Introduction to Metadata

Why Metadata?
• Metadata: cataloging by those paid better than librarians
• Metadata creation = “the art formerly known as cataloging”?
• Metadata: Structured information about an object or collection of objects
• We must become very, very proficient with metadata — creating, harvesting, transforming, serving
• MARC is just the beginning, and unless we’re careful, will be too limiting; we must be proficient with Dublin Core, MODS, METS, etc.
• We never metadata we didn’t like (metadata R Us)
• Metadata can be both mined and enhanced

What is metadata?
• Metadata is cataloging done by men
  – Attributed alternately to Tom Delsey and Michael Gorman

What is metadata?
• The term metadata is used differently in different communities.
• Some use it to refer to machine understandable information, while others use it only for records that describe electronic resources.
• In the library environment, metadata is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital.
• Traditional library cataloging is a form of metadata; MARC 21 and the rule sets used with it, such as AACR2, are metadata standards.
• Other metadata schemes have been developed to describe various types of textual and non-textual objects, including published books, electronic documents, archival finding aids, art objects, educational and training materials, and scientific datasets.

Metadata: Early Example

What is metadata?
Most simply (and literally) …

data about data
What is metadata?

• NISO's Understanding Metadata" (2004) defines metadata as:

  • "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information."

• The American Library Association (ALA) Committee on Cataloging: Description and Access (CC:DA) presented the formal working definitions for the three terms, after a study of 46 potential definitions:

  • Metadata are structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities.
  
  • A metadata schema provides a formal structure designed to identify the knowledge structure of a given discipline and to link that structure to the information of the discipline through the creation of an information system that will assist the identification, discovery, and use of information within that discipline.
  
  • Interoperability is the ability of two or more systems or components to exchange information and use the exchanged information without special effort on either system.

• The usage guide for The Dublin Core explains the term as follows:

  • "Metadata has been with us since the first librarian made a list of the items on a shelf of handwritten scrolls. The term "meta" comes from a Greek word that denotes "alongside, with, after, next." More recent Latin and English usage would employ "meta" to denote something transcendental, or beyond nature. Metadata, then, can be thought of as data about other data. It is the Internet-age term for information that librarians traditionally have put into catalogs, and it most commonly refers to descriptive information about Web resources."

• "metadata record consists of a set of attributes, or elements, necessary to describe the resource in question. For example, a metadata system common in libraries -- the library catalog -- contains a set of metadata records with elements that describe a book or other library item: author, title, date of creation or publication, subject coverage, and the call number specifying location of the item on the shelf."

• Metadata's just another word for…
  – The broad universe of knowledge organization
    • Cataloging
    • Classifying
    • Indexing
    • Creating finding aids
    • Records management
    • Bibliographies
    • Creating museum registries
    • Creating metadata for digital libraries
    • Knowledge management

• "structured data and digital (and non-digital) resources that can be used to support a wide range of operations. These might include, for example, resource description and discovery, the management of information resources (including rights management) and their long-term preservation"

  • U.K. Office for Library and Information Networking (UKOLN)
What is metadata?

• The sum total of what we can say about any information object at any level of aggregation (e.g. in archival processing, dealing with groups (folders), not individual items)
• For a particular purpose or a particular group of users

Metadata and cataloging

• Depends on what you mean by:
  – metadata, and
  – cataloging!
• But, in general:
  – Metadata is broader in scope than cataloging
  – Much metadata creation takes place outside of libraries
  – Good metadata practitioners use fundamental cataloging principles in non-MARC environments
  – Metadata created for many different types of materials

What metadata is not

• Just a new word for cataloging
• Only for Internet resources
• Necessarily in electronic form
• Only created by professionals
• A fundamentally new idea
• A reason to forget everything we know about describing and managing resources

Little Known Facts About Metadata

• Metadata does not have to be digital
• Metadata relates to more than the description of an object
• Metadata can come from a variety of sources
• Metadata continues to accrue during the life of an information object

