
UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN

# Sistema de recomendación

# para el Buscador Académico Venezolano

**Trabajo Especial de Grado presentado ante la ilustre**
**Universidad Central de Venezuela por el**
**Br. José Israel Rodríguez Laguna**
**Tutor: Prof. José R. Sosa**

**Octubre del 2015**

Universidad Central de Venezuela.

Facultad de Ciencias

Escuela de Computación

Sistema de recomendación para el Buscador Académico Venezolano

**Autor**: Br. José Israel Rodríguez Laguna.

**Tutor**: Prof. José R Sosa.

**Fecha**: Octubre del 2015.

## RESUMEN

Los sistemas de recomendación son herramientas que proveen sugerencias basándose en las preferencias del usuario. En el caso de los sistemas de recomendación basados en contenidos aplicados a repositorios digitales institucionales, aparte de las preferencias del usuario se apoyan en los metadatos del recurso a recomendar. En este trabajo, se estudia el proceso de elaboración del primer Sistema de Recomendación basado en contenido para el Repositorio Digital Institucional Buscador Académico Venezolano. Para lograr esto se utilizo un proceso de Text Mining sobre la metadata de los objetos del repositorio, en la cual se pondero las características mediante el índice TF-IDF, para luego categorizar a los objetos mediante el método de Clasificación Jerárquica.

**Palabras Claves:** Metadato, Repositorio Digital Institucional, Mineria de Datos, Text Mining, TF-IDF, Sistema de recomendación, Clasificacion Jerarquica

Universidad Central de Venezuela
Facultad de Ciencias
Escuela de Computación

# ACTA

Quienes suscriben, miembros del jurado designado por el Consejo de la Escuela de Computación para examinar el Trabajo Especial de Grado, presentado por el Bachiller José Israel Rodríguez Laguna C.I.:20.911.444, titulado "Sistema de recomendación para el Buscador Académico Venezolano", a los fines de cumplir con el requisito legal para optar al título de Licenciado en Computación, dejan constancia de lo siguiente:

Leído el trabajo por cada uno de los miembros del jurado, se fijó el día 15 de Octubre de 2015, a las 11:00 am, para que sus autores lo defendieran de forma pública, en el aula PAIII, lo cual estos realizaron mediante una exposición oral de su contenido, y luego respondieron satisfactoriamente a las preguntas que le fueron formuladas por el jurado, todo ello conforme a lo dispuesto en la Ley de Universidades y demás normativas vigentes de la Universidad Central de Venezuela. Finalizada la defensa pública del Trabajo Especial de Grado, el jurado decidió aprobarlo con una nota de 20 puntos.

En fe de lo cual se levanta la presente acta, en Caracas el 15 de Octubre de 2015, dejándose también constancia de que actuó como Coordinador del jurado el Profesor Tutor José R. Sosa.

X
Prof. Jose R. Sosa
Tutor

X
Porf. Jesus Lares
Jurado1

X
Prof. Hector Navarro
Jurado 2

# Agradecimientos

A mis padres, José Luis Rodríguez y Denise Laguna Bagarozza, quienes de tantas formas, y tan repetitivamente prestaron su ayuda y su respaldo en todo momento para permitirme hacer este largo trayecto de formación y estudio que hoy llega a este feliz momento. Un fragmento muy cortó, para reconocerles todo lo han hecho.

Al profesor José R. Sosa por su encomiable labor como tutor y principal fuente de consultas e ideas para la correcta conclusión del presente trabajo. Destacamos de él su respeto, honestidad y disponibilidad para orientar a fin de llevar a feliz término el proceso de investigación.

Al profesor Iván Flores, por ser una puerta franca cuando se requirió una consulta y por ser el primero en brindar la idea de que el problema de incompletitud de datos seria resuelto utilizando Text Mining.

A todos aquellos que con sus ideas y recomendaciones tanto en los aspectos metodológicos como de fondo, buscaron ayudar y favorecer la realización de esta investigación.

A todos mi reconocimiento y eterna gratitud.

## Contenido

# Índice de Ilustraciones

# Índice de Tablas

**Acrónimos**

| IDE | Integrated Development Environment |
|---|---|
| HTTP | Hypertext Transfer Protocol |
| URL | Uniform Resource Locator |
| GB | Gigabyte |
| MB | Megabyte |
| CSV | Comma Separated Values |
| XML | EXtensible Markup Language |
| DC | Dublim Core |
| LOM | Learning Object Metadata |
| BAV | Buscador Académico Venezolano |
| IEEE | Institute of Electrical and Electronics Engineers |
| W3 | World Wide Web Consortium |
| UCV | Universidad Central de Venezuela |
| SMBD | Sistema Manejador de Bases de Datos |
| IDE | Interfaz de Desarrollo |
| TDM | Term Document Matrix |
| DTM | Document Term Matrix |
| OAI-PMH | The Open Archives Initiative Protocol for Metadata Harvesting |
| WAR | Web Application Archive |
| RS | Sistemas de recomendación |
| RDI | Repositorio Digital Institucional |

# INTRODUCCIÓN

Los repositorios digitales institucionales, como el Buscador Académico Venezolano, cumplen la misma función que cualquier biblioteca: una institución cuya finalidad consiste en la adquisición, conservación, estudio y exposición de libros y documentos (Real Academia Española, 2014). Estos repositorios contribuyen en el proceso de enseñanza y difusión del conocimiento, poniendo al alcance recursos educativos que de otra manera seria muy difícil de conseguir.

Sin embargo, en la medida que estos repositorios crecen, al igual que en las bibliotecas, a no ser que el usuario sepa exactamente que está buscando, encontrar recursos relacionados con un área o tema específico puede tornarse complicado. Presentarle a un estudioso, recursos académicos relacionados con aquellos asuntos de su interés, sin duda simplificaría el proceso de búsqueda de información. Para estas situaciones existen los sistemas de recomendación. Estos sistemas, basándose en las preferencias del usuario, le brindan sugerencias oportunas.

En el capítulo 1 de este trabajo se presenta como problemática, la necesidad de implementar un sistema de recomendación basado en contenidos para el Buscador Académico Venezolano. En el segundo se hace un repaso de todos conceptos claves para la correcta lectura de este trabajo, tanto en el área de los repositorios digitales institucionales, como en el uso de métodos y herramientas que permitirán responder las dudas planteadas en el capitulo anterior. En el capítulo 3 se explica la metodología CRISP-DM, empleada durante la investigación la cual será desarrollada en el capítulo 4, a través de un proceso de text mining (minería de texto), para finalmente presentar las conclusiones del trabajo en el capítulo 5.

# CAPÍTULO 1
# EL PROBLEMA

En este capítulo, se presenta la propuesta del trabajo especial de grado, que aborda el tema de los sistemas de recomendación basados en contenido y su relación con los recursos digitales. Se explica, adicionalmente, una solución que atiende unos objetivos determinados para solucionar dicho problema.

## 1.1. Planteamiento del Problema

Los repositorios digitales institucionales son un "sistema de almacenamiento accesible desde la red en la que los objetos digitales pueden ser almacenados para un posible acceso o recuperación posterior", de acuerdo a Kahn y Wilensky (2006, citados por Palavitsinis, 2013). Estos accesos y recuperaciones son posibles gracias a la metadata presente.

La metadata es un conjunto de datos que describe a un objeto. La estructura de los registros de metadata apunta a facilitar el manejo, descubrimiento, y captura de los recursos que ella describe (Palavitsinis, 2013, pág. 22 cita a Al-Khalifay Davis, 2006). Adicionalmente, la existencia de metadata sobre los objetos de aprendizaje permite, a los usuarios potenciales, conocer mejor los recursos sin tener que examinarlos (Palavitsinis, 2013, pág. 22 cita a Haase, 2004).

Gracias a la metadata es posible hacer búsquedas efectivas sobre los recursos así como encontrar descripciones de los mismos. Eso ayuda a los repositorios digitales institucionales a cumplir su principal función, sin embargo si algún repositorio quiere aportar facilidades a sus usuarios, sugiriéndoles recursos que pudieran ser de su interés, se tendría que apoyar en un sistema de recomendación.

Los sistemas de recomendación brindan sugerencias de recursos o ítems de acuerdo a una determinada política. Ellos pueden sugerir algún recurso de acuerdo a que tan similar es este con respecto al que se buscaba originalmente u algún otro esquema de política.

Para un repositorio digital institucional que no almacena las búsquedas de sus usuarios, ni les exige algún tipo de registro, la mejor opción sería un sistema de recomendación basado en contenidos. Este tipo de sistema de recomendación tiene como política hacer sugerencia de recursos en función de que tan similares son estos al recurso original o al que use como modelo.

En este trabajo, estudiaremos el desarrollo de un sistema de recomendación de recursos académicos para el repositorio Buscador Académico Venezolano, haciendo uso de conceptos y técnicas de minería de datos, específicamente text mining sobre las descripciones de los recursos.

## 1.2. Objetivos

El objeto General de este trabajo es el desarrollo de un sistema de recomendación de recursos para el Buscador Académico Venezolano, en el entendido de que este sistema de recomendación está basado en contenidos.

Para lograr este objetivo principal, muchas fases que representan obstáculos y caminos obligatorios tienen que cumplirse primero, muchas preguntas tiene que ser contestadas, como por ejemplo: ¿de dónde y cómo se va a obtener la data para el estudio?, ¿Cuál es el contexto de esta data?, ¿Qué significa cada pieza de datos?, ¿estos datos están listos para ser procesados? De no ser así ¿Qué modificaciones hay que hacer sobre los datos para que puedan ser procesados?, ¿Qué información o patrones existen ocultos dentro de la data? ¿Qué relaciones se pueden descubrir dentro de la data? , ¿Los patrones encontrados, responden a las necesidades del proyecto?, ¿Cómo presentar los resultados, de forma que sea fáciles de entender para cualquier usuario?.

Para responder a estas interrogantes y en consecuencia resolver el objetivo principal, se tiene que atender varios objetivos específicos, pertenecientes a la metodología CRISP-DM (que se describe en Capitulo 3), como son:

- Recuperar la metadata de la base de datos MySQL proporcionada
- Entender el contexto del problema
- Entender y estudiar de los metadatos para encontrar patrones visibles y formular hipótesis de posibles soluciones
- Preparar los datos para ser consumidos por un método de minería de datos determinado
- Modelar los datos con el objetivo de encontrar patrones ocultos que puedan ser de utilidad
- Evaluación del modelo de la fase anterior para corroborar que satisfacen las necesidades del proyecto
- Desplegar los datos de forma tal que sean fácilmente entendibles y verificables por un usuario del negocio

## 1.3. Justificación

El beneficio de este trabajo, es que se proporciona una herramienta que sirve para hacer recomendaciones de recursos basándose en contenidos utilizando el text mining. El potencial beneficiario de este trabajo es el repositorio “Buscador Académico Venezolano”.

Con eso se asienta un primer sistema de recomendación, que a futuro puede evolucionar o permanecer como esta, quedando el presente documento como guía o referencia para la elaboración de sistemas

de recomendación similares o para estudios de text mining el cual será la principal técnica de minería de datos a utilizar.

La presencia de este sistema en el repositorio objetivo, haría que el mismo le de mayor difusión a sus recursos, causando como consecuencia una más amplia y seguramente mejor transmisión de conocimiento.

Los usuarios del repositorio, dependerían mucho menos del buscador, y se simplificaría su paso por el portal, dado que en menos ocasiones se verán en la necesidad de refinar o cambiar las palabras clave con que hacer búsquedas, dado que al encontrar un recurso de su interés, el sistema de recomendación le indicara los otros recursos similares a este.

## 1.4. Antecedentes

Varios autores ya han investigado y estudiado los aspectos de la metadata y los repositorios digitales institucionales en general. Muchos de sus trabajos sirven para comprender mejor todo lo relevante a los sistemas de recomendación. A continuación, en varios puntos, se presentan estudios y esfuerzos previos en el área de los Repositorios digitales institucionales y los sistemas de recomendación

El articulo "Recommendation in repositories of learning Objects: a proactive approach that exploits diversity and navigation-by-proposing" (2009) de Almuadena Ruiz, Guillermo Jimenez-Diaz, Mercedes Gomez-Albarran y Jose Garcia Santesmases se relaciona con este trabajo, porque también estudiaron la implementación de un sistema de recomendación en un Repositorio Digital Institucional. En ese trabajo, la situación problemática era que el buscador podía no dar como primera opción un recurso del interés del estudiante, puesto que el mismo podría no estar al nivel de complejidad esperado por el estudiante, y esto ocasionaría que el usuario buscara otra solución. Para atender este problema, se diseño un sistema de recomendación que considerando las similitudes entre recursos y estudiantes brindara un sistema hibrido de recomendación de unos pocos pero certeros recursos.

El articulo "Personalización en Recomendadores Basados en Contenido y su Aplicación a Repositorios de Objetos de Aprendizaje" (2010) de A. Ruiz-Iniesta, G. Jiménez-Díaz y M. Gómez-Albarrán también trata sobre el tema y además plantea mejoras sobre un sistema de recomendación muy parecido al que se hace en este trabajo. El sistema descrito por estos autores estaba basado en contenidos, como lo es el de esta investigación, y el objetivo de aquel trabajo era la propuesta de añadir información sobre la utilidad pedagógica a largo término de los recursos de aprendizaje, de forma que si dos estudiantes con objetivos a largo plazo diferentes hacían una misma consulta, obtuvieran resultados diferentes dependiendo del objetivo pedagógico de cada uno.

Residencias Pineca, piso 8, apartamento 84, esquina paraíso a pineda detrás del palacio blanco de miraflores

La tesis de maestría de Almudena Ruiz Iniesta titulada "Estrategias de recomendación aplicadas a repositorios de recursos educativos" (2009) toca esta temática al hacer un estudio del estado del arte en los sistemas de recomendación y los repositorios de objetos de aprendizaje. Su estudio al igual que la presente investigacion se dedica a atender el mismo problema: "como darle difusión a recursos más cercanos al interés del usuario, cuando hay muchos recursos disponibles y el buscador no es la solución optima". En ese estudio se incluyen a los sistemas de recomendación basados en contenidos, y también se incluyen a los objetos de aprendizaje. El objetivo de aquel estudio era proponer mejoras sobre los sistemas de recomendación de los RDI, y dar algunas sugerencias.

## 1.5. Fuerzas

Este trabajo tiene relevancia actual, puesto que la temática que aborda es de interés para las áreas de estudio, desarrollo e investigación científica, especialmente en el e-learning. El trabajo presentado se engloba precisamente dentro de esta reciente línea de investigación que aporta técnicas de recomendación al ámbito del e-learning.

### 1.5.1. ONCTI

Citando textualmente a su propio portal, el Observatorio Nacional de Ciencia, Tecnología e Innovación (ONCTI) es una institución adscrita al Ministerio del Poder Popular para la Educación Universitaria, Ciencia y Tecnología (MPPEUCT) de la República Bolivariana de Venezuela, cuya función principal es recopilar, sistematizar, categorizar, analizar e interpretar información con el fin de contribuir en la definición de las políticas públicas que promuevan y fortalezcan el desarrollo científico-tecnológico, impactando económica y socialmente sobre la soberanía de la nación.

Esta institución venezolana, está presente en los esfuerzos para mejorar e impulsar las herramientas educativas disponibles en el país. Entre estos esfuerzos está el de "LAReferencia". LAReferencia es un proyecto internacional de las naciones latinoamericanas por reunir a sus repositorios digitales institucionales en un súper repositorio latinoamericano, en la que se puede ver el portal del buscador de LAReferencia.


**Ilustración 1 - Proyecto LAReferencia**

El 29 de noviembre de 2012 se firma el Acuerdo de Buenos Aires, mediante el cual, varias naciones latinoamericanas (entre ellas Venezuela) acuerdan definir políticas, estándares y servicios comunes para la interoperabilidad de sus repositorios de forma que puedan compartir sus recursos científicos al esfuerzo mancomunado.

Teniendo esta meta, muchas instituciones venezolanas se han plegado al esfuerzo de crear repositorios digitales institucionales que sean interoperables para lograr compartir sus recursos. Muchas instituciones como la Universidad Central de Venezuela (UCV) ya contaban con su propio repositorio institucional (Saber-UCV desde el 2011) sin embargo aun queda la labor de reunir las instituciones venezolanas en un repositorio nacional, así como incrementar la interoperabilidad de cada repositorio para compartir recursos con LAReferencia.

Actualmente el ONCTI está diseñando un repositorio federado nacional que agrupe a los repositorios venezolanos, como parte del esfuerzo por darle difusión e interoperabilidad a los repositorios y recursos venezolanos. Este esfuerzo corresponde al portal Sinamaica, el cual utilizara el motor del Buscador Académico Venezolano como sistema de búsquedas.

### 1.5.2. Repositorios Digitales

Kahn y Wilensky (2006, citados por Palavitsinis, 2013) definen un repositorio digital como un " sistema de almacenamiento accesible desde la red en la que los objetos digitales pueden ser almacenados para un posible acceso o recuperación posterior. El repositorio cuenta con mecanismos para añadir nuevos objetos digitales a su colección (depósito) y para su puesta a disposición (acceso), utilizando, como mínimo, el protocolo de acceso al repositorio. El repositorio puede contener otra información relacionada, servicios y sistemas de gestión”

Estos RDI son el equivalente a las bibliotecas, donde usuarios depositan recursos que son accesibles por otras personas. Esta versión moderna de las bibliotecas nace con el objetivo de darle difusión a los recursos que ellas almacenan, cosa que en el caso del conocimiento y la educación representa un avance fundamental, en la exposición de materiales útiles. Un ejemplo de Repositorio digital podría ser el de la Biblioteca Central de la UCV (Ilustración 2 - Saber UCV), el Repositorio Saber-UCV o Saber-Universidad de Los Andes o Esopo de la Universidad Simón Bolívar, entre otros.

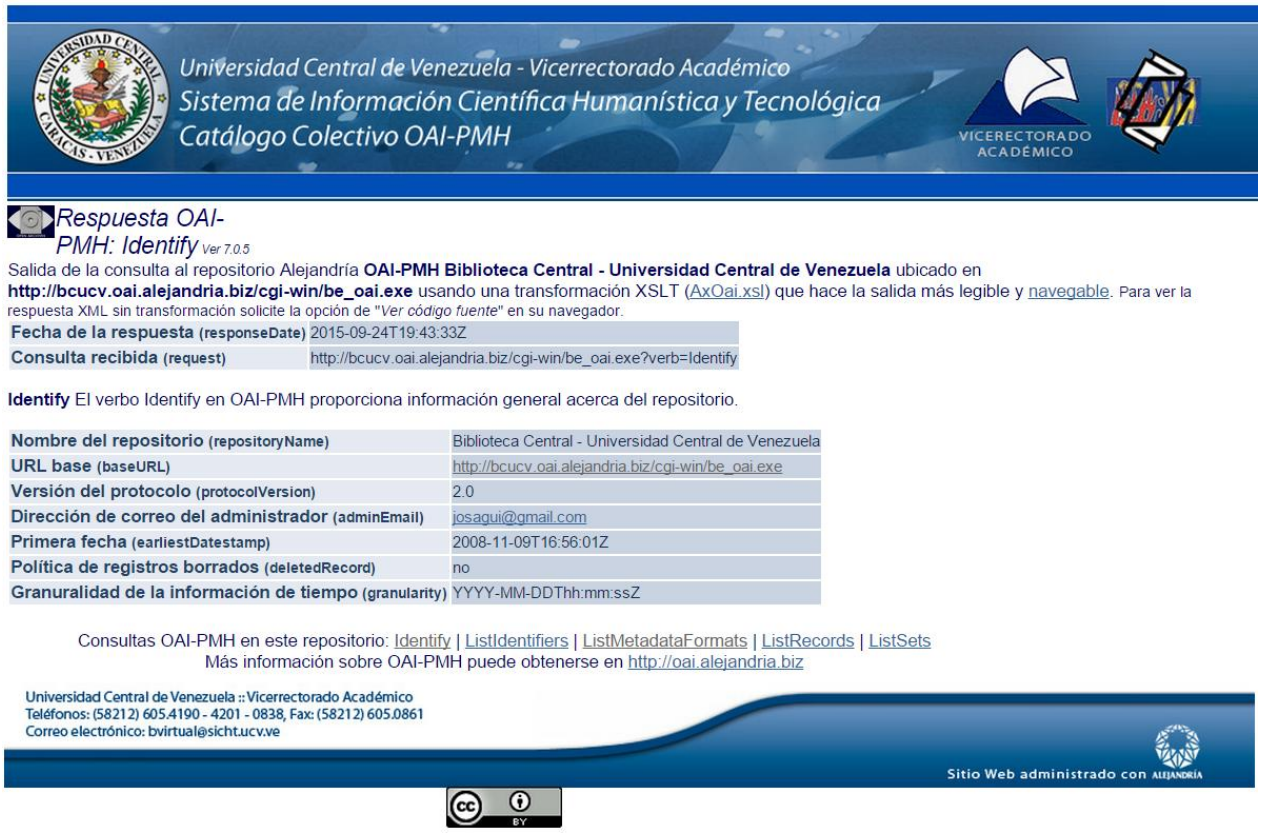


**Ilustración 2 - Saber UCV**

Los repositorios digitales permiten, el almacenaje histórico de documentos y recursos, lo que contribuye positivamente con la preservación del conocimiento, además de ser puntos de difusión para los datos en el almacenados. Estas características hacen que los RPI, al igual que las bibliotecas, sean elementos clave de toda institución académica, permitiendo atravesar las barreras del tiempo y espacio para darle acceso al público usuario a los conocimientos.

### 1.5.3. Sistemas de Recomendación

Antes de los sistemas de recomendación computarizados, se consultaba a expertos o amigos para obtener sugerencias sobre los productos de inquietud. Estos al igual que los sistemas basados en conocimientos o los filtrados colaborativos (véase el capítulo 2.1 de este trabajo) brindaban recomendaciones en función de sus experticias o vivencias previas.

Iniesta (2011) indica que los buscadores no son la solución a la necesidad de sugerencias, ni sustitutos a los sistemas de recomendación, puesto que exigen que se formule muy bien una pregunta. La necesidad de saber a priori que se está buscando dificulta encontrar muchas posibles soluciones, puesto que el usuario no siempre sabe de antemano que es lo que prefiere o necesita ("*Muchas veces, la gente no sabe lo que quiere hasta que se lo enseñas*", Steve Jobs).

Asimismo los buscadores no conocen a sus usuarios. Esto hace que las sugerencias y recomendaciones no sean personalizadas, cosa que es un problema, dado que las sugerencias tienen que ir acordes a las necesidades y gustos del individuo. No es lo mismo indicar que, recursos en su contenido tienen la palabra "minería" que mostrar los artículos que son parecidos a un trabajo de "minería de datos" en el cual el usuario ya ha mostrado interés.

La necesidad de personalizar las sugerencias se puede describir con un ejemplo de la vida real: "Touching The Void" e "Into Thin Air" son dos libros sobre dramas en el alpinismo, pero escritos uno primero que el otro por una separación de 10 años. El primero, cuando fue publicado no logro éxito en las ventas a pesar que fue bien visto por evaluadores. Sin embargo, el segundo cuando salió al mercado fue un éxito. Lo curioso de este caso es que una vez el segundo salió al mercado y empezó a ganar popularidad, el primero comenzó un repunte de ventas, al punto que llego a ser un bestseller e incluso a sobrepasar en ventas al segundo.

¿Qué relación podría haber entre los libros para que el éxito comercial de uno, reviviera e impulsara al otro? El sistema de recomendaciones de Amazon. Este sistema detecto que ambos libros eran similares, y que por lo tanto los usuarios que gustaron de "Into Thin Air" posiblemente también gustarían de "Touching The Void", el sistema estaba en lo correcto.

La mala difusión de un recurso puede convertirlo en un recurso invisible para el usuario, también conocido como recurso oscuro (libros oscuros, páginas web oscuras). Los sistemas de recomendación han ayudado a "desenterrar" y mostrar estos recursos, que de otra forma no hubiesen sido conocidos.

En algunos casos lo que el usuario mas apreciaría son exactamente estos recursos oscuros, ósea aquellos recursos que se parecen más a lo que ellos mediante un proceso de búsqueda finalmente lograron calificar de su gusto.

Hasta los momentos no se conoce ningún repositorio venezolano, que brinde entre sus bondades algún sistema de recomendación.

## 1.6. Alcance

La minería de datos es el área de la computación dedicada a encontrar relaciones aun no descubiertas en un repositorio. Para este trabajo se planea hacer un Text Mining (Minería de datos aplicada a textos) sobre el título, descripción y palabras clave de las instancias de metadata, para determinar su pertenencia a comunidades, que no son más que temas o áreas de estudio.

Es importante mencionar, que se puede hacer minería de datos sobre los datos proporcionados, y no sobre los datos no proporcionados. Esto quiere decir, por ejemplo que es posible imaginarse que existan ciertas comunidades, como una comunidad dedicada a las artes, pero si en el repositorio de objetos de aprendizaje objetivo no existen instancias de dicha comunidad; el proceso de minería de datos no dará como resultado la existencia de una comunidad dedicada al arte, puesto que el proceso de minería de datos omitirá toda la información que no le sea proporcionada.

Entonces este trabajo estaría limitado a unos metadatos, pertenecientes a un RDI, los cuales serán utilizados para un proceso de minería de datos, de cuyo resultado se dará el modelo de un sistema que permitirá asociar para cualquier instancia las que le sean similares.

Estas instancias parecidas serán las recomendaciones que brindara el sistema y como se basan en la metadata de las diferentes instancias, el sistema estará basado en contenidos.

# CAPÍTULO 2
# MARCO CONCEPTUAL

En este capítulo se hará repaso de algunas las definiciones clave para el entendimiento del tema. Asimismo investigaciones, estándares y demás trabajos que varios autores han conducido, darán una perspectiva más amplia del estado, y utilización de los conceptos antes mencionados.

## 2.1. Sistema de recomendación

"Los recomendadores (RS) son sistemas que ayudan a emparejar usuarios con productos" (Iniesta, 2011). Estos sistemas son usados para sugerir e ilustrar recursos/productos que puedan ser de interés del usuario. Los sistemas de recomendación necesitan un mínimo de información del usuario para poderle sugerir algo, que le interesa, que le ha interesado, a quien se parece el usuario.

Los sistemas de recomendación disponen de varias técnicas para lograr sus sugerencias, algunas de estas técnicas son: Filtrado Colaborativo, Basadas en Contenido, Basadas en Conocimiento y Sistemas Híbridos.

### 2.1.1. Filtrado Colaborativo

En esta técnica se buscan usuarios similares de forma tal de proponer productos similares, dicho de otra manera, si un usuario A se parece a uno B, y el B ha declarado que gusto de un producto que el A aun no ha comentado, se le puede proponer a A ese producto que su usuario similar ya ha disfrutado. Como indico Iniesta (2011) "Enséñame lo que es popular entre mis vecinos".

Estos sistemas requieren saber: características de los usuarios, que le gusta a cada usuario y que tan parecidos son los usuarios.

### 2.1.2. Basado en Contenidos

Como Almudena Ruiz Iniesta (2011) declaraba los sistemas de recomendación basados en contenido le dan sugerencias a sus usuarios chequeando la similitud del los recursos que el usuario ha gustado con la de los recursos que pretende sugerir, o como indica Iniesta "muéstrame más de lo que me ha gustado".

Estos sistemas requieren saber: que es lo que le gusta al usuario, y que tan similares es un recurso con respecto a los demás.

### 2.1.3. Basado en Conocimiento

En esta técnica, se diseñan sistemas expertos adaptados a un dominio específico que pueden responder mejor consultas simples de la que se dispone poca información.

Estos sistemas requieren saber: que es lo que le gusta al usuario, y un modelo de conocimientos aplicado al dominio de aquellas cosas que van a recomendar

### 2.1.4. Híbridos

Los sistemas híbridos, como su nombre lo dice, son aquellos que juntan o mezclan las partes de otras técnicas de recomendación para lograr mejores resultados. Las necesidades de estos sistemas dependerán del sistema de recomendación en específico.

## 2.2. Objeto Digital

Un objeto Digital es una instancia de un tipo de dato abstracto que tiene 2 componentes, el dato y la metadata (descripción) (Palavitsinis 2013, pág. 19 cita a KahnyWilensky, 2006). Hay que reconocer la importancia de la metadata, puesto que un archivo que está siendo creando, es un estado de ejecución, más no un objeto digital. Por ejemplo, un documento de un editor de texto no se le identifica como documento hasta que dicho documento es guardado, al guardarlo, se le asocia un autor, un título, una fecha de modificación, un formato de archivo, etc.


Ilustración 3 - Descripción gráfica de componentes de un objeto de aprendizaje

El Contenido de un objeto digital y sus metadatos forman una unidad o dicho de otra manera, un contenido sin metadatos o unos metadatos sin contenidos, no son un objeto de aprendizaje, dicho objeto se forma por la mezcla de ambos. Como la Figura 1 muestra, por lo general el tamaño digital (bytes) de los metadatos es inferior (muy inferior) al tamaño del contenido.

### 2.2.1. Objeto de Aprendizaje

Un objeto de aprendizaje también llamado objeto educativo es "cualquier entidad, digital o no, que pueda ser utilizada para el aprendizaje, la educación o la capacitación" (*Institute of Electrical and*

*Electronics Engineers* (IEEE), 2002, pág 45). Tales objetos podrían ser presentaciones, tareas, informes, artículos, imágenes, audios, videos, tablas de información. Esta definición ha sido criticada por ser muy amplia, sin embargo el Comité de Estandarización de Tecnologías Educativas (patrocinador del estándar de la IEEE (2002) sobre el Estándar de Metadatos para objetos Educativos) ha argumentado, que esa era la idea (tener una definición que abarcará cantidad de objetos).

Otros autores como (Chiappe,2007) hacen foco en la estructura de los objetos de aprendizaje: "Una entidad digital reutilizable y auto-contenida, con un claro propósito educativo, con al menos tres componentes internos editables: contenido, actividades de aprendizaje y elementos de contexto. Los objetos de aprendizaje deben tener una estructura externa de información para facilitar su almacenaje y recuperación: los metadatos.", en donde deja en claro que un objeto de Aprendizaje sin metadatos no sería una definición válida de objeto de aprendizaje.

Wayne Hodgins comenzó el uso del término objeto de aprendizaje en 1992, como nombre para las piezas de aprendizaje que fueran fácilmente reutilizables y ajustables a otras piezas de aprendizaje. La idea de acuñar estos "objetos de aprendizajes" surgió como resultado de reflexionar sobre estrategias de aprendizajes al tiempo que observaba a uno de sus hijos jugar con piezas de LEGO. Recordemos que LEGO es la marca del juego de los bloquecitos de plástico de construcción.

Se puede extraer que un objeto de aprendizaje es un objeto digital, reutilizable, auto-contenido, con propósitos educativos, que debe tener un contexto, un contenido y unos metadatos, estos objetos de aprendizaje, deben poder combinarse con otros objetos de aprendizaje, permitiendo así crear objetos mayores o más complejos.

Autores como Sánchez y Sicilia (2005) también indican que estas combinaciones tienen ciertas reglas, puesto que no se puede presumir que todos los objetos de aprendizaje son combinables con todos los demás. Estas combinaciones deben, cuando menos, establecer unas condiciones previas antes de poder combinarse. Esto es fácilmente ejemplificable para cualquier caso donde se intente combinar un curso básico con uno avanzado y específico, donde la brecha entre el básico y avanzado es tan grande, que el alumno no podría entender del todo la parte avanzada del curso por falta de recursos previos.

## 2.3. Metadatos

Los datos de los datos son conocidos como los Metadatos**.** Son ese conjunto de datos que describen a un objeto. La estructura de los registros de metadata apunta a facilitar el manejo, descubrimiento, y captura de los recursos que ellas describen (Palavitsinis 2013, pág. 22 cita a Al-Khalifay Davis, 2006). Adicionalmente, la existencia de metadata sobre los recursos permite, a los usuarios potenciales, conocer (mejor) a los recursos sin tener que examinarlos (Palavitsinis 2013, pág. 22 cita a Haase, 2004). En la Tabla 1 se puede ver un ejemplo de metadata.

**Tabla 1- Ejemplo de Calidad en metadata**

| | |
|---|---|
| Título | Introducción a Java |
| Autor | IsraelRodríguez |
| Año publicación | 2008 |
| Autor última modificación | IsraelRodríguez |
| Año última modificación | 2015 |
| Tema | Programación |
| Descripción | Curso introductorio básico de programación en Java |
| Palabras clave | Java,programación,básico, principiante |
| Idioma | Español |

Sánchez y Sicilia (2005) afirman que los metadatos buscan brindar información, propiedades y estados de los recursos sin tener que accederlos. Estos metadatos son externos al contenido del documento, se expresan en lenguaje técnico, por ejemplo XML, siguen estándares y normalizaciones para darles interoperabilidad entre sistemas, y dan información general sobre el objeto de aprendizaje.

