The XML Web: a First Study

Laurent Mignet
Univ. of Toronto
mignet@cs.toronto.edu

Denilson Barbosa
Univ. of Toronto
dmb@cs.toronto.edu

Pierangelo Veltri
Univ. of Catanzaro
veltri@unicz.it

ABSTRACT

Although originally designed for large-scale electronic publishing, XML plays an increasingly important role in the exchange of data on the Web. In fact, it is expected that XML will become the lingua franca of the Web, eventually replacing HTML. Not surprisingly, there has been a great deal of interest on XML both in industry and in academia. Despite this interest, to date no comprehensive study on the *XML Web* (i.e., the subset of the Web made of XML documents only) nor on its contents has been made. In this paper we describe the XML Web and study the documents contained in it. Our results are drawn from a sample of a repository of the publicly available XML documents on the Web, consisting of about 200,000 documents. Our results show that, despite its short history, XML already permeates the Web, both in terms of generic domains and geographically. Also, our results about the contents of the XML Web provide valuable input for the design of algorithms, tools and systems that use XML in one form or another.

1. INTRODUCTION

The advent of the Web enabled the sharing of information at an unprecedented scale. Unquestionably, the enormous success and popularity of the Web is due in great part to the introduction of HTML as the standard format for representing content. HTML also enables the exchange of information among Web agents (humans or computer applications) via forms. However, HTML forms are adequate for simple transactions (e.g., online shopping) only, and do not scale to large-scale data exchange among several agents (e.g., B2B transactions). XML [28] is a simple and extremely flexible text format derived from SGML [22] and originally designed as a tool for electronic publishing. Unlike HTML, XML decouples content from presentation instructions, which allows the same XML document to be presented in different ways by using different rendering instructions (e.g., XSL stylesheets [30]), and is a great advantage from a content authoring point of view. From a data exchange point of view, XML allows the definition of domain-specific markup, which lends XML as a natural format for representing and exchanging semistructured data on the Web [1]. In fact, there are already several Web-based data exchange protocols (e.g., SOAP [29]) that use XML as a middleware format between agents.

The reasons for studying the Web extend beyond intellectual curiosity. The Web is a technological phenomenon with several social and economical consequences, and therefore must be studied. There is interest from both academia and industry in understanding the macro properties of the Web (e.g., its shape, size and connectivity). Although such understanding of the Web as a whole exists (see [9] and references therein), to the best of our knowledge, no similar study has been done for the XML subset of the Web, hereafter called the *XML Web*. Unlike HTML, which has a fixed set of markup, the XML Web is stratified into classes of documents defined by conceptual schemata, usually in the form of Document Type Definitions (DTDs) or XML Schema specifications. Characterizing these classes of documents potentially gives a much more accurate picture of the actual content available on the XML Web.

XML has also received considerable attention from the database community, which typically views XML documents as (semistructured) data. There has been an astonishing amount of work in this community aiming at coping with XML content, primarily motivated by data exchange and integration [10]. There has been work on storing XML content, both by developing new technologies (e.g., [17]) and by leveraging mature ones (e.g., [4, 3]); indexing and querying XML content (e.g., [16, 15]); updating XML data (e.g., [25, 19]); and benchmarking XML applications (e.g., [21, 33, 34]). Furthermore, the database industry has also adopted XML aggressively (e.g., [11, 18, 24]). However, the lack of accurate knowledge about the XML documents on the Web, the development of of such tools has been widely guided by folklore (e.g., XML documents are "shallow"), by the few well known publicly available XML documents (e.g., [23, 35]), or by proprietary XML content. In any case, given the amount of algorithms, tools and systems being developed, it becomes crucial to understand the document level or micro properties of the XML Web (e.g., the size and the complexity of typical documents). Furthermore, accurate knowledge about the XML Web is particularly useful for the development of meaningful XML benchmarks.

In this paper we report the results of an analysis of about 200,000 XML documents publicly available on the Web. These documents come from a sample of the Xyleme [37, 36] repository. We gathered several data about these documents, such as their size; number of elements and attributes; their `URL`; other documents they point to; whether they reference a schema; etc. Our results are divided into two categories. First, we study the XML Web by showing how it is distributed across Internet domains and geographically, and by identifying the most common kinds of content in it. Next, we study the contents of the XML Web (i.e., the documents found in it, irrespective of their origin). The results in this paper can be summarized as follows.

Statistics about the XML Web. Our results show that, despite its infancy, XML already permeates the Web: XML documents can be found in all major Internet domains and in all geographic regions of the globe. The ".com" and ".net" domains combined contain 53% of the documents and 76% of the volume of XML content on the Web. Surprisingly, only 48% of the documents reference a DTD, and 0.09% of the documents make reference to an XML Schema specification. WAP and RDF make up 26% and 17% of all documents on the XML Web, respectively. Finally, similarly to HTML documents, the out-degree of the XML documents seems to follow a power law.

Statistics about the XML documents. Our results reveal that typical XML documents on the Web are small: the average document size in our sample is around 4KB. We also found that the volume of markup (i.e., element tags and attributes) is surprisingly high when compared to the actual content of the documents. Confirming the folklore, our results show that XML documents are in fact relatively shallow: 99% of them have less than 8 levels of element nesting. Also, 15% of the documents we analyzed have recursive content, and one can see much regularity in it.

1.1 The Sample of the XML Web

In this section we describe the sample of the XML Web used for obtaining the results we present in this paper. Our sample consists of 190,417 XML documents, randomly chosen from the public repository of Xyleme. This repository is populated by crawling the Web, and, thus contains only documents that are publicly available. Xyleme also has a private repository, which is populated by subscription only. At the time our sample was collected (February 2002), Xyleme's public and private repositories contained approximately 500,000 and 700,000 documents, respectively. The documents in our sample combined represent approximately 843MB of XML content, and come from 19,254 different Web sites.

We note that our sample represents only a snapshot of the publicly available XML documents known to Xyleme, at the time its crawlers fetched these pages, and, unfortunately, there is not much we can say about its representativeness. Given the lack of reliable estimates on the size of the XML Web, and the intrinsic difficulty in obtaining such estimates [2], we do not presume our study is definitive. Nevertheless, our results give an accurate and valuable starting point for understanding the XML Web.

The actual results we present here were obtained by querying a relational database containing several meta-data about the documents in our sample. These data were split into 12 relations, and, altogether correspond to roughly 2.5GB of data.

1.2 Related Work

There are several organizations that periodically release statistics about the size and the shape of the Internet (e.g., [6, 12]). The data collected by these organizations comes primarily from accessing network addresses found in Domain Name Service (DNS) servers, and thus are very accurate. Those reports, however, count the number of computers that belong to a given Internet domain, regardless of how much (if any) Web content (XML or otherwise) is published by them. Our work, on the other hand, focuses on the XML Web only. Also, we give the distributions of the number of sites, number of documents, and volume of published content according to Internet domains.

