3-2008

Symbolic Links in the Open Directory Project

Saverio Perugini
University of Dayton, sperugini1@udayton.edu

eCommons Citation
Perugini, Saverio, "Symbolic Links in the Open Directory Project" (2008). *Computer Science Faculty Publications*. 27.
https://ecommons.udayton.edu/cps_fac_pub/27

Symbolic Links in the Open Directory Project

Saverio Perugini

Department of Computer Science, University of Dayton
300 College Park, Dayton, OH USA 45469–2160
Tel: +001 937 229 4079 Fax: +001 937 229 2193

Abstract

We present a study to develop an improved understanding of symbolic links in web directories. A *symbolic link* is a hyperlink which makes a directed connection from a webpage along one path through a directory to a page along another path. While symbolic links are ubiquitous in web directories such as Yahoo!, they are under-studied and, as a result, their uses are poorly understood. A cursory analysis of symbolic links reveals multiple uses: to provide navigational shortcuts deeper into a directory, backlinks to more general categories, and multiclassification. We investigated these uses in the Open Directory Project (ODP), the largest, most comprehensive, and most widely distributed human-compiled taxonomy of links to websites, which makes extensive use of symbolic links. The results reveal that while symbolic links in ODP are used primarily for multiclassification, only few multiclassification links actually span top- and second-level categories. This indicates that most symbolic links in ODP are used to create multiclassification between topics which are nested more than two levels deep and suggests that there may be multiple uses of multiclassification links. We also situate symbolic links *vis à vis* other semantic and structural link types from hypermedia. We anticipate that the results and relationships identified and discussed in this paper will provide a foundation for (1) users for understanding the usages of symbolic links in a directory, (2) designers to employ symbolic links more effectively when building and maintaining directories and for crafting user interfaces to them, and (3) information retrieval researchers for further study of symbolic links in web directories.

Key words: Cross-references, Hypermedia, Information access, Information hierarchies, Link analysis, Link typing, Navigational search, Open Directory Project, Symbolic links, Taxonomic navigation, Web directories, World Wide Web

---

Email address: `saverio@udayton.edu` (Saverio Perugini).
URL: `http://homepages.udayton.edu/~perugisa` (Saverio Perugini).

Accepted for publication in Information Processing and Management *25 June 2007*

1 Introduction

The problem addressed by this paper is improving the understanding of symbolic links in web directories. Informally, a symbolic link is a hyperlink which makes a directed connection from a webpage along one path through a website to a page along another path. The concept of a symbolic link in general is not unique to electronic environments. Cross-references, such as a 'see also stylists' annotation in the the 'beauty salons' section of the *Yellow Pages*, function similar to symbolic links. In UNIX, soft or symbolic links — those with an `l` to the left of the file permission mode, e.g., `lrwx------` — are used for similar purposes (Marsden and Cairns, 2003).

Symbolic links are most common in large web directories such as Yahoo!. Such directories, which are organized along a hierarchy of categories (e.g, Business, News, or Sports), are the web's analog to the *Yellow Pages*. Web directories represent one of three major paradigms of searching the web and, as a result, serve as a gateway or *starting point* to web resources for many users (Baeza-Yates and Ribeiro-Neto, 1999). Moreover, while 'the web coverage provided by directories is very low (less than 1% of all webpages), the answers returned to the user are usually much more relevant' as such directories often exclude low quality sites (Baeza-Yates and Ribeiro-Neto, 1999). Users interact with these directories by progressively drilling down the categories to find pointers to websites of interest (called *structure guided browsing*; Baeza-Yates and Ribeiro-Neto, 1999). The underlying goal of symbolic links is to improve information access. However, designers of web directories use symbolic links in multiple ways, introduced below, within the scope of improving information access.

1.1 Uses of Symbolic Links in Web Directories

We use the sample web directory shown in Fig. 1 to illustrate the following uses of symbolic links. We call every page in a web directory a *topic* page. We use the words *topic* and *category* interchangeably in this article. Topic pages are divided into non-leaves and leaves. We call a page in a directory with at least one link to another topic page in the directory a *non-leaf*. In Fig. 1, pages 1–7 are non-leaves. Conversely, we call a page in a directory with links only to external webpages (i.e., those which are not part of the directory) a *leaf*. In Fig. 1, pages 8–14 are leaves. We also say that a *path* is an ordered set of hyperlink labels from the root page of a directory to a leaf. For instance, $\prec$computers: software: music$\succ$ and $\prec$arts: theatre: vocal@: jazz$\succ$ are paths in Fig. 1. A *hard path* (consisting of only hard, or non-symbolic, links) does not contain a symbolic link. Therefore, above, the former path is a hard path while the latter is not. We use the term *sequence* to refer to a subset of a path,

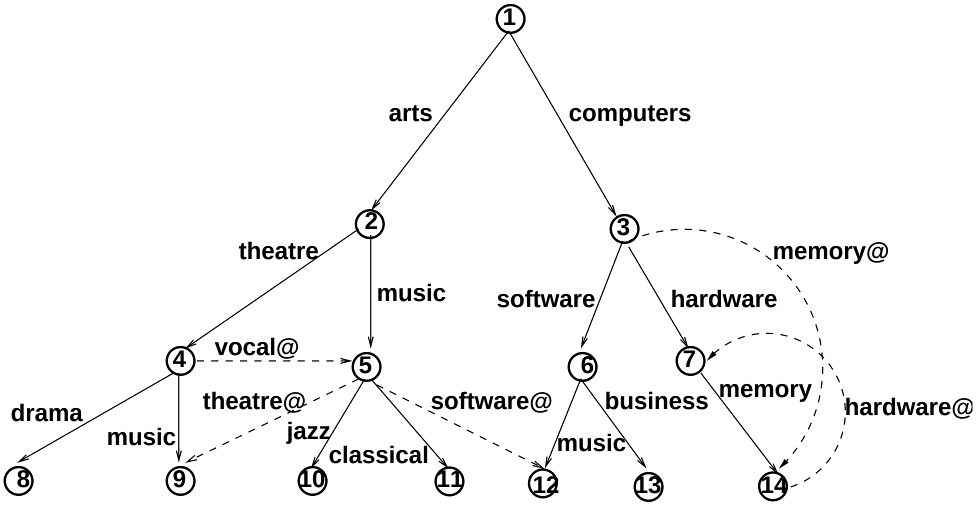


Fig. 1. Sample web directory, simplified for purposes of presentation, with characteristics similar to those in Yahoo!. Nodes correspond to webpages and directed edges correspond to hyperlinks between pages. Symbolic links are indicated by dashed edges and hyperlink labels ending with @.

e.g., ≺computers: software≻ is a sequence in the former path above.

- **Shortcuts** (Scs): One use of a symbolic link is to provide a navigational shortcut to a page nested deeper in a directory. For example, in Fig. 1, the symbolic link labeled ‘memory@’ provides a shortcut from webpage 3 to 14 bypassing the intermediate page 7. Such a usage gets users to the information they desire in fewer steps. We define a *shortcut* as a symbolic link whose target can be reached via a path, without symbolic links, through its source. Notice that a shortcut need not end in a *leaf*.

- **Backlinks** (Bls): Another use of a symbolic link is to provide an avenue back to a more general category. For instance, in Fig. 1, the symbolic link labeled ‘hardware@’ provides a backlink from webpage 14 to page 7. We define a *backlink* as a symbolic link whose target can reach its source through a path without symbolic links.

- **Multiclassification** (Mc): Interacting with any directory entails predicting which section of the directory classifies the desired information, which can be a time-consuming task (Hearst, 1999). Directory designers use symbolic links to accommodate incorrect predictions. Consider the symbolic link labeled ‘software@’ from page 5 to 12 in Fig. 1. It gives the illusion that page 12 is classified under the the ‘arts’ category of the directory, when it is actually classified under the ‘computers’ category. The intended purpose this type of symbolic link is to tighten the gap between an item’s actual placement in the directory and users’ predictions of that placement during information seeking. We define a *multiclassification link* as a symbolic link whose target is classified under a different top-level category than its source. We define a *top-level category of category X* as any hyperlink label at the root page of category X. For example, 'arts' and 'computers' are top-level categories of the root in Fig. 1 while 'software' is a top-level category of the 'computers' sub-category (rooted at page 3). The symbolic link labeled 'vocal@' from page 4 to 5 is another multiclassification link. Multiclassification links allow users to naturally shift between different categories of a directory without traversing manually up and down a hierarchy.

Notice that symbolic links used as backlinks or for multiclassification, by inducing cycles, preclude the underlying graph model of the directory from being a DAG (Directed Acyclic Graph). In the absence of any symbolic links, the model is a tree. Notice further that we define types of symbolic links using purely syntactic notions in the underlying graph-theoretic model of the directory. Also, note that using the definitions above, the shortcuts, backlinks, and multiclassification classes of symbolic links are mutually exclusive, i.e., #symbolic links = #shortcuts + #backlinks + #multiclassification links.

1.2 Our Study

While symbolic links are used copiously in web directories, they are surprisingly under-studied. Many researchers have identified the elementary understanding of links in general as a problem (Bar-Ilan, 2005; Kopak, 1999; Henzinger et al., 2002). Therefore, we investigated the distribution of the preceding uses of symbolic links in the *Open Directory Project* (ODP), a voluminous web taxonomy (over 689 000 topics), to improve the understanding of them in web directories (from a design perspective) and their effect on information retrieval.

The objective of our study was to gain an improved understanding of the uses of symbolic links in web directories. We began our research by studying 'how uses of symbolic links as shortcuts, backlinks, and multiclassification links are distributed?' We studied this question by analyzing the distribution of these uses of symbolic links in ODP using standard link analysis techniques. This analysis revealed how these uses are distributed in the entire directory as well as in each top-level category. However, the results also led to the following additional questions regarding how symbolic links affect the connectivity of top-level categories.

- How are multiclassification links used to bridge topics *between* and *within* top-level categories distributed?
- How well do the multiclassification links used to span topics within top-level categories connect all possible pairs of the top-level categories of those categories?
- How are the sources and targets of symbolic links distributed across top-

level categories?

We answered these questions within the entire directory as well as each top-level category.

The following section surveys related research and situates symbolic links *vis à vis* other semantic and structural link types from hypermedia. We illustrate these three uses of symbolic links in ODP in Section 3 and present the progression of our study resulting from our initial analysis in Section 4. Finally, we discuss our contributions and offer ideas for future work in Section 5.

2 Related Research

Symbolic links evolved from the hierarchical database model and filesystems to hypermedia and web directories.

2.1 Roots in the Hierarchical Database Model and Filesystems

Hierarchical databases used strict hierarchies to model one-to-many relationships (Date, 1993). Such an organization was too rigid for practical purposes which often required records to be classified in more than one way without data duplication (Dourish and Edwards, 2000). Therefore, link records were added to the model and used to create symbolic links across hierarchies (Marsden and Cairns, 2003). Symbolic links have also been similarly supported in filesystems (e.g., UNIX, MacOS, and Windows), albeit in different guises, to help users with file management and access activities (Marsden and Cairns, 2003). Again, such symbolic links are used to create the illusion that files are accessible from multiple locations. Marsden and Cairns (2003) feel that symbolic links used in this manner are kludges which provide only temporary relief for a more systemic multiclassification problem. While the more powerful and flexible relational model obviated the need for symbolic links in databases, similar approaches in filesystems have not been widely embraced.

2.2 Hypermedia

In the hypermedia community (and to some extent web community), symbolic links are often known as *cross-references*[1]. However, symbolic links are the

---

[1] The terms *cross-reference* and *symbolic link* are often, and, in our opinion incorrectly, used interchangeably. We use the term *symbolic link* throughout this pa-

analog of the hypermedia *composites* construct. Composites can be thought of as collection data structures (Helic et al., 1999). 'Just like composites, symbolic links allow containers ([e.g.,] directories) to share the same components without duplication and just like composites, only whole files can be shared among containers' (Bieber et al., 1997). *Transclusions* (or *inclusions*), another hypermedia construct, are similar to symbolic links but the parallel is not as strong as with composites. Transclusions are links to the original copy of data from all of the sources which use it (Bieber et al., 1997). Transclusions support a single access point for data modifications and obviate the need to synchronize updates across potentially many identical copies of data. Transclusions are similar to composites in that both reduce redundancy and simplify updates. However, the two concepts differ in that only whole entities (e.g., files, documents) can be shared among containers, while transclusions can refer to sub-parts of data objects (e.g., paragraphs, clauses). *Server-side includes* are the web analog of transclusions and can be implemented using a web scripting language such as PHP to, e.g., dynamically include the same header and footer automatically on each webpage in a site when retrieved via HTTP (Bieber et al., 1997).

Lastly, we distinguish between *backjumps* (a link to the previously visited node; Bieber et al., 1997) from hypermedia and *backlinks* – a type of symbolic link described above. Not all backlinks are backjumps and not all backjumps are backlinks. Multiclassification links can function as backjumps. Notice that determining whether a link is a backjump depends on context (i.e., if the link's target is a previously visited node). Therefore, we say a backjump is a *semantic*[2] type of link. A backlink, on the other hand, or any symbolic link, is a type of *structural* link by definition. In other words, symbolic links are defined purely by structural or graph-theoretic properties of the underlying graph model of the directory.

2.3 Link Typing

While a node type categorizes the node's *content* (Bieber et al., 1997), a link type describes the relationship between the source and destination of the link (Allan, 1996). The end goal of link typing is to 'make hypermedia's structure and semantics more comprehensible for users' (Noirhomme-Fraiture and Serpe, 1998). "Semantically typed nodes and links help authors organize information more effectively and lend context for readers" (Bieber et al., 1997).

---

per because (1) we advocate that cross-references and symbolic links are different and (2) the ODP RDF *meta*-data used in our study (and described below) uses a `<symbolic>` tag for each symbolic link.

[2] Here we are not referring to relations 'between words at the semantic level, but between relationships at the functional level' (Kopak, 1999).

Specifically, users often desire to know the nature of the target of a link before they invest time in following it (Bieber et al., 1997). Armed with link types, we can study visual cues for implying link types to users (Noirhomme-Fraiture and Serpe, 1998). For example, the standard visual representation for a symbolic link in web directories seems to be (at least in the Yahoo! and ODP) a link label followed by the '@' character. Nonetheless, the primary focus of typing in the hypermedia community has been on nodes rather than links (Kopak, 1999).

Therefore, links are important, but under-studied and under-utilized (Kopak, 1999). In practice, while HTML does support the specification of link types (e.g., the `CLASS`, `REL`, and `REV` attributes of the `A` and `LINK` tags as well as the `TITLE` tag), such features seem to be very rarely used by authors or web browsers (Allan, 1996; Bieber et al., 1997). Given the limited research on links and link types, it is not surprising that, to the best of our knowledge, not much research has been done on the types of symbolic links or symbolic links in general in web directories (or hypermedia systems). Moreover, the little work that has been done on link typing has almost exclusively focused on *semantic* link typing. Even less work on link typing has been done in the web domain:

> "While link analysis has become increasingly important as a technique for web-based information retrieval, there has not been much research into the different types of links on the web" (Henzinger et al., 2002).
>
> "The lack of research in this area is rather surprising, since links are not only a means to link documents; links have been extensively used to improve the performance of IR systems (e.g., Brin and Page, 1998; Kleinberg, 1999; Lempel and Moran, 2001)" (Bar-Ilan, 2005).

While symbolic links are used in nearly every web directory, again surprisingly, but similarly, our literature searches have revealed that very little research has been done to understand how symbolic links are used. We offer our work is a starting point for understanding the types of symbolic links in web directories.

