

XHTML AS AN EMERGING INNOVATION FOR THE WORLD WIDE WEB

by  
Clifton A. Barnett

A Master's paper submitted to the faculty  
of the School of Information and Library Science  
of the University of North Carolina at Chapel Hill  
in partial fulfillment of the requirements  
for the degree of Master of Science in  
Information Science.

Chapel Hill, North Carolina

April 2005

Approved by

---

Brad Hemminger

## **Introduction**

XHTML is a metadata standard that blends the traditional World Wide Web markup language HTML with the less well-known markup language XML. It was developed by the World Wide Web Consortium (W3C). “It’s [W3C] an independent, international organization made up of people and organizations from across the Internet and Web development community – from researchers at universities such as MIT to representatives from major corporations such as Microsoft, IBM, Sun Microsystems, and Netscape.” [1] XHTML is specifically designed to facilitate tasks such as data-integration, data-transformation, and data mining; HTML, on the other hand, focuses mainly on data presentation and document linking. These two differing approaches make XHTML documents much more desirable from an indexing and retrieval viewpoint than HTML documents. A World Wide Web composed of XHTML documents, rather than HTML documents, would provide an opportunity for much greater complexity and sophistication in retrieval processes and tools. It would be one step closer to the realization of the Semantic Web [2]. Tools that take advantage of the possibilities in XHTML exist, and are in use by many organizations and individuals. However, until enough web page developers adopt this newer standard, providing incentive for widespread adoption of these tools, the average user will stick with what is already in use. Of course, unless web page developers can see a return of some sort on any investment they make in XHTML (such as an

increase in user base or user satisfaction), they will be reluctant to make this investment. The problem of how to break this stalemate, and inspire widespread adoption of XHTML, is one whose answer could have a significant impact on the future of the World Wide Web.

### **Problem Statement**

XHTML is a new markup language designed to provide a richer metadata framework for web pages than the original markup language HTML [3]. The process by which XHTML will replace HTML, or fail to replace HTML, as the markup language of choice amongst web page designers is known as diffusion. Rogers defines diffusion as “the process by which an innovation is communicated through certain channels over time among the members of a social system.” [4] This definition identifies four elements of interest in studying the diffusion of an innovation: 1) the innovation itself; 2) the communication channels by which the innovation is spread; 3) the social system among which the innovation is potentially diffused; 4) time.

In studying the diffusion process of XHTML, the social system is readily apparent. Since XHTML is a metadata standard for creating web pages, the social system can be seen to be the community of web page designers. This community includes anyone who is involved in making the decision as to which standard a web page will adhere. Consumers of web pages could also be considered a part of this community, due to their influence on design decisions. No one outside this community is in a position to use XHTML, but everyone inside of it is.

However, this social system is not a homogeneous structure. The individual units, or developers, within the system have widely divergent goals, influences, and motivations. For instance, a web page developer designing a corporation's customer website will probably have different concerns and goals for the website than a teenager that is designing his homage page to a local band. The target audience, required functionality, and content will all be different in different scenarios, and will affect the design parameters of the web page (or collection of pages).

The paths by which information regarding XHTML is transferred amongst members of this community are called communication channels [4]. The presence, or lack, of appropriate channels to convey knowledge of XHTML to potential adopters can have a significant impact on the adoption rate. Different units within the community most likely depend on different communication channels for the dissemination of information. Identifying these channels and the units to which they communicate could provide some insight into the diffusion process for XHTML.

The three units of interest in this study are developers of personal homepages, developers of academic library pages, and developers of online shopping sites. There are many possible channels of communication by which each of these units might learn about XHTML, some of which might be shared, some of which are likely unique to individual units.

There are many different channels of communication to which all three units would have reasonable access. Numerous documents exist on the internet

giving design advice. In fact, there are several communities in the form of user groups that have formed around very specific technologies. A simple web search for advice on a particular design problem would be sufficient to provide a developer with a multitude of data about various solutions to the problem, some of which will involve XHTML. The W3C, for example, has several web pages that discuss XHTML and the various applications that utilize it for web design.

Another channel to which many developers have access is the large number of books, guides, and articles in print that provide information about web design, specifically XHTML. A brief perusal of the technical section of any large bookstore will quickly show a wide selection of web design books that seek to introduce readers to XHTML and its related technologies.

A third channel that would be available to developers in a large organization (such as a corporation, or large library) would be the organizational knowledge. This will include any web design guidelines in place, official knowledge bases maintained by the organization, and the knowledge contained by the other people in the organization. This information can be passed between organizations officially, via shared initiatives, or unofficially, as individuals move between organizations, taking this knowledge with them.

These are just a few of the possible channels of diffusion for XHTML. However, each of these channels will serve to diffuse information about any web design standard, including HTML (with which XHTML is competing for market share). The question becomes, are these channels being utilized to diffuse

XHTML? If not, are there any channels being utilized for this purpose? If so, what are they?

There is a fourth element of interest in the diffusion process; time. Time is of particular interest in this study for its bearing on the rate of adoption. The rate of adoption is simply a plot of the number of individuals that have adopted an innovation over a period of time [4]. XHTML was first proposed in a W3C [5] recommendation paper in January, 2000. So far, the diffusion process for XHTML has had a period of five years to unfold. Since we have no measures of the actual number of individuals (web pages) that have adopted XHTML over this time frame, it is impossible to determine what XHTML's adoption rate curve actually looks like. In order to gain a single snapshot, I will demonstrate below a method for comparing the ratio of XHTML to HTML pages as currently exist on the World Wide Web.

### **The Diffusion Process**

So, just how does an innovation make its way along the communication channels, throughout a social system? The classical diffusion model has the innovation being developed by some core group, which then diffuses the innovation to potential adopters. These potential adopters then either accept the innovation, and diffuse it further across the system, or reject it. XHTML, being first developed by a working group within the W3C, and diffused out from that central source, fits neatly into the classical model.

Once an innovation has been developed, in order to spread, potential adopters that are unaware of the innovation must become aware that the innovation exists. To transition users from this separate state to a linked state, with respect to the innovation, requires the presence of some linking mechanism [6]. This is generally a change agent that seeks to make the potential adopter aware of the innovation, and to some extent, how adoption of the technology can be of benefit over any existing technologies.