### 2.3.1. Estándares de Metadata

En este sub capítulo se explicarán los 2 estándares de metadata más ampliamente aceptados el LOM y el Dublin Core (DC), con 50 y 15 elementos respectivamente son estándares diferentes, que buscan proporcionar una descripción confiable y útil de los objetos de aprendizaje.

### *2.3.1.1. IEEE - Learning Object Metadata (LOM)*

El estándar IEEE 1484.12.1:2002 de metadatos para objetos de aprendizaje (*LearningObjectMetadata* - **LOM**) para la descripción de recursos de aprendizaje fue publicado en 2002 por la IEEE. Se basa en primerizos esquemas de metadatos desarrollados por la ARIADNE *Foundation* y el IMS *Global LearningConsortium* (Ochoa. et al., 2011).

El objetivo principal de LOM es " facilitar la búsqueda, evaluación, adquisición y uso de objetos de aprendizaje, por ejemplo, estudiantes, instructores o procesos de software automatizados" a través de la descripción de un conjunto común de elementos de metadatos (Ochoa. et al., 2011).

El LOM comprende una jerarquía de elementos. En el primer nivel, hay nueve categorías, cada una de las cuales contiene sub-elementos; estos sub-elementos pueden ser elementos simples que contienen datos, o pueden ser ellos mismos los elementos agregados, que contienen más subelementos(IEEE, 2002)

LOM propone alrededor de 50 elementos de metadatos diferentes, agrupados en nueve categorías: General, ciclo de vida, Meta-Metadatos, técnicos, educacionales, derechos, relaciones, anotación y Clasificación(Ochoa. et al., 2011).Estos elementos se representan en forma grafica en la figura 2.

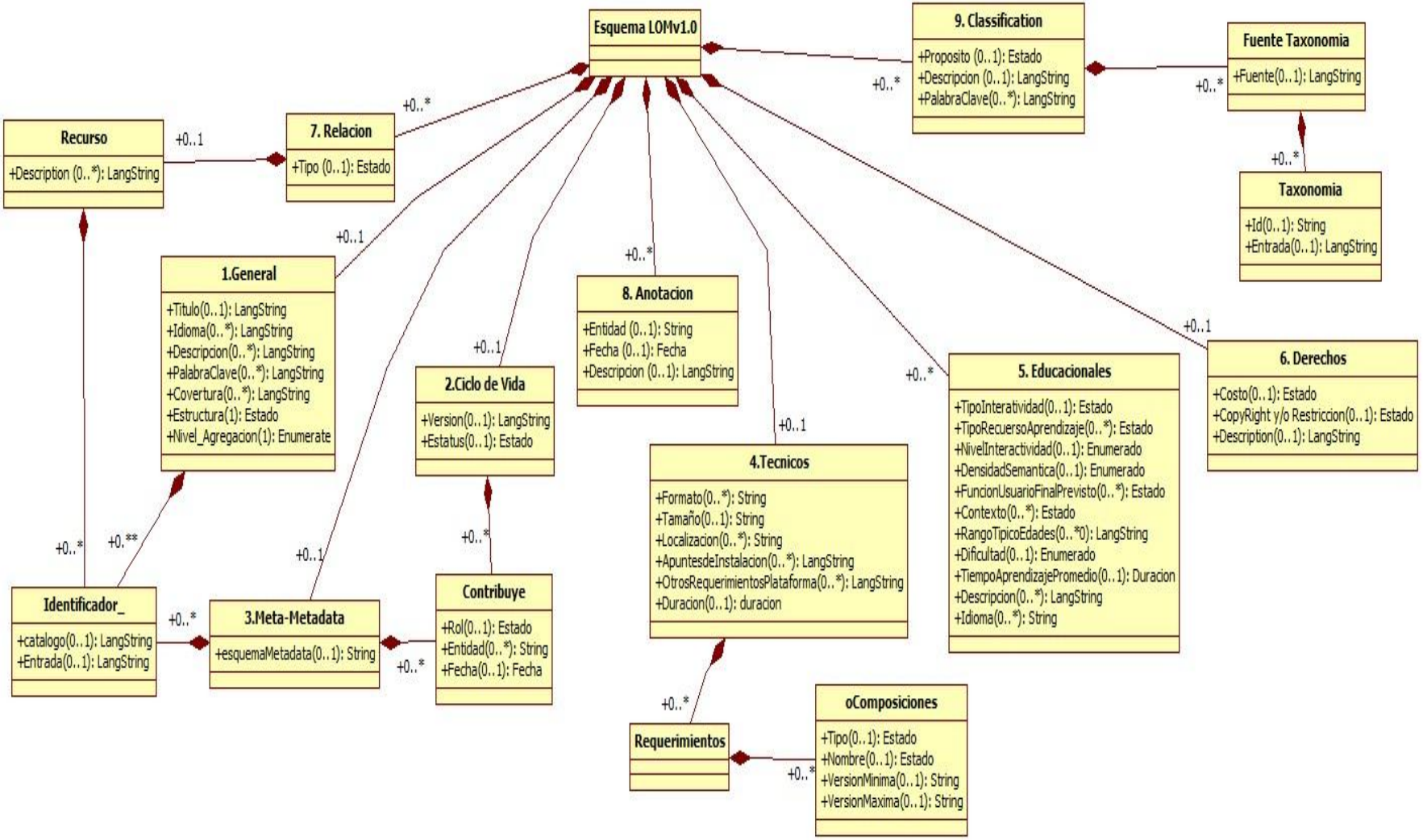


**Ilustración 4 - Representación esquemática de la jerarquía de los elementos de LOM.**

A continuación se explican las categorías del estándar de metadatos LOM, basándose en el trabajo de Paola A Piñero G y Loy E Ramírez P "Construcción de un objeto de aprendizaje Web tipo Simulación Sísmica utilizando tecnologías de dibujo en capas con HTML5" y la traducción del *Institute of Electrical and ElectronicsEngineers* (IEEE) "Estándar para Metadatos de Objetos Educativos".

**Categoría general:** indica información general que describe el objeto de aprendizaje de manera general. (IEEE, 2002)

- *Identificador:* Valor unívoco que permite discriminar el material en su contexto educativo.
- *Título:* nombre del material
- *Entrada en catálogo:*este metadato puede especificarse para seleccionar el recurso cuando este se encuentra indexado en un repositorio.
- *Idioma*: lengua utilizada de manera principal en el documento.
- *Descripción*: texto describiendo el contenido del recurso
- *Palabra clave*: palabra o frase que es representativa del material.
- *Cobertura*: eventos temporales, culturales, o geográficos asociados con el material.
- *Estructura*: define la estructura interna del material. LOM define el siguiente vocabulario para describir la estructura: colección, mixta, lineal, jerárquica, en red, ramificada, atómica. No obstante los autores pueden usar sus propios vocabularios, adaptados a sus necesidades pedagógicas particulares.
- *Nivel de Agregación*: define la granularidad del material.

**Categoría Ciclo de Vida:** características relacionadas con la historia y el estado actual del objeto, así como aquellas que lo han modificado durante su ciclo de vida. (IEEE, 2002)

- *Versión*: edición o versión del recurso
- *Estado*: estado de producción del recurso
- *Contribución*: información del contribuidor de la versión del material

**Categoría meta - Metadatos:** información sobre la instancia de metadatos (no incluye información sobre el objeto al cual la metadata describe) (IEEE, 2002)

- *Identificador*: en este caso el identificador es del conjunto de metadatos del recurso.
- *Catalogo de entrada*: el conjunto de metadatos reside en esta sección
- *Contribuyente*: los creadores de estos metadatos.
- *Esquema de metadatos*: es el patrón de metadatos utilizado, por ejemplo LOMv1.0.
- *Idioma*: la lengua en la que se proporcionan los metadatos.

**Categoría Técnica:** información sobre requerimientos y características técnicas del objeto. (IEEE, 2002)

- *Formato*: indica la codificación del documento (HTML, JPG, entre otros), puede que integre más de un formato (HTML junto con JPG).
- *Tamaño*: dimensión en bytes del recurso
- *Localización*: indica cómo encontrar el recurso, por ejemplo una URL.
- *Requisitos*: requisitos del sistema plataforma para hacer uso del recurso. Los requisitos a su vez están compuestos por:
    - *Tipo de plataforma*: de acuerdo a LOM puede ser browser o Sistema Operativo.
    - *Nombre de plataforma*: nombre del tipo de plataforma (ejemplo: Windows)
    - *Indicación de instalación*: notas para la instalación del recurso
    - *Otros requisitos de plataforma*: adicionales de los requisitos.

**Categoría uso Educativo:** información educativa y pedagógica del objeto. (IEEE, 2002)

- *Tipo de interacción:* LOM propone el siguiente vocabulario para caracterizar este tipo de interacción: activa para los contenidos interactivos, expositivo para contenidos pasivos, mixto para contenidos que comparten ambas características, o indefinido para contenidos para los que no procede especificar el tipo de interacción.
- *Tipo de recurso Educativo:* especifica el tipo de recurso
- *Nivel de interacción:* LOM propone el siguiente vocabulario: muy bajo, medio, alto o muy alto

- *Densidad Semántica:* medida subjetiva de la relación entre la utilidad del recurso contra su tamaño o duración. LOM propone el siguiente vocabulario: muy bajo, medio, alto o muy alto
- *Función Usuario Final Previsto:* indica el rol del usuario. por ejemplo: alumno, profesor, autor
- *Contexto*: entorno esperado donde se usará el material.
- *Rango típico Edades:* edad esperada de los usuarios alos que va dirigidoel recurso.
- *Dificultad*: Complicación de aprendizaje inherente al recurso. LOM propone el siguiente vocabulario para caracterizar dicho grado: muy fácil, fácil, medio, difícil, muy difícil.
- *Tiempo Aprendizaje Promedio:* media de la duración para el aprendizaje del recurso.
- *Descripción*: comentarios sobre el uso del material desde un punto de vista pedagógico.
- *Idioma*: lengua en la que está desarrollado el recurso.

**Categoría Derechos:** indica propiedad intelectual y condiciones de uso del objeto (IEEE, 2002)

- *Costo*: establece si el recurso es pago o no.
- *Derechos de copia y otras restricciones*: indica el estado de derecho de copias y otras restricciones
- *Descripción*: comentarios sobre las condiciones y derechos de autor de este recurso.

**Categoría relación:** brinda información sobre las relaciones del objeto de aprendizaje con otros objetos de aprendizaje (IEEE, 2002) (hágase énfasis en el punto 2.0.1 Objeto de Aprendizaje de este trabajo, en donde se comenta la influencia de LEGO en la definición de los objetos de aprendizaje)

- *La clase de la relación:* el recurso puede ser parte de otro más complejo, o tener a otro como parte integrante, o ser una versión de otro, puede referir a otro, es la base de otro, etc.

**Categoría Anotación:** permite incluir comentarios al objeto, así como información de quien y cuando creo dichos comentarios (IEEE, 2002).

- *Entidad*: Quien hizo la anotación
- *Fecha*: ocasión en la que se realizó la anotación
- *Descripción*: texto de la anotación.

**Categoría Clasificación:** describe el objeto de aprendizaje en relación a un determinado sistema de clasificación (IEEE, 2002).

El modelo de datos LOM especifica qué aspectos de un objeto de aprendizaje deberían ser descritos y que vocabularios pueden ser usados para estas descripciones; también define cómo este modelo de datos puede ser enmendado mediante adiciones o restricciones. Otras partes del estándar son redactadas para definir los formatos del modelo de datos LOM, por ejemplo como los registros LOM debieran ser representados en XML y en RDF (IEEE 1484.12.3 y IEEE 1484.12.4 respectivamente).(Wikipedia ,2015)

El modelo de datos específica que algunos elementos se pueden repetir individualmente o como un grupo; Por ejemplo, aunque los elementos 9.2 (Descripción) y 9.1 (Objetivo) sólo puede ocurrir una vez dentro de cada instancia del elemento contenedor de clasificación, el elemento de clasificación se puede repetir - permitiendo así que muchas descripciones para diferentes propósitos.(Wikipedia ,2015) El modelo de datos también especifica el campo "valor" y tipo de datos para cada uno de los elementos de datos simples. El campo valor define las restricciones, para cada caso, en los datos que se pueden introducir para ese elemento. Para muchos elementos, el campo valor permite cualquier cadena de caracteres Unicode que se introduzca, mientras que las entradas de otros elementos deben ser tomados de una lista específica/declarada (es decir, un vocabulario controlado) o debe estar en un formato especificado (por ejemplo, fecha e idioma códigos) (IEEE, 2002).

Algunos tipos de datos de elementos simplemente permite que una cadena de caracteres a introducir, y otros se componen de dos partes, como se describe a continuación:

LangString: contienen partes del idioma y de cadena, lo que permite la misma información que se registra en varios idiomas (IEEE, 2002).

Elementos de vocabulario: están restringidos de manera tal que sus entradas tienen que ser elegidos de una lista controlada de términos ( compuesto por pares Source-Value ) donde se utiliza la fuente que contiene el nombre de la lista de términos y el valor que contiene el término elegido (IEEE, 2002).

DateTime y duración: contienen una parte que permita la fecha o la duración que debe darse en un formato legible por máquina, y una segunda que permite una descripción de la fecha o la duración (por ejemplo, "a mediados del verano de 1968") (IEEE, 2002).

Como la IEEE se apoyó fuertemente en el IMS LearningResource Meta-data specification 1.2 para la generación de su propio estándar (el LOM), resaltamos como curiosidad que el estándar LOM y el IMS *LearningResource Meta-data Specification* 1.3 son muy parecidos, puesto que este segundo decidió acercar su nueva versión al modelo LOM. (Wikipedia ,2015)

### 2.3.2. Perfiles de Aplicación

Se conoce como perfil de aplicación o AP (por sus siglas en inglés, *ApplicationProfile*) a la combinación de los elementos de un estándar de metadatos con elementos adicionales que son considerados relevantes por el repositorio de objetos de aprendizaje (Ochoa. et al., 2011).

Por ejemplo el repositorio MACE, que atiende recursos de aprendizaje orientados hacia la arquitectura, añadió elementos como geolocalización al estándar LOM, o learningobjectking para discernir entre planos de objetos del mundo real (Ochoa. et al., 2011).

#### *2.3.2.1. Dublin Core (DC)*

El Conjunto de Elementos de Metadatos Dublin Core es un vocabulario de quince propiedades para el uso en la descripción de recursos. El nombre de "Dublín" se debe a su origen en un taller por invitación de 1995 en Dublín, Ohio; "núcleo", porque sus elementos son amplios y genéricos, que puedan utilizarse para describir una amplia gama de recursos (DCMI, 2015).

Desde 1998, cuando estos quince elementos entraron en el camino de la estandarización, las nociones de las mejores prácticas en la Web Semántica han evolucionado para incluir la asignación formal de dominios y rangos, además de las definiciones en lenguaje natural (DCMI, 2015).

Los dominios y rangos especifican qué tipo de recursos descritos y recursos de valor están asociados con una determinada propiedad. Los dominios y rangos expresan los significados implícitos en las definiciones de lenguaje natural en una forma explícita que se puede utilizar para el tratamiento automatizado de inferencias lógicas. Cuando se encuentra una propiedad dada, una aplicación de inferencia puede utilizar la información sobre los dominios y rangos asignados a una propiedad con el fin de hacer inferencias acerca de los recursos descritos en el mismo (DCMI, 2015).

Hasta la versión DC 1.1 los dominios y rangos no estaban implementados en la definición del estándar, para incluir esta nueva especificación se incluyeron como sub-propiedades, en las propiedades ya existentes y homónimas.

Las 15 propiedades (elementos) del Dublin Core se describen en la tabla 2.

Tabla 2 - Propiedades y definiciones de Dublin Core

| **Propiedad** | **Definición** |
|---|---|
| Contributor (Contribuidor) | La entidad responsable de hacer contribuciones al recurso. |
| Coverage (Cobertura) | El tema espacial o temporal del recurso, la aplicabilidad espacial del recurso, o la jurisdicción en la que el recurso es relevante. |
| Creator (Creador) | Una entidad encargada principalmente de crear el recurso. |
| Date (Fecha) | Un punto o período de tiempo asociado a un evento en el ciclo de vida del recurso. |
| Description (Descripción) | Un resumen del recurso |
| Format (Formato) | El formato de archivo, medio físico, o las dimensiones del recurso. |
| Identifier (Identificador) | Una referencia clara al recurso dentro de un contexto dado. |
| Language (Idioma) | Un lenguaje del recurso. |
| Publisher (Publicador) | La entidad responsable de hacer el recurso disponible. |
| Relation (Relación) | Un recurso relacionado. |
| Rights (Derechos) | Información de los derechos sobre el recurso y su uso. |
| Source (Fuente) | Un recurso relacionado a partir del cual se deriva el recurso descrito. |
| Subject (Tema) | El tema del recurso. |
| Title (Título) | Un nombre que se da al recurso. |
| Type (Tipo) | La naturaleza o género del recurso. |

### 2.3.3. Formas de Crear Metadata

La metadata es creada en el proceso de indexación, por los indexadores. Estos últimos pueden ser humanos, expertos en el dominio, o expertos indexadores, o máquinas que de acuerdo a algunas políticas o algoritmos determinan cómo completar la metadata.

Existen 3 formas principales de crear metadata, de forma automática, manual, y semiautomática (Ochoa et al., 2011).La forma automática implica que no hay intervención humana en el proceso de indexación, la manual que sólo los humanos crean la metadata y la semiautomática que algunos campos son rellenados por indexadores humanos y otros por la máquina (ideal para elementos como tipo de archivo o tamaño de archivo).

La mayor parte de la metadata es creada de forma manual, mientras una minoría de forma semiautomática y automática (Ochoa et al., 2011).

### 2.3.4. Calidad de la Metadata

Se puede definir a la Calidad de la Metadata como aptitud para el propósito ("*fitnessforpurpose*"). De acuerdo a Vuorikari R., Manouselis N., Duval E. (2008) hay dos ramas de investigación relacionada a la calidad de la metadata. La "*calidad de la metadata*" y la "*calidad en la metadata*". La primera se encarga de encontrar formas de evaluar y garantizar la calidad de las instancias de metadata, y la segunda de encontrar formas de representar información acerca de la calidad del recurso, usando su metadata.

La "*calidad de los metadatos*" viene dada por su precisión, completitud, proveniencia, conformidad con las expectativas, consistencia, actualidad/relevancia y accesibilidad (Bruce y Hillmann, 2004 citado por Ochoa, 2009)

A continuación, y gracias al trabajo de Ochoa (2009), se explica en detalle las 7 dimensiones de la estructura/modelo Bruce y Hillman:

**Completitud**: la medida de utilización de campos con respecto al total de campos disponibles. Una instancia de metadatos debe describir el recurso lo más posible. Asimismo, los campos de metadatos disponibles deben estar rellenos en la mayor parte posible de los objetos que representan, con el fin de que sean útiles para cualquier tipo de servicio. Aunque la completitud, por si sola, no refleje necesariamente la presencia de información útil en la metadata, se puede utilizar para medir la cantidad de información disponible sobre el objeto.

**Precisión**: la medida de correctitud en la información proporcionada acerca de un recurso. Los errores tipográficos, así como las equivocaciones, afectan esta dimensión de la calidad. Sin embargo, no todos los valores son "correcto" / "incorrecto". Hay campos de metadatos que deben recibir un juicio más subjetivo. Por ejemplo, si bien es fácil determinar si el tamaño o el formato de archivo son correctos o no, la correctitud del título, la descripción o dificultad de un objeto tiene mucho más nivel de dependencia en la percepción de quien evalúa.

**La consistencia lógica y coherencia:** los metadatos deben ser coherentes con las definiciones estándar y conceptos utilizados en el dominio. Debe haber coherencia interna, que es que la información contenida en los metadatos describen al mismo recurso.

**Conformidad con las expectativas**: El grado en que los metadatos satisfacen las expectativas/requerimientos, de una determinada comunidad de usuarios para una tarea determinada, puede ser considerado como conformidad con las expectativas. Si la comunidad puede, sin un cambio relevante en su flujo de trabajo, encontrar, identificar, seleccionar y obtener recursos, se puede decir que la metadata ayudo a cumplir con las expectativas de la comunidad. De acuerdo con la definición de la calidad ("aptitud para el propósito") utilizado en este trabajo, esta es una de las características más importantes de la calidad puesto que hace una relación entre lo que se quería y lo que se obtuvo, y el esfuerzo requerido para obtenerlo.

**Accesibilidad**: si los metadatos que no se puede leer o entender no tienen ningún valor. Hay dos ramas de calidad en accesibilidad claves, la física y la cognitiva. La física se refiere a la compatibilidad de formatos o los enlaces a estos. La cognitiva a la facilidad (o dificultad) para entender los metadatos. Estas dos dimensiones diferentes deben combinarse para estimar qué tan fácil es acceder y comprender la información presente en los metadatos.

**Actualidad/Relevancia:** las modificaciones sobre el objeto descrito debieran cambiar en la misma medida a los metadatos (actualidad). En un enfoque de biblioteca digital, los metadatos sobre un recurso están en constante crecimiento con cada nuevo uso del recurso (como ejemplo "visto unas x veces"). También, se debiera garantizar que los metadatos de un recurso estén disponibles al mismo tiempo que el recurso esté disponible (sin retrasos). El retraso, bajo este punto de vista, se puede considerar como el tiempo que le toma a los metadatos para describir el objeto lo suficientemente bien como para encontrarlo utilizando el motor de búsqueda del repositorio

**Procedencia**: La fuente de los metadatos puede ser otro factor para determinar su calidad. Esta pudiera venir, de medios automáticos o de medios manuales (humanos) o incluso de híbridos entre estos dos. El conocimiento sobre quién creó la instancia, el nivel de experticia del indexador, qué metodologías fueron seguidas durante la indexación y las transformaciones que los metadatos sufrieron, podrían dar una mejor idea de la calidad de la instancia.

La "*calidad en la metadata*" viene dada por la información/evaluación de la calidad del recurso u objeto descrito.

Este enfoque es orientado a aplicación, dicho de otra manera, una vez construido un sistema (con su estructura de metadata) para un determinado dominio no es, para nada, fácil modificar dicho sistema para que con simplicidad se pueda aplicar a otro dominio (Vuorikari et al., 2008).

Por ejemplo: una metadata diseñada para dar información de alta calidad sobre el contenido de investigaciones médicas, facilitando y limitando el espacio de búsquedas de trabajos sobre la materia no puede ser utilizada para obtener información de alta calidad sobre el contenido de análisis de comercio electrónico. Los campos de interés y las piezas clave de información necesaria, no son iguales, puesto que no son del mismo domino.

### *2.3.4.1. Parámetros de calidad en la Metadata*

Dependiendo del autor, la medición de la calidad de la metadata se basa en diferentes parámetros. "Moen et al (1998) identificó 23 parámetros, otros autores como Stvilia et al., (2007) utilizan varios de esos 23 parámetros y agrega otros adicionales, para finalmente agruparlos en 3 dimensiones de Calidad de Información (intrínseca, relacional / contextuales y reputacional) que contienen en total 32 parámetros" (Ochoa, 2009).

Bruce y Hillman (2004) basándose en investigaciones previas sobre Calidad de Información, agruparon varios parámetros en unos menos para así crear una estructura más fácil de aplicar en las evaluaciones, como se puede ver en la Figura 6. Ellos describen 7 dimensiones para la calidad de la metadata, estas son: completitud, precisión, procedencia, conformidad con las expectativas, consistencia lógica y coherencia, actualidad/relevancia y accesibilidad. (Bruce y Hillman, 2004 citado por Ochoa, 2009).

"El análisis de la estructura de Bruce y Hillman es de mucho interés para este trabajo porque sus siete parámetros son fáciles de entender por investigadores humanos, asimismo porque abarcan todas las dimensiones de calidad propuestos en otras estructura y porque su "pequeño tamaño" (pocos parámetros) facilita su evaluación" (Ochoa, 2009).

Sin embargo, la estructura Bruce y Hillman (así como también la Stvilia et al.) fue diseñada pensando en instancias de metadata estáticas (que no se modifican ni varían en el tiempo). Estas estructuras si bien son precisamente lo necesario en el caso de una biblioteca en el caso de repositorios digitales no resulta tan correcto, puesto que la metadata cambia con el tiempo, así que no se puede afirmar que el resultado de una evaluación en algún momento, resulte una correcta evaluación en cualquier

momento. Esto ocurre por la metadata dinámica, la cual puede cambiar cada vez que el recurso es usado o accedido (Ochoa, 2009).

Puesto que aún no se han diseñado estructuras para describir la metadata dinámica, algunos autores utilizan la estructura Bruce yHillman como un primer acercamiento, en algunos casos incluyendo algunas adaptaciones para evaluar las características de calidad cuando sea necesario para las particularidades de las instancias dinámicas (Ochoa, 2009)

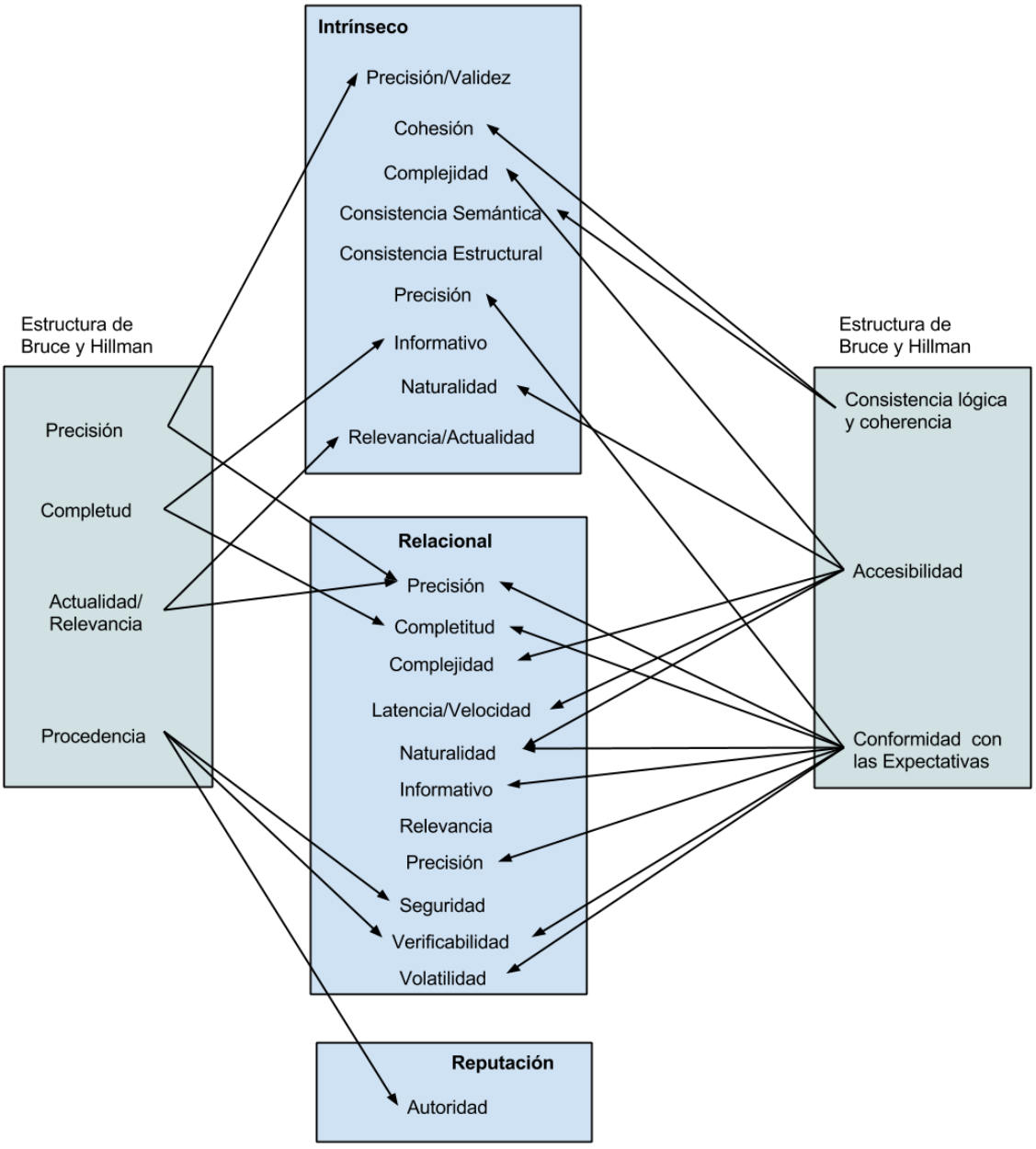


**Ilustración 5 - Correspondencia entre las estructuras de Bruce y Hillman y el Stvilia et al. (Adaptación de Ochoa, 2009)**

Recomendamos revisar el marco conceptual en el capítulo 2.5, para obtener información sobre la precisión, completitud, proveniencia, conformidad con las expectativas, consistencia, actualidad/relevancia y accesibilidad y así facilitar la lectura de este de este trabajo.

### *2.3.4.2. Formas de Estimar la Calidad de la Metadata*

Xavier Ochoa y Erik Duval en la introducción a su trabajo "*Automatic Evaluation of Metadata Quality in Digital Repositories*" (2009) afirman "La calidad de la metadata en las instancias almacenadas de los repositorios digitales es percibida como un asunto importante para sus operabilidad e interoperabilidad".

La principal función de un repositorio digital, la cual es proveer acceso a los recursos, pudiera verse severamente afectada por la calidad de la metadata(Ochoa et al., 2009).

Ochoa nos da como ejemplo lo siguiente: Suponga un recurso de aprendizaje indexado bajo el título "Lección 1 - Curso CS20", eso sin más descripción o palabras claves, con dificultad ser retornado en las búsquedas de materiales sobre "La introducción a la programación en Java" incluso si el recurso antes descrito es, de hecho, un buen texto introductorio a Java. El recurso será parte del repositorio, pero nunca será recuperado en búsquedas relevantes, por carecer de la metadata pertinente.

Existen 2 enfoques generales para la evaluación de la calidad de la metadata: La técnica manual y estadística simple.

En la **técnica manual**, se toma una muestra estadísticamente significativa de metadata y se compara contra un set de parámetros de calidad predefinidos, esta técnica de muestreo es parecida a la utilizada para verificar la calidad del catálogo de bibliotecas. Se promedia las evaluaciones de los evaluadores humanos y se obtiene un estimado de la calidad general de la metadata en el repositorio (Ochoa et al., 2009).

Sin embargo esta técnica presenta algunas desventajas:

1. Esta estimación sólo estará vigente mientras el repositorio no crezca, solo seráválida para el momento en que se toma la muestra, del repositorio crecer en volumen de contenidos o de metadatos, la estimación ya no sería certera y habría que repetirla.

2. Solo se puede estimar la calidad promedio con este método, la calidad de cada instancia individual del repositorio no se podría saber.

3. Esta técnica es costosa, puesto que implica dedicar recursos humanos a estar repetitivamente haciendo estas evaluaciones.

"En el caso de la **técnica estadística simple**, se recolecta información estadística de todas las instancias de metadata para obtener una estimación de su calidad. Este estudio utiliza como principal métrica la completitud (que tantos campos de metadata tienen o no información), aunque también se puede evaluar otras métricas" (Ochoa et al., 2009).

Si bien esta técnica proporciona una estimación de la calidad de la metadata de las instancias del repositorio, sin el elevado costo de una técnica manual, es también menos "significativa" o aporta menos información, puesto que la presencia o no de escrito en un campo no implica que lo escrito en el campo sea relevante o tenga "significancia" para los usuarios. La información que brinda es interesante en términos estadísticos pero podría fallar para alguna otra aplicación real (Ochoa et al., 2009).

Un instrumento ideal para la medición de calidad de metadata en repositorios de rápido crecimiento debiera tener 2 características: que sea escalable, lo cual significa que calcule automáticamente por cada instancia de metadatos y que tenga significación, que es que la información proporcionada sea útil. Las evaluaciones manuales son significativas pero no escalables. Las estadísticas simples son escalables, pero no son significativas (Ochoa et al., 2009).

### 2.3.5. Uso Real de la Metadata

Jehad Najjary Erik Duval en su trabajo "*Actual Use of learning Objects and Metadata: An Empirical Analysis*" (2008) hablan del uso de la metadata en el repositorio de objetos de aprendizaje ARIADNE.

Su estudio tomo 3700 instancias de metadata del repositorio ARIADNE, y produjo unas estadísticas (que podemos apreciar en la Tabla 3) donde se muestra el porcentaje de veces en que cada elemento de datos se rellenó por los indexadores durante el proceso de indexación.