The connectivity and structure of the Web as a whole has also been studied extensively (see [9] and references therein). Those results are primarily aimed at studying Web algorithmics; moreover, results of that nature have been shown to improve the accuracy of search engines [5]. Our work differs from those in the following ways. First, they do not distinguish the XML documents on the Web; therefore, it is not clear whether their results apply to the XML Web at all. Second, by focusing on the XML Web only, our results allow an accurate quantitative analysis of its properties. Third, we consider other aspects of the XML Web, that are not relevant in the context of those works (e.g., the use of conceptual schemata). Finally, we characterize the XML Web both in terms of Internet domains and geographically.

Choi [8] has recently analyzed 60 DTDs found on the Web. Although we were able to find references to 75 different DTDs in our sample, our goals in this work differ from those of [8], in the sense that we are interested on the usage of DTDs (and XML Schema specifications) on the Web, rather than on the quality of these schemata. Furthermore, we present several quantitative results that cannot be derived from analyzing conceptual schemata alone, for obvious reasons.

Outline of the paper. The rest of the paper is organized as follows. We present the statistics about the XML Web in Section 2 and the results about the XML documents found on the Web are presented and discussed in Section 3. Finally, we discuss our results and present directions for future work in Section 4.

2. STATISTICS ABOUT THE XML WEB

The results in this section describe the XML Web and the kinds of documents that are found in it. These results are divided into three groups. First, we show how the sites and the contents of the XML Web are distributed in terms of Internet domains and geographical reasons. Next, we show how the actual content on the XML Web is distributed semantically. Finally, we study the connectivity of the documents in the XML Web.

2.1 Site Distribution

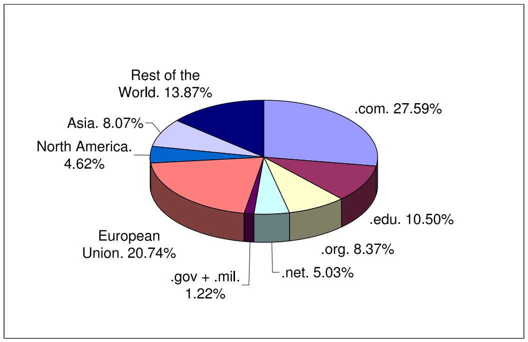


Figure 1: Distribution of XML sites.

We cluster the 19,254 Web sites in our sample by zones, consisting of generic Internet domains (i.e., *.com*, *.edu*, *.net*, *.org*, *.gov* and *.mil*) and geographical regions, defined as follows. The *Asia* zone consists of China, India, Indonesia, Japan, Pakistan, Taiwan, South Korea and Singapore; the *European Union* zone consists of the fifteen European Union countries; the *North America* zone consists of Canada and the United States; finally the *Rest of the World* zone represents sites from any other country not cited before. The distribution of sites according to these zones is given in Figure 1.

Two zones dominate the distribution of sites: the *.com* with 5,312 sites, and the *European Union*, with 3,993 sites. Following those, we have *.edu* with 2,022 sites, *.org* with 1,611 sites, *Asia* with 1,553 sites, *.net* with 968 sites and *North America* with 890 sites. The *Rest of the World* zone is mainly composed by the Russian Federation (314 sites), Switzerland (260), Czech Republic (251) and Norway(199).

In pure geographical terms, the distribution shows that North America contains at least 16% of all sites (corresponding to the zones: *North America*, *.edu*, *.gov* and *.mil*). We cannot distinguish the geographical origin of the other generic domains, and thus give a more accurate geographical distribution of the XML Web. However, we note that at least one country from each other continent is represented in our sample: Brazil (56 sites), Cuba (1), Iran (3), South Africa (83), and the Niue Island, Polynesia, with 39 sites.

2.2 Document Distribution

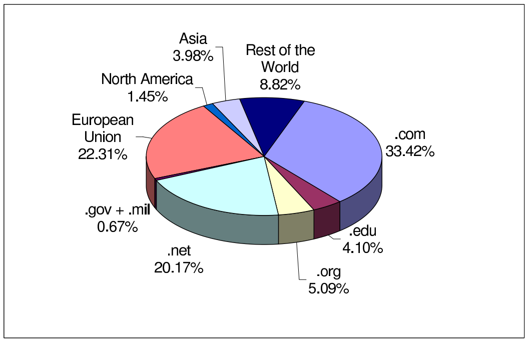


Figure 2: Distribution of XML documents.

We now discuss how the contents of the XML Web are distributed according to the zones discussed in the previous section. First, we consider the distribution of the documents (i.e., the number of documents per site).

As already mentioned, there are 190,417 XML documents and 19,254 sites in our sample. This gives an average of 9.89 documents per site. The sites with the largest number of documents are: `rpmfind.net`, with 12,340 documents (6.5% of the total); `download.sourceforge.net`, with 7,948 documents (4%); and `ludiwap.co.uk`, with 7,029 documents (3.7%). The first two sites have an interesting impact on how the contents of the XML Web are distributed, as shown in Figure 2. One can notice a significant increase in the participation of the *.net* zone, jumping from 5% of sites to 20% of documents in the XML Web. The graph also shows that, similarly to the distribution of sites, the *.com* and *European Union* zones dominate the distribution.

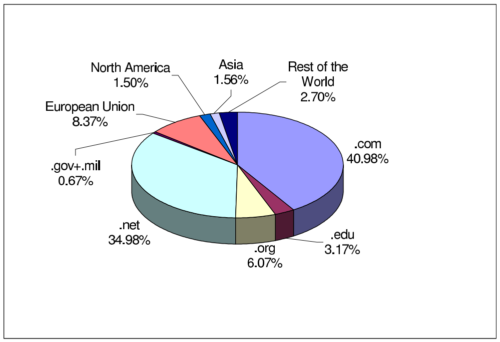


Figure 3: Distribution of volume of XML content.

We now study the distribution of the volume of XML content (i.e., the sum of the sizes of the documents) per zones (see Figure 3). This graph shows the *.com* zone as the dominant zone, but it also shows another increase in the participation of the *.net* zone, now moving to second place with 35%. These two zones alone account for 53% of all documents and 76% of the volume of content on the XML Web.

2.3 Schema Usage

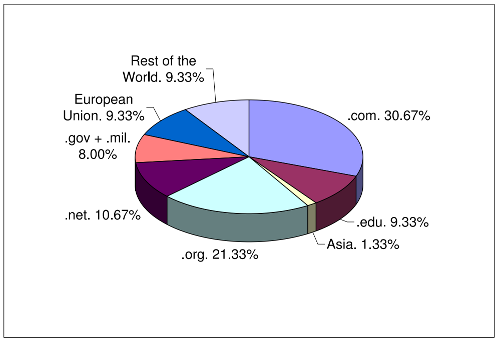


Figure 4: Distribution of DTDs by Web domains.