Once a potential adopter is transitioned to the linked state, there are still obstacles that stand in the way of reaching the adopted state. Potential adopters face many, often conflicting, demands on their resources and loyalties. Time and money invested in a new technology is time and money taken away from some other endeavor. Often, potential adopters have some personal or organizational commitment to competing technologies [7]. Most users, before fully committing to any new technology, will try it out in a testing situation, also referred to as the activated state. If the new technology is not perceived as bringing some advantage to the user at this stage, the user will typically reject full adoption of the technology. This rejection might inspire the now non-potential user to dissuade other linked potential users from transitioning to the activated state. Alternately, a user that enjoys a benefit in the activated state might act to promote the diffusion of the technology by enticing other potential users to begin adopting.

There are several key factors that have been determined to play a role in the decision to adopt or reject an innovation. Characteristics of the innovation itself that influence this decision are 1) relative advantage, 2) compatibility, 3)

complexity, and 4) trialability [4]. An innovation must meet certain criteria in each of these areas before a potential adopter will actually adopt. The necessary criterion for each characteristic is determined by each potential adopter, and thus can vary greatly.

The relative advantage of an innovation is the degree to which it is perceived as being better than any existing solutions. In the case of XHTML, the existing solution is HTML. So, any individual that is considering using XHTML to design a web page must see some benefit to using XHTML that is not present in HTML. In fact, there are several possible benefits. According to the W3C, one of the advantages to XHTML is

“Document developers and user agent designers are constantly discovering new ways to express their ideas through new markup. In XML, it is relatively easy to introduce new elements or additional element attributes. The XHTML family is designed to accommodate these extensions through XHTML modules and techniques for developing new XHTML-conforming modules (described in the forthcoming XHTML Modularization specification). These modules will permit the combination of existing and new feature sets when developing content and when designing new user agents.” [8]

The ability to customize a document based on user-specific needs provides a level of sophistication that allows for much more detailed indexing and retrieval of documents. The benefits of advanced indexing and retrieval capabilities to large organizations, with a huge volume of data, can easily be seen. An example of one such organization is the library system at the University of North Carolina

at Chapel Hill. Part of their stated purpose is to “Provide a system of bibliographic and intellectual access that makes available the Library’s own holdings and relevant materials elsewhere, including information on the World Wide Web.” [9] As a part of providing information on the World Wide Web, they have adopted XHTML as the standard for their web page development. While not explicitly stated, one can only assume that their adoption of XHTML is viewed by the organization as a beneficial step towards achieving their stated goal of providing access to information.

Another benefit to content providers is the ability to easily migrate information stored in XHTML (and thus XML) format. In response to digital libraries’ concerns about technological obsolescence of their holdings, Ludascher, Marciano, and Moore proposed a knowledge-base infrastructure that is based on XML [10]. With such a knowledge base in place, any future advances in the art of information representation could more easily be incorporated into an organization’s existing infrastructure. This capability would also allow organizations to migrate information from one current format to another, in order to provide for different needs amongst content consumers. The ability to reach a wider selection of content consumers easily via data transformation is certainly an attractive possibility to many content providing organizations.

However, any decision to adopt XHTML considers the “relative advantage”, not an objective advantage. If the perceived benefits do not outweigh the perceived costs, the decision will be to reject the innovation. There are several costs to adopting XHTML. Ignoring the costs to go back and migrate

existing HTML documents to the XHTML format, there are also costs associated with any new development. One of the benefits mentioned above is the ability to customize the tag-set for a document. However, there is a cost associated with this, namely the time and effort to develop this customized tag-set. There is also the added complexity of web page development and the learning curve for XHTML versus the (presumed) familiarity with HTML.

Compatibility is not much of a concern for XHTML. Compatibility implies that the innovation is consistent with the potential adopter's needs and beliefs [4]. XHTML can certainly meet any needs that HTML is currently meeting, and does not require any modification of a potential adopter's ideals about web page development. XHTML does provide an adopter with the possibility of an enhanced web page ideology, but does not require it at all.

Complexity is a point on which XHTML could easily founder. One of the attractions of HTML is its simplicity and ease of use. XHTML, by its very nature and purpose, introduces a level of complexity in web page markup that is not present in HTML. It is possible that the benefits of adoption can be somewhat complex to understand. Less web savvy potential adopters might not grasp the concept of richer metadata, much less the idea of a "Semantic Web". Without a demonstration of the concrete possibilities inherent in XHTML, the conceptual benefits might be considered too complex.

Trialability is also a positive characteristic of XHTML. The ability to experiment with just a few web pages provides the potential adopter with less uncertainty about the innovation. If the other factors have not decided the

potential adopter one way or the other, it is a simple matter to try XHTML, and determine on that basis whether or not to adopt.

### **Purpose**

As shown above, XHTML has many characteristics of a useful innovation. The purpose of this study is to gather data on the diffusion process and the identified social system. By gathering data on web pages' document types, I intend to provide a statistical look at the number of web pages that have currently adopted XHTML. Also, collecting information about who has and who has not adopted XHTML might provide insight into which factors are at work in the social system, affecting the adoption rate of XHTML. Are there any characteristics of potential adopters that can be identified as having an impact on the adoption decision? Can these characteristics be used to identify which potential adopters are more likely to adopt? And finally, if a group that is more likely to adopt can be identified, can this information be used to draw any further conclusions about what linking mechanisms are in place to spread the diffusion of XHTML?

### **Categorization Methods**

The basic structure of this study is a binary one: XHTML and non-XHTML web pages. However, in order to impose a more useful structure on the collected data, it is necessary to construct a set of categories into which each collected web page can be placed. Ideally, these categories will be based upon the

characteristics of the member web pages that reflect areas of interest. There have been several studies that examine just which characteristics of a web page are the most useful in correctly identifying the member web pages of a specific category. Some of the most widely examined methods include text classification of the document contents [19][23], clustering techniques to place a web page within a conceptual World Wide Web 'space' [11][17], and looking at structural features that compose the web page[18][21][20]. Each of these broad methods has a variety of specific implementations that have been pursued.

### **Text classification**

Text classification of documents focuses on the textual content of a document. This method predates the web, and has been extensively studied as a method for determining relevance of documents, a task related to classification. As this method has received such attention, I will limit the discussion here to the limitations text classification has with respects to web pages.