**Tabla 3 - Porcentaje de uso hecho a los elementos de datos por los indexadores en ARIADNE (Najjar. et al., 2008).**

| Element | Value provided (%) | Value not provided (%) | Most used Vocab-value(M) | % of M (filled-in) | % M among all cases |
|---|---|---|---|---|---|
| Granularity | 91.9 * | 8.1 | Lesson | 92.7 | 85.2 |
| Didactical Context | 53.3 | 46.7 | University Degree | 69.7 | 37.2 |
| Interactivity Level | 53.2 | 46.8 | Medium | 67.7 | 36.1 |
| Semantic Density | 52.4 | 47.6 | Medium | 76.4 | 40.0 |
| Difficulty Level | 52.2 | 47.8 | Medium | 72.8 | 38.0 |
| Restrictions | 5.2 | 94.5 | Contact Author | 90 | 5.2 |
| Source | 1.3 | 98.7 | - | - | - |
| Version Information | 7.0 | 93.0 | - | - | - |
| Description | 11.2 | 88.2 | - | - | - |
| OS Version | 0.5 | 99.5 | - | - | - |
| Installation remarks | 24.3 | 75.7 | - | - | - |
| Other Constraints | 0.15 | 99.85 | - | - | - |

*: used to be mandatory at the previous version of ARIADNE authoring tools.

A partir de los datos que se muestran en la Tabla 3, se puede observar que sólo un elemento de datos se utiliza casi siempre: el elemento de granularidad. Los demás elementos de datos son utilizados en menor medida, un grupo es utilizado aproximadamente en el 50% de las descripciones (Contexto didáctico, nivel de interactividad, densidad semántica, nivel de dificultad) y el resto rara vez se utilizan (restricciones, fuente, información de la versión, descripción, Versión del OS - OperatingSystem, apuntes de instalación, otras restricciones) en el proceso de indexación (Najjar, J. et al., 2008).

Dicho trabajo también presenta un análisis estadístico (véase figura 7) de los registro de consulta de ARIADNE, los datos disponibles de 4.723 consultas realizadas por aproximadamente 390 usuarios en seis ARIADNE *Local Knowledge Pool Systems*(LKPs) [Génova, Galati, Grenoble-UJF, Lausanne-EPFL,

Lausanne-UNIL y Lovaina-CS] en diferentes períodos de tiempo.

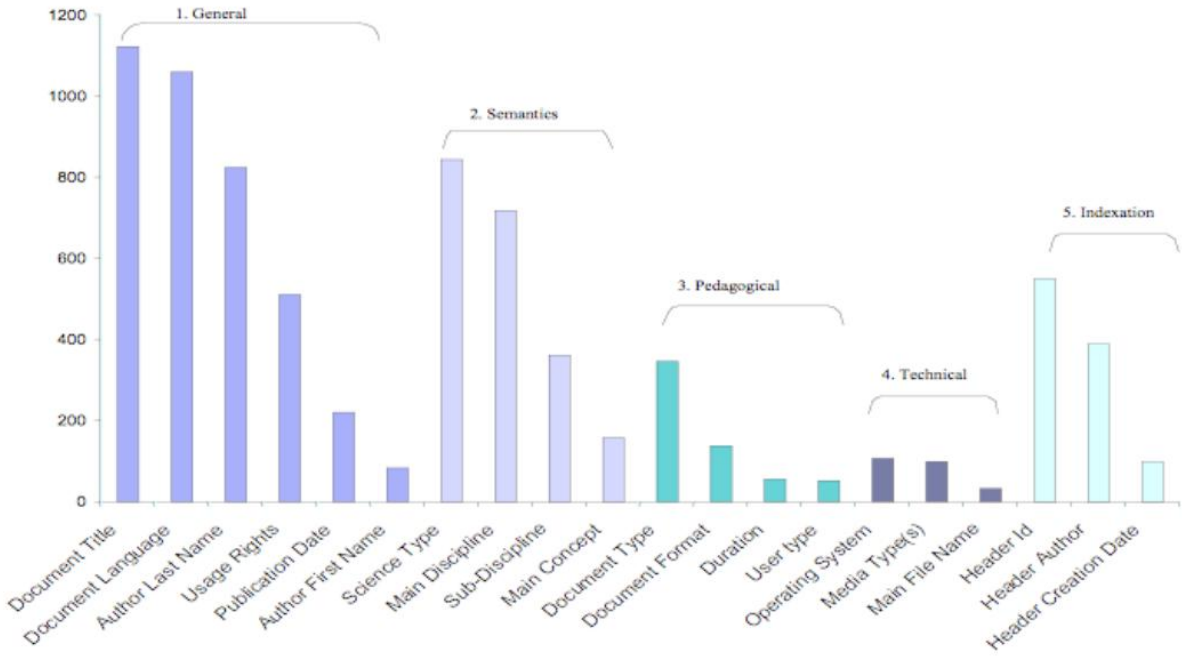


**Ilustración 6- Frecuencia de uso de los elementos de ARIADNE en las consultas de buscadores (Najjar. et al., 2008).**

La Ilustración 2 muestra la frecuencia con la que los diferentes elementos de ARIADNE se han utilizado en las consultas de buscadores. Enfocándose en los valores de la Figura 7 podemos destacar que elementos como el título del documento fue utilizado en unas 1100 consultas (frecuencia).

La mayoría de usuarios aceptaba (2008) los elementos de datos proporcionados por defecto por ARIADNE para las consultas, a tal punto que los 20 elementos de datos más utilizados por los usuarios son elementos proporcionados por defecto. Es destacable que, la herramienta de consulta permitía a los usuarios a cambiar la configuración predeterminada y mostrar toda la lista de elementos sin embargo pocos usuarios cambiaban los ajustes de elemento por defecto proporcionados (Najjar. et al., 2008).

Una comparación entre el uso real de elementos de datos de la indexación y los procesos de búsqueda, revela algunos de los elementos de datos que se han utilizado por más de un 50% de los indexadores, tales como granularidad, contexto didáctico y elementos de densidad semántica, no son utilizados por la mayoría de los buscadores. Dicho de otra manera, son campos que si bien sí se llenaron, los buscadores no les dan utilidad (Najjar. et al., 2008).

Asimismo, el investigador descubrió que tanto usuarios buscadores como indexadores utilizan los mismos valores en los elementos de datos para los objetos de interés. Por lo tanto el contenido de

estos campos tiene información significativa para los usuarios humanos, información que permite hacer búsquedas e indexaciones coherentes sobre los objetos de aprendizaje (Najjar. et al., 2008).

**Tabla 4- Frecuencia de elementos usados en las consultas de usuario (Najjar. et al., 2008).**

| **No. elementos usados en consultas** | **0** | **1** | **2** | **3** | **4** | **>=5** | **Total** |
|---|---|---|---|---|---|---|---|
| **Frecuencia** | 548 | 2488 | 701 | 498 | 258 | 230 | **4723** |
| **Porcentaje** | 11.6 | 52.7 | 14.8 | 10.5 | 5.5 | 4.9 | **100.0** |

La Tabla 4 muestra que los usuarios buscadores están más interesados en la formación de las consultas que contienen relativamente pocos elementos de metadatos. "La mayoría de las consultas (75%) contienen entre uno (1) y tres (3) elementos de datos. Menos de 5% de las consultas contiene cinco o más elementos de datos. La media del número de elementos en consultas es 1,7 elementos y la desviación estándar es de 1.6. Alrededor del 12% de las consultas no contenía elementos de metadatos en absoluto. De hecho, esto está relacionado con algunos problemas de usabilidad con la herramienta de consulta; algunos buscadores lanzaban (en el momento de investigación de aquel trabajo) directamente consultas sin seleccionar ningún elemento de datos" (Najjar. et al., 2008).

### 2.3.6. Presencia de Metadata en Repositorios Digitales Institucionales

Xavier Ochoa, Joris Klerkx, Bram Vandeputte, y Erik Duval en su trabajo "*On the Use of Learning Object Metadata: The GLOBE Experience*" (2011) nos comenta sus descubrimientos al estudiar la metadata de GLOBE (*Global Learning Objects Brokered Exchange*) el cual permite el intercambio y reusó entre varios repositorios de objetos de aprendizaje alrededor del mundo, trabajando sobre el estándar LOM de metadatos. Para Ochoa, GLOBE es un lugar ideal para llevar a cabo un análisis de la utilización real del estándar LOM en el mundo real (Ochoa. et al., 2011).

Sobre el uso de elementos de datos LOM, Ochoa realizó un análisis de frecuencia de los elementos de datos en LOM. Para este análisis, sólo se contaron campos de primer nivel, ósea se contaron las apariciones de elementos de data, no del total de elementos de data dentro de esos elementos. "Por ejemplo, en el formato XML del LOM, el campo General.Structure tiene dos subcampos: General.Structure.source y General.Structure.value. En este caso, sólo se cuenta el número de apariciones de General.Structure. La justificación para esto recae en que no todos los repositorios tienen el subcampo de source y, en la mayoría de los casos, el número de apariciones de campos de primer nivel es igual al número de apariciones de subcampovalue. Si el elemento de datos se llenó más de una vez, se cuenta una sola vez" (Ochoa. et al., 2011). La Figura 8 del anexo A muestra el porcentaje de metadatos que tienen un valor para los diferentes elementos de datos de LOM en GLOBE

El principal hallazgo de ese estudio es que sólo un subconjunto del estándar LOM se utiliza con frecuencia para describir los objetos de aprendizaje. Sólo 20 de los 50 elementos de datos se utilizan más del 60% de las veces. Además, 16 elementos de datos se utilizan menos de 10% del tiempo. (Ochoa. et al., 2011).

Friesen, N. en su trabajo "*The international learning object metadata survey. The International Review of Research in Open and Distance Learning*" (2004) concluía en su trabajo que "la complejidad añadida de LOM no se utiliza en el mundo real", este resultado, pareciera ser corroborado por estas estadísticas (Ochoa. et al., 2011).

Sin embargo, un estudio similar realizado por Wand en casi 1 millón de instancias de metadata de la *Open Archives Initiative*(OAI) la cual utiliza el Dublin Core (DC) ,el esquema de metadatos mucho más simple, encontró que de los 15 elementos de datos que conforman el DC, cinco (creador, identificador, título, fecha y tipo) se utilizan 71% de las veces mientras que los cinco elementos menos usados (idioma, formato, de relación, de colaborador y fuente) se utilizan menos del 6% de las veces (Ochoa. et al., 2011).

El contraste entre estos dos (2) trabajos, brinda la idea que LOM siendo más complejo, ayuda a recoger proporcionalmente más información que un esquema más simple y más general tal como DC. "La desigualdad en la utilización de los distintos elementos de los datos (completitud) parece ser algo inherente a la creación de metadatos. Esta desigualdad merece más investigación, a través de un análisis comparativo de la utilización de diferentes estándares de metadatos" (Ochoa. et al., 2011).

Teniendo en cuenta que LOM fue diseñado específicamente para describir el material educativo, es importante revisar el uso de la sección de Educación. En esta sección 4 de los 11 elementos de datos educativos, (aprendizaje del tipo de recurso, rol previsto del usuario final, rango típico de edad y contexto) se utilizan más del 40% de las veces, 3 elementos (Lengua, nivel de Interactividad y tipo de Interactividad ) se utilizan entre 10% y 20% de las ocasiones, y los 4 elementos restantes (descripción, de dificultad, densidad semántica y tiempo promedio de aprendizaje) se utilizan menos del 10% de las veces. Si bien estos valores sugieren poco interés en utilizar al completo el estándar, presentan la prueba que el LOM si se utiliza en el mundo real para capturar información educativa de los objetos digitales (Ochoa. et al., 2011).

Para encontrar fácilmente las diferencias en la completitud solo de los elementos de datos educativos a través de los diferentes repositorios, una vista de mapa de calor se presenta en la Tabla 5. Se presenta la frecuencia relativa de la completitud de un elemento de datos en cada repositorio estudiado. Los valores mayores que 1 representan que el elemento descrito se utiliza comúnmente más de una vez en

cada caso (un objeto con más de una descripción, por ejemplo). La primera columna es la frecuencia relativa media de todos los repositorios (Ochoa. et al., 2011).

**Tabla 5- Mapa de calor comparando el uso de los elementos educacionales (Ochoa. et al. 2011).**

| TOTAL | LACLO | ARIADNE | LORNET | LRE | OER | KOCW | OUJ | Path |
|---|---|---|---|---|---|---|---|---|
| 0.69 | 0.96 | 0.51 | 0.84 | 0.97 | 1.0 | 0.0 | 1.0 | educational |
| 0.54 | 0.83 | 0.39 | 0.75 | 0.76 | 0.84 | 0.0 | 0.83 | educational.context |
| 0.03 | 0.13 | 0.03 | 0.11 | 0.01 | 0.0 | 0.0 | 0.07 | educational.description |
| 0.03 | 0.0 | 0.03 | 0.0 | 0.05 | 0.0 | 0.0 | 0.0 | educational.difficulty |
| 0.57 | 0.83 | 0.36 | 0.37 | 1.08 | 0.0 | 0.0 | 0.0 | educational.intendedenduserrole |
| 0.15 | 0.0 | 0.24 | 0.01 | 0.03 | 0.0 | 0.0 | 0.0 | educational.interactivitylevel |
| 0.2 | 0.0 | 0.33 | 0.02 | 0.02 | 0.0 | 0.0 | 0.0 | educational.interactivitytype |
| 0.13 | 0.0 | 0.09 | 0.02 | 0.21 | 0.32 | 0.0 | 0.0 | educational.language |
| 0.73 | 0.96 | 0.55 | 0.71 | 1.01 | 1.26 | 0.0 | 0.0 | educational.learningresourcetype |
| 0.02 | 0.0 | 0.03 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | educational.semanticdensity |
| 0.56 | 0.83 | 0.38 | 0.01 | 0.8 | 1.1 | 0.0 | 0.0 | educational.typicalagerange |
| 0.02 | 0.0 | 0.01 | 0.1 | 0.04 | 0.0 | 0.0 | 0.0 | educational.typicallearningtime |

### 2.3.7. Uso de elementos del vocabulario en LOM

"Los elementos de vocabulario son elementos de datos que sólo puede ser llenado con un conjunto limitado de valores establecidos por la norma de metadatos. El objetivo principal de estos elementos de vocabulario es proporcionar un mayor nivel de interoperabilidad semántica"(Ochoa. et al., 2011).

Ochoa. et al., (2011) estudió el uso de los elementos de vocabulario LOM en el repositorio GLOBE. En su trabajo identificaron al menos 3 grupos principales de los elementos de vocabulario.

El grupo "Cola pesada" se caracteriza por crear una curva menos pronunciada donde dos o tres valores que se utilizan en gran medida (más de 25% de las ocasiones), seguido de varios valores con un uso menor, aunque todavía significativos, entre 5% y 25%. Un ejemplo de esta distribución se puede ver en la Figura 9. El lenguaje (del recurso y los metadatos), formato, tipo de recursos y elementos de contexto son parte del grupo "Colas pesadas" (Ochoa. et al., 2011).

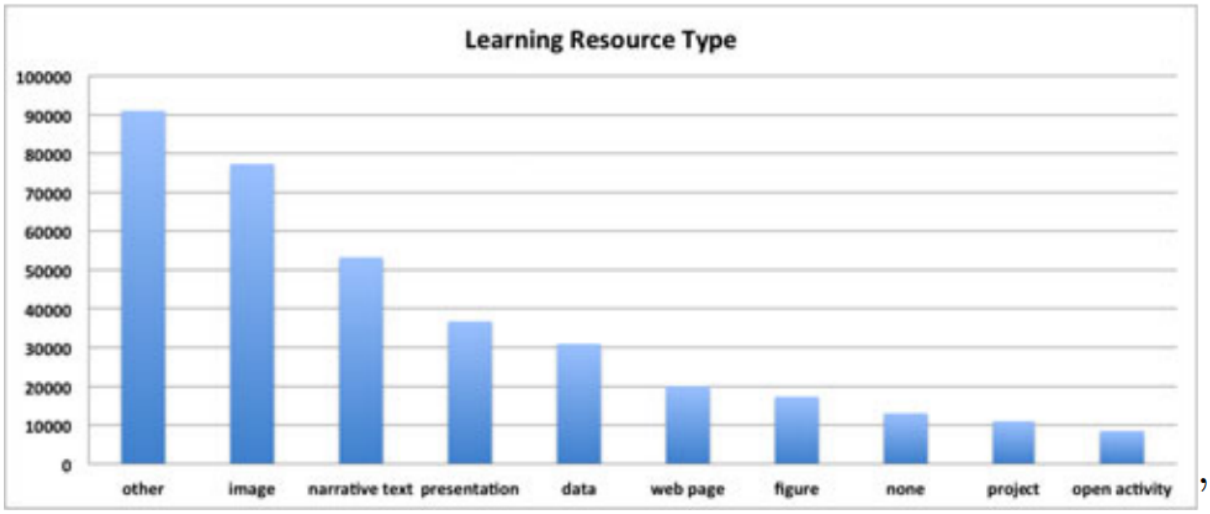


**Ilustración 7 -Distribución "Cola Pesada" en el uso del vocabulario de Metadata (Ochoa. et al., 2011).**

El grupo "cola liviana" se caracteriza por crear una curva mucho más pronunciada donde hay un valor dominante (uso de más de 70% de las ocasiones), seguido de unos pocos (2 o 3) valores que se utilizan más raramente (de 2% a 10% de las ocasiones). Un ejemplo de esta distribución se puede ver en la Figura 10. La estructura, el rol de usuario previsto, tipo de interactividad y estado parecen ser parte del grupo "cola liviana" (Ochoa. et al., 2011).

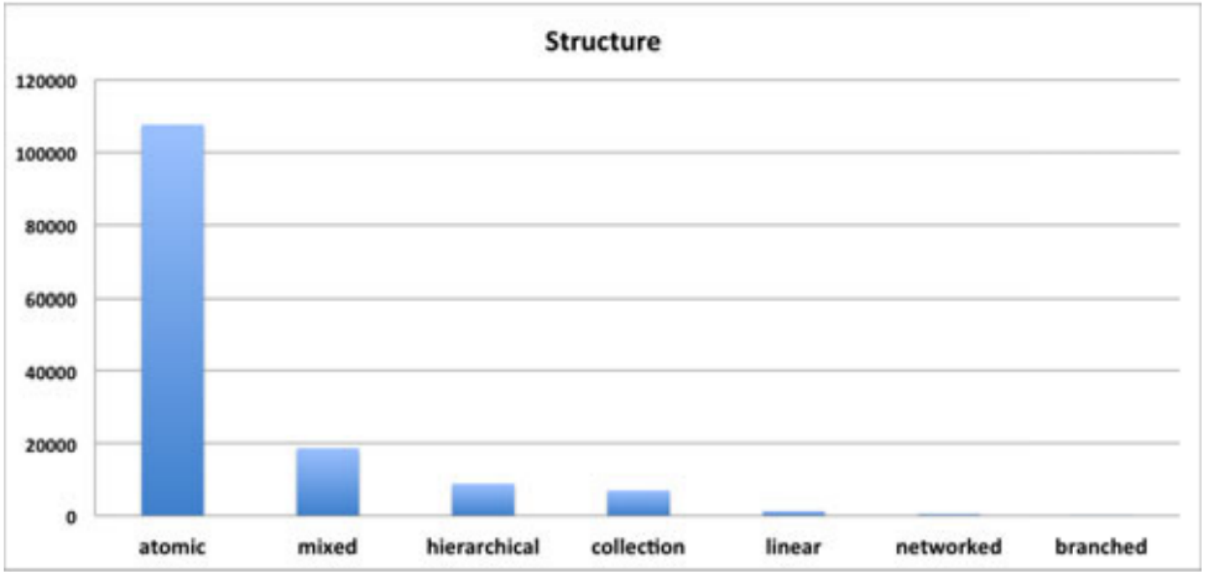


**Ilustración 8 - Distribución "Cola liviana" en el uso del vocabulario de Metadata (Ochoa. et al., 2011).**

El tercer grupo, llamado grupo "simétrico", contiene todos los elementos del vocabulario que se basan en una escala, por ejemplo: para el caso de dificultad: muy fácil, fácil, medio, dificil, muy difícil. En este grupo, el valor central de la escala es, por mucho, el valor más común (60% a 70%). Dificultad, nivel de interactividad, densidad semántica y nivel de agregación son algunos de los elementos que forman parte de este grupo (Ochoa. et al., 2011).

El autor Ochoa. et al., (2011) brinda como posible explicación a estas distribuciones a la objetividad de la información y al uso de los valores por defecto que la herramienta de indexación ofrece.

Un ejemplo en la objetividad de la información, podría darse en el grupo cola pesada, en el cual debiera ser fácil para el indexador determinar el idioma de un objeto, especialmente porque el objeto suele estar en un idioma que él conoce. Como éste otros ejemplos apuntan a que el grupo colas pesadas está formado por elementos en los cuales es fácil obtener información del mismo objeto o su contexto(Ochoa. et al., 2011).

Por otro lado, los elementos de datos en el grupo cola liviana, pudieran seguir esa distribución por valores ofrecidos por defecto de la herramienta de indexación, por ejemplo es seguro suponer que el usuario final previsto es el alumno, aunque en otros casos parece ser más difícil de determinar objetivamente el valor adecuado (tal como la estructura del objeto) (Ochoa. et al., 2011).

La existencia de un valor predeterminado (o por defecto) seguro (por lo general el valor medio) parece ser también una posible explicación para la distribución de valor en los elementos en el grupo simétrico (Ochoa. et al., 2011).

### 2.3.8. Formas de hacer búsquedas usando Metadata

Martin Wolpers, Martin Memmel, Katja Niemann, Joris Klerkx, Marcus Specht, Alberto Giretti y Erik Duval en su trabajo "*Aggregating metadata to improve access to resources*" nos describen su experiencia exploratoria sobre el sistema MACE es cual es un buscador sobre repositorios de objetos de aprendizaje (trabajando en el dominio de la arquitectura).

El sistema de MACE es un repositorio federado, el cual recolecta todos los metadatos y emplea a seres humanos, y técnicas de aprendizaje automático para relacionar semi-automáticamente los recursos dentro de los repositorios y entre ellos. Haciendo esto, en vez de tener que hacer una búsqueda de repositorio en repositorio, el sistema MACE proporciona una interfaz para encontrar y acceder a los objetos de aprendizaje pertinentes de todos los repositorios federados (Wolpers et al., 2011)

El sistema MACE provee varios tipos de búsquedas como: búsqueda facetada, búsqueda orientada geográficamente, búsqueda por clasificación y búsqueda social, proporcionando así una habilidad de obtención de recursos superior. Todas estos tipos de búsqueda se apoyan en la metadata, la cual es combinada y enriquecida por el sistema MACE, su interfaz y su comunidad (Wolpers et al., 2011)

Este estudio es de interés para el presente trabajo, puesto que resalta el hecho que las búsquedas se pueden realizar de más de una forma (orientada geográficamente, clasificación, social), y dependiendo de la forma en que se "oriente" una búsqueda se pueden obtener resultados más satisfactorios para el usuario.

"El portal ofrece facilidades de búsqueda para los diversos recursos de aprendizaje almacenados en los repositorios participantes. Encontrar recursos se facilita a través de búsqueda de palabras clave, búsqueda por clasificación, búsqueda basada en las competencias, la búsqueda social y la búsqueda de facetas" (Wolpers et al., 2011). Gracias al trabajo de Wolpers et al. (2011) lasdiferentes formas de búsquedas serán explicadas a continuación.

**Búsqueda Sencilla por Palabra:** a través del uso de palabras clave y combinaciones de las mismas, el sistema de MACE permite al usuario al usuario buscar a través de los recursos de aprendizaje. En base a, la metadata proporcionada por los repositorios y de las palabras clave proporcionadas por los usuarios, se hace la búsqueda por palabra clave.

**Exploración por Clasificación:** Muchos de los recursos de aprendizaje incluidos dentro del sistema MACE se clasifican utilizando el sistema de clasificación arquitectónica MACE. Este se basa en una visualización árbol hiperbólico de la clasificación, los usuarios pueden ir haciendo clic a través de la clasificación jerárquica hasta que hayan encontrado el término clasificación deseada que luego vinculan a los respectivos recursos de aprendizaje. En esta clasificación se presenta una estructura en forma de árbol donde los nodos y hojas representan términos de clasificación y recursos de aprendizaje asociados.

**Basada en Competencia**: Según el MEC (Marco Europeo de Cualificaciones) los objetos de aprendizaje pueden clasificarse por competencias y niveles, así que los educadores pueden clasificar los objetos de aprendizaje según las mismas. Los usuarios pueden filtrar sus búsquedas en una matriz utilizando diferentes definiciones de competencia y niveles. Inclusive si un recurso fuese clasificado en varias competencias o ámbitos de competencia se podría buscar en el mismo espacio de contenido con diferentes modelos de competencia.

**Búsqueda Facetada:** Basándose en facetas tales como los repositorios en la que desea buscar, el lenguaje de los resultados, el tipo de recursos, la clasificación de los recursos, y la competencia asociada se puede restringir el dominio de los resultados de búsqueda. MACE combina el mecanismo de búsqueda por palabra clave y las facetas, los usuarios son capaces de calificar la palabra clave con varias facetas. Cuando se selecciona un valor para una faceta, la interfaz cambia de forma dinámica y proporciona el número de resultados para cada faceta que coincidan con los criterios seleccionados, tal como hace el actual buscador de ARIADNE.

**Búsqueda Social:** "al usuario se le muestran las etiquetas más populares aportadas por los usuarios de MACE, visualizadas por una nube de etiquetas. Una etiqueta vincula al respectivo recurso (s) de aprendizaje. El uso de la palabra clave de búsqueda permite al usuario buscar a través de los contenidos que ya fueron etiquetados por los usuarios de MACE".

Una variedad de posibilidades de interacción de medios sociales se ofrecen en MACE. Por ejemplo, los usuarios pueden añadir nuevos contenidos a MACE, pueden mantener carteras de recursos personales, es posible mantener listas de contactos y enviar mensajes a otros usuarios, pueden aportar información sobre los recursos (por ejemplo, etiquetas, comentarios y votos), y pueden hacer búsquedas dentro de esta información.

La información sobre el uso de los recursos digitales y de los metadatos sociales que se crean de forma explícita (etiquetas, comentarios, y valoraciones) o implícitamente (mediante la creación de colecciones personales) como consecuencia de las actividades del usuarios se conocen como metadatos sociales. Estos metadatos son diferentes de los metadatos creados automáticamente o metadatos creados en un proceso de arriba abajo por los expertos.

Ordenar por número de visitas, ordenar por calificación u ordenar por número de veces marcado como favorito son nuevas formas de generar puntos de vista sobre los recursos digitales, que nos dan una idea en el juicio de la relevancia de un recurso digital. El dinamismo de los metadatos sociales (etiquetas para añadir, eliminar o modificar los elementos de una taxonomía), ayuda a otros usuarios a navegar por el contenido social permitiendo a los usuarios encontrar nichos que son relevantes para ellos.

Con el objeto de dar una atención más personalizada, mostrando un resumen de las actividades para el usuario, MACE cada vez que un usuario final utiliza el portal, captura los datos de uso en el portal utilizando el esquema Contextual de Atención de Metadatos (CAM).

CAM describe cómo las personas interactúan con el portal (lo que leen, buscan, publican, etc.) y permite un análisis de datos de uso. En MACE, las actividades de los usuarios se utilizan para clasificar los resultados de búsqueda según sus tipos de uso y las cantidades.

## 2.4. Repositorio Digital

Kahn y Wilensky (2006, citados por Palavitsinis, 2013) definen un repositorio digital Institucional como un " sistema de almacenamiento accesible desde la red en la que los objetos digitales pueden ser almacenados para un posible acceso o recuperación posterior. El repositorio cuenta con mecanismos para añadir nuevos objetos digitales a su colección (depósito) y para su puesta a disposición (acceso), utilizando, como mínimo, el protocolo de acceso al repositorio. El repositorio puede contener otra información relacionada, servicios y sistemas de gestión"

En la Figura 3 se muestran los casos de usos propuestos por Khan y Wilensky.

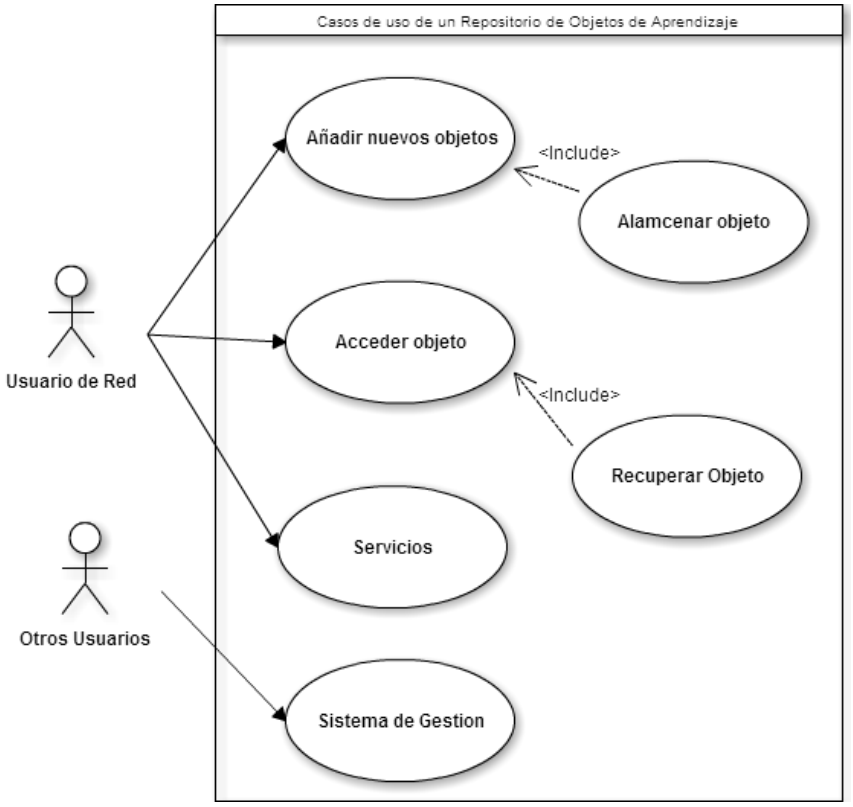


**Ilustración 9- Diagrama de casos de uso de Khan y Wilensky**

Para diferenciar un repositorio digital de cualquier directorio tradicional o base de datos Heery y Anderson (2005, citados por Palavitsinis, 2013) indican que el repositorio tiene que cumplir con los siguientes enunciados:

- El contenido tiene que estar depositado en el repositorio, bien sea por el creador o un tercero.
- La arquitectura del repositorio gestiona tanto el contenido como la metadata
- El repositorio ofrece un conjunto mínimo de servicios, como por ejemplo: poner(subir), obtener, buscar, control de acceso
- El repositorio debe ser sostenible y de confianza , bien apoyado y bien administrado

En este mismo trabajo, los repositorios, sin importar el campo (área) en la que se enfocan (educación, cultura, ciencia, entre otros) comparten objetivos en común que guían su operación y existencia, como por ejemplo:

- Un mejor acceso a los recursos
- Introducir nuevos modos de publicación y revisión por parte de otros autores
- Sistemas de gestión de la información corporativa (sistemas de gestión de documentos y gestión de contenidos )
- Habilita el intercambio de datos ( reutilización de los datos de la investigación , la reutilización de objetos de aprendizaje )
- La ayuda a la preservación de los recursos digitales

### 2.4.1. Repositorio Digital Institucional

Un repositorio digital institucional que en el caso de las instituciones académicas también se lo conoce con el nombre de repositorio objetos de aprendizaje "es una base de datos electrónica que aloja una colección de "pequeñas" unidades de información educativas o actividades que pueden ser accedidas para su obtención y uso" (Lehman, 2007). Un ejemplo de RDI es Saber UCV del cual se puede ver una captura de pantalla en la Ilustración 10.