As mentioned earlier, one distinguishing feature of the XML Web is that it is stratified by classes of documents, defined by conceptual schemata. We now consider the use of the two standard schema languages defined for XML: DTDs and XML Schema. It turns out that 48% of the XML documents in our sample contain a link to a DTD. Surprisingly, only 75 different DTDs are referenced in our sample. These DTDs come mostly from the *.com*, *.org*, and *.net* zones, as shown in Figure 4. Also to our surprise, 92% of all DTD references are made to norms 1.1 or 1.2 of the Wap protocol [27].

The use of XML Schema [32], the new mechanism to specify the schema for XML documents, is insignificant. Indeed, in our sample only 0.09% of the documents (179 documents) use either the attribute label “SchemaLocation” or “noNameSpaceSchemaLocation”.

2.4 File Extension Distribution

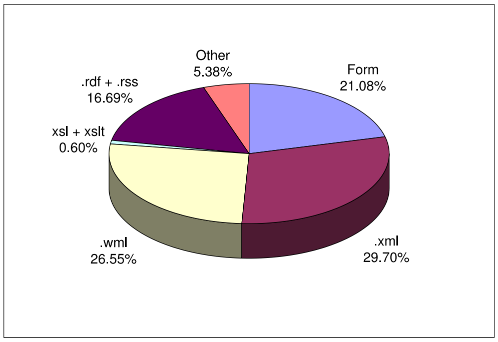


Figure 5: Distribution of files by extension.

Another way of classifying the content of the XML Web is by looking at the extension of the associated files or the method by which they are accessed. We distinguish the following major groups of content in this work: documents from the semantic Web (file extensions “.rdf” and “.rss”); Wireless Application Protocol [27] (WAP) documents (file extension “.wml”); XSL and XSLT documents; form-accessible documents, and indistinguishable “.xml” documents.

The distribution of the documents in our sample according to the groups described above is given by Figure 5. The graph shows that

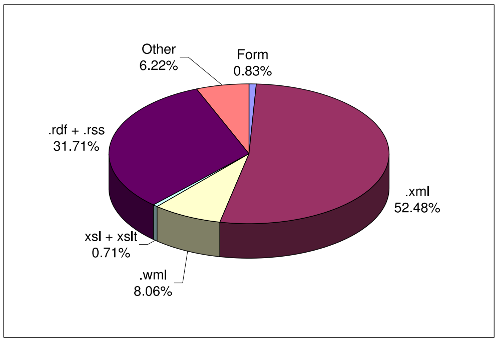


Figure 6: Distribution of volume by extension.

most documents belong to the ".xml", WAP, and form-accessible classes. Documents from the semantic Web community also make up a large fraction of the distribution. We also give the distribution of the volume of content according to these categories is (see Figure 6).

Several observations can be made by comparing the two previous figures. First, we note that although WAP documents account for an expressive number of documents, the combined volume of content in this class is not as significant. This can be explained by the fact that WAP documents are usually viewed in mobile devices, for which memory and communication requirements are very strict. Second, we note a considerable increase in the participation of the semantic Web class. Finally, we observe an almost insignificant volume of XML content obtained from accessing forms (i.e., the "hidden XML Web"). Since we do not know the actual size of the hidden XML Web, we can only speculate that the Xyleme crawler is not designed to retrieve its documents. We note that even estimating the size of the hidden Web (XML or otherwise) is not a trivial task and usually requires special-purpose tools [20] and some human guidance [13].

2.5 Document Out-degree Distribution

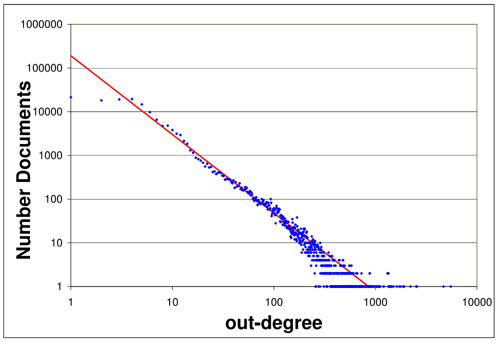


Figure 7: Distribution of documents by their out-degree. The distribution follows a power of law of exponent 1.8.

We conclude the section with an analysis of the connectivity of the XML Web graph. We define the *out-degree* of an XML document as the number of attribute nodes labeled `href`, `xmlhref`, or `xlink:href` in that document. Figure 7 is a log-log plot of the out-degree distribution for the documents in our sample. Similarly to previous observations on the HTML Web [14], we observe that the out-degrees seem to follow a power law: the fraction of XML documents with out-degree i seems to be proportional to $1/i^x$ for $x = 1.8$. This value is derived from the slope of the line providing the best fit to the data. The average out-degree of the documents in our sample is about 11.4, while the out-degree for HTML pages is about 7.2 [14]. However, given the (expected) small size of our sample compared to the (unknown to us) size of the XML Web, we cannot generalize this result.

3. STATISTICS ABOUT THE XML DOCUMENTS

This section studies some structural properties that we gathered from the XML documents. First, we cluster the documents by size and compare the distribution of nodes according to this clustering. Next, we give an overview of the depth of the documents and the distribution of nodes per level. Then, we study the fan-out of the element nodes in terms of element and attribute nodes, for the first three levels. Finally, we characterize the recursive elements found in our sample.

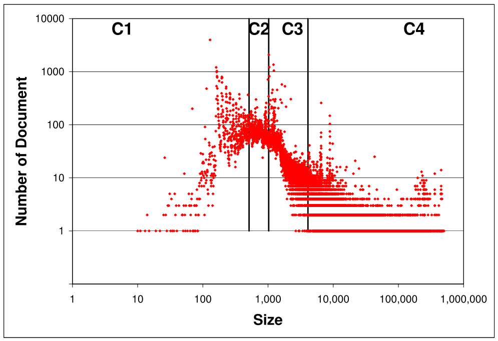


Figure 8: Document distribution by size.

