Erik Mitchell

Metadata Standards and Web Services in Libraries, Archives, and Museums

An Active Learning Resource

Metadata Standards and Web Services in Libraries, Archives, and Museums

METADATA STANDARDS AND WEB SERVICES IN LIBRARIES, ARCHIVES, AND MUSEUMS

An Active Learning Resource

Erik Mitchell

An Imprint of ABC-CLIO, LLC
Santa Barbara, California • Denver, Colorado

Library of Congress Cataloging-in-Publication Data

Mitchell, Erik T., 1972–
Metadata standards and Web services in libraries, archives, and museums : an active learning resource / Erik Mitchell.
pages cm
Includes bibliographical references and index.
ISBN 978-1-61069-449-0 (paperback) — ISBN 978-1-61069-450-6 (ebook) 1. Metadata. 2. Information organization. 3. Web services—Library applications. 4. Museums—Information technology. 5. Archives—Information technology. I. Title.
Z666.7.M58 2015
025.3—dc23 2015015545

ISBN: 978-1-61069-449-0
EISBN: 978-1-61069-450-6

19 18 17 16 15 1 2 3 4 5

This book is also available on the World Wide Web as an eBook.
Visit www.abc-clio.com for details.

Libraries Unlimited
An Imprint of ABC-CLIO, LLC

ABC-CLIO, LLC
130 Cremona Drive, P.O. Box 1911
Santa Barbara, California 93116-1911

This book is printed on acid-free paper ♾

Manufactured in the United States of America

CONTENTS

Chapter 1

INTRODUCTION TO THE WORLD OF DIGITAL INFORMATION ORGANIZATION

WHAT THIS BOOK IS ABOUT

The field of library and information science (LIS) is concerned with the roles, uses, and impact of information in people, across communities and throughout history. Librarianship and other information careers have built a shared understanding of what information is, how it is created and used, and what impact it can have on individuals and communities. In addition, for much of the 20th century, libraries, archives, museums (LAMs) and similar cultural heritage institutions have shared a consistent view of information management, albeit not necessarily consistent systems and standards. Nevertheless, this shared vision has enabled LAM institutions to move quickly into new fields such as information retrieval, information systems analysis, and data curation as the practice of information creation, management, and use changed. Along the way, LAM institutions have had to redefine their information systems to meet shifting technologies and user needs and have had to develop new systems and standards to accommodate digital information resources and networked information communities. To cite a familiar example, library catalogs, initially developed to help users find and gain access to resources, have been re-invented multiple times as information technology changed. Published first on scrolls, later in books, and more recently on catalog cards, in microfiche, and ultimately as web-based applications, library catalogs have sought to provide an important service to library users. In the last decade, however, the needs of information communities have changed as have the formats of the information materials they work with. The development of the Internet into a user-initiated publishing platform and a network of resources turned search engines by Yahoo and Google into world-wide catalogs of information that allowed users to find information in this *de facto* virtual library created by what is generally called the World Wide Web, or simply *the web*. The evolution of the web and the information

services around it meant that issues that drove the creation of LAM catalogs, such as metadata creation and management, user-focused information needs analysis, and the creation and management of digital documents, became important areas of scholarship and development outside of the LIS domain. It also meant that traditional approaches to description needed to be reconsidered for increasingly complex resources.

The growth of prominence of both information and information technology in our daily lives has also meant that librarians, archivists and other information professionals need to be aware of the digital literacies that support information creation, management, and use. In some cases this means being able to design, build, manage, and facilitate use of digital information tools and resources. Every day we use websites that effortlessly allow us to search, browse, select, and use information. We most likely all have our preferred web search engine that prioritizes results for us and may even show us selected e-mails or file results from our cloud-stored personal documents. The simplicity of these sites masks the complexity underneath them. Designing effective information services involves a complex understanding of user needs, technology tools, document and metadata standards, and modes of information interaction. It means understanding the relationship between web design, information organization, and information technology issues. Underneath each of these services there is a world of technology and information-seeking/interaction theory that guides how we interact with the service as well as allowing us understand what information the service can make available to us. These interactions are supported by information systems that store documents, create and manage representations of these documents for search and retrieval, and provide discovery and interaction methods that make the service usable. Information system designers should design systems by thinking about potential users who have information needs, questions, and abilities that guide their use of the system. Between these systems that store and serve resources and a user's information need we have an interaction guided by information-seeking behaviors, human–computer interactions, and personal information-management practices. Many of these systems and services run on computers that are designed to serve thousands or hundreds of thousands of users at the same time; however, the tools and technologies are not so complex that they cannot run on almost any laptop or desktop computer. These two elements combined, the software and the hardware, could be considered to be an "information system."

