

# Analysis of user-supplied metadata in a health sciences institutional repository

Joelen Pastva, MA, MLIS

Galter Health Sciences Library & Learning Center, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA

Launched in October, 2015 by the Galter Health Sciences Library, the **DigitalHub** repository is designed to capture and preserve the scholarly outputs of the Northwestern medical community. A major motivation to deposit in the repository is the possibility of improved citations and discovery of resources, especially for non-traditional materials such as poster presentations and teaching resources that are not consistently made publicly accessible.

## Problem Statement

A lack of user-supplied descriptive metadata hampers the discovery of resources in DigitalHub. Minimal metadata is required upon deposit, though many optional descriptive fields are also available, some using auto-complete suggestions from controlled vocabularies to encourage the consistent and detailed entry of descriptive information. In an effort to improve the discoverability of resources deposited in DigitalHub, Galter Library provides metadata enhancement services for all publicly accessible items. However, the library was curious to evaluate how users were approaching available metadata fields and accompanying instructions prior to the performance of enhancement operations.

## Methods

A CSV export was made of the descriptive metadata in DigitalHub from July 2015 to April 2018. Records previously enhanced by librarians, or records deposited by library staff were excluded from primary consideration. This resulted in a small but representative set of 72 records. The metadata was then evaluated for completeness, choice of terms for resource type, use of controlled vocabulary fields, and any areas that indicated a clear misunderstanding of the intended use of the metadata field.

