



REPÚBLICA BOLIVARIANA DE VENEZUELA  
UNIVERSIDAD DEL ZULIA  
FACULTAD EXPERIMENTAL DE CIENCIAS  
DIVISIÓN DE PROGRAMAS ESPECIALES  
LICENCIATURA EN COMPUTACIÓN



## **Herramienta Web para la Clasificación y Recuperación de Información Digital Basada en Etiquetas**

**Informe de Trabajo Especial de Grado presentado como  
requisito para optar al título de Licenciado en Computación**

Autor: Br. Edinson Padrón Urdaneta

Tutor: MSc. Gerardo Pirela Morillo

Maracaibo, marzo de 2014

**Herramienta Web para la Clasificación y Recuperación de Información Digital  
Basada en Etiquetas**

---

*Br. Edinson Padrón Urdaneta*

*Cl. No.: 19.216.488*

*Teléfono: +58 414 6574149*

*Venezuela, Zulia, Maracaibo, Urb. La Paz, Calle 96D, Avenida 56*

*Correo electrónico: edinson.padron.urdaneta@gmail.com*

---

*MSc. Gerardo Pirela Morillo*

*Cl.: 12.404.565*

*Teléfono: +58 412 1734718*

*Correo electrónico: gepirela@fec.luz.edu.ve*

Padrón Urdaneta, Edinson. “**Herramienta Web para la Clasificación y Recuperación de Información Digital Basada en Etiquetas**”. Trabajo Especial de Grado. Universidad del Zulia. Facultad Experimental de Ciencias. División de Programas Especiales. Maracaibo, Venezuela. 2014. 24p.

## **RESUMEN**

**Palabras claves:** Aplicación web, Etiquetado, Clasificación, Recuperación, Información digital

**Correo electrónico:** edinson.padron.urdaneta@gmail.com

Padrón Urdaneta, Edinson. “**Herramienta Web para la Clasificación y Recuperación de Información Digital Basada en Etiquetas**”. Trabajo Especial de Grado. Universidad del Zulia. Facultad Experimental de Ciencias. División de Programas Especiales. Maracaibo, Venezuela. 2014. 24p.

## **ABSTRACT**

**KeyWords:** Web application, Tagging, Classification, Retrieval, Digital information

**Email:** edinson.padron.urdaneta@gmail.com

## ÍNDICE DE CONTENIDO

	Pág.
Resumen .....	3
Abstract .....	4
Índice de Tablas .....	7
Índice de Figuras .....	8
Introducción .....	9
Capítulo I. El Problema	
Planteamiento del Problema y Justificación de la Investigación .....	10
Alcance del problema .....	12
Objetivos .....	12
Objetivo General .....	12
Objetivos Específicos .....	12
Capítulo II. Marco Teórico	
Antecedentes de la Investigación .....	13
Tags .....	13
TagTool .....	14
Tables .....	14
Bases Teóricas .....	15
Sistema de Archivo .....	15
Transparencia de Localización .....	15
Aplicación Web .....	16
ORM .....	16
AJAX .....	17
JSON .....	17
WSGI .....	17
Capítulo III. Marco Metodológico	

Distribución de Actividades .....	20
Capítulo IV. El título de este capítulo	
Referencias bibliográficas .....	24

## ÍNDICE DE TABLAS

	Pág.
Tabla 1. Cronograma .....	20

**ÍNDICE DE GRÁFICOS**

Pág.



## INTRODUCCIÓN

# **CAPÍTULO I**

## **El Problema**

### **1. Planteamiento del Problema y Justificación de la Investigación**

La época actual es llamada la “era de la información” (Giuliano, 1983) debido a la relevancia de ésta en la sociedad. Furth (1994) escribió: “La información es actualmente tan vital, y tan intangible, como el aire que respiramos, el cual está lleno de ondas de radio”. Hoy en día es el bien más importante y valioso para toda compañía, país y grupo social, por lo que su refinamiento a partir de la abundante cantidad de datos que son creados cada segundo es un proceso primordial y que ofrece una ventaja significativa.