This book explores the world of information organization from the perspective of digital information service design, seeking ways to understand the shift of information management activities in an expanding universe of digital information. In order to do this it combines technical and conceptual content to help the reader better understand what information organization tasks are and how information systems affect how we use information. The book will focus upon the theories and mechanics of information creation, its use, and reuse in a digital age. The goal is to guide the reader through the domains of information organization, information technology, and information interaction and to explore these domains as the three building blocks of information systems and services. In exploring information

system design, this book explores in some detail the connection between metadata and web services, demonstrating how metadata and web services are created, used, and reused.

METADATA AND ITS ROLES IN EVERYDAY LIFE

The simplest definition of *metadata*, and perhaps the easiest to grasp conceptually, is "data about data." Yet, that simple definition fails to reveal what is, in fact, a very complicated and sometimes controversial area of study and practice. Metadata is created every time we add context to information or data, and those who create metadata are often unaware that they are creating it. In the days of the card catalog metadata was something that was created from the title page of a book and captured on a 3x5 card. In our increasingly digital society metadata is generated with every phone call, with every social media post, with every saved document on your computer. The word metadata famously surfaced in 2013 when news of the National Security Agency's collection of information about every phone call made in the United States emerged. It is likely that before the president of the United States gave details about the gathering of metadata for these phone calls people would not have even considered that the simple act of pushing that "send" button on their mobile phone leaves a record of the number called, the number calling, the duration of the call, and the location where the call was made. Perhaps more importantly, the term *metadata* likely would not have entered the popular vernacular.[1]

In the digital world, metadata surrounds every document, every interaction, and every communication. For example, a modern digital camera will attach metadata to the image file it produces using the method provided by the image format's standard. Such information can include a timestamp, geolocation information if the camera has GPS support, image compression technique used, F-stop setting, aperture setting, and whether the camera's flash was used.[2] If a person is somewhat old fashioned and buys music CDs instead of downloading them from Amazon or Apple's iTunes Store, the process of *ripping* the CD—that is, reading each audio track and compressing it to an MP3, AAC, or Og Vorbis format—necessarily requires the creation and attachment of metadata to the new audio file. The metadata is most often obtained by the querying of Gracenote's Compact Disk Database (CDDB) by the ripping program.[3] When the ripping program is, for example, Apple's iTunes, the compressed tracks are stored in the user's iTunes library and the metadata is stored in the library's database. However, the metadata is also stored in the file itself, making it accessible to a portable music player. The act of creating a new written document with office software such as Microsoft Office also offers an opportunity for inserting metadata about the document—information such as the author's name, contact information, document title, subject, keywords, and comments.

Although metadata has become ubiquitous and is often created using automated techniques, the use of metadata creation remain the same as they were in a pre-digital age. For the university student of twenty years ago, conducting

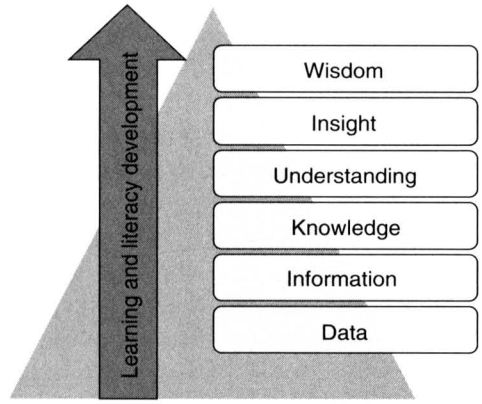
research would require searching the library's catalog as well as journal indexes—both collections of metadata, in this case, data about books and journals in the stacks and the information within. The catalogs and indexes are metadata about the holdings of a library's traditional printed materials and the articles contained within the journals in the stacks. Indeed, a library catalog may be the first example of the creation of metadata, produced by the librarian of the first Sumerian clay tablet archive, perhaps nothing more than a list of holdings and the shelf location of each tablet. University students of today have a vastly wider range of information resources at their disposal, yet the problem remains the same: What information is available, where is it located, and what sources of metadata are available to be searched in order to find it? And, of course, when we use a computer to access a file on some sort of storage device such as a hard disk or a thumb drive, we are using metadata in the form of the directory path and file name so that the program, through the underlying operating system, can determine the file's location on that device and open it. One common thread running through the above scenarios is the need for standards.

