Titulo: Mejora de las búsquedas en CMS donde predomina el contenido no estructurado
Autores: Ramírez Abella Gonzalo y Torres Matías Alejandro.
Director: Luengo, Miguel Ángel
Asesor profesional: Carsen, María Cecilia
Carrera: Licenciatura en Sistemas Plan 2003 y Licenciatura en Informática Plan 2003.

Los organismos de ciencia y tecnología productores de información generan mucha información textual que representa su producción intelectual. En la actualidad esa información es normalmente digital y se expone en sitios web o repositorios al efecto de alcanzar el público objetivo.
En ese contexto es importante su descripción. La descripción mediante metadatos es una forma de consignar los datos extraídos del recurso mismo, o que se le adicionan para su gestión, que facilita que el recurso sea identificado y recuperado en medio de la creciente masa de objetos digitales en Internet. Cada estándar de metadatos existente es un formato de descripción de objetos digitales en Internet que a su vez necesariamente se basa en estándares bibliotecarios al momento de seleccionar los términos que lo describirán y en la forma también establecida por esas normas internacionales.
En el contexto del Instituto Nacional de Tecnologías Agropecuarias se reimplementó el sitio institucional que se encontraba implementado en el manejador de contenidos Plone usando Drupal, haciendo un especial énfasis en el enriquecimiento del contenido y el mejoramiento de las búsquedas tanto internas como externas.

Metadatos, enriquecimiento de contenido, descripción documental, tesauro, reconocimiento de entidades, extracción de tópicos.

La proliferación de documentos digitales expuestos como recursos en entornos web requieren de la creación o asignación de metadatos y la necesidad de describir, etiquetar o definir aspectos informativos del documento electrónico junto al propio documento y que faciliten la recuperación e intercambio de información electrónica. La tesis aporta a la automatización de este proceso apoyándose en herramientas como los tesauros y el reconocimiento de entidades para facilitar la tarea de la descripción documental.

Se estudió el caso específico del INTA y sus necesidades de búsqueda de contenido.
Se evaluó el estado del arte en herramientas de manejo y enriquecimiento semántico de contenido.
El software de enriquecimiento de contenido fue adaptado a la problemática INTA y modificado para la integración con el CMS.
Se analizó el resultado de las herramientas en el contexto.

Ampliación de la extracción de tópicos utilizando otras herramientas de enriquecimiento semántico. Relacionamiento de términos equivalentes y relacionados brindados por el tesauro en las búsquedas.

Febrero 2017

ABSTRACT

Los organismos de ciencia y tecnología productores de información generan mucha información textual que representa su producción intelectual. En la actualidad esa información es normalmente digital y se expone en sitios web o repositorios al efecto de alcanzar el público objetivo. En ese contexto es importante su descripción. La descripción mediante metadatos es una forma de consignar los datos extraídos del recurso mismo, o que se le adicionan para su gestión, que facilita que el recurso sea identificado y recuperado en medio de la creciente masa de objetos digitales en Internet. Cada estándar de metadatos existente es un formato de descripción de objetos digitales en Internet que a su vez necesariamente se basa en estándares bibliotecarios al momento de seleccionar los términos que lo describirán y en la forma también establecida por esas normas internacionales.

En el contexto del Instituto Nacional de Tecnologías Agropecuarias se reimplementó el sitio institucional que se encontraba implementado en el manejador de contenidos Plone usando Drupal, haciendo un especial énfasis en el enriquecimiento del contenido y el mejoramiento de las búsquedas, tanto internas como externas.

Para lograr estos propósitos se probaron diversas estrategias: la descripción del contenido usando los esquemas de metadatos Schema.org y Dublin Core, la detección de entidades nombradas utilizando Stanford NER y OpenNLP, la extracción de palabras claves basadas en tesauros con KEA, y el uso de la plataforma de búsqueda SOLR junto a todas las capacidades que Drupal ofrece.

Como resultado se mejoró la experiencia del usuario al realizar búsquedas internas utilizando SOLR junto con Drupal, se demostró la gran capacidad de KEA para la detección de tópicos y se halló que OpenNLP provee mejores resultados en español que Stanford NER.

Curriculum Vitae de los Autores

Ramirez Abella Gonzalo es estudiante de la carrera Licenciatura en Sistemas de la Universidad Nacional de La Plata, Buenos Aires, Argentina. Trabaja para el INTA en el desarrollo de su sitio institucional y se especializa desarrollo web y transformación de datos.

gonzaloramirezabella@gmail.com, La Plata, Buenos Aires, Argentina

Torres Matías Alejandro es estudiante de la carrera de Licenciatura en Informática de la Universidad Nacional de La Plata, Buenos Aires, Argentina. Trabaja para organizaciones gubernamentales, entre ellas, el INTA como desarrollador, analista e infraestructura.

torresmat@gmail.com. La Plata, Buenos Aires, Argentina

INDICE

1. INTRODUCCIÓN

En este capítulo introductorio se describen los inicios de la web y su crecimiento para transformarse en el contenedor de conocimientos más grande del mundo.

1.1. EL INICIO DE LA "WEB"

No comenzada la década de los 90', Tim Berners-Lee propuso una solución a los problemas que tenían los investigadores para intercambiar y encontrar información en el CERN. El CERN, es una organización que involucra a miles de investigadores en todo el mundo trabajando en conjunto para alcanzar distintas metas, produciendo toneladas de información interconectada que evoluciona constantemente. Dada la gran producción en el CERN, los investigadores sufrían de la difícil la tarea de encontrar la información que previamente había sido registrada. Muchas veces el intercambio de información terminaba siendo en los pasillos de la organización o por llamadas telefónicas. Tim, que había comenzado su carrera en el CERN recientemente, identificó que los problemas de pérdida de información en el CERN era sólo un modelo a miniatura de lo que pasaría en el resto del mundo en un corto plazo y que, era necesario, encontrar una forma de organizar la información que les permitiera continuar.

Él pensó en la "web" como una red de notas con enlaces entre ellas (usando hipertexto o más bien, hipermedia), en donde, una persona siguiendo enlaces podía llegar a encontrar información que ni siquiera sabía que estaba buscando a través de la navegación de estos "vínculos". Las notas o "nodos" podrían representar personas, documentos, conceptos o incluso palabras claves interconectados entre sí. Estableció que este sistema de información debería:

- ser accesible a través de redes remotas,
- ser accesible desde distintos tipos de sistemas,
- ser descentralizado,
- tener acceso a los datos ya existentes usando un "gateway program" que mapee los datos existentes al modelo de hipertexto,
- dar la posibilidad de incluir contenido multimedia a futuro.

Años más tarde el primer sitio del mundo hosteado en el CERN fue dedicado al proyecto World Wide Web en la computadora NeXT de Tim Berners-Lee dando inicio a la world wide web.