Some uses of metadata

• By information specialists
  – Describing non-traditional materials
  – Cataloging Web sites
  – Navigating digital objects
  – Managing digital objects over the long term
  – Managing corporate assets
• By novices
  – Preparing Web sites for search engines
  – Describing Eprints
  – iTunes

Creating descriptive metadata

• “Digital library” systems
  – ContentDM
  – ExLibris Digitool
  – Greenstone
• Library catalogs
• Spreadsheets & databases
• XML
What's an information object?
• A single item or aggregation of items that has
  – Content: what it contains or its subject (traditional cataloging focuses on this)
  – Context: who, what, where of its creation
  – Structure: how it is built, enables searching, manipulation, relating to other information objects

Information communities
• Content emphasis: libraries
• Context emphasis: archives, museums
• Structure emphasis: IT staff, computing centers

Metadata - Who needs it?
• Impact of metadata on collection access
  – Without metadata there is no service to users
  – Metadata provides the means for resource discovery, grouping, filtering, matching user needs
  – Keyword searching works only for resources that are text-based - excludes photographs, data sets, objects, maps, audio, video…
• Metadata itself as valuable content
  – Item descriptions, Finding aids, Reviews

Metadata
• Description vs. discovery
  – Full description is important for collection inventory and management - less so for discovery
  – Full description of a resource includes much information that will never be part of a user’s search key
• Deep vs. shallow
  – Basic discovery metadata supports broad, cross-domain searching that can lead users to more complete search mechanisms and descriptions

Metacrap (Cory Doctorow)
• “People lie”
• “People are lazy”
• “People are stupid”
• “Mission: Impossible—know thyself”
• “Schemas aren’t neutral” (he is referring to classification schemes)
• “Metrics influence results”
• “There’s more than one way to describe something”

The development of metadata: Pre-Internet Era of Metadata
• MAchine Readable Cataloging (MARC).
• Developed at the Library of Congress in 1960s.
• In terms of specificity, structure and maturity, it is a highly structured and semantically rich metadata.
• Purposes:
  – (1) to represent rich bibliographic descriptions and relationships between and among data of heterogeneous library objects; and
  – (2) to facilitate sharing of these bibliographic data across local library boundaries.
• The emphasis is on the entire document;
  – the surrogates are MARC records;
  – the records are produced by human catalogers;
  – MARC does not fare well with regard to management needs (e.g., intellectual property, preservation), or evaluative needs (e.g., authenticity, user profiles, and grade levels).
The development of metadata: The Internet Arena and Evolving Metadata Traditions

- Since the early 1990s,
  - distributed repositories on the Internet have had an exponential growth
  - repositories are contributed by different communities
  - there is a need to describe, authenticate, and manage these resources
  - therefore, new guidelines and architectures are developed among different communities.
- Priscilla Caplan described the metadata movement as "a blooming garden, traversed by crosswalks, atop a steep and rocky road" (Caplan, 2000).

This metadata "blooming garden" can be viewed from different perspectives:

- (1) There is no limit for the type or amount of resources that can be described by metadata.
  - For any area that shows a demand for electronic resource discovery and sharing, a metadata standard can be developed or proposed.
  - Today, the resources described by metadata consist of:
    - bibliographical objects (e.g., as represented by MARC metadata),
    - archival inventories and registers (e.g., EAD metadata),
    - geospatial objects (e.g., FGDC metadata),
    - museum and visual resources (e.g., CDWA, VRA Core, CIMI metadata),
    - educational materials (e.g., LOM),
    - software implementation (e.g., CORBA),
    - and many others.
  - The use of these metadata standards is not limited by language or country boundaries.

This metadata "blooming garden" can be viewed from different perspectives:

- (2) There is no limit for the number of overlapping metadata standards for any type of resources or any subject domain.
  - Variant systems are often found even within a single subject community.
  - In describing museum and visual resources, for instance, there are at least nine well-structured and well-documented metadata schemas, ranging from very comprehensive and detailed ones to the more general and open cores.