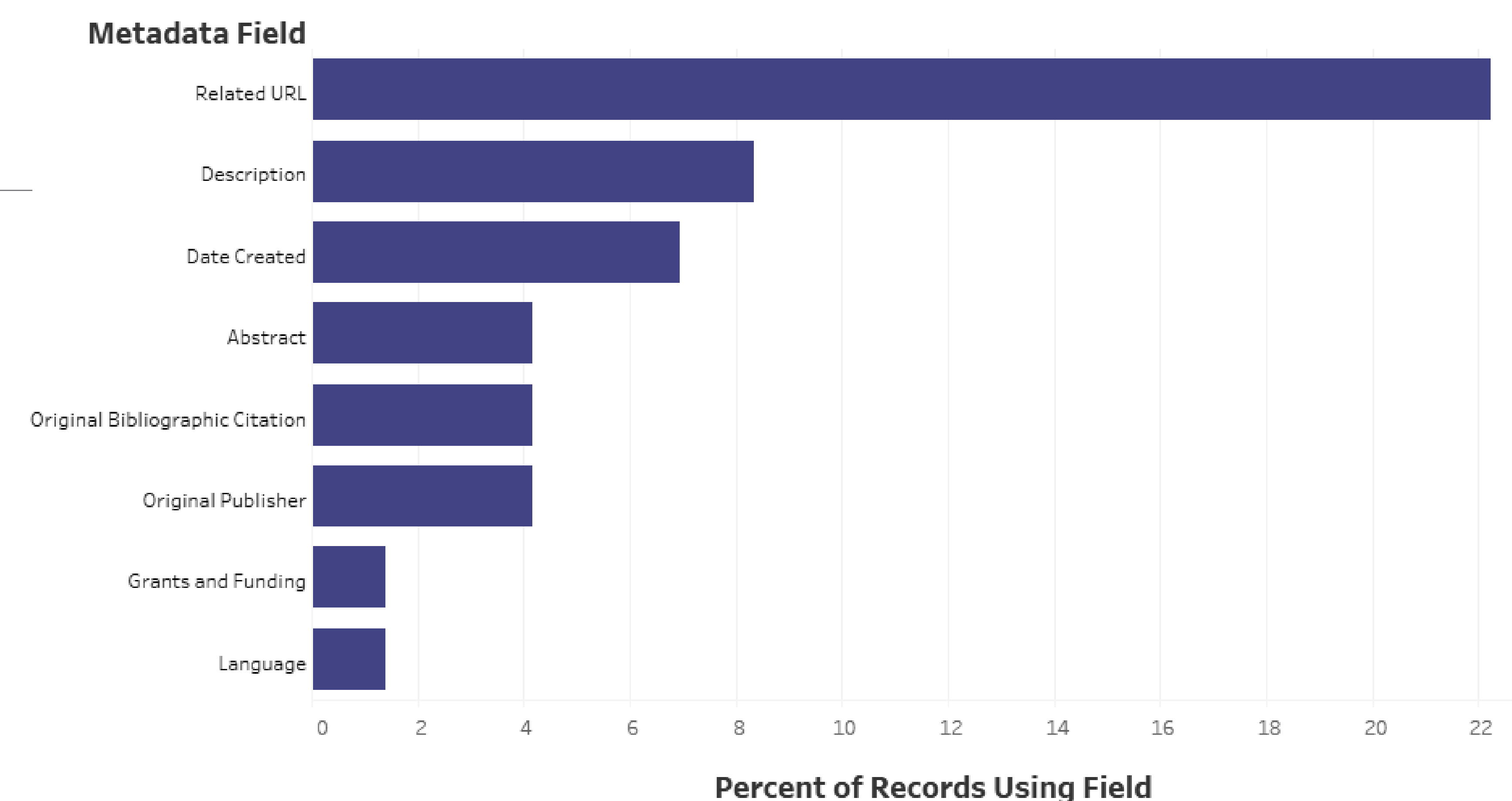
## Results

- 55% of all optional metadata fields were left blank, including all auto-complete subjects
- Depositors were comfortable using Keyword tags, with over half supplying 2 or more
- 25% of optional fields were underutilized
- 93% of depositors did not supply Date, often considered core information
- 38% of depositors used the resource filename, a system-supplied default value, for Title, another core field
- Only 17 unique Resource Type values were used, out of 159 options
- 57% of depositors selected "All rights reserved," the most restrictive Rights option, and a required field; CC BY-NC-ND, the most restrictive Creative Commons license, was the 2<sup>nd</sup> most popular option, at 17% of deposits

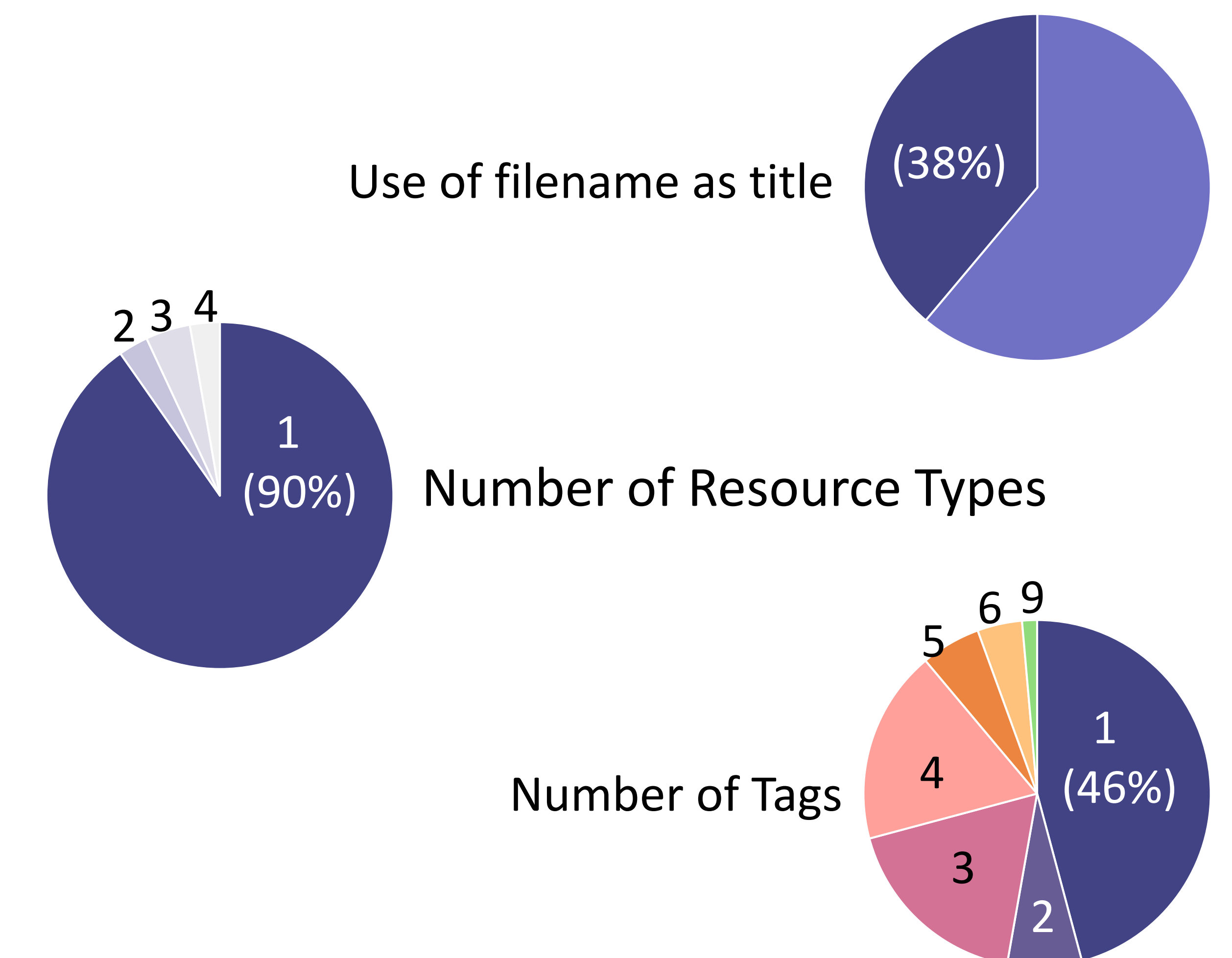
### Fields Left Blank

- Contributor
- Digital Publisher
- Original Identifier
- Subject: MeSH
- Subject: LCSH
- Subject: Geographic
- Subject: Name
- Location
- Page Number
- Acknowledgements
- Private Note

