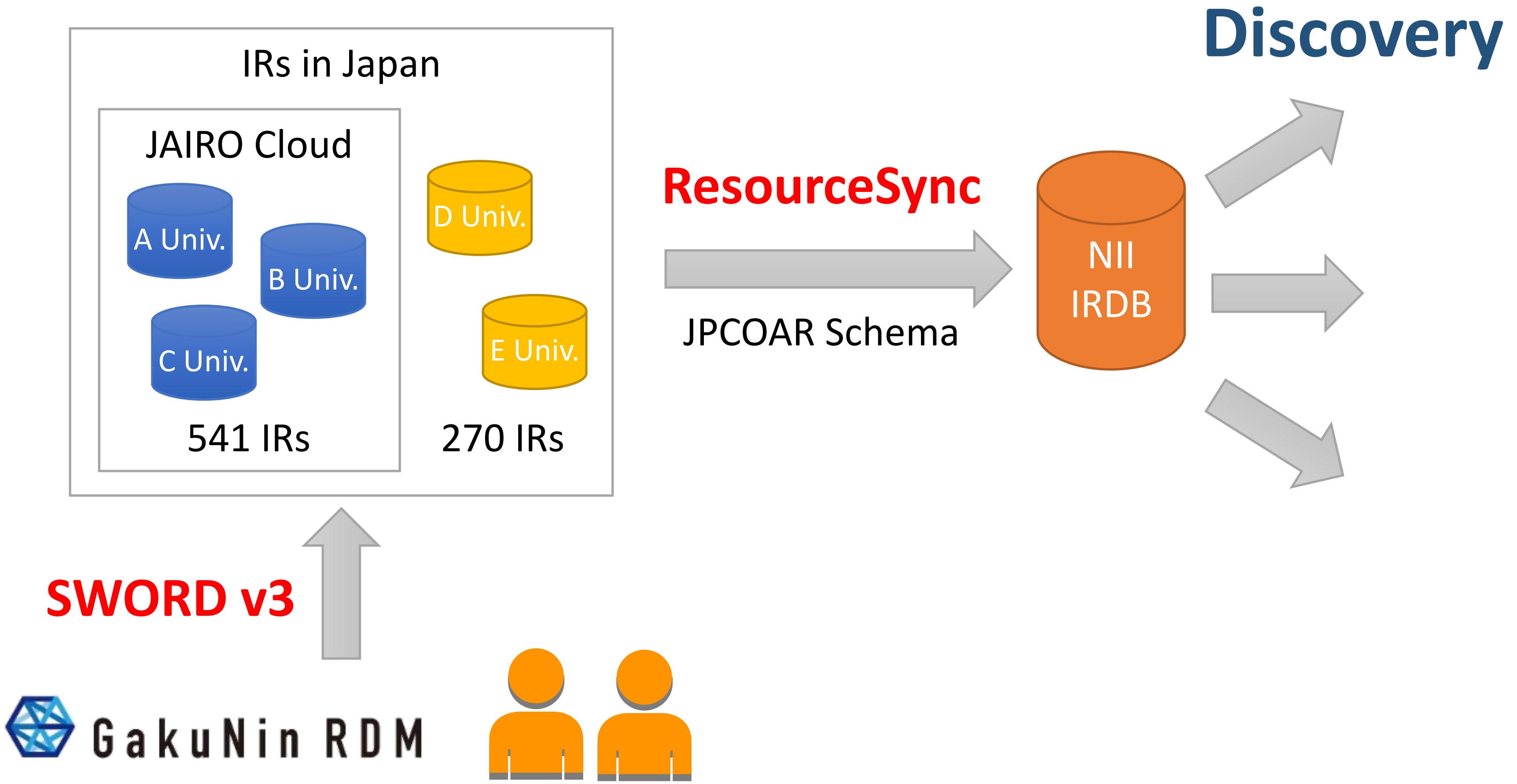
Basic Concept for NGR Implementation

- COAR Next Generation Repository
  - To position repositories as the foundation for a distributed, globally networked infrastructure for scholarly communication
- Our approach:
  - To develop the NGR functions with users requirements and make the use cases of them
  - To develop the NGR functions based on the context of resource centric
  - To combinate suitable services (e.g. repository, discovery and so on) for realizing NGR functionality

Development of NGR recommended functions by NII

- SWORD3
  - Starting to develop reference implementation of SWORDv3 on Invenio
    - https://github.com/swordapp/invenio-sword/
    - https://github.com/swordapp/sword3-client.py
    - https://github.com/swordapp/sword3-common.py
- ResourceSync
  - Starting to develop ResourceSync functions on WEKO3
  - https://github.com/RCOSDP/weko/tree/develop/modules/invenio-resourcesyncclient
  - https://github.com/RCOSDP/weko/tree/develop/modules/invenio-resourcesyncserver

Expected Case Study at NII



RCOS

mhaya@nii.ac.jp