*Malaysian Journal of Library and Information Science, Vol.11, no.2, Dec 2006: 1-20*

# INFORMATION DESCRIPTION AND DISCOVERY METHOD USING CLASSIFICATION STRUCTURES IN WEB

**Mohammad Nasir Uddin[1], Muhammad Mezbah-ul-Islam[2], Kazi Mostak Gausul Haque[2]**
[1]Dept. of Library & Information Science, University of Rajshahi
[2]Dept. of Information Science & Library Management, University of Dhaka, BANGLADESH
email: nasiruddin26@gmail.com; mezbah2000@yahoo.com, kazimostaq@yahoo.com

***ABSTRACT***
*Classification structures are useful for organising and finding information. The right use of classification structure in the web information architecture provides a user friendly interface and can serve as an effective tool for information retrieval. This paper defines the concepts related to the recent development in classification and gives an overview with case studies of using classification structure in information description and discovery in Web. The study found that hierarchical-enumerative structure is used mostly in directories, subject gateways and in cataloguing electronic contents. Faceted structure is used in the commercial sites to effectively organise and retrieve the web document through a multidimensional taxonomy. A third approach known as Folksonomy has emerged as a user oriented classification on the Web without maintaining any explicit relationship of classifying document. This study suggests using these three approaches in the appropriate context to ensure the optimum use of corporate information.*

**Keywords**: Classification structures; Web information organisation; Hierarchical-enumerative structure; Faceted structure; Folksonomy

## INTRODUCTION

Although the web information structure is steadily improving, studies show that search and browsing is still the primary usability problem in website design (English, et al., 2003; Sullivan, 2000). Most users face significant problems in website searching either due to their lack of knowledge on how to use the search engines or the unorganised information and navigation structure of the websites. The rapid growth of electronic information in organisations, and the divergent patterns of storage, distribution, and usage of that information (through databases, data warehouses, web servers and ERP packages) have made it difficult to create a usable web interface that can properly articulate the information content of an organisation. This results to the improper use of companies' electronic information because of the imperfect structure and knowledge representation in the Web. One of the main reasons behind the problems, specifically, in the eyes of information manager, is that most websites and search engines do not allow any classification or clustering of metadata keywords, and controlled vocabulary, while many words have multiple meanings depending on the context. Therefore, search engines often provide results that are imprecise and unnecessary. A study on 69 websites (King, 2004) found that the most common usability problem was poorly organised search results, affecting 53 percent of sites studied, followed by poor information architecture, affecting 32 percent of sites. This shows a major challenge for the organisations in efficiently finding content from the Web.

The central research issue of this study is how the use of classification system in Web can reduce the problem of information overload and searching. The application of formal classification schemes to aid retrieval in a network environment, specifically in developing the information architecture on the Internet is comparatively a recent phenomenon. However, the library community, over many years, have been using subject indexing systems (the use of a controlled vocabulary to assign indexing terms to documents) along with the use of hierarchical classification schemes (grouping documents into a hierarchical structure of subject categories) to describe and retrieve document resources. During the first period of the development of networked information services, many specialists especially those from the computer science community, explored the value of library subject description systems for the accomplishments of full-text indexing software. Since the 1990s, the use of classification structures for organising information and browsing architecture of websites for a better navigation has become important to website designers. In addition, the structure of the enumerative classification system and thesaurus used in libraries has encouraged the computer science professionals to develop domain based ontology to enhance searching. Now classification systems are used widely in classifying information products, industry products, in organising web portals, web directories (such as Yahoo! and Google), Intranet content management and in knowledge management.

Nevertheless, there is a wide gap in researching the best possible use of classification system in the electronic environment, which can significantly improve the information structure of the Web and can serve as a retrieval tool. Given the ground, this study examines some of the issues such as the impact and value of using classification structures in the Web, the current practice utilizing the classification systems, and how classification systems aid in information retrieval, searching and browsing. This paper also describes the conceptual understanding of the related terms and finally develops a standard to use classification/taxonomy structures for designing a better web information architecture.

## CONCEPT OF CLASSIFICATION

The term classification is often used interchangeably with the terms cataloguing, categorizing, taxonomy and ontology. These are the ways of organising information into categories according to likeness or differences. Classifications are systems that cluster (group/categorise) entities or concepts based on shared attributes or divide an entity or concept into smaller/ narrower entities/concepts based on some rule of division. It is an act of placing an object or concept into a set or sets of categories based on its properties, and external attributes such as colour or geography with underlying semantic relationship (Rees, 2003). Usually, the relationships expressed in classifications are not essential, but arbitrary. Stefik (1995) states, "To classify something is to identify it as a member of a known class." However the main purpose of classification is to guide users to a body of information.

Many refer to classifications as 'taxonomies' in the context of Knowledge Management (KM) and enterprise information portals. In taxonomies, generally the relationships expressed between concepts are essential. In other words, taxonomies group on the basis of internal properties of the related pieces of information (Rees, 2004). For instance, all books are published items, but not all published items are books. In the relationship, published item is a super-ordinate category and the book is a subcategory. Taxonomies are commonly created from the bottom up from actual content by multidisciplinary teams, and tend to be more concise, reusable, more updatable than classifications (Conway & Sligar, 2002). The main purpose of taxonomies is to utilize the structure of thesauri to guide the user or computer through a body of information, most often hierarchically (Argo, 2004). The two most common relationships in a taxonomy are described by 'is a' and 'part of' relationship.The use of 'is a' relationships creates a 'linear' taxonomy; a tree-like structure that clearly depicts how entities relate to one another. In such a structure, each entity is the 'child' of only one 'parent class' as depicted in Figure 1. If we introduce 'part of' relationships, the structure may turn to complex and more difficult to interpret. The relationship can better be described by an ontology. Ontologies are defined as "an explicit formal specification of a shared conceptualization" (Borst, 1997). As discussed in Studer et al (1998),