1.2. WEB 1.0

Fue entonces que la década de los ‘90 marcó el inicio de una nueva forma de compartir información. En pocos años, desde que el primer servidor web, W3C httpd, fue iniciado, la cantidad de sitios creció estrepitosamente compartiendo así una cantidad abrumadora y creciente de información.

En estos primeros años, la WWW, fue un conjunto de sitios estáticos en donde el usuario que visitaba el sitio solamente consumía la información en ella dispuesta sin poder contribuir a la misma. Algunos puntos característicos de los sitios de principios de los ‘90 eran:

- Páginas estáticas por el usuario que la visita: artículos, libros, documentos científicos, guías de personas, páginas personales.
- No se podían añadir comentarios u opiniones ni contribuir al contenido expuesto.
- Todas sus páginas se creaban de forma fija y muy pocas veces se actualizaban.

1.2.1. Estadísticas sobre la cantidad de sitios

Mirando como la web fue creciendo en su primera década podemos darnos cuenta de lo rápida de su evolución gracias a la inmediata adopción del protocolo de Tim Berners-Lee por investigadores, empresas y luego por usuarios de computadoras personales (Internet Live Stats, s.f.).

Año	Websites	Crecimiento	Usuarios de internet	Usuarios por sitio	Sitios lanzados
2001	29,254,370	71%	500,609,240	17	Wikipedia
2000	17,087,182	438%	413,425,190	24	Baidu
1999	3,177,453	32%	280,866,670	88	PayPal
1998	2,410,067	116%	188,023,930	78	Google
1997	1,117,255	334%	120,758,310	108	Yandex
1996	257,601	996%	77,433,860	301	
1995	23,5	758%	44,838,900	1,908	Altavista, Amazon, AuctionWeb
1994	2,738	2006%	25,454,590	9,297	Yahoo
1993	130	1200%	14,161,570	108,935	
1992	10	900%			

1991	1				World Wide Web Project

Vemos cómo, a partir de la tabla previamente detallada, en los años 1993 y 1994 nacieron los primeros robots y crawlers de páginas completas (Wordstream, s.f.), solo 2 años después del nacimiento del primer navegador portable, indexando más de 2500 servidores (Cailliau, 1995) cubriendo la necesidad de los usuarios de la web navegar la información con los primeros robots de contenido. La web se empezaba a convertir en un repositorio de información basto de contenido y encontrar este contenido empezó a ser un problema importante a resolver por los informáticos.
A continuación, describiremos cómo comenzaron las búsquedas web de manera cronológica.

1.2.2. Evolución de los buscadores en internet

Los primeros buscadores fueron simplemente un directorio de sitios, a veces generado manualmente y, con el tiempo, generado usando bots que guardaban en bases de datos de distinto tipo la información del sitio. En un principio se indexó solamente la URL, a esto a veces se le agregaba una descripción manual, con el tiempo el tag TITLE fue indexado hasta que comenzaron a aparecer los indexadores de texto complejos con sistemas de ranking de resultados para sus listados.

1.2.2.A. Cronología de buscadores (Wordstream, s.f.)

- 1990: Archie
    - Primer buscador: fue un servidor FTP con los listados de directorios “descargables”.
    - Solo se indexaba la referencia al sitio y no su contenido.
- 1991: Verónica and Jughead
    - Índice de nombres de archivos y títulos construido con Gopher.
- 1992: VLib
    - Una librería virtual realizada por Tim Berners-Lee.
- 1993: World Wide Web Wanderer
    - En un principio era un robot que recorría la web para calcular el “crecimiento de la web”. El robot fue luego actualizado para que, además de contar sitios, almacene URLs. Consumía demasiado ancho de banda ya que ingresaba varias veces por día al mismo sitio.
- 1993: ALIWEB

    - ALIWEB es considerado el primer buscador de la Web ya que sus predecesores eran construidos por diferentes propósitos (Wanderer, Gopher) o simplemente eran indexadores (Archie, Veronica and Jughead). Permitía a los usuarios ingresar las ubicaciones de los índices de sus páginas para agregarlos a los sitios indexados por el motor de búsqueda y, a la vez, les permitía agregar palabras claves y descripciones de cada uno de sus sitios. Esto incentivó a los webmasters a definir términos y descripciones para sus páginas y, además, evitar el uso de robots que consumían ancho de banda. ALIWEB no fue usado ampliamente ya que muy pocas personas ingresaban sus sitios en el motor de búsqueda.
- 1993: Primeros crawlers de contenido
    - Jumpstation: Usaba búsqueda lineal para guardar el título y la cabecera de la página indexada.
    - World Wide Web Worm: listaba el título y la URL de la página en el orden en que fueron indexados (sin ranking).
    - RBSE Spider (Gate, 2016):
        - Mostraba los resultados de búsqueda según un ranking.
        - Indexaba el tag “TITLE” y la URL.
        - La búsqueda era realmente difícil ya que se debía poner exactamente el título de contenido para que aparezca.
        - Fuente:
    - Webcrawler:
        - Primer buscador que indexaba toda la página.
        - Demasiado popular como para ser usado durante horas pico de uso.
    - Yahoo Search:
        - Comenzó siendo un listado de páginas favorables que al crecer demasiado rápido debieron indexar el directorio.
        - Manualmente se ingresaba una descripción del sitio.
    - Looksmart:
        - Debe ser la única compañía que depende de humanos para realizar las indexaciones.
        - Hace unos años compró el crawler WiseNut para complementar los resultados de los humanos.

        - Muy poca gente utiliza looksmart para realizar búsquedas. Hoy en dia actúa como un proveedor.
    - Google
        - Lanzado en 1998 como un proyecto de la universidad de Stanford.
        - Por su habilidad de analizar los enlaces a través de la web, produjo una nueva generación de resultados. Por varios motivos, es hoy en día el buscador más utilizado.

1.3. WEB 2.0

El término "Web 1.0" surgió al mismo tiempo que el término "Web 2.0" y se usa en relación con este segundo para comparar ambos ya que, en realidad, la web 2.0 se basa en los estándares que definió la "Web 1.0" anteriormente y por ende no debe ser usado en contraposición sino más bien que la "Web 2.0" es la consecuencia de una mejor implementada "Web 1.0". (Graham, 95)

Paul Anderson en su reporte "What is Web 2.0? Ideas, technologies and implications for education" (Anderson, 2007) dice que detrás de lo que parece ser un nuevo conjunto de tecnologías y servicios en la Web 2.0, existen 6 ideas poderosas que cambiaron la forma en que las personas interactúan con ella:

