# Exercise 2: Comparing repository systems

## Introduction

As you know from the DP lecture… it is impossible to make a decent comparison of repository software without getting a hands-on experience. Several studies exist (see References), but they are superficial and mostly based on marketing materials and not real experience.

In this exercise, you are going to compare two repository systems. You will have to install them and migrate data between them. Thus, you will get practical experience that will allow you to create a credible comparison of repository software. The outcome will be a report in which you present results, discuss comparison criteria, and provide supporting evidence.

You are supposed to collaborate in **groups of two**. For questions please use TUWEL forum. When there is no answer within few days (unlikely), please write to tmiksa@sba-research.org

## Deadline and Groups

The start for the exercise is 26.04.2018. The **deadline** is **07.06.2018 at 23:55** local Vienna time.

There are 6 options of this exercise, choose **only one**:

1. DSpace[1] -> CKAN[2]
2. DSpace -> Fedora
3. CKAN -> DSpace
4. CKAN -> Fedora
5. Fedora -> DSpace
6. Fedora -> CKAN

'->' denotes the direction of a migration. For example, in the first case you need to migrate data from DSpace to CKAN, not the other way round.

Fedora is repository backend software. It is different than CKAN and DSpace which consist of backend and frontend. Nevertheless, Fedora provides minimal functionality allowing viewing and managing objects. Groups working with Fedora can either use the Fedora Commons[3] or the Islandora[4] which is repository software based on the Fedora Commons that has additionally Drupal as a frontend.

In case you would like to use a different repository (see lecture slides for a list), then please let us know beforehand – this can be arranged.

[1] http://www.dspace.org
[2] https://ckan.org
[3] https://fedorarepository.org
[4] https://islandora.ca/

# Detailed instructions

In this exercise you are going to compare two repository systems: repository A and repository B. You need to:

1. install on your local machine repository A and repository B
2. populate repository A with content
3. migrate objects from repository A to repository B
4. compare repositories and write a report

## Installation

Go to the official websites of repository software to get specific installation instructions. You can also find many tutorials and mailing lists on the internet. You do not have to install everything from scratch – you can use available virtual machines, containers, or packages. Important is that both repositories are operated by you and you have all privileges needed to configure them.

## Content and metadata

Repository A is empty (B as well). You need to upload some content into repository A, before you will migrate it. Please add **at least two files for each** content type.

- Text (e.g. PDF, DOCX)
- Data (e.g. CSV, NetCDF)
- Image
- Audio
- Video
- Source Code

Create also a collection of objects. Collection groups objects in a similar way as folders. Each of the repository systems has a collection or an equivalent. Example of a collection: https://phaidra.univie.ac.at/detail_object/o:378098

Provide metadata for each uploaded file. Use the default metadata scheme of a repository.

However, for Audio and Video objects extend the metadata to include information on bitrate and length of the recording. For this purpose you need to define your own metadata standard and configure the repository to use it.

## Object migration

Use APIs of both repositories to migrate automatically objects from repository A to repository B. Write a tool that:

- lists contents of the repository A
- lists contents of the repository B
- migrates all objects from the repository A to the repository B

The tool must be run from a command line and must accept parameters allowing choosing each of the options. Configuration details, such as repository address, must be provided in a separate configuration file.

You can implement the tool using any programming language. Publish the tool on GitHub – provide a readme and provide a proper license.

## Comparison and reporting

Compare both repositories and discuss differences. Base your comparison on the experience you have gathered by installing, configuring, and working with the systems. Provide evidence, describe specific challenges. Please consider following points in your comparison:

- Installation
  - What are the available options?
  - How is the process automated?
  - What are the main tasks that need to be completed?
  - How big is the supporting community and how easy it is to find solutions to problems?
  - What is the quality of available documentation?
- Content Organization
  - What are the supported content types?
  - Is it possible to define new content types?
  - Does a repository support collections?
  - How repository supports versioning?
  - What persistent identifiers are used by the repository?
- Metadata
  - What are the default standards?
  - How to add another standard?
- Content presentation
  - How are contents presented?
  - Is there a preview?
  - Can the file contents be directly viewed?
- Content Search and Discovery
  - Can all metadata fields be used for searching?
  - Are there classifications that support faceted search?
  - Can the contents of data objects be searched (e.g. text search)?
- User management and access rights
  - What are the mechanisms to authenticate users?
  - Can different roles be defined for users?
  - Can embargo periods be defined for data?
  - How privacy and security are ensured?
- Reporting and administration
  - Does the system provide statistics, e.g. size of data, active users, etc.?
  - Is there a central administration panel?
  - Can the system be configured and managed in one place or changes are needed in several places?
- Interoperability
  - What is the coverage of the API?
  - Can data be easily migrated?

- Can metadata schemas, content type definitions, etc. be exported?
- Was there any information lost due to migration?

For each point provide evidence supporting your statements. You can use tables to summarise your findings. Refer to specific examples you have collected when migrating different content types.

In conclusions provide advantages and disadvantages of each system and explain which system you would choose and why.

## Submission

PDF report describing what you did:

- indicate which repositories you have compared
- Tool description
  - provide link to GitHub to the software you have developed
  - provide screenshots demonstrating how listing and migration work
  - describe how you implemented it and what were specific challenges
- Comparison (address points outlined in the previous section)
  - present comparison criteria
  - provide evidence
  - provide summary (e.g. a table)
  - provide recommendation

## References


1. Lecture slides
2. Institutional repository software comparison: DSpace, EPrints, Digital Commons, Islandora and Hydra https://open.library.ubc.ca/cIRcle/collections/graduateresearch/42591/items/1.0075768#downloadfiles
3. Research Data Repositories: Review of current features, gap analysis, and recommendations for minimum requirements https://www.rdc-drc.ca/wp-content/uploads/Review-of-Research-Data-Repositories-2015.pdf
4. OPEN SOURCE SOFTWARE FOR DIGITAL PRESERVATION REPOSITORIES: A SURVEY https://arxiv.org/ftp/arxiv/papers/1707/1707.06336.pdf
5. Institutional Repository Software Comparison https://works.bepress.com/jean_gabriel_bankier/22/