Search and Discovery: OER's Open Loop

Abstract

Open educational resources (OER) promise increased access, participation, quality, and relevance, in addition to cost reduction. These seemingly fantastic promises are based on the supposition that educators and learners will discover existing resources, improve them, and share the results, resulting in a virtuous cycle of improvement and re-use. By anecdotal metrics, existing web scale search is not working for OER. This situation impairs the cycle underlying the promise of OER, endangering long term growth and sustainability. While the scope of the problem is vast, targeted improvements in areas of curation, indexing, and data exchange can improve the situation, and create opportunities for further scale. I explore the way the system is currently inadequate, discuss areas for targeted improvement, and describe a prototype system built to test these ideas. I conclude with suggestions for further exploration and development.

Keywords: linked data, semantic web, search, metadata harvesting, OER quality, distributed curation

Nathan R. Yergler <nathan@yergler.net>, Creative Commons, 171 2nd Street, Suite 300, San Francisco, CA 94105, United States of America

Nathan R. Yergler is Chief Technology Officer at Creative Commons. Mr. Yergler joined Creative Commons as a software engineer in 2004, and has been responsible for developing the technical infrastructure which supports Creative Commons licenses, as well as tools for users of the licenses. Previously, he held a faculty position at Canterbury School. While at Canterbury, Mr. Yergler developed learning technology to connect teachers, students, and parents. He pioneered the use of Python in Canterbury School's Computer Science courses, developing both introductory and advanced elective curricula. Mr. Yergler holds a B.S. in Computer Science from Purdue University.

Introduction

The phrase "Open Educational Resource" (OER) was first introduced in 2002 at the UNESCO Forum on the Impact of Open Courseware for Higher Education In Developing Countries ("What is OER?," n.d.). Since its introduction, the phrase "OER" has come to encompass more than simply the availability for use without royalty. OER has come to describe the environment in which the resource is developed. This environment includes sharing materials created, using or adapting materials created by others, and sharing back modifications so that others may benefit ("Why OER?," n.d.). The success of this environment requires that educators can find materials to work with, they can make changes to the materials found, and they can publish the modified resource in a manner which makes it available to others. When all of these assumptions are true, a self-reinforcing cycle exists which allows the best materials to be discovered and continuously improved within communities of interest.

However, these assumptions are not all true today, and the cycle of discovery, improvement, and publication is impeded at every level. Educators are unable to find resources which are appropriate for their use, and when they do find them, they are often unable to adapt and improve them, due to either format, permissions, or licensing issues. While more research is needed to

establish baseline metrics, it is clear from conversations within the OER community that both educators and publishers view discovery as a hurdle to adoption. When educators do find resources and improve them, the opportunities for contributing back may be limited (i.e., by institutional policy), or the republished resource may not be discoverable by downstream users. Search and discovery underlies all of these issues.

An Ideal Search Tool

An ideal search tool for educators would return materials that are relevant, usable, and from a diversity of sources. Web scale search tools generally accomplish relevance through the use of full text indexing and link analysis. While some are adding support for structured data, the present level of adoption is limited to specific use cases and vocabularies. The reliance on a full text index and link analysis casts a broad net when searching, but impedes the process of discovery by including resources which are not necessarily educational. Increasing the relevance of the resources returned by a search can minimize the time educators need to spend exploring irrelevant resources.

The usability of a resource refers to several characteristics, including but not limited to its copyright status. For example, a resource released under a Creative Commons Attribution 3.0 License is very usable from a copyright perspective, but if the resource is only provided in Portable Document Format (PDF), it is less usable (editable) than one provided as Open Document Text (ODT). If the format requires proprietary, commercial tools for editing, it is less usable in a broad sense than one which can be edited using a variety of tools (i.e., ODT, which allows users to choose between open source and commercial tools for editing). The usability of resources impacts every stage of the cycle. Discovery takes longer if an educator needs to manually explore whether resources can be adapted for their classroom use, or edited with the tools available to them. Improvement may require specific software tools, or not be possible at all. Finally, publication of improved resources may not be permitted by the license. When looking for educational resources, an ideal search tool would provide easy filters for format and license information, allowing educators to choose resources which they can adapt for their own needs, and ideally, re-share.