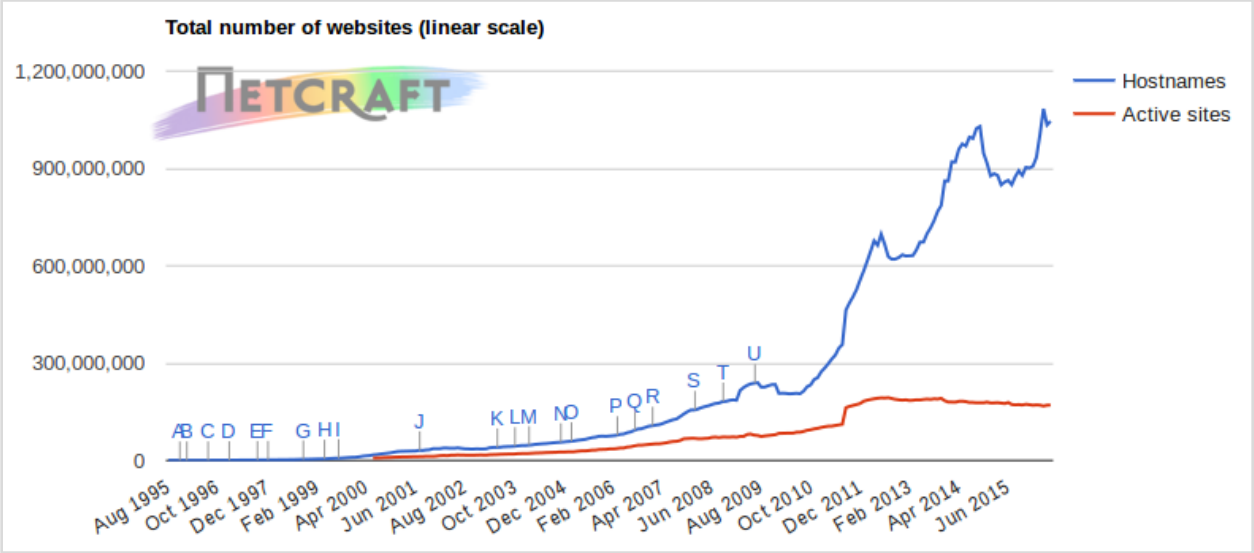
- Producción individual y contenido generado por el usuario: hoy en día los usuarios de la web tienen el poder de crear contenido a través del cual pueden expresarse libremente. Ya sea a través de un blog, videos, o un sitio para exponer sus creaciones musicales, fotografías, pinturas, opiniones sobre artículos, noticias, etc, etc en sitios y tecnologías cambiantes constantemente. La explosión de los blogs de hace unos años atrás fue reemplazada por otras nuevas redes sociales; la web evoluciona constantemente y cada vez incluye más gente a participar en su contenido.
- Concentrar el poder de la multitud: Se utilizan herramientas que obtienen datos o toman decisiones a partir de la participación de muchos usuarios. La idea principal es que una decisión que toman muchos es mejor o más fuerte que una decisión individual. Como ejemplo se puede nombrar el "Folksonomy". El cual es el resultado del etiquetado libre por parte de los usuarios tanto de contenidos como de objetos (cualquier cosa con una URL) para la propia recuperación. El etiquetado es realizado en un ambiente social, por las personas que consumen la información. Esto puede "filtrar" personas que poseen el mismo vocabulario y hacer filtros de usuarios.

- Arquitectura de participación: esto significa que la forma en que un servicio está diseñado realmente puede mejorar para el usuario, como efecto lateral del mismo uso, ya que fue diseñado para con las interacciones del usuario sean utilizadas para esto.
- El efecto de la red: es un término utilizado en la economía y se refiere al aumento del valor del producto cuando más usuarios lo utilizan. Por ejemplo Facebook, la red social fue aumentando y brindando mejores servicios para los usuarios a medida que mas y mas usuarios comenzaron a utilizarlos.
- Openness: trabajar con estándares abiertos, el uso de software de código abierto, el uso de datos libres, la reutilización de datos y trabajar con un espíritu de innovación abierta. Por ejemplo el navegador Firefox y su sistema extensible de plugins que permiten la experimentación por parte de los usuarios.

1.3.1. La web como contenedor de conocimiento

En el año 2015 el número de sitios en Internet fue 53059 veces mayor a aquel registrado en el año 1994 y, año tras año, crece enormemente. ¿Puede imaginarse esa cantidad de información apilada? Realmente dudo que pueda; es prácticamente imposible imaginarse todos los contenidos que abarca la web. Cada vez que un buscador nos devuelve resultados, ¿Cómo sabemos que es el mejor resultado? ¿Cómo sabemos que este buscador entendió todos los contenidos que la web tiene sobre la búsqueda realizada? ¿Cómo es posible que un agente de software, ya sea un bot o un crawler, entienda realmente lo que está indexando y es, siquiera posible, que pueda vincularlo con otros “conocimientos” similares?

En el gráfico siguiente (NetCraft, 2016), vemos el crecimiento de hostnames y websites activos entre los años 1995 y 2015.



El crecimiento exponencial que demuestran las estadísticas nos está mostrando como la web ha dejado de ser una guía estática de información, cambiando rápidamente a ser ***el contenedor de conocimiento más importante*** del mundo. Este crecimiento está impulsado por la industria, la educación y la comunidad en general. Hoy en día la web debe soportar la interacción entre una gran cantidad de vendedores independientes de fuentes de información y aplicaciones obsoletas corriendo en plataformas heterogéneas y redes de información distribuidas.

Es claro que las ciencias informáticas han tenido que enfrentar una serie de nuevos problemas en un corto lapso de tiempo con algo que se volvió tan masivo que no ha detenido su crecimiento desde que por primera vez apareció. Los buscadores de internet han tenido que, primero, aparecer y, luego, evolucionar rápidamente desde un listado manual de resultados sin ningún tipo de ordenamiento hasta la búsqueda que hoy en día existe donde se le brinda al usuario la información que necesita. Para que un indexador pueda entender un documento y ponderar bien los resultados de búsqueda se necesitan entender las técnicas que describen información sobre la información -metadato- lo que permite hacer una descripción legible por agentes de software.

Así como los motores de información e indexado deben soportar lo arriba mencionado también deberán hacerlos las futuras aplicaciones. Estas nuevas aplicaciones deberán poder consultar y consumir información relacionada entre todas las fuentes de datos ya que es este el patrón que la comunidad está pidiendo.

Todos estos cambios son el desafío que las aplicaciones deben, hoy en día, enfrentar. Este tipo de sistemas son llamados “sistemas cooperativos de información” y deberán hacer uso de las tecnologías semánticas de la web dado que la información debe hacerse disponible junto

con sus metadatos de forma tal que un autómata o agente de software pueda leerla y hacer que la web sea accesible para las computadoras.

1.3.2. Motivaciones de la web semántica y la web 3.0

En el libro "A semantic web primer" se define el papel de las computadoras en la web de hoy en día de la siguiente manera: "La visión general de la web semántica puede ser resumida en una sola frase: hacer la web accesible a las computadoras. Hoy en día la web es un contenedor de texto e imágenes. Esta información es muy útil para las personas, pero las computadoras juegan un papel muy limitado en la web actual: indexan palabras y envían información de servidores a clientes. Todo el trabajo inteligente de seleccionar, combinar y agregar información debe ser realizado por el lector humano." (Grigoris & Frank, 2004).