This metadata "blooming garden" can be viewed from different perspectives:

- (3) There is no limit for the types of profession or subject domain that can be involved in metadata standard development and application.
  - Metadata and Organizing Educational Resources on the Internet (Greenberg, 2000) documents the experiences of those who are actively engaged in projects that organize Internet resources for educational purposes, including metadata creators (both catalogers and indexers), library administrators, and educators.
  - The National Science Digital Library (NSDL) established a Metadata Repository based on the metadata records harvested from nearly 100 digital collections funded by the National Science Foundation. The collections and the metadata for the collections and items were built by educators of K-12, undergraduate, and graduate schools, together with publishers, scientists, engineers, medical doctors, professional associations, and so on.

Metadata records

- THE RELATIONSHIP BETWEEN METADATA (data used for resource description and retrieval) AND THE KNOWLEDGE ARTIFACTS THESE DATA REPRESENT (or, for which metadata serve as surrogates) is direct. In most cases, metadata are transcribed inherent data; that is, the data are taken directly from the resource and then reassembled according to the schema in such way as to create a representation of the resource.
- Caplan says metadata are “structured information about an information resources of any media type or format.” Key terms here are “structured” and “information resource.”

This metadata "blooming garden" can be viewed from different perspectives:

- KINDS OF METADATA
  - Citations
  - ISBD
  - Markup languages
  - MARC Coding and tagging
  - Webpage metadata

- Example:
  - A journal article and its citation.
  - An electronic resource and its metadata.
Metadata records

- Metadata may be either:
  - **Extrinsic**: Existing independently of the primary data being described, usually in an indexable metadata base
  - or
  - **Intrinsic**: Existing as a part of the primary data being described

Metadata records

- **Metadata embedded in digital images**
  - Some image software allows metadata records about an image to be recorded and attached in the image. When an image is viewed from the software application, it looks as if a record is embedded in the digital image. Values in some elements are automatically captured by the software while others are controlled by metadata creators.

Metadata records

- **Metadata records displayed from databases**
  - Bibliographic databases, digital collections, and digital repositories store metadata records in databases and display the records with a more user-friendly interface.
    - Library bibliographic catalogs
    - Digital collections
    - Digital repositories

Metadata types and functions

- Descriptive metadata describes a resource for purposes such as discovery and identification. It can include elements such as title, abstract, author, and keywords.
  - All about discovery
  - Catalog records, finding aids, indexes
  - Usually publicly accessible

Metadata types and functions

- Functions of Descriptive Metadata
  - **Representation**
    - Represent the resource to the user
    - Serve as a surrogate for resource itself
    - Provide descriptive information
    - Help user identify, evaluate and select
  - **Retrieval**
    - Provide means for search, browse, navigation
    - Known item searches and exploratory searches
    - Retrieve sets of results, not just individual items
      - grouped according to one or more common characteristics

Note: In many websites, metadata records are embedded in the source code of a webpage. Users usually will not see the metadata when they access and browse a website unless they choose to view the source code.
**Metadata types and functions**

- **Administrative metadata** provides information to help manage a resource, who can access it. There are several subsets of administrative data; two that sometimes are listed as separate metadata types are:
  - Rights management metadata, which deals with intellectual property rights
  - Preservation metadata, which contains information needed to archive and preserve a resource.

- **Physical condition of resource**
- **Data refreshing**
- **Technical metadata**
  - Hardware and software requirements
  - Digitization, microfilming formats/ratios
  - Encryption, passwords
  - Often not publicly accessible

- **Structural metadata** indicates how compound objects are put together, for example, how pages are ordered to form chapters.
  - How something can be used
  - Glue for compound digital objects
  - Used for machine-processing
  - Defines internal organization (structure) of object
  - Defines object types
  - Links synchronous files (audio with score)
  - Helps reconstruct distributed resources
  - Used for navigation
  - Enables use of the resource

- **Schema semantics**: Meaning ascribed by a community to a metadata element or to the values for that element. Organized into a "vocabulary."
  - Names
  - Definitions
  - Required, conditional required, or optional?
  - Repeatable?
- **Content semantics**: Content rules determine how the elements are selected and recorded (e.g., AACR2, DACS, CCO).
  - Formatting
  - Controlled vocabularies/Thesauri
  - Classification
  - Identifiers

- **Syntax**: Provides a means to represent one or more structures in a flexible, extensible manner. Provides underlying mechanism for encoding, exchange, display and machine processing of metadata. Example: HTML
  - Record structure based on specified rules
  - Constructed with search and retrieval in mind
  - Complexity may vary:
    - Independent (no prescribed syntax)
    - Medium complexity (HTML, XML)
    - Complex (MARC, SGML, etc.)