Finally, a search tool which only provides results from a single site or repository is less useful than one which provides access to the wealth of OER sites available. The development of the OER ecosystem resembles the development of early data networks which eventually became the internet. Educators are asked to join multiple networks and sites to publish content there, and the ability to “connect the dots” between resources on different platforms is limited. An ideal search tool could address this by aggregating information from multiple sites and multiple authorities (curators), providing users with a single view on a large pool of OER which can then be explored and dissected.

Areas for Targeted Improvement

Looking at the description of an ideal search tool (one which provides results that are relevant, usable, and from multiple sources), we can begin to see how web scale is presently inadequate. While it excels at providing information from multiple sources, it does so at the expense of relevance and usability. There are two problems that must be addressed to improve current OER search tools: the size of the search pool (what resources are relevant), and the ability to filter by resource properties (i.e., license, subject, etc), which is also referred to as faceted search.

Curation

The present situation for OER mirrors the situation when Creative Commons launched its licenses in 2002. Creative Commons licenses are decentralized: there is no centralized database of licensed works, and no registration is required to use them. Creative Commons provides tools which generate our suggested marking, but ultimately authors and publishers are responsible for marking works with a CC license. Like OER, Creative Commons licenses suggest a cycle of re-use: creators make their work available, and other creators can find materials they can re-use. As the licenses became more widely used, questions about how to find Creative Commons licensed works increased. What was needed was an approach to search that limited the size of the search pool (to only licensed works), and added the ability to filter within that pool by the specific license permissions (i.e., those which allow derivative works, or commercial use).

Creative Commons addressed this issue by building a prototype search tool based on Nutch (http://nutch.apache.org), now a project of the Apache Software Foundation. Nutch provides the basic tools needed to develop a search engine, including a crawler and document processing and indexing support. Creative Commons' prototype indexed resources with a CC license, and added the ability to restrict searches by license type. The use of an existing open source platform allowed Creative Commons to more rapidly develop the prototype, and to demonstrate a viable approach to indexing licensed materials. It is worth noting that Creative Commons' search tool was eventually decommissioned, as search vendors saw the value of providing support for Creative Commons licenses in their core offering.

Creative Commons was able to limit the set of resources to search by using the Creative Commons license metadata to identify resources as members of the set or not. Unfortunately no similar mark exists for open educational resources, significantly because there are standing questions about what qualifies a resource as educational, as well as what qualifies it as "open".

One way to limit the set of resources is to adopt a curatorial model, which allows individuals or organizations to specify a set of resources they believe are educational. These resources may also meet some additional criteria, such as having passed a review for quality or relevance to a particular domain. A curatorial approach leverages the nascent OER ecosystem by allowing domain experts to focus on their particular area of expertise and pushing the need to normalize data into an infrastructural layer.

Organizations and individuals are already acting as distributed curators, although they may not consider their work as such. OER publishers, such as university open courseware (OCW) platforms, are acting as de facto curators. Aggregators which identify resources and add metadata or other value (such as the website OER Commons), are acting as more formal curators, developing an index of OER and allowing their community to comment on and annotate it. Leveraging this curation process fully means that resources identified by a curator are indexed, and that users may exclude specific curators or limit their search to a subset of curators. A tool which operates in this manner would allow users to search across a wide diversity of sites, as well as offer the ability to discover new communities that may be relevant to their area of interest.

Indexing

Providing access to specific properties of resources through the search index may also offer dramatic improvements to the search utility. Many OER publishing platforms allow authors to add metadata about their work, such as educational level, subject area, and language. As mentioned previously, existing curators are also adding or updating metadata about resources. This information may be indexed by a web scale search platform, but is usually simply considered as additional text. Allowing users to search by a specific property (i.e., education level) allows much more precise refinement. An improved OER search tool should offer users the ability to refine and filter searches based on metadata provided by the creator, or by another curator.

Provenance is an important issue to consider when determining how to index metadata. In order to maximize flexibility for users, metadata will need to be indexed in a manner which allows the exclusion of, or limitation to, specific curators. A naive approach which does not store the source of metadata will only offer incremental improvements over existing systems.