1.3.2.A.Comenzando a describir los contenidos

Durante la década de los '90 y principios del milenio las páginas web estaban escritas para humanos más que para programas. HTML sigue siendo el lenguaje predominante en que las páginas web son escritas y, para el lector humano, es satisfactorio. Para una máquina, en cambio, no es posible entender una página escrita sólamente en HTML; esta necesita de un lenguaje más descriptivo, un lenguaje que permita dar información sobre la información misma, en otras palabras, **metadatos**. Como se describe más adelante, un metadato es un dato que provee información sobre uno o más aspectos del contenido. Si estuviésemos leyendo una obra literaria los metadatos de ésta podrían describir su autor y colaboradores junto con sus propósitos y fechas de creación. La idea de adjuntar los metadatos a un contenido cualquiera sea es permitir que un agente de software pueda entender el contenido sin la necesidad de utilizar, por ejemplo, inteligencia artificial.

Un primer paso en esta dirección es eXtensible Markup Language, que permite definir la estructura de la información de la web. Por ejemplo:

```
<obraLiteraria>
  <título>Como perros y gatos</titulo>
  <autor>José Luis Herrera de la Fuente</autor>
  <palabra clave>reino animal</palabra clave>
  <palabra clave>perros</palabra clave>
  <palabra clave>gatos</palabra clave>
  <fecha de edición>10 de julio de 1978</fecha de edición>
</obraLiteraria>
```

Su representación es mucho más fácilmente procesable que el lenguaje HTML, sobre todo, para el intercambio de información. Sin embargo, una descripción XML de un contenido sigue siendo una descripción sintáctica y de la estructura de la información. Un computador

no podría entender las diferentes “secciones” o “elementos” de un XML; las palabras “autor”, “obraLiteraria” y “palabra clave” no tienen significado alguno para un analizador de contenidos. Para esto se implementó un lenguaje para describir la semántica o significado de un contenido. Este lenguaje es llamado RDF y es el lenguaje básico de la web semántica. RDF nos permite realizar declaraciones sobre secciones de información usando XML como base. En nuestro ejemplo anterior, la descripción RDF del contenido sería:

- El documento es un libro
- El nombre del libro es “Como perros y gatos”
- José Luis Herrera de la Fuente es un escritor.
- José Luis Herrera es el autor de “Como Perros y gatos”

Los metadatos descritos por el ejemplo anterior son información semántica del contenido descrito en la página que usamos de ejemplo. Con el tiempo varias organizaciones fueron creadas para el desarrollo de extensiones al lenguaje RDF.

Resource Description Framework o RDF es un idioma para presentar información sobre recursos en la Web. Su concepción es principalmente para presentar metadatos sobre recursos en la web, tales como título, autor, y fecha de modificación de una página web, información sobre derechos del autor o condiciones de uso de un documento web, u horario de disponibilidad de un recurso compartido. Pero, por generalización del concepto del "recurso en la Web", se puede usar RDF también para presentar información sobre cosas que se puede identificar en la web, aún que no se puede descargarlas de la web. Ejemplos incluyen información sobre cosas disponible por e-commerce (por ejemplo, información sobre especificación, precios y disponibilidad) o descripción de las preferencias de un usuario de la Web para la manera de recibir información. (W3C, 2014)

2. ARQUITECTURA DE INFORMACIÓN

En el presente capítulo se aborda una disciplina llamada arquitectura de información, se da una introducción sobre la misma, y la necesidad por parte de los desarrolladores de sitios web, de obtener una visión más amplia sobre los desarrollos.

2.1. ARQUITECTURA DE INFORMACIÓN

La arquitectura de la información es una disciplina de diseño focalizada en hacer que la información será entendida y localizable. Esta disciplina nos permite pensar en estos problemas a través de dos perspectivas: Que los productos y servicios de información son percibidos por las personas a través de sitios hechos de lenguaje, y estos ambientes de información pueden ser organizados para una óptima capacidad de búsqueda y entendimiento.

Como se indicó previamente la web pasó a convertirse en el contenedor de información más grande del mundo en donde se encuentran sitios con dominios de datos específicos al conocimiento que les confiere. Este conocimiento necesita estructurarse de forma tal que se le pueda dar uso a la información a través de la aplicación que la presenta. Aquí es donde entra el juego la arquitectura de información para que estos sitios puedan evolucionar respondiendo a las necesidades de los usuarios.

La mayoría de las aplicaciones de software son diseñadas para resolver problemas específicos y, a su vez, las aplicaciones que triunfan tienden a abarcar más y más funcionalidades con el tiempo. Como resultado pierden claridad y simplicidad. Un ejemplo de esto es la plataforma multimedial “iTunes” que empezó su vida como una herramienta para habilitar digitalización y manejo de colecciones de música en computadoras personales, creció y se transformó en una plataforma de medios completa donde no solo se manejan colecciones de música sino también películas, audio de libros, podcasts, compras, alquileres, streaming y abarcando varios sistemas operativos y paradigmas. En otras palabras, pasó de ser una herramienta a ser un ecosistema.

La proliferación de información y dispositivos mencionados es un problema con el que muchas organizaciones se encuentran y luchan contra él. Lo que se necesita es un enfoque sistemático, comprensivo y holístico para estructurar la información de una forma que sea fácil de encontrar y entender sin importar el contexto, canal o medio que el usuario utilice para acceder. En otras palabras, alguien, tiene que dar un paso afuera del desarrollo del producto, mirar la foto grande en abstracto desde afuera para entender cómo encaja en

conjunto y que la información pueda ser encontrada más fácil y entendida. La arquitectura de información puede ser utilizada como un lente para ayudar a equipos e individuos a obtener esta perspectiva.

2.2. Lenguaje

Como se indicó previamente, la experiencia de la utilización de productos digitales y servicios, crece constantemente, a través de múltiples dispositivos en diferentes tiempos y lugares. Es importante reconocer que estamos interactuando a través de estos productos y servicios mediante el uso de lenguajes, etiquetas, menús, descripciones, elementos visuales, contenido y sus relaciones con otros contenidos. Estos crean un ambiente que diferencia la experiencia y las facilidades de entendimiento -o no-, por parte del usuario. Por ejemplo, el lenguaje empleado para una aplicación de recetas de un teléfono es diferente a un sitio web de una compañía de seguros. Estas diferencias en el lenguaje ayudan a definir diferentes "lugares" donde la gente puede visitar para saciar ciertas tareas específicas. Se crea un marco de información conveniente, permitiéndonos entender los conceptos relativos que ya conocemos.