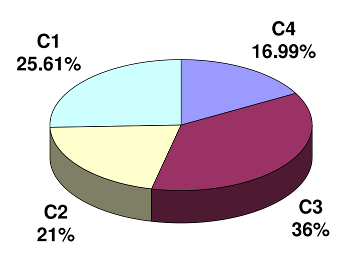
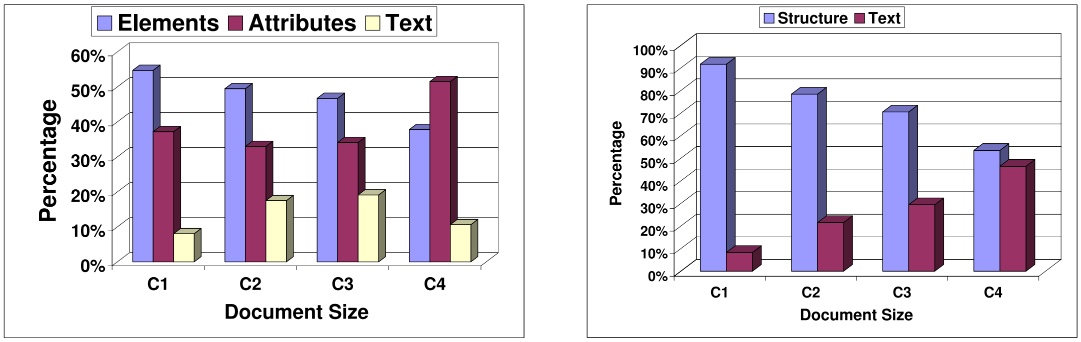


Figure 9: Document clusters by size.

3.1 Document Clusters by Size

The sizes of the XML documents vary from 10 to 500,608 bytes, for an average of 4,641 bytes. In Figure 8, we show how the documents are distributed according to their sizes, in a log-log scale. The vertical lines in the figure represent, from left to right, 512, 1024 and 4096 bytes. We use these values as boundaries for clustering the documents by size. These values are common candidates for disk page sizes in secondary-memory storage systems, and, thus, natural candidates for our clustering purposes. We name the clusters from C_1 the cluster containing documents with a size



(a) Distribution of element, attribute and text nodes.

(b) Size distribution.

Figure 10: Structural vs content distribution.

lesser than 512 bytes to C_4 for documents with a size greater than 4096 bytes. Figure 9 gives the distribution of documents per cluster.

A closer look at the document clusters shows that their content is distributed as follows:

- C_1 (48671 documents in total): 12% ".wml" (5,871 documents), 60% ".xml" (29,556 documents), 1% ".rdf + .rss", and 37% for other types;
- C_2 (39449 documents in total): 62% ".wml" (24,500 documents), 20% ".xml" (8,035 documents), 1% ".rdf + .rss" (681 documents), and 17% for other types;
- C_3 (69846 documents in total): 30% ".wml" (21,403 documents), 36% ".xml" (25,115 documents), 16% ".rdf + .rss" (11,765 documents), and 18% for other types;
- C_4 (32361 documents in total): 1% ".wml" (356 documents), 37% ".xml" (12,156 documents), 58% ".rdf + .rss" (18,733 documents), and 4% for other types.

The clustering above reveals that most wml documents are relatively small (88% of the total belong to either C_1 or C_2), while rdf-rss documents are larger (96% of all such documents belong to either C_3 or C_4). There is no such apparent classification for other kinds of documents.

3.2 Node Distribution

This section compares the amount of markup, which we call structural content and consists of element and attribute nodes, versus the amount of textual content (i.e., PCDATA nodes), based on the clustering defined in the previous section. For these results, we do not keep track of "empty" text nodes (by empty, we mean text nodes with no characters except the different blank characters as defined in [28]). First, we compare the distribution of nodes of each type (Figure 10(a)). Several observations can be made from this figure:

- For documents of up to 4096 bytes, the number of element nodes dominates the distributions (54.74%, 49.51%, 46.74%), although this dominance declines gradually as the proportion of text nodes increases (8.05%, 17.50%, 19.13%).
- For documents of up to 4096 bytes, the proportion of attribute nodes seems constant (37.21%, 32.99%, 34.13%).
- For documents larger than 4096 bytes, there are proportionally more attribute nodes than element nodes (51.13% vs. 37.83%), and the proportion of text nodes seems to decline (10.64%). The inversion of proportions between attribute and element nodes has a strange consequence on the number of nodes contained in our sample: on a total of 36,498,256 nodes 14,514,673 are element nodes; 4,381,442 text nodes; and 17,602,141 attribute nodes, which yields 3,087,468 more attribute than element nodes!

We also compare the size (in bytes) of the structural content versus the size of the textual content, as shown in Figure 10(b). For the text size we count the number of characters contained in each non-empty text node. The size of the structural part of the document is simply the size of the serialized form of the document minus the total size of the textual information in the document. Note that the tags of empty text nodes are counted as structural information. As we can see, in all clusters, the structural information dominates the size of the documents.

The main conclusion one can draw from both previous graphs is that the structural information found in XML documents is in fact dominant over the textual content. This comes as no surprise for small documents, since XML (fortunately) requires explicit closing tags for all elements in the document. However, although our results show that the content/markup ratio increases with the size of the documents, the dominance of the markup over the content and the high number of attribute node show that the notions of data and meta-data are somewhat blurred in the XML web.

3.3 Depth

This section studies the depth (i.e. the longest path from the root node to a leaf node inside the tree view of an XML document) of the documents. The distribution of documents according to their depth is given in Figure 11. As one can see, most documents are relatively shallow: 99% of the documents have less than 8 levels. The average depth is 4, and the deepest document has 135 levels. There are 1,986 documents whose depth is zero: 1,671 documents which consist of a single empty element node, and 377 other documents that have a single element with some textual content.

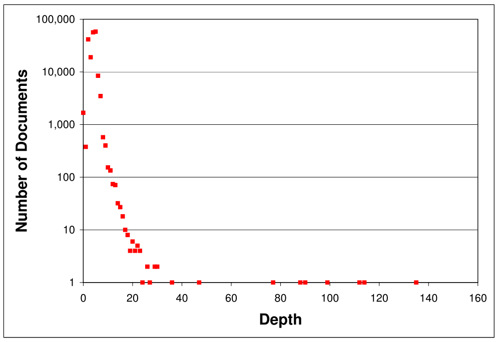


Figure 11: Distribution of documents by depth.

Figure 12 gives the distribution of the different node types per level in the XML tree, where level 0 is assigned to the root node. For instance, Figure 12(a) shows that the second level in the XML tree contains more attributes than any other level. In fact, 89% of all attributes are found in the first 3 levels of the documents. A similar pattern is also observed for element nodes and text nodes: 77% of all element nodes and 61% of all text nodes are found in the first 3 levels of the documents (see Figures 12(b) and 12(c), respectively).

These distributions can be extended to study the fan-out of the element nodes in term of attributes and child elements. The two next section give these informations.

