ccLearn

About this report:

This report was researched and written by ccLearn, comprised in part of Ahrash Bissell (Executive Director) and Jane Park (Research Assistant and Communications Coordinator), and Creative Commons, comprised in part of Nathan Yergler (CTO) and Mike Linksvayer (VP). Many people contributed their time and expertise to this report, spanning early conceptual phases to final edits; in particular, we would like to recognize Ben Adida and Hal Abelson for their contributions. We gratefully acknowledge the feedback and insights of other members of the Creative Commons staff, Google colleagues, and other participants from throughout the OER movement. We would especially like to thank the William and Flora Hewlett Foundation for providing support for this research and activities to follow.

This report is available for download and distribution in several different formats. Please visit http://learn.creativecommons.org/productions/ for all versions and additional details.

ccLearn

Creative Commons

171 Second St, Ste 300

San Francisco, CA 94105

cclearn-info@creativecommons.org

Table of Contents

Executive Summary

Users of search tools who seek educational materials on the Internet are typically presented with either a web-scale search (e.g., Google or Yahoo) or a specialized, site-specific tool. The specialized search tools often rely upon custom data fields, such as user-entered ratings, to provide additional value. As currently designed, these systems are generally too labor-intensive to manage and scale up beyond a single site or set of resources.

However, custom (or structured) data of some form is necessary if search outcomes for educational materials are to be improved. For example, design criteria and evaluative metrics are crucial attributes for educational resources, and these currently require human labeling and verification. Thus, one challenge is to design a search tool that capitalizes on available structured data (also called metadata) but is not crippled if the data are missing. This information should be amenable to repurposing by anyone, which means that it must be archived in a manner that can be discovered and leveraged easily.

In this paper, we describe the extent to which DiscoverEd, a prototype developed by ccLearn, meets the design challenge of a *scalable*, *enhanced search* platform for educational resources. We then explore some of the key challenges regarding enhanced search for topic-specific Internet resources generally. We conclude by illustrating some possible future developments and third-party enhancements to the DiscoverEd prototype.

Background

The hurdle for those who seek educational resources on the Internet is not a lack of materials, but the difficulty of discovery of *appropriate* and *desired* materials. The tool often used to discover these resources is a search engine. The success of any search engine depends on search and ranking algorithms that return web sites that are relevant to what the user wants. Most popular search engines use an index of the text and links found on pages to return results. This works extremely well for most searches for general information. However, educators are often interested in specific types of materials or materials that have certain *attributes*, such as the types of audiences for which the materials were designed, the amount of time it takes to apply a lesson, or different state-education standards that the materials are designed to meet. Searches for materials with these attributes are often suboptimal for several reasons, including:

•   There is usually a smaller audience for targeted resources, and therefore a smaller audience of users publishing links to the resources. This leads to a smaller dataset with which to establish authority (ranking), leading in turn to the relevant, desired results being "lost in the crowd"[1].

---

1   Saeid Asadi and Hamid R. Jamali, "Shifts in Search Engine Development: A Review of Past, Present and Future Trends in Research on Search Engines," Webology 1, no. 2 (December 2004), www.webology.ir/2004/v1n2/a6.html. Accessed 16 March 2009.

- Some attributes of resources are ill-suited to the full-text search models used by most general-purpose search engines. For example, a lesson in American History for students in the ninth grade satisfying California State Standard X may not contain all of those queried words or phrases in the actual text of the website or resource itself. As a result, a text-only search for "Ninth Grade American History Lessons" may not yield the most relevant resources available.
- Educators tend to distrust materials that do not appear to be authoritative. Thus, even though educators are interested in web-scale search, they still tend to rely on trusted sites and sources as arbiters of authority for the majority of their materials.

While most web-scale search engines rely on computer algorithms to extract meaning from written text, more targeted search applications can take advantage of *structured data* to provide more flexible refinements and targeted search – otherwise known as *enhanced search*. Structured data is information which has a "label" attached that tells software exactly what it means. For example, a camera review website might attach the label "megapixels" to the value "10". This allows users to search for specific attributes of a camera and easily filter out the results that meet their criteria. A site that only used text-based search would need to ensure that everyone – authors and users both – knew to refer to a camera's resolution using a single, agreed upon piece of text. For example, searching for "10MP" when "10 megapixels" was the site standard could lead to no or incorrect results. The use of structured data, which comprise information *about* the resource as opposed to the resource itself, can avert this problem. In the case of our hypothetical camera review site, this information is created *in addition to* the actual resource (the review) and requires some manual intervention. Methods of automatically deriving structured data from resources are continually being developed, but in the education space most structured data are still associated with resources manually through human effort and interpretation.

Unfortunately, manually creating structured information about every resource available is an expensive and time-consuming proposition. This fact means that many potentially valuable resources are associated with little or no structured data. Moreover, those structured data that exist can be quite variable in content and specifications. Despite these challenges, we still maintain that *any* structured data are likely to enhance the discoverability of appropriate and desired materials; therefore, a search tool that queries multiple stores of educational resources must somehow capitalize on available data while remaining robust to their variability. This background is sufficient to understand some of the design principles for DiscoverEd; however, we further discuss some of the challenges and promising developments for widespread and effective use of structured data (for any area of interest) later in this paper.

Enhanced Search for OER

Even if completely and properly specified, the presence of structured data only partially solves the challenges to enhanced search and discovery of educational resources. This

section examines key issues ccLearn encountered while building a working prototype of an education-specific search tool that leverages structured data. In each case, we also describe the solutions we chose and the manner in which those solutions were incorporated into the search prototype.