2.4 Link Taxonomies

There have been several classifications of link types inspired by the field of discourse analysis. One the earliest attempts to classify links was by Trigg (1983) who created a taxonomy of 80 links types for a scientific publication system. DeRose (1989) considers the structure of a hypertext, in addition to semantics and, therefore, provides a more holistic link taxonomy. *Inclusion* (or one-to-many) links represent super-ordinate/sub-ordinate relationships between documents. *Sequential* links are inclusion links whose target

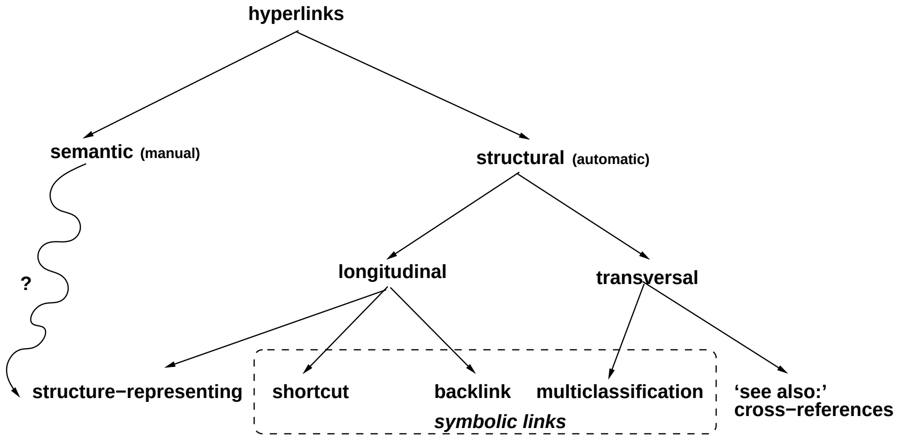


Fig. 2. A conceptual taxonomy of hyperlinks designed to situate symbolic links in relation to various other types of structural links.

locations are ordered. 'The most important example of a sequential link is the *structure-representing* [(or *s-r*)] link' (DeRose, 1989). 'A major identifying feature of these s-r links is that they provide a basis for presenting the text in linear form when needed' (DeRose, 1989). '[S]-r links express many useful and standard hierarchies' (DeRose, 1989).

Overall, there is no agreement in the literature on taxonomies of link types on either structure or nomenclature (DeRose, 1989). Fig. 2 is a hierarchical classification of link types we developed for purposes of understanding how symbolic links relate to various other structural link types. We adopt the philosophy of DeRose and agree that '[the] precise divisions could be expressed in different ways', but suggest that this hierarchy provides a conceptual illustration of the three types of symbolic links studied in this paper *vis à vis* other structural link types. We make a distinction between *longitudinal* and *transversal* structural links (Noirhomme-Fraiture and Serpe, 1998). Longitudinal links represent parent-child relationships whereas transversal links are used for cross-referencing (Noirhomme-Fraiture and Serpe, 1998). Longitudinal links are similar to *organizational* (Conklin, 1987) and *inclusion* links (DeRose, 1989). *Structure-representing* links (DeRose, 1989) represent the strict parent-child relationships in web taxonomies, i.e., they lead to a child or parent. They are the links in web directories which induce the hierarchy and those we refer to as *hard* links above. Shortcuts and backlinks, on the other hand, lead to non-child descendants and non-parent ancestors, respectively. In ODP, there are two types of transversal links: multiclassification links and 'see also:' cross-reference links. We shall have more to say about the distinction between multiclassification links and 'see also:' cross-references in Section 3. Longitudinal links take users up and down a hierarchy while transversal links take users across a hierarchy. There are various types of semantic links in the literature and semantic links types are still an open area of current research (Bar-Ilan, 2005). Fig. 2 is not comprehensive or standard.

While link taxonomies depend on applications, there have been attempts to reconcile these differences (Kopak, 1999; Noirhomme-Fraiture and Serpe, 1998). Regardless of the disagreement in these taxonomies, symbolic links are the primary focus of none of the link taxonomy articles cited above. Furthermore, none of them recognize the possibility that there might exist multiple types of symbolic links.

3 The Open Directory Project

The Open Directory Project (ODP) at dmoz.org (formerly *GnuHoo* and then *NewHoo*) is the largest, most comprehensive, and most widely distributed human-compiled taxonomy of links to websites (The Open Directory Project, 2002). Since it uses symbolic links extensively, it provides a fertile data source from which to analyze their uses. In addition, ODP 'powers the core directory services for the web's largest and most popular search engines and portals, including *Netscape Search*, *AOL Search*, *Google*, *Lycos*, *HotBot*, *DirectHit*, and hundreds of others' (The Open Directory Project, 2002). All of the directory's data is also available free to the public at `http://rdf.dmoz.org` in RDF, a common format for describing web data[3]. Therefore, the choice to study symbolic links in ODP was motivated by its relevance, availability, and the applicability of our results to the most important web search engines and portals. Also, note that the graph-theoretic notions and definitions developed above are applicable to any hierarchical directory employing symbolic links. We computed our results from the `structure.rdf.u8.gz` RDF dump file, downloaded on 11/4/2005, which contains the category hierarchy information.

3.1 Structural Characteristics

Table 1 captures the relative volume of each category of ODP across a variety of structural characteristics. The parenthesized numbers in this table as well as all following tables in this article represent the values from the sample directory in Fig. 1 corresponding to the given characteristic. The column labeled **#Topics** provides the sum of the values from the corresponding entries of the columns labeled **#Non-leaves** and **#Leaves**. We define *term* as a string labeling a hyperlink, i.e., the complete text between the `<a href="">` and `</a>` HTML tags. In Fig. 1, 'arts' is a term. Notice that while each term in Fig. 1 contains only one word (i.e., any string of characters except space), our definition permits a term to consist of more than one word (e.g., 'Products and Shopping' is *one* term in ODP) and this viewpoint is reflected in our terms

---

[3] ODP is the only major directory whose data is 100% free to the public.

Table 1
Structural characteristics of ODP. **Legend**: μ = mean; C/N children per node; σ = standard deviation.

Category	**#Non-leaves**	**#Leaves**	**#Topics**	**#Terms**	μ **C/N**	**Depth range**
Adult	2 085	5 887	7 972	2 191	3.82	[2–11]
Arts	(3) 8 344	(4) 38 525	(7) 46 869	(8) 28 595	(2.33) 5.62	([3–3]) [2–11]
Business	1 977	9 207	11 184	4 311	5.66	[2–10]
Computers	(3) 1 637	(3) 6 478	(6) 8 115	(6) 3 730	(1.66) 4.96	([3–3]) [2–10]
Games	2 775	8 652	11 427	7 013	4.12	[2–11]
Health	999	5 534	6 533	3 276	6.54	[2–9]
Home	493	2 026	2 519	1 667	5.11	[2–8]
Kids & Teens	837	3 220	4 057	3 346	4.85	[2–11]
News	266	1 583	1 849	338	6.95	[2–7]
Recreation	1 782	8 770	10 552	3 539	5.92	[2–10]
Reference	2 747	8 798	11 545	6 650	4.20	[2–12]
Regional	87 262	210 130	297 392	48 429	3.41	[2–14]
Science	2 205	9 423	11 628	7 965	5.27	[2–11]
Shopping	1 173	4 153	5 326	3 132	4.54	[2–10]
Society	4 532	23 083	27 615	12 370	6.09	[3–12]
Sports	2 329	15 446	17 775	7 473	7.63	[2–10]
World	42 478	164 746	207 224	105 498	4.88	[3–14]
ODP	(7) 163 922	(7) 525 661	(14) 689 583	(13) 218 640	(1.86) 4.21	([3–3]) [2–14]
μ	9 642.41	30 921.24	40 562.65	14 677.82	5.27	[2.12–**10.65**]
σ	22 321.01	60 125.68	81 986.75	26 316.08	1.13	[0.33–1.80]

counts. Term counts in Table 1 omit duplicates. We compute the average (μ) number of children per node as the total number of children, i.e., the total number of topics minus one (because the root is a child of no node), divided by the total number of parents, i.e., non-leaves, in the category/directory. Values under the column labeled **Depth range** indicate the minimum and maximum length of any path starting from the root to reach a leaf.

Table 1 reveals that there are many more leaves than non-leaves – on average, more than 3 ($\approx$ 30 921.24/9 641.41) times as many per category. Also, note that the standard deviation of the number of non-leaves, leaves, topics, and terms per category is high while the standard deviation of the average number of children per node per category is low.

The reader will notice that the row labeled **ODP** does not provide the exact sum of the previous 17 rows from the columns labeled **#Non-leaves**, **#Topics**, and **#Terms**. The values in the entries in the row labeled **ODP** from the columns labeled **#Non-leaves** and **#Leaves** are one greater than the sum of the values from the preceding 17 rows of the same columns to account for the root of the entire directory. In addition, the value of the column labeled **#Terms** from the row labeled **ODP** does not equal the sum of the previous 17 rows from the same column due to duplication in terms between categories.

Since symbolic links do not affect the count of non-leaves, leafs, and topics, they are not considered in Table 1. In other words, the numbers in Table 1 reflect a tree model of ODP and therefore the values in the column labeled **#Leaves** represent the number of (hard) paths (not involving a symbolic

Table 2
Effect of symbolic links on the average number of children per node. **Legend**: C/N = children per node; (w/o) = without symbolic links; (w/) = with symbolic links; $r^{(w/:w/o)}$ = ratio of average number of children per node with symbolic links to that without.

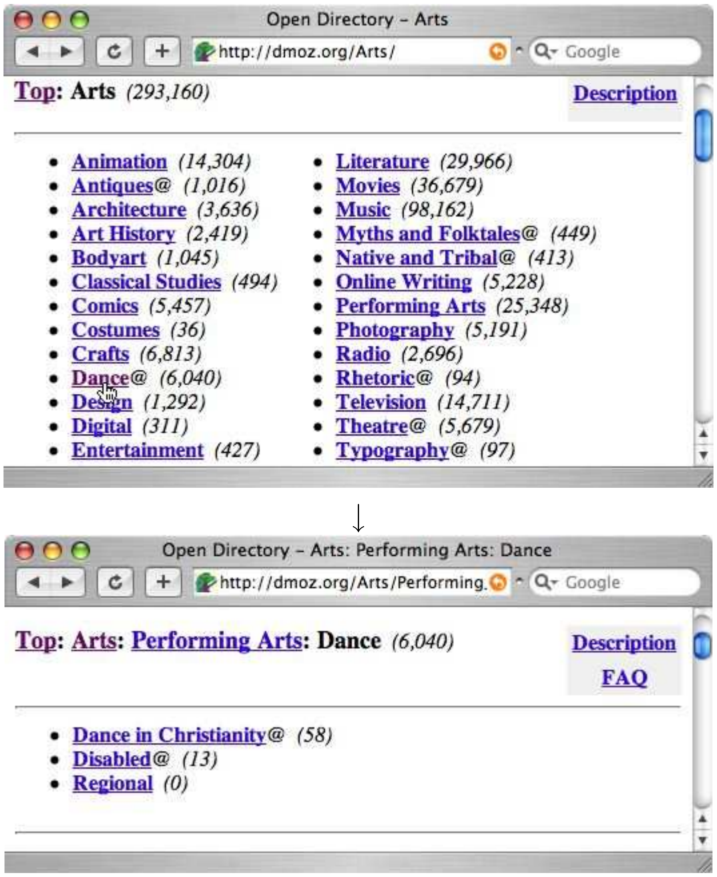
Category	μ **C/N**$^{(w/o)}$	μ **C/N**$^{(w/)}$	**R**$^{(w/:w/o)}$
Adult	3.82	5.81	1.52
Arts	(2.33) 5.62	(3.00) 16.22	(1.29) 2.89
Business	5.66	8.81	1.56
Computers	(1.66) 4.96	(2.33) 9.63	(1.40) 1.94
Games	4.12	11.12	2.70
Health	6.54	12.88	1.97
Home	5.11	9.56	1.87
Kids & Teens	4.85	7.63	1.57
News	6.95	10.89	1.57
Recreation	5.92	9.87	1.67
Reference	4.20	10.66	2.54
Regional	3.41	7.42	2.18
Science	5.27	9.01	1.71
Shopping	4.54	7.30	1.61
Society	6.09	11.44	1.88
Sports	7.63	13.39	1.75
World	4.88	9.27	1.90
ODP	(1.86) 4.21	(2.57) 8.77	(1.38) 2.08
μ	5.27	10.05	1.93
σ	1.13	2.54	0.42

link) from the root of the directory to each leaf. Symbolic links also do affect term counts (owing to the presence of the term labeling the symbolic link), average number of children per node, and depth range per category. Table 2 describes the effect of symbolic links on the average number of children per node in the categories of ODP as well as the entire directory. In the presence of symbolic links, the total number of children equals the total number of children in the absence of symbolic links plus the number of symbolic links whose source resides in the particular category of the directory considered. As can be seen in Table 2, symbolic links more than double the average number of children per node in the directory with a small standard deviation across all top-level categories.

In addition to the distinctions made by progressive hyperlinks between pages to more specific topics, ODP makes within-page distinctions. These distinctions are illustrated visually in the webpages through the use of horizontal rule tags (i.e., <`hr/`>s) in the HTML source to partition a page into levels. Tags corresponding to hard[4] (<`narrow/`>) and symbolic (<`symbolic/`>) links in the RDF structure file may contain a trailing numeral to indicate the partition of the page from which the link originates. For instance, a symbolic link annotated with the tag <`symbolic/`> indicates that it resides on the top-most partition of the page while one annotated with a <`symbolic1/`> tag origi-

---

[4] Note that we refer to such non-symbolic links in Section 2 and Fig. 2 as *structure-representing* or *s-r* links.



Top: **Arts** *(293,160)*

- **Animation** *(14,304)*
- **Antiques**@ *(1,016)*
- **Architecture** *(3,636)*
- **Art History** *(2,419)*
- **Bodyart** *(1,045)*
- **Classical Studies** *(494)*
- **Comics** *(5,457)*
- **Costumes** *(36)*
- **Crafts** *(6,813)*
- **Dance**@ *(6,040)*
- **Design** *(1,292)*
- **Digital** *(311)*
- **Entertainment** *(427)*
- **Literature** *(29,966)*
- **Movies** *(36,679)*
- **Music** *(98,162)*
- **Myths and Folktales**@ *(449)*
- **Native and Tribal**@ *(413)*
- **Online Writing** *(5,228)*
- **Performing Arts** *(25,348)*
- **Photography** *(5,191)*
- **Radio** *(2,696)*
- **Rhetoric**@ *(94)*
- **Television** *(14,711)*
- **Theatre**@ *(5,679)*
- **Typography**@ *(97)*

↓
Open Directory - Arts: Performing Arts: Dance
http://dmoz.org/Arts/Performing.
Google
Description
FAQ

Top: **Arts**: **Performing Arts**: **Dance** *(6,040)*

- **Dance in Christianity**@ *(58)*
- **Disabled**@ *(13)*
- **Regional** *(0)*

Fig. 3. A shortcut, labeled 'Dance@,' in ODP of length 2 from topic 'Arts' to topic 'Arts: Performing Arts: Dance.'

