



Preparing Your Educational Resources for DiscoverEd

What is this document?

This is a basic guide to increasing the discoverability of online educational resources by preparing them for inclusion into search engines like DiscoverEd that utilize structured data.

Who is it for?

This guide is targeted at people or institutions interested in making their digitally published educational resources more discoverable. This document contains technical language and sample XHTML and RDFa.



This document is licensed using a [Creative Commons Attribution 3.0 Unported License](http://creativecommons.org/licenses/by/3.0/).
Attribute to CC Learn with a link to <http://learn.creativecommons.org>.

Note that Creative Commons Corporation is not a law firm and does not provide legal services. Distribution of this document does not create an attorney-client relationship. Creative Commons provides this information on an "as-is" basis. Creative Commons makes no warranties regarding the information provided, and disclaims liability for damages resulting from its use. The information provided below is not exhaustive—it may not cover important issues that may affect you. We recommend that you familiarize yourself with our licenses before applying them. Please read more at: creativecommons.org/about/licenses/.

Introduction

DiscoverEd¹ is an experimental project from Creative Commons intended to explore how structured data² may be used to enhance the search experience. Metadata about the resources, including the license and subject information available, are exposed in the search result set. We are particularly interested in open educational resources (OER) and are collaborating with other open education projects to improve search and discovery capabilities for OER, using DiscoverEd and other available tools. For in-depth details, read the our white paper³ that describes the goals and design of DiscoverEd.

This document is meant to be a quick checklist for maximizing the discoverability of your resources in DiscoverEd and similarly designed search engines. Not all of these steps are necessary for inclusion into DiscoverEd. For example, structured data are not technically required for resources to be included in search results, but without them users of the search engine will be provided with very little information about your resources.

¹ <http://discovered.creativecommons.org/search/>

² The difference between structured and unstructured data is like the difference between a table of information and a free-form paragraph of text: you know which bits of text correspond to which fields. A free-form paragraph of text requires a human mind to be interpreted, while a table of information can be automatically manipulated by helpful software.

³ "Enhanced Search for Educational Resources: A Perspective and a Prototype from ccLearn," <http://learn.creativecommons.org/wp-content/uploads/2009/07/discovered-paper-17-july-2009.pdf>

I. Resource Feed

DiscoverEd uses resource feeds to direct its resource crawl. In order to index your educational resources, DiscoverEd will need the URL to an RSS or Atom feed that is limited to your educational resources. It is not likely that a site is composed entirely of educational materials, instead consisting of “About” pages, links to staff profiles, and so on, in addition to the educational resources. An index of educational resources should be composed of *only* actual educational materials, thereby reducing or eliminating clutter that typically accompanies web-scale queries.

DiscoverEd consumes the feeds for each site that has been listed for inclusion. Your feed essentially provides a URL “road map” of your resources, which can then be used to run a directed crawl of the resources you curate. In other words, the crawler knows where the relevant resources are located because you, the curator, have pointed at them directly using the feed.

Many curatorial sites already have feed functionality (RSS or Atom) or support the Open Archive Initiative's Protocol for Metadata Harvesting (OAI-PMH)⁴. The MIT OpenCourseWare site, for example, allows you to subscribe to a feed of the courses, which means that you can get an update every time a course is added, deleted, or changed. This type of feed also usually contains a list of the URLs for every course already on the site. Both feeds and OAI-PMH also provide a convenient method of polling, allowing the system to periodically check for new resources. Once a feed is set up, the DiscoverEd system can be kept up to date with minimal oversight.

Adding Your Feed to DiscoverEd

Once you have RSS/Atom feed functionality on your site delimited to educational resources, you can get it included into DiscoverEd in one of two ways:

1. Email cclearn-info@creativecommons.org with the feed URL.
2. Add your organization and feed into ODEPO.

ODEPO is a database within OpenEd, which is a site for the open education community hosted by Creative Commons⁵. OpenEd is a source for information or guidance on open education; a way to discover other organizations, projects, people, educational resources; a source for data about the OER movement; and a way to connect, discuss, and coordinate. ODEPO (Open Database of Educational Projects and Organizations) is an open, collaborative collection of pages on OpenEd, each of which provides information about educational sites and projects that are accessible on the Internet⁶.

By adding your site to ODEPO you increase its visibility and at the same time enable search and discovery of your resources or resource pages by filling in the Resource Feed URL property box. In this field, you would put the URL to your feed of resources so that Creative Commons, or any other project, can see the list of feeds in the database and add them to a search index.