3.4 Element Fan-Out

In this section we study the *element fan-out* (i.e., the number of children per element) of the element nodes. Our goal is to correlate the number of nodes for the first three levels in Figure 12(b) to study the structure of the subtrees rooted by these nodes. Intuitively, one can expect large element fan-out for "collection" documents containing several similar items. For instance, in a document like DBLP [35], one would expect a large fan-out for elements representing conferences. Small fan-out, on the other hand, intuitively indicates the document represents a single object (say, a single conference paper).

Figure 13(a) is a log-log plot of the element fan-out of the root element; i.e., the distribution of child nodes of the root elements. As already mentioned, 1,986 documents consist of a single root node with no children. Also, 53,401 documents have exactly 2 nodes (a root node with a single child). The distribution of the element fan-out seems to follow a power law (of degree 1.85). The same observation (with a degree of 3.1) can be made for the distribution of the element fan-out of element nodes at the second level, as shown in Figure 13(c).

The distribution for the element fan-out of element nodes at the first level (Figure 13(b)) is not as easy to characterize, however. Although one can notice that part of the distribution seems to follow a power law (of degree 2.8), there is also a considerable number of elements nodes that have element fan-out of around 10,000. These elements are found in 518 documents that come from only two sites: 514 from the `ibm.com/developerworks` site and 4 from the `w3.org/TR` site. The label distribution for the children of the elements with large fan-out is as follows: in 135 documents, a single label is found; in 301 documents, two distinct labels are found; and in the remaining 82 documents, exactly three labels are found. Surprisingly enough, all 518 documents are character encoding maps for various different languages.

Another observation we make is that the average values for the element fan-out at levels zero, one and two are, respectively, 8.57, 5.76 and 0.18. This not only reinforces the previous observation that XML documents are shallow but also suggests that "tall" documents (i.e., documents with large depth) are not wide.

3.5 Attribute Fan-Out

A similar analysis with respect to the number of attribute nodes per element (i.e., the *attribute fan-out*) for the first three levels of the documents is shown in Figures 14(a),14(b), and 14(c), respectively. The shape of these distributions seem to follow also power laws with different slopes. We draw the attention to the element's fan-out in term of attribute nodes in the second level, which corresponds to the level where most attributes are found, as discussed earlier (see Figure 12(a)).

Other statistics are: in total, 2,588,286 element nodes (18% of the total) have no attributes, and, thus, were not counted in this analysis. The average number of attributes per element in each level are 0.18, 1.06 and 1.4 for levels zero, one and two, respectively.

3.6 Recursion

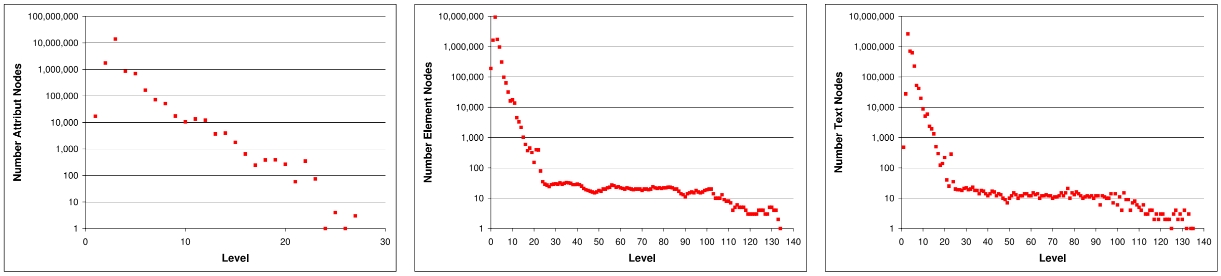
Our final study is an analysis of the 28,208 XML documents (14.81% of the documents) that contain recursive elements. The reasons for studying recursion in the XML Web are simple. First, recursion is naturally captured by XML documents and schema specifications such as DTDs. However, recursion can have a considerable impact on the performance of query processors and storage mechanisms for XML. This study is complementary to the work of [8], which characterizes recursive DTDs found on the Web.

For our purposes here, we say an element e is recursive if there exists at least one element d in the same document such that d is a descendant of e and d has the same label as e. For simplicity, we call an element-descendant association an ed pair. A *recursive XML tree* is an XML tree that is rooted at a recursive element and whose leaves are recursive descendants of the root (e.g., the tree in Figure 15(a)). For reasons that will become apparent shortly, we use two different interpretations on what to count as ed pairs. Consider the XML tree in Figure 15(a). For the *All-Descendants* interpretation (AD), shown in Figure 15(b), elements `3`, `4` and `5` are the recursive descendants of element `1`. In the *Closest-Descendants* interpretation (CD) in Figure 15(c), only elements `3` and `4` are considered to be recursive descendants of element `1`. In both interpretations, element `5` is a recursive descendant of element `4`.

We now present some statistics about the ed pairs found in our sample. In total, there are 66,139 recursive XML trees (i.e., elements that contain at least one recursive descendant); there are 213,507 ed pairs in the AD interpretation, and 147,557 ed pairs in the CD interpretation. Among all recursive elements, only 260 different labels are found. In 27,577 of the documents with recursive content (98% of the total), a single label is used for all recursive elements, and in 307 documents (1% of the total), 2 labels are found among all recursive elements. The maximum number of labels used for recursive elements in a single document is 9. The most popular labels for the recursive elements in all documents are:

- `ae`, which labels 68,930 elements (32.28%) and is found in 77 documents (0.27% of all documents with recursive content);
- `description`, which labels 30,509 elements (14.28% of the total) and is found in 25,368 documents (89.93%);
- and `page`, which labels 30,429 elements (14.25%) and is found in 19 documents (0.06%).

Among all documents with recursive content, 26,920 (95% of the total) do not reference a DTD. We also observe that the Wap

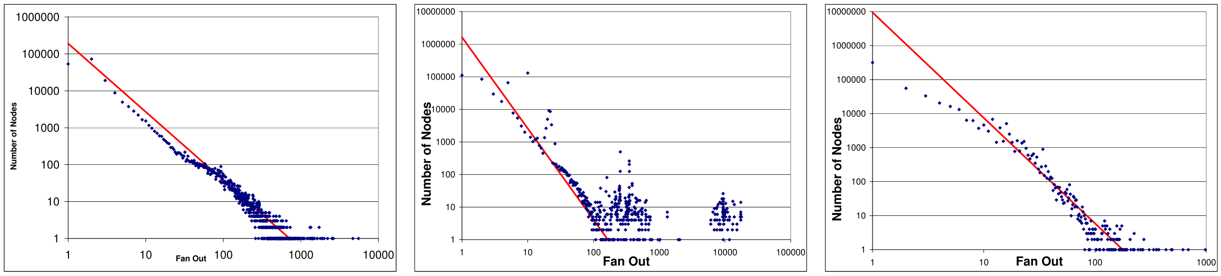


(a) Attribute nodes.