- "explicit" means that "the type of concepts used and the constraints on their use are explicitly defined";

- “formal” refers to the fact that “it should be machine readable”;
- “shared” reflects that the knowledge represented in an ontology “captures consensual knowledge, that is, it is not private to some individual, but accepted by a group”;
- “conceptualization” refers to “an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon.

```
                          Class A
                             |
              ------------------------------------------
                   |                           |
              Sub-Class B                 Sub-Class C
                   |                           |
        ---------------------------      --------------------------
          |              |                        |           |
Sub-Sub-Class D   Sub-Sub-Class E   Sub-Sub-Class F  Sub-Sub-Class G
```

Figure 1: Class Subclass Relationship in a Taxonomy

Ontology consists of a set of connected (networked) terms on specific domain knowledge. The terms are the representation of concepts, which allows unambiguous interpretation of terms for use by machines. Ontologies have been developed to solve the problems that arise from using different terminology to refer to the same concept or using the same term to refer to different concepts (Beck and Pinto, 2002). Ontologies can include many sub-class based taxonomies connected together (Edols, 2001). However, it creates a common vocabulary of concepts that can be used consistently for communication about specific domains, and guides the user in becoming proficient in the retrieval and understanding of a particular body of information. An example of sample crop-pest management ontology, which illustrates ‘part of’ relationship, is given in Figure 2. It shows that “Beet Armyworm” is an object within the “insect pest” class. There are several taxonomic superclass-class-subclass relationships, for example “crop” – “agronomic crop” – “soybeans”. “Beat Armyworm” is associated with “Soybeans”, specifically it is an “Insect Pest” of soybeans. Thus the concepts in ontology structure are networked, which can map from one concept to another and help in web searching. To represent ontologies in machine-readable way standard languages such as RDF, DAML/OIL and OWL can be used.

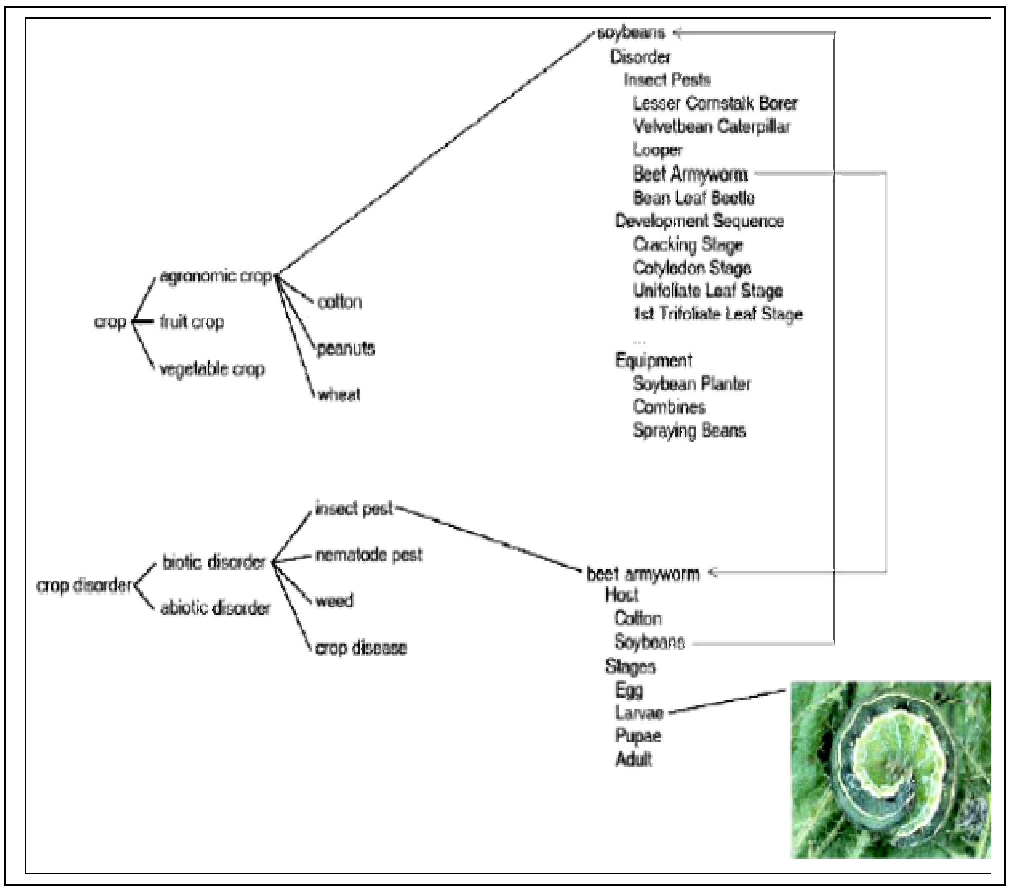


Figure 2: A Sample Ontology for Crop-Pest Management (Beck and Pinto, 2002)

**Types of Classification Structure**

There are four major classification structures, which form the foundation of different classification schemes. They are hierarchies, trees, paradigms and facet analysis (Kwasnik, 1999) and described as follows:

- ***Hierarchies*** are formed by dividing a whole into classes and sub-classes, using specific rules of division and subdivision. Concepts/entities organised into hierarchies follow 'is-a kind of ' relationship. Examples include: Genus-species divisions in plant taxonomy; division of different diseases in Medline/MESH. A key feature of hierarchies is inheritance, where entities in a sub-class share all or most of the attributes of the higher-level class.