WHAT IS INFORMATION?

Information is both a very concrete concept—for example, the contents of a printed book—and an ephemeral concept, such as the process of communicating to others, your internal understanding of data, and your memory of what others said to you. Information theory has been a central part of the work of 20th-century information theorists[4] and is a common thread through many works on the evolution of computers and digital information environments. Several detailed histories of information and information technology discuss the roles of key individuals, technologies, and movements in the evolution of information across human history and consider the thread of information as it transforms societies.[5] An overriding theme in these histories is an interest in how *information objects*—including books, scrolls, web pages, and other forms of physical and digital documents—are created, organized, and used. Each of these histories also explores other types of information, including that which forms the foundation of internalized knowledge—information that is collectively shared and understood in communities, and information as an element of a cultural value system. Wright,[6] for example, suggests that documents need not be understood to have impact in communities. Legal documents, for example, are rarely fully understood outside the small communities that create and make use of them and generally contain a jargon not often used by ordinary people; yet these documents and the information contained within them govern interpersonal relationships, business relationships, and have real-world legal outcomes.

The field of information science views the abstract concept of information through a multifaceted lens. Michael Buckland's seminal work exploring the facets of information[7] studied information in three forms: objects, memory, and communication. Buckland's work in this area dates back to the 1990s, when the world

Figure 1.1.
The Data-to-Wisdom Hierarchy



of information was much different but the underlying theory is as relevant in the digital world as it was when applied to physical information objects.

While Buckland's article views information through a form and formats lens, another popular view of information is as a building block in the continuum from data to wisdom. Figure 1.1 shows an adapted view of this hierarchy that is popularly represented in education literature. This view of information places "information" as a stepping-stone on a hierarchy of increasingly contextualized information. From an information science perspective, knowledge, insight, and wisdom are derivative products of the state of being informed, and data is simply information without context, for example, using a number without specifying any units.

The Data-to-Wisdom Hierarchy

In the education field, questions about the origin and nature of information focus on differentiating information from data, knowledge, and wisdom, and on aligning the concept of learning with different cognitive outcomes (e.g., remembering, understanding, analysis). This hierarchy was first described by Benjamin Bloom as *Bloom's Taxonomy*[8] and has enjoyed a number of updates to reflect current thinking in educational practice. For example, the initial Bloom's taxonomy viewed evaluation as the highest level of learning and differentiated synthesis from analysis. In updated versions, the highest level of learning is represented as "creating," with "evaluating" serving the penultimate role.[9]

Bloom's Taxonomy can be a useful tool in evaluating user information engagement and learning. In addition to defining cognitive approach to understanding the hierarchy of information engagement (i.e., to remember and/or understand it), information use (i.e., to apply, analyze, and/or evaluate it), and information creation, Bloom's taxonomy discusses affective (e.g., emotional) and psychomotor

(e.g., skills and abilities) characteristics of individuals. These areas map to similar focus areas in the information community, such as Kuhlthau's exploration of the connection between information-seeking activities and affective impact. Her Information Seeking Process (ISP) model is built around the cognitive and affective experience of uncertainty and the effect of increased confidence on the information seeker as the search and the collection of wanted information progress.[10] Kuhlthau and Bloom's works are very much directed at exploring Buckland's cognitive and communicative forms of information. Both Bloom's Taxonomy, which puts structure around the amorphous cognitive state of being informed, and Kuhlthau's synthesis of cognitive, affective, and communicative roles into a holistic view of information seeking help sequence the activities of becoming informed.

Information systems play an important role in these interactions, often serving as a way to search for and filter through information, as a structure to put around information (e.g., a classification system), or a storage mechanism for keeping track of complicated data. In fact, information objects have a strong symbiotic relationship with people and communities of practice. An information object's meaning is reflected in the language used to create it, in the structural cues and prompts used to represent key information (e.g., title, author, tables, pictures, graphs), and in the organizational concepts used to deliver information. These structures serve as a shared, common framework through which information is shared and preserved, elevating the details of the document from data, to information, to wisdom.