Pure-text classification on web pages has been seen to be less accurate than other methods of web page categorization, although when used in combination with other methods, improvements are seen over either individual method [11]. Web pages often contain information objects that are not found in a typical textual document. They often involve banner advertisements and navigation bars, which may contain text that is unrelated to the topic of the web page. This text can provide white noise that may bias a pure-text classification algorithm, leading it to provide inaccurate results [12].

Important information about a web page may also be contained in the link structure of the document [13]. Of particular use in this study is the idea of comparing the ratio of text to links. For instance, a library home page is likely to have a high ratio of links to text, as it is intended to direct users to a multitude of informational resources, rather than contain this information itself. The same applies to an online shop, although it is directing users to goods/services, rather than informational resources.

### **Clustering techniques**

Many studies have focused on classifying web documents by examining their semantic space, or relationship to other documents that are near them on the World Wide Web. This can be defined as a function of the documents to which they link, or which link to them [14]. An augmentation to this method is to include the text of the neighboring (linked) documents. In [16], the text of neighboring documents that linked into the target document was considered, and found to be of some use in improving classification accuracy. However, when the text of a neighboring document that was linked to from the target document was considered, it was found to reduce accuracy [15].

Another method of considering a document's semantic space is by examining the path a user took to reach this page, referred to as the browsing path [17]. The various paths by which users travel to a page help to define the page's relationship to all the other pages contained in each path.

For this study, I utilize the idea of considering a page's semantic space in determining its classification in only a basic manner. Online shops are likely to link to pages within the company's domain only, as most companies are not eager to send business to their rivals. However, personal homepages are much more likely to link to a widely spaced document set, as there is not financial incentive to not do so, and most personal homepages serve as a quasi-portal to the various subjects in which the publisher is interested. Academic library pages fall into the middle, linking to many resources within their own site, but also linking to sources outside their site when this is more convenient or effective.

### **Structural features**

A third method of categorizing a web page is to identify similar web pages as pages that have similar structural characteristics contained in the page.

Matsuda and Fukushima identified seven such structural characteristics:

KEYWORD, LINK, URL, STRUCTURE, IMAGE, OCR, and PLUGIN [18].

Keyword, link, url, and image are familiar terms to anyone involved in web page information retrieval. Structure, Ocr, and Plugin, however, require a brief explanation.

In the methodology proposed by Matsuda and Fukushima, 'Structure' refers to the order inherent in the markup tags that define the web page. For instance, the fact that a web page organizes information using the <table> tag could be captured using the Structure element. 'Ocr' is very similar to a 'Keyword', only differing in scope. A 'Keyword' can refer to any combination of

terms that are found on the web page. ‘Ocr’, for optional character recognition, refers specifically to text paired with an image, such as a caption or <alt> designation. A ‘Plugin’ designates a neighboring document that links into the target document. With these elements, the structural based approach takes into account the utility of both of the two previously discussed approaches, namely text classification (Keyword and Ocr) and clustering (Link and Plugin). In addition, however, this approach also considers a characteristic that is an inherent element of any web page, namely the tag structure of its markup language (‘Structure’).

I utilize the tag element extensively in my classification schemas, particularly in combination with keywords. For instance, most online shops will have a link, denoted by the <a href> tag, that directs the user to a shopping cart/basket. It is highly unlikely that either of the other two categories would have such a link.

The structural element URL has also been found to contain a surprisingly rich amount of data about a web page [19]. The top level domain, such as .com or .edu, can give broad clues as to the category of web page. While perhaps not an absolute rule, the vast majority of university library sites are going to be under the .edu domain. Alternately, most commerce sites (such as an online shopping site) will be under the .com domain. Also, the majority of library pages will have “lib”, “library”, or some variation on one of the two in its URL.

## **Methodology**

I constructed three categories using a strategy that incorporates structural features of the document as well as the hypertext linkage patterns. This strategy is based on the assumption that websites with similar functionalities will have similar structural features similarly exhibited on the page [24]. The importance of constructing categories based on functionality comes from my underlying assumption that web pages with similar functionality will have been developed to meet similar needs and goals. If the organizations/individuals responsible for the web pages have similar needs and goals, there is a good chance that there will be other common characteristics shared by these entities. By examining these characteristics, I hope to arrive at some useful conclusions about the role these entities play in the diffusion process for XHTML.

Each of the categories I constructed aims at representing one of three broad categories of development entity – individuals, corporations, and academic institutions. I began by intuitively listing features in each of the categories that I felt should ideally appear on a web page of that type. Once I had developed the initial category definitions, I then refined them by doing a manual comparison to a few samples of each. For example, I refined the online-shop category by comparing my initial definition to amazon.com, barnesandnoble.com, and walmart.com. The final definitions at which I arrived are specified in Appendix A.

Once I had the category definitions, I began collecting data. To do this, I utilized the corpus of the Open Directory Project (DMOZ). In order to facilitate a

manual search, I ran a search against “online shop”, “personal homepages”, and “university library”, comparing the result sets to the appropriate categories. A page was considered to fit into a category if and only if it exhibited each of the features listed in the category definition. A page was considered to be XHTML compliant if its document type definition, located in the source header, was XHTML (!DOCTYPE XHTML), or if the document linked to the XHTML DTD in the header. While this does not reflect true compliance to the XHTML standard, it does give sufficient cause to assume adoption of XHTML. A document was considered to be HTML if its document type definition was HTML, or if it had no document type definition, but was merely well-formed HTML. I collected 75 sample pages for each category.

In an effort to begin looking at XHTML over time, I utilized the Wayback machine ([www.archive.org](http://www.archive.org)) to look at archived copies of each sample page that was determined to be XHTML compliant. Some of the pages were not archived, due to restrictions on automatic indexing of these pages. By looking at these archived pages, I was able to determine the approximate date that most of the adopters first implemented XHTML. These dates reflect the first date that an XHTML compliant version of the page was archived, not necessarily the actual first date an XHTML compliant version of the page was implemented. However, the frequency of the archives provides an uncertainty period of less than a month, which is an acceptable uncertainty level.