En el año 2010 se produjo y almacenó un estimado de dos exabytes de información (esto incluye todos los medios: libros, revistas, documentos, Internet, fotografías, televisión, radio, música, entre otros) de los cuales, el 93% era digital (Burgin, 2010). Dos años después, la cantidad de información digital generada, se elevó a 2.5 exabytes al día (IBM, 2012), evidenciando un crecimiento exponencial, con la Internet como principal catalizador de este fenómeno.

A la Internet se le suma el creciente uso de dispositivos móviles (teléfonos inteligentes, tabletas, lectores electrónicos, entre otros) que permiten la accesible y cómoda creación y consumo de información gracias a la web, lo cual provoca una cuantiosa producción de datos, además de contribuir con la heterogeneidad y descentralización de los mismos, ocasionando así que su tratamiento sea una tarea penosamente difícil de realizar.

En la actualidad, los distintos repositorios presentan al usuario la información digital en ellos almacenados mediante una estructura jerárquica de directorios en forma de árbol invertido. Esta representación, nacida en 1969 con el lanzamiento del sistema operativo Multics en respuesta a la necesidad de un sistema de almacenamiento secundario en un ambiente de multiprogramación (Multics, 1968), rígida por causa de su inherente propiedad taxonómica, limita la manera en la que puede ser organizada la información, mediante una relación padre-hijo, y hace de la búsqueda de la misma un proceso arduo debido al uso de engorrosas rutas de directorios. Además, esta aproximación dista en gran medida del método natural usado

por el hombre para almacenar y recuperar información de su memoria a largo plazo, el cual consiste en codificarla semánticamente para su almacenamiento (Baddeley, 1966), asociando los datos que percibe (Atkinson y Shiffrin, 1968). Por lo tanto, el ser humano no es una máquina que trabaje en base a datos aislados, éste busca interrelacionarlos, describiéndolos de manera inherentemente subjetiva, breve y dinámica; haciendo necesario un método semejante para clasificar y recuperar la información digital, siendo éste más flexible, intuitivo y natural.

Afortunadamente, la web ha dado origen a un nuevo sistema de clasificación. Originalmente diseñado e implantado por el servicio del.icio.us (Mathes, 2004), el sistema ofrece la posibilidad de describir recursos mediante un conjunto de palabras clave llamadas etiquetas (*tags*, por su término en inglés), las cuales son consideradas por los usuarios como relevantes para caracterizar dichos recursos de acuerdo a sus necesidades sin depender de un vocabulario controlado o de una estructura previamente definida, estableciendo así una relación entre el recurso y un concepto en su mente, con el objeto de organizar el contenido para uso futuro de una manera fácil y flexible (Specia y Motta, 2007). Este esquema de catalogación se ha extendido en uso por una gran variedad de servicios web debido a su popularidad entre los usuarios, quienes ven en éste una manera más natural, sencilla, rápida y personal de clasificar la información digital que les es de interés, hallándola posteriormente con mínimo esfuerzo, siendo ésta una mejor alternativa al problema de darle tratamiento a una gran cantidad de datos, contrastada con la estructura jerárquica que ofrecen los sistemas de archivos modernos.

Hoy en día, existe una serie de herramientas que extienden las capacidades de los sistemas de archivos al permitir el uso de etiquetas; sin embargo, estas herramientas presentan algunas limitaciones, tales como: el tratamiento exclusivo de una fracción de la información digital existente, su operación bajo un número reducido de plataformas, la prestación de un conjunto restringido de funcionalidades, entre otros.

Con este trabajo de investigación se buscó desarrollar una herramienta web integral que garantizara la interoperabilidad entre los distintos sistemas operativos y dispositivos al permitir un acceso convenientemente centralizado a la información digital sin importar la distribución física de ésta, brindando así transparencia de localización mediante el uso de un servidor central que contenga la herramienta y aplicaciones que enlacen los repositorios de los usuarios con dicho servidor. Además, se implantó una manera sencilla, versátil y personal de clasificar

y recuperar la información digital contenida en los mencionados repositorios al permitir su descripción y recuperación mediante el uso de etiquetas definidas a conveniencia. Adaptando así la tecnología nacida en la web al ámbito de los actuales sistemas de archivo.