---

**Standards Landscape for Descriptive Data**

"The nice thing about standards is that there are so many of them to choose from."

- **Data Structure Standards**: MARC, EAD, DC, MODS, VRA Core, CDWA
- **Data Content Standards**: AACR2, APPM, CCO, DACS
- **Data Value Standards**: LCSH, MeSH, AAT, TGM, ULAN

"Standards are like toothbrushes, everyone agrees they're a good thing but nobody wants to use anyone else's."

--Rachel Frick
Metadata types and functions

• Structure
  – Overall containing architecture for metadata record content and syntax
  – Forms the foundation for the metadata’s transmittal and use
  – Metadata can be contained in a variety of architectural structures
    • Resource Description Framework (RDF)
    • Metadata Encoding & Transmission Standard (METS)
    • Voyager Library Catalog

Metadata types and functions

• Schema: Identifies, defines, organizes and constrains the elements in a set, their characteristics and descriptions. Involves both semantics and structure. Examples: TEI, Dublin Core, EAD, CDWA, VRA Core

Metadata schemas

• A metadata schema is
  – A set of elements (tags, fields, categories, etc.) (semantics), and the
  – Rules for their use (content)
  – For a particular purpose (syntax)

Metadata Schema Characteristics

• A set of elements
  – discrete units of data or metadata
  – may be mandatory or optional
• A name for each element
• A definition or meaning for each element
• A registry where information about each element in a metadata set is recorded

Metadata functions

• Resource discovery
  – Allowing resources to be found by relevant criteria;
  – Identifying resources;
  – Bringing similar resources together;
  – Distinguishing dissimilar resources;
  – Giving location information.

Metadata “Buzzwords”

• Interoperability
  – the ability of software and hardware on different machines from different vendors to share data
    • Crosswalks
    • Harvesting OAI-PMH
• Modularity
  – constructed with standardized units or dimensions for flexibility and variety in use
• Extensibility
  – capable of being increased in scope or range
Metadata functions

• Organizing e-resources
  – Organizing links to resources based on audience or topic.
  – Building these pages dynamically from metadata stored in databases.

• Facilitating interoperability
  – Using defined metadata schemes, shared transfer protocols, and
crossover between schemes, resources across the network can
be searched more seamlessly.
  • Cross-system search, e.g., using Z39.50 protocol;
  • Metadata harvesting, e.g., OAI protocol.

• Digital identification
  – Elements for standard numbers, e.g., ISBN
  – The location of a digital object may also be given using:
    • a file name
    • URL
  – Some persistent identifiers, e.g., (PURL (Persistent
    URL); DOI (Digital Object Identifier)
  – Combined metadata to act as a set of identifying data,
differentiating one object from another for validation
purposes.

Metadata functions

• Archiving and preservation
  – Challenges:
    • Digital information is fragile and can be corrupted or altered;
      It may become unusable as storage technologies change.
    • Format migration and perhaps emulation of current hardware
      and software platforms are strategies for overcoming these
      challenges.
    • Metadata is key to ensuring that resources will survive and
      continue to be accessible into the future. Archiving and
      preservation require special elements:
      • to track the lineage of a digital object,
      • to detail its physical characteristics, and
      • to document its behavior in order to emulate it in future
        technologies.