Metadata Exchange

Adding structured data to works (i.e., RDFa + [X]HTML) provides a structure which allows emergent tools and applications to be built with the data in ways not previously expected (Abelson, Adida, Linksvayer, & Yergler, 2008). While the ideal scenario is one which relies solely on linked data, there are many incumbent platforms which do not support linked data, and are unlikely to adopt it without a clear benefit. For this reason, different approaches to the exchange of data between sites will be required to fully utilize resource metadata from different curators and communities.

At a meeting of organizations interested in OER search and discovery in July 2009, participants agreed that search and discovery tools could be improved without end to end agreement about format and schema of metadata. The recommendation from this meeting (Duval & Yergler, 2010) suggests some baseline practices for publishers to adopt which will enable tools to build upon their work. An improved OER search tool should leverage the existing behavior of publishers and users, without requiring the adoption of specific technologies. By leveraging existing behavior, tools can demonstrate utility and provide guidance for developing standards and practices by consensus.

A Prototype System

In 2008 Creative Commons began developing a search prototype focused on OER and on testing the feasibility of these approaches. This prototype, DiscoverEd (http://wiki.creativecommons.org/DiscoverEd), is also based on Apache Nutch, and attempts to address the shortcomings of existing search tools (Bissell, Park, Yergler, & Linksvayer, 2009). DiscoverEd addresses both of the identified shortcomings: limiting the pool of resources to be searched and providing faceted search, and incorporates improvements in all three targeted areas. The result is a search platform which can be adapted to a variety of domains, and which provides users an improved ability to find resources which are relevant, usable, and from a diversity of sources.

DiscoverEd utilizes a distributed curation process to address the issue of limiting the set of resources to search. The list of resources from all curators is used to direct a crawl for traditional

full text analysis, providing a baseline search experience for resources without additional metadata. The curator (or curators) of each resource is displayed in the search results. A user may choose to limit their search to specific curators, or exclude one or more curators from a search in order to find resources most relevant to their needs.

DiscoverEd's curatorial process is distributed because it assumes curators will be publishing their selections (and possibly metadata) on their own sites, and DiscoverEd will ingest them. This is in contrast to requiring curators or publishers to deposit or register materials with a central authority. By adopting a distributed process, DiscoverEd encourages curators to take ownership of their work, and allows other applications to be built using the data, without permission or mediation.

In order to support curation, DiscoverEd adds an additional step to the typical crawl-index process, aggregation (Figure 1). This step polls curators for new resources, and aggregates the metadata about them in an RDF store using Jena (http://openjena.org). When Nutch crawls the resources, additional structured data (RDFa) may be extracted from the resource as well. The index generated by Nutch includes all of the known information about each resource, including curator provided information and information from the resource itself.

Curators can provide their list of resources to DiscoverEd in several ways. DiscoverEd has the ability to consume Atom and RSS feeds describing resources, harvest resources and metadata from Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH, a protocol implemented by many repository platforms which enables clients to harvest metadata about resources), and can discover additional feeds through the use of Outline Processor Markup Language (OPML, a format often used to describe a list of feeds). DiscoverEd's architecture utilizes extensions to support different interchange formats, which allows for the new formats to be added without impact on other parts of the application.

The adoption of curation also allows DiscoverEd to test improvements to indexing. Curators may simply identify resources, or they may also provide additional metadata about resources. This metadata is combined with structured data found in the resource (i.e., RDFa in the resource), and is searchable through DiscoverEd's web interface, allowing users to further refine their search results. DiscoverEd accommodates varying descriptions and classifications of resources by displaying all the information found, and allowing users to further refine their search by curator or other property. For example, different curators might identify the same resource as educational, but have differing perspectives on the subject or education level. If a user learns that specific curators' perspectives match their own, search results can be limited to those curators, excluding metadata and resources from others.

Areas for Further Research

DiscoverEd demonstrates how the overall search experience may be improved with targeted improvements. It does this by leveraging the existing behavior of publishers and the OER ecosystem. DiscoverEd also provides a platform for additional testing and experimentation, which is necessary to determine if these solutions improve OER search at scale. Additional curation and publishing of linked data that describes resources will encourage the development of additional tools which leverage this information. Based on experience to date, there are several areas which require further exploration.