## Results

*Table A – Total counts for each category*

Total pages/category =75	Total # of XHTML Pages	Total # of HTML Pages	% XHTML
<b>Online Shops</b>	6	69	8.00%
<b>Personal Homepages</b>	12	63	16.00 %
<b>University Library Homepages</b>	19	56	25.33%

*Table B-1: Adoption dates for Online Shops*

<b>URI of Online Shop</b>	<b>Approximate date of adoption</b>
<a href="http://www.marksandspencer.com/">http://www.marksandspencer.com/</a>	June 11, 2004
<a href="http://www.tesco.com/">http://www.tesco.com/</a>	August 21, 2004
<a href="https://www.europe.redhat.com/shop/en/">https://www.europe.redhat.com/shop/en/</a>	November 2, 2004
<a href="http://www.kylieshop.com/mall/departmentpage.cfm/kmen/">http://www.kylieshop.com/mall/departmentpage.cfm/kmen/</a>	November 10, 2004
<a href="http://www.instore.com/instore/CategoryPage?id=1">http://www.instore.com/instore/CategoryPage?id=1</a>	November 14, 2004
<a href="https://shop.mysql.com/">https://shop.mysql.com/</a>	No archived data

*Table B-2: Adoption dates for Personal Homepages*

<b>URI of Personal Homepage</b>	<b>Approximate date of adoption</b>
<a href="http://homepages.ihug.co.nz/~dhbayne/">http://homepages.ihug.co.nz/~dhbayne/</a>	October 10, 2002
<a href="http://faculty.washington.edu/naosok/">http://faculty.washington.edu/naosok/</a>	December 8, 2003
**	**

\*\* The remaining ten adopting units within the personal homepage category had no archive data available.

*Table B-3: Adoption dates for University Academic Library Homepages*

<b>URI of Library Homepage</b>	<b>Approximate date of adoption</b>
<a href="http://www.lib.cmich.edu/">http://www.lib.cmich.edu/</a>	June 03, 2002
<a href="http://library.ttu.edu/ul/">http://library.ttu.edu/ul/</a>	December 06, 2002
<a href="http://info.lib.uh.edu/">http://info.lib.uh.edu/</a>	February 03, 2003
<a href="http://library.ust.hk/">http://library.ust.hk/</a>	July 19, 2003
<a href="http://stauffer.queensu.ca/">http://stauffer.queensu.ca/</a>	January 01, 2004
<a href="http://www-sul.stanford.edu/">http://www-sul.stanford.edu/</a>	February 04, 2004
<a href="http://infolib.berkeley.edu/">http://infolib.berkeley.edu/</a>	March 21, 2004
<a href="http://www.lib.unc.edu/">http://www.lib.unc.edu/</a>	May 24, 2004
<a href="http://www.lib.ksu.edu/">http://www.lib.ksu.edu/</a>	June 11, 2004
<a href="http://library.duke.edu/">http://library.duke.edu/</a>	July 26, 2004
<a href="http://www.lib.utulsa.edu/">http://www.lib.utulsa.edu/</a>	August 21, 2004

<a href="http://www.lib.utexas.edu/">http://www.lib.utexas.edu/</a>	August 27, 2004
<a href="http://www.lib.ua.edu/">http://www.lib.ua.edu/</a>	August 28, 2004
<a href="http://www.lib.ucdavis.edu/">http://www.lib.ucdavis.edu/</a>	September 20, 2004
<a href="http://wwwlib.gsu.edu/">http://wwwlib.gsu.edu/</a>	September 23, 2004
<a href="http://www.lib.muohio.edu/">http://www.lib.muohio.edu/</a>	September 24, 2004
<a href="http://www.libraries.psu.edu/">http://www.libraries.psu.edu/</a>	March, 2005
<a href="http://library.usask.ca/">http://library.usask.ca/</a>	March, 2005
<a href="http://www.ull.ac.uk/">http://www.ull.ac.uk/</a>	March, 2005

### **Study Limitations**

There are a few major limitations to this study. The biggest one is the limitation on sample size imposed by the manual collection of the data. An application that automatically crawled the DMOZ corpus, comparing each page to all three category definitions would greatly enhance the statistical usefulness of this study. Of course, the design and implementation of such an application would require extensive testing to ensure that it provided an acceptable level of classification accuracy compared to the manual approach.

Another limitation to this study is the binary model of category fit. A web page is either in or out of a category, there is no provision made for determining a range of probabilities of fit. Ranking the features in terms of importance, reflected by attaching a weight to each feature, would allow for the determination

of such a range. This would greatly enhance the study, especially when included into an automated application.

A third limitation of the study is the language barrier. Due to my lack of linguistic ability, I was forced to only consider web pages whose content was in English. This narrows the sampling frame significantly, and could overlook national or cultural trends in the adoption of XHTML.

## **Conclusions**

Table A gives a breakdown of the actual numbers of XHTML compared to HTML pages by category, as determined by this study. Looking at this table, it is plain that XHTML has not penetrated very far into any of the three developer communities. The largest market share achieved was in university libraries, with just slightly over one quarter of the pages surveyed having adopted XHTML. Online shops brought up the end, with roughly one twelfth of the pages having adopted XHTML. Personal homepages fell in between these two, with about one sixth of the pages being adopters.

University libraries have enough adopters within the category to assume that there is some change agent at work, promoting the adoption of XHTML amongst this community. Given the nature of academia, it is not surprising that there are sufficient channels of communication between libraries to propel the adoption of XHTML. However, examining the data does not provide easy clues as to what these channels might be. The adopter libraries are geographically diverse, with several on either coast of the American continent, some in the

middle, and a few overseas. Two of the adopter libraries, UNC-Chapel Hill and Duke, are members of a library network, the Triangle Research Library Network (TRLN). However, the other two members of this network, North Carolina State University and North Carolina Central University, are non-adopters. Some of the universities in the California system are adopters, whereas others are not (see Appendix B). So, unsurprisingly, it does not appear that geography plays a role in the linking mechanisms between adopters. Somewhat more surprisingly, it also does not appear that membership in larger organizations (such as TRLN) has provided any successful linking mechanisms between adopting units. Presumably, this indicates that the decision to adopt or not is based upon characteristics intrinsic to each unit, rather than the outside influence of other units.

