

System Analysis of Open Directory Project (ODP)

<http://dmoz.org>

Katherine T. Haack

University of Texas School of Information

1385 TKMS – Dr. Turnbull

1 April 2003

Austin, Texas

“The Open Directory Project is the largest, most comprehensive human-edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors.”

-- <http://dmoz.org/about.html>

What is Open Directory Project?

The Open Directory Project (ODP) is a massive endeavor that intends to catalog the world-wide-web through the dedicated service of (presently) over 50,000 volunteer editors. The net continues to grow at unprecedented rates, thus disorganization is rabid as there is no central organization to the web. Profit-based groups have emerged to create search engines and other knowledge management tools for the organization of the web, however these groups cannot keep up with the ever-changing organic nature of the web. Thus, their search engine sites often return search results that are saturated with link rot and topically invalid site recommendations.

In 1998, Rich Skrenta began the volunteer-run GnuHoo (name inspired by GNU freeware group), which would later become the ODP. On June 5, 1998 GnuHoo went live and within two weeks “there were 200 editors, 27,000 sites, 2,000 categories”. Within five months the site garnered the attention of Netscape and was acquired for their Netcenter site. However, the site’s name had now been changed from GnuHoo to NewHoo due to naming conflicts with the freeware group called GNU who was in the process of developing a free UNIX inspired operating system. Soon the name changed one final time, becoming the Open Directory Project (<http://www.laisha.com/zine/odphistory.html>).

Inspired by the Debian Social Contract, which was created by the group who developed the Debian GNU/Linux system, the ODP issued a Social Contract that has seven primary edicts:

1. The Open Directory Will Remain 100% Free
2. We Give Back to the Web Community
3. We Don't Hide Our Official Editorial Policies
4. We Provide an Open Invitation to Join
5. We Encourage a Self-Regulating Community
6. Our Priorities are Our Data Users and the Community
7. Users Not Meeting The Free Use License (*are not permitted*)¹

<http://dmoz.org/socialcontract.html>

The social contract of the ODP is interesting because it openly sets forth what the ODP expects to be to the wider Web community while simultaneously how they expect the Web community to treat the ODP's resources.

The ODP is also known as DMOZ, which is the acronym for Directory Mozilla. The DMOZ name demonstrates the association that the ODP has with Netscape's Open Source browser, Mozilla.

How does the ODP function?

¹ Author's italics

In similar, but electronic, fashion to the development of the Oxford English Dictionary (OED) creators of the ODP envisioned a not-for-profit, volunteer group that would review, edit and index websites that are submitted for entry to the OPD. One of the primary differences (besides length of existence) between the OED and the ODP is that the OED is a printed edition. Thus, the OED can be updated only periodically, which is acceptable because the English language, while dynamic, is static enough that its comprehensive dictionary need not be frequently updated. Alternatively, the ODP's goal is to have a directory: 1) that is developed by Web searchers, for Web searchers, 2) that is filled with current, relevant topic driven site recommendations. This goal requires the virtually real-time ability to update the directory because the Web is a dynamic entity that grows and morphs by the hour, necessitating a group of human editors that can observe sites, site submission and complaints in a real-time environment and make updates as quickly as humanly possible. Now that the directory has become so large the human element does cause more of a lag time in updates simply due to the imbalance between the ratio of sites in the directory, continued submissions, and number of editors.

How do the editors function?

Although the ODP is “the most widely distributed data base of content classified by humans” with “over 3.8 million sites” included in the directory (<http://dmoz.org>) there are only approximately 50,000 volunteer editors working to keep those sites indexed properly. The directory continues to grow on a daily basis, presently (27 March 2003), it is divided into a total of 460,000 categories that are comprised of sixteen top-level categories, including broad areas as Arts, Society, and Science and their subsequent sub-categories that are appropriate to the top-level topic. Some categories, for example Arts: Television, have over 15,000 entries while other

categories, such as Health: Alternative: Acutouch, have 4 entries. Each one of the categories/sub-categories has at least one editor, but many editors work on more than one category area.