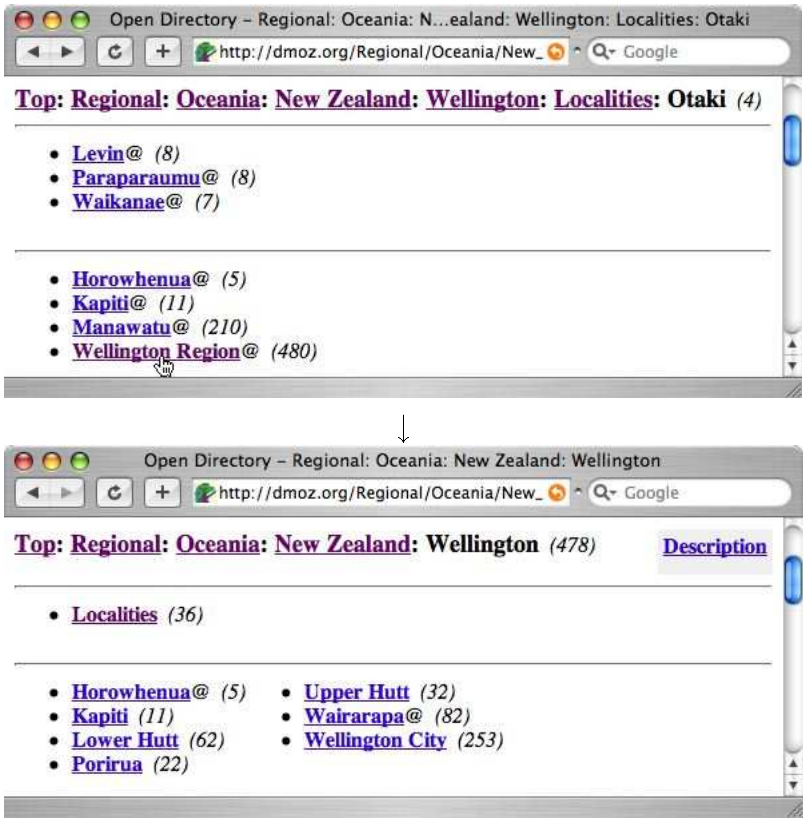
nates from the second vertical partition of the page, and so on. We did not consider within-page distinctions in our study

According the definitions given in Section 1, the category page labeled 'Arts' in ODP shown in Fig. 3 (top) is a *non-leaf*. Conversely, the page at 'Recreation: Antiques: Books' in ODP is a leaf. 'Arts' and 'Recreation' are each top-level categories of the ODP root. The *length* of a shortcut or backlink is the number of non-symbolic links it bypasses on the (hard) path, without symbolic links, in which its source and target reside. The length of the shortcut shown in Fig. 1 is 2 while that of the backlink there is 1.

3.2 Symbolic Links

Shortcuts

Fig. 3 illustrates a symbolic link used as a shortcut in ODP. The shortcut is labeled 'Dance@' and it connects the page labeled 'Arts' to the page labeled 'Arts: Performing Arts: Dance.' Therefore, this shortcut bypasses the page in



Open Directory – Regional: Oceania: N...ealand: Wellington: Localities: Otaki

http://dmoz.org/Regional/Oceania/New_

Top: Regional: Oceania: New Zealand: Wellington: Localities: **Otaki** *(4)*

- Levin@ *(8)*
- Paraparaumu@ *(8)*
- Waikanae@ *(7)*

- Horowhenua@ *(5)*
- Kapiti@ *(11)*
- Manawatu@ *(210)*
- Wellington Region@ *(480)*

↓

Open Directory – Regional: Oceania: New Zealand: Wellington

http://dmoz.org/Regional/Oceania/New_

Top: Regional: Oceania: New Zealand: **Wellington** *(478)*

Description

- Localities *(36)*

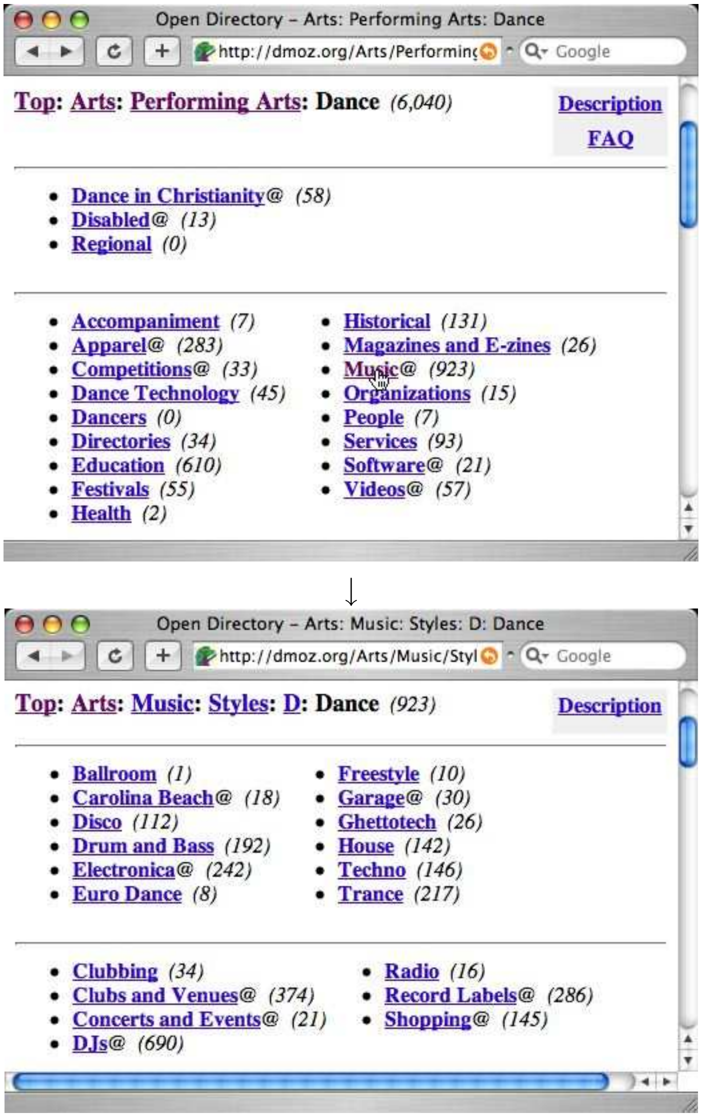
- Horowhenua@ *(5)*
- Kapiti *(11)*
- Lower Hutt *(62)*
- Porirua *(22)*
- Upper Hutt *(32)*
- Wairarapa@ *(82)*
- Wellington City *(253)*

Fig. 4. Illustration of a backlink, labeled ‘Wellington Region@,’ of length 2 in ODP from topic ‘Regional: Oceania: New Zealand: Wellington: Localities: Otaki’ to topic ‘Regional: Oceania: New Zealand: Wellington.’

ODP labeled ‘Arts: Performing Arts’ which is a child of the page labeled ‘Arts’ (Fig. 3, top) and is of length 2. This shortcut obviates the need to click on the series of hard links in ODP labeled ‘Performing Arts’ and ‘Dance.’ This shortcut, like all shortcuts, brings the user deeper into the directory.

Backlinks

Fig. 4 illustrates a symbolic link used as a backlink in ODP. The backlink is labeled ‘Wellington Region@’ and it connects the page labeled ‘Regional: Oceania: New Zealand: Wellington: Localities: Otaki’ in ODP to the page labeled ‘Regional: Oceania: New Zealand: Wellington.’ Therefore, this backlink bypasses the page in ODP labeled ‘Regional: Oceania: New Zealand: Wellington: Localities’ and is of length 2. This backlink obviates the need to traverse back through the series of non-symbolic links in ODP labeled ‘Localities’ and ‘Otaki.’ This backlink, like all backlinks, brings the user to a higher level of the directory. As can be seen in Figs. 3 and 4 and their descriptions above, symbolic links used as shortcut and backlinks transition the user vertically down or up a directory, respectively.


Open Directory – Arts: Performing Arts: Dance
http://dmoz.org/Arts/Performing
Google
Description
FAQ
↓
Open Directory – Arts: Music: Styles: D: Dance
http://dmoz.org/Arts/Music/Styl
Google
Description

Top: **Arts**: **Performing Arts**: **Dance** *(6,040)*

- **Dance in Christianity**@ *(58)*
- **Disabled**@ *(13)*
- **Regional** *(0)*

- **Accompaniment** *(7)*
- **Apparel**@ *(283)*
- **Competitions**@ *(33)*
- **Dance Technology** *(45)*
- **Dancers** *(0)*
- **Directories** *(34)*
- **Education** *(610)*
- **Festivals** *(55)*
- **Health** *(2)*
- **Historical** *(131)*
- **Magazines and E-zines** *(26)*
- **Music**@ *(923)*
- **Organizations** *(15)*
- **People** *(7)*
- **Services** *(93)*
- **Software**@ *(21)*
- **Videos**@ *(57)*

Top: **Arts**: **Music**: **Styles**: **D**: **Dance** *(923)*

- **Ballroom** *(1)*
- **Carolina Beach**@ *(18)*
- **Disco** *(112)*
- **Drum and Bass** *(192)*
- **Electronica**@ *(242)*
- **Euro Dance** *(8)*
- **Freestyle** *(10)*
- **Garage**@ *(30)*
- **Ghettotech** *(26)*
- **House** *(142)*
- **Techno** *(146)*
- **Trance** *(217)*

- **Clubbing** *(34)*
- **Clubs and Venues**@ *(374)*
- **Concerts and Events**@ *(21)*
- **DJs**@ *(690)*
- **Radio** *(16)*
- **Record Labels**@ *(286)*
- **Shopping**@ *(145)*

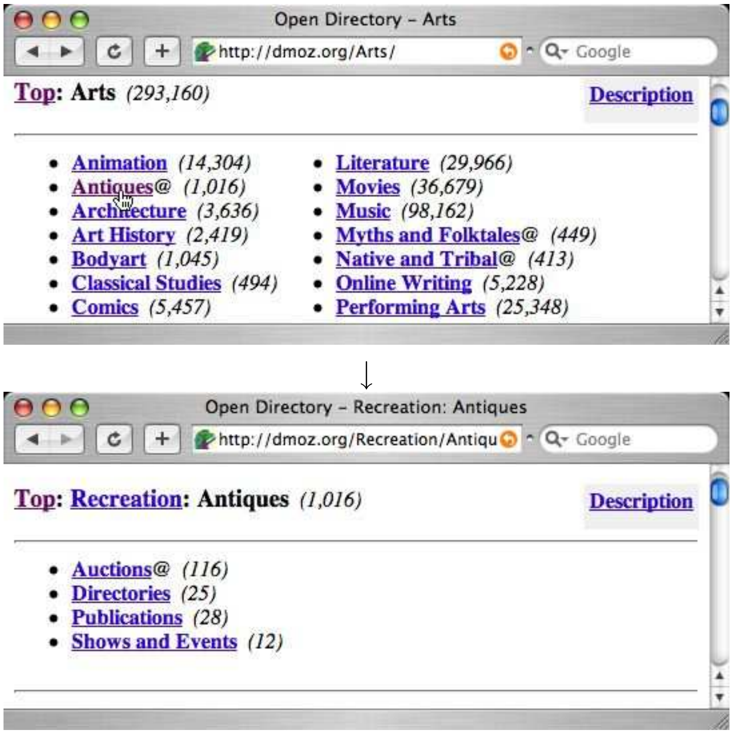
Fig. 5. Illustration of a symbolic link in ODP, labeled ‘Music@,’ creating multiclassification between two top-level categories of the ‘Arts’ category, namely, ‘Performing Arts’ to ‘Music.’

Multiclassification Links

Fig. 5 shows a symbolic link used for multiclassification in ODP. It is labeled ‘Music@’ and connects the page labeled ‘Arts: Performing Arts: Dance’ in ODP to the page labeled ‘Arts: Music: Styles: D: Dance.’ This multiclassification link, like all multiclassification links, brings the user across the directory into a different category. However, note that the multiclassification link illustrated in Fig. 5 connects two distinct sub-categories of the ‘Arts’ category, namely, ‘Performing Arts’ to[5] ‘Music’ rather than two distinct top-level categories of the entire (root) directory. This multiclassification prevents the user

---

[5] Multiclassification links, like all symbolic links, are unidirectional.


Open Directory – Arts
http://dmoz.org/Arts/
Google

Top: **Arts** *(293,160)*

Description

- **Animation** *(14,304)*
- **Antiques@** *(1,016)*
- **Architecture** *(3,636)*
- **Art History** *(2,419)*
- **Bodyart** *(1,045)*
- **Classical Studies** *(494)*
- **Comics** *(5,457)*
- **Literature** *(29,966)*
- **Movies** *(36,679)*
- **Music** *(98,162)*
- **Myths and Folktales@** *(449)*
- **Native and Tribal@** *(413)*
- **Online Writing** *(5,228)*
- **Performing Arts** *(25,348)*

↓
Open Directory – Recreation: Antiques
http://dmoz.org/Recreation/Antiqu
Google

Top: **Recreation**: **Antiques** *(1,016)*

Description

- **Auctions@** *(116)*
- **Directories** *(25)*
- **Publications** *(28)*
- **Shows and Events** *(12)*

Fig. 6. Illustration of a symbolic link, labeled 'Antiques@,' creating multiclassification between top-level categories from 'Arts' to 'Recreation: Antiques' in ODP.

at the page labeled 'Arts: Performing Arts: Dance' in ODP from having to traverse back (or up) 2 steps to the page labeled 'Arts' and drill down from there 4 steps to the 'Arts: Music: Styles: D: Dance' topic. Therefore, this multiclassification links helps the user bypass 6 steps. Since the depth of the target of this multiclassification link is greater than that of its source, this multiclassification link brings the user deeper into the directory as it transitions them across it from one sub-category to another. Here, the *depth* of a topic page is defined as the number of hard links which must to followed from the root of the entire directory to reach it.

Fig. 6 illustrates another symbolic link used for multiclassification ODP. It is labeled 'Antiques@' and connects the page labeled 'Arts' in ODP to the page labeled 'Recreation: Antiques.' Again, this multiclassification link, like all multiclassification links, brings the user across the directory into a different category. However, unlike the multiclassification link shown in Fig. 5, this one connects two top-level categories of the entire (root) directory, namely 'Arts' to 'Recreation,' while the former bridges two distinct sub-categories of the 'Arts' category, namely, 'Performing Arts' to 'Music.' This multiclassification link prevents the user at the page labeled 'Arts' in ODP from having to traverse back (or up) one step to the root of the directory and drill down from there 2 steps to the 'Recreation: Antiques' topic. Therefore, this multiclassification link helps the user bypass 3 steps. Akin to Fig. 5, since the depth of the target of this multiclassification link is greater than that of its source, this multiclassification link brings the user deeper into the directory as it transitions them across it.