- A ***tree*** divides and subdivides its classes based on specific rules for distinction just as in a hierarchy but does not assume the rules of inheritance. In a tree, entities have systematic relationships but not the generic 'is-a kind of' relationship. Examples: Chain of command in the army: generals - colonels - captains - lieutenants - sergeants - privates; Partitive trees for example geographic areas, organs of a body and parts of a vehicle.
- ***Paradigms*** (or matrix) are a classificatory structure in which entities are described by the intersection of two attributes at a time. The resulting matrix reveals the presence or absence and the nature of the entity at the intersection. Example: gender (male, female) and kinship (parent, sibling, parent's sibling, parent's sibling's child) relationship.
- ***Faceted Analysis*** is not a representation structure but an approach to the classification process, which is first stated by SR Ranganathan. Any complex entity could be viewed from a number of perspectives or facets. Faceted approach has been reinterpreted and used in several fields for example computer software (for reuse), patents, books, and art objects ('art and architecture thesaurus').

The use of classification schemes in websites provides logical organisation of information resources and thereby efficient access to web resources. A site that organises its information architecture and knowledge with a classification scheme demonstrates the following advantages over sites which do not (Koch et al, 1997).

- ***Content Browsing***: Browsing is particularly helpful for the inexperienced users who are not familiar with the content of a site, its structure and terminology. The structure of the classification scheme can be used to develop browsing by organising the knowledge materials and display them in different ways as a navigation aid.
- ***Broadening and Narrowing Searches***: Classification schemes are hierarchical and therefore can be used to broaden (for improved recall) or narrow a search when required. Questions can be limited to individual parts of a collection (filtering) and the number of false hits can be reduced for improved precision.
- ***Multilingual Access to a Collection***: Since classification systems often use notations independent from a specific language, indices in different languages can offer multilingual access to the same resources without any further changes to the collection. A searcher could enter search terms in a given language and those terms would then relate to the relevant parts of the classification system (as a switching language).

- ***Context***: Classification scheme gives context to the search terms used. For example, by using an ontology-based search the problem of homonyms (words which have the same form and spelling but a different meaning) can be partly overcome.
- ***Partitioning and Manipulation of Database***: Large classified lists can be divided logically into smaller parts if required. Besides, the use of an agreed classification scheme could enable improved browsing and subject searching across databases.
- ***Machine-Readable Format*****:** Many classification schemes are available in machine-readable form, which also ensures interoperability. For example DDC (Dewey Decimal Classification) is distributed by Machine Access Readable catalogue (MARC)

However there is also some criticism in using classification systems in the Web, described as follows:

- ***Illogical subdivision of classes***: Some popular schemes do not always subdivide classes in a logical manner (Buchanan, 1979; Rowley, 1987). This can make them difficult to use for browsing purposes.
- ***Assimilating new areas of interest****:* Classification schemes, since they are usually updated through formal processes by organised bodies, often reveal difficulty in reacting to new areas of study.
- ***Cost and Complexity****:* Use and maintenance of established classification systems may involve huge cost. Besides, the sophistication of the classification scheme may produce complexity among the user of the information system.

**Current Use of Classification Structure in Web**

Most of the classification structure used in the Web is either hierarchical-enumerative or analytical synthetic (Faceted Approach). Apart from the library classifications used for organising books and other materials in libraries, classifications have been developed and used in variety of contexts. These include various standard industry taxonomies and ontologies, some of which are: Product and Service classification such as UNSPSC (United Nations Standard Product and Services Classification), Industrial Classification such as NAICS (North Americal Industrial Classification System), Plant taxonomies (Biosystematic codes in BIOSIS), software classifications (ACM Computing Classification) and Medical sciences (MeSH). Most of these classification schemes in different disciplines follow more or less the same kind of structure: clustering the entities based on their properties; structuring them in some level of hierarchy; and use symbols to maintain the hierarchy. In the web environment, the structures of these classification systems are used as an effective tool of information retrieval and a cheaper way of improving search and navigation.

**a) Hierarchical-enumerative classification structure**

This structure refers to a taxonomic top-down scheme, in which knowledge is divided into progressively narrower and more specific categories (a hierarchy), where one object is typically located to one category. Enumerative classification assigns names to every subject and enumerates them, typically in a systematic order. There are different 'universal classification scheme' such as Subject specific scheme(NAICS and UNSPSC), and Home Grown Scheme (Ontology of Yahoo search service) that follow hierarchical enumerative structure. The term 'universal' schemes is used for schemes which aim to include all subjects, are global geographically, and multilingual in scope. The most widely used universal classification schemes are Dewey Decimal Classification (DDC), the Universal Decimal Classification (UDC) and the classification scheme devised by the Library of Congress (LCC). In DDC and UDC, the whole universe of knowledge is firstly divided into ten major classes, then each class is divided again into ten divisions forming a hundred divisions which are again divided into thousands sub-divisions. Thus the knowledge structure of DDC and UDC cover the whole field of knowledge and organised by hierarchies. In contrast, LCC is an enumerative system built on 21 major classes, each class being given an arbitrary capital letter between A-Z, with 5 exceptions (IOWXY). LCC notations are composed of repeated letters and numbers. Capital letters are used for main class and subclass notations while Arabic numerals are used for subdivisions further down the hierarchies (for example HB1-3840 Economic theory).

The DDC and UDC hierarchy in organising browsing sections is mostly used by websites, which provide cataloguing resource services, Internet subject gateways, and directory services. Examples are BUBL (bubl.ac.uk/), NISS Directory of Networked Resources (www.lub.lu.se/tk/demos/ao/niss.htm), GERHARD in the Deutsche Forchungsgemeinschaft (DFG) funded project GERHARD (German Harvest Automated Retrieval and Directory), OMNI (Organising Medical Networked Information) and SOSIG (The Social Science Information Gateway – www.sosig.as.uk). Figure 3 and 4 illustrate the use of DDC and UDC in the browsing section of BUBL and SOSIG respectively.