The OPD asks that people who choose to apply to become an ODP editor have a defined, educated interest in the content area they apply to edit. Prospective editors applications are reviewed and they are informed via e-mail if they have been accepted to edit a content area. Often a prospective editor will be denied because the area they have applied to is too broad, or it is apparent that they are not meticulous in their editing skills due to a poorly edited application. Some of the qualities that are looked for in editors are that they are committed to building a directory that is “free of commercial interests and favoritism” (<http://dmoz.org/become.html>) As a junior editor a person will be assigned to a small topic area and they can advance to edit larger areas as they gain experience cataloging for the ODP.

Editors have access to many tools that they can use in cataloging work. By performing an ODP search using the keyword “editors” one will find that there are numerous knowledge management tools available for editors of the ODP as well as a plethora of editing tools for other areas ranging from ‘copyediting’ to ‘binary editing.’ Access to these tools creates a closer community for the editors to work within because they have access to the knowledge and resources that their fellow editors are also using to create and improve the ODP. I attempted to access several of the ODP/DMOZ editor links but was presented a password-protected screen, thus I cannot explicate further on the knowledge management tools available to editors. I find this an interesting barrier due to the fact that the ODP/DMOZ speaks so highly about their open editorial policies.

How is ODP data accessed?

From a free-form convergence of a few people with a vision of creating a catalog of reliable, accessible websites, the ODP has become a large, regulated, rule-bound group of volunteer editors who operate the largest human-organized Web directory on the Net. As the project's organic nature led to exponential growth, editors began reviewing thousands of submitted sites. This extreme growth necessitated the creation of guidelines for the project, although the group is still more self-regulating than rule bound. If guidelines had not been implemented, it is possible that internal disorganization could have destroyed what might be the most important indexing development in the vast World-Wide-Web community. This project continues to be an organic, volunteer-based venture that relies on both the current, internal community of editors as well as its users and potential, future editors to thrive.

As is it is not a search engine, but rather a directory the ODP does not have the same algorithmic or hit-based ranking systems that create the ranking structure for many search engines such as Google or Excite. Due to the reliability and quality of the sites that are listed in the ODP many of the major for-profit search engines converge their services with ODP results which are completely free and available to anyone who wants to use the information in the ODP catalog (provided users of the catalog credit the source of the information). The ODP files are accessed for use by other sites through a Resource Description Framework (RDF) dump. This dump is a large download of metadata file information from ODP's system into another group's system so that the ODP's catalog of the WWW can be shared freely throughout the WWW.

What is RDF to the ODP?

A brief explanation of RDF is deemed necessary so that one might understand the knowledge sharing that goes on between ODP and those who access the ODP's data through RDF dumps. RDF came to be developed because "everything on [the Web] is *machine-readable*, this data is not *machine-understandable*" (<http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>) and there needed to be a way for data about the data on the web, or metadata, to be accessible across applications. RDF uses a metadata vocabulary, which allows people to use different software to exchange a variety of meta-data vocabularies. In the W3C (World Wide Web Consortium) recommendations for RDF it is explained that RDF is a tool that can describe resources without making "assumptions about a particular domain, nor defines (a priori) the semantics of any application domain". Furthermore, the recommendations state that RDF should be neutral, yet "suitable for describing information about any domain." RDF is based on a hierarchical schema of properties and classes that allow an object to be cataloged by the subject's metadata.