This supposition is supported by the patterns of adoption seen in the other categories. The adopting units in the online shops category are open source technology based companies, such as MySQL and Red Hat, which could be expected to embrace W3C open source standards. Marks and Spencer is a large clothing retailer that is an adopter, the only non-technology based large corporation to do so. Several smaller representatives of online shops, which are more likely to be influenced by an individual's decision than their larger counterparts, were also adopters.

However, the vast majority of the larger companies sampled, where an individual would have much less influence than the corporate culture, were non-adopters. In libraries, where an individual can have a much greater influence, we

see a much greater percentage of adoption. But what is it that causes these individuals to become change agents? What channel of communication do these individuals all have access to that others either do not, or do not find convincing?

A clue to this question can be found by examining the patterns of adoption in the personal homepages category. Ten of the twelve homepages that were adopters belonged to members of a university, as evidenced by their URLs and examination of their pages. So, it would seem that the most common adoption units are those that exist in the academic community.

The channels of communication that are most closely associated with academia are web design courses and research articles. However, members of other communities also have access to these resources. Therefore, it seems unlikely that the disparity of adopters between academic and corporate communities could be explained solely by relying on communication channels. While corporations may have access to information resources universities do not (e.g., corporate knowledge bases), universities do not have access to resources that corporations do not. In fact, most larger corporations recruit employees heavily from amongst the best and the brightest universities have to offer.

So, if communication channels are not the cause of this disparity, what is? My belief is that this disparity can be explained by the different motivations that govern each group. Corporations are motivated by a desire for profit, where academicians are motivated by a desire to promote human knowledge, or by a desire for respect within the academic community. Individuals, in regard to their

homepages, are more likely to be motivated by a desire for respect, or a desire to promote knowledge, than by any expectation of profit.

Based upon this initial study, it would seem that the adoption of XHTML has sufficient channels to communicate knowledge of the diffusion throughout the various communities involved. It would seem to be the motivations, or change agents, behind the adoption that differ from community to community. A purely financial assessment of XHTML, as one would expect a corporation to make, apparently favors remaining with HTML. However, when considering more intangible rewards, such as respect or intellectual advancement, as one might expect motivates academicians or individuals, XHTML is apparently seen as preferable to HTML.

Consideration of these motivations as they might affect a unit's eagerness or reluctance to innovate also leads to the likelihood of academicians being early adopters, with corporations being later adopters. Assuming that academicians are motivated by a desire for respect from their peers would tend to make them early adopters, given Rogers' description of early adopters' social position [4]. Corporations, however, are bound by more restrictions regarding how they risk their resources, and thus might be less likely to invest until a technology is proven. They are more likely to be skeptical of a new innovation, fitting with Rogers' description of late adopters.

In fact, the data found in tables B-1 and B-3 supports this hypothetical classification of academia and corporate culture. About one half of the university libraries that adopted XHTML adopted before the first online shop. The small

number of personal homepage units from table B-2, are not enough on which to base any conclusions, but do support the idea of corporate units being later adopters. The only two developers of personal homepages for whom archived data was available both fit into the early category of the collected dates.

Therefore, it would seem that XHTML is still in the early phases of its diffusion process. Likely early adopters, such as university libraries, appear to have begun embracing XHTML, recently at an increasing rate. Over 50% of the adopters in the university library category have adopted within the last year, some within the last few months.

Likely later adopters, such as corporations, appear to have yet to be convinced in significant numbers of the value of adopting. Only a small percentage of these units have already implemented the adoption stage. However, given the very recent increase in the rate of adoption amongst our earlier adopters, it is possible that the small percentage of corporate adopters seen here represent the initial wave of adopters amongst the majority adopter category. If so, this might mean we can expect to see a much more rapid rate of adoption for XHTML in the near future.

### **Future Work**

There are many questions that need to be answered to gain a truly accurate picture of the diffusion process of XHTML. Future data points along the time line will need to be collected. This data will be necessary to determine the changing rate of increase in the adoption rate. It will also provide a framework in

which to judge the accuracy of my hypothesis about the adoption phase in which XHTML is currently. If XHTML is successfully diffusing, one would expect to see larger numbers of individuals and corporations adopting in the future. One would also expect to see some group representing the laggards appear, although I suspect this will not be a new group, merely a subset of one of the three groups identified by this study. Alternately, if a number of the identified early adopters discontinue implementation of XHTML, it will indicate that XHTML is likely to be an unsuccessful innovation. One method of determining this would be to revisit the pages already examined to see if there has been a change in their status.

Another issue that needs to be explored is the idea of other adoption groups. Are there other groups not related to these three that are involved in the diffusion process of XHTML? If so, what communication channels and linking mechanisms are there between these unknown groups and the ones already examined? Any future studies would do well to keep these questions in mind.

## Appendix A -- Web Page Category Structural Definitions

**Personal Homepages** – this category represents individual web page providers. The pages themselves generally serve as the individual’s World Wide Web presence, presenting to the cyber world the characteristics the individual would like to emphasize.

KEYWORD : <title> person’s name

IMAGE: person’s picture – Optional, but highly indicative of type if it appears

LINK: a personal section

LINK: <mailto: >Email address | contact me

Will have a high ratio of outgoing links-to-inbound links

The outgoing links will be widely-spread over web-space, if not topic

**Online Shop** – this category represents business organizations as web page providers. The purpose of these pages is to provide users with some basic information about the available products, and also to provide a convenient method for purchasing these products.