(b) Element nodes.

(c) Text nodes.

Figure 12: Distribution of nodes per level.

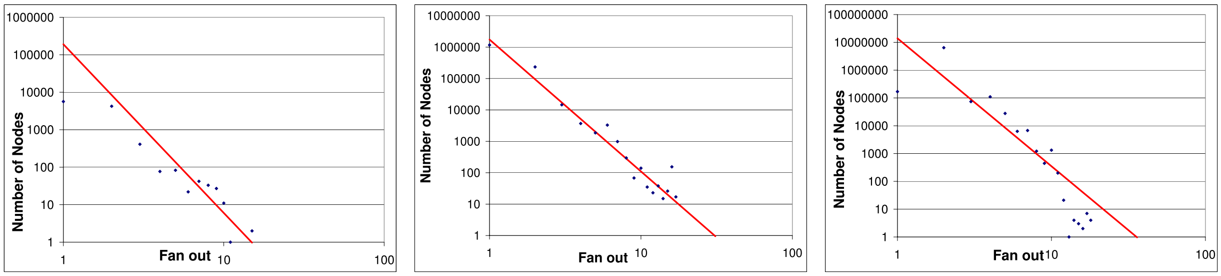


(a) Number of children per element for the level 0. The distribution follows a power of law of degree 1.85.

(b) Number of children per element for the level 1. The distribution follows a power of law of degree 2.8.

(c) Number of children per element for the level 2. The distribution follows a power of law of degree 3.1.

Figure 13: Element fan-out for the first three levels.

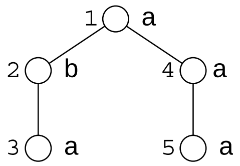


(a) Number of attributes per element for the level 0. The distribution follows a power of law of degree 4.5.

(b) Number of attributes per element for the level 1. The distribution follows a power of law of degree 4.2.

(c) Number of attributes per element for the level 2. The distribution follows a power of law of degree 4.6.

Figure 14: Attribute fan-out for the first three levels.



(a) An XML tree containing recursive elements.

e	d
1	3
1	4
1	5
4	5

(b) ed pairing in the all-descendants interpretation.

e	d
1	3
1	4
4	5

(c) ed pairing in the closest-descendants interpretation.

Figure 15: The all-descendants and closest-descendants interpretations for recursive elements. The numbers in the left of the XML nodes in the tree are the node identifiers; the letters at the right are the node labels. The tables in Figures (b) and (c) show the pairing of elements (e column) and their recursive descendants (d column).

protocol DTDs are the most popular among recursive documents as well: 876 documents (3%) reference the DTD for the Wap 1.2 protocol while 338 documents (1%) reference the DTD for the Wap 1.1 protocol. Also, 25,226 documents (89%) with recursive content have either the "rdf" or "rss" suffixes. Cross-referencing the tag and document frequencies we find that most of the recursive documents come from the semantic Web community [26]. Finally 89% of these documents come massively from the *.net* Web domain with 36% from `rpmfind.net` and 22% from `download.sourceforge.net`. It appears that a huge amount of these document describe the content of rpm file (file extension used in some Linux community to deploy softwares).

Distance. Our first study of the recursive XML content concerns the distance in the XML tree between elements and their recursive descendants. We measure distance by counting the number of edges separating the two nodes in the XML tree. For instance, the distance between elements `1` and `3` in Figure 15(a) is equal to 2.

Figures 16(a) and 16(b) show the distribution of the ed pairs in each distribution according to the distance of the paired elements. The graphs have only one plot in common: the 115,622 ed pairs of elements whose distance is 1; a quick look at the AD plot shows that there are recursive XML trees of depth up to 119 levels, while the CD plot shows that there are recursive elements separated by a path of length 20 that does not contain other elements with the same label.

These results alone already justify the need for the AD and CD interpretations: the AD interpretation describe "global" properties of the recursive XML trees, while the CD interpretation describe "local" properties related to each recursive element and its children only. Evidently, some observations can be derived from either interpretation; for instance, both graphs above show that most elements in ed pairs have distance of 5 or less.

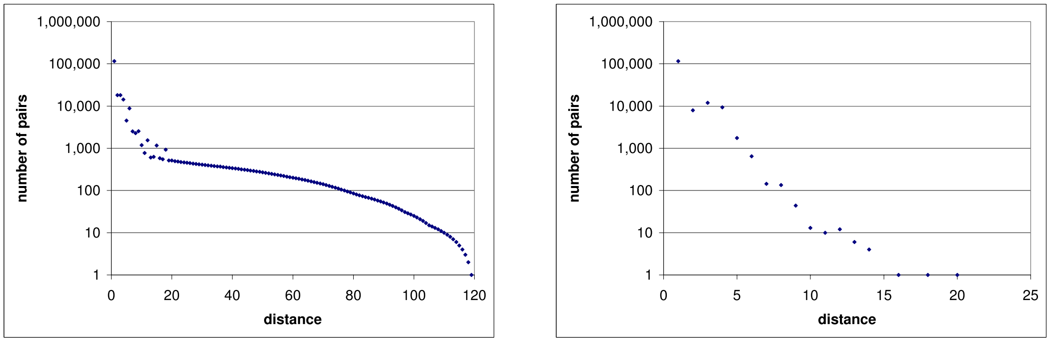
Regularity. A natural question about the recursive XML trees is whether there is any regularity in their shape. A simple notion of regularity can be the average distance between the elements in all ed pairs in the document. For this study, the CD interpretation provides a better reading. To see why, consider again the recursive tree in Figure 15(a); its average distance using the CD interpretation is $(1 + 1 + 2)/3 = 1,33$. Now, consider the subtree obtained by deleting element `2` from the tree in the figure; the average distance of the new tree is $(1 + 1)/2 = 1$. Intuitively, we can say that the recursion in the second tree is more regular, because the distance between the elements in all ed pairs is constant and, thus, equals the average.

The notion of regularity we describe above has the advantage of being extremely simple to compute. However, it is easy to see that it can be misleading if different labels are present in the same recursive tree: consider an XML tree containing 3 ed pairs with elements labeled a, all with distance of 1, and 2 ed pairs with elements labeled b whose distances are 2. The regularity of this tree is $(1+1+1+2+2)/5 = 1.4$ despite the fact that the distance of the elements in ed pairs for each given label is constant. Therefore, a better notion of regularity would take different labels into account. However, since 98% of the documents with recursive content have a single label for all of their ed pairs, our simple notion of regularity can be used without incurring in significant error.