En su Libro "Understanding Context, information" (Hinton, 2014), el arquitecto Andrew Hinton argumenta que damos sentido a estas experiencias tanto como lo hacemos en espacios físicos: "eligiendo palabras particulares e imágenes, que definen qué se puede y que no se puede hacer en el ambiente" siendo en un campo abierto idílico o en una búsqueda web. Las experiencias digitales son nuevos sitios creados a partir de estos tipos de lenguaje; el desafío en el diseño reside en que sean coherentes a través de múltiples contextos. Como dice Andrew: "La arquitectura de la información es una disciplina para atender estos desafíos. Una disciplina que ha estado trabajando de una forma u otra por décadas."

2.3. Pensando los sistemas

Teniendo en cuenta este énfasis en abstraer soluciones a desafíos complejos, la arquitectura de información también requiere que el diseñador piense sistemáticamente sobre los problemas cotidianos.

Mientras otras disciplinas se enfocan en diseñar artefactos particulares, la arquitectura de información tiene que tener en cuenta sistemas complejos que integran distintos artefactos individuales (aplicaciones, sitios web, e interfaces de voz, etc).

El libro de Peter "Interwingled" menciona los peligros de pensar con bajo nivel los diseños de productos y servicios.

“En la era de los ecosistemas, ver la gran imagen es más importante que nunca, y menos probable. No es algo simple, ya que las grandes organizaciones y la exigencia de la especialización nos sometió a trabajar bajo la presión de lo ‘micro’. Nos gusta estar ahí, nos sentimos seguros, pero no lo estamos. Este no es un tiempo para cerrarse, sino un tiempo para salir de la caja y conectarse.”

No podemos diseñar productos y servicios que sean efectivos y coherentes a través de varios canales de interacción si no entendemos cómo influyen e interactúan con otros sistemas que son afectados por ellos. Como mencionamos antes, cada canal de interacción trae diferentes limitaciones y posibilidades que deben contemplar el todo. Una comprensión y entendimiento de alto nivel del ecosistema ayuda a que los elementos trabajen juntos para presentar experiencias coherentes a los usuarios. Como una disciplina, la información de arquitectura es especial para la tarea.

El foco de la información de arquitectura no son solo los modelos abstractos de alto nivel, el diseño de productos y servicios que sean accesibles y entendibles requieren la creación de muchos artefactos de bajo nivel también. Habitualmente la gente cree que la arquitectura de la información es meramente la estructura de la navegación web, y esta visión no es total. Los menús de navegación y los de su clase, son ciertamente resultados del proceso, pero no se puede llegar a ese resultado sin haber explorado el territorio abstracto primero. Ambientes efectivos de información tienen un balance entre la estructura (Alto nivel) y flexibilidad (bajo-nivel).

Obteniendo una visión de niveles que se nutre (y que informa) actividades del día a día es también una forma de asegurar que se están resolviendo los problemas correctos. En el libro “Introduction To General Systems Thinking” Gerald Weinberg (M.Weinberg, 2001) utiliza la siguiente historia para ilustrar lo que él llama errores de pensamientos absolutos:

Un cura estaba caminando y pasó por una construcción, vio a dos hombres levantando una pared. -Que están haciendo? le preguntó al primero,

-Estoy poniendo ladrillos, contestó con brusquedad.

-¿Y usted?, le preguntó al otro.

-Estoy construyendo una catedral, contesto feliz.

El Cura quedó impresionado con el idealismo de este hombre y su sentido de participación en el gran plan de Dios. Compuso un sermón sobre el sujeto y regresó al día siguiente para hablar con el inspirado albañil. Solo el primer hombre estaba trabajando.

-¿Dónde está tu amigo?, preguntó el cura.

-Lo despidieron.

-Que terrible. ¿Por qué?

-El pensó que estábamos construyendo una catedral, pero estamos construyendo un garaje.

Entonces preguntémonos, ¿estamos diseñando una catedral o un garaje? La diferencia entre los dos es muy importante, y a menudo es difícil saber de ellos cuando la atención está puesta en la colocación de los ladrillos. A veces -como en el caso de iTunes- los diseñadores comenzaron diseñando un garaje y antes de darse cuenta, tenían un altar, un coro y ventanas de cristal, lo cual lo hizo difícil de entender y manejar.

La arquitectura de información, ayuda a asegurarse que estás haciendo planes para un gran garaje (El mejor del mundo) o una catedral.

2.4. Definiendo una arquitectura de información

No se puede definir de una forma simple y completa una arquitectura de información, por este motivo existe complejidad a la hora de diseñar productos y servicios digitales.

La arquitectura de información se comprende un poco más definiendo los conceptos semánticos básicos, pero si bien esto nos da un acercamiento a esta definición, debemos saber que es meramente una aproximación a lo que es realmente este concepto, porque las definiciones son imperfectas y limitantes al mismo tiempo. Una definición de una arquitectura de información en sí es una gran ilustración de este paradigma.

Vamos a descender a lo básico, explayándose en cuatro conceptos.

2.4.1 Información

Utilizamos el término información para distinguir arquitectura de información entre datos y gestión de conocimientos. Los datos son hechos y figuras, las bases de datos relacionales son altamente estructuradas y producen respuestas específicas a preguntas específicas, son a su vez poderosas pero limitadas. El conocimiento son las “cosas” que tienen en la cabeza la gente. Los gerentes del conocimiento desarrollan herramientas, procesos e incentivan a la gente a compartir su saber, y la información existe en ese medio desordenado. En los sistemas de información a diferencia de las bases estructuradas, no siempre existe una respuesta correcta a una pregunta dada. Sino que estamos interesados en información de todo tipo y forma, por ejemplo:

Sitios web, documentos, aplicaciones de software, imágenes, etc. También estamos interesados en los metadatos, estos son:

Términos usados para describir y representar objetos como documentos, personas, procesos y organizaciones.

2.4.2. Estructuración, organización y etiquetado

La estructuración consiste en determinar los niveles apropiados de granularidad para los "átomos" de información de su producto o servicio, y la decisión de cómo relacionarlos entre sí.
Organización implica un agrupamiento de estos componentes en categorías significativas y distintivas, creando contextos apropiados para que los usuarios entiendan el ambiente en el que están y qué están viendo.
Etiquetar es la acción de nombrar estas categorías y definir una estructura de navegación de elementos para conducir hacia ellas.

2.4.3. Búsqueda y gestión

La buscabilidad es un factor crítico para el éxito que pueda tener un sitio. Si los usuarios no pueden encontrar lo que ellos necesitan a través de una combinación de navegación, búsquedas y preguntas, el sistema falla. Pero diseñar para las necesidades de los usuarios no es suficiente, ya que las organizaciones y personas que manejan la información son importantes también. Una arquitectura de información debe balancear las necesidades de los usuarios con los objetivos de negocio. El manejo eficiente de contenidos, políticas y procesos claros son esenciales. (P., Jorge, & L.)