## 2. Alcance del problema

La investigación estuvo delimitada al desarrollo de un sistema cliente-servidor donde el cliente, desarrollado para trabajar solo bajo los entornos Windows®, OS X® y Linux®, se comunica con una aplicación web, manteniendo sincronizadas las etiquetas asociadas a los archivos ubicados en los repositorios del usuario, permitiéndole así clasificarlos y recuperarlos mediante el uso de las mencionadas etiquetas.

## 3. Objetivos

### 3.1. Objetivo General

- Desarrollar una herramienta web para la clasificación y recuperación de información digital basada en etiquetas

### 3.2. Objetivos Específicos

- Realizar una revisión documental sobre herramientas de clasificación y recuperación de información digital
- Implementar la aplicación servidor para el almacenamiento y gestión de etiquetas de recursos digitales
- Desarrollar el módulo de la aplicación cliente encargada del rastreo personalizable de la información a etiquetar
- Desarrollar el módulo de la aplicación cliente correspondiente a la interfaz web de usuario
- Realizar las pruebas de caja gris de la herramienta

## Marco Teórico

En la actualidad, los sistemas de clasificación basados en etiquetas son ampliamente usados por innumerables aplicaciones web, sin embargo, su uso como complemento a los sistemas de archivo modernos presenta una reducida gama de opciones, dentro de la cual se pueden mencionar:

Tags es una aplicación que permite etiquetar archivos y directorios en *finder* (explorador), correos electrónicos en *Mail* (gestor de emails), fotos en *iPhoto* (gestor de imágenes) y enlaces en *Safari* (navegador web); con el objetivo de mantener organizados dichos recursos y poder hallarlos de manera rápida y sencilla.

## Gráfica 1. Tags



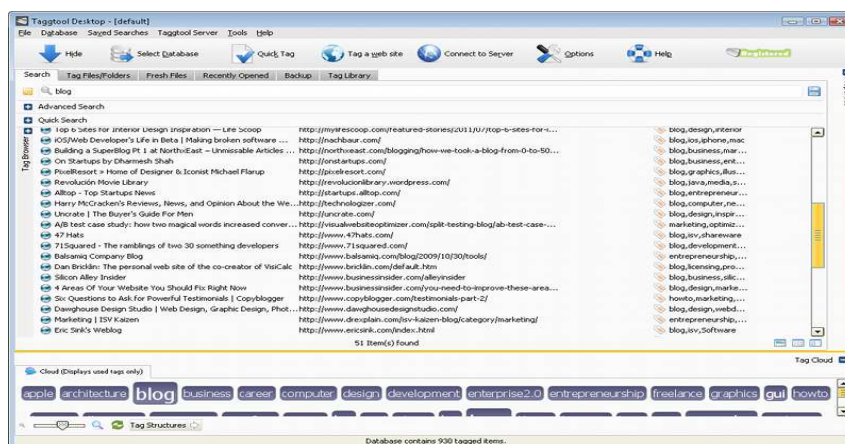
## 1.2. TagTool

TagTool permite añadir etiquetas a archivos, directorios y «bookmarks». Adicionalmente, el usuario puede proporcionar una descripción del recurso, además de una calificación cuantitativa (*rating*). Esta herramienta ofrece la posibilidad de importar metadatos pre-existentes en los recursos y realizar búsquedas en base a estos.

TagTool también posee una alta integración con el sistema operativo y brinda una vista previa de los recursos gestionados. Además, facilita un sistema de etiquetado automático basado en la naturaleza de los recursos, y muestra una nube de etiquetas para la visualización de las mismas.