Finding, evaluating, and archiving educational resources on the web

The first issue when developing a specialized search system is to decide which resources to include. One of the most valuable elements of any good archive is the fact that someone took the time to verify that the information deserves to be included. This is a task that the major search organizations have not incorporated and which has been done to some degree by those creating third party structured data for resources in silos. However, this process is time-consuming and costly, so a scalable discovery tool will have to leverage the expertise of the broader community. Unfortunately, allowing "just anyone" to contribute to the archive can be rife with quality-control and political issues, especially for sensitive topics. Our solution was to build the archive at the level of "curators" (Figure 1). A curating organization (or perhaps person) is responsible for ensuring that all of the content they have curated meets whatever standards they have set for themselves. That organization accepts both blame and credit for the quality of their contributions. Tying the curator to the resources also enables users to better control the type of archive they are interested in – if there are materials from a source a user does not trust, then those curators can be excluded from the search. Conversely, if there are materials in one or more sites that a user prefers, then the search can be limited to those curators alone. It is important to emphasize that educators and educational organizations already expend substantial time and effort evaluating and curating educational resources.

Figure 1: Different curators of OER will include different types of structured data, if any at all. But any curator might have resources that one would like to examine.
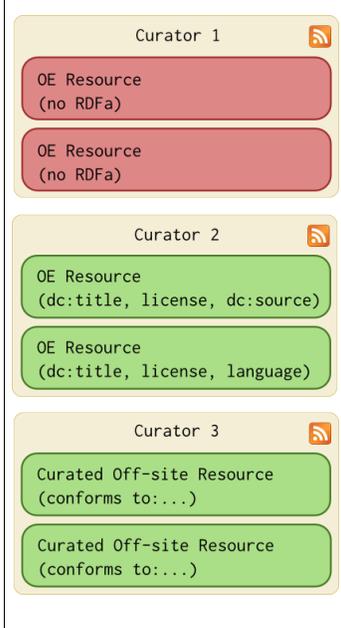
There are many positive outcomes of this approach. First, it removes the burden of evaluating materials from anyone other than the curating organizations themselves. This is the task that hosts of online resources already perform; this model simply capitalizes on the work they already do. Second, most curators are interested in specific subsets of materials, since their expertise can be effectively leveraged without requiring them to adhere to generalized (and perhaps inappropriate) standards. Third, management of the search index can occur at the level of the curating organizations, rather than at the level of individual resources, greatly reducing the burden to the host of the index. Fourth, organizations currently authoring structured data about resources on other sites can also become involved and free their data from its "silo".

Curators have an interest in enhancing and protecting their reputations, so it is likely that accuracy and quality-control will be high priorities for curatorial activities. Furthermore, the explicit rationale for being a curator is to enhance discovery and use of those educational resources to improve learning. Therefore, it is in the curating organization's interest to engage in and promote the addition of structured data to enhance the discoverability of their resources.

In short, many people and organizations need to be involved in the task of finding, evaluating, and describing educational resources. As already mentioned, many involved in education already do these things, so it was incumbent on us to design a search tool to leverage that work and encourage it to continue.

Delimiting the scope of the index for educational resources

Even if participation is limited to self-identified or recruited curators, the problem of generating an index that consists of nothing but *bona fide* educational resources remains. For example, one could perform a full-text crawl on all of the pages in the MIT OpenCourseWare[2] (OCW) site (directing the crawler to the home page and to all pages within), but of course a substantial fraction of the site isn't composed of educational materials at all, consisting instead of "About" pages, links to staff profiles, and so on. Ideally, our index should be composed of *only* actual educational materials, thereby reducing or eliminating the irrelevant clutter that typically results from web-scale queries.

The solution to this problem presented itself once we adopted the curator model. Most curatorial sites have feeds (RSS or Atom) or support the Open Archive Initiative's Protocol for Metadata Harvesting (OAI-PMH)[3]. The MIT OCW 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. The feed should also contain a list of the URLs for every course on the site.

We designed a system that consumes the feeds for each curatorial site that has been integrated into the search prototype. The feeds essentially provide a "road map" of URLs, which we then use to run a *directed crawl* of the resources within each site. In other words, the crawler knows where the relevant resources are located because the curator has pointed at them directly using the feed. The crawler is a piece of software which retrieves each resource and adds its contents to an index. This index can be used to return relevant results for search terms.

Both feeds and OAI-PMH also provide a convenient method of polling, allowing the system to periodically check for new resources. Once the feed is set up, the system can be kept up to
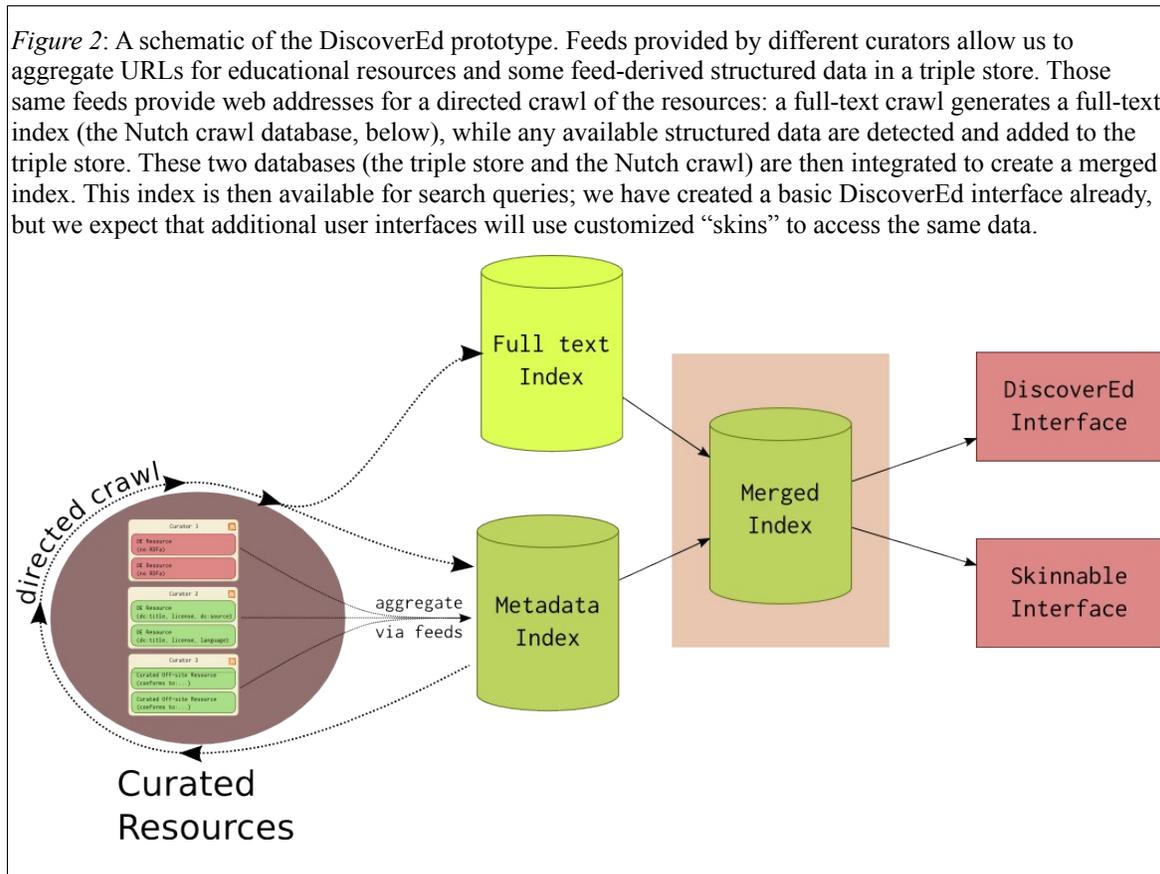
---