2.4.4. Arte y ciencia

Disciplinas como la ingeniería en usabilidad y metodologías como etnografía traen rigor al método científico para analizar las necesidades de los usuarios y los comportamientos de búsqueda de información.
Somos cada vez más capaces de estudiar patrones de uso y consecuentemente implementar mejoras en nuestros sitios web, pero la práctica de arquitectura de información nunca puede ser reducida a números. Hay mucha ambigüedad y complejidad, los arquitectos de la información deben confiar en la experiencia, la intuición y la creatividad. Tenemos que ser capaces de tomar riesgos y de confiar en nuestra intuición. Ese es el arte de la información de la arquitectura.

2.5 Diseñando para encontrar

La arquitectura de la información no está restringida a taxonomías, motores de búsqueda y otras herramientas que ayudan a los usuarios a encontrar cosas en un ambiente de información. La arquitectura de información comienza con la gente y la razón por la que van al sitio o utilizan el servicio. Ellos tienen una necesidad de información.

Es importante entender las necesidades y comportamientos de los usuarios y darle forma al diseño para corresponder acordemente estos requerimientos, no hay un mejor objetivo de diseño de una arquitectura que el de poder satisfacer estas necesidades.

2.5.1. El modelo "too simple"

Existen diferentes modelos de que es lo que ocurre cuando la gente realiza una búsqueda de información. El modelado necesita y nos fuerza a hacernos preguntas sobre el tipo de información que las personas quieren, cuánta información es suficiente y como se interactuara con esa información.

Por desgracia, el modelo "too simple" es el modelo más común y también el más problemático, es algo similar a esto:

1. Usuario hace una pregunta
2. Algo sucede (Búsqueda o navegación)
3. Usuario Recibe respuesta
4. FIN.

Los usuarios no siempre saben exactamente qué es lo que quieren, entonces sucede que uno visita un sitio solo para navegar, explorando el sitio uno trata de obtener información de cierto tipo, aun cuando uno no sabe qué es lo que busca, ni cómo expresarlo.

Este modelo esencialmente ignora el contexto, porque rara vez se enfoca en que es lo que pasa mientras el usuario interactúa con la información. La información necesita de un contexto, este es todo lo ocurrido desde antes y después que el usuario deja el teclado.

El modelo es erróneo porque está basado en una mala concepción; donde las búsquedas son un problema sencillo que pueden ser resueltos por un simple algoritmo. Después de todo ya resolvimos el problema de devolver datos.

Con las tecnologías como SQL, entonces pensamos que podemos tratar estas ideas abstractas y conceptos embebidos en nuestras bases, documentos semi-estructurados de la misma manera.

Esta actitud ha llevado a malgastar millones de dólares en motores de búsqueda y otras tecnologías que funcionan si estas afirmaciones fueran correctas.

2.5.2. Necesidades de información

Cuando alguien visita un sitio web para encontrar algo, ¿Qué es lo que realmente quiere? En el modelo simple el usuario quiere la "respuesta correcta"; ¡es cierto!, las respuestas correctas

se encuentran buscando en las bases de datos que guardan hechos y responden preguntas que tienen “respuestas concretas”. Por ejemplo, ¿Cuántos habitantes tiene La Plata?

Pero los sistemas digitales abarcan más información que solamente datos estructurados, no sorpresivamente, el texto es el dato más común guardado, y el texto en sí puede ser ambiguo, ideas y conceptos mezclados.

¿Qué es lo que la gente busca? Utilizamos una analogía con la pesca para explicarlo.

- La pesca perfecta, a veces los usuarios realmente están buscando una respuesta concreta, cuántos habitantes tiene la ciudad de La Plata, un sitio como Wikipedia, brindara el dato y se lograra la respuesta (521.936 - 1991). El modelo simple lo obtendrá.

Cuando se espera encontrar la pesca perfecta, generalmente el usuario sabe qué busca, cómo llamarlo y donde buscarlo.

- Pesca de langosta, ¿qué pasa cuando no se está buscando una simple respuesta? Por ejemplo: Buscar hoteles en Bariloche. Son búsquedas en las que no sabemos bien qué es lo que deseamos hallar y no esperamos encontrar la respuesta definitiva porque no sabemos si la obtuvimos o no. Similar a la pesca, con trampas para langostas, donde lo que obtenemos desde el mar es lo suficientemente bueno. Esto es una búsqueda exploratoria donde el usuario no está totalmente seguro de que es lo que busca, pero va aprendiendo algo en el proceso de exploración, que lo hace volver a hacer una búsqueda, no hay un claro objetivo de una búsqueda correcta, sino que es feliz obteniendo buenos resultados utilizándolos como trampolín hacia la nueva interacción de esta, no es simple determinar cuándo una búsqueda de este tipo llega a su fin.
- Redes de derivación indiscriminada, hay momentos en los que no se quiere dejar ninguna piedra de la mar suelta en la búsqueda de algún tópico, por ejemplo, haciendo una búsqueda para un trabajo de tesis o estudiando alguna condición médica. En estos casos el usuario quiere obtener toda la información posible sobre un tema en particular; el usuario tiene varias formas de expresar lo que está buscando y tiene la paciencia para construir las búsquedas utilizando todas las variaciones de términos.

3. BÚSQUEDAS

3.1 BÚSQUEDAS EN LA WEB

En la actualidad las búsquedas son una parte integral de la sociedad. Con más de 197.9 billones de búsquedas realizadas cada mes, aproximadamente 6.6 billones de búsquedas diarias, esto implica unas 7500 búsquedas son realizadas cada segundo de cada día (Comscore, 2014).

La demanda de las búsquedas continua, la gente puede encontrar en cuestión de segundos lo que hace 20 años atrás implicaba un viaje a la librería, el uso de un catálogo de fichas y el sistema decimal de Dewey y una búsqueda física a través de volúmenes impresos, lo que habría consumido al menos un par de horas. A través de los nuevos canales de búsqueda las personas pueden llevar a cabo muchas de sus compras, banca y transacciones sociales en línea. Algo que sin duda ha cambiado la forma de interacción de la población mundial.

Este cambio dramático en el comportamiento sitúa a los motores de búsqueda en el centro de la escena. La forma de aprender, trabajar, compartir, jugar, comprar e investigar ha cambiado para siempre. Tanto las organizaciones, las empresas, las marcas, las personas tienen que tener presencia de algún tipo en la web. Y necesitan la de los motores de búsqueda y las funciones de búsqueda en todas las plataformas para generar la exposición y facilitar el acoplamiento.

A medida que nuestra sociedad se mueve cada vez más a una economía de consumo profesional ("prosumer"), las formas en que las personas crean, publican, distribuyen, y en última instancia, buscan información y recursos en la Web. Vamos a investigar más a fondo cómo la búsqueda, y por lo tanto buscar la optimización del motor interno del sistema como externo, está en el centro del ecosistema web y por lo tanto es la clave del éxito en la economía digital en constante evolución.