THE INFORMATION LIFECYCLE

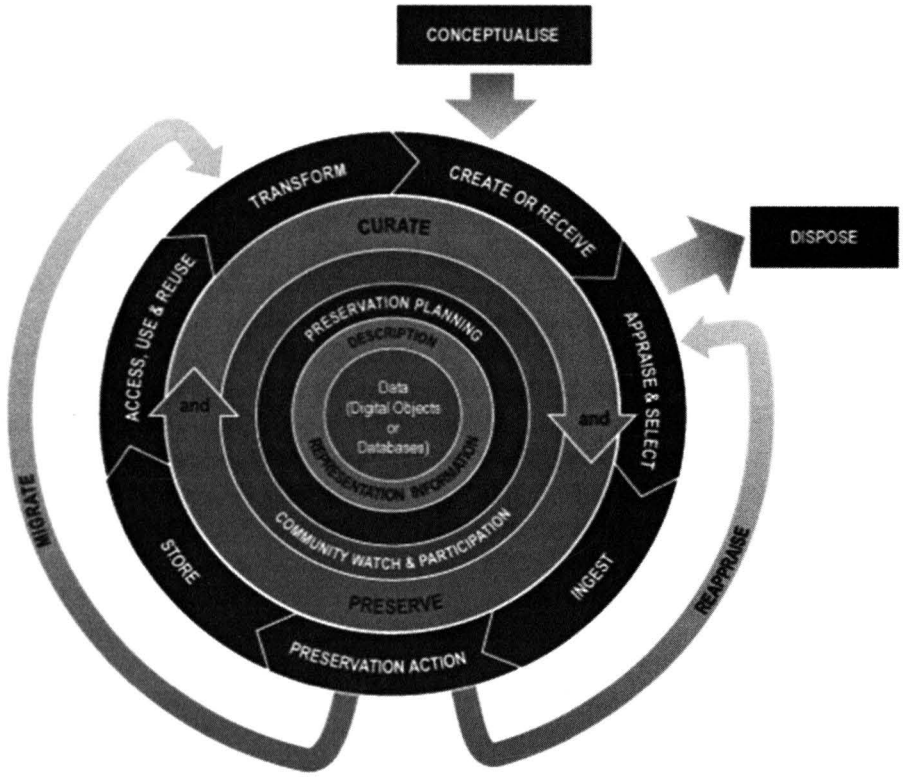
When information is recorded and becomes an object to be managed, it is important to consider the steps through which this resource progresses. In LIS this model is often referred to as an information lifecycle. The development and refinement of these models has been a subject of great interest and study. As a result, several lifecycles have been identified, including information creation lifecycles, curation lifecycles, and even information-seeking lifecycles. The lifecycle model fits the concept of information in part because of the connection to Buckland's views of information being a thing, a process, and knowledge, which agree with the notion that information, is created, managed, accessed, and preserved.

Lifecycle models describe the movement of information through a process (as a thing), emphasize the need to support dissemination and use (again, as a process), and result in the creation of new information about what the next steps users of the information will take as part of their information-seeking activity (as knowledge).

While each of these lifecycle models features different activities, they all share a number of commonalities in that each of these models tends to follow the "plan —> acquire —> manage —> provide access —> analysis" approach.[11]

Each of these activities ought not to be thought of as being performed sequentially for a given object or collection but rather as a process of its own that may have multiple step dependencies and may be active at different stages during the information object management life cycle. For example, it is reasonable to expect that

Figure 1.2.
Digital Curation Centre Curation Lifecycle (http://www.dcc.ac.uk)



data gathering and analysis related to the information object will occur throughout the object's lifecycle.

In addition, it should not be assumed that activities such as discovery, storage, and preservation are bounded or one-time events. Each of these activities may involve one or more information systems, will likely involve multiple information organization activities and standards, and will likely require analysis of different users' information needs and their respective literacies.