2   ocw.mit.edu/OcwWeb/web/home/home/index.htm. Accessed 16 March 2009.
3   www.openarchives.org/pmh/. Accessed Feb 25, 2009.

date with minimal oversight. One can set up a regular crawling schedule to add resources to the index and crawl new or updated resources, and the feeds provide all of the information one needs. Clearly this greatly reduces the management burden.

Combining full-text and metadata indexes

After assembling the list of resources for inclusion and crawling them, we have two indexes: one index of structured data collected from feeds, OAI-PMH, and scraped by the crawler, and another index of full text generated by the crawler. These two indexes are then merged to create an integrated index. Any search query capitalizes on this joint index, so the result-set is informed by both the full-text index and any available structured data, which can be used to allow for more refined search queries. Note that this architecture allows for the inclusion of resources that have no machine-readable data. Similarly, one can also include resources that have structured data but no text (e.g., images, videos, etc). The design of the entire system can be seen in Figure 2.

Figure 2: A schematic of the DiscoverEd prototype. Feeds provided by different curators allow us to aggregate URLs for educational resources and some feed-derived structured data in a triple store. Those same feeds provide web addresses for a directed crawl of the resources: a full-text crawl generates a full-text index (the Nutch crawl database, below), while any available structured data are detected and added to the triple store. These two databases (the triple store and the Nutch crawl) are then integrated to create a merged index. This index is then available for search queries; we have created a basic DiscoverEd interface already, but we expect that additional user interfaces will use customized "skins" to access the same data.

Our Search Prototype: DiscoverEd

We tested out these ideas by building a working prototype. Because this is a prototype, we intentionally built a sparse user-interface with a minimum of refinements. Different decisions and design points are illustrated below.

Educational sites indexed

We are currently only indexing the following curated sites. Hardware and computing requirements increase with increasing numbers of curators, but there is otherwise no particular limit on the number of curators that could be included.

Figure 3: A partial screen shot of the DiscoverEd prototype. The five refinement fields are visible, including those fields with no structured data available.
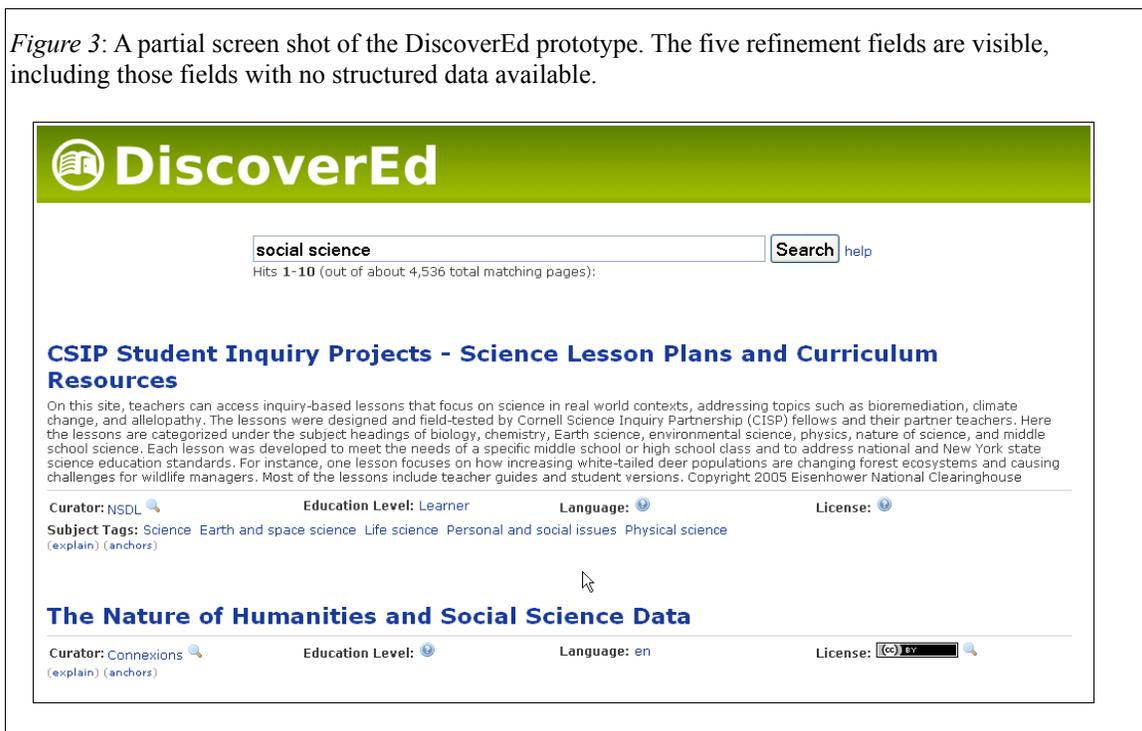
- Connexions (http://cnx.org)
- National Science Digital Library (http://nsdl.org)
- OER Commons (http://oercommons.org)
- OpenCourseWare Consortium institutions (e.g., MIT, Johns Hopkins School of Public Health, the Open University UK, etc) (http://ocwconsortium.org)

