Duplicates in a repository: remediation and reconciliation in three systems, including DataCite

Suni Wong, Ask A Librarian Intern; Esther Jackson, Scholarly Communication Technologies Librarian; Frederic Duby, Programmer/Analyst; Kathryn Pope, Digital Repository Manager

**Problem**

- Content added to the repository from various sources introduced duplicates over time in multiple systems.
- Each system had slightly different (and valuable) information about these items.
- The project team needed to remove duplicates from users' view while retaining the valuable information about these records.

**Process**

1. Identify and flag duplicate items on Academic Commons (AC) over time.
2. Create a list to track and map relationships between AC item PIDs and DOIs of duplicate resources.
3. Using Python, locate all child assets of each duplicate item, and map them to the equivalent assets that will be retained.
4. Unpublish duplicates, delete duplicates, and merge usage stats to the kept resource on Hyacinth and Academic Commons. (Fedora preservation copy is retained but hidden.)
5. Using a Python script sending requests to the DataCite REST API, update DataCite DOI registry to redirect DOIs of each duplicate (i.e., deleted) resource to the corresponding retained resource.

**GOALS**

1. Remove duplicates from users' view to improve search efficiency and to avoid confusion.
2. Merge existing usage stats of duplicate resources to prevent data loss.
3. Update DataCite's information so that necessary DOIs would redirect users to the most current resources.
4. Develop duplicates remediation workflow for the future.

**CHALLENGES**

1. Deciding which item to keep and which to delete is a bit subjective. Different items have different metadata strengths and weaknesses. There could not be a simple rubric to assess items.
2. Child assets may be slightly different, e.g., PDFs may be generated in different ways.
3. The review process doesn’t catch every duplicate, since we are matching on metadata fields, not checksums.

**Outcome**

- 689 duplicate records were deleted and/or merged across multiple systems.
- 713 DOIs were updated.
- A new standing workflow for remediation was created.