While DiscoverEd focuses on leveraging existing technology and tools to improve OER search, scientifically rigorous research about educators' search habits and success rates will enable more thorough evaluation of success. Such research will establish baseline metrics regarding efficacy of web scale search. The creation of a testing suite/protocol for measuring efficacy of experimental search tools could be an additional benefit of completing this research.

DiscoverEd currently makes no attempt to normalize or rationalize metadata from different curators. Operating DiscoverEd at scale may reveal that this leads to fractured search results where two curators have used similar, but not identical terms. One approach to addressing this may be the application of domain specific thesauruses, which would allow indexing by the curator provided terms, as well as synonyms. Such an approach has the advantage that it does not require publishers or curators to change their existing behavior. However, a successful experiment should also attempt to draw conclusions and provide feedback to curators so that they can see emergent behavior and possibly reach consensus on how to label specific terms.

While current curation models largely center around identifying existing resources and optionally adding additional metadata, this is not the only model for curation. Curators may also work directly with creators to review, vet, and ensure the quality of their work. In this scenario, it is mutually beneficial for creators to indicate that their work has passed review: it provides the both parties with additional credibility, and may increase adoption and reuse of the curated works. The curator, however, may be understandably concerned about misappropriation of any badge or mark used.

Creative Commons developed technology for describing copyright registrations in 2008-2009 as part of the CC Network project (Yergler, 2009). The CC Network model does not rely on a central authority; rather, it utilizes reciprocal assertions about a work's status. The adaptation of this work to support quality and review marks would provide a flexible model for stronger curation, as well as additional linked data about works.

DiscoverEd currently relies on a polling model: the DiscoverEd site administrator needs to execute an aggregation and crawl, which will find new resources and add them to the index. Protocols like PubSubHubbub (PuSH) (Fitzpatrick, Slatkin, & Atkins, 2010) describe how feeds can be augmented with push notifications. To fully utilize PuSH, curators would need to ping a hub when they update their content. However, by supporting PuSH, curators could ensure that aggregators and search tools are as up-to-date as possible. The development of a prototype to test this approach should include implementation with a publication/curation platform, as well as in DiscoverEd.

Finally, DiscoverEd provides a search tool which exposes structured data and curation to users. Additional, complementary tools can help increase the impact and adoption. Tools such as validators, structured data generators, and tools which help users publish information about their source works would all complement an enhanced search tool.

Conclusion

An ideal OER search tool will provide results which are relevant, usable, and from a diversity of sources. Such a tool would help close the loop of discovery, improvement, and publication, allowing open educational resources to fulfill their promise and continue to scale. DiscoverEd demonstrates how these can be achieved through targeted improvements to indexing, and the

addition of curation. While further development is needed, it is clear that improvements to search and discovery can help open educational resources fulfill their promise.

Acknowledgements

Creative Commons is grateful for the generous support of organizations including the Center for the Public Domain, Google, the John D. and Catherine T. MacArthur Foundation, the Mozilla Foundation, Omidyar Network, Red Hat, and the William and Flora Hewlett Foundation, as well as members of the public. Creative Commons' work on open education search and discovery has been supported by the William and Flora Hewlett Foundation and the Open Society Institute. Thanks to Ahrash Bissell, Mike Linksvayer, and Timothy Vollmer for providing review of this paper.

Bibliography

Abelson, H., Adida, B., Linksvayer, M., & Yergler, N. (2008, May 1). ccREL: The Creative Commons Rights Expression Language. Retrieved September 12, 2010, from http://www.w3.org/Submission/ccREL/

Bissell, A., Park, J., Yergler, N., & Linksvayer, M. (2009). *Enhanced Search for Educational Resources - A Perspective and a Prototype from ccLearn*. Retrieved from http://learn.creativecommons.org/wp-content/uploads/2009/07/discovered-paper-17-july-2009.pdf

Duval, E., & Yergler, N. (2010, January 28). Towards a Global Infrastructure For Sharing Learning Resources. Retrieved September 12, 2010, from http://wiki.creativecommons.org/Towards_a_Global_Infrastructure_For_Sharing_Learning_Resources

Fitzpatrick, B., Slatkin, B., & Atkins, M. (2010, February 8). Draft: PubSubHubbub Core 0.3 -- Working Draft. Retrieved September 12, 2010, from http://pubsubhubbub.googlecode.com/svn/trunk/pubsubhubbub-core-0.3.html