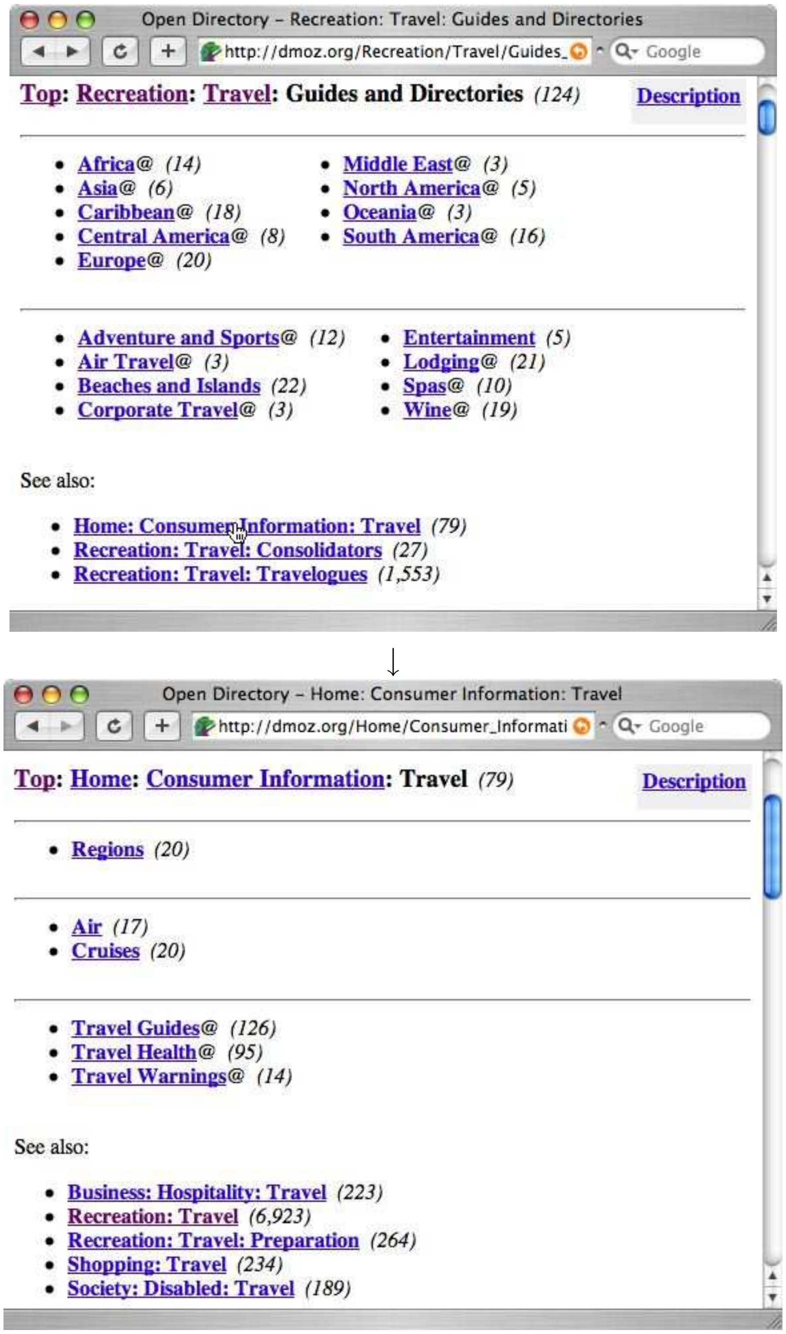
3.3 Cross-references

Multiclassification links are a type of transversal link (see Fig. 2). As mentioned in Section 2, another type of transversal link are cross-references such as those annotated with ‘see also:’ and frequently used in ODP. Fig. 7 (top) illustrates the use of a ‘see also:’ cross-reference link. Here the ‘Recreation: Travel: Guides and Directories’ category of ODP has 3 ‘see also:’ cross-reference links. The user selects the ‘see also:’ cross-reference link labeled ‘Home: Consumer Information: Travel’ which transfers the user to the ‘Home: Consumer Information: Travel’ category of ODP. Such links are not considered symbolic links in the ODP RDF markup where they are represented with the `<related/>` tag. These links act more as pure cross-references, and, specifically *implicit* or *isomorphic* links (DeRose, 1989), since, unlike symbolic links, they do not have an explicit label other than the label of the category to which they link. For example, the label of the ‘see also:’ cross-reference link followed in Fig. 7 (top) is ‘Home: Consumer Information: Travel’ while the label of the symbolic link followed in Fig. 6 (top) is ‘Antiques@’ even though it transfers the user to the ‘Recreation: Antiques’ category of ODP. In other words, the label of a ‘see also:’ cross-reference link informs the user of the link target *a priori*, i.e., before the user commits to following it. Symbolic links, on the other hand provide no such preview. The user cannot be sure as to where any symbolic link will lead until they arrive at the destination category. Perhaps this is a reflection of the fact that symbolic links used for multiclassification are intended to provide an illusion that information is classified in more than one category.

4 Results

4.1 Shortcuts and Backlinks

Table 3 provides the distribution of shortcuts and backlinks in ODP and within each of its top-level categories. Numbers typeset in bold font in the tables of results are those which we discuss in the body of the text. Since shortcuts and backlinks, by definition, start and end in the same top-level category, we do not specify whether the categories listed in Table 3 contain the source or target of each symbolic link counted there. Such detailed analysis will become relevant when we examine symbolic links used for multiclassification. As can be seen


Open Directory – Recreation: Travel: Guides and Directories
http://dmoz.org/Recreation/Travel/Guides_
Google

Top: Recreation: Travel: Guides and Directories *(124)* **Description**

- **Africa@** *(14)*
- **Asia@** *(6)*
- **Caribbean@** *(18)*
- **Central America@** *(8)*
- **Europe@** *(20)*
- **Middle East@** *(3)*
- **North America@** *(5)*
- **Oceania@** *(3)*
- **South America@** *(16)*

- **Adventure and Sports@** *(12)*
- **Air Travel@** *(3)*
- **Beaches and Islands** *(22)*
- **Corporate Travel@** *(3)*
- **Entertainment** *(5)*
- **Lodging@** *(21)*
- **Spas@** *(10)*
- **Wine@** *(19)*

See also:

- **Home: Consumer Information: Travel** *(79)*
- **Recreation: Travel: Consolidators** *(27)*
- **Recreation: Travel: Travelogues** *(1,553)*

↓

Open Directory – Home: Consumer Information: Travel
http://dmoz.org/Home/Consumer_Informati
Google

Top: Home: Consumer Information: Travel *(79)* **Description**

- **Regions** *(20)*

- **Air** *(17)*
- **Cruises** *(20)*

- **Travel Guides@** *(126)*
- **Travel Health@** *(95)*
- **Travel Warnings@** *(14)*

See also:

- **Business: Hospitality: Travel** *(223)*
- **Recreation: Travel** *(6,923)*
- **Recreation: Travel: Preparation** *(264)*
- **Shopping: Travel** *(234)*
- **Society: Disabled: Travel** *(189)*

Fig. 7. Illustration of a 'see also:' cross-reference, labeled 'Home: Consumer Information: Travel,' in the 'Recreation: Travel: Guides and Directories' category of ODP. This cross-reference indicates to the user that the current category, 'Recreation: Travel: Guides and Directories' (top), is related to the 'Home: Consumer Information: Travel' category (bottom).

from the table, shortcuts and backlinks account for a very small percentage ($< 3\%$) of the total number of symbolic links in the entire directory. Backlinks are nearly non-existent in the directory and account for only 51 of the total 748 205 symbolic links present (percentage omitted since it is $< 0.05\%$ in the

Table 3
Distribution of shortcuts and backlinks in ODP and within its top-level categories. Numbers typeset in bold font are discussed in the text. **Legend**: μ = mean; σ^2 = variance; σ = standard deviation.

				Sc Length			**Bl Length**	
Category	**#Scs**	**#Symlinks**	**%Scs**	μ	σ^2	**#Bls**	μ	σ^2
Adult	384	4 143	9.27%	2.04	0.13	0	–	–
Arts	(0) 584	(3) 88 513	(0%) 0.66%	2.54	3.86	(0) 0	–	–
Business	82	6 228	1.32%	1.90	0.19	2	2.00	2.00
Computers	(1) 221	(2) 7 656	(50%) 2.89%	(2) 2.25	1.05	(1) 1	(1) 1.00	–
Games	1 822	19 442	9.37%	2.81	3.30	0	–	–
Health	471	6 333	7.44%	1.52	1.74	1	1.00	–
Home	165	2 193	7.52%	3.84	2.37	1	2.00	–
Kids & Teens	251	**2 332**	**10.76%**	2.13	2.25	0	–	–
News	**4**	**1 049**	**0.38%**	1.75	0.25	0	–	–
Recreation	214	7 038	3.04%	4.20	3.75	3	1.00	0.00
Reference	380	17 747	2.14%	6.16	8.15	**1**[a]	**8.00**	–
Regional	**10 943**	**349 692**	**3.13%**	2.01	0.26	**22**	1.14	0.12
Science	495	8 241	6.01%	2.91	6.53	1	1.00	–
Shopping	154	3 239	4.75%	2.73	1.15	1	1.00	–
Society	760	24 217	3.14%	4.39	9.43	4	1.25	0.25
Sports	406	13 418	3.03%	3.98	4.88	1	1.00	–
World	3 906	**186 724**	2.09%	3.30	6.55	**13**	1.08	0.08
Total/ODP	(1) 21 242	(5) **748 205**	(20%) **2.84%**	(2) **2.58**	3.32	(1) **51**	(1) **1.29**	1.09
μ	1 249.53	44 012.06	4.53%	2.97	3.28	3.00	–	–
σ	2 667.41	91 449.53	3.26%	**1.21**	2.93	5.79	–	–

[a] indicates a backlink to the root of the entire directory: `http://dmoz.org`.

entire directory as well as in each category). There are no backlinks to the root of any top-level category. However, there is exactly one backlink to the root of the entire directory at topic 'Reference: Education: Colleges_and_Universities: North_America: United_States: Indiana: Bethel_College: Athletics.' This symbolic link is labeled '/NAIA/B@' and is the only backlink found in the 'Reference' category.

The columns labeled μ and σ^2 provide the average and variance, respectively, of the length of shortcuts and backlinks. Recall that the *length* of a shortcut or backlink is the number of non-symbolic links it bypasses on the (hard) path, without symbolic links, in which its source and target reside. Shortcuts have a greater length than backlinks on average (2.58 vs. 1.29) with low standard deviation (1.21) between the categories. This means that shortcuts in ODP bypass about two and a half hard links on average. Considering that ODP is a directory whose leaves have an average maximum depth of 10.65 (see Table 1), its shortcuts are not skipping past many hard links. In the entire directory, shortcuts range in length from 1–14 while most backlinks range from 1–3. The only backlink with a length greater than 3 is the only backlink to the root; it has length 8. A shortcut of length 1, of which there are 2 462 (11.59%), does not make the access path to its target any shorter than the hard link that ends there. We speculate that these shortcuts are merely provided to offer synonyms for the label of the corresponding hard link.

Table 4
Distribution of multiclassification links within and between the top-level categories of ODP. **Legend**: b/t = between; w/i = within; $R^{(w/i:b/t)}$ = ratio of the number of symbolic links within the specified top-level category to the number between top-level categories.

Category	**#Mc**	**%Mc**	**#b/t**	**%b/t**	**#w/i**	**%w/i**	$R^{(w/i:b/t)}$
Adult	3 759	90.73%	52	**1.38%**	3 707	**98.62%**	**71.29**
Arts	(3) 87 928	(100%) 99.34%	(1) 3 990	(33%) **4.54%**	(2) 83 938	(67%) **95.46%**	**21.04**
Business	6 144	99.65%	2 340	38.09%	3 804	61.91%	1.63
Computers	(0) 7 432	(0%) 97.07%	(0) 1 326	(0%) 17.84%	(0) 6 106	(0%) 82.16%	4.60
Games	17 614	90.60%	342	**1.94%**	17 272	**98.06%**	**50.50**
Health	5 861	92.55%	891	15.20%	4 970	84.80%	5.58
Home	2 027	92.43%	500	24.67%	1 527	75.33%	3.05
Kids&Teens	2 081	89.24%	0	**0.00%**	2 081	**100.00%**	–
News	1 045	99.62%	874	**83.64%**	171	**16.36%**	**0.20**
Recreation	6 820	96.90%	2 177	31.92%	4 643	68.08%	2.13
Reference	17 366	97.85%	14 477	**83.36%**	2 889	**16.64%**	**0.20**
Regional	338 627	96.84%	38 339	11.32%	300 288	88.68%	7.83
Science	7 744	93.97%	1 781	23.00%	5 963	77.00%	3.35
Shopping	3 083	95.18%	480	15.57%	2 603	84.43%	5.42
Society	23 431	96.75%	8 601	36.71%	14 830	63.29%	1.72
Sports	13 008	96.94%	3 148	24.20%	9 860	75.80%	3.13
World	182 801	97.90%	156	**0.09%**	182 645	**99.91%**	**1170.80**
Total/ODP	(3) 726 771	(60%) **97.14%**	(1) 79 474	(20%) **10.94%**	(2) **647 297**	(40%) **89.06%**	**8.14**
μ	42 751.24	**95.44%**	4 674.94	24.32%	38 076.29	75.68%	**84.53**
σ	88 861.71	**3.25%**	9 437.49	**25.46%**	81 668.38	**25.46%**	**290.37**

While it is difficult to draw any meaningful conclusions from such a small percentage of shortcuts and backlinks, we can make some cursory remarks. The topic with the smallest percentage (0.38%) of shortcuts and least number of shortcuts (4) and symbolic links (1 049) is 'News.' This result may be an indication that news aficionados do not wish to skip through their topic, but rather prefer a more lengthy and thorough treatment of the subject. Conversely, the topic with the greatest percentage of shortcuts (10.76%) is 'Kids and Teens.' This result may indicate that youngsters have a short attention span and, thus, prefer to jump around this topic, as opposed to more mature browsers. The 'Kids and Teens' topic also has the third smallest number of symbolic links (2 332). The topic with the greatest number of shortcuts (10 943), 'Regional,' also has the greatest number of symbolic links (349 692). However, the percentage of shortcuts (3.13%) in the 'Regional' topic is consistent with the average across all topics (2.84%). Clearly, the 'Regional' and 'World' categories are outliers with respect to the number of backlinks they contain (22 and 13, respectively). However, they also are the topics which contain the greatest (349 692) and second greatest (186 724) number of symbolic links, respectively. Since we see that only very few symbolic links are being used for shortcuts and backlinks, we know that a very high percentage are used for multiclassification.

4.2 Multiclassification Links

Table 4 provides the distribution of multiclassification links throughout ODP. Over 97% of the symbolic links in the entire directory are used for multiclassification. Moreover, greater than 95% of symbolic links in each category are for multiclassification on average with a very small standard deviation (3.25%). This result indicates that symbolic links in ODP are primarily available when users' initial predication of item placement is incorrect (as opposed to shortcuts and backlinks which always begin and end in the same hard path and, thus, same top-level category). While the percentages of multiclassification links is much higher than those of shortcuts and backlinks, we have a similar interpretation problem. The very high percentage of multiclassification links in ODP as well as the consistency among the very high percentages of multiclassification links in each of its top-level categories make it challenging to draw conclusions. Therefore, given this result, we turned our attention to studying the distribution of multiclassification links *within* and *between* top-level categories for a deeper analysis. We were interested in determining whether more multiclassification links reside within each of the given top-level categories or between top-level categories. Table 4 provides the results of this analysis as well.

Topic Connectivity within and between Top-Level Categories

We factored multiclassification links into those that end in a different top-level category than that in which they begin and those that end in the same in which they originate. The counts of multiclassification links in Table 4 are of those whose source resides in the specified category. From the column labeled **%w/i** of Table 4, we see that many more multiclassification links connect pages within the same category than between different categories (89.06% and 10.94%, respectively, for 8.14 times more in the whole directory and 84.53 times more on average in each category, though with a high standard deviation of 290.37). This means that while we know that a very high percentage of symbolic links are used for multiclassification, a large portion of these multiclassification links induce *local*, rather than *global*, connectivity. Note however, that the standard deviation of the percentage of multiclassification links within and between categories is high (25.46% in both cases).

These high standard deviations mean that there are some outliers. For instance, the categories 'News' and 'Reference' have a much higher percentage of multiclassification links to other top-level categories (83.64% and 83.36%, respectively) than to pages within their own category (16.36% and 16.64%, respectively); both categories have over 5 times more multiclassification links to other categories than within their category. We surmise that the reason

Table 5
Local degree of connectivity created by multiclassification links.