Refinements for the user interface of DiscoverEd

Refinements are the structured data fields that are displayed along with the query results and allow users of the prototype to easily "drill down" into a given set of results with more

refined queries. Search refinements of nearly any type are possible, provided the structured data are provided in a machine-readable format. In our case, we chose to reveal the following refinements:

- License   What is the copyright license of the resource?
- Curator   Which organization is curating this resource? Note that there may be more than one curator since many of these resources are aggregated by secondary curators (e.g., OER Commons) who then enhance the value of those resources by adding more structured data. Our user-interface will display every curator that is associated with any given resource.
- Education level   What grade-level or age is targeted by this resource?
- Language   What is the language of the resource?
- Subject area   Any keywords or tags associated with the resource.

We chose these refinements (Figure 3) because we feel that they are key attributes that are of value to anyone interested in open educational resources. As content creators publish structured data alongside their resources, their information will be included in the index although it may not be revealed in the prototype. Each organization that wishes to host its own search portal can alter the user interface to suit and can pre-set certain filters that favor

their audiences. Note too that the refinements will only show information when the information is available. If the information does not exist, or is not provided in a machine-readable format, then we cannot display it. We chose to leave all of the refinement fields in place for every result, whether or not there is a value to display. Our purpose was to give visible evidence that the information *could* be displayed if the curating organization gave it to us in a usable format.

Working with the results of a DiscoverEd query

If a user clicks on a resource of interest, they are then linked to the canonical URL of that resource on the Internet. If the curator is also the host for that resource, then the user will be sent to the resource on the curator's site. If the curator simply links to a resource of interest, then the user will be sent to the site where that resource actually resides. This is one way of resolving where a user is sent when there is more than one curator for a resource[4].The current architecture of the data store allows users to refine results by limiting ("only") or excluding ("not") certain properties (Figure 4). This means that you can pro-actively delimit the results you see based on those curators, refinements, or tags that you prefer. A user can click on a curator name, thereby limiting the results to show only those resources that came from that

---

4   Note that some curators incorrectly report the canonical resource URL; in these situations the result may link to the resource page at the curator's site. In these situations users may also see resources repeated in the search results, with one instance linked to the curator site and another to the actual resource.

curator. Similarly, you can limit by language, license, grade-level, or any keyword. A user may also choose to exclude a certain language, grade-level or keyword[5].

Figure 4: Another partial screen shot of the DiscoverEd prototype, showing a more refined search (delimited by a specific curator, OER Commons).
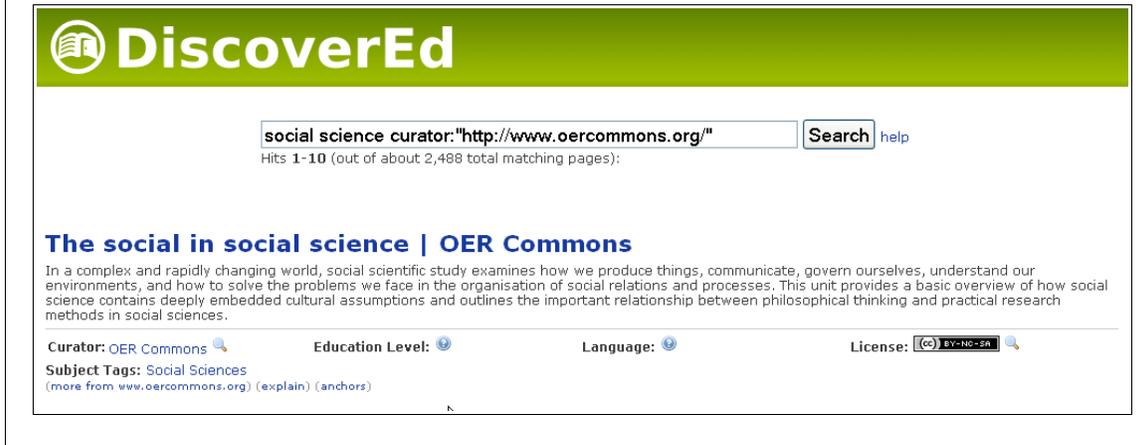
Connecting to the curatorial sites

One or more curators are listed with every resource that is displayed, which obviously indicates where that resource came from. In the cases where the same resource is curated by several curators, the structured data from all of the curators are pooled and all available values are displayed. Users can refine the search by clicking on the magnifying glass next to the name of the curator. This will re-run the query returning only results from that curator. If a user clicks on the name of the curator itself, they will go to the website (home page) of the curator. *As a result, this search tool has the potential to greatly enhance the site traffic for any of the curators.* Users can discover educational resources from all across the Internet, giving a greater sense of the diversity and depth of the available materials in the emerging global learning commons.

Syndicating DiscoverEd queries and results

Just as we consume feeds, we provide OpenSearch[6] feed capabilities for any query. This means that a user can submit queries to the search engine in order to discover resources of interest, and if a particular query brings up useful results, that person can subscribe to a feed of that particular query which will then deliver automated results to a web-based reader or any other software that consumes feeds. There is no limit to the number of different feeds to

---

5   Exclusions can be applied to a query using the minus sign ("-") in front of the search term; for example, adding "-tag:biology" to a query excludes resources tagged with "biology".

6   www.opensearch.org/ Accessed Feb 25, 2009.

11

which a person can subscribe. We believe that the search prototype we have developed is an excellent precursor to "personal search" designs (see below), but in the meantime, the feeds provide a form of personalization and automation that is very easy to use.