While each of these models treats information organizational activities as part of a much larger system, the actual processes of organization, description, and metadata creation have their own processes and cycles. For example, the preservation processes in these models often follow the Open Archival Information System (OAIS) model.[12] Each step in a lifecycle model has corresponding standards relating to the digital documents, document management, and document use. Furthermore, a close alignment exists between the steps in an information lifecycle and the standards that are useful in guiding a resource through the process. These standards are drawn from a number of communities and standards bodies, and all of them represent best practice in information management in their respective

Figure 1.3.
Library and Archives Canada Records Management Model (http://www.bac-lac.gc.ca)

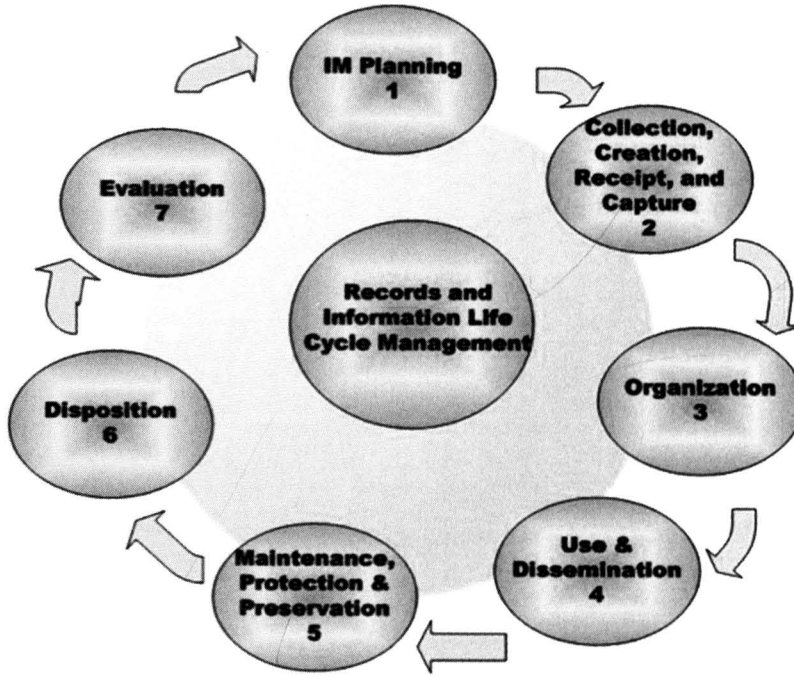
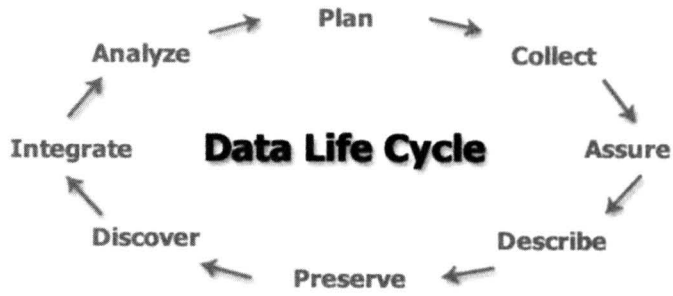
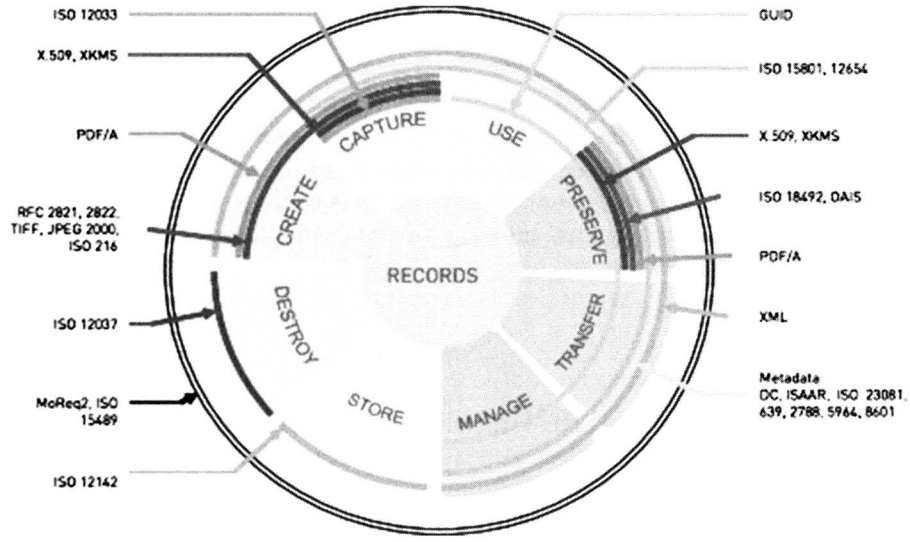


Figure 1.4.
DataOne Data Life Cycle Model (https://www.dataone.org)



fields. For example, the PDF/A (Portable Document Format/Archiving) standard is a more open version of the (PDF) standard, which enables long-term archiving, hence the additional "/A" at the end of its abbreviation. This standard is the preferred document representation format for the creation process in the MoReq2[13] digital records management specification.