• Metadata schemas (also called schemes)
  generally specify names of elements and their
    semantics.
  Optionally, they may specify:
  – rules for how content must be formulated (for example, how to identify the main title),
  – representation rules for content (for example, capitalization rules), and
  – allowable content values (for example, terms must be
    used from a specified controlled vocabulary).
  Many metadata schemas are being developed in
    a variety of user environments and disciplines.

Metadata standards

• Metadata for resource description
  – Metadata such as catalog records and index citations
    have been used now for thousands of years (literally
    since antiquity). Always there has been a yearning
    among knowledge organization professionals to find
    more efficient and accurate means for providing
    resource description. Yet, even now, metadata are
    mostly compiled by lone individuals working with
    loosely defined standards.

• Standards are developed to:
  – Create durable, persistent metadata records that precisely define
    the asset so that exactly-relevant assets are identified and
    retrieved in response to a query.
  – Create metadata that is flexible, extensible, and scalable to
    support the needs of any organization, any type of asset, and
    varying skill and interest levels of metadata creators.
  – Allow the metadata records from many schemas with differing
    levels of complexity to interoperate for data discovery.
  – Enable machine-intervention for automatic interpretation of
    metadata and data discovery, particularly among disparate
    search and retrieval platforms.
Metadata Standards: Bibliographic Description

• MARC (MAchine-Readable Cataloging)
  – MARC provides the mechanism by which computers exchange, use, and interpret bibliographic information, and its data elements make up the foundation of most library catalogs used today. MARC became U.S.MARC in the 1980s and MARC 21 in the late 1990s.
  – MODS (Metadata Object Description Schema): MODS includes a subset of MARC fields and uses language-based tags rather than numeric ones, in some cases regrouping elements from the MARC 21 bibliographic format. MODS is expressed using the XML schema language of the World Wide Web Consortium.

• DUBLIN CORE: The Dublin Core metadata element set is a standard for cross-domain information resource description. It is now a U.S. national and international standard.

• Text Encoding Initiative (TEI): An international standard for representing all kinds of literary and linguistic texts for online research and teaching.

• TEI Header: In addition to specifying how to encode the text of a work, the TEI Guidelines for Electronic Text Encoding and Interchange also specify a header portion, embedded in the resource, that consists of metadata about the work.

• Archives
  – EAD (Encoded Archival Description) DTD
  – ONIX (Online Information Exchange): Built on the INDECS Framework, developed and maintained by EDItEUR jointly with book industries. The ONIX for Books Product Information Message is the international standard for representing and communicating book industry product information in electronic form. It has elements to record a wide range of evaluative and promotional information as well as basic bibliographic and trade data.

• E-Commerce
  – The INDECS project: Created to address the need, in the digital environment, to put different creation identifiers and their supporting metadata into a framework where they could operate side by side, especially to support the management of intellectual property rights. The main focus of INDECS is on the use of what is commonly (if imprecisely) called content or intellectual property.

• Archives
  – EAD (Encoded Archival Description) DTD

• E-Commerce
  – ONIX (Online Information Exchange): Built on the INDECS Framework, developed and maintained by EDItEUR jointly with book industries.

Metadata standards

• Visual Objects
  – Categories for the Description of Works of Art (CDWA): For describing works of art, architecture, groups of objects, and visual and textual surrogates.
  – VRA Core Categories: For creating records to describe works of visual culture as well as the images that document

• Geospatial Data
  – Content Standards for Digital Geospatial Metadata (CSDGM)

Design Criteria for a Metadata System:

• Durable - independent of changes to hardware, software and network infrastructure

• Interoperable - Can be seamlessly shared across the web with disparate hardware, software, network infrastructure and search engines

• Precise - Enables the creation of customized "virtual collections"—pulling objects together seamlessly from any digital space to meet exact information requirements.

• Media-Specific
  – MPEG-4: A standard for multimedia for the fixed and mobile web.
  – MPEG-7: A standard for description and search of audio and visual content.
Design Criteria for a Metadata System:
• Flexible - Supports any search engine, search strategy, transport or display option
• Efficient - Provides immediate access to the most appropriate asset for the searcher.
• Controlled - Insures digital assets are from a trusted source to an authorized end user.
• Granular - Able to search the top page, subsequent pages, or drill down to an underlying database of objects.