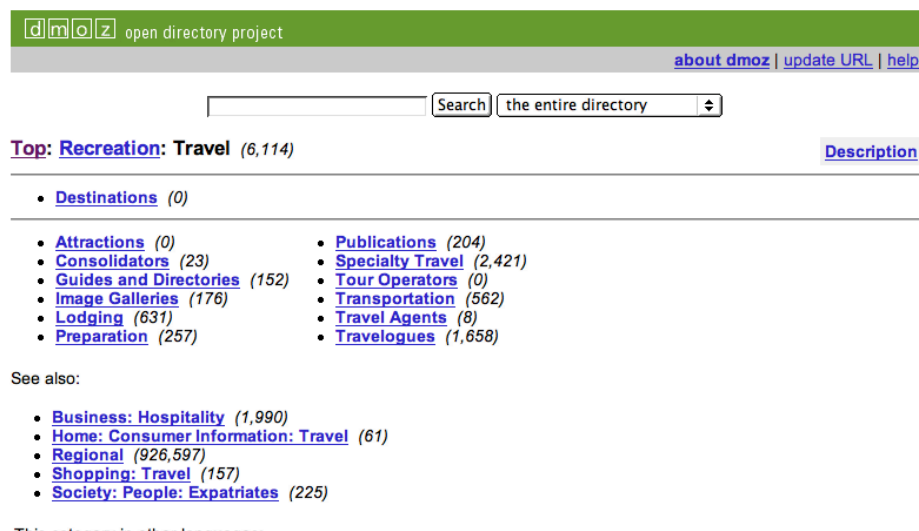
RDF's use of subject hierarchy allows catalogs, such as the ODP, to be broken down into numerous categories and sub-categories so that a user does not enter a broad search term, such as Travel, only to be bombarded with a million site results. Instead, through the ODP's use of RDF metadata the searcher would find that Travel is an initial sub-category of the top-level category Recreation (*see* screenshot 1).



Recreation as top category with Travel as sub-category.

screenshot 1: Category levels

When they click on the Travel link they will find there are 6,114 links for Travel and an additional twelve sub-categories ranging from Attractions to Travelogues, which have a total of over 6,000 entries (*see* screenshot 2). On the travel sub-category page one will also find a See also: section that lists links that pertain to travel but are about such things as Business: Hospitality, People: Expatriates, and Regional. These links will open up even more sub-categories, leaving the user with endless possibilities from which to begin their search for the most broad or specific information that they choose to locate.



Screenshot 2: Additional sub-categories reached from Recreation: Travel

One can see how this hierarchical RDF schema can make the ODP exceptionally well organized, but also how it is vital that the human element of editors are meticulous in their cataloging abilities and that they can see the broad overview of the structure they are working within as well as being clear as to what the focus of their narrow editorial chasm is so that sites they index are truly aligned with the area the site is indexed in on the ODP.

The Future of the altruistic ODP

The volunteers at ODP continue to work hard to keep their project free from commercial interests and unbiased towards personal opinions of site relevance. However, there are many profit-based search engine companies that appear to be more concerned with the bottom-line dollar revenue than becoming a valuable, integrity driven knowledge management tool. The for-profit tools, such as Looksmart and Excite, are often good for finding what one might need for general information, but the results are not combed through and indexed for validity in the way that the indexed results are at ODP. As google becomes a verb that is analogous to research the concepts of reliability and validity, that are so integral to strong research, are being lost to a generation that is constantly pounded with commercially based 'information' that may or may not be accurate. Due to the fact that for-profit search engines, such as Google, piggy back their services on top of the results they derive from the ODP index makes deciphering results more difficult. The difficulty occurs because for-profit search engine results are mixed with 'good' and 'bad' results, which can make recognizing what is reliable versus what is spam difficult. Due in part to the fact that companies like Google access the RDF dumps that ODP so generously makes available to anyone who wants the knowledge they hold in their indexes the validity of sites may not be noticed by all users, especially young adults or people who are new to the Net, of search engines.

I do not believe the misleading appearance of information is intentional on part of the for-profit sites, as I do not believe that the ODP cannot say a well-done spam site has never foiled them and made it into their directory. The greatest hurdle that the ODP might have in its continued strength is making itself recognized on the waves of the commercial Net. However, the group has such an enormous base of sites and since for-profit sites like Google and Yahoo access their stores of indexes I believe the ODP will continue to thrive as a place of knowledge where search engine users can retrieve information free from the pounding of Internet sales pitches. Finally, in the completion of my research I often had a screen such as follows, which leads me to believe the only real problem the ODP might have, insofar as the question of its future existence is concerned, is the need for more server space.

The Open Directory search is currently under a heavy load. Please try back later.
No **Open Directory Project** results found

Try your search on:

[Google](#)

[Google Directory](#)

[Google Groups](#)

[Google Image](#)

New Search

[Advanced Search](#) [Help on Search](#)

Works Cited

The following references were retrieved on or before 31 March 2003.

<http://dmoz.org>

<http://dmoz.org/about.html>

<http://www.laisha.com/zine/odphistory.html>

<http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>