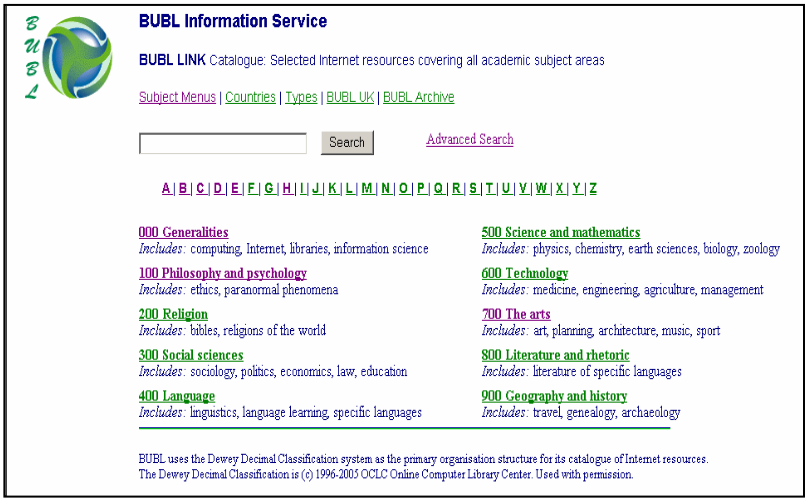


Figure 3: DDC Use in BUBL Information Service Browsing (bubl.ac.uk)

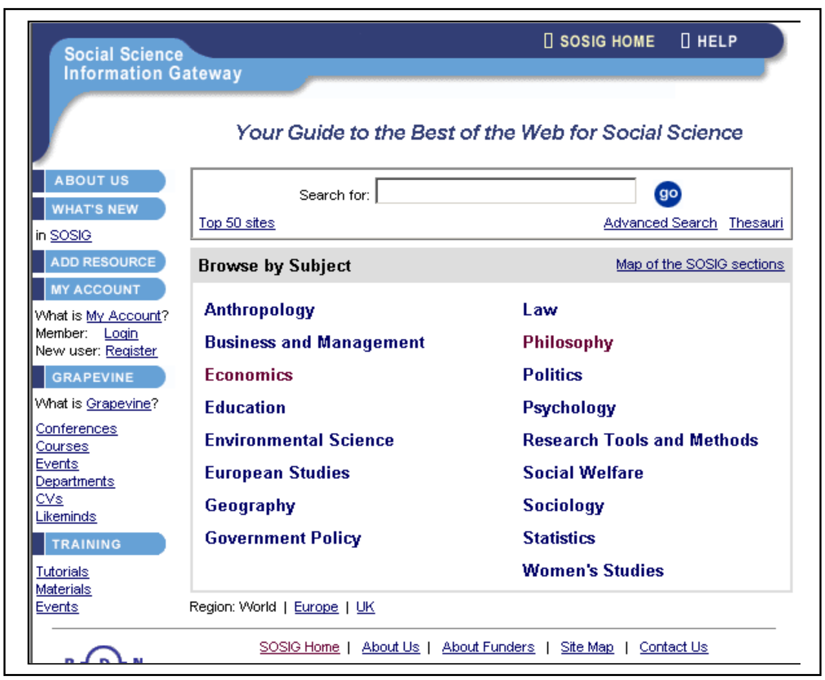


Figure 4: Use of UDC in the Browsing Section of SOSIG (www.sosig.ac.uk)

There are several information services which use LCC for classification of resources. The WWW Virtual Library (vlib.org) and *CyberStacks (*www.public.istate.edu/ ~CYBERSTACKS/) are two fine examples. In *CyberStacks*, resources are organised under one or more relevant LC class numbers and an associated publication format and subject description (Figure 5).

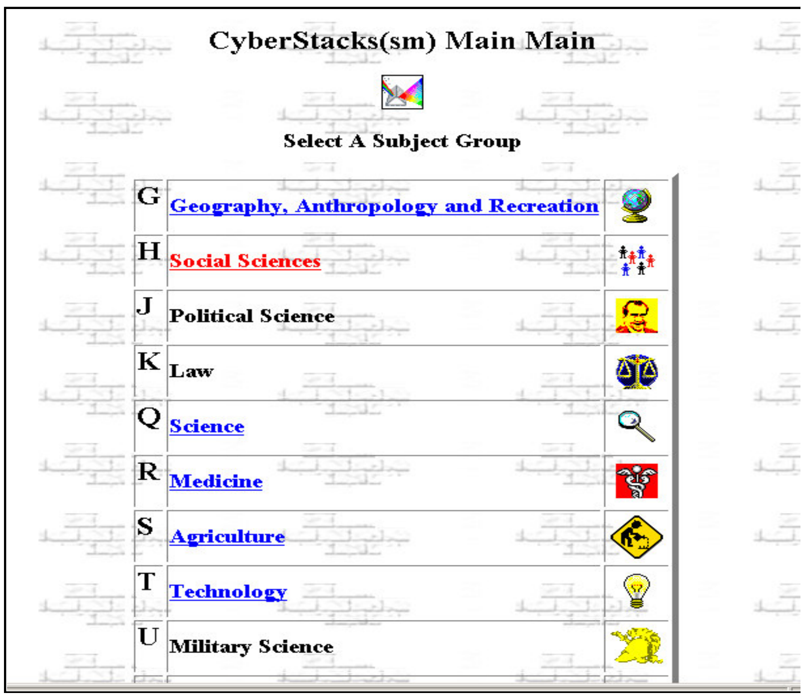


Figure 5: LCC Use in CYBERSTACKS
(www.public.istate.edu/~CYBERSTACKS/)