KEYWORD: <title>Name of store

STRUCTURE: <input type=”text”> (a search box for searching products)

OR <li><a href> (a list of links representing a product menu)

STRUCTURE: <a href> checkout | cart | basket

Will have very few links that refer outside the shop domain

**University Academic Library Homepage** – this category represents academic organizations as web page providers. The purpose of these pages is to act as a portal to information provided by a University’s library

KEYWORD: <title> Library

URL: the host will be within the .edu domain

URL: will contain “lib”, “library”, or some variation of one of these

LINK: <a href>Libraries | Collections (link to other libraries)

Will have a large ratio of links-to-text

## Appendix B -- Collected data

### Online Shops

X - <http://www.marksandspencer.com/> -- June 11, 2004  
 X - <https://shop.mysql.com/> -- No data  
 X - <http://www.tesco.com/> - August 21, 2004  
 X - <http://www.kylieshop.com/mall/departmentpage.cfm/kmen/> -- November 10, 2004  
 X - <http://www.instore.com/instore/CategoryPage?id=1> -- November 14, 2004  
 X - <https://www.europe.redhat.com/shop/en/> -- November 2, 2004

H - <http://www.lillianvernon.com/home.jsp?bs=1>  
 H - <http://www.chessbase.com/shop/index.asp>  
 H - <http://www.hawaiiflowerlei.com/>  
 H - <http://www.inkjetcartridges.com/>  
 H - <http://www.masterg.com/supplies.html?src=ssn>  
 H - <http://shop.upromise.com/browse.php>  
 H - <http://www.indiangiftsportal.com/>  
 H - <http://www.mountwashington.org/shop/>  
 H - <http://www.dancingwind.com/>  
 H - <http://www.smartbargains.com/>  
 H - <http://www.belkin.com/index.asp>  
 H - <http://www.officedepot.com/>  
 H - <http://shop.npr.org/>  
 H - <http://www.sinnfeinbookshop.com/>  
 H - <http://www.buy4now.ie/>  
 H - <http://www.iee.org/shop/>  
 H - <http://www.macys.com/?bhcp=1>  
 H - <https://www.iataonline.com/Store/default.htm?cookie%5Ftest=1>  
 H - <http://shop.borland.com/>  
 H - <http://www.ems.com/>  
 H - <http://secure.www.oldnavy.com/asp/home.html?wdid=0>  
 H - <http://www.cathaypacific.com/intl/pretrip/cxcitement/0,,00.html>  
 H - <http://www.target.com/gp/homepage.html/601-2810009-2656110?>  
 H - <http://idjnow.com/>  
 H - <http://www.djtools.com/>  
 H - <http://www.jbsmusic.co.uk/>  
 H - <http://www.jws-uk.co.uk/acatalog/index.html>  
 H - <http://www.storedj.com.au/site.htm>  
 H - <http://www.aardvarkstore.com/cart/Cigarsintro.cfm>  
 H - <http://www.absolutecigars.com/>  
 H - <http://cigarprices.com/>  
 H - [http://www.h-s.co.uk/cgi-bin8/web\\_store.cgi](http://www.h-s.co.uk/cgi-bin8/web_store.cgi)  
 H - <http://www.bellini-baskets.com/>  
 H - <http://www.caribbeantastes.com/>  
 H - <http://basketsoftreasuresbl.com/>  
 H - <http://www.justgifts.co.za/catalog/>  
 H - <http://www.amazon.com/exec/obidos/subst/home/home.html/102-4942527-4643321>  
 H - <http://shopping.yahoo.com/>  
 H - <http://shop.abc.net.au/>  
 H - <http://www.about.com/shopping/>  
 H - <http://www.over2u.com/shop/>  
 H - <http://www.overstock.com/>  
 H - <http://www.zappos.com/welcome3.zhtml?0328>  
 H - <http://www.soleas.com/shoes>

H - <http://www.canadiantire.ca/index.jsp>  
 H - <http://www.microcenter.com/>  
 H - <http://onlineshop.rnib.org.uk/>  
 H - <http://www.qvc.com/>  
 H - <http://www.mfa.org/shop/>  
 H - [http://www.basspro.com/servlet/catalog.OnlineShopping?CMID=MH\\_HOME](http://www.basspro.com/servlet/catalog.OnlineShopping?CMID=MH_HOME)  
 H - <http://www.shopnbc.com/>  
 H - <http://www.dragonweave.com/>  
 H - <http://www.le-shop.ch/>  
 H - <http://www.tate.org.uk/shop/>  
 H - [http://shop.usps.com/cgi-bin/vsbv/postal\\_store\\_non\\_ssl/home.jsp](http://shop.usps.com/cgi-bin/vsbv/postal_store_non_ssl/home.jsp)  
 H - <http://www.mind.org.uk/osb/showitem.cfm/Category/0>  
 H - <http://www.rei.com/>  
 H - <http://www.kodak.com/eknec/>  
 H - <http://siemens.letstalk.com/brands/siemens/>  
 H - <http://www.cafepress.com/utchsshop>  
 H - <http://www.roughtrade.com/site/index.lasso>  
 H - <http://shop.indiainfo.com/Layouts/Templates/Default/index.asp>  
 H - <http://www.cafepress.com/seahorses>  
 H - <http://www.jcpenney.com/jcp/default.aspx>  
 H - <http://www.artinstituteshop.org/>  
 H - <http://www.durrellwildlife.org/index.cfm?a=3>  
 H - <http://www.vegansociety.com/catalog/default.php>  
 H - <http://www.llbean.com/>  
 H - <http://www.kmart.com/home.jsp>

## Personal Homepages

X - <http://faculty.washington.edu/naosok/> -- December 8, 2003  
 X - <http://www.trinity.unimelb.edu.au/theology/homepages/mcgowan.shtml> -- No data  
 X - <http://www.trinity.unimelb.edu.au/theology/homepages/cruickshank.shtml> -- No data  
 X - <http://www.trinity.unimelb.edu.au/theology/homepages/devos.shtml> -- No data  
 X - <http://www.trinity.unimelb.edu.au/theology/homepages/fishburn.shtml> -- No data  
 X - <http://www.trinity.unimelb.edu.au/theology/homepages/kpower.shtml> -- No data  
 X - <http://www.trinity.unimelb.edu.au/theology/homepages/french.shtml> -- No data  
 X - <http://www.trinity.unimelb.edu.au/theology/homepages/obrien.shtml> -- No data  
 X - <http://www.trinity.unimelb.edu.au/theology/homepages/treloar.shtml> -- No data  
 X - <http://homepages.ihug.co.nz/~dhubayne/> -- October 10, 2002  
 X - <http://www.faculty.iu-bremen.de/hjaeger/> -- No data  
 X - <http://www.edgore.com/> -- No data  
  
 H - <http://www.logic.at/people/terwijn/>  
 H - <http://www.cecilw.com/>  
 H - <http://homepages.paradise.net.nz/kliomuse/index1.html>  
 H - <http://www.comp.nus.edu.sg/~luajooaha/home.htm>  
 H - <http://theory.ipm.ac.ir/~mahdi/>  
 H - <http://www2.gsb.columbia.edu/faculty/jstiglitz/>  
 H - <http://www.one-eyed-alien.net/~janet/>  
 H - <http://www.stallman.org/>  
 H - <http://lachlan.bluehaze.com.au/>  
 H - <http://alandlew.mysite.wanadoo-members.co.uk/>