Al igual que Tags, TagTool está sujeto a una licencia privativa y no es capaz de sincronizar recursos localizados en dispositivos diferentes. Otra desventaja que presenta este sistema es que está disponible solo para el sistema operativo *Microsoft Windows®*.

**Gráfica 2. TagTool**



**Fuente: TagTool (2014)**

## 1.3. Tables

Tabbles ofrece la posibilidad de clasificar, buscar, organizar y compartir (a través de una LAN o la Internet) archivos, carpetas y «bookmarks», mediante el uso de etiquetas. Adicionalmente, la herramienta ofrece un sistema de auto-etiquetado, además de una excelente integración con el sistema operativo.

A pesar de ser capaz de gestionar, compartir y sincronizar recursos localizados en dispositivos distintos, Tabbles está disponible únicamente para el sistema operativo Microsoft Windows® bajo una licencia privativa.

**Gráfica 3. Tabbles**



**Fuente: Tabbles (2014)**

## 2. Bases Teóricas

### 2.1. Sistema de Archivo

Un sistema de archivo es un módulo central de todo sistema operativo moderno que consiste de estructuras lógicas y rutinas que controlan la creación, modificación, eliminación y acceso a los datos que residen en una unidad de disco, partición o volumen lógico. Dicho sistema organiza los datos en una estructura jerárquica mediante el uso de directorios, los cuales funcionan como contenedores de punteros a múltiples archivos (EMC, 2012). Algunos de los sistemas de archivo más ampliamente usados en la actualidad son: FAT, FAT32, NTFS, HFS Plus, ext2, ext3 y ext4.

### 2.2. Transparencia de Localización

Se refiere a una consideración de diseño de los sistemas operativos distribuidos. Se habla de transparencia de localización cuando un usuario accede a un recurso en particular sin que sea necesario conocer la ubicación del mismo dentro de la red de trabajo a fin de acceder a él. El no conocer la localización de un recurso implica que todo acceso a éste es realizado por medio de

un nombre que no depende de la ubicación dónde reside actualmente, ni la ubicación dónde fue creado (Borghoff y Schlichter, 2000).

### 2.3. Aplicación Web

Es una aplicación desarrollada mediante el uso de tecnologías web (html, css, javascript, entre otros) cuyo entorno de ejecución (*runtime enviroment*) está conformado por un motor web (*web engine*). Como ejemplos de este tipo de aplicaciones se pueden mencionar: sistemas de reservación, sitios de compras en línea, juegos, aplicaciones multimedia, mapas, aplicaciones interactivas de diseños, sistemas de correo, entre otros (W3C, 2012).

Los cuatros componentes típicos de una aplicación web son: un navegador, un servidor web, programas de aplicación y un servidor de base de datos. Las responsabilidades del navegador incluyen: presentar una interfaz de usuario, comunicarse con un servidor web, ejecutar *scripts* embebidos, gestionar la *cache* y las *cookies*. El servidor web se encarga, entre otras cosas, de comunicarse con los clientes, proveer contenido estático, invocar programas de aplicación para la generación de contenido dinámico y gestionar las conexiones (Grove, 2009).

*CGI*, *servlets*, lenguajes de plantilla (*template languages*) y lenguajes de guión (*script languages*) son algunos de los entornos de desarrollo para la construcción de los programas de aplicación mencionados anteriormente, y están típicamente acompañados por un sistema de gestión de base de datos relacional (*RDBMS*, por sus siglas en inglés) accedido por los programas de aplicación a través de un *ODBC / ORM* (Grove, 2009).

### 2.4. ORM

Es una herramienta que permite conectar un objeto, a veces llamado “modelo de dominio”, con una base de datos relacional de manera automática mediante el uso de metadatos como descriptores del objeto y los datos en sí mismos. Algunas de las ventajas que ofrece el uso de un ORM sobre otras técnicas de acceso de datos se listan a continuación: en primer lugar, se automatiza la conversión bidireccional entre objetos y tablas; además, se provee el almacenamiento en *cache* de los objetos de manera transparente en el lado de la aplicación, mejorando así el rendimiento del sistema; y por último, brinda una API de acceso a la base de datos, abstrayendo por completo el sistema de gestión de bases de datos relacional



(Mehta, 2008). *ODB*, *JDBC*, *Peewee*, *SQLAlchemy* y *Yii* son algunos de los ORM usados en la actualidad.