Among the industrial taxonomy and ontologies**,** the UNSPSC classification structure provides an open global multi-sector standard coding system for accurate and efficient classification of products and services for e-business. UNSPSC is a hierarchical classification, having five levels which help to search products more precisely and allow company-wide visibility of spend analysis. It classifies more than 8,000 products and services around the world. UNSPSC is an eight-digit code and contains two-character numerical value and a textual description in each of the level, illustrated as follows:

- XX: *Segment*:
  The logical aggregation of families for analytical purposes
  - XX: Family:
    A commonly recognized group of inter-related commodity categories
    - XX: Class:
      A group of commodities sharing a common use or function
      - XX: Commodity:
        A group of substitutable products or services
        - XX: Business Function:
          The function performed by an organisation in support of the commodity

The following hierarchy allows looking for lead refills to use the higher level terms to narrow search to the relevant domain that will most likely lead to the desired item. For example if we want to classify the good "Lead refills", UNSPSC classification number will be 44-12-19-02 *(*Figure 6)

| **Hierarchy** | **Category Number and Name** |
|---|---|
| *Segment* | **44** Office Equipment, Accessories and Supplies |
| | **10** Office Machines and their supplies and accessories |
| | **11** Office and desk accessories |
| *Family* | **12** Office supplies |
| | **15** Mailing supplies |
| | **16** Office supplies |
| | **17** Writing instruments |
| | **18** Correction media |
| *Class* | **19** Ink and lead refills |
| | **01** India ink |
| | **02** Lead refills |
| *Commodity* | **03** Pen refills |

Figure 6: UNSPSC Classification Hierarchy (Granada Research, 1999)

**(b) Faceted classification Structure**

Faceted classification is a bottom-up approach where each object (such as document or image) is tagged with certain set of attributes and values, called facets (orthogonal sets of categories), and the organisation of these objects emerges how a user may choose to access them. It is a method of multidimensional description and arrangement of information resources by their subject, attributes, or "aboutness"(Louie et al, 2003). It addresses the fact that users may look for a document resource from any number of angles corresponding to its rich attributes and multidimensionality. By encapsulating those distinct characteristics or dimensions as "facets", faceted classification system may provide multiple facets, or main categories of information object to identify a resource with greater flexibility in access. For example, in a shopping centre a user may look for items by product name, or brand name, or he may be interested in particular ingredients of the product and their price, or any of the combination of these information. Therefore, facets of a shopping centre can be "Product Name', 'Brand Name' Ingredients, and 'Price' under which particular information such as 'T-shirt', 'Nike', 'Cotton', and '$20' can be arranged. Each of the facets contain isolates (content oriented metadata/facet element), or subcategories arranged in a hierarchy, which is used to classify, organise, and access the resource by browsing or querying in the Web.

A more simplified example can be given from Broughton (2004) in Table 1 to organise a collection of socks by representing all the characteristics of a particular sock in five facets: colour, pattern, material, function, and length. If applied to the Web, the structure is suitable for browsing and searching by combining any of the given attributes under the facets, such as a sock can be black plain cotton sock or blue striped silk sock for hiking. The combination of the item in searching can be made by providing an interface with multiple selection box or checkbox

Table 1: Classification of Object Attributes by Facets (Broughton, 2004)

| **Color** | **Pattern** | **Material** | **Function** | **Length** |
|---|---|---|---|---|
| Black | Plain | Wool | Work | Ankle |
| Grey | Striped | Polyester | Evening | Calf |
| Brown | Spotted | Cotton | Football | Knee |
| Green | Checkered | Silk | Hiking | |
| Blue | Novelty | Nylon | Protective | |

Ranganathan,(1960) was the first to introduce the word "facet" in library and information science, and the first to develop the theory of facet analysis in his Colon Classification (CC) scheme. He describes five fundamental facets, known as PMEST in CC:

- Personality (Something in question or what the object is, e.g. a person, an animal or event)
- Matter (what something is made of, physical matter)
- Energy (how something changes, is processed, evolves, or action)
- Space (where something is)
- Time (when it happens)

For example, if a document discusses on "the design of wooden furniture in 18th century America." the facets would be as follows: Personality—furniture, Matter - wood, Energy - design, Space - America, Time - 18th century. The resource is described by aggregating and combining the information element under each facet. "Wood" is a piece of that description which covers an area that none of the other pieces cover. Thus, the classification strength comes through combining the pieces together to form the whole (Taylor, 1999). The logical and predictable structure of faceted system makes it compatible with the requirements of electronic environment in a way that enumerative and pre-coordinated systems are not and can serve as the basis of all methods of information retrieval (Broughton, 2006). The system is compatible to develop graphical interfaces by exposing the domain concepts and objects attributes in the faceted taxonomies. Specifically, faceted classification can be used to:

- Directly find information by drilling down taxonomy of concepts.
- Refining or broadening the search within particular topics and with several different dimensions simultaneously.
- Group search results within browsable topics.
- Serve as a good basis for thesaurus creation

An example of how Ranganathan's Colon Classification impact in the organisation of web information according to several facets (personality, matter, energy, space and time) simultaneously, and how the shape of knowledge can more faithfully be rendered in the shape of the collection of these sites can be seen in Epicurious (www.epicurious.com). It is a website on cooking and basically provides recipe information. Recipes are examples of information for which hierarchical faceted metadata is used to provide search interface. Epicurious provides three interfaces for searching: a basic keyword search, an enhanced search to expose the faceted metadata search in the form of checkbox and a browser interface based on metadata facets to allow navigating through the collection. In the browse interface following facets are used which can be treated in view of PMEST.