H - <http://www.gmu.edu/departments/economics/bcaplan/>  
H - <http://www.esm.vt.edu/~danko/>  
H - <http://www.electricpenguin.com/ohi/main.html>  
H - <http://www.dwheeler.com/>  
H - <http://www2.bitstream.net/~krajewsk/>  
H - <http://www.ifa.hawaii.edu/faculty/barnes/barnes.html>  
H - <http://www.magres.nottingham.ac.uk/~mansfield/>  
H - <http://rford.home.igc.org/>  
H - <http://polaris.gseis.ucla.edu/pagre/>  
H - <http://www.physics.arizona.edu/~fanglz/>  
H - <http://incolor.inetnebr.com/gaskell/gaskell.html>  
H - <http://www.econ.ucy.ac.cy/~echalias/>  
H - <http://www.ece.cmu.edu/~pueschel/>  
H - [http://www.phyast.pitt.edu/People/Faculty/A\\_Connolly.htm](http://www.phyast.pitt.edu/People/Faculty/A_Connolly.htm)  
H - <http://www.stanford.edu/~duffie/>  
H - <http://www.lub.lu.se/netlab/staff/koch.html>  
H - <http://www.klenow.com/>  
H - <http://cowles.econ.yale.edu/faculty/bergemann.htm>  
H - <http://www2.am.uni-erlangen.de/~kocvara/personal/index.shtml>  
H - <http://www.emanator.demon.co.uk/bigclive/>  
H - <http://www.davelane.ca/>  
H - <http://www.pcr.uu.se/personal/anstalda/wallensteen.htm>  
H - <http://irmen.razorvine.net/>  
H - <http://www.cs.iit.edu/~xli/>  
H - <http://www.ifa.hawaii.edu/~meech/>  
H - <http://wwwhome.math.utwente.nl/~endrayantoai/>  
H - <http://tigger.uic.edu/~pdoran/home.htm>  
H - <http://theory.itp.ucsb.edu/~deholz/>  
H - <http://www.cs.man.ac.uk/~zhangy/>  
H - <http://pemoreau.neuf.fr/>  
H - <http://www.cogs.susx.ac.uk/users/blayw/>  
H - <http://lsb.scu.edu/~dklein/>  
H - <http://www.haskell.org/~simonmar/>  
H - <http://chandra.as.utexas.edu/~kormendy/>  
H - <http://www.cs.bham.ac.uk/~jxb/>  
H - <http://www.phon.ucl.ac.uk/home/hchung/>  
H - <http://web.mit.edu/spb/www/home.html>  
H - <http://www.math.ntnu.no/~norsett/>  
H - <http://www.math.ku.dk/~solovej/>  
H - <http://graphics.ethz.ch/~grossm/>  
H - <http://csmr.ca.sandia.gov/~krlong/>  
H - <http://www.zpr.uni-koeln.de/~schliep/>  
H - <http://members.globule.org:8041/~elth/http-index.html>  
H - <http://paulav.com/>  
H - <http://ira.stojanovic.online.fr/>  
H - <http://www.public.iastate.edu/~vardeman/>  
H - <http://www.ida.his.se/~sanny/>  
H - <http://www.zerocut.com/als/als.html>  
H - <http://www.math.udel.edu/~sturm/>  
H - <http://tony.veggiedude.com/>  
H - <http://www.rulequest.com/Personal/>  
H - <http://www.math.ias.edu/~misha/personal.html>  
H - <http://zap.to/helmer>

## University Libraries

X - <http://www.lib.unc.edu/> -- May 24, 2004  
 X - <http://www.lib.ua.edu/> -- August 28, 2004  
 X - <http://www-sul.stanford.edu/> -- February 4, 2004  
 X - <http://infolib.berkeley.edu/> -- March 21, 2004  
 X - <http://www.lib.utexas.edu/> -- August 27, 2004  
 X - <http://library.duke.edu/> -- July 26, 2004  
 X - <http://info.lib.uh.edu/> -- February 3, 2003  
 X - <http://library.ust.hk/> -- July 19, 2003  
 X - <http://www.libraries.psu.edu/> -- March, 2005  
 X - <http://library.usask.ca/> -- March, 2005  
 X - <http://www.lib.ksu.edu/> -- June 11, 2004  
 X - <http://www.lib.muohio.edu/> -- September 24, 2004  
 X - <http://www.ull.ac.uk/> -- March, 2005  
 X - <http://stauffer.queensu.ca/> -- January 1, 2004  
 X - <http://wwwlib.gsu.edu/> -- September 23, 2004  
 X - <http://www.lib.ucdavis.edu/> -- September 20, 2004  
 X - <http://library.ttu.edu/ul/> -- December 06, 2002  
 X - <http://www.lib.cmich.edu/> -- June 03, 2002  
 X - <http://www.lib.utulsa.edu/> -- August 21, 2004

H - <http://spirit.lib.uconn.edu/>  
 H - <http://www.lib.virginia.edu/>  
 H - <http://www.lib.cam.ac.uk/>  
 H - <http://www.lib.umich.edu/>  
 H - <http://www.library.yale.edu/>  
 H - <http://www.usm.maine.edu/~maps/>  
 H - <http://www.library.uq.edu.au/pse/index.html>  
 H - <http://www.library.okstate.edu/>  
 H - <http://www.nccu.edu/library/shepard.html>  
 H - <http://library.brandeis.edu/>  
 H - <http://www.library.cornell.edu/>  
 H - <http://www.coaps.fsu.edu/lib/>  
 H - <http://www.lib.auburn.edu/>  
 H - <http://panther.bsc.edu/~libref/>  
 H - <http://library.samford.edu/>  
 H - <http://www.eocc.edu/library/index.htm>  
 H - <http://www2.una.edu/library/>  
 H - <http://www.uah.edu/library/>  
 H - <http://www.lib.uaa.alaska.edu/>  
 H - <http://thorplus.lib.purdue.edu/>  
 H - <http://www.lib.uchicago.edu/e/index.html>  
 H - <http://www.library.cmu.edu/>  
 H - <http://www.library.ucsf.edu/>  
 H - <http://www.library.northwestern.edu/>  
 H - <http://www.library.utoronto.ca/>  
 H - <http://www.library.wisc.edu/>  
 H - <http://www.libs.uga.edu/>  
 H - <http://lib.harvard.edu/>  
 H - <http://www.lib.umn.edu/>  
 H - <http://www.lib.utk.edu/>  
 H - <http://www.lib.uci.edu/>  
 H - <http://drseuss.lib.uidaho.edu/>