3.2. FUNCIONAMIENTO Y ETAPAS DE UN MOTOR DE BÚSQUEDA

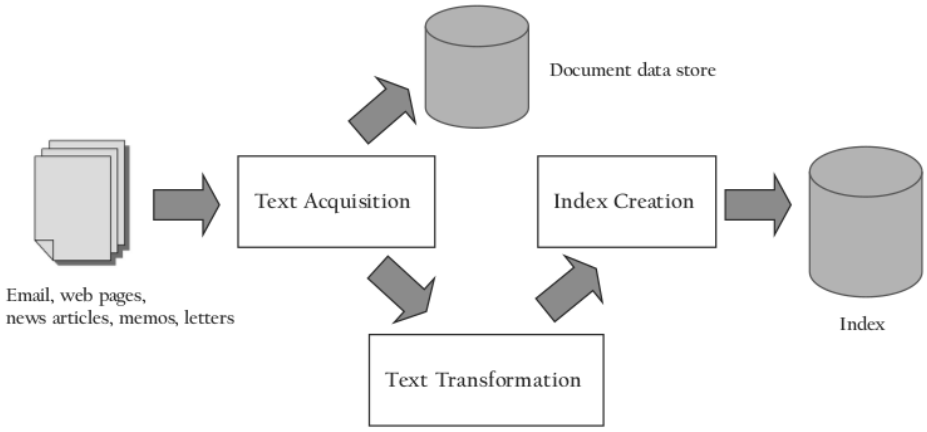
3.2.1 Breve introducción

Para ofrecer mejores resultados, los motores de búsqueda deben primero intentar descubrir todas las páginas publicadas en la web para luego poder devolver las que mejor atienden la consulta del usuario. El primer paso en este proceso es el rastreo de la web, junto con el rastreo viene la indexación la cual, usando técnicas diversas, la clasifica para luego permitir la consulta por parte de los usuarios. Un motor de búsqueda empieza por un conjunto de sitios

que son conocidos por su calidad y luego recorren los enlaces descubriendo las páginas relacionadas. Siguiendo estos enlaces los llamados rastreadores o arañas pueden alcanzar los trillones de documentos interconectados que existen en la web descartando incluso aquellos sitios que no son aptos para el indexado.

Además del rastreado y el indexado está el proceso de consulta; los buscadores como Google y Bing, no solo devuelven resultados correspondientes a un query del usuario, sino que además la búsqueda debe estar ordenada por relevancia.

3.3 DESCOMPONIENDO UN MOTOR DE BÚSQUEDA



3.3.1 Obtener el texto

3.3.1.A Crawler

El componente crawler tiene la responsabilidad principal de identificar y rastrear documentos para el motor de búsqueda. Hay muchos tipos de crawlers, pero el más común es un agente de software diseñado para la web el cual sigue enlaces en las páginas web descubriendo y descargando nuevas páginas, aunque esto suene simple existen diferentes retos al diseñar un crawler web que eficientemente pueda manejar grandes volúmenes de nuevas páginas en la web y, además, páginas ya rastreadas que probablemente hayan cambiado. Un crawler web puede estar restricto a un solo sitio, como podría ser una universidad o entidad gubernamental.

3.3.1.B Conversión

Los documentos encontrados por un crawler o un feed comúnmente no están en formato de texto simple, generalmente están en otros formatos variados como pueden ser HTML, XML, Adobe PDF, Microsoft Word, Microsoft Power Point, etc. La mayoría de los motores de

búsqueda requieren una conversión de estos formatos a un texto formateado con metadatos considerando que el formato particular del documento puede ser relevante en un procesamiento posterior. En esta conversión las secuencias de control y los contenidos que no son datos asociados con algún formato particular son eliminados o guardados como metadatos.

3.3.1.C Almacenamiento de datos

Base de datos utilizada para el manejo de grandes volúmenes de documentos y datos estructurados asociados a ellos. Los datos estructurados consisten en metadatos del documento y otros extractos de información como enlaces o palabras claves. Estas bases pueden ser bases relacionales o en casos de mayor volumen otros tipos de sistemas de almacenamiento más eficientes. Hoy en día, los productos de motores de indexación incluyen una base de datos para guardar sus índices.

3.3.2 Transformación del texto

3.3.2.A Análisis gramatical

El análisis gramatical procesa la secuencia de palabras en el contenido para reconocer elementos estructurales como títulos, enlaces, y encabezados. ***Tokenizar*** el texto es el primer paso de este proceso, en la mayoría de los casos los tokens son palabras. Tanto los contenidos como la consulta realizada por el usuario, deben ser "tokenizadas" de la misma manera para una eficaz comparación. Hay varias decisiones que potencialmente afectan la recuperación de datos, lo que hace no trivial este paso.

A modo de explicación una definición simple sería separar en tokens las cadenas de caracteres separadas por un espacio. Esto no nos dice, sin embargo, como afrontar los casos especiales como los caracteres mayúsculas, los guiones, las apóstrofes, etc. ¿Debemos tratar de la misma forma la palabra "Apple" que "apple"? En algunos lenguajes, la tokenización es aún más interesante; en el lenguaje chino no hay una separación entre las palabras clara como en el castellano.

Al hacer el análisis gramatical hay que tener en cuenta, también, la estructura de los contenidos que en el caso web está generalmente en lenguaje HTML o XML. Tanto el HTML como el XML utilizan etiquetas para definir elementos, por ejemplo, la etiqueta "<h2></h2>" define un segundo nivel de encabezado o importancia en el texto. Las etiquetas y otras secuencias de control deben tratarse apropiadamente al tokenizar. Otros tipos de documentos,

como el email y las presentaciones, poseen sintaxis y métodos específicos para una estructura específica. Pero muchos de estos serán removidos.

3.3.2.B Stopping -stopwords-

Este componente tiene la simple tarea de remover palabras comunes de los tokens. Las palabras más comunes como “de”, “a”, “el”, “para” o “como”; son tan comunes, que removerlas puede reducir el tamaño de los índices considerablemente. Dependiendo en el modelo de recuperación elegido para los aspectos básicos del ranking, eliminar estas palabras no tienen impacto en la efectividad de la búsqueda, y hasta las mejora.

El problema de estas listas de palabras a eliminar, es que resulta imposible para los motores encontrar resultados a consultas como “to be or not to be”. Para evitar esto, las aplicaciones de búsqueda utilizan listas pequeñas (quizás solo conteniendo “el”) cuando se procesa el texto, pero utilizando listas más largas cuando procesa la consulta.