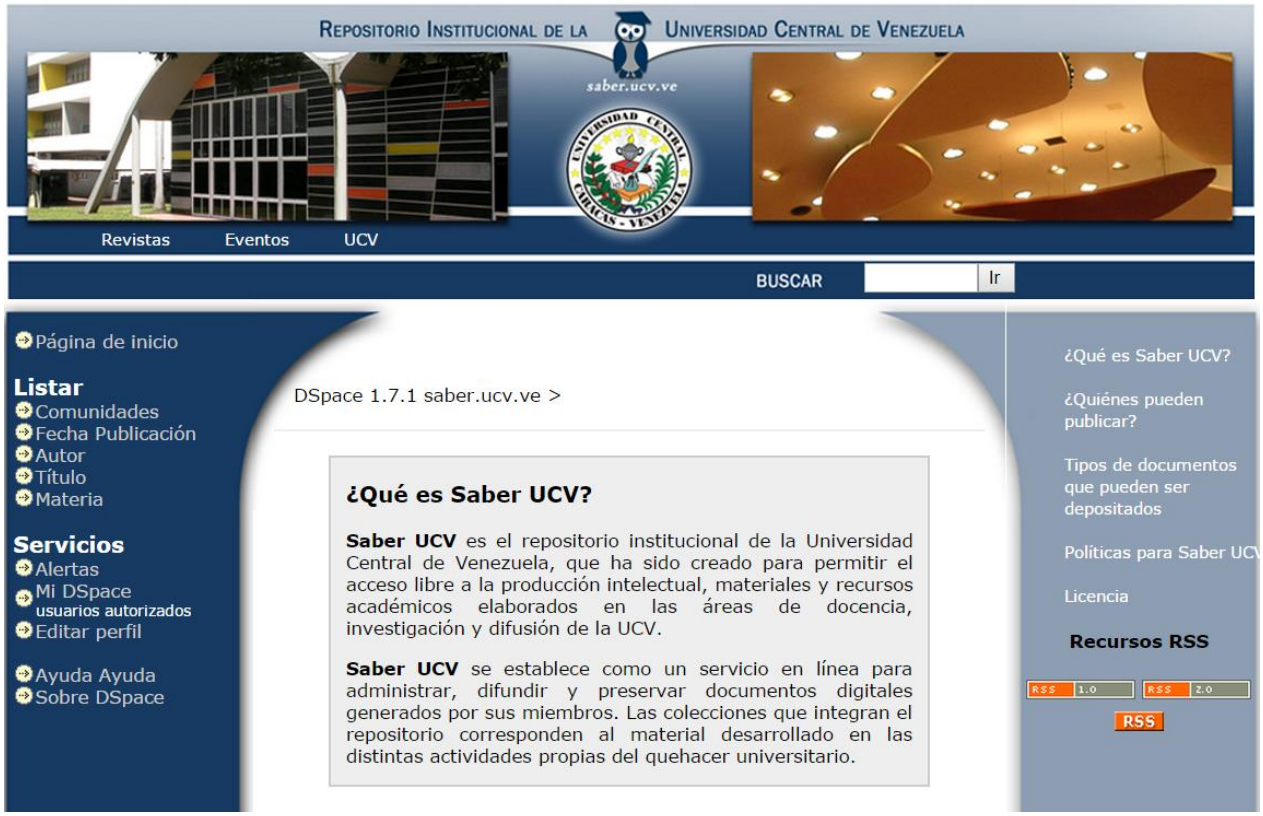


**Ilustración 10 - Saber UCV Portal**

Los repositorios digitales institucionales permiten la organización de objetos de aprendizaje, incrementa la eficiencia, facilita los trabajos colaborativos y la reutilización del objeto de aprendizaje, y da apoyo a las oportunidades de aprendizaje (Lehman, 2007).

Las organizaciones que operan repositorios digitales toman la responsabilidad del mantenimiento a largo plazo de estos recursos digitales, así como hacer de los repositorios disponibles para las comunidades que el depositante y repositorio hayan acordado (British Library, 2006 citada por Lehman, 2007).

Los repositorios pueden basarse en una base de datos o varias bases de datos unidas por un motor de búsqueda en común. Cuando son varias bases de datos unidas por un motor de búsqueda se trata de un repositorio federado.

Existen 2 tipos, populares, de repositorios federados. El que se basa en búsquedas federadas, que básicamente funciona como interfaz para hacer una búsqueda distribuida. Y el que se basa en cosecha de metadatos en el cual un repositorio central almacena la información de metadata, para así poder contestar las búsquedas de los usuarios (Ochoa, 2009).

En la búsqueda federada un nodo central recibe la consulta del usuario y seguido a que el nodo realice la consulta distribuida, cada repositorio enviará una lista de resultados que serán ensamblados para crear una lista (respuesta final). Este tipo de repositorios federados requieren tener establecidos varios estándares previos, para las comunicaciones, para las consultas, para los resultados (Ochoa, 2009).

Uno de los estándares utilizado para implementar búsquedas federadas es SQI (Simple Query Interface o traducido al español Interfaz de Consultas Simples). SQI es un API (Application Programming Interface traducido como Interfaz de Programación de Aplicaciones) que puede trabajar con repositorios de metadata muy heterogéneos que permite establecer listas de comandos en algún lenguaje de consulta con algún formato de resultado para facilitar la comunicación entre repositorios (Ochoa, 2009).

La Figura 4 brinda un esquema de la arquitectura de SQI.

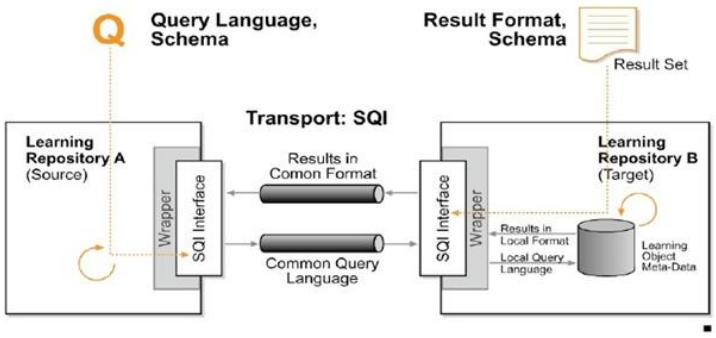


**Ilustración 11 - Arquitectura SQI. (Ochoa, 2009)**

En la cosecha de metadatos un nodo central envía requerimientos a los nodos repositorios para que estos le envíen su metadata. Teniendo esta metadata este nodo queda empoderado para responder consultas sobre el total de la federación (Ochoa, 2009).

El protocolo más difundido para la cosecha de metadatos es el *Open Archive Initiative Protocol for Metadata Harvesting* (OAI-PMH) (traducido al español como Protocolo para la Cosecha de Metadatos Iniciativa Archivo Abierto). Como se ve en la Figura 5, en este protocolo hay 2 actores, los repositorios y los cosechadores, donde los repositorios deben soportan al protocolo OAI-PMH, mientras los cosechadores utilizan OAI-PMH (Ochoa, 2009).

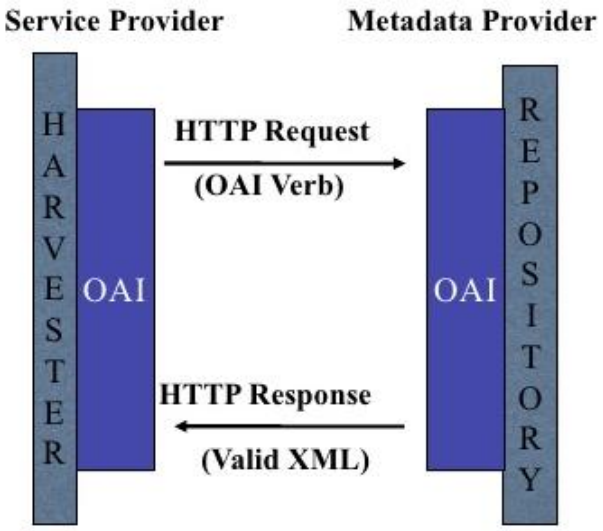


Ilustración 12 - Como Funciona OAI PMH (Ochoa, 2009).

Otra forma de ver los repositorios de objetos de aprendizaje es dividiéndolos en tres (3) tipos: general, específicos a una disciplina, y comerciales ("Learning Objects: Collections", 2003 citado por Lehman, 2007). Factores importantes a ser tomados en consideración cuando se selecciona el tipo de repositorio son el tipo de concordancia, la accesibilidad, flexibilidad y facilidad de uso para los usuarios finales: instructores, diseñadores instruccionales, o los alumnos (Lehman, 2007).

### 2.4.2. Proceso de Indexación

El proceso de Indexación es aquel en el que se identifica al objeto de aprendizaje, en este proceso se instancia la metadata. El objetivo de este proceso es hacer "buscable" al objeto, hacerlo visible ante el buscador.
El proceso de indexación debiera ser aplicado cada vez que se introduce un objeto al repositorio. Puede hacerse de manera manual (por humanos) o automático.

## 2.5. Minería de Datos

De acuerdo a Han (2012) la minería de datos es el proceso de descubrir patrones interesantes de grandes cantidades de datos. Según la Universidad Central de Venezuela (2014) se puede definir como: "Conjunto de técnicas y herramientas aplicadas al proceso no trivial de extraer y presentar conocimiento implícito, previamente desconocido, potencialmente útil y humanamente comprensible, a partir de conjuntos de datos". Otras definiciones se acercan bastante a estas dos. Sera la segunda la definición que se utilizara para minería de datos, en el presente trabajo.

Todo resultado de minería de datos tiene que presentar 4 características deseables (UCV, 2014):

1. Valido: el resultado debe ser correcto, no solo para responder preguntas de la muestra utilizada sino de cualquier muestra de la misma fuente (siempre y cuando conserven las mismas características).
2. Novedoso: el resultado debe ser novedoso, no debe presentarse como resultados cosas que ya se sabían o que son obvias y evidentes.
3. Potencialmente útil: el resultado tiene que tener utilidad o beneficio, la información proporcionada debe ayudar a tomar mejores decisión.
4. Comprensible: los seres humanos deben poder entender los resultados, validarlos y darles utilidad en el proceso de toma de decisiones.

La minería de Datos tiene 5 eslabones, que hay que reconocer (UCV, 2014):

1. Datos: la minería de datos se aplica a algo, y ese algo son los datos, estos son los que se procesan y estudian para producir conocimiento.
2. Preparación de los Datos: es común, que los datos "crudos" como se obtienen de la fuente de datos no estén en condiciones optimas, o siquiera buenas para ser utilizados. Ello conlleva a la necesidad de preparar los datos para ser procesados.
3. Minería de Datos: los datos procesados pueden ser tratados en los procesos de minería de datos, y de ellos se pueden obtener varios resultados. Dependiendo de qué resultado se quiera y con qué datos se cuenta se puede aplicar una técnica u otra de minería.
4. Evaluación e interpretación: los algoritmos de minería de datos no van a presentar resultados en lenguaje natural, es necesario interpretar los resultados para darles sentido completo. Por ejemplo: Indicar que la medición de la distancia euclidiana de tales vectores da como resultados asociaciones de distancia fuertes entre algunos y débiles con otros, no dice nada útil para quien toma decisiones , por el contrario, indicar que transformar la data a valores numéricos que pueden ser tratados como vectores a los que se les puede medir su similitud con otras instancias de data, da como resultados el descubrimiento que las instancias se pueden agrupar de una forma "X", si es información comprensible, utilizable, novedosa y correcta.
5. Conocimiento: Los descubrimientos de la fase anterior deben ser presentados de forma comprensible y útil, se debe tener cuidado que brindar siempre información novedosa.

Existen dos grupos de tareas principales en la minería de datos, las tareas predictivas y las tareas descriptivas. Las tareas predictivas buscan pronosticar algún valor en función de otros, se les asocia con el aprendizaje supervisado, lo que quiere decir que se dispone de un atributo que representa la respuesta del problema. Las tareas descriptivas buscan asociarle detalles a las instancias de data, se les asocia al aprendizaje no supervisado, lo que quiere decir que no se dispone de algún atributo respuesta.

Entre las tareas predictivas se cuenta la clasificación y la regresión, mientras que en las descriptivas se cuenta con tareas como la clusterizacion (agrupación), análisis de asociación (reglas) y la detección de anomalías.

### 2.5.1. K-medias

El k-medias es un método utilizado en la minería de datos para el agrupamiento. Se basa en particionar un conjunto de N vectores (conocidos como instancias, también llamadas observaciones) en k grupos, donde cada instancia pertenece al grupo de cuyo centro esta instancia posea la distancia más corta, como se puede ver en la figura x.

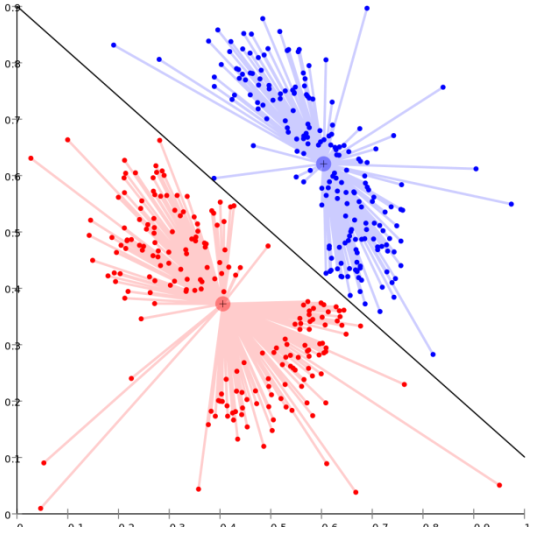

Ilustración 13 - Representación grafica de un k-medias con 2 grupos (Wikipedia, 2015).

Este método posee ciertas variantes, como por ejemplo: k-mediodes utiliza como centro a una instancia de la muestra, otro caso es que los algoritmos de k-medias pueden crean los centros y los "mueven" (modifican), en un proceso iterativo, para intentar que ellos resulten ser el valor promedio del grupo/partición que representan.

### 2.5.2. Agrupamiento Jerárquico

El agrupamiento jerárquico es un método de minería de datos que busca particionar N instancias de forma que se pueda establecer distintos niveles de agregación o agrupamiento. El resultado de un agrupamiento jerárquico, puede expresarse en un dendograma, que no es más que una representación grafica en forma de árbol de los niveles de agregación, donde las raíces serian los grupos y las hojas las instancias, como se puede ver en la figura X.

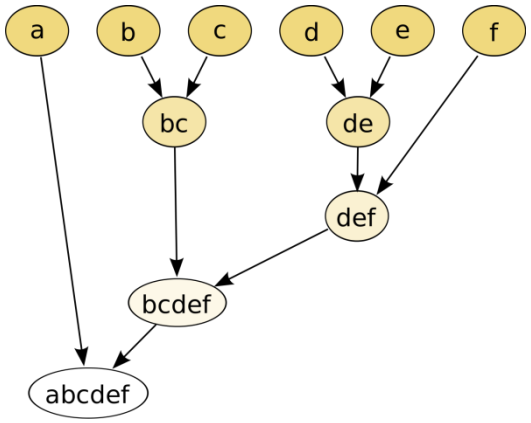


Ilustración 14 - Ejemplo de Dendograma (Wikipedia, 2015).

El agrupamiento jerárquico posee ciertas variantes, como por ejemplo: métodos aglomerativos, los cuales partiendo del total de instancias las va agrupando; o divisivos, donde partiendo de un grupo conformado por todas las instancias las va dividiendo.

Entre los método aglomerativos se pueden mencionar al método del mínimo, método del máximo, método de la media, método del centroide, método Ward, entre otros. La diferencia entre ellos es como calculan cuales individuos debieran agruparse en cada paso.

Si bien todos los métodos utilizan el cálculo de las distancias entre los individuos, como escoger quienes se unen en un grupo, en una determinada iteración, varía. En el caso del método de los centroides, ideado por Sokal y Micherner, se calcula la distancia entre los centroides de cada instancia, cosa que sin ninguna otra consideración, da como resultado que los grupos grandes influyen de manera especial en el cálculo del centroide. En el método Ward, ideado por el estadista Joe H. Ward, se emplea el criterio de la varianza mínima, el cual busca hacer la "mezcla" de instancias de tal forma que la varianza de la nueva agrupación formada represente el incremento mínimo de varianza (Universidad de Valencia, 2015).

### 2.5.3. Text Mining

El text mining (minería de texto) es una rama de la Minería de Datos dedicada al estudio de los escritos. El text mining se puede aplicar sobre data recogida de redes sociales, de encuestas de opinión, de discursos, de producciones literarias, y básicamente de cualquier entidad con texto. El text Mining se puede aplicar para hacer análisis de sentimiento del consumidor, al evaluar sus comentarios sobre un producto, se puede hacer clusterizacion de textos, para saber el tema de un determinado documento, o brindar información sobre un determinado documento.

La característica principal de los métodos de text mining es que trabaja con data no estructurada, puesto que no es posible forzar una estructura para la producción de textos, esta dificulta convertida en habilidad permite hacer estudios que de otra manera no serian posibles.

Una de las principales herramientas utilizadas en el text mining, es la matriz de términos por documento (o su equivalente matriz de documentos por términos), esta matriz indica por termino (palabra) y por documento (objeto del cual se saco el termino), un ponderaje. Este ponderaje se obtiene de una función, y hay varias funciones para tal objetivo, como por ejemplo frecuencia de termino, la cual cuenta el numero de apariciones de una palabra en el documento, o frecuencia de termino por el inverso de la frecuencia entre documentos (term frequency – inverse document frequency), la valora de manera especial aquellas palabras que abundan en un documento o en un grupo pequeño de documentos, y escasean en los demás.

### *2.5.3.1. Term Frequency – Inverse Document Frequency*

Frecuencia de término - Inverso de la frecuencia entre documentos (llamado de aquí en adelante como Tf-idf por sus siglas en ingles) es una medida numérica de cuan relevante es una palabra para un documento de entre un conjunto de documentos.

Su formula se puede dividir en 2 partes que se multiplican la parte tf (term frequency) y la parte idf (inverse document frequency). Si bien hay variantes de la formula, aquí se explicara una.

$$tf = \frac{|t|}{w:d}$$

La cual se puede leer como apariciones de una palabra entre el número de palabras del documento.

$$idf = \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$$

La cual se puede leer como el logaritmo de la cardinalidad de documentos entre el número de documentos donde aparece el término t.

Donde t es el termino, d el documento y D el conjunto de documentos. Finalmente la formula de Tf-idf se escribe como:

$$\mathrm{Tf} - \mathrm{idf} = \mathrm{tf}(\mathrm{t}, \mathrm{d})\ \mathrm{x}\ \mathrm{idf}(\mathrm{t}, \mathrm{D})$$

Un ejemplo de esta formula la encontramos en la página http://www.tfidf.com/ que indica:

"Considere un documento que contiene 100 palabras donde la palabra gato aparece 3 veces. La frecuencia del término para gato es entonces (3/100)= 0.03. Ahora, asuma que tenemos 10 millones de documentos y que la palabra gato aparece en mil de ellos. Entonces el inverso de frecuencia entre documentos es calculado como log(10.000.000/1.000) = 4. Finalmente el ponderaje tf-idf es el producto de estas cantidades: 0.03*4= 0.12".

### 2.5.4. Método del codo

El método del codo es un método de la minería de datos, que se utiliza para identificar el número de grupos presentes en una muestra. El método utiliza cualquier función que permita tener alguna métrica de error por grupo, por lo general se utiliza el algoritmo de kmedias y se utiliza como métrica la distancia promedio de los individuos al centro del grupo. La lógica es que se pueda construir una grafica donde se muestre la métrica de error (de distancia con respecto al centro del grupo) promedio de los individuos en la medida que se aumenta el k (número de grupos). Esto se hace con la esperanza de encontrar un k (número de grupos) donde la disminución de la métrica de error sea notable antes de ese punto y de menor categoría a partir de ese punto, como se puede ver en la figura X.

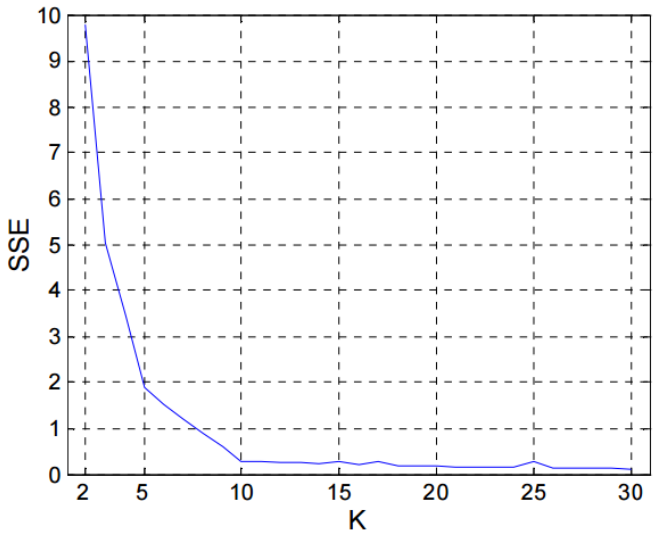


**Ilustración 15 - Ejemplo del método del codo (Fuente: http://elvex.ugr.es/decsai/intelligent/slides/dm/D3%20Clustering.pdf)**

## 2.6. XML

De acuerdo a w3, XML es el acrónimo en ingles de EXtensible Markup Language (Lenguaje de marcado extensible) el cual fue diseñado para describir la data, es una herramienta independiente del

hardware o software para el transporte de información. Su uso es amplio y se ha convertido en un estándar. Un ejemplo de XML es presentado a continuación:

```
<?xml version="1.0" encoding="UTF-8"?>
<note>
  <to> Tove</to>
  <from>Jani</from>
  <heading>Reminder</heading>
  <body>Don't forget me this weekend!</body>
</note>
```

Este lenguaje, no realiza operaciones sino que solo es un contenedor de data, sin embargo ofrece bondades, como por ejemplo, es posible crear sus propias etiquetas xml, por lo que cada campo de información tiene el nombre que se le considere más apropiado. El xml ofrece una estructura en árbol (similar a la de html) que puede ser de infinita profundidad.

XML tiene algunas reglas como son:

- Todo etiqueta de elemento que se abra, tiene que cerrarse
- Las etiquetas de apertura y clausura deben ser idénticas en nombre, se diferencia incluso las mayúsculas de minúsculas.
- La estructura de árbol debe ser respectada, no se debe cerrar un padre sino se ha cerrado un hijo
- Debe haber una etiqueta raíz
- Si la etiqueta posee atributos estos deben ser colocados entre comilas
- En el contenido de un etiqueta ciertos caracteres especiales no pueden estar presentes
- Los comentarios se escriben de la siguiente manera <!--comentario -->

Todo documento XML que cumpla con las reglas antes mencionadas es un documento XML bien formado.

## 2.7. JSON

De acuerdo a su propia documentación, JSON (JavaScript Object Notation) es un formato de intercambio de data de poco peso. Está basado en JavaScript Programming Language, Standard ECMA-262 3rd Edicion - Diciembre 1999. JSON es un formato de textos que es independiente del lenguaje utilizado, y sin embargo es bastante similar a estándares que resultan familiares para los programadores de lenguajes como C, Java, Python, Perl, JavaScript, entre otros. Esto hace de JSON una herramienta ideal para el intercambio de datos.

Algunos ejemplos y características de Json serian:

{"clave":"valor"}

- Todo JSON están encerado en llaves ( {} )
- Toda clave y todo valor JSON tiene que estar encerrado entre comillas ( "" )

["valor1", "valor2","valor3"]

- Los arreglos en JSON se expresan con corchetes.
- Los valores en JSON se serparan con coma ( , )

## 2.8. CSV

CSV es el acrónimo en ingles de Comma Separated Values (Valores separados por coma) y es un formato de almacenaje de data, comúnmente utilizado en las hojas de cálculo y bases de datos en primera forma normal (una sola tabla). Como la Internet Society indica no hay un estándar que lo defina, sin embargo es común que se respetan las siguientes reglas:

- Cada línea corresponde a una instancia de datos
- Todas las líneas deben tener el mismo número de campos
- El separador para cada campo es la coma (,).
- Los espacios en blanco antes y después de cada instancia son ignorados
- Caso que el valor de un campo lleve comas en su contenido, se debe hacer mecanismos para no confundir esta coma con la de los separadores.

CAPÍTULO 3

# MÉTODO DE DESARROLLO

Para dar luz a esta propuesta de trabajo especial de grado se utilizaron herramientas ya existentes, con las cuales realizar el estudio del comportamiento.

Existen varias metodologías para el desarrollo de proyectos de Minería de Datos como lo son KDD, CRISP-DM y SEMMA. Todas tienen pasos que son equivalentes y son de hecho parecidas en su procedimiento. Para este trabajo especial de grado se investigo bajo la metodología CRIPS-DM por ser neutral con respecto a las herramientas a utilizar, es independiente del tipo de problema a tratar, y no tiene restricciones de uso tipo propietario.

## 3.1. Metodología CRIPS- DM

CRISP-DM o El Estándar para Procesos de Entre las Industrias para la Minería de Datos (*Cross Industry Standard Process for Data Mining* en inglés), es una metodología de desarrollo para los proyectos de Minería de Datos (así como también algunos proyectos de Ciencias de Datos y Datawarehousing) que posee gran participación y relevancia en el mercado de los proyectos de minería de datos como metodología de desarrollo, tal como se puede constatar en la Ilustración 16. (Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer yRüdiger Wirth ,2000)

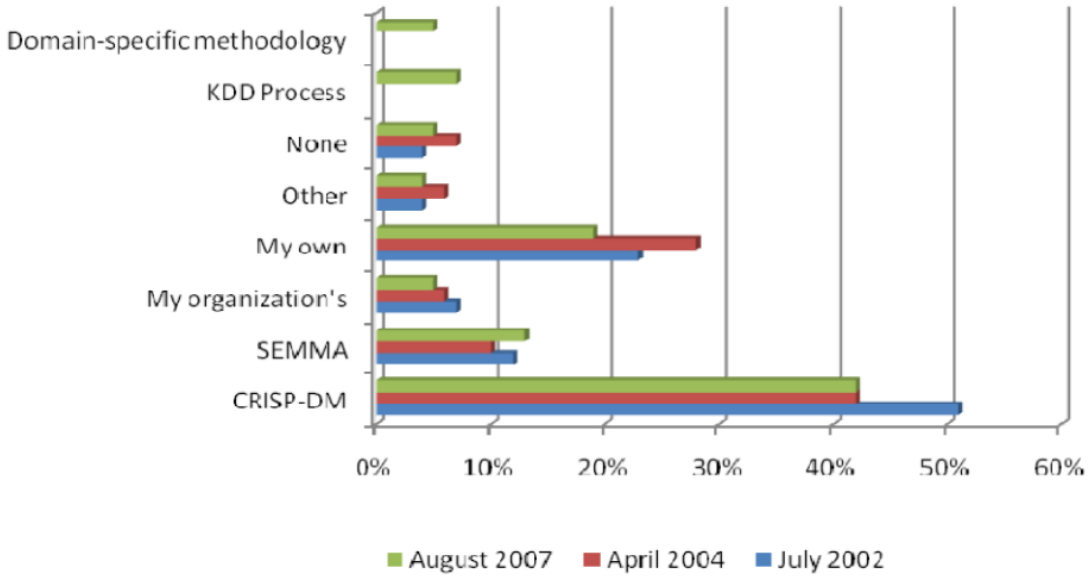


Ilustración 16 - Metodologías utilizadas en Minería de Datos (Oldemarrodriguez.com, 2015)

Como se puede apreciar en la Ilustración 17. CRISP-DM está definido por 6 pasos iterativos.

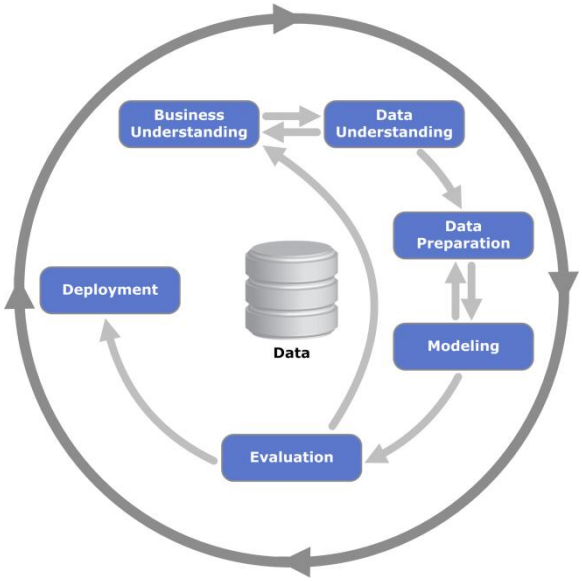


Ilustración 17 - Diagrama de proceso que muestra la relación entre las diferentes fases de CRISP-DM

Estos 6 pasos permiten el desarrollo de proyectos de minaría de datos. A continuación se describen los mismos:

- **Comprensión del negocio:** se establecen los objetivos y planes del proyecto de minería de datos atendiendo las necesidades y requisitos empresariales, determinando objetivos, requisitos, supuestos, restricciones, entre otros.

- **Comprensión de Datos:** esta fase se basa en comprender la data, sus fuentes, sus tipos, su creación, su utilidad, su calidad, entre otros. Esta fase permite saber de que se dispone y permite identificar posibles puntos clave para encontrar relaciones entre los datos. Esta fase suele llamarse también fase de exploración de datos.

- **Preparación de datos:** Esta fase cubre, entre varias cosas, todo lo referido al proceso de ETL (*Extraction, Tranformation and Load* o Extracción, Transformación y Carga en español). En esta fase se hace la búsqueda de valores atípicos, se busca estadísticas entre los datos, se normalizan los datos si hiciera falta, se seleccionan las instancias de data (observaciones) más útiles para el proyecto, se seleccionan las variables (o dimensiones) mas útiles para el proyecto, se limpian los datos de ruidos o outliners que puedan entorpecer el estudio, entre otros procesos.

- **Modelado:** se seleccionan y aplican varios algoritmos de minería de datos y se calibran los parámetros para obtener óptimos resultados. Hay varias técnicas que tienen requerimientos

específicos para la forma de los datos, por lo que frecuentemente es necesario volver a la fase de preparación de datos. De esta fase se suele obtener modelos, que debe ser evaluados (indicar que tan bueno/preciso es el modelo)

- **Evaluación:** de los resultados del estudio se estudian los resultados desde una perspectiva de análisis de datos y se determina si los resultados satisfacen los requisitos, si el proceso fue llevado a cabo correctamente y se determinan cuales eran los próximos pasos.

- **Despliegue:** Se dan a conocer los resultados, presentándolos en un formato legible y útil para la parte(s) interesada(s), puede ir acompañado de un plan de implementación, monitoreo y mantenimiento. En esta fase se entrega el informe final del proyecto y se valora al proyecto.

## 3.2. Herramientas a utilizar

Para la obtención de la data, se empleo el sistema manejador de bases de datos (SMBD) MySQL. Esta decisión fue tomada en virtud de que la base de datos con la muestra del estudio es MySQL. Se utilizo la versión para Linux del SMBD dado que los tiempos de respuesta en maquinas Linux con dicho SMBD son mejores que su parte en Windows.

De esta base de datos se extrae un archivo con toda la data, la cual fue divida en trozos del tamaño de una instancia de metadatos. Este proceso se hizo utilizando el lenguaje de programación Java y una interfaz de desarrollo (IDE) llamada Netbeans.

Finalmente, el grueso del trabajo, el análisis exploratorio, transformación, modelado, evaluación y despliegue de todo el proyecto se realizo en lenguaje R. Dicho lenguaje cuenta con una IDE llamada RStudio, un repositorio de paquetes (equivalente a librerías) llamado CRAN, y un paquete llamado Shiny con el cual se construyo un prototipo para probar el sistema una vez concluido.

# CAPÍTULO 4
# DESARROLLO DE LA SOLUCIÓN

El Buscador Académico Venezolano (BAV) es un proyecto de ONCTI de Venezuela cuyo objetivo es brindar un punto central desde el cual hacer consultas a una cantidad de repositorios de objetos de aprendizaje venezolano. Actualmente el BAV es la federación de varias instituciones las cuales suman 49 repositorios de objetos de aprendizaje. Ellos se listan aquí:

1. Revistas Saber ULA
2. Tesis de Posgrado de la Universidad del Zulia
3. Universidad de Oriente
4. Biblioteca Digital Revichluz Serbiluz
5. Saber UCV
6. Memoria Cientifica y Academica de la Universidad de Carabobo, todo el...
7. DSpace de la Universidad de los Andes - Merida – Venezuela
8. UCV - Biblioteca D.F. Maza Zavala - Facultad de Ciencias Económicas y…
9. UCLA Biblioteca Virtual
10. UCV - Biblioteca Belarmino Lares - Facultad de Humanidades y Educación
11. UCV - Biblioteca Boris Bunimov Parra - Facultad de Ciencias Jurídicas y...
12. UCV - Biblioteca Centro de Estudios Integrales del Ambiente (CENAMB)
13. UCV - Biblioteca Dr. Alberto Rivero - Facultad de Medicina - Instituto...
14. UCV - Biblioteca Dr. Oswaldo Enríquez Isava - Facultad de Farmacia
15. Biblioteca Central - Universidad Central de Venezuela
16. UCV -Biblioteca Instituto de Biología Experimental
17. Acta Bioclínica ;UCV
18. Acta Botánica Venezuelica
19. Revista de Pedagogía
20. Revista Anales Universidad Metropolitana
21. Boletín del Centro de Investigaciones Biológicas
22. Capítulo Criminológico
23. COEPTUM
24. Cuestiones Jurídicas
25. EDUCARE
26. Interacción y Perspectiva
27. INVESTIGACION Y POSTGRADO
28. LAURUS
29. LETRAS

30. Omnia
31. Opción
32. PARADIGMA
33. Politeia
34. Portafolio
35. REDHECS
36. Revista de Ciencias Sociales
37. Revista de la Facultad de Ciencias Veterinarias, UCV
38. REDIP - Revista Digital de Investigación y Postgrado
39. Revista Latinoamericana de Metalurgia y Materiales
40. Revista Tecnocientífica URU
41. Revista Venezolana de Gerencia
42. Universidad de Oriente - Venezuela. Consejo de Investigación. Revistas...
43. SÍNDROME CARDIOMETABÓLICO
44. TELEMATIQUE
45. Telos
46. Anuario ININCO / Investigaciones de la Comunicación
47. Terra. Nueva Etapa
48. TIEMPO Y ESPACIO
49. DSpace de la Universidad de los Andes - Mérida – Venezuela

Dado que el repositorio BAV da soporte al protocolo OAI PMH, y que el estándar de metadatos utilizado es Dublin Core, se ha seleccionado al BAV como la fuente de datos para este estudio. Para la cual fue muy impórtate la colaboración del especialista Profesor José R. Sosa dado que labora en dicha institución.