Our results for regularity using the simple metric described above are shown in Figure 17(b). We draw the reader's attention to two observations: (1) most documents have average distance of 1, which is nothing more than a re-statement of the results presented above; (2) there is a lot of regularity in the recursion among the documents. The highest values in Figure 17(b) are: 26,388 documents (93% of the total) with average distance 1; 1,141 documents (4%) with average distance 2; and 124 documents (0.44%) with average distance 3. This shows that more than 97% of all documents with recursive content exhibit high regularity. The equivalent plot using the AD (All Descendant) interpretation is given by Figure 17(a); we note that this figure permits a similar reading, as expected.

Recursive fan-out. Another important parameter for studying the recursion in XML documents is what we call the *recursive fan-out* of an element, which is the number of recursive descendants of that element (or, the number of ed pairs in which the element appears in the e column). Again, the AD and CD interpretations provide complementary readings. The AD interpretation measures the total number of recursive elements in a given recursive XML tree. This is precisely the semantics of an Xpath [31] expression of the form $e//e$, where e is the label of a recursive element. The recursive fan-out under the CD interpretation, on the other hand, can be viewed as a "branching factor" of the recursive XML trees: intuitively, it measures how wide the XML tree gets as a function of the distance of the root of the tree. We note that a recursive fan-out of 1 means that the recursive tree gets only taller, but not wider, as the distance from the root increases.

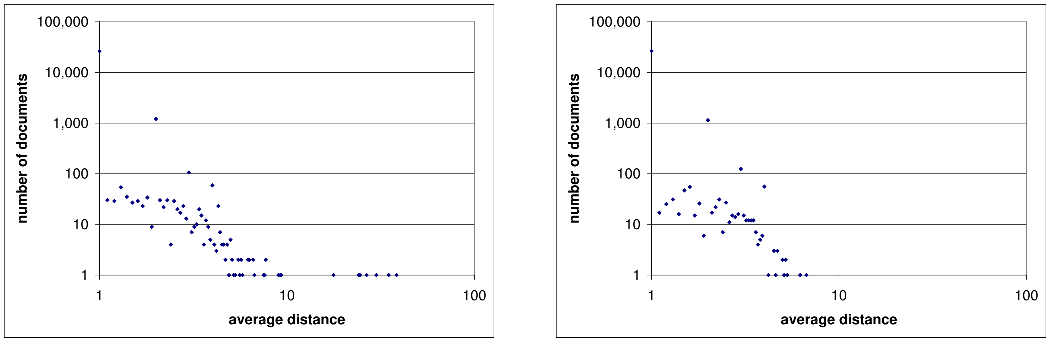
Figure 18(b) shows the distribution of the recursive elements w.r.t. their average recursive fan-out, using the CD interpretation. Figure 18(a) shows the equivalent plot using the AD interpretation. As one can see, both distributions seem to follow power laws. We note that a similar notion of regularity applies here, and, again using the CD interpretation, we observe that the most popular aver-



(a) AD interpretation.

(b) CD interpretation.

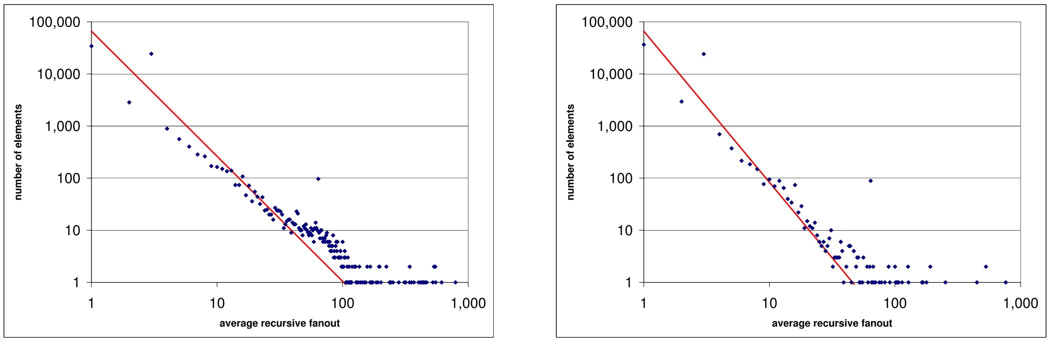
Figure 16: Distance between recursive elements and their descendants.



(a) AD interpretation.

(b) CD interpretation.

Figure 17: Distribution of documents according to the average distance between elements in all ed pairs in the document.



(a) AD interpretation. The distribution follows a power of law of degree 2.4.

(b) CD interpretation The distribution follows a power of law of degree 2.9.

Figure 18: Recursive fan-out of the elements.

age fan-outs are 1, found in 36,498 elements (60% of the total); 3, found in 24,177 elements (37%); and 2, found in 2,951 (4%). Also, the average recursive fan-out of all elements is 2.23; the largest recursive fan-out is 752, found in one element.

Several observations can be made from our results in this section. First, the fraction of documents containing recursive elements is not negligible (14.81%). Second, both the width and the height of recursive XML trees can grow relatively large, and vary considerably. The final, and perhaps, most important observation that we make is that there is a considerable amount of regularity in the recursion found in the XML documents of the Web.

4. CONCLUSION

In this paper we presented the results of a statistical analysis of a sample of the XML Web, consisting of about 200,000 XML documents. Our results can be classified into two broad categories: macro-level results, describing the XML Web and the kinds of contents in it, and document-level results, describing typical XML documents on the Web. Our results can be summarized as follows.

We showed that, despite its short history, XML is already pervasive: XML content can be found in all major internet domains and also in all continents of the globe. We also showed that 75% of all documents and 85% of the volume of XML content are provided by the *.com*, *.net* and by the different countries of the European Union. We gave the distribution of the contents of the XML Web based on several kinds of content, which revealed an expressive amount of content related to the semantic Web initiative. Next, we showed that the use of conceptual schemas on the XML Web is not yet widespread: only 48% of the documents reference DTDs while the number of documents that reference XML Schema specifications is insignificant (0.09% of all documents). These statistics can be viewed as empirical evidence for motivating the work on techniques for discovering semantic information from data (e.g., schema discovery, web mining and clustering, data integration, etc.). We also showed that, similarly to HTML documents, the outdegree of the XML documents seems to follow a power law.

The second part of our paper discussed structural properties of the XML documents on the Web. We showed that the average size of typical XML documents on the Web is around 4KB. We also found that the volume of markup is surprisingly high when compared to the actual content of the documents. On similar lines, we showed that the number of attributes exceed the number of element nodes by a large margin. This phenomenon suggests that content providers might use attributes to represent actual content, thus “blurring” of the concepts of data and meta-data. Also, our results confirmed the folklore that XML documents on the Web are shallow: 99% of them have less than 8 levels. We also showed that such documents can be very wide: their element fan-out can be as high as 10,000. Finally, we showed that 15% of the documents on the XML Web have recursive content, although one can identify much regularity in it. Our results provide valuable insight for developing algorithms, tools and systems that use XML in one form or another. In particular, our results have direct application in the development of meaningful benchmarks for XML applications.