Ranking query results

A search application's ranking algorithm determines the order of resources in the search results. We are currently using the default ranking algorithm for the underlying search engine, Nutch[7]. There are a number of different ways that the ranking can be customized or optimized, depending on our needs or the needs of any other organizations deploying the prototype code. Again, our intention was not to build the "best" search engine, but rather to build a search prototype that enters a previously unavailable part of the search landscape, and which is amenable to customization for all of the specialized communities and needs that might exist. Future iterations of the search prototype would probably benefit from some customization of this algorithm, and we encourage any organization that decides to host their own copies of the index to experiment and share the results of their efforts.

Structured Data – Towards Decentralization and Interoperability

As currently designed, feeds and OAI-PMH allow the DiscoverEd software to collect some structured data, but these data delivery mechanisms also suffer from some significant disadvantages. For example, while OAI-PMH has proved itself to be a capable protocol for harvesting and exchanging structured data about resources, it requires the deployment of specialized software to serve it and requires specialized software that understands the protocol's operation. For the purpose of this project, these hurdles are both easily cleared: many institutions utilize content management systems that provide OAI-PMH as a feature, and software libraries exist that allow developers to connect to them. However, when applying criteria that ensure at least minimal interoperability, OAI-PMH falls short. Specifically, the use of OAI-PMH requires tools to be aware of other, secondary locations for the structured, machine-readable information about a resource.

We believe that structured data are more useful when widely exposed and linked. Furthermore, such data are more likely to be provided once they are parsed by useful and popular web-based tools, such as search engines. A key benefit pertains to auto-discovery: if you've got the URL, then you have also got access to the structured data over HTTP and by following HTML links. There is no longer any need for a specialized protocol to benefit from the virtually linked structured data.

Structured data can gain the widest exposure and opportunities for linking when published in [X]HTML, visible to both people as well as software. Creative Commons addressed this issue as part of our work with licenses and identified the following as important principles for structured data in HTML documents:
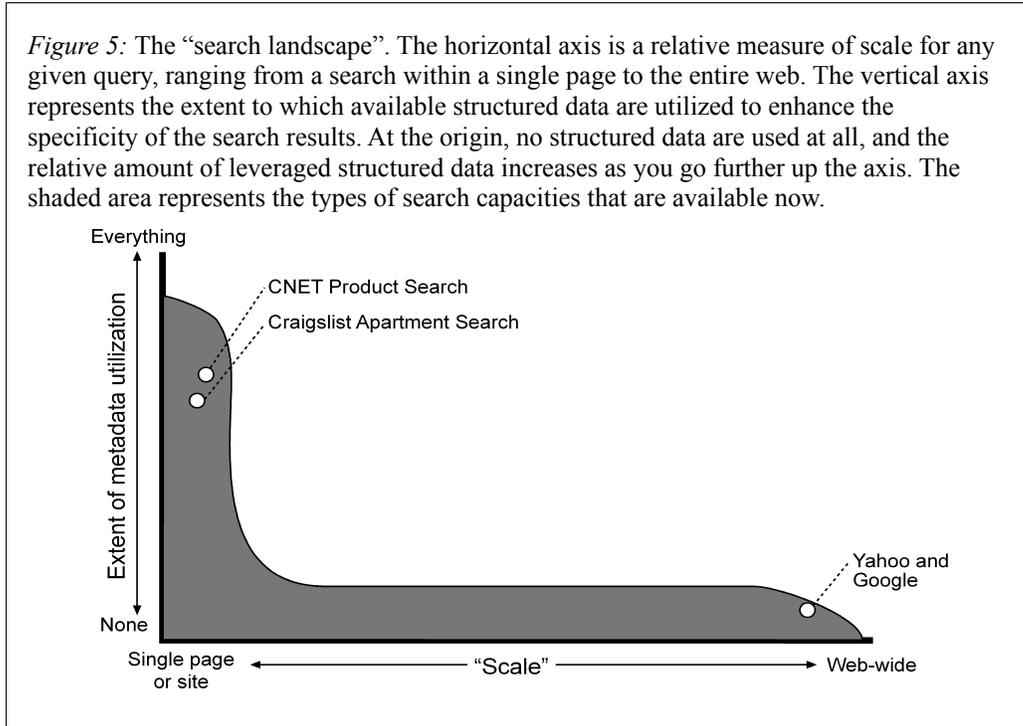
---

7   lucene.apache.org/nutch/ Accessed Feb 25, 2009.

• **Independence and Extensibility**: The means of expressing information in HTML should be (1) independent of any central authority and (2) extensible, i.e., enabling the reuse of existing data models and the addition of new properties by anyone. Adding new properties should not require extensive coordination across communities or approval from a central authority. Tools should not suddenly become obsolete when new properties are added, or when existing properties are applied to new kinds of data sets.

• **Don't Repeat Yourself**: Providing machine-readable structure should not require duplicating data in a separate format. Notably, if the human-readable links or text are changed, a machine processing the page should automatically note this change without the publisher having to update another part of the HTML file to keep it "in sync" with the human-readable portion. This helps reduce the overall load of creating structured data after the fact.

• **Visual Locality**: An HTML page may contain multiple items, for example a dozen photos, each with its own structured data. It should be easy for tools to associate the appropriate structured data with their corresponding visual display.

• **Remix Friendliness**: It should be easy to copy an item from one document and paste it into a new document with all appropriate structured data included. In a world where people constantly remix old content to create new content, copy-and-paste, widgets, and sidebars are crucial elements of the remixable Web.

Through Creative Commons' work with the W3 Consortium, these principles developed into RDFa (Resource Description Framework -in-attributes)[8]. RDFa uses a set of HTML attributes to enable the expression of structured data in HTML documents. Software developed to consume RDFa does not need to be aware of the vocabularies that will be used beforehand in order to extract the information and make it available.

An important improvement on the DiscoverEd prototype would consume RDFa from the resources themselves, reusing the content where possible, instead of requiring curators to publish specialized representations of the structured data.