## 4.1. ARQUITECTURA DE LA SOLUCIÓN

El proyecto fue dividido en 2 partes. La obtención de la data por un lado y la minería de datos sobre los mismos. Para la primera fase fue utilizando el sistema manejador de bases de datos (SMBD) MySQL el cual permitió extraer la tabla y columna exactas donde se alojaba la data de una base de datos proporcionada por el Profesor José R. Sosa. Posteriormente se preparo la data para el consumo del programa de minería usando lenguaje Java y haciendo uso de una interfaz de desarrollo (IDE) llamada NetBeans; mientras que para la segunda parte, la minería de datos, se realizo en lenguaje R,

en un segundo programa, carente de interfaz, que permitía el estudio de la metadata, haciendo base en una interfaz de desarrollo (IDE) llamada RStudio.

Sabemos que el repositorio es una federación de repositorios que emplean OAI, así que como indica en su documentación Open Archives Iniciative (2015) el protocolo utiliza peticiones y respuestas HTTP para la obtención de la metadata, la cual viene en formato XML y debe seguir un estándar determinado. Como se explica en el marco conceptual, es deducible que la data esta en XML, y bajo un estándar de metadata conocido, en este caso Dublin Core.

La data esta dentro de un archivo .sql de la base de datos MySQL, por lo que se debe cargar en el mismo sistema manejador de bases de datos y un vez cargada, se estudia la data disponible, extrayéndose a un archivo xml el fragmento de data que es de interés, específicamente.

Este nuevo archivo de datos de formato xml, tendrá toda la data en un solo archivo, y es conveniente tenerla en archivos separados, por lo que un programa, en lenguaje Java, que separe las instancias de metadata a una por archivo .xml resultara ventajoso.

Para atender la arquitectura del programa encargado de la minería de datos, de estableció un funcionamiento local, utilizando el sistema de archivo de Microsoft en su sistema operativo Windows versiones 7 Home y 8.1. Las instancias de metadata, en formato XML serian leídas por el programa en R, el cual haciendo uso de sus librerías y virtudes, en el área estadística y minería de datos, se sacaría provecho en el manejo de grandes volúmenes de información.

Parte de este provecho vino de la mano de las librerías de las que dispone R, como por ejemplo: doParallel de Revolutions Analytics, que permiten el procesamiento paralelo de información, así como de la librería Matrix de Douglas Bates and Martin Maechler que permite trabajar con estructuras de datos de matriz tipo sparce (con baja densidad) o la librería tm de Ingo Feinerer, Kurt Hornik y Artifex Software, que brinda cantidad de métodos indispensables en las labores de text mining. Asimismo otras librerías contribuyeron sino de manera indispensable, de utilidad para presentar los resultados o crear los modelos de minería, como fueron las librerías wordcloud que permite crear nubes de palabras, plyr que permite la unión de matrices con auto-completación de valores en las columnas, dendextend que permite darle mayor significancia a los dendogramas y rattle que sirvió de enlace entre las tareas de agrupamiento jerárquico y clasificación por distancias.

Las maquinas utilizadas para el estudio, fueron principalmente:

- Una PC Windows 8, 20 Gb en RAM, i5 de 4 núcleos
- Una laptop con Windows 7 Home y con Linux Mint 12 de 4 Gb de RAM, i5 de 4 núcleos.

En la laptop con Linux se ejecutaron las labores sobre la base de datos, aprovechando que es más rápida la carga y descarga de data a la base de datos desde ese sistema operativo. En la PC, siendo la maquina más potente, se ejecuto el trabajo de la minería de datos así como la aplicación de Java. La

laptop en su presentación Windows fue utilizada para mostrar progresos y hacer consultas sobre el trabajo.

## 4.2. ANÁLISIS Y DISEÑO DE LA SOLUCION

El archivo sql proporcionado por el profesor Sosa, es de 2 GB de tamaño, por lo que se sabe que la base de datos es grande. Los sistemas operativos de la familia Linux suelen operar a mayor velocidad que los Windows, posiblemente por simplificaciones de funciones. Esta simplificación que brinda como consecuencia mayor velocidad en las operaciones es útil y explotable en situaciones donde la muestra de data es demasiado grande.

Cargada la data en el SMBD MySQL, se podrá explorar la base de datos para localizar el punto exacto o puntos exactos donde se aloja la información, la cual deberá ser descargada en formato xml. El SMBD solo puede crear un archivo de salida por instrucción ejecutada, así que en primera instancia se descargaría toda la información a un solo archivo XML.

Es cómodo, y más fácil de trabajar la data si esta viene presentada en la modalidad una instancia de metadata por cada archivo .xml, así que utilizando un programa en Java se dividirá ese único archivo xml en varios.

Por su parte, el programa en R para la minería de datos también necesito su propio diseño

Se podría pensar que la data obtenida, utilizaría el estándar de metadatos a un nivel suficiente para que se pudiera hacer un estudio utilizando los campos de metadata del estándar, sin embargo esto no fue así. Hay 6 campos con un 90% de escases de data y otro con más de 60 %, esto significa que en esos campos de metadata no se tiene suficiente información para hacer algún estudio que diera resultados sobre toda la muestra.

En la minería de datos, esto es un problema, porque el exceso de escases en los datos hace difícil determinar patrones, al no haber datos que comparar. Asimismo, el exceso de datos diferentes hace difícil encontrar patrones puesto que no hay coincidencias, este segundo caso se puede apreciar en las columnas como fecha, autor o identificador.

Un ejemplo sencillo del problema antes mencionado seria intentar agrupar personas, basándose solamente en el numero de cedula. No se puede hacer mucho con estos datos, más que ordenarlos entre números mayores que X  y menores que X. Con ese dato, diferente para todos, y sin ningún otro dato, no se pueden encontrar patrones.

Como no es posible sacarle ninguna utilidad a un dato que no existe, era necesario sacarle el máximo provecho a los campos de metadata que si fueron proporcionados y como estos campos son del tipo cadena de caracteres, se hace necesario un estudio de text mining.

Dicho estudio permitiría agrupar los documentos en función de las palabras que se utilizan para describirlos. Es necesario que dicha agrupación sea correcta, y que brinde información verificable y replicable, por lo que se tienen que usar métodos de agrupamiento de alta precisión tal como el agrupamiento jerárquico. A pesar que utilizar dicho método resulte oneroso en términos computacionales, dará resultados certeros y replicables.

Una vez se pueda agrupar a los documentos en comunidades, se podrá determinar cuántas comunidades hay. En este punto el criterio del investigador entrara en juego, porque cada instancia puede ser su propia comunidad, y todas las instancias pueden ser una comunidad. Encontrar el punto de equilibrio entre la granularidad deseada y los resultados encontrados deberá ser calculado subjetivamente por el investigador.

Puede escoger un nivel de granularidad de facultades universitarias, donde cada grupo equivale a una facultad de la universidad, asimismo se puede escoger un nivel de granularidad de escuela o una granularidad de mención dentro de una carrera.

Se tiene que tomar en cuenta que unos grupos estarán mejor descritos que otros, por lo que no todos pueden llegar al mismo nivel de detalle; así como también se tiene que tomar en cuenta que algunas aéreas del conocimiento si bien están separadas son muy parecidas o hermanas. Se puede brindar como ejemplo el caso de enfermería y medicina, matemáticas y física, estos ejemplos y muchos otros son casos de riesgo donde conseguir la separación de estos grupos hermanos podría causar un nivel de granularidad tan bajo, que primero se consiga dividir a un grupo en 2 o más partes que no debieran estar separadas como es el caso de arquitectura y urbanismo o administración y contabilidad.

Definido el número de grupos, cuales son y qué términos los caracterizan, se puede idear el modelo para clasificar a cualquier instancia en alguno de esos grupos.

## 4.3. DESARROLLO

El desarrollo de este proyecto estuvo caracterizado por el ensayo y error, en el cual se probaron muchos caminos antes de dar con el correcto. A lo largo del proyecto, este ensayo y error permitió dar certezas y conocimientos que serian utilizadas en el experimento acertado.

El desarrollo de la investigación se realizo en dos partes principales y en secuencia. Primero la recuperación de datos y luego la minería de datos. En la primera parte, es decir la recuperación de

datos, se invirtieron una tercera parte del tiempo total de investigación, mientras que para la segunda, dos tercios del mismo.

### 4.3.1. Recuperación de datos

Los datos vienen en un archivo .sql y por lo tanto se trabajo sobre la base de datos. La base de datos contiene 56 tablas, y aquella que guarda la data se llama records, específicamente en su columna contents (records significa instancias y contents, contenido).

La información contenida en ese campo de tabla de la base de datos fue enviada a un archivo xml, utilizando al mismo SMBD para hacer dicha descarga. Este archivo seria dividido en tantos mini-archivos como instancias existieran dentro del original. Como se podía leer en una mínima muestra del xml con todas las instancias, estas seguían el dublin core como estándar, con lo que se sabía cuáles eran los limites de cada instancia y se pudieron dividir.

Originalmente se desconocía cuantas instancias estaban almacenadas en dicha base de datos, sin embargo luego se conocería que habían 161.125 instancias en perfecto dubin core junto a otras 2.714 que estaban en lo que aparenta ser dublin core también pero sin etiquetas que identificaran cada campo del estándar.

Esta data, donde cada archivo es una instancia de metadata, está en perfecto estado para ser consumida por el programa de minería de datos.

### 4.3.2. Análisis Exploratorio de Datos

Teniendo toda la metadata era necesario pasarla de su estado en xml a una estructura tipo tabla llamada dataframe, que pudiera ser utilizada por el programa en R. El dataframe es una tabla que cumple las mismas propiedades de los csv (coma separate values). Para eso se creó una función que transforma cada XML individual en una tabla dataframe de una sola fila (una sola instancia), donde las columnas serian los campos de metadata. Creado ese conjunto de 161.124 dataframes se unirían todos en un solo dataframe con las 161.124 instancias (filas).

Las operaciones de lectura de 161.124 archivos son lentas puesto que son del tipo entrada/salida (con respecto al disco duro), así que se hacía necesario un medio de acelerar los procesos. Se recurrió a la computación paralela, explotando las características de la maquina con respecto a sus 4 núcleos, utilizando el paquete doParrallel de R y en especifico la instrucción foreach.

**Tabla 6 - Tiempos de ejecución de lectura de los XML a CSV.**

| test | replications | elapsed | relative | user.self | sys.self | user.child | sys.child |
|---|---|---|---|---|---|---|---|
| FOREACH1 | 1 | 632.51 | 2.224 | 183.53 | 8.89 | NA | NA |
| FOREACH2 | 1 | 392.37 | 1.380 | 177.70 | 7.36 | NA | NA |
| FOREACH3 | 1 | 293.86 | 1.033 | 168.55 | 6.39 | NA | NA |
| **FOREACH4** | **1** | **284.39** | **1.000** | **176.31** | **6.74** | **NA** | **NA** |

Leyenda:

- Test indica el nombre de la prueba. Cada FOREACHX donde X es un número del 1 al 4, es el conjunto de instrucciones FOREACH ejecutada con 1 a 4 núcleos respectivamente.
- Replication es el número de veces que se repitió esa prueba ( 1 sola ocasión en todos los casos)
- Elapsed es el tiempo calculado en segundos que se percibe en el mundo natural la duración de ejecución del conjunto de instrucciones
- Relative es una comparación basada en regla de 3 con respecto al mejor valor obtenido en la prueba
- User.self es el tiempo que el cpu gasto en la instrucción
- Sys.self es el tiempo que el sistema tardo en adecuarse para ejecutar la instrucción (cargar los datos en los registros, etc)
- Sys.child no están disponibles en maquinas windows e indica una unidad de medida de tiempo.

Como se puede ver, explotando el poder del paralelismo se podía convertir todos los XML a dataframe en 5 minutos (300 segundos), y de no hacerlo se podía tardar 10 minutos esta operación. Dicha actividad de lectura y transformación del xml al dataframe fue realizada varias veces, pero en específico se necesitaban 3 corridas, una para hacer un análisis exploratorio de datos, otra para el proceso de text mining y la final que serviría como estructura de índice para el sistema de recomendación.

Para el caso de análisis exploratorio, el valor real era sustituido por verdadero o falso dependiendo de si la data está presente o no. En el proceso de text mining, se hacia un data frame de una sola columna, donde esa columna contenía la concatenación del título, tema y descripción del trabajo (title, subjet y description). Y el índice fue también una sola columna que indicaba el valor del identifier de la instancia, o sea es una clave primaria. Es de suma importancia identificar que todos los dataframes son construidos partiendo de la misma data, tratada de forma diferente.

El análisis exploratorio de la metadata cosechada revelo la inmensa escases en la data en la muestra. El análisis exploratorio se abarcaba sobre: cantidad de datos y la variedad en los datos disponibles. A continuación se presenta a manera de gráficos algunas de las estadísticas obtenidas.

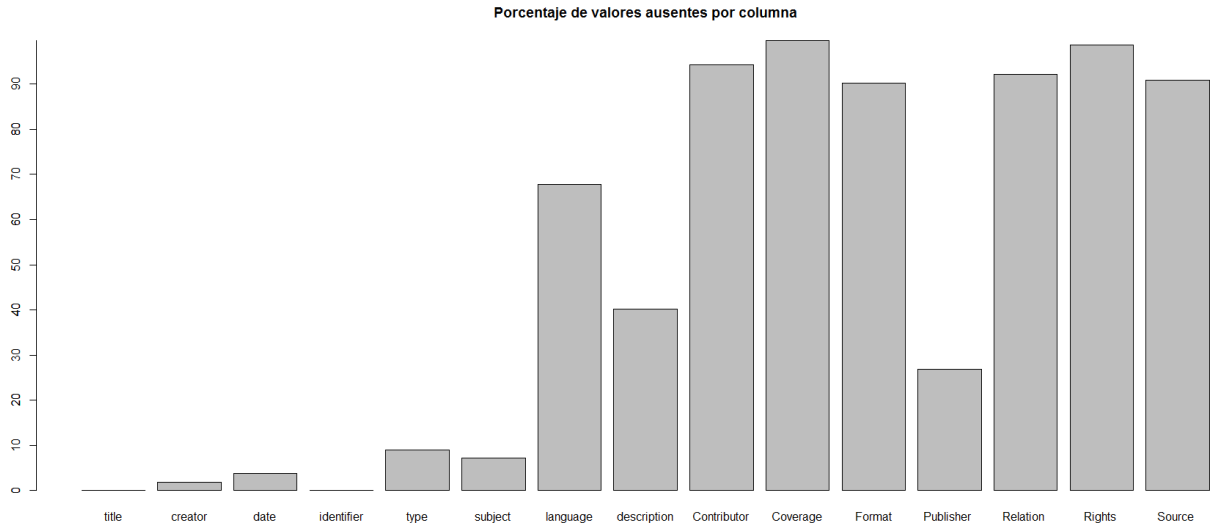


**Ilustración 18 - Porcentaje de valores ausentes por columna**

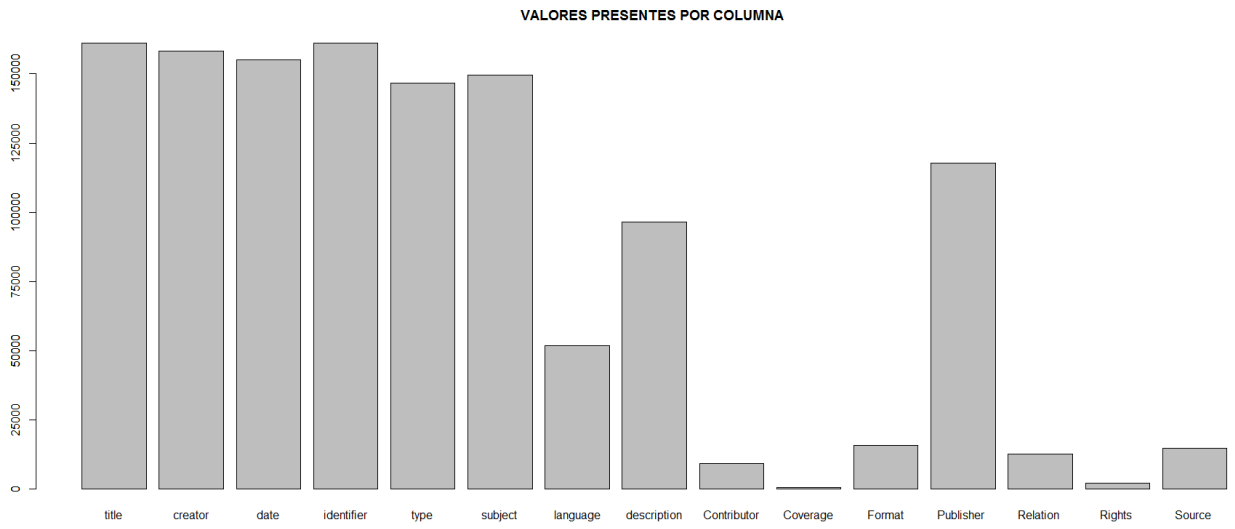


**Ilustración 19 - Valores presentes por columna**

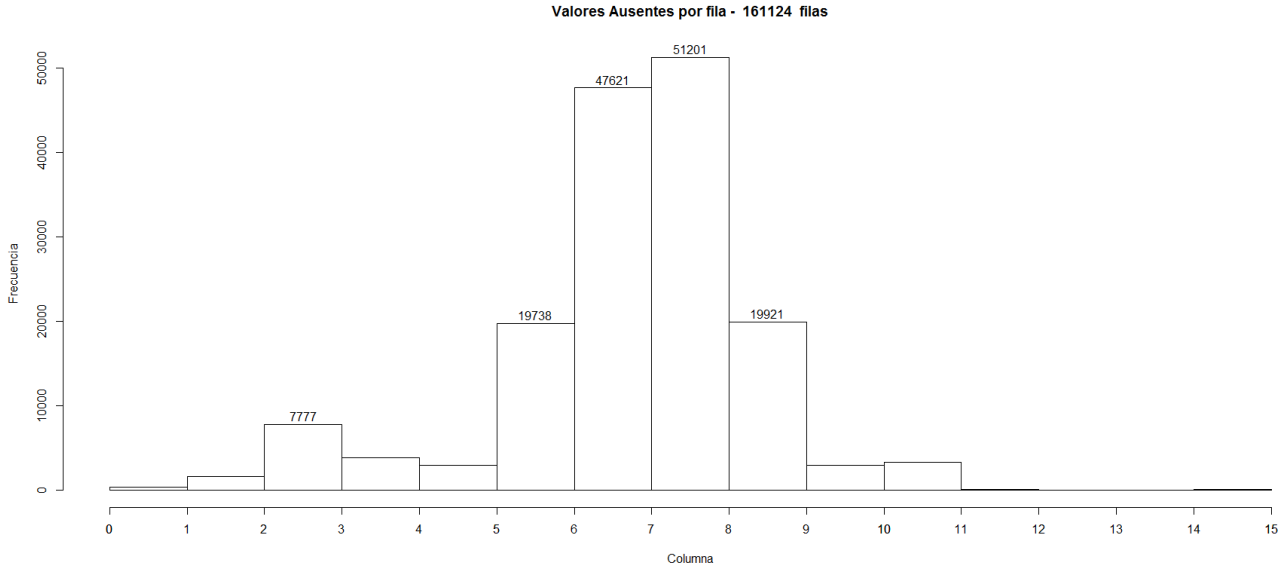


**Ilustración 20 - Numero de valores ausentes por fila**

Como se puede apreciar en las Ilustración 18 - Porcentaje de valores ausentes por columnaIlustración 18 y Ilustración 19, las mismas son complementarias, y señalan lo mismo, pero una midiendo la ausencia de valores y otra la presencia. Esas graficas nos indican en donde hay datos y en donde no. La Ilustración 20, nos dice que tanta data falta por instancia.

Podemos establecer relaciones entre las ilustraciones, por ejemplo: por instancia suelen faltar 6, 7, 8 u 9 campos, y existen 6 campos prácticamente inutilizados junto a otros 3 con poco uso. Así sabemos que los campos más utilizados son los 6 restantes, que son: titulo, autor, fecha, identificador, tipo y tema.

Utilizando esas 3 graficas, se pueden descartar ideas como hacer minería sobre los datos en bruto. La necesidad de transformar los datos antes del modelado es que valores que debieran seguir una distribución "cola liviana" como derechos, formato cobertura, están ausentes, para casi todos los casos; mientras los valores de titulo, autor, fecha, identificador, que si están presentes, son campos con muchos valores diferentes, donde no cabe esperar que existan muchas coincidencias que indiquen patrones. El único campo que por sí solo ayudaría es el de subjet (tema) y esa alternativa no es usable puesto que basar todo el estudio en un solo campo seria limitarse.

Por lo que para el trabajo de text mining se decidió "crear" un campo que mezclara el título, el tema, y la descripción, para que fungiera todo como un solo texto que luego sería estudiado.

Puesto que el valor con menos ausencias es el identifier, y este concuerda con el identificador del recurso el cual debe ser una clave primaria, se decidió construir el índice sobre este campo. De esa manera la misma instancia proveerá un identificador univoco con respecto a las demás.

### 4.3.3. Preparación de la Data

De moda hay un 90% de escases de data por campo de metadata, esto significaba que la mayoría de los campos de metadata no tenían suficiente información para hacer algún estudio que diera resultados sobre toda la muestra.

En la minería de datos, esto es un problema, porque el exceso de escases en los datos hace difícil determinar patrones, al no haber datos que comparar, y el exceso de datos diferentes hace así mismo difícil encontrar patrones, puesto que no hay coincidencias. Un ejemplo sencillo seria intentar agrupar personas, basándose solamente en el número de seguro social. Este dato no arrojaría gran información, más que ordenarlos entre número mayores que X y menores que X. Con ese dato diferente para todos, y sin más ningún otro dato, no se pueden encontrar patrones.

Como no es posible sacarle ningún beneficio a un dato que no existe, era necesario sacarle provecho a los 4 campos de metadata que si eran proporcionados: el titulo, la descripción, tema, y la localización. Para esto habría que estudiar esos 4 campos, dando como resultado que el titulo, la descripción y tema son del tipo cadena de caracteres, mientras la localización es una dirección web que suele terminar en un número de identificación; eso es equivalente a una clave primaria, como una cedula, por lo que ese campo quedaba descartado.

Se sabe que los 3 campos útiles, eran del tipo cadena de caracteres, se podía hacer un estudio de Text Mining. El text mining es una forma de minería de datos que utiliza como pieza principal una matriz de términos por documentos (o su equivalente de documentos por término), donde las relaciones entre los documentos (instancias de metadata) vienen dadas por la utilización de determinadas palabras en el mismo.

Para hacer la matriz de términos por documento (la llamaremos tdm, por sus siglas en ingles term document matrix) más útil posible, es necesario hacer una limpieza y transformación de la data. Los artículos, conjunciones y números debieron ser eliminados por ser muy frecuentes y no aportar información, las palabras tendrán que no presentar diferencias de mayúsculas o minúsculas y las cadenas de caracteres deben no tener espacios en blanco para así no indicar como diferentes a la misma palabra, se deben retirar los signos de puntuación y además reducir las palabras a su raíz para no limitar su influencia como verbos conjugados.

Este conjunto de temimos deben ser ponderados, para entonces poder empezar a trabajar. Hay varias formas de calificar los términos, la más común y fácil es contar el numero de apariciones de una palabra, sin embargo la utilizada en este trabajo fue tf-idf (tern frequency – inverse document frequency o en español, frecuencia de términos por el inverso de la frecuencia en documentos). Tf-idf se puede explicar de forma sencilla, como la cuenta de apariciones de un término en un documento, multiplicado por el inverso de la aparición de dicho termino en otros documentos, esto permite diferenciar a cada documento por el uso de sus términos. Un término que aparezca en muchos documentos obtendrá una baja ponderación; un término que aparezca en pocos, obtendrá una mayor ponderación y puesto que la ponderación nace de una fórmula matemática se permiten valores intermedios.

Con esta matriz es posible realizar labores de minería de datos, encontrando relaciones entre los documentos (instancias) a partir de las palabras empleadas (columnas o dimensiones). Esta matriz puede que sea muy grande pero poco densa y eso acarreara problemas de procesamiento de información y manejo de la misma, lo que ocasionara la utilización de varios GB de memoria y de días procesando la data. Para afrontar este problema de volumen (Big Data) se puede adoptar una de dos resoluciones, o se prepara un cluster de computadoras para el procesamiento distribuido, o se reduce la complejidad del problema.

Puesto que la intención es encontrar grupos a partir del contenido de las características que describen a las instancias, y que estos grupos sean hechos basándose en la mayoría, y no en la totalidad, se puede optar por un enfoque de reducir la complejidad del problema, sin incurrir en un exceso de pérdida de información. Puesto que la tdm son matrices con poca densidad (debido a que no todas las palabras aparecen en todos los documentos) se tendría que estudiar cuan densa es la matriz y que términos no ayudan significativamente a agrupar las instancias en conjuntos.

Para estudiar la densidad de utilización de términos, se decidió presentar la matriz con puntos negros en donde se utilice un término y puntos blancos donde no. Como las dimensiones (complejidad) del problema podían ser de gran tamaño (muchas palabras) se decidió quitar las palabras con una densidad inversa de 0.999, esto quiere decir que toda palabra que sea seleccionada debería aparecer en al menos 0,001 de los documentos. Para una muestra de 160.000 instancias esto significaría que aparezca en al menos 160 documentos, sino seria descartada. Haciendo esto pasamos de unas 300.000 palabras a 4269, cosa que es un gran avance, sin embargo esto tiene un precio, puesto que 958 instancias de datos perdieron la totalidad de las palabras que los definían.

A pesar de la gran reducción de la complejidad del problema, este sigue siendo muy grande. La maquina que se utilizo para este estudio, no tenía el poder para procesar esta data. Así que una vez más se hace necesario reducir la complejidad del problema.

Buscando por instancias las palabras más importantes, se constato que prácticamente para todos los documentos las palabras mejor ponderadas no eran al mismo tiempo las palabras más importantes de otros documentos. Esto significaba en términos prácticos que reducir la dimensión del problema partiendo de la ponderación de cada palabra en cada documento no daría resultados significativos, puesto que muy pocos términos serian descartados de esta forma.

El plan para reducir la complejidad del problema (4269 dimensiones) tendría entonces que venir de la suma de las ponderaciones por término entre todos los documentos, dicho de otra manera, buscar las palabras más importantes para todos los documentos, y no para cada documento individual. Esta suma indicaría cuales eran los términos más valiosos, aquellos que se repetían dentro de un mismo documento, y en varios otros en numerosas ocasiones; sin embargo no se repetían lo suficiente como para descalificarlos de poco significativos. De estos se seleccionaron los 500 cuya suma dio mayor valor. Con esto se tenía un problema manejable por un computador. En la Ilustración 211 se muestra una grafica de densidad de utilización de términos, esta vez con los 500 cuya suma arrojo el mayor valor.

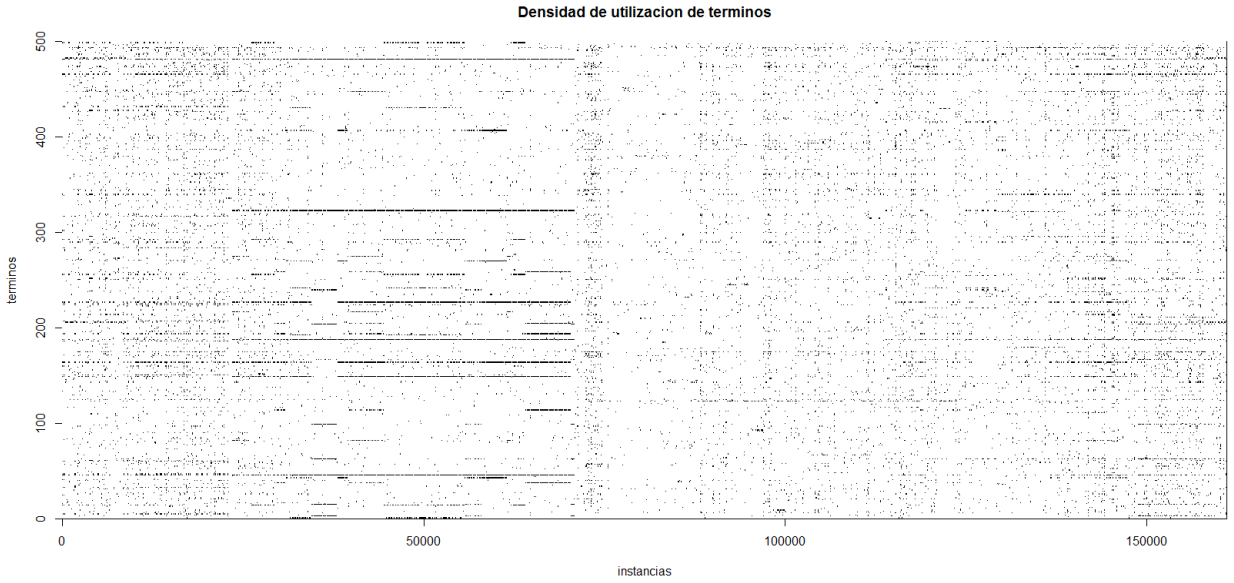


**Ilustración 21 - Densidad de utilización de términos, con las 500 palabras más valiosas.**

Gracias a la reducción de la complejidad del problema se pueden pensar en hacer tareas de minería de datos. Se pasó de tener 161124 instancias de metadata que conformaban unos 300.000 términos diferentes en la tdm(ocupando 103 MB como matriz sparse, no calculado como dataframe), a tener las mismas 161124 instancias con 4269 términos (56.5 MB de matriz sparse o 5.1 GB de DataFrame) después de quitar los términos que se repetían menos de 160 veces y finalmente las mismas 161124 instancias con 500 términos solamente (32.6 MB de matriz sparce y 654 MB de Dataframe).

Es necesario verificar que es posible alcanzar los objetivos (obtener grupos) con los datos que se tienen. Una técnica para saber de qué datos se dispone, siendo las dimensiones las palabras, es la nube de palabras (wordcloud), la cual podemos apreciar en la Ilustración 2222.

.

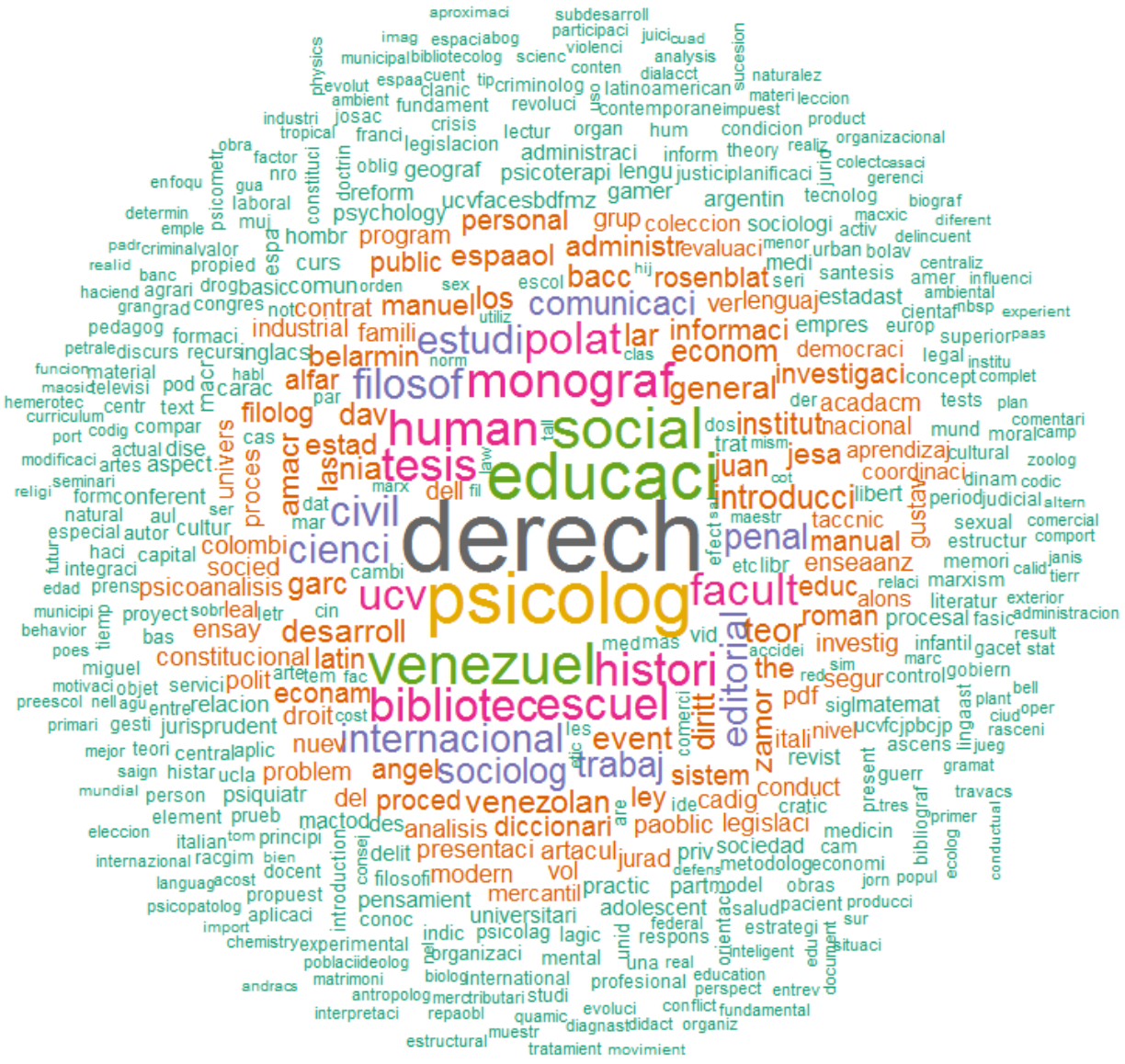


**Ilustración 22 - WordCloud de los 500 términos utilizados.**

Un rápido estudio exploratorio de los términos, nos demuestra que muchos de ellos coinciden con aéreas del estudio y saber humanos, con áreas de estudios universitarios y más. Esto, da pie a la continuación de la investigación, puesto que los resultados que se pueden obtener de dicha muestra encajan con comunidades que pueden ser comprendidas por los seres humanos.