Standards
• Increase interoperability
• Lower use and participation barriers
• Build larger communities of users which can drive creation of a wider range of relevant services and tools (Windows vs Mac)
• Improve chances of long term survival of materials
• Prefer open over proprietary

Primary Functions of Metadata
• Creation, multiversioning, reuse and recontextualization of information objects
• Organization and description
• Validation
• Searching and retrieval (a.k.a. discovery)
• Utilization and preservation
• Disposition

Why is Metadata Important?
• Increased accessibility
• Retention of context
• Expanding use
• System development and enhancement
• Multiversioning
• Legal issues
• Preservation and persistence
• System improvement and economics

CATALOGING IN PUBLICATION
• In the early twentieth century (1901 in fact) the Library of Congress began to make copies of its catalog cards available for purchase by librarians. This was the real beginning of cooperative cataloging. For any book for which the Library of Congress had prepared cataloging, you the local librarian were freed from that effort. All you had to do was buy the cards, type added entries on top of them and call numbers in the upper left corner, and then file the cards.
• Savings were dramatic. As a result, standardization of cataloging spread across the United States, then North America, then throughout the English-speaking world, as cooperation grew among the Library of Congress, the British Library (then the library of the British Museum) and the National Library of Canada.

CATALOGING IN PUBLICATION
• In the 1950s there were many projects undertaken to provide copies of proof sheets for LC cards in the books libraries were buying as new acquisitions. This meant that, if your jobber participated in the program, the mere act of buying the book also brought with it the professional and standardized cataloging. This was pretty close to in-source metadata for the time.
CATALOGING IN PUBLICATION

• Beginning in 1961 publishers and librarians in the U.S. (and later worldwide) began to cooperate on a larger scale, implementing a project known as Cataloging in Publication, or CIP. You've surely seen CIP copy on the verso of title pages of books you've acquired.
• Here is metadata literally in the resource. Now if only we could teach resources to describe themselves.

MARKUP LANGUAGES

• Markup languages provide vocabulary and syntax, which, when entered into a document, provide cues for computer manipulation of the text.
• It is markup language that turns normal text into a website.

MARKUP LANGUAGES

• International Standard for Bibliographic Description (ISBD): Punctuation as Markup
• Framework for the descriptive portion of a bibliographic record (the title transcription, through the series transcription and annotations). Disseminated in 1974 in the first generic ISBD (International Standard for Bibliographic Description), these conventions quickly became the norm worldwide.

MARKUP LANGUAGES

• The punctuation, which always precedes an element, delineates the parts of this record. The title is followed by a statement of responsibility, which must be preceded by a space-slash-space, thus the title must be:
Rhksow fjkslw bf ksjk jsiousol
because the statement of responsibility is:
w Hfuysce can Lqzx.

MARKUP LANGUAGES

• The conventions of ISBD punctuation can be found in AACR2. A summary:
  • . – (full-stop, space, dash, space) precedes a new area of description
  • / (space, slash, space) precedes a statement of responsibility
  • : (space, colon, space) precedes the second element of an area (the publisher in area 4, the illustrations in area 5)
  • ; (space, semi-colon-space) precedes the third element of an area (a second author in area 1, a second city or publisher in area 4, the dimensions in area 5)
Machine-Readable Cataloging (MARC)

- No discussion of "mark-up" would be complete without a nod to the MARC coding language, which has fueled the great international effort to make catalogs electronic and to share catalog data worldwide via computer transmission.
- Essentially, catalog data are compiled according to standards (mostly AACR2) then marked up with MARC. The MARC tags, which one can view on OCLC or in "full" displays in online catalogs, but which are not visible to the searching public, designate for the computer the contents of fields and subfields. Their function is similar to that of the ISBD punctuation, but the language of MARC is much more complex.

Here is a MARC markup of the bibliographic record from the preceding example:

245 10 Rhkjsow fjkslw bf ksjk jsiouso / $c w Hfuysc can Lqzx.
250 2c pj.
300 232 p. ; $c 28 cm.