- Main Ingredients (e.g. Beans) ……. Similar to Matter
- Cuisine, i.e. ethnic origin (e g. African)……. Similar to Space
- Special consideration (e.g. Kid Friendly)
- Preparation Method (e.g. Bake) …………… Similar to Energy
- Season/ Occasion (e.g. Christmas, Easter) Similar to time
- Course/Dish (e.g. Bread) …………………… Similar to personality

Another example is Flamenco (FLexible Access to MEtadata in NOvel Combinations, http://orange.sims.berkeley.edu/cgi-bin/flamenco), an Image

Browser that provides access to an architectural image collection consisting of 36,000 images. The interface uses hierarchical faceted metadata in a manner that allows users to both refine and expand the current query, while maintaining a consistent representation of the collection's structure. The interface framework has the primary design goal of allowing users to move through large information spaces in a flexible manner without feeling lost. The site uses the facets some of which are as follows:

- Media (e.g. Painting, Photographs)…………………… Personality
- Date (e.g. 19th century) ………………………………... Time
- Location (E.g. Asia) ……………………………………. Space
- Occupation (e.g. Entertainer) …………………………… Energy
- Animals and Plants (Birds, Flowers)…………………… Personality
- Built Places ( e.g. Bridge) ……………………………… Matter

Thus it can be seen that faceted classification in the Web can serve as a better information retrieval tool, which provides a more legible and understandable knowledge structure to develop an information architecture that can closely imitate the user need in the web. Faceted classification with multidimensional organisation of content can fit very well in the web information architecture for the ease of navigation and searching. As mentioned by Kwasnik (1999) and quoted by Denton (2003) "The notion of facets rests on the belief that there is more than one way to view the world ... Facets make a multi-dimensional organisational scheme, and web browsers are an easy and familiar tool for navigating many dimensions. All of the benefits of faceted classifications can be realized and provided on the web."

**(c) User-Oriented Classification**

An informal classification called Folksonomies recently comes into the practice to provide Web-specific classification issues. Folksonomy is the practice of collaborative categorization of web resources using freely chosen tags or keywords, that is, storing and retrieving the web content by using one's own descriptors. A folksonomy is a user-generated classification, emerging through bottom-up consensus (Wright, 2005). An important aspect of a folksonomy is that it comprises terms in a flat namespace – that is, there is no hierarchy, and no directly specified parent-child or sibling relationships between these terms. It generates "related" tags automatically, which cluster tags based on common URLs. This is unlike formal taxonomies and classification schemes where there are "multiple kind of explicit relationships between terms" (Mathes, 2004). Folksonomies require people to associate keywords with content and generate a list of popular keywords by tagging all the related documents that other users store in a particular website. Using popular keywords gives the users the "reward of visibility, to see one's own content gravitate in evidence in the system" (Emanuele, 2005), and folksonomies lead to a shared collection of web documents through user's choice of classification.

In contrast to formal classification methods, this phenomenon typically arises in non-hierarchical communities, such as public websites and weblogs, as opposed to multi-level teams. Since the organisers of the information are usually its primary users, folksonomy produces results that reflect more accurately the population's conceptual model of the information. The practice is also known as free tagging, open tagging, ethnoclassification, faceted hierarchy, or distributed classification. Figure 7 shows the Del.icio.us site where its user community collaboratively organise a shared set of resources by assigning classifiers or tags.

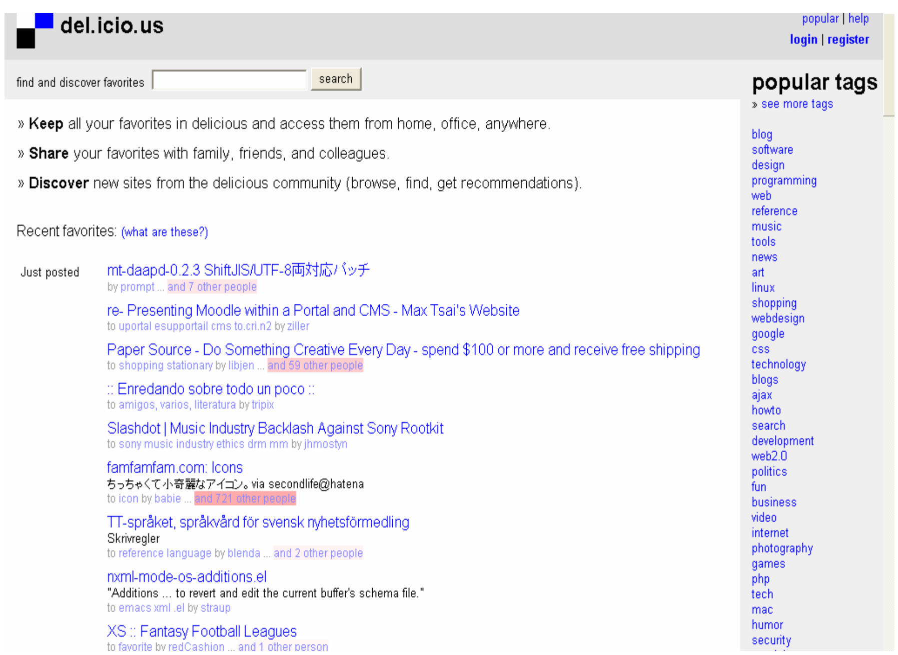


Figure 7: The Delicious Page Showing the Use of Folksonomy (http://del.icio.us/)

**Recommendation of Using Classification Structure in Web**

The implementation of classification structure in Web information service should depend upon the scope of its service. The audience-based classification schemes make sense when the informational domain caters to clearly delineated audiences (Rosenfeld & Morville, 2002) for which document resources can be arranged by topic, task, alphabet, chronology, geography, or a hybrid scheme. Normally, for the cataloguing and subject services, hierarchical-enumerative structure from a well-known classification scheme can be chosen to develop the web interfaces. By harmonizing the corporate need and users' convenience in searching and browsing, a specific universal or a subject specific scheme, or even a home-grown scheme can be used to create the information structure. But often a simple hierarchical structure when applied to web design treats web knowledge as an integrated whole, which is divided and subdivided into specific group in a tree-like structure. Therefore, it is not capable of expressing the multi-dimensional properties and relationships of digital objects where no specific ordering is dominant. To alleviate the limitation of hierarchical structure, specifically in designing an Internet selling or commercial websites, faceted classification system can be used to classify and organise the web document in multiple independent taxonomies to provide multiple ways of locating and accessing to a document. On the other hand the approach of folksonomy as user generated classification is not an alternative to the formal classification systems but can be a powerful and innovative tool if apply only under the right circumstances, most often in developing community sites, along with other classification structures.