	Distinct		**Duplicate**		**Distinct**		
Category	**#b/t**	**%b/t**	**#w/i**	**%w/i**	**#w/i**	**#poss**	**%w/i**
Adult	**7**	**43.75%**	264	7.02%	38	132	28.79%
Arts	**15**	**93.75%**	49 301	**56.07%**	349	1 482	23.55%
Business	14	87.50%	1 411	22.97%	497	1 980	25.10%
Computers	16	100.00%	1 949	26.22%	282	1 806	15.61%
Games	**11**	**68.75%**	441	2.50%	83	552	15.04%
Health	13	81.25%	1 204	20.54%	248	1 260	19.68%
Home	13	81.25%	67	3.31%	39	342	11.40%
Kids & Teens	**0**	**0.00%**	397	19.08%	66	182	**36.26%**
News	14	87.50%	112	10.72%	35	380	9.21%
Recreation	14	87.50%	246	3.61%	81	992	8.17%
Reference	14	87.50%	**169**	**0.97%**	**31**	**506**	**6.13%**
Regional	14	87.50%	1 681	0.50%	167	1 260	13.25%
Science	14	87.50%	1 483	19.15%	395	2 550	15.49%
Shopping	13	81.25%	1 331	43.17%	325	1 190	27.31%
Society	15	93.75%	2 357	10.06%	292	870	33.56%
Sports	14	87.50%	2 221	17.07%	605	10 920	5.54%
World	**10**	**62.50%**	**28**	**0.02%**	**9**	**5 700**	**0.16%**
ODP	**211/272**	**77.57%**	–	–	–	–	–
μ	**12.41**	77.57%	3 803.65	**15.47%**	208.35	1 888.47	**17.31%**
σ	3.84	24.01%	11 751.83	**15.60%**	182.98	2 671.55	**10.40%**

behind this result is that ‘News’ and ‘Reference’ are general categories. Arguably, there can be ‘News’ and ‘Reference’ of all of the other categories of ODP. On the other hand, the ‘Adult,’ ‘Arts,’ ‘Games,’ ‘Kids and Teens,’ and ‘World’ categories each have a much higher percentage of multiclassification links within their category (98.62%, 95.46%, 98.06%, 100%, and 99.91%, respectively) than to other top-level categories (1.38%, 4.54%, 1.94%, 0%, and 0.09%, respectively) and are, thus, very isolated topics from the rest of the directory. The ‘Kid and Teens’ category is completely isolated; it has no multiclassification links to other top-level categories. This result may indicate that these categories represent highly-specialized topics. It should not come as a surprise that the ‘Adult,’ ‘Arts,’ and ‘Kids and Teens’ categories are highly isolated from the rest of the directory. However, we find it odd that the category ‘Games’ is isolated as it intuitively seems to have much overlap with the ‘Recreation’ category. We find the isolation of the category ‘World’ interesting. Its isolation may indicate that the categories of ODP are factored into a domestic vs. world dichotomy, i.e., all categories except ‘World’ deal with domestic items, while the ‘World’ category deals with international items. Studying isolated categories and sub-categories further may help us identify emerging cyber-communities (Kumar et al., 1999).

Local Connectivity

To gain a more detailed view of connectivity, we measured the number of distinct connections made between top-level categories as a fraction of the total possible. Table 5 provides this analysis. The column labeled **Distinct**

#b/t contains values representing the number of distinct categories to which a multiclassification link originating in the specified category reaches. Since there are 17 top-level categories in ODP, the percentages given in the column labeled **Distinct %b/t** are computed by dividing the associated number in the column labeled **Distinct #b/t** by 16 (discounting self-connections). These columns indicate how well the multiclassification links of ODP *cover* all possible connections between its top-level categories. There are 272 (= $(17 \times 17) - 17$) possible top-level category–category connections in ODP and multiclassification links induce 211 (77.57%) of those connections. Therefore, we see that while there are very few multiclassification links that span more than one top-level category (10.94%), those that do, cover over 77% of the connections possible in the entire directory. The 'Computers' category is connected to all other categories. Conversely, and again, the 'Kids and Teens' category is connected to no other categories. Each category is connected to 12.41 other categories on average.

For the most part, the trends in Table 4 are also seen in Table 5. For instance, the few multiclassification links in the categories 'Adult,' 'Games,' 'Kids and Teens,' and 'World' to other top-level categories do not cover many of the possible top-level category–category connections (7/16=43.75%, 11/16=68.75%, 0/16=0%, and 10/16=62.50%, respectively). The one category which does show a difference with the analysis in Table 4 is 'Arts.' The few multiclassification links in the 'Arts' category induce 15 of 16 (93.75%) possible top-level category–category connections. Perhaps this is because of all of the isolated categories identified above, 'Arts' is the least isolated (4.54%; see Table 4).

We conducted the same connectivity analysis as above within in each top-level category of ODP. In other words, we examined the large majority of multiclassification links which originate and terminate in the same category X to determine how well they cover the possible connections among the top-level categories of top-level category X. The column in Table 5 labeled **Duplicate #w/i** provides the absolute number of connections made between top-level categories of each top-level category. The column labeled **Distinct #w/i** provides the number of unique connections made between top-level categories of each top-level category. Here, *unique* means that a connection from sub-category X to sub-category Y (both with the same parent category) is counted only once even if there are $n \geqslant 1$ multiclassification links from sub-category X to sub-category Y. The column labeled **Distinct #poss** provides the number of $(n \times n) - n$ possible connections between the immediate sub-categories of each top-level category, where n is the number of top-level sub-categories in the specified top-level category. The columns labeled **Duplicate** provide an idea of how *densely* populated each top-level category is with multiclassification links spanning two distinct and immediate sub-categories. The columns labeled **Distinct** indicate how well these multiclassification links *cover* all possible connections between the top-level categories of each top-level category.

This analysis reveals that many of the sub-categories within a given category are poorly connected; only 17.31% of all possible distinct immediate (sub-category, sub-category) pairs within a given top-level category are connected on average. The percentage of multiclassification links spanning more than one immediate sub-category of any given top-level category is never greater than 56.07% (in the 'Arts' category). Many categories have a much lower percentage of multiclassification links reaching a different immediate sub-category than that in which they originated; the average percentage is 15.47%. Moreover, other than all being less than 56.07%, there is not much consistency in the percentages of these multiclassification links within each category; the standard deviation is 15.60%. The sum of the column in Table 5 labeled **Duplicate #w/i**, 64 662, represents the number of multiclassification links which connect two different top-level categories of each top-level category. These multiclassification links represent 10% of the total 647 297 multiclassification links originating and terminating within the same top-level category (see Table 4) and 9% of the total 748 205 symbolic links in the entire directory. Therefore, 90% of multiclassification links starting and ending in the same top-level category connect categories which share at least the first two levels of topic specificity; these links account for 77% of the total number of symbolic links in the entire directory.

The coverage of all possible distinct immediate sub-category–sub-category connections induced by these sparse multiclassification links in each top-level category is never greater than 36.26% (in the 'Kids and Teens' category). Again, other than being low, there is not much agreement in these percentages across all top-level categories; the standard deviation is 10.40%. While low overall, the relatively high percentage of multiclassification links connecting different immediate sub-categories of the 'Kids and Teens' category reinforces its isolation and cohesion in the directory. Moreover, this analysis provides additional evidence of the isolation of the 'World' category. It has only 28 (0.02%) multiclassification links connecting its immediate sub-categories of which only 9 (0.16%) connect distinct immediate sub-category–sub-category pairs of the 5 700 connections possible. The multiclassification links in the 'Reference' category follow a similar, but less extreme, trend, i.e., 169 (0.50%) multiclassification links connect its distinct immediate sub-categories of which only 31 (6.13%) connect distinct immediate sub-category–sub-category pairs of the 506 connections possible. Overall, these results indicate that while many of the symbolic links in ODP provide multiclassification, they are doing so at deeper levels of the directory, akin to shortcuts, but connecting two different hard paths, unlike shortcuts. Specifically, most (77%) of the symbolic links in ODP are used to create multiclassification between two specific topics within the same top-level category which share at least the first two levels of topic specialization.

Additional work is necessary to determine the specificity of the multiclassifi-

Table 6
Distribution of non-leaf and leaf pages as sources and targets of symbolic links. **Legend**: $R^{(s:t)}$ = ratio of total (non-leaf + leaf) symbolic link sources to targets.

		Source			**Target**		
Category	**#Non-leaf**	**#Leaf**	**#Total**	**#Non-leaf**	**#Leaf**	**#Total**	$\mathbf{R^{(s:t)}}$
Adult	4 143	0	4 143	4 104	0	4 104	1.01
Arts	(3) 88 363	(0) 150	(3) 88 513	(1) 109 542	(1) 181	(2) 109 723	(1.50) 0.81
Business	6 225	3	6 228	6 684	0	6 684	0.93
Computers	(1) 7 651	(1) 5	(2) 7 656	(1) 7 604	(2) 0	(3) 7 604	(0.66) 1.01
Games	19 440	2	19 442	19 718	0	19 718	0.99
Health	6 333	0	6 333	7 898	0	7 898	0.80
Home	2 193	0	2 193	2 473	0	2 473	0.89
Kids & Teens	2 332	0	2 332	2 647	0	2 647	0.88
News	1047	2	1 049	412	0	412	**2.55**
Recreation	7 038	0	7 038	8 434	0	8 434	0.83
Reference	17 746	1	17 747	11 540	0	11 540	1.54
Regional	349 647	45	349 692	320 272	0	320 272	1.09
Science	8 241	0	8 241	9 243	0	9 243	0.89
Shopping	3 229	10	3 239	5 344	23	5 367	0.60
Society	24 209	8	24 217	25 834	0	25 834	0.94
Sports	13 391	27	13 418	19 618	49	19 667	0.68
World	186 691	33	186 724	186 550	33	186 583	1.00
Total/ODP	(4) **747 918**	(1) **286**	(5) **748 205**	(2) **747 917**	(3) **286**	(5) **748 203**[a]	(1.00) 1.00
μ	43 995.24	16.82	44 012.06	43 995.12	16.82	44 011.94	**1.03**
σ	91 433.62	36.86	91 449.53	86 032.03	44.70	86 042.57	**0.44**

[a] Since the target of one symbolic link is the root of the directory and the target of another is the hidden category 'Netscape,' there are two more sources (748 205) than targets (748 203).

cation source/targets topics to develop a better understanding of the effect of symbolic links on topic connectivity. Table 6 provides a start to such analysis. The table factors the source and target of all symbolic links, within each top-level category as well as the entire directory, into those which start/end at a non-leaf and leaf. Nearly all symbolic links (99.96%) start and end at non-leafs. Furthermore, we see that each category has about the same number symbolic link sources as targets; the ratio of sources to targets is 1.03 on average in each category, though with a high standard deviation of 0.44. The 'News' category is an outlier; it has more than two and a half (2.55) as many symbolic link sources than targets. The 'Reference' category also has a high ratio of sources to targets (1.55). This is because, as shown in Table 4, the 'News' and 'Reference' categories have a much higher percentage of multiclassification links to other top-level categories than to pages within their own category. Bear in mind that the ratios given in Table 6 imply nothing about topic connectivity within and between top-level categories; Table 4 provides such information. We shall have more to say about multiclassification source/target topic specificity when we discuss future work below.

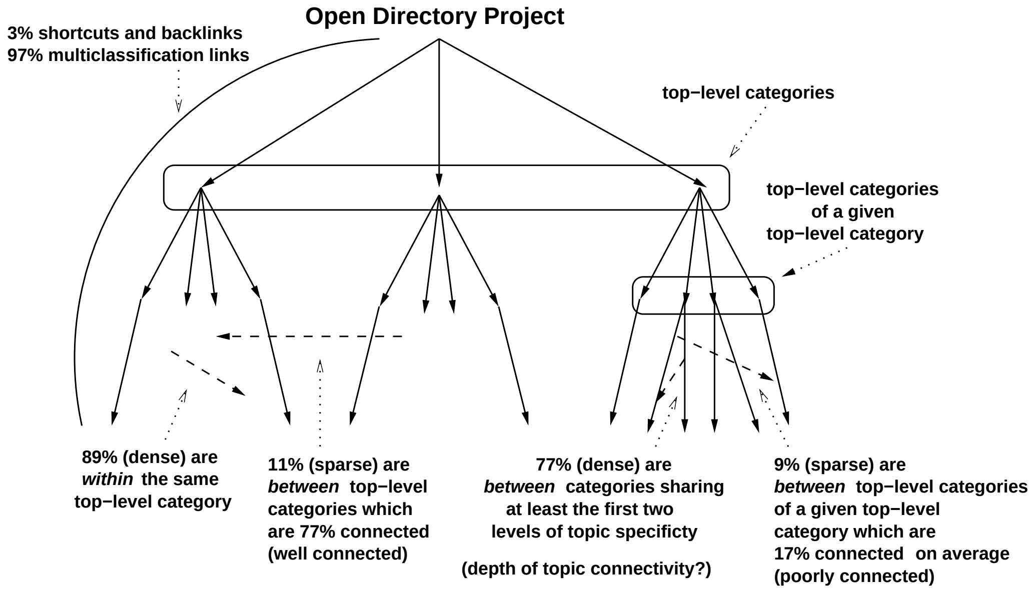

Open Directory Project
3% shortcuts and backlinks
97% multiclassification links
top−level categories
top−level categories of a given top−level category
89% (dense) are *within* the same top−level category
11% (sparse) are *between* top−level categories which are 77% connected (well connected)
77% (dense) are *between* categories sharing at least the first two levels of topic specificty
(depth of topic connectivity?)
9% (sparse) are *between* top−level categories of a given top−level category which are 17% connected on average (poorly connected)

Fig. 8. Graphical summary of the results. Solid lines represent hard links/paths, dashed lines represent multiclassification links, and dotted lines are for annotation purposes.

4.3 Summary

In summary, our results indicates the following:

- nearly all (> 97%) of the symbolic links in ODP create multiclassification,
- most (> 89%) of those multiclassification links connect topics within the *same* top-level category of the root rather than bridging two distinct top-level categories,
- while the fraction of total multiclassification links that connect two distinct top-level categories is very small (< 11%), those links cover over 77% of the possible, distinct top-level category–category connections,
- while the fraction of total multiclassification links that connect two distinct topics (on different hard paths) within the same top-level category is very large (> 89%), only a small percentage (10%) of those connect two distinct immediate sub-categories of the same top-level category (< 9% of all symbolic links),
- therefore, the majority of symbolic links (> 77%) are multiclassification links which connect two categories which share at least the first two levels of topic specificity.

Fig. 8 provides a graphically summary of our results.

5 Discussion

Many researchers have identified link typing as an under-studied, but important, research topic for the web (Bar-Ilan, 2005; Kopak, 1999; Henzinger et al., 2002). However, most work focuses on node typing and the little work that does concentrate on link typing, covers semantic link typing. Similarly, though symbolic links are ubiquitous in web directories, surprisingly, our literature searches revealed no studies of symbolic links, a structural type of link, their uses, or effect on information retrieval, and only few articles which explicitly mentioned symbolic links at all (Balasubramanian et al., 1997; Bieber et al., 1997). Therefore, the elementary understanding of the uses of symbolic links on the web, albeit the frequency of their use in important web directories to reduce redundancy, including Google (`http://directory.google.com`) and Yahoo!, which are used by many as hubs to sites of interest, provides the motivation for our work. We anticipate that our research will serve as a starting point to help address this long unfilled void in the literature. For these reasons we feel that our study is worthwhile and particularly timely.

5.1 Contributions

We have illustrated that symbolic links are distinct from cross-references and situated symbolic links *vis à vis* other semantic and structural link types from hypermedia. We identified 3 uses (or types) of symbolic links in web directories, namely *shortcuts*, *backlinks*, and *multiclassification* links, and established their distribution in the Open Directory Project (ODP), the largest, most comprehensive, and most widely distributed human-compiled taxonomy of links to websites (The Open Directory Project, 2002). Since it uses symbolic links frequently, ODP provides an attractive data source from which to analyze these uses. Our results indicate that symbolic links in ODP are used most frequently for multiclassification. However, only few multiclassification links span top- and second-level categories. This indicates that most symbolic links in ODP are used to create multiclassification between two topics which are nested more than two levels deep and suggests that there may be multiple uses of multiclassification links.