## 2.5. AJAX

Es una tecnología empleada para prevenir la carga de una página en su totalidad al permitir la transferencia de datos entre el cliente y el servidor web de manera asíncrona, logrando que las páginas se carguen más rápidamente al retrasar la descarga de grandes recursos (Zakas, 2010).

## 2.6. JSON

Es un formato ligero y fácil de analizar sintácticamente para representar datos, basado en el uso de la sintaxis dispuesta para los literales de tipo arreglo y objeto en el lenguaje de programación JavaScript, usado generalmente para transmitir información entre distintos sistemas de software (Zakas, 2010). La serialización de información estructurada, tarea que se solía realizar con XML, es actualmente llevada a cabo mediante el uso del formato JSON, tendencia que crece con el pasar del tiempo.

## 2.7. WSGI

Es una interfaz simple y universal entre servidores y aplicaciones web para el lenguaje de programación Python (Eby, 2010). En términos simples, esta especificación establece la CGI (*Common Gateway Interface*) para el lenguaje de programación Python.

## CAPÍTULO III

### Marco Metodológico

#### 1. Primer Lanzamiento

##### 1.1. Primer Sprint

Durante este *sprint* se culminó la revisión de literatura relacionada con los sistemas de clasificación y recuperación de información digital basados en etiquetas, con la recopilación de los antecedentes de esta investigación. Además, se comenzó a diseñar e implementar la interfaz web de usuario a través del diseño de bocetos físicos y digitales; y la base de datos, mediante el esquema fundamental de la misma. Adicionalmente, se dio inicio al proceso de codificación e implementación de la aplicación servidor, diseñando la funcionalidad base para recibir y responder peticiones HTTP. Todo esto junto a las respectivas pruebas de caja gris de los componentes desarrollados, acción que se mantuvo durante todo el ciclo de desarrollo de la herramienta.

##### 1.2. Segundo Sprint

Durante este sprint se continuó con las actividades iniciadas en el anterior refinando el diseño de la base de datos, añadiendo nuevas características a la aplicación servidor y mejorando la interfaz web de usuario con el fin de hacerla más intuitiva y funcional.

##### 1.3. Tercer Sprint

Al igual que en el anterior, durante este sprint se dio continuidad a las labores anteriormente descritas, añadiendo un nuevo conjunto de características a la aplicación servidor, reflejando éstas mejoras en la interfaz web del usuario y dándole los toques finales a la base de datos, finalizando el proceso de diseño e implementación de la misma.

## 2. Segundo Lanzamiento

### 2.1. Cuarto Sprint

En este sprint se comenzó el diseño de una API pública, implementada en la aplicación servidor, para permitir una comunicación sencilla con la misma, potenciando así la posibilidad de desarrollar una aplicación cliente que se comunique con la aplicación servidor de manera efectiva. También se hicieron un par de cambios menores a la interfaz web de usuario.

### 2.2. Quinto Sprint

Es en este sprint que se da inicio a la codificación e implementación de la aplicación cliente, conformada, al final de este ciclo, solo por el mecanismo de rastreo de recursos del sistema. Por otro lado, se sigue con la expansión de la API mencionada anteriormente y el refinamiento de la interfaz web de usuario, acciones que se repiten hasta el final del tercer lanzamiento.

## 3. Tercer Lanzamiento

### 3.1. Sexto Sprint

Se desarrolló el módulo que permite al usuario seleccionar a potestad cuáles directorios serán incluidos en el proceso de rastreo, así como el módulo que se encarga de mantener sincronizada con el servidor la información generada a partir del mencionado rastreo.