In developing the classification taxonomy in any of the hierarchical or faceted approach, three key factors such as context, users, and information content of the system need to be focused to make an appropriate articulation of information architecture. Context is the organisation environment and the coverage of information based on which taxonomy will be developed. This considers the organisational objectives, geographical coverage, types of applications where taxonomy will be used, corporate culture, and artifacts within the organisation or across the organisation boundary. Users refer to the target audience for the taxonomy, user profiles, and user characteristics in terms of information usage patterns. Content is the type of information that will be covered by the taxonomy or that the taxonomy will be built upon. These three factors provide a "trinity compass" in the road of developing classification structure (Morrison, 2003):

Following the trinity compass, the taxonomy development process typically involves the steps of forming a team with people of different expert group, defining the scope in the context of corporate goals, content and user interest, and then creating the taxonomy either manually or automatically or a using combination of both, which follows implementation by tagging the content. Figure 8 illustrates the process.

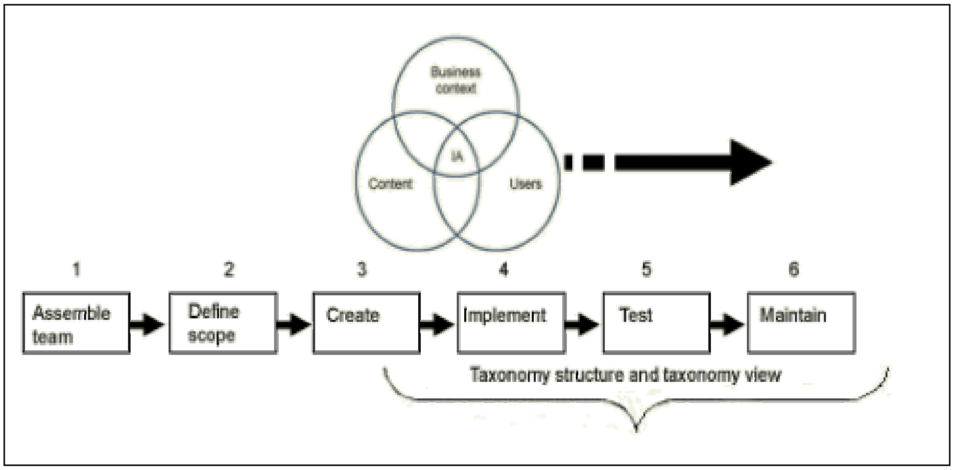


Figure 8: Taxonomy Development Process (Morrison, 2003)

There are some rules that can be followed to design a good taxonomy. First, it would be fine if the hierarchical categories are mutually exclusive to provide balance between exclusivity and inclusivity within the single organisational scheme. Mutually exclusive means each of the categories describes one single aspect of information object and should not contain the element of other categories. Second, balance should be provided between breadth and depth in the taxonomy. Breadth refers to the number of options at each level of the hierarchy. Depth refers to the number of levels in the hierarchy. If a hierarchy is too narrow and deep, users have to go through many levels to find what they are looking for, which is often frustrating**.** Figure 9 illustrates the narrow-and-deep hierarchy in which users are faced with six steps to reach the deepest content. In the broad-and-shallow hierarchy, users must choose from ten categories to reach ten content items. In designing breadth, we should also take into consideration the cognitive limits of the human mind, as users will have to many options if a hierarchy is too broad and shallow (Rosenfeld and Morville, 2002).

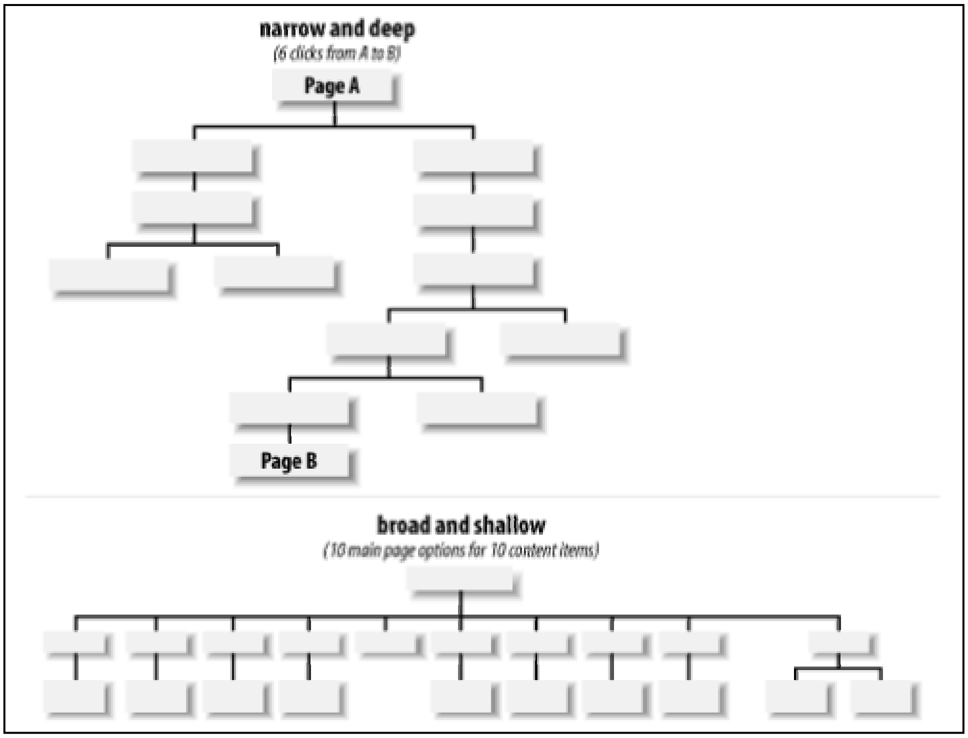


Figure 9: Balancing Depth and Breadth of Taxonomy
(Rosenfeld and Morville, 2002)