---

8   http://www.w3.org/TR/xhtml-rdfa-primer/ Accessed June 3, 2009.

Figure 5: The "search landscape". The horizontal axis is a relative measure of scale for any given query, ranging from a search within a single page to the entire web. The vertical axis represents the extent to which available structured data are utilized to enhance the specificity of the search results. At the origin, no structured data are used at all, and the relative amount of leveraged structured data increases as you go further up the axis. The shaded area represents the types of search capacities that are available now.
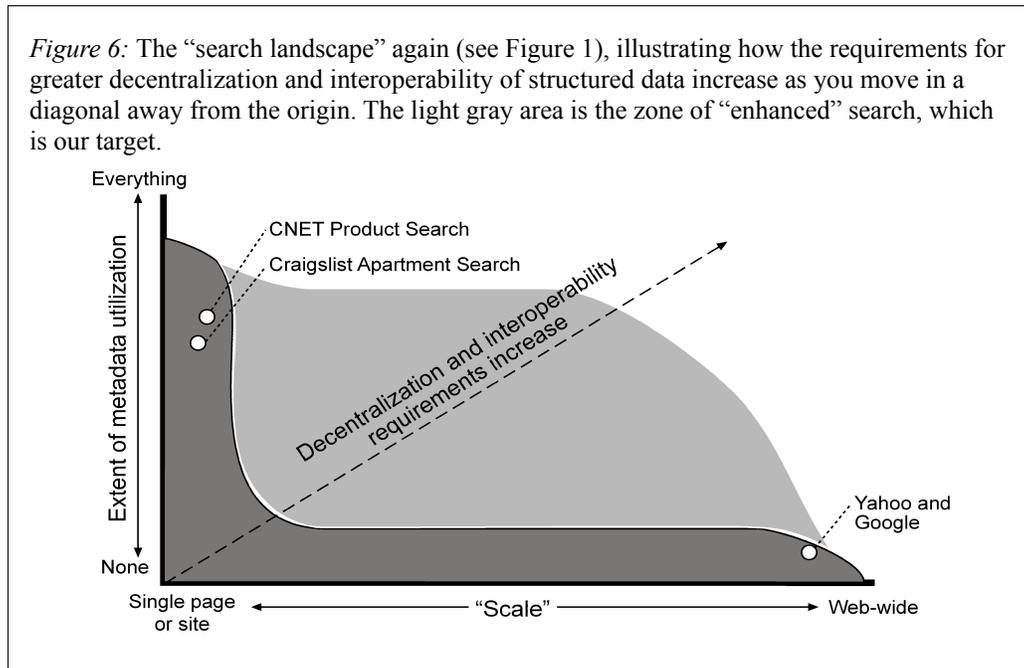

Visualizing the current "search landscape"

The costs and complications associated with publishing structured data have constrained the types of enhanced-search tools that are available. To better understand these constraints, it may help to examine the online search landscape as illustrated in Figure 5. The horizontal axis is a relative measure of *scale* of any given query, ranging from a search within a single page to the entire web. The vertical axis represents the extent to which structured data are utilized to enhance the specificity of the search results. At the origin, no structured data are utilized. The relative amount of leveraged structured data increases as you go further up the vertical axis. The shaded area represents the types of search capacities that are available now. For example, Craigslist Apartment Search is site-specific but leverages a rich body of structured data for ease of discovery, including the location, size, and price of the units. These data had to be input by the people who posted advertisements for apartments, and Craigslist provides the forms and infrastructure for the data to be rendered correctly to apartment hunters. However, the data are highly specific to their application to the Craigslist site, and only limited amounts of data are collected in order to lessen the burden to the person posting the advertisement. Furthermore, the site only gathers data provided by the person listing the apartment. Relevant information that might be provided by other people – opinions from prior tenants, comments about the neighborhood, ratings of the efficiency of the appliances – are generally not available.

14

It should be clear that there is currently a trade-off between the scale of the search and the extent to which structured data can be utilized. This trade-off means that there is a significant proportion of the theoretical search landscape which is not yet available to the general public. Globally, there are many different efforts underway to improve the diversity and accuracy of automatically generated structured data, but many types of structured data do not appear to be amenable to automation. Data about age-targets, implementation times, state standards, impact on learning, licensing rights, and many other crucial attributes are not likely to be automatically generated any time soon[9,10].

The case for decentralized and interoperable structured data

For our purposes, the crucial aspect of structured data is that it allows relevant information to be specified in a *decentralized* yet *interoperable* manner. 'Decentralization' refers to the fact that the data may be provided by a multitude of users from any number of locations on the web. In other words, one cannot presume that all of the structured data about the items of interest will be collected and managed within a single site. When you consider a global phenomenon such as open education, it becomes obvious that all interested learners and educators are not going to contribute their resources and insights to a single location on the web. 'Interoperable' refers to the notion that equivalent types of data, regardless of origin, should be recognized and treated equivalently by the search software. For example, if two

Figure 6: The "search landscape" again (see Figure 1), illustrating how the requirements for greater decentralization and interoperability of structured data increase as you move in a diagonal away from the origin. The light gray area is the zone of "enhanced" search, which is our target.

---

9   Jane Greenberg, Kristina Spurgin, and Abe Crystal, Final Report for the AMeGA (Automatic Metadata Generation Applications) Project (Library of Congress, 2005), ils.unc.edu/mrc/amega.htm, www.loc.gov/catdir/bibcontrol/lc_amega_final_report.pdf. Accessed 23 Mar 2009.

10 Jane Greenberg, "Metadata Extraction and Harvesting: A Comparison of Two Automatic Metadata Generation Applications," Journal of Internet Cataloging 6, no. 4 (2004): 59-62.

different apartments are listed on separate websites, but the neighborhood is the same for both units, then it should be possible to automatically discover this fact and display both units for any query by location. If we revisit the search landscape, shown again in Figure 6, we can see that interoperability and decentralization of relevant structured data must increase in tandem as one goes from the origin towards web-scale, enhanced search.