MARKUP LANGUAGES IN PUBLISHING

- In the early automation of publishing, markup was used to set cues within an author's text, which would tell a type-setting program how to set the type when it printed out the book (article, etc.).
- A simple version might look like this:
  `<b><t>Introduction to Markup Languages</t></b> by John Smith
  <pl>Chicago</pl><pu>Silly Press</pu><b><d>2001</d></b>

- This markup (which I also just invented) might turn that text into a title page something like this:
  Introduction to Markup Languages
  By John Smith
  Chicago
  Silly Press
  2001

  Note that each element is marked on both ends; that is text is enclosed between a start tag `<a>` and an end tag `</a>`.

STANDARD GENERALIZED MARKUP LANGUAGE (SGML)

- SGML was the first "meta" markup language.
- Developed to serve as a standard platform for the development of other languages, SGML provides conventions for naming the logical elements of documents, and syntax for expressing the logical relations among document components.
- SGML was intended to be used by specific communities to develop specific markup languages, known as Document Type Definitions or DTDs.
- Most of the metadata schema that we will be studying in this course, are in fact, SGML-derived DTDs.

HYPERTEXT MARKUP LANGUAGE (HTML)

- HTML is an SGML DTD that underlies the World Wide Web. HTML is the source code that resides behind the displayed website, telling browsers how to display the text to the viewer, and serving as source data for search engines.
- According to Ian S. Graham's 1995 HTML Sourcebook (New York: Wiley) requires a document to be constructed with sections of text marked as logical units, such as titles, paragraphs, or lists, and leaves the interpretation of these marked elements up to the browser displaying the document.

- An HTML document is composed of elements, which are marked by tags. Some elements do not affect a block of text (such as a paragraph command); these are called empty elements, and do not require end tags. Element names and attributes (which instruct the browser but do not display) are case-insensitive. But the attribute value (the text that will display) is case-sensitive.
**Hypertext Markup Language (HTML)**

- An HTML document has two main elements: `HEAD` and `BODY`. Each main element has sub-elements. The `TITLE` sub-element is the only required element of `HEAD`.
- The `BODY` has many sub-elements, such as:
  - **Headings**, which come in six levels:
    - `<H1>` ... words ... `<H1>`
    - `<H2>` ... words ... `<H2>`
    - `<H3>` ... words ... `<H3>`
    - `<H4>` ... words ... `<H4>`
    - `<H5>` ... words ... `<H5>`
    - `<H6>` ... words ... `<H6>`
  - These tags cause headings to display in different sizes of type, from large, bold-face (h1) to small type (h6).

- **Highlighting**, which gives special emphasis:
  - `<EM>` ... words ... `<EM>` will render the phrase in *italics*
  - `<STRONG>` ... words ... `<STRONG>` will render the phrase in **bold**.

- **Paragraphs**, an empty element, causes the text to break into paragraphs `<P>`
- **Break** is similar `<BR>`

- **Lists** cause a list to appear indented and bulleted. Lists may be unordered (`ul`) or ordered (`ol`):
  - `<UL>`
  - List items, each tagged with `<LI>`
  - `<UL>`
- **Horizontal Rule** draws a horizontal line across the page `<HR>`

- **Hypertext Links** can be used to move between documents:
  - `<A HREF="http://smiraglia.org">Click here for my Vita</A>`

- **Images** can be embedded in a webpage. For instance, a still image in the form of a graphical interface file (gif) can appear to be embedded in the website by using a hyperlink:
  - `<IMG SRC="portrait.gif">`

- **Tables** format text into tabular form. The following code creates a table with three columns and two rows:
  - `<TABLE>`
  - `<TR><TD>first data</TD><TD>second data</TD><TD>third data</TD></TR>`
  - `<TR><TD>fourth data</TD><TD>fifth data</TD><TD>sixth data</TD></TR>`
  - `<TABLE>`
- "Markup" per se is structural metadata that tell the browser how to display otherwise normal text