We would like to mention that the full version of the paper presents several other results, such as the distribution of words in `#PCDATA` nodes; the distribution of element tag names and attribute names according to several criteria; and the use of namespaces in the XML Web. We also chose to present structural statistics for the whole data set even when some results are clearly biased towards some classes of documents (e.g., most of the documents with recursive content belong to the semantic web); or by the number of replicas in the sample (26,989 documents, nearly 20% of the total in our sample, were exact replicas of other documents). We note that previous results on the Web as a whole show that 36% of the documents in a large crawl were exact replicas of other documents [7]. We decided to present our results as we did in order to provide a whole vision of our sample. Providing such statistics for specific classes of documents will be done in a second step.

Future work. We identify several opportunities for extending this work. First, we plan to fetch new snapshots of the XML Web, in order to see how the it evolves over the time.

We also plan to investigate the percentage of semantic metadata that is effectively used by the XML documents on the Web. For instance, we want to check whether the documents that declare schemas do in fact conform with them, and how much of the schema is effectively used. A more sophisticated analysis would be testing the quality of the schemas based on how general they are. The adoption of XML Schema [32] motivates such qualitative studies notably to deal with several datatypes and the use of namespaces. Other interesting studies that fall in this category are determining the use of `ID`/`IDREF(S)` attributes, for instance.

Another interesting study we identify is comparing the distribution of content of the Web as a whole to the contents of the XML Web, in terms of document and volume distribution by zones (as in Section 2). Such a study could help identify which communities on the Web are the “driving forces” behind XML as a technology.

Acknowledgments. We would like to thank Guy Ferran and Sophie Cluet for giving us access to the Xyleme repository. We would like to thank Mariano Consens, Sergio Greco, Alberto Mendelzon and Tova Milo for their comments and useful suggestions in improving this paper. D. Barbosa and L. Mignet are supported in part by grants from the National Science and Engineering Research Council of Canada and Bell University Laboratories. D. Barbosa is supported in part by an IBM PhD Fellowship since September 2002.

5. REFERENCES

[1] Serge Abiteboul, Peter Buneman, and Dan Suciu. *Data on the Web*. Morgan Kauffman Publishers, Inc., 1999.

[2] Serge Abiteboul and Victor Vianu. Queries and Computation on the Web. In *ICDT*, pages 262–275, 1997.

[3] S. Al-Khalifa, H. V. Jagadish, N. Koudas, J. Patel, and Y. Wu. Structural joins a primitive for efficient XML query pattern matching. In *ICDE*, pages 141–152, 2002.

[4] Philip Bohannon, Juliana Freire, Prasan Roy, and Jerome Simeon. From XML Schema to Relations: A Cost-Based Approach to XML Storage. In *ICDE*, 2002.

[5] Sergey Brim and Larry Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *WWW*, 1998.

[6] Cooperative Association for Internet Data Analysis. `http://www.caida.org/`.

[7] Junghoo Cho and Hector Garcia-Molina. Finding Replicate web collections. In *SIGMOD*, pages 355–366, 2000.

[8] Byron Choi. What Are Real DTDs like. In *WebDB*, 2002.

[9] Stephen Dill, Ravi Kumar, Kevin S. McCurley, Sridhar Rajagopalan, D. Sivakumar, and Andrew Tomkins. Self-similarity in the Web. In *VLDB*, pages 69–78, 2001.

[10] Ronald Fagin, Phokion G. Kolaitis, Reneé J. Miller, and Lucian Popa. Data exchange: Semantics and query answering. In *ICDT*, to appear. 2003.

[11] IBM DB2 v8.1. `http://www.ibm.com`.

[12] Internet Domain Survey. `http://www.isc.org/ds/`.

[13] Panagiotis Iperiotis, Luis Gravano, and Mehran Saham. Probe, Count, and Classify: Categorizing Hidden Web Databases. In *SIGMOD*, pages 67–78, 2001.

[14] S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, D. Sivakumar, Andrew Tomkins, and Eli Upfal. The Web as a Graph. In *PODS*, pages 1–10, 2000.

[15] Quanzhong Li and Bongki Moon. Indexing and Querying XML Data for Regular Path Expressions. In *VLDB*, pages 361–370, 2001.

[16] Ioana Manolescu, Daniela Florescu, and Donald Kossmann. Answering XML Queries on Heterogeneous Data Sources. In *VLDB*, pages 241–250, 2001.

[17] data ex machina. http://www.dataexmachina.de/natix.html.

[18] Oracle 9i. http://www.oracle.com.

[19] Y. Papakonstantinou and V. Vianu. Incremental validation of XML documents. In *ICDT*, to appear. 2003.

[20] Sriram Raghavan and Hector Garcia-Molina. Crawling the Hidden Web. In *VLDB*, 2001.

[21] A. R. Schmidt, F. Waas, M. L. Kersten, D. Florescu, I. Manolescu, M. J. Carey, and R. Busse. The XML Benchmark Project. Technical Report INS-R0103, CWI, April 2001.

[22] ISO 8879 - standard generalized markup language (SGML), 1986.

[23] The Plays of Shakespeare in XML. http://metalab.unc.edu/bosak/xml/eg/shaks200.zip.

[24] Tamino XML Server. http://www.softwareag.com/tamino.

[25] I. Tatarinov, Z. Ives, A. Halevy, and D. Weld. Updating XML. In *SIGMOD*, pages 413 – 424, 2001.

[26] Semantic Web. http://www.w3.org/2001/sw.

[27] Wireless Application Protocol. http://www.wapforum.org/.

[28] World Wide Web Consortium. eXtensible Markup Language (XML) 1.0 . http://www.w3.org/XML/.

[29] World Wide Web Consortium. Simple Object Access Protocol (SOAP) 1.1 . http://www.w3.org/TR/SOAP/.

[30] World Wide Web Consortium. The Extensible Stylesheet Language (XSL) . http://www.w3.org/Style/XSL/.

[31] World Wide Web Consortium. XML Path Language (XPATH). http://www.w3.org/TR/xpath/.

[32] World Wide Web Consortium. XML Schema. http://www.w3.org/XML/Schema.

[33] The XOO7 Benchmark. http://www.comp.nus.edu.sg/~ebh/XOO7.html.

[34] The XML benchmark project. Available from http://www.xml-benchmark.org/.

[35] DBLP XML. http://dblp.uni-trier.de/xml/.

[36] Xyleme S.A. http://www.xyleme.com/.

[37] Lucie Xyleme. A Dynamic Warehouse for XML Data of the Web. *IEEE - Data Engineering Bulletin*, 24(2):40–47, 2001.