DPLA Rights Statement Validation

[DRAFT]

Background

DPLA's rights information may be found in two fields—*dcRights* and *edmRights*. At least one of these is required for all items. *edmRights* contains the standardized rights statement, which means it must be a valid statement, given in the form of a URI, found in the rightsstatements.org set or a Creative Commons license. Using a rights statement or license allows rights information to be machine-readable and standardized. This standardization opens up new possibilities for DPLA, such as enabling the user to filter the aggregation on rights.

Current process

Some basic normalization and validation mechanisms exist to ensure every value ingested *resembles* a valid rights statement or license URI. For example, only domains of *rightsstatements.org* or *creativecommons.org* are accepted.

Current normalizations:

- Replace `/page/` with `/vocab/`
- Replace `https://` with `http://`
- Remove `www.`
- Drop all query parameters (e.g. `?` and everything after it)
- Trailing slashes added, where missing (*new*)
-

Current validation checks:

- If more than one value is mapped
- If value is not a URI
- If domain of URI is not creativecommons.org or rightsstatements.org

These normalizations correct some of the common errors. Any value which requires normalization due to these errors will be reported as a warning to the provider so they can be fixed, but the normalized value *is* ingested.

The validations filter out obvious invalid values. They prevent a provider from erroneously using words to describe rights in this field, or putting a non-standard URI, such as a local terms of use web page link. A record can still be accepted if an edmRights value fails the validation, as long as there is another valid edmRights value provided, or a dcRights value.

Problem

In order to fully realize the potential of standardized right statements, edmRights values need to be controlled. Despite the normalization and validation measures described above, there is currently no guarantee that the final accepted value is actually valid according to the definition of the edmRights field.

Many invalid values still pass the current validation because they follow the correct format for a valid URI, even though there may be a typo or some other issue within the string. These include errors such as "http://rightsstatements.org/" (correct domain, but points to no specific statement), http://rightsstatements.org/vocab/NoC-US/1.0/q" (random characters included in URI), and "http://creativecommons.org/licenses/by-nc/4.0/legalcode" (links to legal document instead of canonical license URI). Additionally, a wrong statement/license code, version number, or jurisdiction could be used and accepted, as long as the format of the URI is correct.

As a result, while edmRights is intended to contain only standardized rights information, in practice, it can still contain many non-standard values.

Solution

DPLA is introducing new validation measures that would ensure only a valid Creative Commons or rightsstatements.org URI is ingested as a value of the edmRights field. This will be enforced by rejecting records—no longer just warning—where the provided edmRights value, after any normalization, is not one of the finite number of valid URIs (*even if a dcRights value is provided*). In order to accomplish this, DPLA will maintain an authoritative listing of all URIs considered valid for this field, and all edmRights values will be checked against it. Rejected records must be fixed in order to be ingested. In summary:

1. Current normalization procedures will remain in place
2. If, after any normalizations, a value in edmRights does not match any of DPLA's accepted valid rights URIs, it will be rejected
3. Any record with multiple values for edmRights will be rejected

Valid edmRights URIs

The following are 59 URIs that would be considered valid rights statements or Creative Commons licenses.

> *Note: This list contains only unported (or Generic or International) versions of Creative Commons licenses; DPLA's full list of URIs includes several hundred additional URIs, which constitute all the valid ported versions of the licenses below. The use of ported*

*Creative Commons licenses is very rare in DPLA, and no non-US ports are currently used in the aggregation, so this has been excluded for readability, but the full list of all valid edmRights URIs **can be found here**.*

- http://rightsstatements.org/vocab/InC/1.0/
- http://rightsstatements.org/vocab/InC-OW-EU/1.0/
- http://rightsstatements.org/vocab/InC-EDU/1.0/
- http://rightsstatements.org/vocab/InC-NC/1.0/
- http://rightsstatements.org/vocab/InC-RUU/1.0/
- http://rightsstatements.org/vocab/NoC-CR/1.0/
- http://rightsstatements.org/vocab/NoC-NC/1.0/
- http://rightsstatements.org/vocab/NoC-OKLR/1.0/
- http://rightsstatements.org/vocab/NoC-US/1.0/
- http://rightsstatements.org/vocab/CNE/1.0/
- http://rightsstatements.org/vocab/UND/1.0/
- http://rightsstatements.org/vocab/NKC/1.0/
- http://creativecommons.org/licenses/MIT/
- http://creativecommons.org/licenses/BSD/
- http://creativecommons.org/licenses/GPL/2.0/
- http://creativecommons.org/licenses/LGPL/2.1/
- http://creativecommons.org/licenses/by/1.0/
- http://creativecommons.org/licenses/by/2.0/
- http://creativecommons.org/licenses/by/2.5/
- http://creativecommons.org/licenses/by/3.0/
- http://creativecommons.org/licenses/by/4.0/
- http://creativecommons.org/licenses/by-sa/1.0/
- http://creativecommons.org/licenses/by-sa/2.0/
- http://creativecommons.org/licenses/by-sa/2.5/
- http://creativecommons.org/licenses/by-sa/3.0/
- http://creativecommons.org/licenses/by-sa/4.0/
- http://creativecommons.org/licenses/by-nc/1.0/
- http://creativecommons.org/licenses/by-nc/2.0/
- http://creativecommons.org/licenses/by-nc/2.5/
- http://creativecommons.org/licenses/by-nc/3.0/
- http://creativecommons.org/licenses/by-nc/4.0/
- http://creativecommons.org/licenses/by-nc-sa/1.0/
- http://creativecommons.org/licenses/by-nc-sa/2.0/
- http://creativecommons.org/licenses/by-nc-sa/2.5/
- http://creativecommons.org/licenses/by-nc-sa/3.0/
- http://creativecommons.org/licenses/by-nc-sa/4.0/
- http://creativecommons.org/licenses/by-nd/1.0/
- http://creativecommons.org/licenses/by-nd/2.0/
- http://creativecommons.org/licenses/by-nd/2.5/

- http://creativecommons.org/licenses/by-nd/3.0/
- http://creativecommons.org/licenses/by-nd/4.0/
- http://creativecommons.org/licenses/by-nc-nd/2.0/
- http://creativecommons.org/licenses/by-nc-nd/2.5/
- http://creativecommons.org/licenses/by-nc-nd/3.0/
- http://creativecommons.org/licenses/by-nc-nd/4.0/
- http://creativecommons.org/licenses/by-nd-nc/1.0/
- http://creativecommons.org/licenses/by-nd-nc/2.0/
- http://creativecommons.org/licenses/nc/1.0/
- http://creativecommons.org/licenses/nc/2.0/
- http://creativecommons.org/licenses/nc-sa/1.0/
- http://creativecommons.org/licenses/nc-sa/2.0/
- http://creativecommons.org/licenses/nd/1.0/
- http://creativecommons.org/licenses/nd/2.0/
- http://creativecommons.org/licenses/sa/1.0/
- http://creativecommons.org/licenses/sa/2.0/
- http://creativecommons.org/licenses/sampling/1.0/
- http://creativecommons.org/licenses/publicdomain/
- http://creativecommons.org/publicdomain/mark/1.0/
- http://creativecommons.org/publicdomain/zero/1.0/