Returning to our Craigslist example, one can imagine that other data relevant to the apartment query might exist outside of the Craigslist site, such as tenant comments, proximity to neighborhood attractions, repair histories, etc. Craigslist would become unwieldy if it always required (and displayed) such information, but because the data are interoperable, someone who was interested in such data could get benefit from discovering and including such data in a query. Users would benefit even more if they could run searches across multiple apartment-listing sites, where all of the decentralized data were sufficiently interoperable to lend themselves to a single query.

Current trends in structured data adoption

Along with other leaders in this field[11], Google and Yahoo have recognized the value of structured data in web pages and have developed tools to take advantage of this. Yahoo is now indexing RDFa in pages and is making it available through a variety of tools. RDFa is a decentralized, extensible and interoperable way of expressing structured data and is Creative Commons' preferred way of expressing structured data. Yahoo has integrated RDFa into Search Monkey, their open search platform[12].

Search Monkey makes use of RDFa into two ways. The first is the use of structured data to describe objects on the web, such as Flash videos or slideshows. When Yahoo finds information describing these resources they use it to generate a small preview of the resource in search result listings[13]. This provides an incentive for creators to include the information – it enhances their appearance in search results. By using RDFa, Yahoo has clearly indicated they want others to be able to benefit from this information as well.

Search Monkey also allows users to enhance their search results with additional information through small applications they add to their Yahoo account. These applications have access to the RDFa Yahoo extracts from resources and can use this to provide resource-specific information in the search result listing. For example, a Creative Commons application could display the license URL associated with a resource via RDFa.

Google is also using RDFa to provide enhanced search results. Rich Snippets allow authors to annotate pages about people, businesses, products or reviews with structured data that

---

11 Review by Ivan Herman: http://ivan-herman.name/2009/06/19/semtech2009-impression/
12 http://developer.yahoo.com/searchmonkey/
13 http://developer.search.yahoo.com/help/objects/documents

Google can include in search results[14]. For example, a review might include a star rating from 1 to 5 stars; Google will display the stars under the result title, making it clearer to users what sort of resource the result contains.

It is worth noting that once search engines parse structured data as RDFa, it gives further incentive to publish more data, thereby completing a virtuous circle for search enhancement. It is our expectation that these trends will continue and that widespread publication and use of structured data on the Internet will become commonplace in the next few years.

Future directions

The current DiscoverEd software is a prototype. We believe that this tool provides access to an area of the theoretical "search landscape" which has not been previously easily accessible, and hope that people will find it immediately useful. However, ongoing development and maintenance of the tool will require additional collaboration and resources It is our hope that the broader education and technology communities will contribute to this effort.

All of the software code is open source, and ccLearn will continue to populate and maintain the resource index for the foreseeable future. We are going to be evaluating usage and impact of the DiscoverEd tool and examining the extent to which it does or does not meet the needs of the various educational communities who are seeking solutions to identifying relevant resources on the Internet. It may be that the ideas put forth in this white paper lead to alternative approaches that prove more effective; we welcome such developments, and we hope that participants in such projects see fit to share their developments with the education community.

In the meantime, there are a number of clear opportunities for further development of the search prototype that we have already identified. These opportunities vary substantially in the level of technical difficulty (or even feasibility), which we try to clarify in the discussion below. This list is not intended to be exhaustive or prescriptive, but rather to spark further discussion and action among people and organizations that are interested in these issues and are perhaps in a position to pursue some of these suggestions. We welcome commentary and questions from anyone.

Further enhancements to the semantic architecture: tracking provenance

One of the great challenges people face for information flow on the Internet, and not just in education, is maintenance and determination of the provenance of the data[15,16]. Specific to our

---

14 http://www.google.com/support/webmasters/bin/answer.py?answer=99170

15 Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan, "Data Provenance: Some Basic Issues," in *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*, vol. 1974, Lecture Notes in Computer Science (Springer Berlin / Heidelberg, 2000), 87-93, db.cis.upenn.edu/DL/fsttcs.pdf. Accessed 16 March 2009.

16 Martin Szomszor and Luc Moreau, "Recording and Reasoning over Data Provenance in Web and Grid Services," in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, vol. 2888, Lecture Notes in

search prototype, we have a data store that associates resource URLs with specific curators, so we maintain that relationship and display it on the user-interface. But what if information about a given resource comes from more than one place, such as when more than one curator is involved, or if we were to include user-generated tags and reviews which were not published by the resource curator? Currently, the data store simply integrates all of the information to keep it logically coherent, but the origin – or *provenance* – of the archived data is not retained. While indexing the source of the information is a straight-forward solution, integrating it with the underlying search platform has proven challenging using our available resources.

There are many benefits to tracking the provenance of structured data. For example, to revisit the issue of working with the search results, what if users want to search only the keywords provided by a specific curator, excluding all others? Tracking provenance would allow us to perform this sort of search.

Some of the additional suggestions below are either enabled or improved by our ability to track provenance, and we make mention of these advantages where appropriate.

Customization

The prototype is already designed to be highly customizable. However at this time most customizations require a high degree of technical sophistication. It is not difficult to imagine that many customizations could be simplified into a few built-in tools or perhaps a third-party application that eases the process of creating and hosting a localized search portal. We are hoping to showcase the diversity of any customized applications that are developed.

There are two distinct types of customization that are likely to be of greatest interest. First, there is the option to bias the search results in a particular direction. One way to do this, as mentioned previously, is to simply limit the result set according to pre-determined criteria (e.g., a specific subset of refinements). In this way, users at a French-language site might only get results where the resources are in French. Alternatively, it is also possible to change the scoring algorithm used in order to alter the rank-order of the result set. Altering the results in this manner requires more technical sophistication and a greater commitment of hardware resources, but it also allows for highly refined degree of control over the function of the search engine.