### 3.2. Séptimo Sprint

A lo largo de este sprint, se implementó la gestión de usuarios en la aplicación servidor, se adaptó la interfaz web de usuario para reflejar dichos cambios y se implementó en ambas aplicaciones un mecanismo de comunicación sencillo, robusto y seguro, basado en *tokens*.

## 4. Cuarto Lanzamiento

#### 4.1. Octavo Sprint

Durante este sprint se desarrolló e integró un mecanismo de etiquetado automático de la información rastreada.

#### 4.2. Noveno Sprint

Este sprint cerró el ciclo de desarrollo de esta investigación, durante el cual se llevaron a cabo las pruebas más rigurosas del sistema, con la finalidad de detectar y posteriormente corregir errores funcionales que residían en la aplicación.

**Tabla 1. Distribución de Actividades**

Año	2013							2014					
Mes	Junio	Julio	agosto	septiembre	octubre	noviembre	diciembre	enero	febrero	marzo	abril	mayo	junio
Lanzamiento				1º			2º		3º		4º		
Sprint				1º	2º	3º	4º	5º	6º	7º	8º	9º	
Asesorías con el tutor													
Revisión de literatura													
Diseño e implementación de la base de datos													
Diseño e implementación de la interfaz web de usuario													
Codificación e implementación de la aplicación servidor													
Codificación e implementación de la aplicación cliente													
Pruebas de caja gris de la herramienta													
Elaboración del informe final del TEG													

**Fuente: Padrón (2014)**

## CAPÍTULO IV

### El título de este capítulo

#### 1. El título de esta sección

En primer lugar, se describirán las clases que definen los modelos usados por el módulo servidor. **BaseModel** pauta la interfaz mínima que debe poseer todo modelo, obligando la implementación de tres métodos fundamentales, *exist*, que determina si un registro existe o no; *get\_by\_id*, el cual realiza una búsqueda de registros a partir de un conjunto de identificadores; y finalmente, *json*, el cual transforma un registro al formato *json*.

La clase **User** aglomera la información relacionada con el nombre e información de acceso de un usuario del sistema. Las clases **Device**, **Search** y **Tag**, son usadas para representar los dispositivos, búsquedas y etiquetas de los usuarios, respectivamente; estas clases dependen directamente de la clase **User**.

Por su lado, la clase **Resource** sirve para representar los recursos almacenados en los distintos dispositivos de los usuarios. Las clases **TagResources** y **TagSearch** son empleadas para relacionar etiquetas con recursos (en el primer caso) y búsquedas (en el segundo caso).

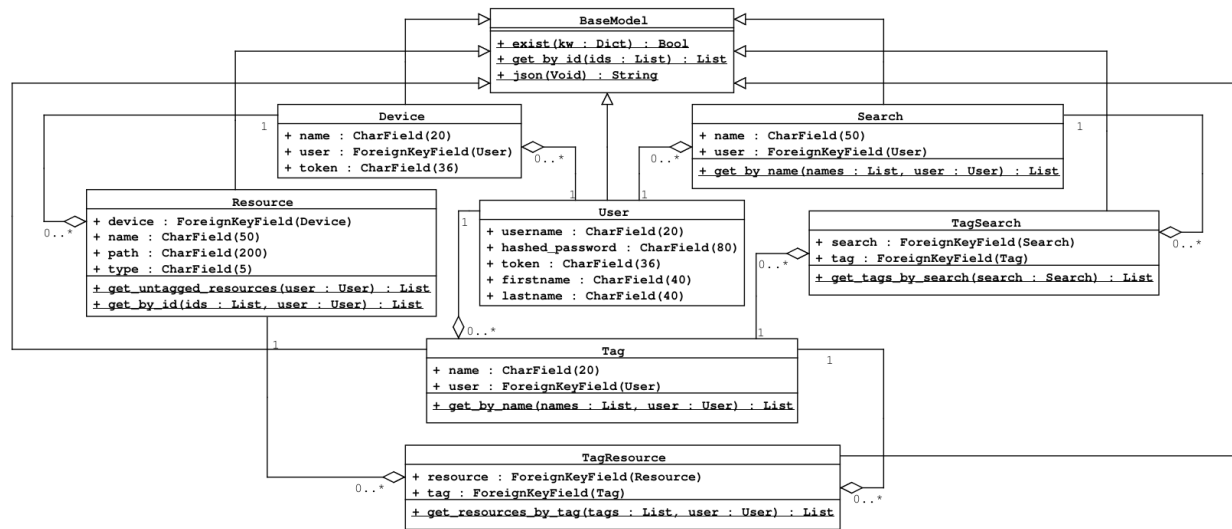
Las anteriores, están plasmadas en el diagrama de clases de la gráfica 4, presentada más abajo; y a excepción de **BaseModel**, corresponden a una y sola una tabla en la base de datos, describiendo en totalidad los modelos implementados en el sistema.