## CONCLUSION

Classification systems involve structuring knowledge in different ways, although mostly hierarchical is in a tree structure, and the classification taxonomies emerge in relation to the need of different domain. This paper defines the concepts related to the recent development in classification and gives an overview with case studies of using classification structure in information description and discovery in the web. The study found that hierarchical structure is used mostly in directories, subject gateways and in cataloguing the e-content while faceted structure is used in the commercial sites to effectively organise and retrieve the web document through a multidimensional hierarchy. A third approach known as Folksonomy has emerged as a user oriented classification in Web without maintaining any explicit relationship of classifying document. The authors suggest using this approach as a supplement of formal classification by integrating it into community websites. The right use of classification structure in the web information architecture can provide a user-friendly interface that can serve as a cheaper method of information retrieval and can ensure the optimum use of corporate information.

## REFERENCES


Argo, G. 2004. *Classification and taxonomies*. School of Information, University of Texas at Austin. Available at: http://www.ischool.utexas.edu/~i385e/resources.html

Beck, H and Pinto, H. S. 2002. *Overview of approach, methodologies, standards, and tools for ontologies.* Agricultural and Biological Engineering Department of Information Systems, University of Florida, The Agricultural Ontology Service, UN FAO.

Borst, P. 1997. *Construction of engineering ontologies for knowledge sharing and reuse.* Ph.D. Dissertation. Tweente University.

Broughton, V. 2004. Faceted classification. *Essential Classification*, Facet, London, pp. 257-83.

Broughton, V. 2006. The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings: New Information Perspectives*, Vol. 58, no. 1/2 : 49-72.

Buchanan, B. 1979. *Theory of library classification*. London: Bingley.

Conway, S., and Sligar, C. 2002. *Unlocking knowledge assets*. Redmond, Washington: Microsoft Press.

Denton, W. 2003. *How to make a faceted classification and put It on the web*. Available at http://www.miskatonic.org/library/facet-web-howto.html

Edols, L. 2001. Taxonomies are what? *FreePint, 97*, 9-11. Available at http://www.freepint.com/issues/041001.pdf

Emanuele, Q. 2005. Folksonomies: power to the people. Paper presented at *The ISKO Italy-UniMIB Meeting* : Milan, Available at http://www.iskoi.org/doc/folksonomies.htm#limits

English, J., Hearst, M., Sinha, R., Swearington, K., and Yee, P. 2003. *Examining the usability of web site search.* Available at http://bailando.sims.berkeley.edu/papers/epicurious-study.pdf

Granada Research.1998 *Using the UNSPSC: Why coding and classifying products is critical to success in electronic commerce*. Available at: http: www.unspsc.org/AdminFolder/ Documents/UNSPSC_White_Paper.doc

King, M. 2001. *A tangled web.* Available at http://www.etailersdigest.com/resources/Specials/Tangled_Web.htm

Koch, T., Day, M., Brümmer, A., Hiom, D., Peereboom, M., Poulter, A., and Worsfold, E. 1997. *The role of classification schemes in Internet resource description and discovery,* Work Package 3 of Telematics for Research project DESIRE (Development of a European service for information on research and education). Available at: http://www.ukoln.ac.uk/metadata/desire/classification/

Kwasnik, B. H.1999. The role of classification in knowledge representation and discovery. *Library Trends,* Vol. 48, no. 1: 22-47.

Louie, A.J., Washington, W. and Maddox, E.L. 2003. *Using faceted classification to provide structure for information architecture*. Available at http://depts.washington.edu/pettt/presentations/conf_2003/IASummit.pdf

Mathes, A. 2004. Folksonomies: Cooperative classification and communication through shared metadata. *Computer Mediated Communication - LIS590CMC*, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign. Available at: http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Morrison, J. 2003.*Understanding information taxonomy helps build better apps.* Available at: http://www.steptwo.com.au/columntwo/archives/cat_information_architecture.html.

Ranganathan, S.R. 1960. *Colon Classification Basic Classification*. 6th ed. Asia Publishing House: New York, NY.

Rees, V. 2003. Clarity in the usage of the terms ontology, taxonomy and classification. *CIB73 2003* Available at http://vanrees.org/research/papers/cib2003.pdf

Rosenfeld, L., and Morville, P. 2002. *Information architecture for the World Wide Web.* Cambridge: Sebastopol, CA: O'Reilly.

Rowley, J.E. 1987. *Organising knowledge: an introduction to information retrieval*, Aldershot: Gower.

Stefik, M. 1995. *Introduction to knowledge systems*. San Francisco: Morgan Kaufmann.

Studer, R., Richard, B. and Dieter F. 1998. Knowledge Engineering: Principles and Methods. *Data and Knowledge Engineering*, Vol. 25, no. 1-2: 161-197.

Sullivan, D. 2000. *Npd search and portal site study*. Available at: http://searchenginewatch.internet.com/reports/npd.html,

Taylor, A. G. 1999. *The Organization of Information*. Libraries Unlimited, Englewood.

Wright, A. 2005. *Folksonomy*. Available at http://www.agwright.com/blog/archives/000900.html