An improved understanding of the uses of symbolic links is a necessary step toward improving information access in web directories. Our results can be used by designers of directories to evaluate how close the actual uses of symbolic links in ODP meet their intended uses, to more effectively employ symbolic links to meet those uses during directory construction and maintenance, and to better organize information in the directory for improving information access. Moreover, ‘many web users complain that sometimes they do not really know where a link will take them' (Bieber et al., 1997). Armed with our results, taxonomy designers can better annotate symbolic links to imply type, and, thereby, aid navigation akin to (Noirhomme-Fraiture and Serpe, 1998). For instance, authors can augment the label of a backlink, shortcut, and multiclassification link with '@up,' '@down,' and '@across,' respectively. We can even convey more information by additionally indicating how many steps the symbolic link obviates, e.g., '@up2' (backlink of length 2), '@down10' (shortcut of length 10), and '@across(up2_and_down4)' (multiclassification link up 2 steps and then down 4). This approach can give the user a preview of the target's location and generality/specificity (relative to the current page) before following the symbolic link. Like Bieber, Vitali, Ashman, Balasubramanian, and Oinas-Kiukkonen (1997), 'we view this, in part, as both a user-interface design question and a hypermedia design question.' Lastly, an improved grasp of the uses of symbolic links on the web helps us to better relate them to faceted browsing and search techniques, such as *zoom* (Sacco, 2000) and *out-of-turn interaction* (Narayan et al., 2004), for interacting with taxonomies which serve related purposes.

Our results are widely applicable since ODP 'powers the core directory services for the web's largest and most popular search engines and portals' (The Open Directory Project, 2002) and all of its data is available free to the public. We anticipate that the results and relationships identified and discussed in this paper will provide a foundation for (1) users for understanding the usages of symbolic links in a directory, (2) designers to employ symbolic links more effectively when building and maintaining directories and for crafting user interfaces to directories, and (3) information retrieval researchers for further study of symbolic links in web directories.

5.2 Future Work

This research can be extended in multiple directions.

Specificity of Multiclassification Sources and Targets

We can study the specificity/depth of the topics connected as described above for a more detailed analysis of topic connectivity. Specifically, we need to study the specificity of the topics where multiclassification links start and end to develop a better understanding of the effect of symbolic links on topic connectivity. Further, since nearly all symbolic links start and end at non-leaves (see Table 6), additional work is necessary to study the depth of those non-leaf sources and targets. To conduct such an analysis, we can plot two histograms: one for multiclassification sources and one for targets, where the x-axis describes the depth of the page and the y-axis provides the number of multiclassification links which start and end, respectively, at that depth. From the results in Table 4, we know that multiclassification links in ODP are not *top-heavy*. This analysis helps us determine if the multiclassification links in ODP are *middle-heavy* or *bottom-heavy*.

Alternate View of Multiclassification

An alternate way to study the multiclassification created by multiclassification links is through the number of additional paths they induce from the root of a directory to each page containing links to external sites. We can then think of each such path as an index to a particular leaf. In this setting, multiclassification links, by increasing the number of paths which index a particular leaf, increase opportunity for information access. Now a leaf is multiclassified by several paths to it from the root. For instance, the leaves numbered 9, 12, and 14 in Fig. 1 are each indexed by 2 paths, while all the other leaves are indexed by 1 path. To study this effect of multiclassification links we could plot a histogram where the x-axis describes the number of paths (x) to a leaf and the y-axis measures the number of documents/leaves indexed by x paths. Such an analysis might provide an alternative way to identify *authoritative* websites (Kleinberg, 1999), i.e., those indexed by the greatest number of paths. In the absence of symbolic links, this histogram would contain a vertical line from the point (1, 0) to the point (1, 525,661)[6].

Human-Computer Interaction

The reader will have noticed that the usages of symbolic links showcased here are presented from a design, rather than user, perspective. Determining how users employ symbolic links in information seeking and whether users prefer to interact with web directories through symbolic links or faceted browsing or search interfaces to achieve the same ends are still open research issues.

Our work has laid the foundation for the study of these issues and fits into the larger and long-term goal of accommodating multiple user mental models of information seeking through the design of information, faceted browsing and search techniques, and user interfaces.

---

[6] This is the number of leaves in ODP.

References

Allan, J., 1996. Automatic Hypertext Link Typing. In: Proceedings of the Seventh Annual ACM Conference on Hypertext. ACM Press, New York, NY, pp. 42–52.

Baeza-Yates, R., Ribeiro-Neto, B., 1999. Modern Information Retrieval. Addison-Wesley, Boston, MA.

Balasubramanian, V., Bashian, A., Porcher, D., 1997. A Large-scale Hypermedia Application using Document Management and Web Technologies. In: Bernstein, M., Carr, L., Østerbye, K. (Eds.), Proceedings of the Eighth Annual ACM Conference on Hypertext. ACM Press, New York, NY, pp. 134–145.

Bar-Ilan, J., 2005. What Do We Know about Links and Linking? A Framework for Studying Links in Academic Environments. Information Processing and Management 41, 973–986.

Bieber, M., Vitali, F., Ashman, H., Balasubramanian, V., Oinas-Kiukkonen, H., 1997. Fourth Generation Hypermedia: Some Missing Links for the World Wide Web. International Journal of Human-Computer Studies 47, 31–65.

Brin, S., Page, L., 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems 30, 107–117.

Conklin, J., 1987. Hypertext: An Introduction and Survey. IEEE Computer 20 (9), 17–41.

Date, C., 1993. An Introduction to Database Systems. Addison-Wesley, Boston, MA.

DeRose, S., 1989. Expanding the Notion of Links. In: Proceedings of the Second Annual ACM Conference on Hypertext. ACM Press, New York, NY, pp. 249–257.

Dourish, P., Edwards, W., 2000. Extending Document Management Systems with User-Specific Active Properties. ACM Transactions on Information Systems 18 (2), 140–177.

Hearst, M., 1999. Modern Information Retrieval. Addison-Wesley, Boston, MA, Ch. 10: User Interfaces and Visualization, pp. 257–323.

Helic, D., Maurer, H., Scherbakov, N., 1999. Introducing Hypermedia Composites to WWW. Journal of Network and Computer Applications 22, 19–32.

Henzinger, M., Motwani, R., Silverstein, C., 2002. Challenges in Web Search Engines. SIGIR Forum 36 (2), 11–22.

Kleinberg, J., 1999. Authoritative Sources in a Hyperlinked Environment. Journal of the ACM 46 (5), 604–632.

Kopak, R., 1999. Functional Link Typing in Hypertext. ACM Computing Surveys 31 (4es), Article No. 16.

Kumar, R., Raghavan, P., Rajagopalan, S., Tomkins, A., 1999. Trawling the Web for Emerging Cyber-Communities. In: Mendelzon, A. (Ed.), Proceedings of the Eighth International World Wide Web Conference (WWW). Elsevier Science, New York, NY.

Lempel, R., Moran, S., 2001. SALSA: The Stochastic Approach for Link-structure Analysis. ACM Transactions on Information Systems 19 (2), 131–160.

Marsden, G., Cairns, D., 2003. Improving the Usability of the Hierarchical File System. In: Eloff, J., Engelbrecht, A., Kotzé, P., Eloff, M. (Eds.), Proceedings of the Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement through Technology (SAICSIT). South African Institute for Computer Scientists and Information Technologists, Republic of South Africa, pp. 122–129.

Narayan, M., Williams, C., Perugini, S., Ramakrishnan, N., 2004. Staging Transformations for Multimodal Web Interaction Management. In: Feldman, S., Uretsky, M., Najork, M., Wills, C. (Eds.), Proceedings of the Thirteenth ACM International World Wide Web Conference (WWW). ACM Press, New York, NY, pp. 212–223.

Noirhomme-Fraiture, M., Serpe, V., 1998. Visual Representation of Hypermedia Links According to Their Types. In: Catarci, T., Costabile, M., Santucci, G., Tarantino, L. (Eds.), Proceedings of the Working Conference on Advanced Visual Interfaces (AVI). ACM Press, New York, NY, pp. 146–155.

Sacco, G., 2000. Dynamic Taxonomies: A Model for Large Information Bases. IEEE Transactions on Knowledge and Data Engineering 12 (3), 468–479.

The Open Directory Project, 2002. About the open directory project. Retrieved May 30, 2007, from `http://dmoz.org/about.html`.

Trigg, R., 1983. A Networked Approach to Text Handling for the Online Scientific Community. PhD dissertation, University of Maryland.

Tables

Table 1

Category	**#Non-leaves**	**#Leaves**	**#Topics**	**#Terms**	μ **C/N**	**Depth range**
Adult	2 085	5 887	7 972	2 191	3.82	[2–11]
Arts	(3) 8 344	(4) 38 525	(7) 46 869	(8) 28 595	(2.33) 5.62	([3–3]) [2–11]
Business	1 977	9 207	11 184	4 311	5.66	[2–10]
Computers	(3) 1 637	(3) 6 478	(6) 8 115	(6) 3 730	(1.66) 4.96	([3–3]) [2–10]
Games	2 775	8 652	11 427	7 013	4.12	[2–11]
Health	999	5 534	6 533	3 276	6.54	[2–9]
Home	493	2 026	2 519	1 667	5.11	[2–8]
Kids & Teens	837	3 220	4 057	3 346	4.85	[2–11]
News	266	1 583	1 849	338	6.95	[2–7]
Recreation	1 782	8 770	10 552	3 539	5.92	[2–10]
Reference	2 747	8 798	11 545	6 650	4.20	[2–12]
Regional	87 262	210 130	297 392	48 429	3.41	[2–14]
Science	2 205	9 423	11 628	7 965	5.27	[2–11]
Shopping	1 173	4 153	5 326	3 132	4.54	[2–10]
Society	4 532	23 083	27 615	12 370	6.09	[3–12]
Sports	2 329	15 446	17 775	7 473	7.63	[2–10]
World	42 478	164 746	207 224	105 498	4.88	[3–14]
ODP	(7) 163 922	(7) 525 661	(14) 689 583	(13) 218 640	(1.86) 4.21	([3–3]) [2–14]
μ	9 642.41	30 921.24	40 562.65	14 677.82	5.27	[2.12–**10.65**]
σ	22 321.01	60 125.68	81 986.75	26 316.08	1.13	[0.33–1.80]

Table 2

Category	μ **C/N**$^{(w/o)}$	μ **C/N**$^{(w/)}$	**R**$^{(w/:w/o)}$
Adult	3.82	5.81	1.52
Arts	(2.33) 5.62	(3.00) 16.22	(1.29) 2.89
Business	5.66	8.81	1.56
Computers	(1.66) 4.96	(2.33) 9.63	(1.40) 1.94
Games	4.12	11.12	2.70
Health	6.54	12.88	1.97
Home	5.11	9.56	1.87
Kids & Teens	4.85	7.63	1.57
News	6.95	10.89	1.57
Recreation	5.92	9.87	1.67
Reference	4.20	10.66	2.54
Regional	3.41	7.42	2.18
Science	5.27	9.01	1.71
Shopping	4.54	7.30	1.61
Society	6.09	11.44	1.88
Sports	7.63	13.39	1.75
World	4.88	9.27	1.90
ODP	(1.86) 4.21	(2.57) 8.77	(1.38) 2.08
μ	5.27	10.05	1.93
σ	1.13	2.54	0.42

Table 3

				Sc Length			Bl Length	
Category	**#Scs**	**#Symlinks**	**%Scs**	μ	σ^2	**#Bls**	μ	σ^2
Adult	384	4 143	9.27%	2.04	0.13	0	–	–
Arts	(0) 584	(3) 88 513	(0%) 0.66%	2.54	3.86	(0) 0	–	–
Business	82	6 228	1.32%	1.90	0.19	2	2.00	2.00
Computers	(1) 221	(2) 7 656	(50%) 2.89%	(2) 2.25	1.05	(1) 1	(1) 1.00	–
Games	1 822	19 442	9.37%	2.81	3.30	0	–	–
Health	471	6 333	7.44%	1.52	1.74	1	1.00	–
Home	165	2 193	7.52%	3.84	2.37	1	2.00	–
Kids & Teens	251	**2 332**	**10.76%**	2.13	2.25	0	–	–
News	**4**	**1 049**	**0.38%**	1.75	0.25	0	–	–
Recreation	214	7 038	3.04%	4.20	3.75	3	1.00	0.00
Reference	380	17 747	2.14%	6.16	8.15	**1**[a]	**8.00**	–
Regional	**10 943**	**349 692**	**3.13%**	2.01	0.26	**22**	1.14	0.12
Science	495	8 241	6.01%	2.91	6.53	1	1.00	–
Shopping	154	3 239	4.75%	2.73	1.15	1	1.00	–
Society	760	24 217	3.14%	4.39	9.43	4	1.25	0.25
Sports	406	13 418	3.03%	3.98	4.88	1	1.00	–
World	3 906	**186 724**	2.09%	3.30	6.55	**13**	1.08	0.08
Total/ODP	(1) 21 242	(5) **748 205**	(20%) **2.84%**	(2) **2.58**	3.32	(1) **51**	(1) **1.29**	1.09
μ	1 249.53	44 012.06	4.53%	2.97	3.28	3.00	–	–
σ	2 667.41	91 449.53	3.26%	**1.21**	2.93	5.79	–	–

Table 4

Category	**#Mc**	**%Mc**	**#b/t**	**%b/t**	**#w/i**	**%w/i**	$\mathbf{R}^{(\mathbf{w/i:b/t})}$
Adult	3 759	90.73%	52	**1.38%**	3 707	**98.62%**	**71.29**
Arts	(3) 87 928	(100%) 99.34%	(1) 3 990	(33%) **4.54%**	(2) 83 938	(67%) **95.46%**	**21.04**
Business	6 144	99.65%	2 340	38.09%	3 804	61.91%	1.63
Computers	(0) 7 432	(0%) 97.07%	(0) 1 326	(0%) 17.84%	(0) 6 106	(0%) 82.16%	4.60
Games	17 614	90.60%	342	**1.94%**	17 272	**98.06%**	**50.50**
Health	5 861	92.55%	891	15.20%	4 970	84.80%	5.58
Home	2 027	92.43%	500	24.67%	1 527	75.33%	3.05
Kids&Teens	2 081	89.24%	0	**0.00%**	2 081	**100.00%**	–
News	1 045	99.62%	874	**83.64%**	171	**16.36%**	**0.20**
Recreation	6 820	96.90%	2 177	31.92%	4 643	68.08%	2.13
Reference	17 366	97.85%	14 477	**83.36%**	2 889	**16.64%**	**0.20**
Regional	338 627	96.84%	38 339	11.32%	300 288	88.68%	7.83
Science	7 744	93.97%	1 781	23.00%	5 963	77.00%	3.35
Shopping	3 083	95.18%	480	15.57%	2 603	84.43%	5.42
Society	23 431	96.75%	8 601	36.71%	14 830	63.29%	1.72
Sports	13 008	96.94%	3 148	24.20%	9 860	75.80%	3.13
World	182 801	97.90%	156	**0.09%**	182 645	**99.91%**	**1170.80**
Total/ODP	(3) 726 771	(60%) **97.14%**	(1) 79 474	(20%) **10.94%**	(2) **647 297**	(40%) **89.06%**	**8.14**
μ	42 751.24	**95.44%**	4 674.94	24.32%	38 076.29	75.68%	**84.53**
σ	88 861.71	**3.25%**	9 437.49	**25.46%**	81 668.38	**25.46%**	**290.37**