Second, the user-interface itself can be customized. In this case, the goal would be to appeal to the particular interests and needs of the users, or to more closely associate the look and feel of the search portal with the host site. Customization of this sort includes everything from color schemes to the arrangement of the information to the different refinements or

---

Computer Science (Springer Berlin / Heidelberg, 2003), 603-620, www.springerlink.com/content/5a4bk24wc47elk7f/. Accessed 16 March 2009.

other data displayed. Easy-to-use tools for customized web sites are already common on the Internet, so we expect that this functionality will be both simple to enable and popular among people and organizations.

Easy-to-use tool for adding third-party metadata

In our opinion, the future of targeted, domain-specific search depends on the existence and improvement of relevant structured data. That assumption is one of the core drivers behind the design of the DiscoverEd prototype. However, as already mentioned, structured data tend to be incomplete and can be costly or time-consuming to add to resources. We intentionally designed the prototype so that structured data are not required, but it remains the case that machine-readable, structured data are highly desirable.

How can we encourage the publication of more (and more relevant) structured data? One decision we made was to reveal when useful data are missing, by leaving the placeholders for every data field in the search results even if there are no data to display. We hope that curating organizations will take it upon themselves to start providing those structured data, or provide them in a machine-readable manner, thereby improving the discoverability of their own resources.

Another tactic is to engage the broader education community in the collective task of adding structured data to the resources they discover. Many curators have community web sites where users can rate resources or add other information, all of which can be very helpful. But the participation rate in such activities is nearly always extremely low. We believe that the lack of participation is due to the fact that the feedback tools only function on the web sites that contain the resources. This arrangement makes no sense, as users cannot be expected to return to a specific site in order to provide feedback about the resources in that archive. Community sites that encourage users to tag resources and provide feedback are already acting as de facto curators. If their user reviews were rendered with RDFa, a DiscoverEd crawl could consume and integrate the additional structured data. Allowing DiscoverEd to consume user-generated data provides additional motivation for users to participate in communities and could drive new participants to the community as well.

Interestingly, the non-standardized (at least in terms of ontologies) user-feedback that one might obtain in this manner is actually one of the great strengths of the project, rather than a concern (as it is usually assumed). The reason is that users who tag resources in ways that make sense to them (due to their culture, context, or whatever), are then enhancing the utility of the DiscoverEd tool for anyone else who shares those same perspectives. In contrast, a one-size-fits-all solution necessarily results in a product that doesn't really work for anyone. With time and use, our scheme may result in a product that works very well for a great diversity of people who are all interested in finding the same basic types of information.

Assuming such a feedback tool is built, there will be substantial outreach and field-building tasks associated with encouraging people to submit the data. However, there are many educational and professional groups that are ideally poised to facilitate this effort, especially considering the value of more relevant structured data to the communities at large.

Some people have questioned the validity and usefulness of community-acquired data. The concern is that non-experts and antagonists may do more harm than good by contributing useless or inaccurate information. As currently architected, this issue is indeed a concern, since there is no way to distinguish the origins of different forms of information about any given resource. Tracking the provenance of structured data resolves this problem. Users would then have the power to include or exclude information based on whom they trust or any other metric. This model should theoretically encourage greater attention to the quality of the structured data since the information will be tightly coupled to the person or organization that supplied it in the first place. If a particular curator is abusing the system, it can be identified and barred from further participation. As a general rule, the solutions we seek to these types of problems strive towards transparency and user-empowerment, rather than exclusion and restrictions on the flow of information. Anonymity has no place in such a system and we believe that anyone who contributes useful information should be willing to stand by that information so that everyone else can evaluate it honestly.

Personal search

Perhaps the ultimate realization of the capacities of this prototype would be its possible application to enabling "personal search." In short, personal search refers to the notion that people can identify particular needs and preferences that they have and the software will automatically adjust the way it functions according to those settings. Limited forms of personal search already exist, so the idea is not new, though it has not perhaps become as fundamental to the operation of the Internet as many have thought.

The DiscoverEd prototype allows for a significant amount of customization, as already described above. If the underlying data store can maintain information provenance, then it will be possible to exercise an extraordinary degree of personal control over the way the search tool functions. It should not be difficult to store personal preferences, which can either be archived online or on a personal computer – the user can decide, depending on the extent to which the search portal is accessed from one place or from various locations. A user can also choose when and whether to save preferences, and of course any preferences could be changed as needed.

Expert-Directed Search

An implementation of personal search opens other opportunities to collaborate and share information. If an "expert" user chooses to restrict her searches to a specific set of curators, or to a specific set of refinements, that information is likely to be of interest and value to other members of the community as well. For example, the preferences designated by an

20

expert in one field can greatly simplify the search for relevant resources for someone who has less expertise in that field. Users could choose whether or not to share such information. We believe that these types of activities will enhance the value of existing expertise in a manner that enables greater recognition of such in the eyes of peers and the public at large. As the sheer quantity of information increases, both on the Internet generally and in any enhanced search index, people will need all the help they can get to sort and evaluate among the materials that are available.

Conclusions

In this paper we provided a perspective on the search landscape with particular focus on domain-specific search for open educational resources. We described the role of structured data and how they can help build a richer, more accurate search experience. We provided an overview of our search prototype, DiscoverEd, and a discussion of possible future directions.

Throughout, the design considerations took into account both the technical considerations and the social contexts in which both creators and users of educational resources are likely to operate. We intentionally designed things to encourage greater awareness of copyright issues and of structured data standards, both of which are part of the core mission for ccLearn. It was important for us to create a tool that capitalized on the hard work and expertise of others, rather than trying to duplicate efforts. All of these issues needed to be resolved in a manner that did not close off access or engagement for anyone, thereby empowering both creators and users of educational resources to work together to simplify and improve the task of resource discovery.

We believe that we have created a compelling prototype, one of several steps towards the realization of *enhanced* search. The fundamental design principles are not restricted to the educational domain, but we believe that education is a perfect testing ground for these and comparable social-technical developments. Depending on the levels of additional support available, we will be promoting this work and engaging with others to help us think about next steps, collaborative opportunities, and communities of practice that can benefit from the tool and improve its function. We are eager to see what innovative ideas emerge from these activities.

Acknowledgements

ccLearn