What is OER? (n.d.). *OER Africa*. Retrieved September 11, 2010, from http://www.oerafrica.org/understandingoer/UnderstandingOER/WhatisOER/tabid/1097/Default.aspx

Why OER? (n.d.). *OER Commons Wiki*. Retrieved September 11, 2010, from http://wiki.oercommons.org/mediawiki/index.php/Why_OER%3F

Yergler, N. (2009, May 12). Describing Work Registrations. Retrieved September 12, 2010, from http://labs.creativecommons.org/~nathan/oscri/describing-registrations.html

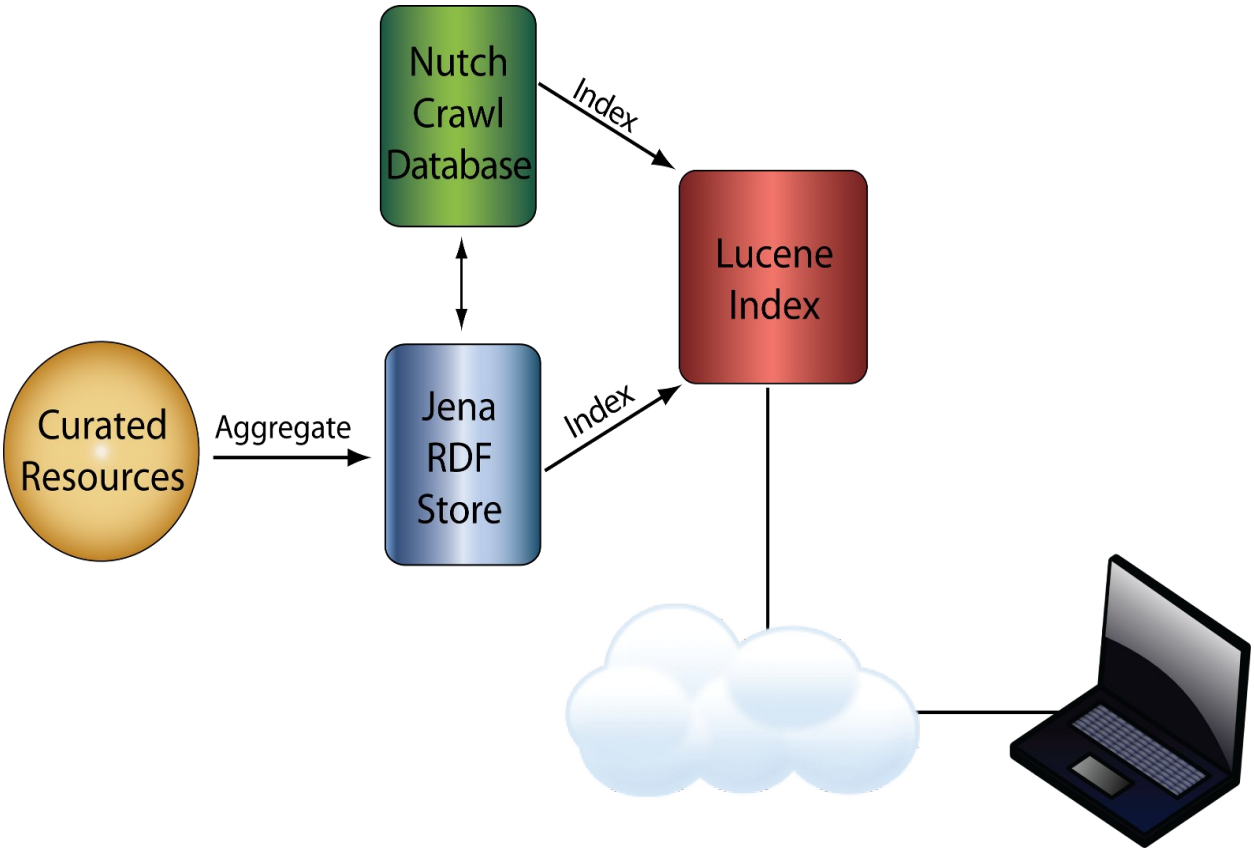

Figure 1: DiscoverEd system architecture.