Como último paso de transformación de la data, contamos aquellas instancias que quedaron sin datos durante el proceso de reducir dimensiones (quitar palabras) del dtm. Esta selección es útil puesto que se descubrió durante la fase de exploración de la data que habían 85 instancias sin ningún campo de

metadata, ósea instancias vacías o eliminadas. Al respecto, quedaron 958 instancias sin palabras después del proceso de quitar los términos que aparecían menos de 160 veces, y finalmente 1716 instancias también se quedaron sin palabras en el proceso de reducir dimensiones hasta 500 palabras solamente, lo que significa que hay 2759 instancias que no pueden ser descritas por no tener data y que no hay forma de agruparlas correctamente en algún grupo. Lo cual deja 158365 instancias útiles.

Estas instancias no fueron retiradas de la muestra, y ello es la decisión óptima, puesto que el sistema de recomendación siempre tendrá que dar una respuesta, aun cuando la respuesta sea "no hay respuesta". El sistema de recomendación no puede dejar al usuario sin contestación, y siendo ese el caso, lo más congruente con el sistema, es permitir respuestas incluso para esas eventualidades. Así que hay que recordar que de las 161124 instancias hay 2759 sin data lo que deja 158365 instancias útiles; todo lo cual representa el 98.28% del original de la muestra.

### 4.3.4. Modelo de Datos

Teniendo esta solida muestra, se intento determinar el número de comunidades presentes en la misma, mediante el método del codo. Sin embargo como se puede constatar en la Ilustración 233, los resultados no fueron conclusivos.

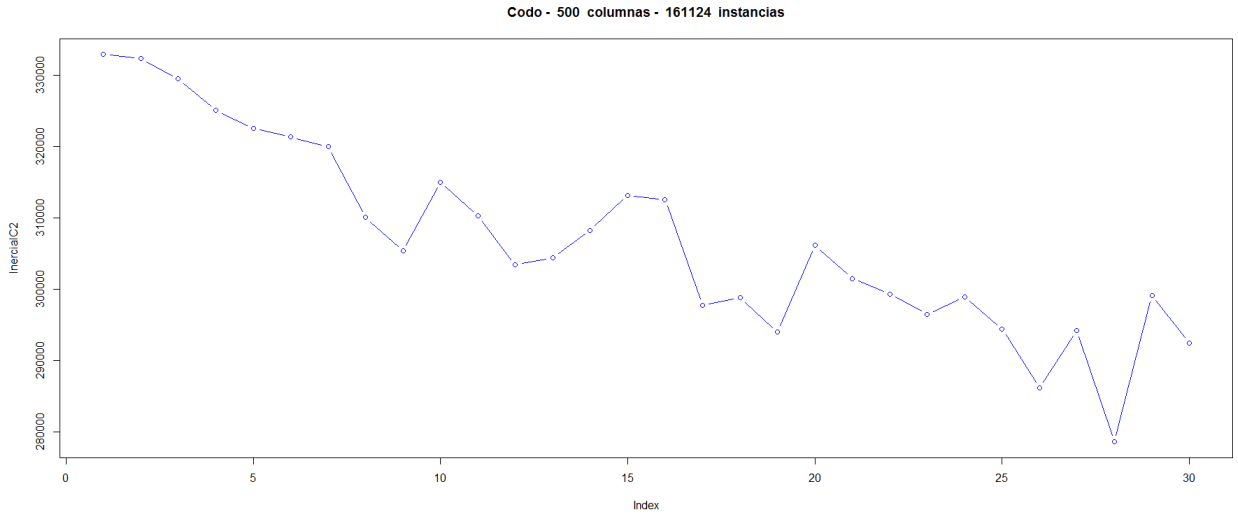


**Ilustración 23 - Método del Codo para repositorio, 500 columnas.**

Tomando toda la muestra de 161124 instancias se calculo un codo con un k máximo de 30 (eje X), sin embargo la reducción de las distancias hacia el centro de cada grupo nunca fue vertiginosa. Recordemos que en los codos la bajada debe ser muy fuerte de unos 80 grados de inclinación. Esto significaba que todos los grupos se parecen entre sí, y que hay instancias difusas (que pertenecen a más de un grupo al mismo tiempo). Estos casos son posibles, cuando en la descripción de un trabajo

se emplean términos de otras áreas, lo cual hace que los individuos se parezcan los unos a los otros. Un trabajo puede estar fuertemente relacionado con 2 áreas, por ejemplo, una aplicación digital para estudios biológicos es un trabajo del área de la computación, pero va a hablar mucho de biología en su contenido, esto hará que el trabajo se calificado como un intermedio entre computación y biología.

Se podría argumentar que el número de columnas eliminadas fue tan grande que se incurrió en una pérdida de información, por lo que intentar hacer el codo con 4269 columnas pudiera dar mejores resultados, sin embargo la Ilustración 24, demuestra lo contrario. El aumento de la complejidad del problema no induce mejores resultados. De hecho si se observa el eje Y de la grafica, que corresponde con la distancia promedio de los individuos al centro del grupo, se puede observar, que incluso dicha distancia aumento, y que el aumento de la complejidad del problema incurre en un sobre-entrenamiento que hace peores modelos.

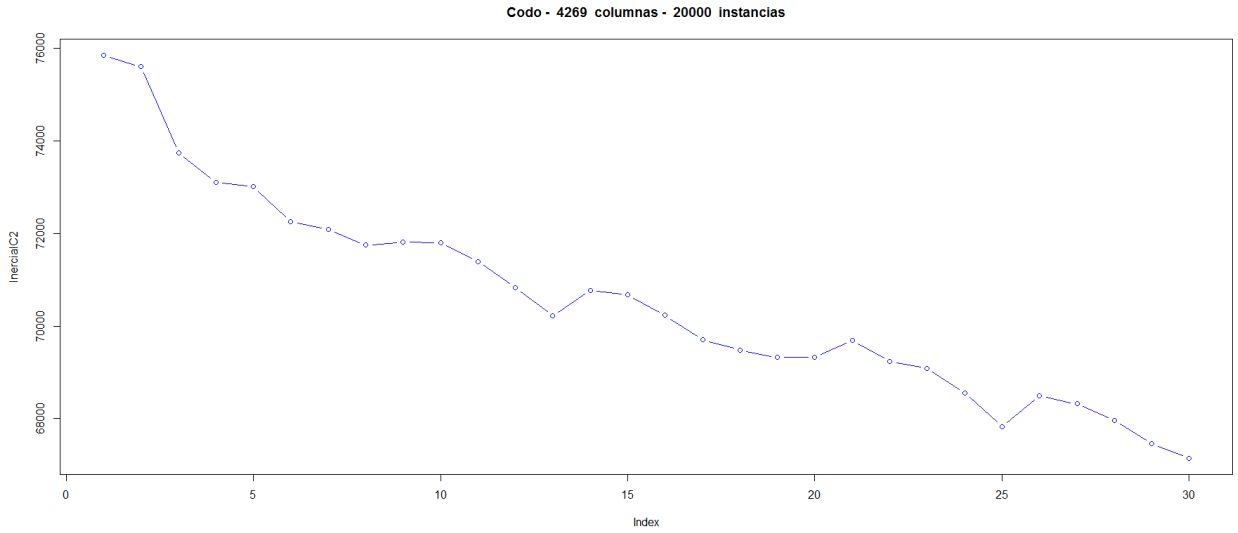


**Ilustración 24 - Método del Codo para repositorio, 4269 columnas.**

Todo esto implica que el investigador tiene que determinar el k (número de grupos) de manera "manual", probando diferentes k y observando las características de los grupos hasta lograr comunidades coherentes, tomando en cuenta en nivel de granularidad deseado.

El método del codo, el cual se basa en kmedias, no arrojo resultados satisfactorios. Hacer el estudio utilizando el algoritmo de kmedias tampoco arrojo resultados satisfactorios. La razón de eso, es que si bien todos los grupos "tendrían sentido", en cada corrida del algoritmo se obtendrían grupos distintos, y esto ocurre porque el kmedias crea centros aleatorios, y los va reacomodando para ser un mejor promedio de grupo. El lugar donde nace el centro de un grupo es un factor determinante a la hora de indicar que grupos existen.

Lo mas que se puede lograr con el método del codo en estas situaciones es intentar acotar el problema, para poder determinar, por lo menos, cuantos grupos no pueden haber, o dicho de otra manera, encontrar un k tal que a partir de él, se evidente que es excesivo. En la Ilustración 25, se puede observar el codo de un experimento donde se aumento a 500 la complejidad del problema y se estudio con hasta un k de 350 grupos.

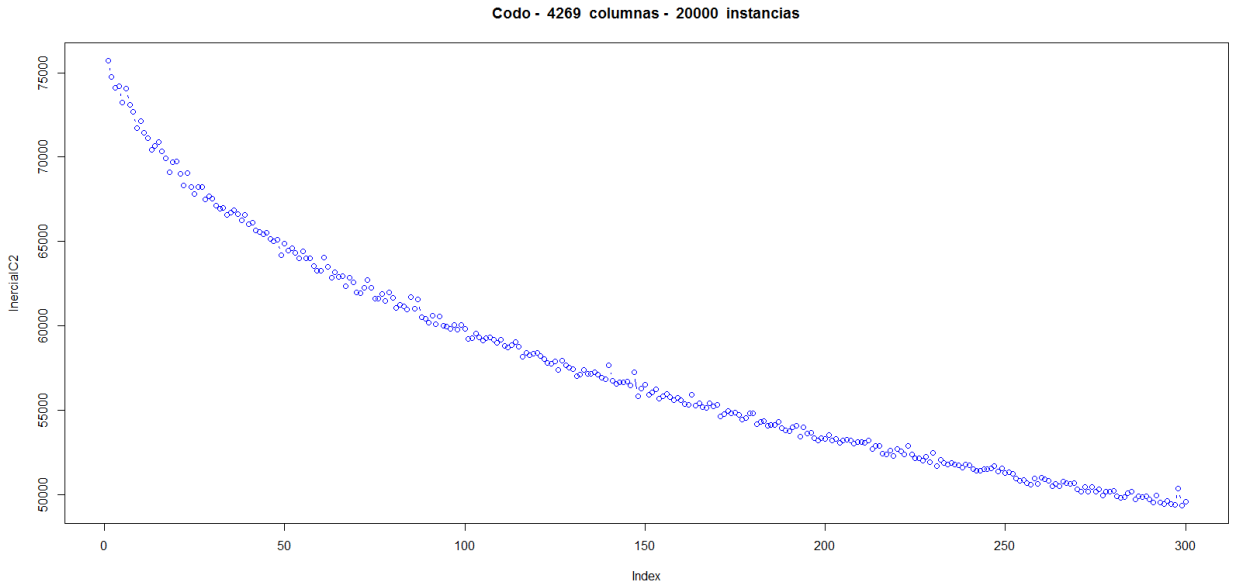


**Ilustración 25 - Método del Codo para repositorio, 500 columnas, k<350.**

La grafica, levemente sugiere un k cercano a 100. Si bien eso es un progreso, el grado de incertidumbre no es aceptable, y se vuelve necesario pasar a algoritmos más costosos en términos computacionales, pero más certeros, como el agrupamiento jerárquico.

Para la agrupación jerárquica se han probado todos los métodos que ofrece R, estos son: centroids (Ilustración 26 - Dendograma del método centroid.Ilustración 26), complete (Ilustración 27), mcquitty (Ilustración 28), median (Ilustración 29), avarage (Ilustración 30), single (Ilustración 31), Ward.d2 (Ilustración 32) y ward.d (Ilustración 33). En cada dendograma (grafica del árbol) se pueden ver unos rectángulos morados. Estos rectángulos son los puntos de corte entre diferentes grupos, para k = 110 grupos. Estos 110 grupos son los 110 grupos definitivos de las comunidades. Esta división de los individuos por grupos siguiendo la distribución del dendograma, se hizo para tener una idea superflua de la distribución de instancias por grupo.

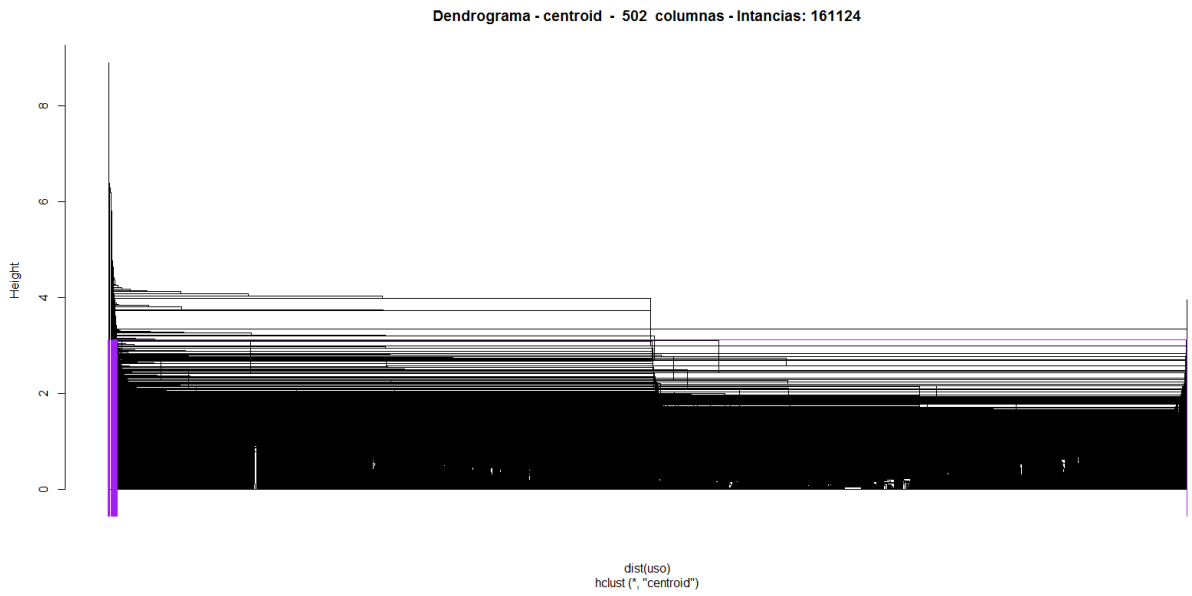


**Ilustración 26 - Dendograma del método centroid.**

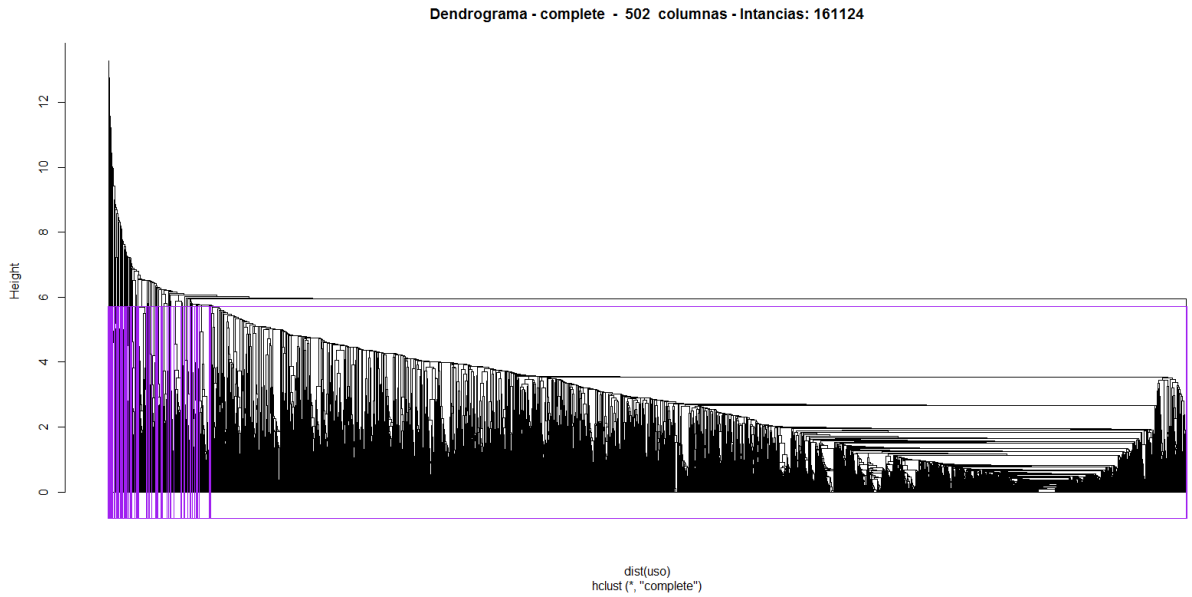


**Ilustración 27 - Dendograma del método complete.**

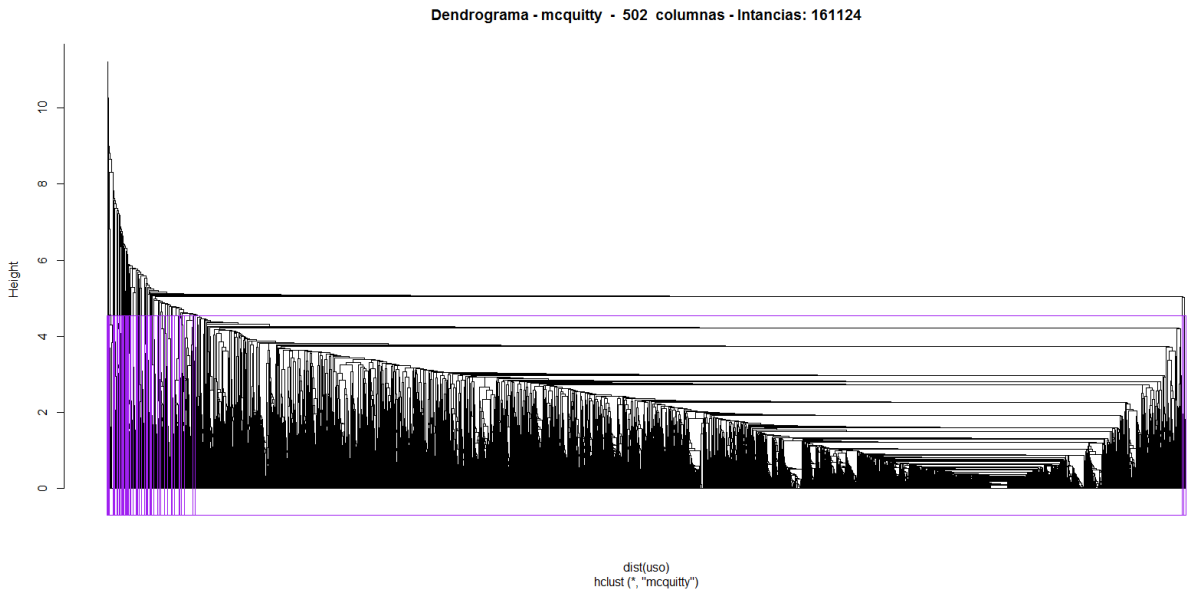


**Ilustración 28 - Dendograma del método mcquitty.**

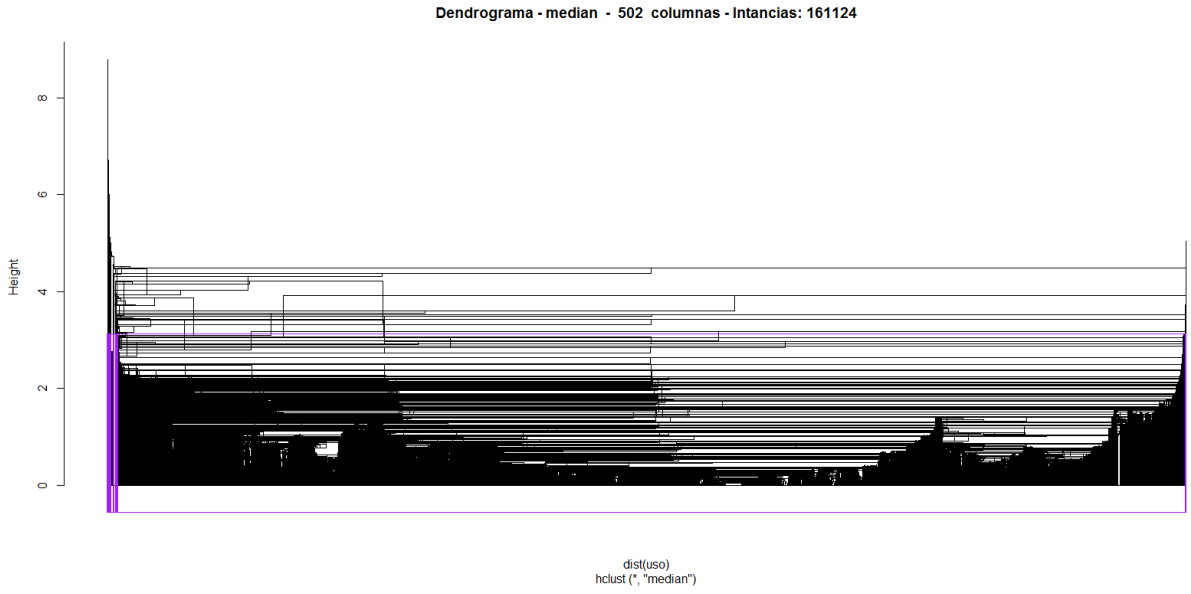


**Ilustración 29 - Dendograma del método median.**

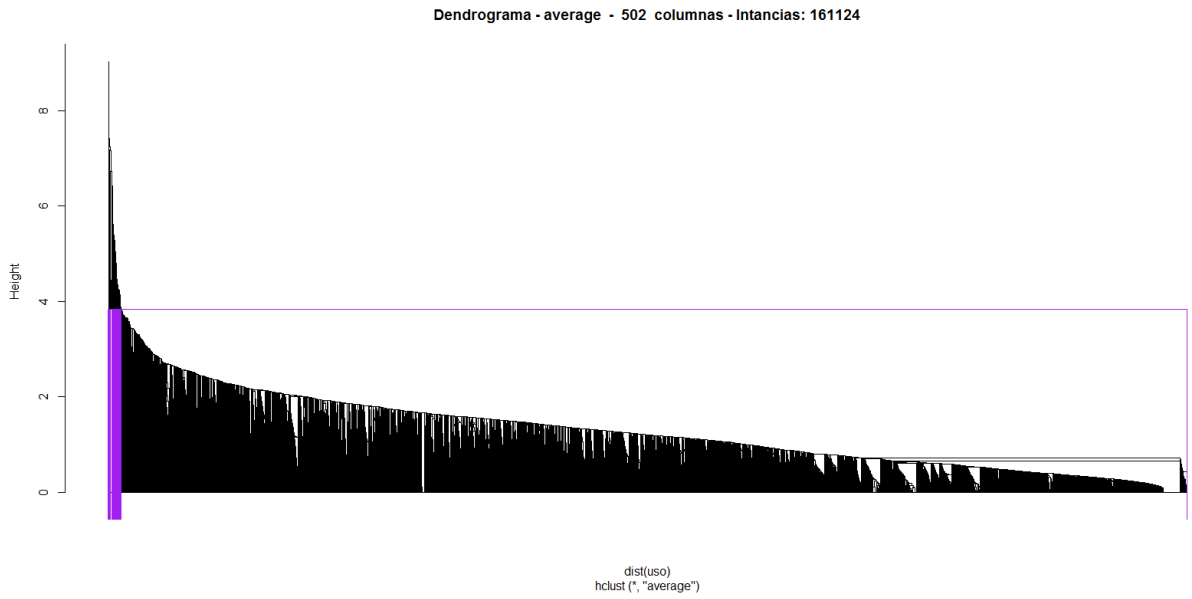


Ilustración 30 - Dendograma del método average.

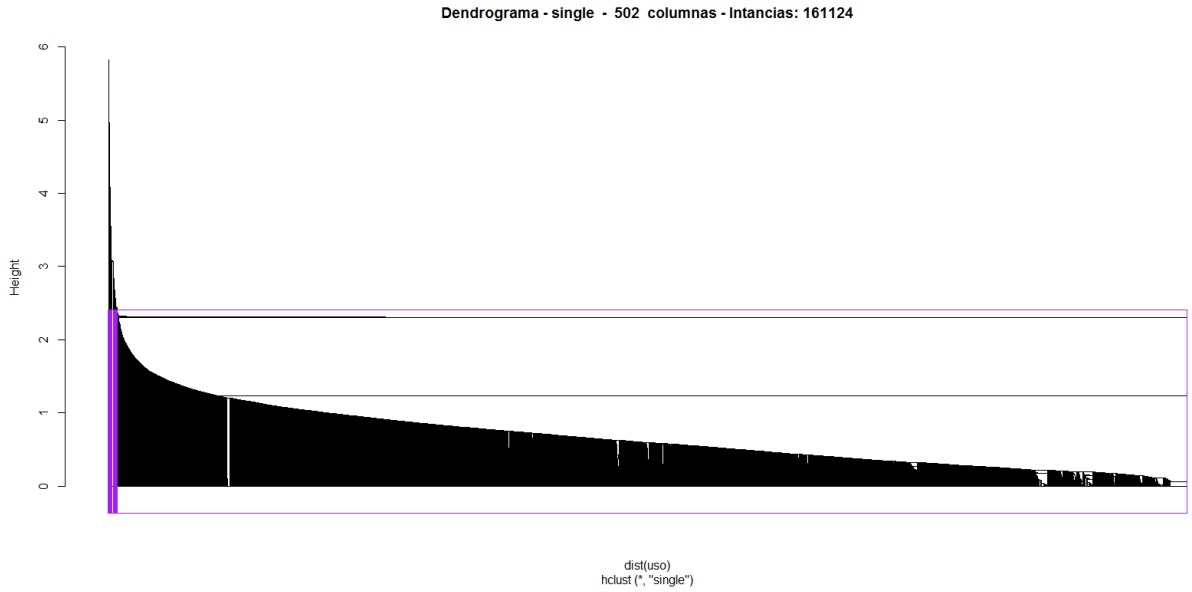


Ilustración 31 - Dendograma del método single.

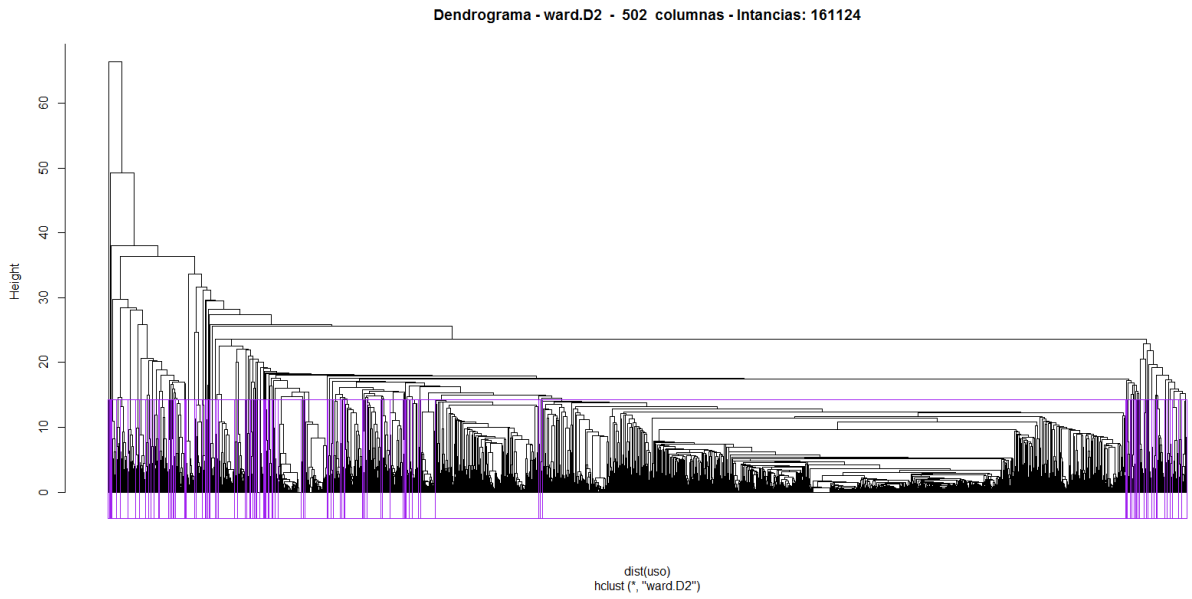


**Ilustración 32 - Dendograma del método Ward.D2.**

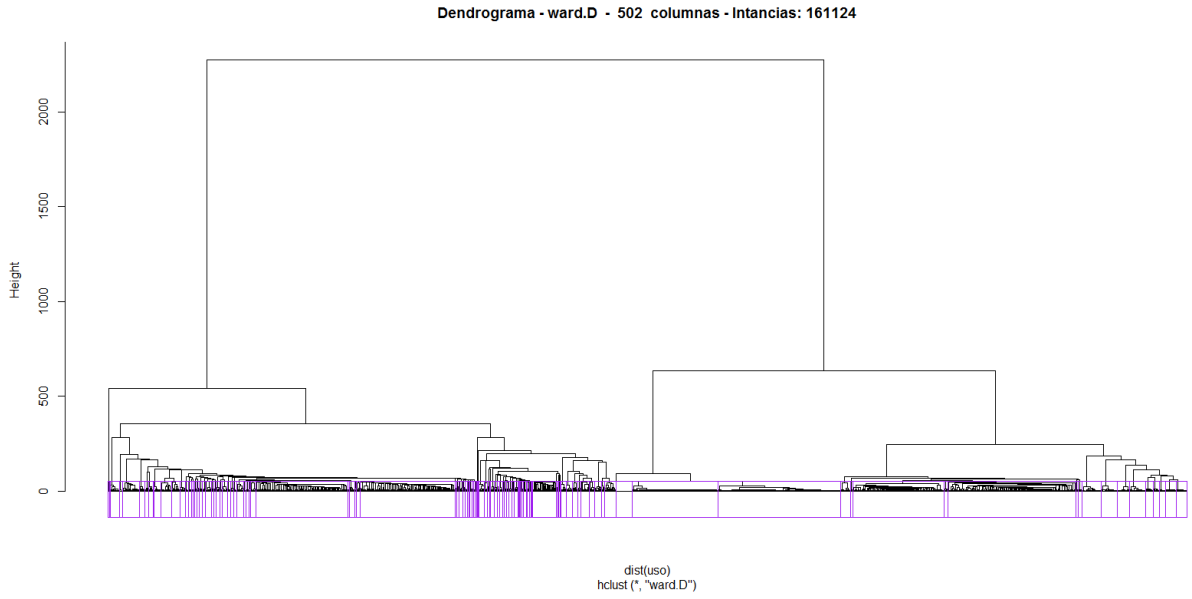


**Ilustración 33 - Dendograma del método Ward.D**

Aunque no lo parezca, todas las graficas muestran en rectángulos morados la distribución para 110 grupos. Como se puede apreciar, en todos los métodos a excepción de los Ward, hay un grupo enormemente dominante y los otros 109 de tamaño mínimos.

Es sumamente improbable que exista un grupo tan dominante en la muestra, o que las distancias (recordemos que agrupación jerárquica agrupa por distancias mínimas) permitan hacer agrupaciones

tipo escalera, donde en cada peldaño se separa un individuo de la muestra grande. De ser eso real, significaría que cada instancia es su propio grupo, cosa que no es válida. En los casos donde "la escalera" no es tan radical, la distribución de instancias por grupo, hace muchos grupos muy pequeños, lo que sugiere que hay demasiados datos por individuo.

Por la distribución de instancias de metadata entre los grupos de manera más equitativa (que los otros métodos), pareciera que Ward.D es el mejor método para agrupar, de hecho al aumentar el número de grupos (K), Ward.D sigue dando mejores resultados, inclusive superiores que su método hermano Ward.D2.