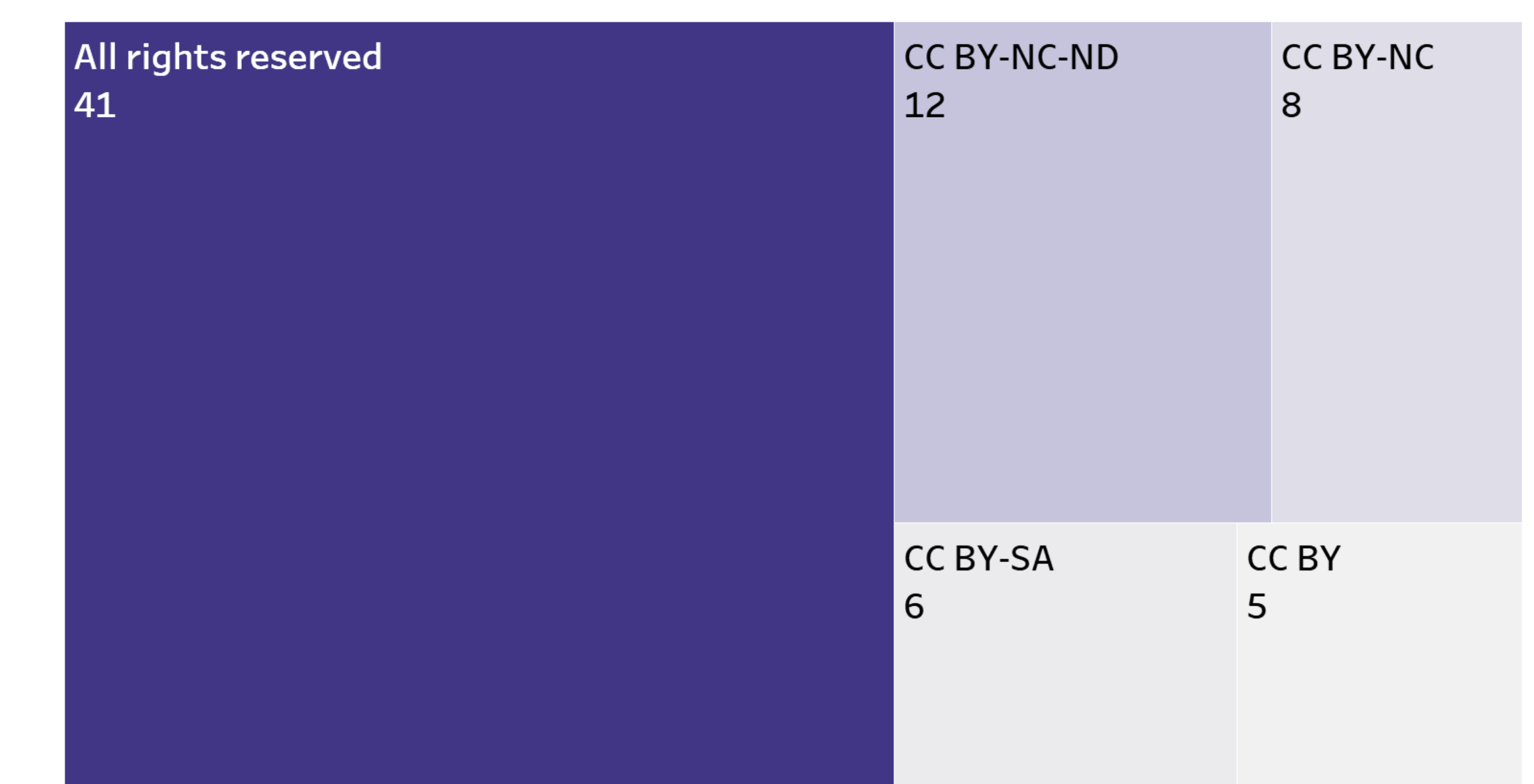
### Use of Optional Fields



### Resource Types



### Rights



## Conclusions

The lack of complete metadata supplied by depositors indicates the continued need for library metadata enhancement for improved discovery, and opportunities to pre-populate standardized fields and investigate automated metadata extraction solutions. Unused Resource Type options should be pared down, and the need for inclusion of underutilized fields should be reevaluated. As an aside, the tendency of users to select restrictive Rights terms indicates the possible need for improved outreach efforts regarding open access scholarship. Ultimately, the goal is to provide a usable repository interface to encourage depositors to supply an optimal amount of descriptive metadata up front, and to encourage continued repository use in the future. These results should be of interest to repository developers and managers who rely on users to supply descriptive metadata, especially for health sciences disciplines.