To add your organization to ODEPO, follow these steps:

⁴ OAI-PMH is currently a supported protocol within DiscoverEd, but the recommended method for metadata markup is [X]HTML+RDFa, and is discussed in Section II.

⁵ <http://opened.creativecommons.org>

⁶ More information can be found on the ODEPO page: <http://opened.creativecommons.org/ODEPO>

1. Create or log in to your account on OpenEd (the “Log in / create account” link is at the top of every page).
2. Navigate to the ODEPO page: <http://opened.creativecommons.org/ODEPO>
3. Use the “Browse ODEPO” link or the search box at the top of the page to check if your organization already exists in the database.
4. *If it already exists:* On the page for your site you can click the “Edit with form” link at the top of the page to easily edit/add/correct any information.
5. *If it doesn't exist:* Click “Add an organization” on the main page of ODEPO. Fill in as many of the form fields as possible, including the Resource Feed URL field which will indicate the URL to your feed of educational resources.

Please note that if you do not fill in contact information for your page, it will be difficult for us to contact you with questions, comments, or updates about your feed in DiscoverEd. Consider adding in that information, or if that is not an option, sending an email to cclearn-info@creativecommons.org.

II. Resource Metadata

Once you have located the URL to a feed that is limited to your educational resources, a good next step to increasing their discoverability would be to provide metadata about those resources. We recommend XHTML+RDFa for metadata encoding and transport.

As a curator, you have certain goals for the resources you curate. Generally, you want curated resources to be as easy to find as possible. Core to this goal is enabling machines to detect and interpret metadata about the resources, such as title, language, or licensing terms, in a way that is interoperable with as many detection and interpretation methods as possible. Interoperability here means not only that different programs can read particular metadata properties, but also that the vocabularies themselves, which are sets of related properties, can evolve and be extended. It is also important that potential extensions be backward compatible: existing tools should not be disrupted when new properties are added. If possible, existing tools should even be able to handle basic aspects of new properties. This is precisely the kind of "interoperability of meaning" that RDF⁷ is designed to support.

Therefore, the ideal method for metadata encoding/transport is XHTML+RDFa⁸. We believe this has the broadest possible exposure for current and future software agents. For more information as to why we recommend and require RDFa for metadata transport, see the ccREL W3C specification and our white paper⁹. For technical information on XHTML and RDFa, see the W3C RDFa Primer¹⁰.

This section outlines some of the RDFa metadata Creative Commons is collecting for the DiscoverEd project and gives some examples of using RDFa in XHTML documents. These metadata are extracted from the document at crawl time. While our metadata store may include additional metadata information from resources, these fields are exposed by default in the search results:

- **Title**
- **Summary**

7 http://en.wikipedia.org/wiki/Resource_Description_Framework

8 There are more reasons to use XHTML+RDFa for metadata transport, including the co-location of metadata with documents. For more information, see “Enhanced Search for Educational Resources: A Perspective and a Prototype from ccLearn,” <http://learn.creativecommons.org/wp-content/uploads/2009/07/discovered-paper-17-july-2009.pdf>

9 <http://learn.creativecommons.org/wp-content/uploads/2009/07/discovered-paper-17-july-2009.pdf>

10 <http://www.w3.org/TR/xhtml1-rdfa-primer/>

- **License**
- **Education level**
- **Language**
- **Subject**

Note about RDFa Vocabularies

Notice that each metadata label is preceded by a prefix of either `dc` or `xhtml`. In the RDFa specification, these are indicators of which vocabulary defines the properties, or metadata terms. We recommend the Dublin Core vocabulary for the majority of properties because of its widespread adoption.¹¹ For `license`, we recommend using the `xhtml` namespace because it's built in to the XHTML specification and is equivalent to other definitions of the property.¹²

Title (`dc:title`)

A brief descriptive title for the resource.

Summary (`dc:description`)

A relatively short summary or synopsis of the resource.

License (`xhtml:license`)

The stable URL of the work's license; e.g., <http://creativecommons.org/licenses/by/3.0/>. If you are using Creative Commons licenses, we also recommend following the ccREL specification for identifying further CC license metadata. For more information, see the W3C ccREL publication.¹³

Education level (`dc:educationLevel`)

What grade(s) or age-level(s) this material is suitable for. The education level should indicate all levels (student ages) for which the resource is deemed appropriate. Though we accept any descriptions that seem appropriate to you, please consider using one of the following schemas:

- `primary`, `secondary`, `tertiary`, `adult`;
- `K, 1, 2, 3, . . . , 20` (where the number refers to the actual grade-level).