Sin embargo, una distribución más o menos homogénea no es sinónimo que la distribución es correcta, y si no es adecuada el modelo no es correcto. Para determinar si el modelo y las distribuciones son correctos, hay que analizar los grupos que se pueden formar, y a criterio del investigador determinar si son validos. Sera en este punto donde se determine el número de grupos, basándose en la granularidad deseada, y donde se determinen las características de cada grupo, así como el tamaño en individuos de los mismos.

Este estudio se condujo básicamente aumentando y disminuyendo el número de grupos y estudiando las palabras que los caracterizan hasta dar con una distribución que tiene relación con el mundo real, es explicable y por lo tanto ha sido probada como correcta.

Para determinar que palabras caracterizan a cada grupo, se suman las ponderaciones tf-idf de cada término por grupo, y los valores más elevados resultaban ser los términos representativos del grupo. Este proceso se repitió para cada K, por lo que fue un proceso muy extenso. El resultado final fue el siguiente: se puede agrupar la muestra en 110 grupos principales. Estos 110 grupos siguen una distribución de instancias como se ve en la Ilustración 34, indicandose en el eje Y el número de miembros por grupo.

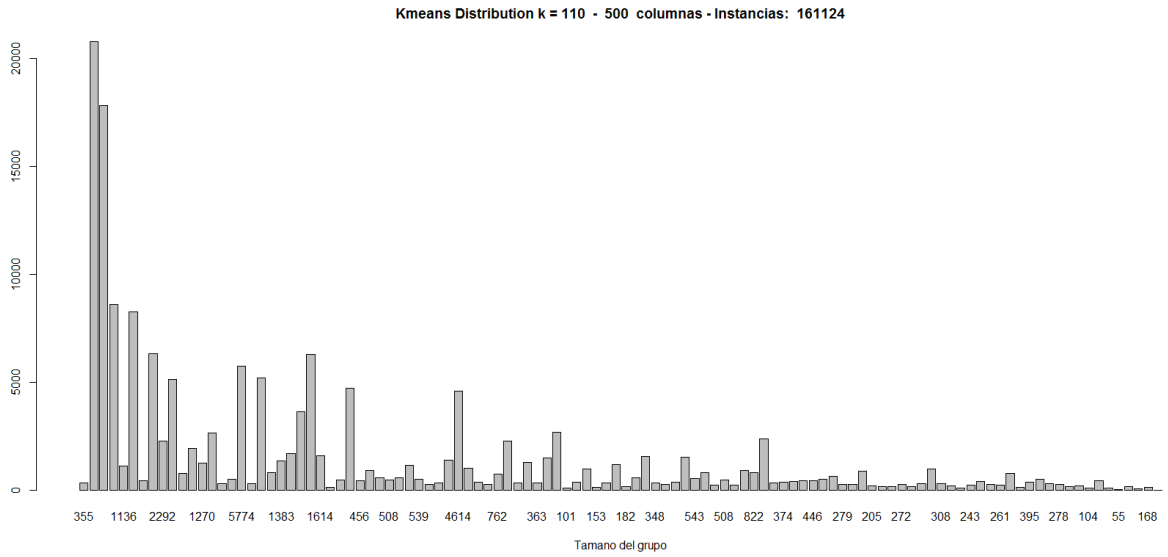


**Ilustración 34 - Instancias por grupo en agrupamiento en comunidades.**

En todos de los casos, los grupos son bastante lógicos, y se puede decir que la investigación en ese sentido dio resultado, puesto que el grupo concuerda con áreas de estudio universitario.

A continuación se muestra una lista con los grupos, el número de instancias por grupo, y las 2 palabras más importantes del grupo, teniendo entre paréntesis las sumas de ponderaciones tf-idf por palabra en ese grupo. Cabe destacar que las palabras no están completas, sino reducidas a su raíz, por el proceso de transformación de data del text mining.

Grupo 1 ---------- 355 instancias ---------- ascens ( 23 ) - memori ( 467 ) -

Grupo 2 ---------- 20786 instancias ---------- venezuel ( 276 ) - present ( 274 ) -

Grupo 3 ---------- 17833 instancias ---------- estudi ( 836 ) - investigacon ( 461 ) -

Grupo 4 ---------- 8629 instancias ---------- tesis ( 679 ) - venezuel ( 865 ) -

Grupo 5 ---------- 1136 instancias ---------- derech ( 413 ) - penal ( 1647 ) -

Grupo 6 ---------- 8296 instancias ---------- derech ( 115 ) - itali ( 114 ) -

Grupo 7 ---------- 452 instancias ---------- hombr ( 579 ) - psicologi ( 27 ) -

Grupo 8 ---------- 6344 instancias ---------- social ( 744 ) - trabaj ( 565 ) -

Grupo 9 ---------- 2292 instancias ---------- informacion ( 825 ) - investig ( 847 ) -

Grupo 10 ---------- 5162 instancias ---------- teori ( 468 ) - tesis ( 516 ) -

Grupo 11 ---------- 803 instancias ---------- introduccion ( 135 ) - sociologi ( 1472 ) -

Grupo 12 ---------- 1943 instancias ---------- comerci ( 292 ) - justici ( 389 ) -

Grupo 13 ---------- 1270 instancias ---------- pdf ( 1206 ) - ver ( 1125 ) -

Grupo 14 ---------- 2681 instancias ---------- letr ( 371 ) - tesis ( 375 ) -

Grupo 15 ---------- 312 instancias ---------- cin ( 402 ) - event ( 56 ) -

Grupo 16 ---------- 533 instancias ---------- aul ( 520 ) - entre ( 480 ) -

Grupo 17 ---------- 5774 instancias ---------- econom ( 409 ) - polit ( 435 ) -

Grupo 18 ---------- 320 instancias ---------- scienc ( 354 ) - social ( 48 ) -

Grupo 19 ---------- 5214 instancias ---------- educacion ( 1046 ) - zamor ( 756 ) -

Grupo 20 ---------- 844 instancias ---------- modern ( 581 ) - mund ( 313 ) -

Grupo 21 ---------- 1383 instancias ---------- politic ( 1218 ) - polit ( 155 ) -

Grupo 22 ---------- 1711 instancias ---------- dell ( 996 ) - itali ( 466 ) -

Grupo 23 ---------- 3670 instancias ---------- alons ( 532 ) - gamer ( 531 ) -

Grupo 24 ---------- 6303 instancias ---------- bacc ( 1132 ) - filosofi ( 1169 ) -

Grupo 25 ---------- 1614 instancias ---------- codig ( 473 ) - civil ( 1908 ) -

Grupo 26 ---------- 138 instancias ---------- biologi ( 297 ) - fundament ( 18 ) -

Grupo 27 ---------- 490 instancias ---------- segur ( 702 ) - social ( 139 ) -

Grupo 28 ---------- 4754 instancias ---------- belarmin ( 776 ) - psicologi ( 1124 ) -

Grupo 29 ---------- 456 instancias ---------- ingles ( 36 ) - theory ( 460 ) -

Grupo 30 ---------- 921 instancias ---------- conduct ( 730 ) - modificacion ( 299 ) -

Grupo 31 ---------- 604 instancias ---------- psicologi ( 218 ) - psychology ( 893 ) -

Grupo 32 ---------- 508 instancias ---------- contrat ( 910 ) - mercantil ( 93 ) -

Grupo 33 ---------- 586 instancias ---------- lengu ( 139 ) - lenguaj ( 662 ) -

Grupo 34 ---------- 1178 instancias ---------- ley ( 931 ) - organic ( 205 ) -

Grupo 35 ---------- 539 instancias ---------- administr ( 819 ) - derech ( 269 ) -

Grupo 36 ---------- 274 instancias ---------- criminal ( 229 ) - delincuent ( 289 ) -

Grupo 37 ---------- 339 instancias ---------- fil ( 313 ) - ideologi ( 321 ) -

Grupo 38 ---------- 1402 instancias ---------- amaric ( 978 ) - latin ( 995 ) -

Grupo 39 ---------- 4614 instancias ---------- filologi ( 1071 ) - rosenblat ( 1059 ) -

Grupo 40 ---------- 1025 instancias ---------- derech ( 164 ) - introduccion ( 1174 ) -

Grupo 41 ---------- 391 instancias ---------- aplic ( 44 ) - estadist ( 569 ) -

Grupo 42 ---------- 287 instancias ---------- guerr ( 446 ) - mundial ( 34 ) -

Grupo 43 ---------- 762 instancias ---------- diccionari ( 1150 ) - espanol ( 108 ) -

Grupo 44 ---------- 2280 instancias ---------- event ( 1709 ) - monograf ( 619 ) -

Grupo 45 ---------- 348 instancias ---------- histori ( 37 ) - revolucion ( 459 ) -

Grupo 46 ---------- 1293 instancias ---------- tropical ( 325 ) - zoologi ( 288 ) -

Grupo 47 ---------- 363 instancias ---------- psicoanalisis ( 830 ) - psicologa ( 39 ) -

Grupo 48 ---------- 1528 instancias ---------- psicologi ( 75 ) - the ( 855 ) -

Grupo 49 ---------- 2711 instancias ---------- comunicacion ( 1095 ) - gustav ( 442 ) -

Grupo 50 ---------- 101 instancias ---------- criminologi ( 389 ) - introduccion ( 10 ) -

Grupo 51 ---------- 375 instancias ---------- educacion ( 65 ) - education ( 328 ) -

Grupo 52 ---------- 996 instancias ---------- conferent ( 336 ) - ensay ( 499 ) -

Grupo 53 ---------- 153 instancias ---------- consej ( 189 ) - gacet ( 426 ) -

Grupo 54 ---------- 353 instancias ---------- colect ( 61 ) - psicoterapi ( 589 ) -

Grupo 55 ---------- 1201 instancias ---------- derech ( 352 ) - diritt ( 1094 ) -

Grupo 56 ---------- 182 instancias ---------- editorial ( 513 ) - med ( 66 ) -

Grupo 57 ---------- 592 instancias ---------- derech ( 417 ) - roman ( 1106 ) -

Grupo 58 ---------- 1581 instancias ---------- histori ( 1127 ) - venezuel ( 170 ) -

Grupo 59 ---------- 348 instancias ---------- delit ( 607 ) - penal ( 62 ) -

Grupo 60 ---------- 282 instancias ---------- boliv ( 358 ) - simon ( 174 ) -

Grupo 61 ---------- 399 instancias ---------- mercantil ( 198 ) - sociedad ( 642 ) -

Grupo 62 ---------- 1558 instancias ---------- psicologi ( 1590 ) - social ( 260 ) -

Grupo 63 ---------- 543 instancias ---------- industrial ( 421 ) - propied ( 455 ) -

Grupo 64 ---------- 817 instancias ---------- des ( 627 ) - franci ( 48 ) -

Grupo 65 ---------- 268 instancias ---------- public ( 510 ) - tecnologi ( 30 ) -

Grupo 66 ---------- 508 instancias ---------- muj ( 357 ) - sexual ( 341 ) -

Grupo 67 ---------- 253 instancias ---------- inglas ( 41 ) - physics ( 335 ) -

Grupo 68 ---------- 931 instancias ---------- derech ( 532 ) - internacional ( 1036 ) -

Grupo 69 ---------- 822 instancias ---------- constitucional ( 904 ) - derech ( 409 ) -

Grupo 70 ---------- 2384 instancias ---------- derech ( 1988 ) - mercantil ( 434 ) -

Grupo 71 ---------- 372 instancias ---------- mental ( 496 ) - salud ( 101 ) -

Grupo 72 ---------- 374 instancias ---------- adolescent ( 509 ) - psicologi ( 79 ) -

Grupo 73 ---------- 417 instancias ---------- amiric ( 34 ) - democraci ( 683 ) -

Grupo 74 ---------- 450 instancias ---------- histori ( 42 ) - pensamient ( 540 ) -

Grupo 75 ---------- 446 instancias ---------- personal ( 631 ) - teori ( 67 ) -

Grupo 76 ---------- 540 instancias ---------- colombi ( 609 ) - politic ( 60 ) -

Grupo 77 ---------- 661 instancias ---------- economi ( 841 ) - economi ( 145 ) -

Grupo 78 ---------- 279 instancias ---------- introduccion ( 78 ) - logic ( 530 ) -

Grupo 79 ---------- 286 instancias ---------- coleccion ( 787 ) - polit ( 27 ) -

Grupo 80 ---------- 896 instancias ---------- derech ( 255 ) - droit ( 964 ) -

Grupo 81 ---------- 205 instancias ---------- sociologi ( 80 ) - sociologi ( 500 ) -

Grupo 82 ---------- 173 instancias ---------- antropologi ( 355 ) - hombr ( 16 ) -

Grupo 83 ---------- 188 instancias ---------- psicometri ( 288 ) - tests ( 322 ) -

Grupo 84 ---------- 272 instancias ---------- complet ( 221 ) - obras ( 392 ) -

Grupo 85 ---------- 175 instancias ---------- coleccion ( 80 ) - jurid ( 472 ) -

Grupo 86 ---------- 310 instancias ---------- marx ( 237 ) - marxism ( 551 ) -

Grupo 87 ---------- 1008 instancias ---------- nin ( 927 ) - psicologi ( 234 ) -

Grupo 88 ---------- 308 instancias ---------- civil ( 198 ) - procesal ( 435 ) -

Grupo 89 ---------- 222 instancias ---------- editorial ( 1536 ) - med ( 1 ) -

Grupo 90 ---------- 133 instancias ---------- sintesis ( 499 ) - ucla ( 505 ) -

Grupo 91 ---------- 243 instancias ---------- infantil ( 45 ) - psiquiatri ( 649 ) -

Grupo 92 ---------- 409 instancias ---------- doctrin ( 78 ) - jurisprudent ( 618 ) -

Grupo 93 ---------- 300 instancias ---------- abog ( 304 ) - seri ( 215 ) -

Grupo 94 ---------- 261 instancias ---------- chemistry ( 367 ) - inglas ( 44 ) -

Grupo 95 ---------- 790 instancias ---------- derech ( 67 ) - manual ( 785 ) -

Grupo 96 ---------- 146 instancias ---------- pais ( 31 ) - subdesarroll ( 273 ) -

Grupo 97 ---------- 395 instancias ---------- comunicacion ( 58 ) - period ( 402 ) -

Grupo 98 ---------- 519 instancias ---------- inglas ( 52 ) - introduction ( 599 ) -

Grupo 99 ---------- 311 instancias ---------- educacion ( 57 ) - pedagogi ( 417 ) -

Grupo 100 ---------- 278 instancias ---------- mar ( 396 ) - ucvfacesbdfmz ( 38 ) -

Grupo 101 ---------- 197 instancias ---------- drog ( 372 ) - las ( 13 ) -

Grupo 102 ---------- 235 instancias ---------- andris ( 206 ) - bell ( 308 ) -

Grupo 103 ---------- 104 instancias ---------- mund ( 5 ) - nbsp ( 281 ) -

Grupo 104 ---------- 459 instancias ---------- derech ( 65 ) - famili ( 692 ) -

Grupo 105 ---------- 127 instancias ---------- presentacion ( 902 ) -

Grupo 106 ---------- 55 instancias ---------- port ( 336 ) - vol ( 65 ) -

Grupo 107 ---------- 195 instancias ---------- edu ( 348 ) - ucvfacesbdfmz ( 39 ) -

Grupo 108 ---------- 93 instancias ---------- autor ( 9 ) - indic ( 325 ) -

Grupo 109 ---------- 168 instancias ---------- derech ( 105 ) - tributari ( 329 ) -

Grupo 110 ---------- 26 instancias ---------- cot ( 235 ) - ucvfacesbdfmz ( 2 ) -

Como se puede apreciar en la lista, las palabras están incompletas, esto se debe a que fueron reducidas a su raíz durante el proceso de creación de la tdm.

Dado que el agrupamiento se hace por instancias, cada instancia pertenece o no a un grupo dependiendo de los valores de cada una de sus columnas. Partiendo de este principio es correcto afirmar que las columnas que den mayor suma, son las palabras clave o representativas de ese grupo.

Entonces si se sabe que el modelo es correcto, es necesario crear las estructuras que permitan usarlo. Hay muchas formas de hacer esto, por ejemplo: reglas o arboles de decisión. Sin embargo, en este trabajo se opto por indicar los centros de cada grupo, de forma tal que cada instancia pueda ser clasificada a un grupo en función de su distancia con respecto a los centros de grupo.

Esto fue posible gracias al paquete rattle de R, creado por Graham Williams, Mark Vere Culp, Ed Cox, Anthony Nolan, Denis White, Daniele Medri, Akbar Waljee, Brian Ripley. Este paquete ofrece una función "centers.hclust" que permite generar una matriz de centros partiendo de un agrupamiento jerárquico. Esta matriz es útil para calcular las distancias de cada individuo a un centro y agruparlo en dicho grupo, lo cual es lo que hace el algoritmo de kmedias.

Teniendo esos centros es posible agrupar a toda la muestra y a cualquier nueva instancia, utilizando una función que indique con cual centro hay menor distancia, o utilizando el algoritmo de kmedias sin movimiento de centros.

Se debe mencionar respecto a la clasificación por distancias a partir de centros creados con agrupamiento jerárquico, que la misma no será idéntica a la de agrupación jerárquica por sí sola. Esta diferencia suele medirse y ejemplificarse con matrices de confusión, como el ejemplo de la tabla 7, la cual tiene las instancias bien clasificadas en la diagonal y las mal clasificas en los demás puntos. En relación a la misma hay 110 grupos, 110 columnas y 110 filas, lo cual es mucha información para retener. Sin embargo, es importante recordar que la clasificación jerárquica dará un resultado ligeramente diferente al de clasificación por distancias, y el tamaño del error vendrá dado por la calidad del modelo.

**Tabla 7- Matriz de confusión entre hclust y clasificación por distancias**

| Grupos | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | **6000** | 1550 | 273 | 0 | 19 |
| 2 | 299 | **5000** | 57 | 0 | 0 |
| 3 | 83 | 68 | **1000** | 0 | 1 |
| 4 | 14 | 21 | 0 | **500** | 0 |
| 5 | 4 | 110 | 0 | 0 | **1000** |

La clasificación por distancias no es idéntica al agrupamiento jerárquico, esto se debe a que algunas instancias pueden ser clasificadas en más de un grupo. Eso es posible si se considera que hay trabajos inter-áreas, por ejemplo una aplicación computarizada para trabajos en el área de la biología, o en el área de los negocios, o un proyecto de ingeniera evaluado desde su enfoque social. El alcance de este proyecto indica una comunidad por instancia, es importante tener eso en consideración.

El grueso de las instancias, están clasificadas acorde a la diagonal, y el error de clasificación por grupo es pequeño para todos los casos. Prueba de esto se puede ver con el BoxPlot de la Ilustración 355.

Es necesario, para todo modelo, indicar un grado de error. Todo modelo comete errores, y es necesario poder cuantificar ese error, para determinar una confianza en dicho modelo. En este proyecto el error se midió utilizando la herramienta diagrama de caja, también llamada “boxplot”.

Boxplot mide las distancias entre las instancias, e indica los valores outliers. En la Ilustración 35 se puede ver el boxplot del modelo, el cual indica tener 502 columnas (500 de data, 1 que corresponde a que comunidad fue agrupado y una que señala la distancia de la instancia hacia el centro del grupo).

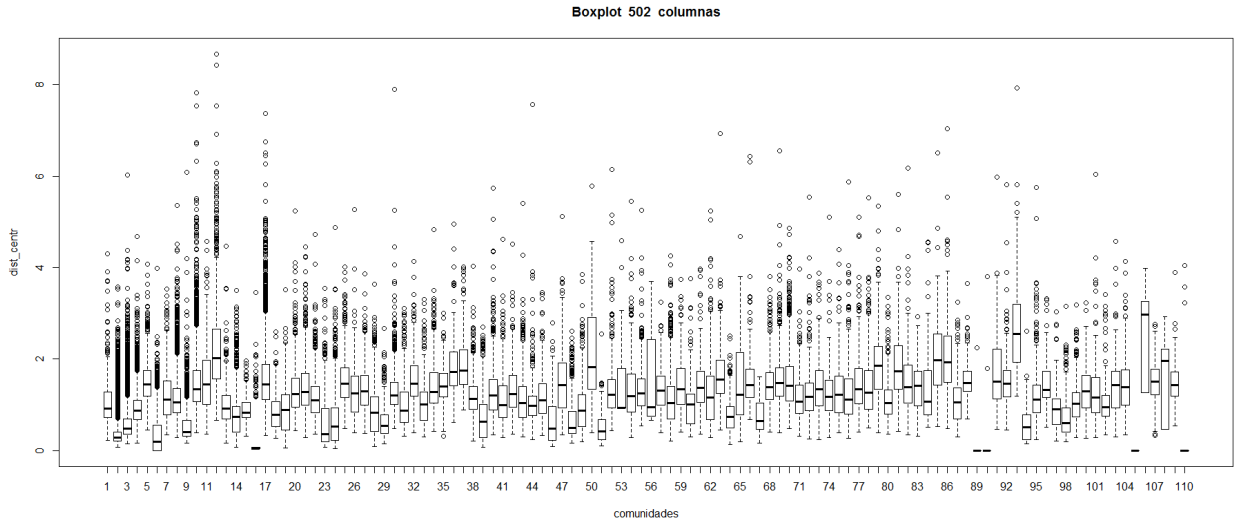


**Ilustración 35 – Boxplot**

Leyenda:

- La grafica mide distancias
- La línea negra en el medio de cada rectángulo representa la mediana del grupo.
- El recuadro de cada grafica representa la "zona" donde radica el 50% de las instancias del grupo.
- Las líneas punteadas verticales que terminan en una línea continua horizontal, indican la "zona" donde radica el otro 45% de las instancias.
- La línea continua horizontal (donde termina la vertical) se llama bigote, y representan el mínimo y máximo valor que aun se considera bien agrupado en la comunidad.
- El 5% restante es graficado a manera de puntos, y son los puntos más alejados del centro, por los que se les ve como outliers.

Se considera que el modelo esta correcto cuando el rectángulo es lo más compacto posible (las instancias están lo más cerca del centro posible, la mediana (la línea horizontal dentro del cuadro) está cerca de la mitad del rectángulo (demostrando que la distribución es simétrica y que el centro está bien posicionado para ese grupo colocado), que exista el menor número posible de outliers, y que si estos existen estén posicionados lo más cerca posible del bigote (si la mayoría de outliers están muy alejados del bigote eso significaría que ese grupo en realidad contiene 2 grupos distintos o que el centro está muy alejado del nucleo del grupo).

Las líneas negras, que denotan muchos outliers juntos, en este caso están ubicadas exactamente donde los grupos son de mayor tamaño; a mayor población aumenta el número de outliers. En total hay 6908 outliers de una muestra de 161124, lo cual representa el 4,28 % del total.

En las investigaciones sociológicas, como esta, un trabajo contribuye a la toma de decisiones si brinda una confianza de 80%, y se toma como verdaderas las afirmaciones del trabajo si rondan el 95% de confianza. Solo en el área de la medicina es común que para considerarse correcto un trabajo se necesite una confianza del 99% o superior. Esto quiere decir que todos los indicios apuntan a que el resultado presentado es correcto para la agrupación, y en consecuencia que no es un error indicar que existen 110 comunidades en la muestra.

### 4.3.5. Función de Recomendación

La función de recomendación recibe como entrada un string que corresponde al URL del recurso al cual se le desea la sugerencia, y devuelve como salida una lista de 3 posiciones, donde cada posición tiene el índices que corresponden a las ubicaciones dentro de la tabla de URLs de los recursos recomendados.

Lo que hace dicha función es buscar dentro de la tabla de URLs cuál es el índice del recurso solicitado. Este índice coincide con el índice del dtm, por lo que sabiendo la URL de un recurso se pueden saber los valores de la fila del dtm que corresponden a esa URL. A esta fila del dtm la llamaremos instancia.

Sabiendo la instancia, podemos calcular su distancia en relación con todas las demás instancias. Este cálculo, se puede hacer en un "todos contra todos", y tener así almacenado las distancias para todos los casos a fin de no tener que calcularlas en tiempo de ejecución. Sin embargo, para el caso de 161124 recursos con 500 filas + las filas de clasificación esta tabla sería de unos 90 GB, por lo que no es una opción viable. Otra opción será calcular las distancias solamente en los grupos más grandes, que debieran ser lo que más se demoren, sin embargo para el grupo de 20.000 recursos esta tabla pesa 3.2 GB lo que lo hace poco práctico.

No siendo posible tener almacenada las distancias a priori es necesario calcularlas en tiempo de ejecución. Para agilizar el cálculo, se reduciría el tamaño de la muestra a solamente instancias del mismo grupo. Puesto que se desean trabajos similares en naturaleza, no es conveniente intentar mezclar grupos de diferente naturaleza. No hacer esto brindaría como consecuencia que se tuviera que calcular la distancia contra todos los individuos, cosa que sería muy costosa en términos computacionales y por lo tanto muy largo el tiempo de espera para los resultados, brindando como único beneficio, que para las instancias outliers (que representan el 4,28%) se obtendrian resultados mejores, puesto que la instancia más cercana de un outlier puede estar "al otro lado de la frontera" del grupo.

Para agilizar aun más el cálculo, se utilizo computación paralela de forma que cada cálculo se ejecutara de manera simultánea e intentar reducir aún más la demora para un resultado.

Con todo esto se pudo finalmente probar una primera versión, llamada Frankenstein, del sistema de recomendación, la cual dio resultados satisfactorios. Efectivamente para la mayoría de casos, las recomendaciones si tenían relación con el recurso. Sin embargo, el sistema era de velocidad variable, dependiendo del recurso a consultar la respuesta podía darse entre 5 minutos y 20 segundos, tiempos estos inaceptables para una aplicación de producción. Esto obligo a replantear el modelo y sistema, para brindar rapidez a la consulta. Se recomienda leer el capitulo 4.3.6.

### 4.3.6. Mejoras al modelo

Dado que los tiempos de ejecución más cortos se lograban con los grupos más pequeños, se decidió convertir a todos los grupos, en unos  más pequeños. Se decidió aplicar agrupación jerárquica, por segunda ocasión, sobre los grupos de mayor tamaño, si el resultado seguía teniendo grupos grandes, se volvía a aplicar, y se volvía aplicar inclusive una cuarta vez. Con esto se creó un agrupamiento de cuatro niveles, donde el primer nivel corresponde al agrupamiento jerárquico original y los otros 3 a las mejoras incorporadas.

De esta manera cada instancia tenía 4 campos que correspondían a su agrupamiento, estos campos siguen un esquema jerárquico, interrelacionados entre sí para poder identificarlos. Por ejemplo, no se puede saber a qué grupo pertenece el subgrupo 5, sin establecer cuál es el grupo y luego indicar cuales el subgrupo.

Se crearon 102 subgrupos de clasificación jerárquica. Todos utilizando el método Ward con las distancias calculadas con la función euclidiana. De los grupos originales se subdividieron los grupos de:

- grupo 2 - 20786 instancias
- grupo 3 - 17833 instancias
- grupo 4 - 8629 instancias
- grupo 6 - 8296 instancias
- grupo 8 - 6344 instancias
- grupo 9 - 2292 instancias
- grupo 10 - 5162 instancias
- grupo 12 - 1943 instancias
- grupo  13 - 1270 instancias
- grupo 14 - 2681 instancias
- grupo 17 - 5774 instancias
- grupo 19 - 5214 instancias
- grupo 21 - 1383 instancias
- grupo 22 - 1711 instancias
- grupo 23 - 3670 instancias

- grupo 24 - 6303 instancias
- grupo 25 - 1614 instancias
- grupo 28 - 4754 instancias
- grupo 38 - 1402 instancias
- grupo 39 - 4614 instancias
- grupo 44 - 2280 instancias
- grupo 48 - 1528 instancias
- grupo 49 - 2711 instancias
- grupo 70 - 2384 instancias

Todos los grupos antes listados tienen 3 niveles a excepción del 2 y 3 que tiene 4 niveles. Los grupos originales que no aparecen en la lista solo tienen un nivel. Esta división se hizo con el propósito de que ningún grupo ni subgrupo tuviera 1500, más o un tanto menos de esa cifra de instancias. La razón para que los grupos tuviesen como cota esa cifra es que, en el proceso de optimizar la función de recomendación, se hizo evidente que algunas funciones trabajan a mucha mayor velocidad por debajo de ciertas cotas.

### 4.3.7. Resultados

Mientras se hacían las mejoras al modelo, se probaba el sistema, y se hicieron modificaciones al sistema buscando mayor velocidad de respuesta. Se aplicaron conceptos de arboles de optimización de consultas en bases de datos, también se estudiaron estructuras de datos diversas que planteaban velocidad, como el paquete data.table, que es una estructura muy similar en uso al data.frame pero que incorpora índices para la ejecución a mayor velocidad de las consultas. Inclusive se investigo como escribir funciones de forma tal que fueran procesadas por el interprete de R lo mejor posible, cual fue el caso del subset y el [ .

Después de mucho ensayo y error, se perfilo una función de ejecución en un solo núcleo que, bajo ciertas condiciones en la data, brinda resultados a mayor velocidad. La función dist que calcula la distancia de cada instancia contra todas las demás de un data.frame originalmente fue descartada, por consumir demasiados recursos y no ser rápida. Sin embargo, luego se descubrió que la misma podía ser muy rápida para matrices menores a 1500 observaciones (instancias). Esta combinación de la data correctamente preparada, con las instrucciones de la función de recomendación brindaba un mejor tiempo de respuesta, como se puede apreciar en la Tabla 8.

Tabla 8 - Tiempos de ejecución de versiones de la función de recomendación

| | test | replications | elapsed | relative | user.self | sys.self | user.child | sys.child |
|---|---|---|---|---|---|---|---|---|
| 1 | Versión Original | 3 | 957.86 | 30.809 | 45.85 | 5.31 | NA | NA |
| 2 | Segunda Versión | 3 | 89.37 | 2.875 | 10.77 | 0.64 | NA | NA |
| **3** | **Versión Final** | **3** | **31.09** | **1.000** | **30.86** | **0.19** | **NA** | **NA** |

Leyenda:

- Test indica el nombre de la prueba.
- Replication es el número de veces que se repitió esa prueba ( 3 ocasiones en todos los casos)
- Elapsed es el tiempo en segundos que se percibe en el mundo natural que duro la ejecución del conjunto de instrucciones
- Relative es una comparación basada en regla de 3 con respecto al mejor valor obtenido en la prueba
- User.self el tiempo que el cpu gasto en la instrucción
- Sys.self el tiempo que el sistema tardo en adecuarse para ejecutar la instrucción (cargar los datos en los registros, etc)
- Sys.child no están disponibles en maquinas windows

Como se puede ver en la tabla 8, entre la primera y la última versión hay una diferencia de hasta 30 veces en la duración, esto se tradujo que se paso de un peor caso de 5 minutos a uno de 10 segundos, de un caso promedio de unos 3 minutos (la mitad de los casos/instancias, pertenecen a un grupo grande que dura 5 minutos y el otro a un grupo pequeño de unos 40 segundos) a un promedio de 2 segundos de duración por consulta.

Con una función que finalmente responde a velocidad aceptable se creó una primera página web utilizando el paquete shiny que permite la creación de páginas web sin necesidad de saber html, css,

js, ni tener algún conocimiento en un servidor especifico. La interfaz de dicha página se puede ver en la Ilustración 366.

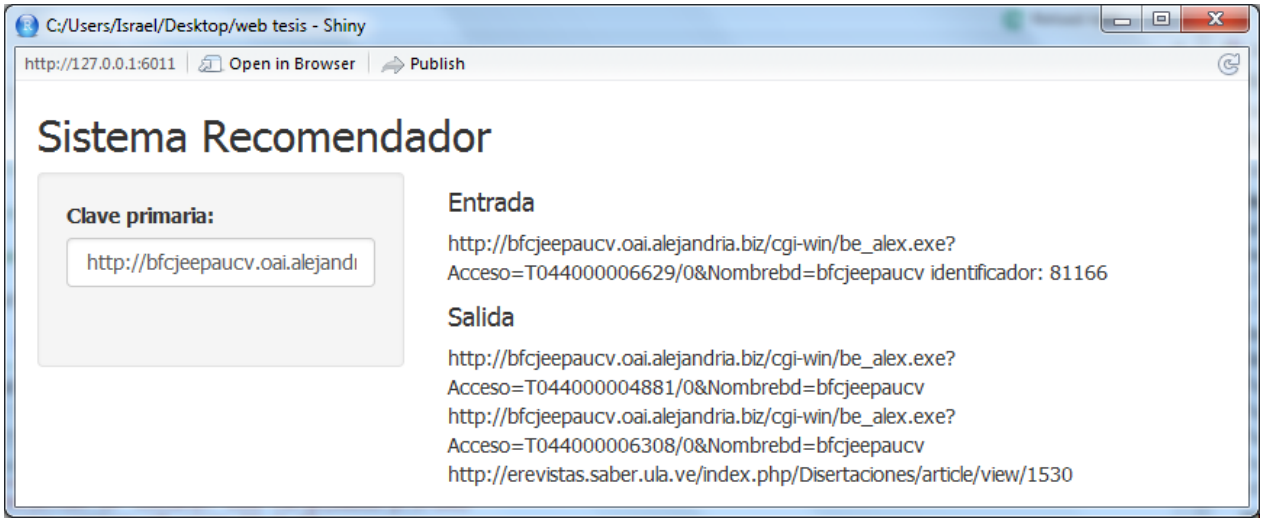


**Ilustración 36 – Primera Interfaz con Shiny**

La interfaz que se puede apreciar en la Ilustración 366 es muy sencilla. Hay un espacio para escribir texto del lado izquierdo, donde se coloca una dirección URL que corresponde con el identifier del recurso, del lado derecho. Debajo de entrada se indica que recurso se solicito y cuál es su número dentro del índice de URLs y debajo de Salida, se indican los recursos sugeridos.