H - <http://www.lib.ncsu.edu/>  
H - <http://www.library.ucsb.edu/>  
H - [http://www.brown.edu/Facilities/University\\_Library/](http://www.brown.edu/Facilities/University_Library/)  
H - <http://www.library.uq.edu.au/>  
H - <http://www.uic.edu/depts/lib/>  
H - <http://www.library.ohiou.edu/index.htm>  
H - <http://gulib.lausun.georgetown.edu/>  
H - <http://lib.nmsu.edu/>  
H - <http://library.tamu.edu/portal/site/Library>  
H - <http://www.library.kent.edu/page/10000>  
H - <http://www.asu.edu/lib/>  
H - <http://www.uky.edu/Libraries/>  
H - <http://www.lib.ed.ac.uk/>  
H - <http://www.library.okstate.edu/>  
H - <http://www-lib.iupui.edu/>  
H - <http://www.libraries.iub.edu/>  
H - <http://www.sc.edu/library/>  
H - <http://ublib.buffalo.edu/libraries/>  
H - <http://www.lib.fsu.edu/>  
H - <http://www.lib.flinders.edu.au/>  
H - <http://www.wfu.edu/Library/>  
H - <http://www.lib.ndsu.nodak.edu/>  
H - <http://elibrary.unm.edu/>  
H - <http://www.uakron.edu/libraries/index.php>

## References

- [1] Holzschlag, M., *Special Edition Using XHTML*. Que, 2001.
- [2] Berners-Lee, T., Fischetti, M., *Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor*, HarperSanFrancisco, 1999.
- [3] Musciano, C. *HTML and XHTML, the definitive guide*. O'Reilly, 2000.
- [4] Rogers, Everett. *Diffusion of Innovations*. Macmillan Publishing Co., Inc. 1983.
- [5] W3C. *XHTML 1.0: The Extensible HyperText Markup Language* (2005) <http://www.w3.org/TR/xhtml1/>.
- [6] Jenkins, R., Chapman, R., *The process of adoption*. December 1998 Proceedings of the 30<sup>th</sup> Conference [of the ACM] on Winter Simulation.
- [7] Mark, G., Poltrock, S., *Shaping technology across social worlds: groupware adoption in a distributed organization*. Proceedings of the 2003 International ACM SIGROUP conference on Supporting Group Work.
- [8] W3C. *XHTML<sup>TM</sup> 1.0: The Extensible HyperText Markup Language* (2004) <http://www.w3.org/TR/2000/REC-xhtml-20000126/#why>.
- [9] University of North Carolina at Chapel Hill. *University Library System-Administration-Mission of the Library System* (2004) <http://www.lib.unc.edu/aoffice/about/mission.html>.
- [10] Ludasher, B., Marciano, R., Moore, R., *Special section on advanced XML data processing: Preservation of digital data with self-validating, self-instantiating knowledge based archives*. ACM SIGMOD Record, September 2001, v. 30 issue 3.
- [11] Denoyer, L., Vittaut J., Gallinari, P., Brunessaux Sylvie, Brunessaux Stephan, *Structured Multimedia Document Classification*, November 2003 Proceedings of the 2003 ACM symposium on Document engineering.
- [12] Shen, D., Chen, Z., Yang, Q., Zeng, H., Zhang, B., Ma, Y., *Text Classification: Web-page Classification through Summarization*, July 2004 Proceedings of the 27th annual international conference on Research and development in information retrieval.

- [13] A. Kolcz, V. Prabaharmurthi, J.K. Kalita. *Summarization as feature selection for text categorization*. Proc. Of CIKM01, 2001.
- [14] Kwon, O., Lee, J., *Text categorization based on k-nearest neighbor approach for Web site classification*, Information Processing & Management v.39, issue 1
- [15] Chakrabarti, S., Dom, B., Indy, P. *Enhanced hypertext categorization using hyperlinks*. June 1998 ACM SIGMOD Record , Proceedings of the 1998 ACM SIGMOD international conference on Management of data, v.27 Issue 2
- [16] Glover, E., Tsioutsouluklis, K., Lawrence, S., Pennock, D., Flake, G. *Using Web Structure for Classifying and Describing Web pages*. May 2002 Proceedings of the eleventh international conference on World Wide Web.
- [17] Grotzky, W. I., Sreenath, D. V., Fotouhi, F. *Special section on semantic web and data management: Emergent semantics and the multimedia semantic web*. December 2002 ACM SIGMOD Record, v. 31 issue 4.
- [18] Matsuda, K., Fukushima, T., *Task-Oriented World Wide Web Retrieval by Document Type Classification*. November 1999 Proceedings of the eighth international conference on Information and knowledge management.
- [19] Yu, H., Han, J., Chang, K., *PEBL: Positive Example Based Learning for Web page Classification Using SVM*. July 2002 Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.
- [20] Asirvatham, A., Ravi, K., *Web page Categorization based on Document Structure* (2005), <http://gdit.iiit.net/~arul/paper.pdf>
- [21] Kan, M., *Web page Categorization without the Web page*. May 2004 Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters
- [22] Yu, H., Chang, K., Han, J. *Heterogeneous Learner for Web page Classification*, December 2002 Proceedings of the 2002 IEEE International Conference on Data Mining.
- [23] Joachims, T., *A Statistical Learning Model of Text Classification for Support Vector Machine*, September 2001 Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval.
- [24] Amitay, E., Carmel, D., Darlow, A., Lempel, R., Soffer, A. *The Connectivity Sonar: Detecting Site Functionality by Structural Patterns*. August 2003 Proceedings of the fourteenth ACM conference on Hypertext and hypermedia.