Figure 1.5.
MoReq2 Digital Records Management Model



"ABOUTNESS" AND THE ROLE OF CLASSIFICATION IN INFORMATION

The information lifecycle is a useful framework for thinking the aspects of information and information services. Certain steps in the lifecycle, notably the process of description of discovery, require the creation of metadata that helps represent the "aboutness" of a resource in a discovery system. This process of assigning the "aboutness" of an item is also known as classification. Classification often involves the assignment of a series of subject headings or other categorizing data (e.g., place, date, format) to an item, in many cases also assigning a unique identifier (e.g., a call number in the Library of Congress schema) based on the content of the resource.

Classification is an important part of cataloging but is also a common activity in indexing and abstracting work. The process uses *conceptual analysis* of a work to identify the *aboutness* of a resource with the goal of enabling recall either through the search of information systems such as card catalogs or the accurate placement of a physical resource in context of other resources on the same topic. A sufficiently cataloged item should be cataloged down to the most *specific* level appropriate for a given classification, must be described thoroughly, or *exhaustively*, to ensure that all of the appropriate topics are captured in the classification, and often uses a *controlled vocabulary* to map topics in the resource onto a common classification schema. Classification often uses semantic and syntactic analysis as part of the content analysis process, extracting meaning from the actual content, or *semantics* of a work as well as from the content structure, or *syntactics* of a resource.

There are a number of classification systems commonly used in LAM communities including The Library of Congress system, the Colon Classification System, the Universal Decimal System, Bliss Bibliographic Classification system, and the Dewey Decimal system. Each of these systems focus on identifying the "aboutness" of a document and coding of that aboutness into a classification number. Broadly stated, these classification systems can be grouped into three broad types, Enumerative, Faceted, and Analytico-synthetic.

- Enumerative: Subjects are pre-defined and listed in a hierarchical notation. Application of the classification system involves finding the appropriate class in the classification system and applying the class without modification.
- Faceted: Faceted systems are non-hierarchical and involve the combination of multiple categorization areas (or facets) to create a classification. One of the most popular faceted classification systems is Ranganathan's colon classification. Ranganathan's system features five facets: Personality, Matter, Energy, Space, and Time (PMEST).
- Analytico-synthetic: Analytico-Synthetic systems are hierarchical but rather than relying on a completely pre-defined hierarchy, they allow the cataloger to add refining concepts to classification such as geographic, temporal, and topical refinements. In addition, an Analytico-Synthetic system allows the classifier to build a classification number using the combination of hierarchical and refining concepts.

These three types of classification systems (Enumerative, Faceted, and Analytico-synthetic) are the most common in LAM communities. In addition to these systems there are social classification systems known as folksonomies, which rely on the aggregation of tags assigned by users of information resources. Folksonomies are often represented in Tag Clouds, a visual representation of tags with emphasis based on tag occurrence.

Generally speaking, the Library of Congress Classification System and Dewey Decimal Systems are considered Analytico-Synthetic because they blend a hierarchical subject analysis (Enumerative) with refining classification schedules (quasi-faceted). For example, the LCC system allows you to assign geographic and time facet refinements to a subject classification and the Dewey system features 10 main divisions that are hierarchically arranged to create a classification.

Classification is not always a manual process or even a process that relies on human interpretation of meaning. Computational linguistic techniques including document clustering, topic modeling, and keyword clustering are each techniques that rely primarily on computer analysis of texts and the derivation of classification. While classification of bibliographic works in libraries is important in order to assign a unique call number for a resource, classification does not always require the assignment of a unique call number. In fact in the case of digital resources, the process of classification as well as the utility derived from it is a topic of research.

THE INTERNET AND ITS IMPACT ON INFORMATION USE

The world of information is a very different place from what it was before the introduction of the Internet to the general public. The web has changed the way