En el caso del módulo cliente, se tiene la clase **Manager**, la cual sirve como pilar fundamental al encargarse de la generación, lectura y respaldo de los archivos de configuración y estado del sistema, así como también de la comunicación con el módulo servidor. El respaldo y comunicación con el servidor se lleva a cabo gracias a las clases **Cache** y **SyncAgent**, respectivamente. Estas tareas corresponden a dos hilos de ejecución independientes, ambos instancias de la clase **DaemonThread**. La clase **Crawler** encapsula las rutinas relacionados con el rastreo de recursos en el dispositivo anfitrión, mientras que la clase **AutoTagger** es la encargada del etiquetado automático de los mencionados recursos.

Por último, **SettingsForm** es una clase empleada en la verificación de los datos enviados por

el usuario en el formulario dispuesto para la actualización de los parámetros de configuración del sistema.

**Gráfica 4. Diagrama de Clases de los Modelos**



**Fuente: Padrón (2014)**

Gráfica 5. Diagrama de Clases del Módulo Cliente



Fuente: Padrón (2014)

## REFERENCIAS BIBLIOGRÁFICAS

- Atkinson, R.C.; Shiffrin, R.M. (1968). Chapter: Human memory: A proposed system and its control processes. The psychology of learning and motivation. pp. 89-195.
- Baddeley, A.D. (1966). The influence of acoustic and semantic similarity on long-term memory for word sequences. The Quarterly Journal of Experimental Psychology. pp. 302-309.
- Burgin, Mark (2010). Theory of Information: Fundamentality, Diversity and Unification. World Scientific. p. VI.
- Furth, J. (1994). The Information Age in Charts, Fortune International.
- Giuliano, V.E. (1983). The United States of America in the Information Age, en Information Policy and Scientific Research, Elsevier, Amsterdam, pp. 59-76.
- IBM (2012) (Página consultada el 25 de junio de 2013) [On-line]. Dirección: [www-01.ibm.com/software/data/bigdata/](http://www-01.ibm.com/software/data/bigdata/)
- Mathes, Adam (2004) (Página consultada el 25 de junio de 2013) [On-line]. Dirección: [www.adammathes.com/academic/computer-mediated-communication/folksonomies.html](http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html)
- Multics (1968) (Página consultada el 25 de junio de 2013) [On-line]. Dirección: [www.multicians.org/fjcc4.html](http://www.multicians.org/fjcc4.html)
- PC Magazine (2014) (Página consultada el 23 de enero de 2014) [On-line]. Dirección: <http://www.pcmag.com/encyclopedia/>
- Schwaber, Ken; Sutherland, Jeff (2011). The Definitive Guide to Scrum: The Rules of the Game.
- Specia, Lucia; Motta, Enrico (2007). Integrating Folksonomies with the Semantic Web. Knowledge Media Institute - The Open University.
- Guenther, Rebecca; Radebaugh, Jaqueline (2004). Understanding Metadata. NISO Press.
- wise GEEK (2014) (Página consultada el 20 de mayo de 2014) [On-line]. Dirección: <http://www.wisegeek.com/what-is-location-transparency.htm>