Para probar el sistema de recomendación se amplió la página web anterior para crear un prototipo del sistema de recomendación, otra vez con el paquete Shiny de R, esta nueva interfaz buscar simular como se verá el funcionamiento del sistema de recomendación a instalar en el servidor del Buscador Académico Venezolano.

La interfaz básicamente recibe una URL a buscar y un numero de recomendaciones deseadas con lo que despliega información básica del recurso identificado por esa URL. Recordemos que para los recursos del Buscador Académico Venezolano el estándar es dublin core el cual cuenta con el campo identifier que sirve de clave primaria del recurso y en el caso especifico del Buscador Académico Venezolano es su URL. Cargada esta información básica, se despliega tantas recomendaciones como se haya indicado sobre la URL especificada. Esta interfaz se puede apreciar en la Ilustración 367.

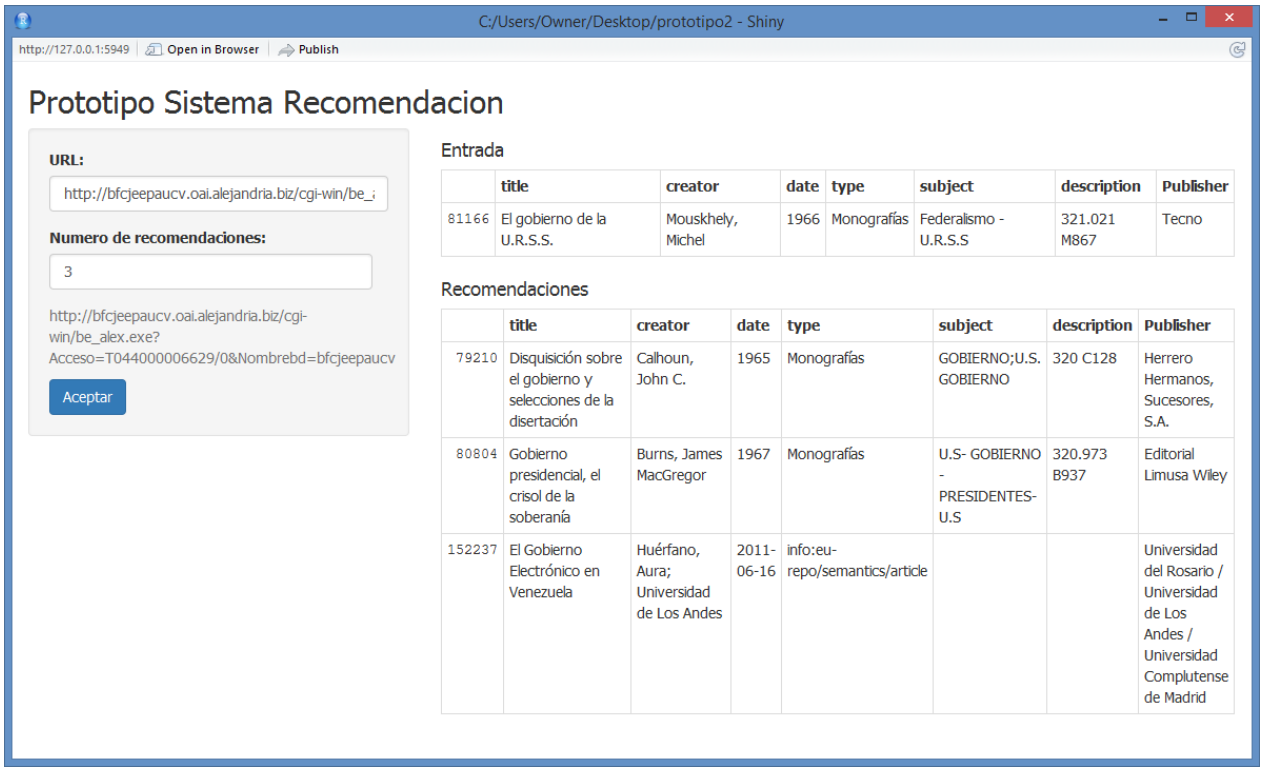


**Ilustración 37 - Prototipo Sistema de Recomendacion**

El paquete Shiny provee un servidor limitado, el cual solo emite solicitudes y respuestas sobre los widget de su interfaz. El portal dispone de un widget tipo campo de entrada de texto para recibir la URL, y 2 widgets, de los cuales uno es para mostrar la información básica del recurso solicitado (el indicado en la URL) y otro para la información de los recursos recomendados.

Por detrás, en el lado del servidor, el sistema lo que hace es cargar un conjunto mínimo y necesario de datos para su funcionamiento, que contiene:

- Una lista que sirve como índice de recursos que contiene todas las URL
- Un dataframe que contiene las ponderaciones tf-idf de los recursos indexados (dtm)
- Un dataframe que contiene los valores encontrados por cada campo del dublin core para cada recurso indexado
- Una función de recomendación, la cual dada una URL y un numero de sugerencias deseadas, retornan tantas recomendaciones como el numero de sugerencias indicadas. En total 4 elementos de datos.

La función de recomendación recibe 2 parámetros, un número que corresponde al número de recomendaciones deseadas y una cadena de caracteres que corresponde a la URL del documento del cual se desean sugerencias. La función de recomendación devuelve un objeto tipo JSON que contiene

en un arreglo con tantas sugerencias como el número de recomendaciones solicitadas. Esta función vivirá como un proceso cgi dentro del servidor de producción del Buscador Académico Venezolano.

## 4.4. PRUEBAS

Para medir el tiempo de respuesta del sistema, se utilizo la siguiente lógica "si un caso es el peor, un caso más leve, no puede durar más que el peor". El peor caso para la muestra presentada, era un grupo de 20.000 instancias que duraba 5 minutos y medio en conseguir una recomendación. Casos como ese y otros fueron atacados durante el proceso de mejoras al modelo, para reducir el tiempo de respuesta. El peor caso encontrado en la versión final del sistema es de 1200 instancias el cual dura 10 segundos. Un caso con más instancias resultaría en peores resultados, y un caso con menos instancias resultaría en mejores resultados. Sin embargo se indica que no es una regla de 3 exacta, pues para un grupo de 600 instancias el tiempo de ejecución es de 2 segundos, es decir un 150% menos de lo que cabria esperarse por regla de 3.

Para calcular la calidad de la recomendación, se utilizaron medidas subjetivas y empíricas, tomando una muestra mínima, no aleatoria, y no representativa para establecer a criterio del evaluador si la recomendación era buena

Específicamente se realizon pruebas sobre 7 URLs, solicitando 3 recomendaciones en cada caso. En la Tabla 9 se puede apreciar para cada experimento, y a juicio subjetivo del investigador si los resultados guardan alguna relación con lo solicitado o no.

**Tabla 9 - Calidad de Recomendacion**

| Prueba | Solicitado | Recomendación 1 | Recomendación 2 | Recomendación 3 |
|---|---|---|---|---|
| 1 | Identidad, Modernidad y Familia | Mal | Mal | Mal |
| 2 | CAMBIOS ECONÓMICOS Y PRIMICIA URBANA | Bien | Bien | Bien |
| 3 | Como aprender para lograr el dominio de lo aprendido | Bien | Bien | Bien |
| 4 | Sobre las maneras de tratar científicamente el derecho | Bien | Bien | Bien |

| | | | | |
|---|---|---|---|---|
| | natural su lugar en la filosofía práctica y su relación constitutiva con la ciencia positiva del derecho. | | | |
| 5 | An Investigation of Phonological and Syntactic Variation in Spoken Chilean Spanish | Mal | Mal | Mal |
| 6 | Evaluación de estrategias de un control de tránsito aéreo | Mal | Mal | Mal |
| 7 | Cómo elaborar un proyecto: guia para diseñar proyectos sociales y culturales | Bien | Bien | Bien |

Como se puede ver en la Tabla 9, en la mayoría de casos el resultado si guarda relación con lo solicitado. Por supuesto el sistema de recomendación no es perfecto y en varios casos el resultado proporcionado si bien fue la mejor opción basándose en el estudio de text mining, no es un resultado deseado.

Así mismo, cabe destacar que los resultados parecieran seguir una distribución "binaria", es decir, para todos los casos la recomendación es buena, o para todos los casos la recomendación es mala. Se desconoce y no es el alcance de este trabajo el porqué ocurre esto, pero una buena hipótesis seria que hay recursos poco o mal descritos, donde las pocas palabras que los caracterizan alientan sugerencias no validas.

En la prueba se obtuvo un 60% de recomendaciones que son claramente deseables mientras que un 40% de las recomendaciones no aportan valor adicional. Esto quiere decir que en 4 de cada 7 consultas del usuario se le sugerirán recursos de su interés, mientras que en los otros 3 casos, no. Obtener estadísticas de mayor confianza ayudaría a establecer con mayor precisión la exactitud de las sugerencias.

A pesar de lo antes mencionado, sin duda, el sistema de recomendación no es un objeto carente de valor adicional, pues en la mayoría de los casos facilita el proceso de búsqueda de información, con lo que apoya a la difusión de recursos y por lo tanto la difusión de conocimiento.

# CAPÍTULO 5
# CONCLUSIONES

Con todo lo antes indicado, se deja constancia del correcto funcionamiento del sistema. El sistema es eficaz (proporciona recomendaciones validas) y eficiente en tiempo (responde rápido), utilizando un modelo creado a partir del procesamiento de textos en un desarrollo de minería de datos que permite medir las distancias entre los recursos indexados, con lo que el sistema recomendar da sugerencias basándose en la similitud de los recursos; dicho de otra manera está basado en contenidos. Con todo ello se cumplió el objetivo general de este trabajo especial de grado.

Para lograr este objetivo general, fue necesario completar cada uno de los objetivos específicos, los cuales representan la metodología de desarrollo. Los datos fueron capturados desde una base de datos MySQL proporcionado por el tutor de este trabajo, Profesor Sosa. El repositorio, es una federación de repositorios digitales institucionales dedicados a la academia y educación. Los metadatos vienen en estándar Dublin Core y hablan sobre recursos académicos.

Estos metadatos fueron transformados de la siguiente manera: el titulo, la descripción y el tema se juntaron en un solo campo, para el procesamiento de texto. Durante el procesamiento de texto, la palabras de parada (el, la, the, to, para, entre otras ) fueron retiradas, y las demás fueron reducidas a su raíz, no sin antes haber sido re-escritas todas las palabra en minúsculas, sin números y sin espacios en blanco. Transformado el data.frame de un único campo existente (que contenía todo el texto de los 3 campos dublicore originales) a tantos campos como palabras existentes hubiera, a cada instancia se le indico la ponderación resultado de la función tf-idf. Finalmente, se escogieron aquellas 500 palabras cuya suma de ponderaciones tf-idf fuera mayor, puesto que estas serian las palabras más valiosas para la mayoría de los documentos.

Con la labor de transformación concluida, se perfilo un modelo creado a partir de agrupación jerárquica. Se utilizo este método puesto que permite la clasificación más certera, la cual era necesaria al comprobarse que algoritmos más débiles como kmedias no lograrían un buen primer acercamiento. Obtenido un modelo basado en agrupación jerárquica, se paso a clasificar las instancias basándose en la distancia mínima con respecto al punto central de cada grupo, lo cual paradójicamente se podía resolver de forma más rápida con el algoritmo de kmedias, forzando por parámetros a no mover los centros.

Obtenida la clasificación se idearon formas de evaluar el resultado. Dicha evolución arrojo que si bien el resultado era correcto, era obtenido de forma muy lenta, cosa que motivo un revisión del modelo, a fin de lograr un modelo, y unas funciones que procesaran todo más rápido y brindaran sugerencias a mayor velocidad. Las mejoras dieron resultados, se redujo a 1/30 el tiempo de ejecución promedio de una consulta.

Llegado a la conclusión que el resultado era válido y bueno, el esfuerzo se centro en hacer una función lo más robusta posible, y una interfaz agradable con la cual se pudiera desplegar y mostrar el sistema de recomendación a los usuarios interesado.

## 5.1. CONTRIBUCIÓN

El presente trabajo contribuye con el Buscador Académico Venezolano, proporcionándole herramientas e insumos para su sistema de recomendación, permitiéndole de esa manera ser un portal más completo y útil para sus usuarios. Con ese sistema de recomendación el proceso de búsqueda de información se verá mejorado y simplificado, lo que facilitara la difusión del conocimiento.

Así mismo, sirve de demostración que es posible confiar en el lenguaje R y el paquete Shiny para la creación de servidores y páginas web. Pruebas como esta, ayudan a darle difusión al uso de lenguajes estadísticos, brindando confianza a sus usuarios en la amplitud de sus herramientas. Estadísticos, científicos de datos, mineros de datos tiene en este trabajo otra fuerza que los alienta a compartir sus investigaciones y descubrimientos utilizando el paquete Shiny que no demanda mucho tiempo de aprendizaje y ni siquiera conocimientos en programación web.

Si bien es lógico y posible hacer un agrupamiento jerárquico de un agrupamiento jerárquico, valga la redundancia, no es común, y este trabajo, brinda confianza a estudiantes y desarrolladores a aplicar técnicas de este estilo para lograr sus objetivos.

Así mismo, este trabajo sirve de referencia para indicar que en los estudios de text mining, no es para nada recomendable intentar agrupar por kmedias, sino que más bien lo conveniente es utilizar agrupación jerárquica. Como consecuencia de no ser recomendable el kmedias, no se pueden hacer codos útiles en text mining.

De igual manera, se comprueba que para cada instancia de un dtm que utilice ponderaciones tf-idf la palabra/columna más valiosa será aquella que no sea igual de valiosa para otra instancia. Como consecuencia de esto, intentar hacer limpieza de datos descartando palabras que no son valiosas para ninguna instancia, no arrojaría resultados satisfactorios. Específicamente en este trabajo se buscaron las 20 palabras más valiosas por instancia mientras se marcaba con ponderación 0 desde la palabra 21 en adelante. Reescrita la matriz de esta manera se busco que columnas/palabras habían desaparecido, y se encontró, que ninguna palabra había desaparecido. El tf-idf esta tan perfeccionado en su definición que incluso para casos grandes como este, no hay muchos donde una misma palabra sea la más valiosa para 2 o más documentos.

## 5.2. RECOMENDACIONES

Una consideración esencial en el tiempo de respuesta del sistema, es que lo grupos sean de relativamente pocas instancias; menos de 1500 instancias seria lo preferible. Crear una función que automáticamente ejecute agrupación jerárquica sobre los grupos o subgrupos con más de 1000 instancias, permitiría hacer agrupaciones más confiables, con lo que se podría aumentar el tamaño de la muestra sin temor a dedicar mucho tiempo humano a la reducción del tamaño de los grupos.

Este trabajo se puede ampliar, con sistemas de calificación de resultados y perfiles de usuario, cosa que sin duda incidiría favorablemente en el Buscador Académico Venezolano. En la sección de trabajos Futuros ahondamos esta misma sugerencia.

## 5.3. LÍMITES

Este desarrollo se realizo sobre una muestra del repositorio, y no sobre la totalidad del mismo, así que hay que tomar en cuenta que los tiempos, el modelo y en general todo el sistema esta adecuado para un muestra de tamaño 161124 instancias. Obviamente se puede extender a muestras de mayor tamaño, y con ello igual se conserva el mismo modelo y mismo sistema. Sin embargo, puede haber un empobrecimiento de los tiempos, si el modelo no se adapta a la creación de grupos de menos de 1500 instancias, puesto que los tiempos de respuesta podrían ser mayores.

El trabajo midió la calidad del modelo con boxplot, y la calidad del sistema en tiempos de respuesta, pero para medir la calidad de la recomendación, se utilizaron medidas subjetivas y empíricas, como tomar una muestra mínima, no aleatoria, y no representativa para medir a criterio del evaluador si la recomendación era buena. De haberse dispuesto de un sistema de calificaciones para las respuestas, como el sistema de estrellas, y un grupo de usuarios para experimentos, se podría tener una medida más confiable de la calidad de la misma.

## 5.4. TRABAJOS FUTUROS

Este trabajo abre las puertas para la creación de otros mas, dejándose a sí mismo como referencia para desarrollos futuros. Entre estos posibles desarrollos futuros esta:

De crearse un sistema de calificación de respuestas se podría medir la calidad de las recomendaciones. Eso aportaría más confianza en el sistema, daría medidas formales de la calidad de la recomendación, y alertaría de cualquier posible desperfecto en el sistema. Un sistema de calificación muy ampliamente

utilizado es el sistema basado en estrellas, donde el usuario califica de 1 a 5 estrellas el objeto que se vaya a calificar, como se puede ver en la Ilustración 38.

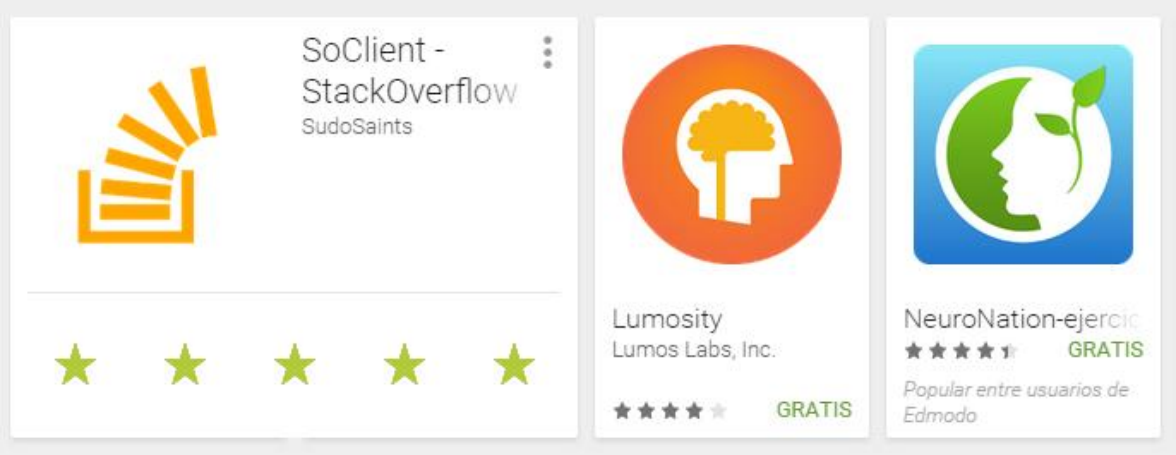


Ilustración 38- Sistema de Estrellas

El crear un sistema de recomendación basado en perfiles de usuario, le daría a cada usuario una experiencia más personal, lo cual debiera incidir en una mejor didáctica. Para lograr hacer un sistema de recomendación basado en perfiles de usuario, se necesita registrar a los usuarios, así que una incorporación importante e interesante al Buscador Académico Venezolano seria un sistema de registro de usuarios.

Así mismo, se podría brindar más atención a los documento outliers que por tener fuerte vinculación con 2 o más aéreas terminan obteniendo un clasificación mejorable. Una posible solución para estos casos, sería crear una estructura tipo red, donde los nodos serian los recursos. En esta red se permitiría asociar grados de cercanía a diferentes grupos e indicar efectivamente la existencia de documentos que pertenecen a más de un grupo. Esta misma red podría ser una opción para aumentar aún más la velocidad de respuesta del sistema de recomendación.

# REFERENCIAS BIBLIOGRÁFICAS Y DÍGITALES

A. Ruiz-Iniesta, G. Jiménez-Díaz y M. Gómez-Albarrán (2010) . "Personalización en Recomendadores Basados en Contenido y su Aplicación a Repositorios de Objetos de Aprendizaje" .URL : http://rita.det.uvigo.es/201002/uploads/IEEE-RITA.2010.V5.N1.A6.pdf

Al-Khalifa, H. S., y Davis, H. C. (2006). "The evolution of metadata from standards to semantics in elearning applications".Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, Odense, Denmark.

Almuadena Ruiz, Guillermo Jimenez-Diaz, Mercedes Gomez-Albarran y Jose Garcia Santesmases (2009)."Recommendation in repositories of learning Objects: a proactive approach that exploits diversity and navigation-by-proposing" .URL : http://www.researchgate.net/profile/Mercedes_Gomez-Albarran/publication/224572784_Recommendation_in_Repositories_of_Learning_Objects_A_Proactive_Approach_that_Exploits_Diversity_and_Navigation-by-Proposing/links/0a85e53328edc7e371000000.pdf

Almudena Ruiz Iniesta (2009). "Estrategias de recomendación aplicadas a repositorios de recursos educativos" .URL : http://eprints.ucm.es/9908/1/RuizIniesta-proyectoMaster2009.pdf

Almudena Ruiz Iniesta (2011). "Sistemas de recomendación. Presente y futuro de la web" URL : http://gaia.fdi.ucm.es/files/people/almudena/seminario/recsys-dia1.pdf

Anonimo, (2015). "What does tf-idf mean?" URL http://www.tfidf.com/

British Library. Recuperado en Enero 6 del 2006, desde http://www.bl.uk/about/strategic/glossary.html

Bruce y Hillmann. (2004). "The Continuum of Metadata Quality: Defining, Expressing, Exploiting". Metadata in Practice, D. Hillmann& E Westbrooks, eds. ISSN: 0-8389-0882-9. URI: http://hdl.handle.net/1813/7895

Carl Lagoze, Herbert Van de Sompel, Michael Nelson, Simeon Warner (2015). "The Open Archives Initiative Protocol for Metadata Harvesting". URL: http://www.openarchives.org/OAI/openarchivesprotocol.html

Chiappe, A.; (2007)."Toward an instructional design model based on learning objects."Educational Technology Research and Development.Diciembre 2007, Volume 55, Issue 6, pp 671-681

Chris Anderson (2004).”The Long Tail”. URL: http://archive.wired.com/wired/archive/12.10/tail.html

El Instituto Federal de Tecnología de Zúrich (ETH) (2015). “R-manual”. URL: https://stat.ethz.ch/R-manual/R-devel/library/

Elizabeth León Guzmán, (2015). “Metodologías Aplicadas al proceso de Mineria de Datos” URL http://disi.unal.edu.co/~eleonguz/cursos/md_2014/presentaciones/Sesion5_Metodologias.pdf

Fernando Berzal (2006). “Ejemplo de método del codo” Recuperado de: http://elvex.ugr.es/decsai/intelligent/slides/dm/D3%20Clustering.pdf

Friesen, N. (2004). “The international learning object metadata survey. The International Review of Research in Open and Distance Learning.Vol 5, No 3.Recuperado de http://www.irrodl.org/index.php/irrodl/article/view/195

Han, Jiawei , Micheline Kamber, Jian Pei (2012) . Data mining : concepts and techniques / Jiawei Han, Micheline Kamber, Jian Pei. – 3rd ed.p. cm. ISBN 978-0-12-381479-1

Institute of Electrical and Electronics Engineers (IEEE) (2002). “IEEE Standard for Learning Object Metadata”.E-ISBN :0-7381-3298-5

Institute of Electrical and Electronics Engineers (IEEE) (2002). “Estándar para Metadatos de Objetos Educativos”.E-ISBN :0-7381-3298-5

J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

JehadNajjar y Erik Duval (2008). “Actual Use of learning Objects and Metadata: An Empirical Analysis”. IEEE Technical Committee on Digital Libraries Bulletin, volume 2, issue 2, p. 1-12

JSON, (2015). “Introducing JSON” . URL : http://www.json.org/

JSON, (2015). “The JSON Data Interchange Format”. URL: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

Kahn y Wilensky (2006). “A framework for distributed digital object services” URL http://www.doi.org/topics/2006_05_02_Kahn_Framework.pdf

Karina De Sousa y Carlos Mora (2015), Predicción de Enlaces y Aplicación de Sistemas de Recomendación. Paper, Universidad Central de Venezuela – Venezuela

Kdnuggets.com (2015).” Poll Data Mining Methodology 2007” URL: http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm

Learning Objects: Collections. RecuperadoenEnero. 25, 2006, desde http://www.uwm.edu/ Dept/CIE/AOP/LO_collections.htm.

Lehman, R. (2007).” Learning Object Repositories”.New Directions for Adult and Continuing Education” Capitulo 6. DOI: 10.1002/ace

Martin Wolpers, Martin Memmel, KatjaNiemann, JorisKlerkx, Marcus Specht, Alberto Giretti, Erik Duval (2011). “Aggregating metadata to improve access to resources”. 12th IEEE International Conference on Information Reuse and Integration, August 2011, Las Vegas, USA

Moen, W.E., Stewart, E.L., McClure, C.R. (1998).“Assessing metadata quality: Findings and methodological considerations from an evaluation of the u.s. government information locator service (gils)”. T.R. Smith (ed.) ADL ’98: Proceedings of the Advances in Digital Libraries Conference, pp. 246–255. IEEE Computer Society, Washington, DC, USA

Nikos Palavitsinis (2013). “Metadata Quality Issues in Learning Repositories“.Doctoral Thesis, Universidad de Alcalá, España. URL: http://dspace.uah.es/dspace/handle/10017/20664

Ochoa, X y Duval, E. (2009). “Automatic Evaluation of Metadata Quality in Digital Repositories” International Journey on Digital Libraries 10(2-3),67-91.

Oldemarrodriguez.com (2015). “Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM”. URL: http://oldemarrodriguez.com/yahoo_site_admin/assets/docs/Documento_CRISP-DM.2385037.pdf

ONCTI (2015), “Misión y Visión”. URL: http://www.oncti.gob.ve/index.php?option=com_content&view=article&id=22&Itemid=34

Paola A Piñero G, Loy E Ramírez P (2012) “Construcción de un Objeto de Aprendizaje Web tipo Simulación Sísmica utilizando tecnologías de dibujo en capas con HTML5”.Tesis de Grado, Universidad Central de Venezuela – Venezuela

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, ColinShearer y RüdigerWirth (2000). “CRISP-DM 1.0”. URL: http://the-modeling-agency.com/crisp-dm.pdf

Real Academia Española.”Diccionario de la Lengua Espanola”.URL http://lema.rae.es/drae/?val=biblioteca.

Stvilia, B., Gasser, L., Twidale, M. (2007) A framework for information quality assessment. Journal of the American Society for Information Science and Technology 58(12), 1720–1733

The Dublin Core Metadata Initiative (DCMI). (2015). URL: http://dublincore.org/documents/dces/

The Dublin Core Metadata Initiative (DCMI). (2015). URL: http://dublincore.org/documents/2001/04/12/usageguide/

The Internet Society (2005). "Common Format and MIME Type for Comma-Separated Values (CSV) Files". URL : https://tools.ietf.org/html/rfc4180

The Internet Society (2005). "Common Format and MIME Type for Comma-Separated Values (CSV) Files". URL : https://tools.ietf.org/html/rfc4180

Universidad de Los Andes (2015). "Minería de datos: Modelos de Predicción, Modelos de Descripción, y usos en IN". URL: http://www.ing.ula.ve/~aguilar/actividad-docente/IN/transparencias/clase41.pdf

Universidad de Valencia. "MÉTODOS DE ANÁLISIS CLUSTER". URL : http://www.uv.es/ceaces/ multivari/cluster/metodos.htm

Universidad Nacional de Colombia (2015). "Descubrimiento de conocimiento en bases de datos". URL: http://www.bdigital.unal.edu.co/2037/2/germanaugustoosoriozuluaga_Parte2.pdf

Vuorikari R., Manouselis N., Duval E. (2008). "Using Metadata for storing, sharing, and reusing Evaluations in Social Recommendation: the Case of learning Resources".GoD.H. &Foo S. (Eds.)"Social Information Retrieval Systems: Emerging Technologies and Applications for Searching the Web Effectiv ely" , Hershey, PA: Idea GroupPublishing

Wikipedia. 2015 "Learning object metadata".Recuperado de : http://en.wikipedia.org/wiki/Learning_object_metadata

World Wide Web Consortium (2015). "XML Tutorial".URL: http://www.w3schools.com/xml/

Xavier Ochoa (2009). "Arquitectura SQI". Recuperado de:http://es.slideshare.net/xaoch/federacion-repositorios-objetos-de-aprendizaje

Xavier Ochoa (2009)."Federación Repositorios Objetos de Aprendizaje". URL: http://es.slideshare.net/xaoch/federacion-repositorios-objetos-de-aprendizaje

Xavier Ochoa, JorisKlerkx, Bram Vandeputte, y Erik Duval (2011). "On the use of learning object metadata: The GLOBE experience". Lecture Notes in ComputerScience, 6964, 271–284.

# Anexo A

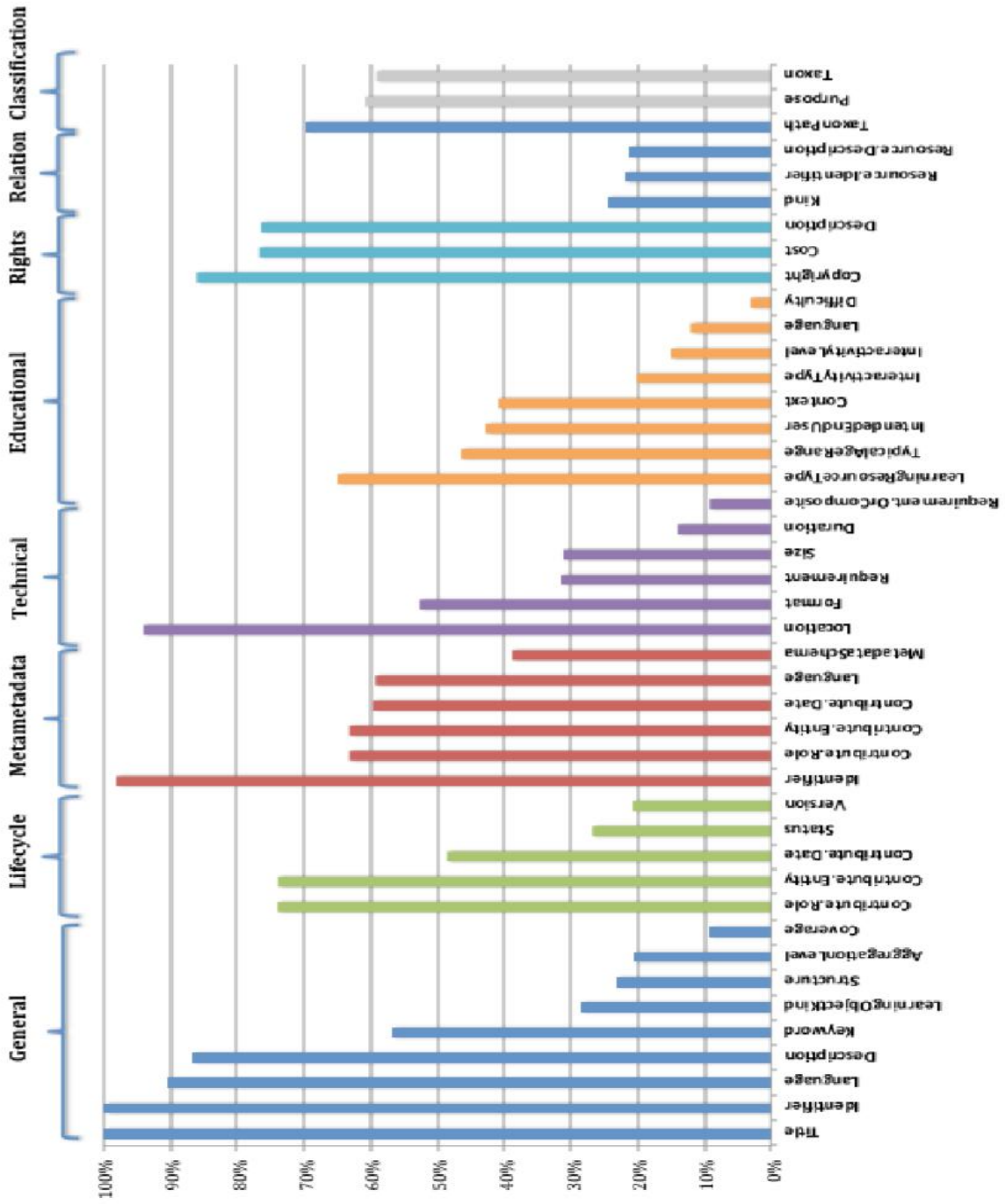


Ilustración 39 -Porcentaje de uso de los diferentes elementos de datos LOM en GLOBE (Ochoa et al. 2011).