Automatic Ingestion of Large-scale Binary Files, Data and Associated Metadata using REST API - Scope of DSpace

Nushrat Khan & William Ingram
Scholarly Communication & Repository Services, University of Illinois at Urbana-Champaign

Introduction

Institutional repository systems built on DSpace, such as IDEALS have been used to store smaller-scale research data so far. However, the compatibility of such frameworks to ingest large-scale scientific data using the most recent REST API is not well explored yet.

Current Workflow for Data Ingestion on IDEALS

- Time consuming
- Requires external support
- Needs separate Packaging system

Overview of DSpace Model for Metadata Ingestion

Supported metadata format: Qualified Dublin Core
Supported JSON Format: {"key": "dc.field", "value": "field value", "language": "value"}
Supported object data types for Community, Collection, Item and Bitstream Objects:
Community Object: id, name, handle, link, countItems
Collection Object: id, name, handle, link, numberItems
Item Object: id, name, handle, link, lastModified, bitstreams, title, description, author
Bitstream Object: id, name, link, format, sizeBytes

Methodology

Automation of Data Ingestion Process

Data Sending and Receiving System

Matchmaker is used to select and send research objects to certain repository. JSON-LD is used for ORE-map.

Data Extraction and Sharing with Repository

Received ORE Map is parsed and extracted metadata is used to create JSON objects to ingest into DSpace using Rest Client.

Results & Future Directions

As a part of the SEAD project, we have developed a service to automatically receive data and metadata from users and ingest them into the repository.

The system is currently built to test on Dspace but can be modified to use with any other repository system, such as Fedora. We have found that the current Dspace REST API does not support all metadata formats, especially at the bitstream level. Therefore, further development is necessary with Dspace and other repository systems should be tested as well.