Table 5

	Distinct		**Duplicate** (Density)		**Distinct** (Coverage)		
Category	**#b/t**	**%b/t**	**#w/i**	**%w/i**	**#w/i**	**#poss**	**%w/i**
Adult	**7**	**43.75%**	264	7.02%	38	132	28.79%
Arts	**15**	**93.75%**	49 301	**56.07%**	349	1 482	23.55%
Business	14	87.50%	1 411	22.97%	497	1 980	25.10%
Computers	16	100.00%	1 949	26.22%	282	1 806	15.61%
Games	**11**	**68.75%**	441	2.50%	83	552	15.04%
Health	13	81.25%	1 204	20.54%	248	1 260	19.68%
Home	13	81.25%	67	3.31%	39	342	11.40%
Kids & Teens	**0**	**0.00%**	397	19.08%	66	182	**36.26%**
News	14	87.50%	112	10.72%	35	380	9.21%
Recreation	14	87.50%	246	3.61%	81	992	8.17%
Reference	14	87.50%	**169**	**0.97%**	**31**	**506**	**6.13%**
Regional	14	87.50%	1 681	0.50%	167	1 260	13.25%
Science	14	87.50%	1 483	19.15%	395	2 550	15.49%
Shopping	13	81.25%	1 331	43.17%	325	1 190	27.31%
Society	15	93.75%	2 357	10.06%	292	870	33.56%
Sports	14	87.50%	2 221	17.07%	605	10 920	5.54%
World	**10**	**62.50%**	**28**	**0.02%**	**9**	**5 700**	**0.16%**
ODP	**211/272**	**77.57%**	–	–	–	–	–
μ	**12.41**	77.57%	3 803.65	**15.47%**	208.35	1 888.47	**17.31%**
σ	3.84	24.01%	11 751.83	**15.60%**	182.98	2 671.55	**10.40%**

Table 6

		Source			Target		
Category	**#Non-leaf**	**#Leaf**	**#Total**	**#Non-leaf**	**#Leaf**	**#Total**	$\mathbf{R}^{(\mathbf{s:t})}$
Adult	4 143	0	4 143	4 104	0	4 104	1.01
Arts	(3) 88 363	(0) 150	(3) 88 513	(1) 109 542	(1) 181	(2) 109 723	(1.50) 0.81
Business	6 225	3	6 228	6 684	0	6 684	0.93
Computers	(1) 7 651	(1) 5	(2) 7 656	(1) 7 604	(2) 0	(3) 7 604	(0.66) 1.01
Games	19 440	2	19 442	19 718	0	19 718	0.99
Health	6 333	0	6 333	7 898	0	7 898	0.80
Home	2 193	0	2 193	2 473	0	2 473	0.89
Kids & Teens	2 332	0	2 332	2 647	0	2 647	0.88
News	1047	2	1 049	412	0	412	**2.55**
Recreation	7 038	0	7 038	8 434	0	8 434	0.83
Reference	17 746	1	17 747	11 540	0	11 540	1.54
Regional	349 647	45	349 692	320 272	0	320 272	1.09
Science	8 241	0	8 241	9 243	0	9 243	0.89
Shopping	3 229	10	3 239	5 344	23	5 367	0.60
Society	24 209	8	24 217	25 834	0	25 834	0.94
Sports	13 391	27	13 418	19 618	49	19 667	0.68
World	186 691	33	186 724	186 550	33	186 583	1.00
Total/ODP	(4) **747 918**	(1) **286**	(5) **748 205**	(2) **747 917**	(3) **286**	(5) **748 203**[a]	(1.00) 1.00
μ	43 995.24	16.82	44 012.06	43 995.12	16.82	44 011.94	**1.03**
σ	91 433.62	36.86	91 449.53	86 032.03	44.70	86 042.57	**0.44**

Figures

Figure 1

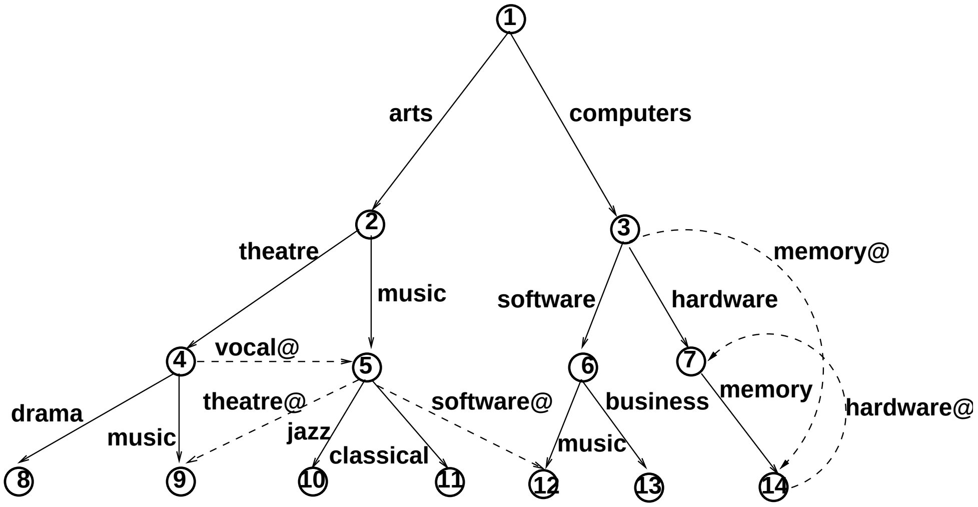


Figure 2

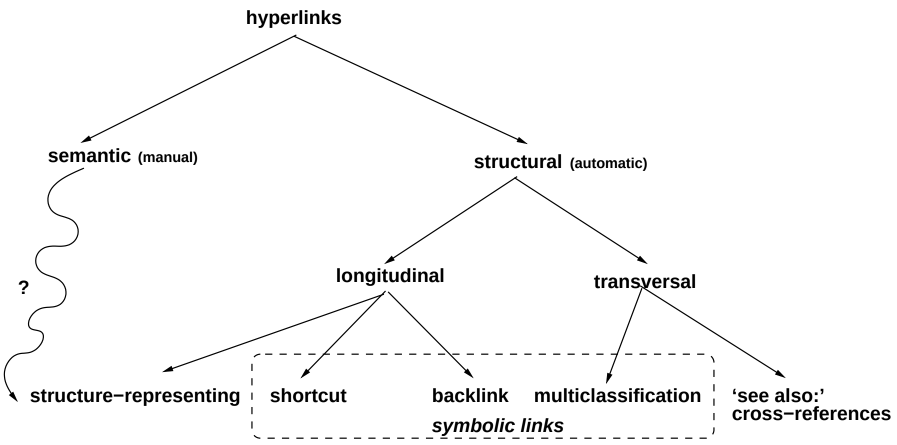
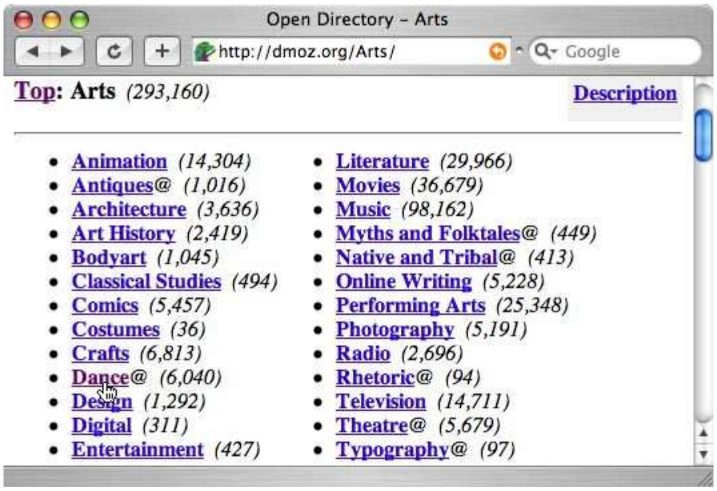

hyperlinks
semantic (manual)
structural (automatic)
?
longitudinal
transversal
structure−representing
shortcut
backlink
multiclassification
'see also:' cross−references
symbolic links

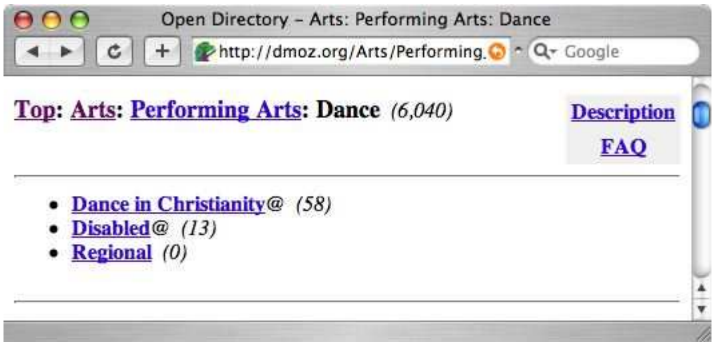
Figure 3



Top: **Arts** *(293,160)*

- **Animation** *(14,304)*
- **Antiques**@ *(1,016)*
- **Architecture** *(3,636)*
- **Art History** *(2,419)*
- **Bodyart** *(1,045)*
- **Classical Studies** *(494)*
- **Comics** *(5,457)*
- **Costumes** *(36)*
- **Crafts** *(6,813)*
- **Dance**@ *(6,040)*
- **Design** *(1,292)*
- **Digital** *(311)*
- **Entertainment** *(427)*
- **Literature** *(29,966)*
- **Movies** *(36,679)*
- **Music** *(98,162)*
- **Myths and Folktales**@ *(449)*
- **Native and Tribal**@ *(413)*
- **Online Writing** *(5,228)*
- **Performing Arts** *(25,348)*
- **Photography** *(5,191)*
- **Radio** *(2,696)*
- **Rhetoric**@ *(94)*
- **Television** *(14,711)*
- **Theatre**@ *(5,679)*
- **Typography**@ *(97)*

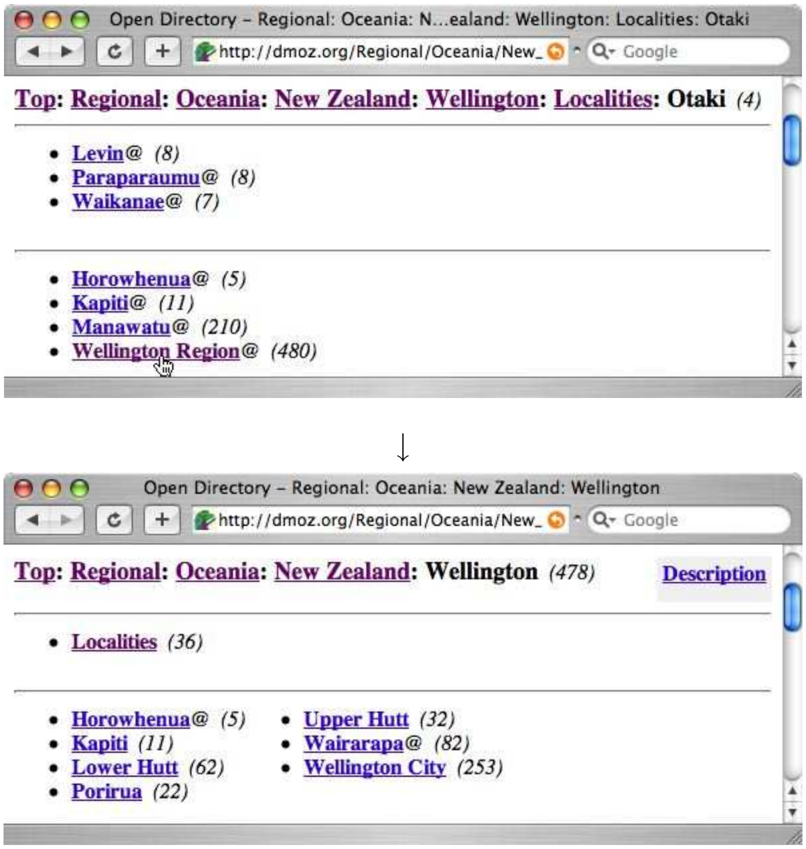
↓


Open Directory - Arts: Performing Arts: Dance
http://dmoz.org/Arts/Performing
Google
Description
FAQ

Top: **Arts**: **Performing Arts**: **Dance** *(6,040)*

- **Dance in Christianity**@ *(58)*
- **Disabled**@ *(13)*
- **Regional** *(0)*

Figure 4


Open Directory – Regional: Oceania: N...ealand: Wellington: Localities: Otaki
http://dmoz.org/Regional/Oceania/New_
Google

Top: **Regional**: **Oceania**: **New Zealand**: **Wellington**: **Localities**: **Otaki** *(4)*

- **Levin**@ *(8)*
- **Paraparaumu**@ *(8)*
- **Waikanae**@ *(7)*

- **Horowhenua**@ *(5)*
- **Kapiti**@ *(11)*
- **Manawatu**@ *(210)*
- **Wellington Region**@ *(480)*

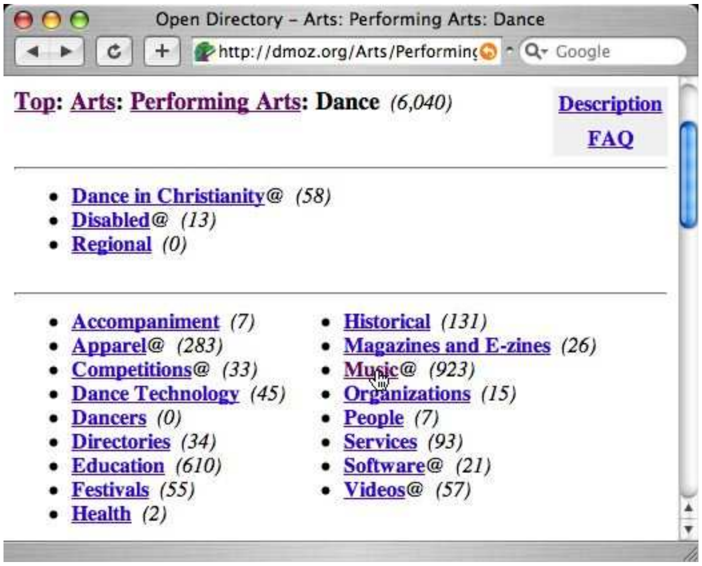
↓
Open Directory – Regional: Oceania: New Zealand: Wellington
http://dmoz.org/Regional/Oceania/New_
Google

Top: **Regional**: **Oceania**: **New Zealand**: **Wellington** *(478)* **Description**

- **Localities** *(36)*

- **Horowhenua**@ *(5)*
- **Kapiti** *(11)*
- **Lower Hutt** *(62)*
- **Porirua** *(22)*
- **Upper Hutt** *(32)*
- **Wairarapa**@ *(82)*
- **Wellington City** *(253)*

Figure 5


Open Directory – Arts: Performing Arts: Dance
http://dmoz.org/Arts/Performing
Google