You may include equivalent terms as well by specifying more than one value for `DCT:educationLevel`. For example, you might include a separate `DCT:educationLevel` tag for `9, 10, and secondary`.

Language (`dc:language`)

The language(s) of the referenced resource (not of your site). When specifying the language for a resource, the value should be specified as described by RFC-4646.¹⁴ For example, `en` for English. To distinguish English (United States) from English (United Kingdom), the language

¹¹ <http://purl.org/dc/terms/>

¹² If you define your default namespace as `xmlns="http://www.w3.org/1999/xhtml/"`, you do not need to use an `xhtml:` prefix for the license property. See the examples in Section III and the W3C RDFa Primer at <http://www.w3.org/TR/xhtml-rdfa-primer/>.

¹³ <http://www.w3.org/Submission/ccREL/>

¹⁴ <http://tools.ietf.org/html/rfc4646>

would be specified as `en-US` and `en-GB`, respectively.

Subject (`dc:subject`)

The subject(s) of the resource; e.g., `mathematics`. The subject refers to the actual content in the resource; i.e., what the resource is about. For many resources, more than one subject will be necessary; in these cases, simply specify multiple subject elements. Ideally you should try to limit the contents of the subject to only those subjects that are objectively reflective of the entire resource. Other types of categories (opinions, metrics, etc.) may have other vocabularies available which are more appropriate.

III. [X]HTML + RDFa Examples

The following is an example of how a resource at `http://ocw.example.org/math/101` could be annotated with machine-readable metadata, including license and attribution information. This is our preferred manner for encoding this information as it exposes the metadata to a much wider range of clients.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
"http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<html xmlns="http://www.w3.org/1999/xhtml/"
      xmlns:dc="http://purl.org/dc/terms/"
      xmlns:cc="http://creativecommons.org/ns#">
  <head>
    <title>OER Site</title>
  </head>

  <body>
    <h1 property="dc:title">Math 101</h1>
    <h2>by <a href="http://example.org/~johnq" property="dc:author
cc:attributionName" rel="cc:attributionURL">John Q.
Public</a></h2>
    <p property="dc:description">Basic mathematics for 5th
graders</p>
    <p>Subjects: <span property="dc:subject">Math</span></p>
    <p>Grade level: <span property="dc:educationLevel">5</span></p>
    <p>Language: <span property="dc:language"
content="en">English</span></p>
    <p>License: <a href="http://creativecommons.org/by/3.0/"
rel="license">Attribution 3.0</a></p>

    <p>Lorem ipsum, etc, etc.</p>

  </body>
</html>
```

If a site aggregates resources such that the metadata appear on a page other than the actual resource, the about attribute can be used to indicate that the metadata are about a different resource. For example, the following page could be published at <http://commons.oer.example.org/math/101> and still refer to the same resource as the previous example:

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN"
"http://www.w3.org/MarkUp/DTD/xhtml-rdfa-1.dtd">
<html xmlns="http://www.w3.org/1999/xhtml/"
      xmlns:dc="http://purl.org/dc/terms/">
  <head>
    <title>OER Site</title>
  </head>

  <body>
    <div about="http://ocw.example.org/math/101">
      <h1 property="dc:title">Math 101</h1>
      <h2>by <span property="dc:author">John Q.
Public</span></h2>
      <p property="dc:description">Basic mathematics for 5th
graders</p>
      <p>Subjects: <span property="dc:subject">Math</span></p>
      <p>Grade level: <span
property="dc:educationLevel">5</span></p>
      <p>Language: <span property="dc:language"
content="en">English</span></p>
      <p>License: <a href="http://creativecommons.org/by/3.0/"
rel="license">Attribution 3.0</a></p>
    </div>

    <p>Lorem ipsum, etc, etc.</p>

  </body>
</html>
```

IV. Additional information

If you have further questions, please browse the DiscoverEd FAQ (http://wiki.creativecommons.org/DiscoverEd_FAQ) or contact Creative Commons (cclearn-info@creativecommons.org) for clarification.

You can also review a more detailed explanation of DiscoverEd in the white-paper entitled “Enhanced Search for Educational Resources: A Perspective and a Prototype from ccLearn.”¹⁵

15 <http://learn.creativecommons.org/wp-content/uploads/2009/07/discovered-paper-17-july-2009.pdf>