Top: Arts: Performing Arts: **Dance** *(6,040)*

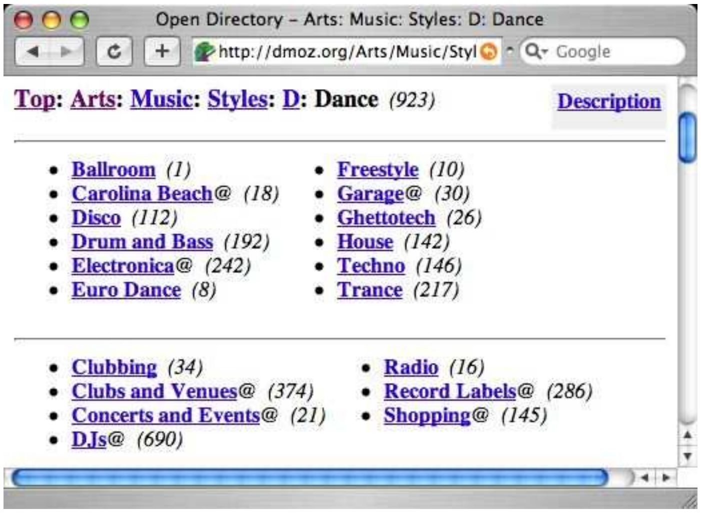
Description

FAQ

- Dance in Christianity@ *(58)*
- Disabled@ *(13)*
- Regional *(0)*

- Accompaniment *(7)*
- Apparel@ *(283)*
- Competitions@ *(33)*
- Dance Technology *(45)*
- Dancers *(0)*
- Directories *(34)*
- Education *(610)*
- Festivals *(55)*
- Health *(2)*
- Historical *(131)*
- Magazines and E-zines *(26)*
- Music@ *(923)*
- Organizations *(15)*
- People *(7)*
- Services *(93)*
- Software@ *(21)*
- Videos@ *(57)*

↓


Open Directory – Arts: Music: Styles: D: Dance
http://dmoz.org/Arts/Music/Styl
Google

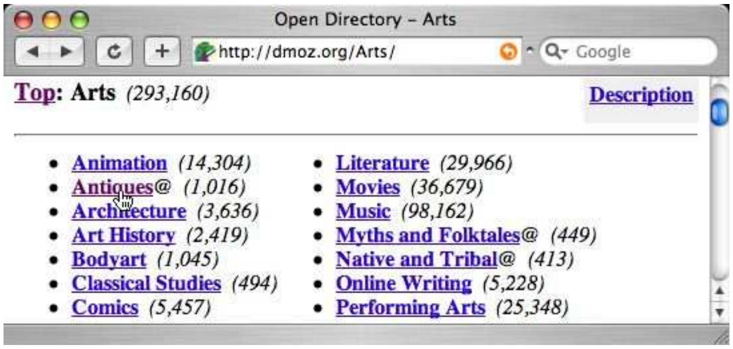
Top: Arts: Music: Styles: D: **Dance** *(923)*

Description

- Ballroom *(1)*
- Carolina Beach@ *(18)*
- Disco *(112)*
- Drum and Bass *(192)*
- Electronica@ *(242)*
- Euro Dance *(8)*
- Freestyle *(10)*
- Garage@ *(30)*
- Ghettotech *(26)*
- House *(142)*
- Techno *(146)*
- Trance *(217)*

- Clubbing *(34)*
- Clubs and Venues@ *(374)*
- Concerts and Events@ *(21)*
- DJs@ *(690)*
- Radio *(16)*
- Record Labels@ *(286)*
- Shopping@ *(145)*

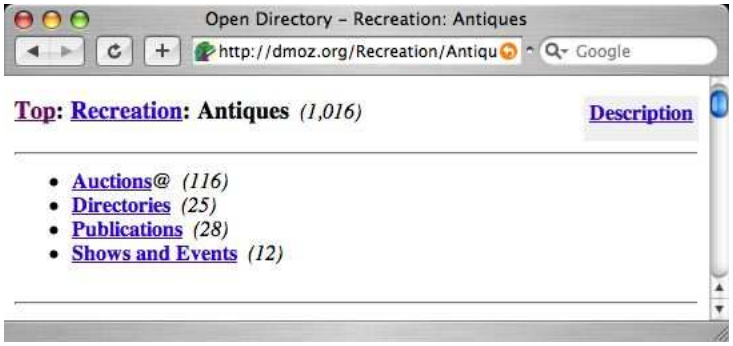
Figure 6



Top: **Arts** *(293,160)* **Description**

- **Animation** *(14,304)*
- **Antiques**@ *(1,016)*
- **Architecture** *(3,636)*
- **Art History** *(2,419)*
- **Bodyart** *(1,045)*
- **Classical Studies** *(494)*
- **Comics** *(5,457)*
- **Literature** *(29,966)*
- **Movies** *(36,679)*
- **Music** *(98,162)*
- **Myths and Folktales**@ *(449)*
- **Native and Tribal**@ *(413)*
- **Online Writing** *(5,228)*
- **Performing Arts** *(25,348)*

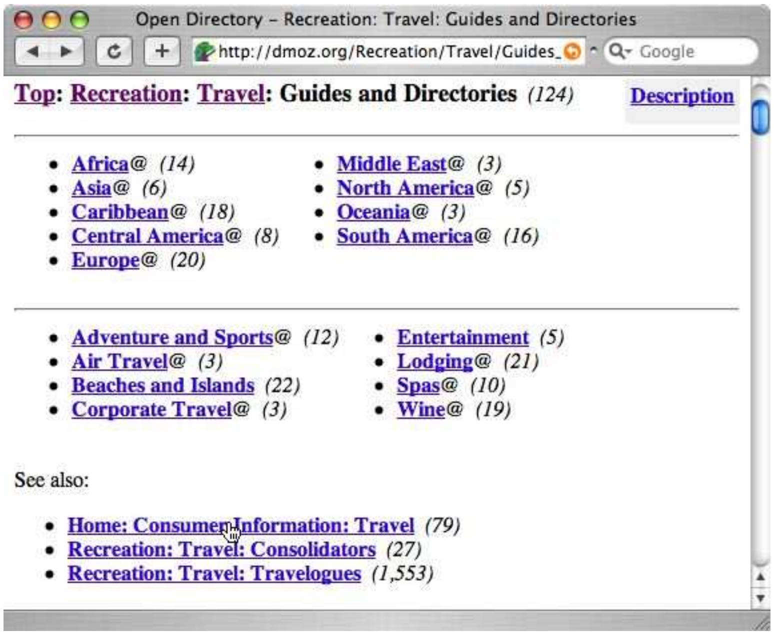
↓


Open Directory – Recreation: Antiques
http://dmoz.org/Recreation/Antiqu
Google

Top: **Recreation**: **Antiques** *(1,016)* **Description**

- **Auctions**@ *(116)*
- **Directories** *(25)*
- **Publications** *(28)*
- **Shows and Events** *(12)*

Figure 7


Open Directory - Recreation: Travel: Guides and Directories
http://dmoz.org/Recreation/Travel/Guides_
Google

Top: **Recreation**: **Travel**: **Guides and Directories** *(124)* **Description**

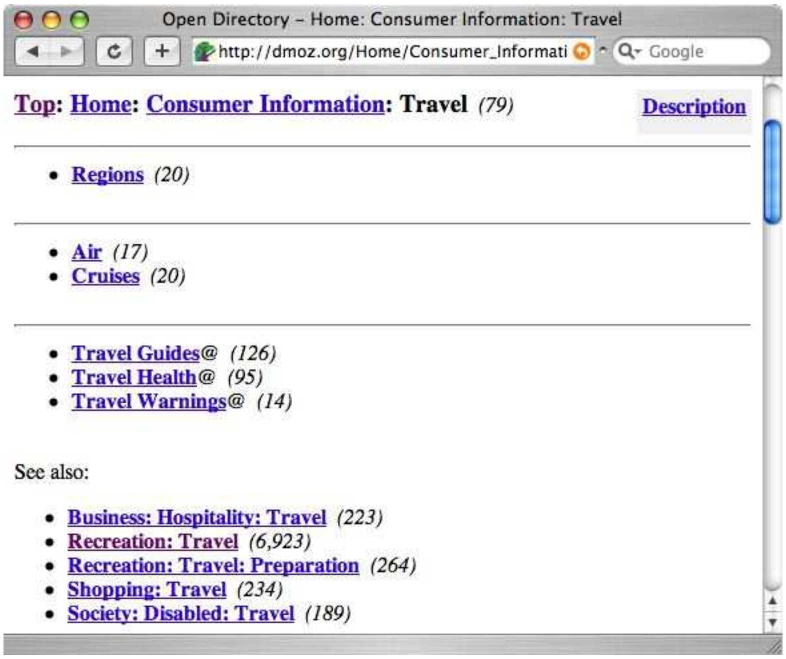
- **Africa**@ *(14)*
- **Asia**@ *(6)*
- **Caribbean**@ *(18)*
- **Central America**@ *(8)*
- **Europe**@ *(20)*
- **Middle East**@ *(3)*
- **North America**@ *(5)*
- **Oceania**@ *(3)*
- **South America**@ *(16)*

- **Adventure and Sports**@ *(12)*
- **Air Travel**@ *(3)*
- **Beaches and Islands** *(22)*
- **Corporate Travel**@ *(3)*
- **Entertainment** *(5)*
- **Lodging**@ *(21)*
- **Spas**@ *(10)*
- **Wine**@ *(19)*

See also:

- **Home: Consumer Information: Travel** *(79)*
- **Recreation: Travel: Consolidators** *(27)*
- **Recreation: Travel: Travelogues** *(1,553)*

↓


Open Directory - Home: Consumer Information: Travel
http://dmoz.org/Home/Consumer_Informati
Google

Top: **Home**: **Consumer Information**: **Travel** *(79)* **Description**

- **Regions** *(20)*

- **Air** *(17)*
- **Cruises** *(20)*

- **Travel Guides**@ *(126)*
- **Travel Health**@ *(95)*
- **Travel Warnings**@ *(14)*

See also:

- **Business: Hospitality: Travel** *(223)*
- **Recreation: Travel** *(6,923)*
- **Recreation: Travel: Preparation** *(264)*
- **Shopping: Travel** *(234)*
- **Society: Disabled: Travel** *(189)*

Figure 8

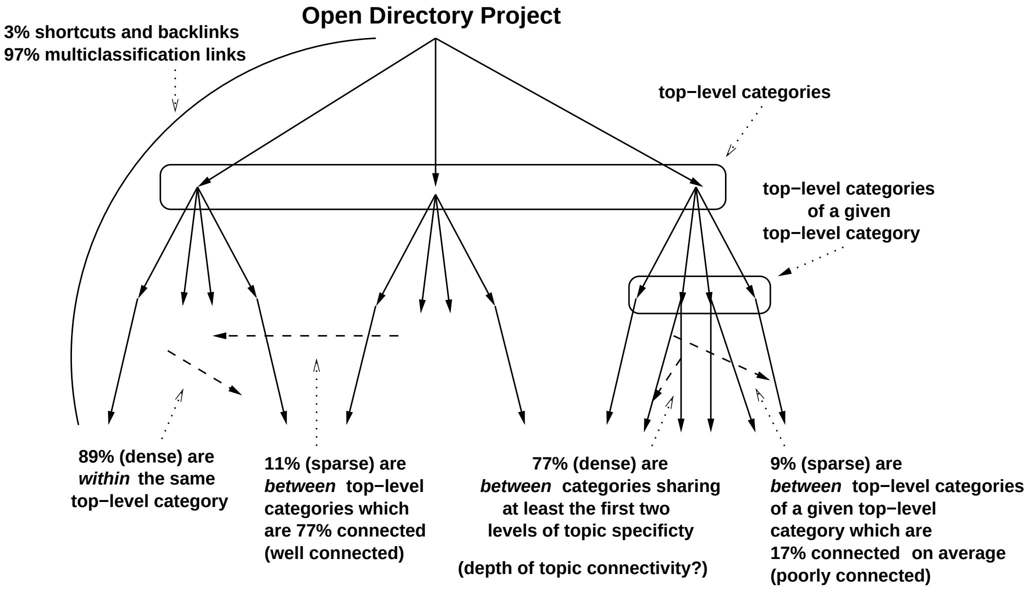

Open Directory Project
3% shortcuts and backlinks
97% multiclassification links
top−level categories
top−level categories of a given top−level category
89% (dense) are *within* the same top−level category
11% (sparse) are *between* top−level categories which are 77% connected (well connected)
77% (dense) are *between* categories sharing at least the first two levels of topic specificty
(depth of topic connectivity?)
9% (sparse) are *between* top−level categories of a given top−level category which are 17% connected on average (poorly connected)

Table legends

Legend for Table 1

Legend: μ = mean; C/N children per node; σ = standard deviation.

Legend for Table 2

Legend: C/N = children per node; (w/o) = without symbolic links; (w/) = with symbolic links; $\mathrm{r}^{(\mathrm{w/:w/o})}$ = ratio of average number of children per node with symbolic links to that without.

Legend for Table 3

Legend: μ = mean; σ^2 = variance; σ = standard deviation.

Legend for Table 4

Legend: b/t = between; w/i = within; $\mathrm{R}^{(\mathrm{w/i:b/t})}$ = ratio of the number of symbolic links within a category to the number b/t categories.

Legend for Table 6

Legend: $\mathrm{R}^{(\mathrm{s:t})}$ = ratio of total (non-leaf + leaf) symbolic link sources to targets.

Table captions

Caption for Table 1

Structural characteristics of ODP. **Legend**: μ = mean; C/N children per node; σ = standard deviation.

Caption for Table 2

Effect of symbolic links on the average number of children per node. **Legend**: C/N = children per node; (w/o) = without symbolic links; (w/) = with symbolic links; $\mathrm{r}^{(\mathrm{w/:w/o})}$ = ratio of average number of children per node with symbolic links to that without.

Caption for Table 3

Distribution of shortcuts and backlinks in ODP and within its top-level categories. Numbers typeset in bold font are discussed in the text. **Legend**: μ = mean; σ^2 = variance; σ = standard deviation.

Caption for Table 4

Distribution of multiclassification links within and between the top-level categories of ODP. **Legend**: b/t = between; w/i = within; $\mathrm{R}^{(\mathrm{w/i:b/t})}$ = ratio of the number of symbolic links within the specified top-level category to the number between top-level categories.

Caption for Table 5

Local degree of connectivity created by multiclassification links.

Caption for Table 6

Distribution of non-leaf and leaf pages as sources and targets of symbolic links. **Legend**: $\mathrm{R}^{(\mathrm{s:t})}$ = ratio of total (non-leaf + leaf) symbolic link sources to targets.

Figure captions

Caption for Figure 1

Sample web directory, simplified for purposes of presentation, with characteristics similar to those in Yahoo!. Nodes correspond to webpages and directed edges correspond to hyperlinks between pages. Symbolic links are indicated by dashed edges and hyperlink labels ending with @.

Caption for Figure 2

A conceptual taxonomy of hyperlinks designed to situate symbolic links in relation to various other types of structural links.

Caption for Figure 3

A shortcut, labeled ‘Dance@,’ in ODP of length 2 from topic ‘Arts’ to topic ‘Arts: Performing Arts: Dance.’

Caption for Figure 4

Illustration of a backlink, labeled ‘Wellington Region@,’ of length 2 in ODP from topic ‘Regional: Oceania: New Zealand: Wellington: Localities: Otaki’ to topic ‘Regional: Oceania: New Zealand: Wellington.’

Caption for Figure 5

Illustration of a symbolic link in ODP, labeled ‘Music@,’ creating multiclassification between two top-level categories of the ‘Arts’ category, namely, ‘Performing Arts’ to ‘Music.’

Caption for Figure 6

Illustration of a symbolic link, labeled ‘Antiques@,’ creating multiclassification between top-level categories from ‘Arts’ to ‘Recreation: Antiques’ in ODP.

Caption for Figure 7

Illustration of a 'see also:' cross-reference, labeled 'Home: Consumer Information: Travel,' in the 'Recreation: Travel: Guides and Directories' category of ODP. This cross-reference indicates to the user that the current category, 'Recreation: Travel: Guides and Directories' (top), is related to the 'Home: Consumer Information: Travel' category (bottom).

Caption for Figure 8

Graphical summary of the results. Solid lines represent hard links/paths, dashed lines represent multiclassification links, and dotted lines are for annotation purposes.