3.3.2.C Stemming

La tarea de este componente es la de agrupar palabras que provienen de un mismo tronco, agrupando “pez”, “pescados” y “pescar” podría ser un ejemplo válido. Reemplazando cada miembro del grupo por alguna de las palabras designadas (por ejemplo, la más corta), podemos incrementar probabilidad de que las palabras utilizadas en las consultas y los contenidos coincidan.

De hecho generalmente se producen pequeñas mejoras en la efectividad del ranking. El stemming puede ser ejecutado de tres maneras.

Stemming agresivo, puede causar problemas en las búsquedas, podría no ser apropiado, por ejemplo, recuperar los documentos de diferentes variedades de “peces” en respuesta a la consulta “pescar”. Algunas aplicaciones de búsqueda utilizan un método más conservador, como solamente identificar las palabras plurales que terminan con “s”, o no hacer ningún stemming al procesar los contenidos y enfocar en agregar variantes apropiadas a una consulta.

3.3.3 Extracción de información

Extracción de información es utilizada para identificar términos que son más complejos que simplemente palabras. Esto podría ser palabras en negrita o palabras en encabezados, pero en general puede requerir un procesamiento significativo. Extraer sintagmas nominales requiere un análisis sintáctico y un etiquetado gramatical complejo. Investigaciones en esta área han focalizado en técnicas para extraer algunos contenidos específicos, como puede ser

reconocimiento de entidades que pueden identificar en forma fiable así como nombres de personas, nombres de empresas, fechas y localizaciones.

3.3.3.A Clasificador

Este componente identifica clases de metadatos relacionados en documentos o en partes de documentos. La técnica de la clasificación asigna clases de etiquetas predefinidas a los documentos. Estas etiquetas representan categorías de tópicos como pueden ser “deportes”, “política” o “negocios”. Dos ejemplos importantes de otro tipo clasificación puede ser identificar un documento como un spam e identificar publicidad.

3.3.4. Creación del índice

3.3.4.A Documentar estadísticas

La tarea de este componente es la de simplemente reunir y guardar información estadística de palabras, características y documentos. Esta información es usada por el componente de ranking para computar las puntuaciones de los documentos. Los tipos de datos generalmente requeridos son, la cantidad de ocurrencias de los términos de índice (tanto palabras como características más complejas) en contenidos individuales, las posiciones en el contenido donde se ubican estas ocurrencias, etc. Los datos realmente requeridos son determinados por el modelo de recuperación y el algoritmo de rankeo. Las estadísticas de los documentos están almacenadas en tablas de búsqueda, que contienen estructuras diseñadas para una recuperación veloz.

3.3.4.B Ponderación

La ponderación de los términos indexados, reflejan la relativa importancia de las palabras en los documentos, y son utilizadas para el cómputo del ranking. La forma específica de una ponderación es determinada por el modelo de recuperación. El componente calcula una ponderación utilizando las estadísticas de los documentos y la almacena en las tablas de búsqueda. La ponderación puede ser calculada como parte del proceso de la consulta, y algunos tipos de ponderaciones requieren información de la búsqueda, pero haciendo muchos cálculos sobre el proceso de consulta inevitablemente la eficiencia será afectada.

Uno de los tipos más utilizados en los modelos de recuperación es conocido como tf.idf. Hay muchas variaciones de estas ponderaciones, pero todas están basadas en una combinación de la frecuencia o cantidad de ocurrencias de un término dentro de un documento (Term frencuency tf). Y la frecuencia o cantidad de ocurrencias del término en una colección entera

de documentos (inverse document frequiency 'idf'). Este caso es llamado inverso porque determina más ponderación a los términos que ocurren en pocos documentos. Una fórmula típica es idf log N /n, donde N es el total de documentos indexados por el motor de búsqueda y n es el número de documentos que contienen este término particular.

3.3.4.C Recambio

El recambio es el componente núcleo del proceso de indexación. La tarea consiste en cambiar el flujo de información del documento, por los índices invertidos. El desafío es hacer esto de forma eficiente, no solo para documentos grandes cuando los índices invertidos son recién creados, sino también cuando los índices son actualizados por los crawls. El formato de estos índices es diseñado para procesar consultas velozmente y depende en cierta forma del algoritmo de rankeo utilizado. Los índices son comprimidos para mejorar su eficiencia.

3.3.5. Interacción de usuarios

3.3.5.A Ingreso de consultas

Provee una interfaz y un analizador al lenguaje de consulta. Los lenguajes más simples, como los utilizados en la mayoría de las interfaces web, poseen algunos pocos operadores. Un operador es un comando en el lenguaje de consulta que es utilizado para indicar el texto que debe ser tratado de una manera especial. Un ejemplo de un operador puede ser la doble comilla, indicando que las palabras entre comillas deben aparecer como una frase en el documento, en vez de palabras individuales sin relación.

Una búsqueda típica web consiste en un grupo pequeño de palabras claves sin operadores. Una palabra clave es simplemente una palabra que es importante para especificar el tema de una consulta. Como los algoritmos de rankeo están diseñados para realizar consultas con palabras claves, las consultas largas generalmente no arrojan buenos resultados. Por ejemplo si buscamos "search engine" va a producir mejores resultados que si buscamos "'¿Cuáles son las implementaciones técnicas y las estructuras de datos de los motores de búsqueda ?" Uno de los objetivos de una ingeniería de búsqueda es la de otorgar buenos resultados a un rango de consultas, y mejores resultados a consultas más específicas.

Algunos lenguajes de consultas más complejos están disponibles, sobre todo para personas que quieren tener control sobre los resultados de búsqueda. En la misma manera que el lenguaje SQL (Elmasri & Navathe, 2006) el cual no está diseñado para un usuario final. Boolean query lenguajes tienen una historia larga en la recuperación de información. Los

operadores que incluyen pueden ser AND, OR y NOT y algunos operadores de proximidad que especifican que las palabras tienen que ocurrir con cierta distancia.

3.3.5.B Transformación de la consulta

La transformación de la consulta incluye un rango de técnicas que son diseñadas para mejorar la consulta inicial, antes y después de producir un ranking de documentos. El proceso es simple, incluye algunas de las técnicas antes utilizadas en la transformación del documento. Tokenizing, stopping y steamming son procesos aplicados al texto de la consulta para producir términos de índice comparables a los términos del documento.

Corrección ortográfica y la sugerencia de consulta son técnicas de transformación que producen una salida similar. En ambos casos, al usuario se le presentan alternativas a la consulta inicial que no poseen errores ortográficos o poseen descripciones más específicas a las necesidades de información. Estas técnicas aprovechan los logs almacenados por las aplicaciones web. Las técnicas de extensión de consulta también sugieren o agregan información adicional a la consulta, pero usualmente basados en análisis de las ocurrencias de los términos en los documentos.

3.3.5.C Resultados de salida

Este componente es el responsable de mostrar los resultados, esto puede incluir tareas como generar resúmenes de los documentos, resaltar las palabras claves que coinciden con la búsqueda, agrupar la salida para identificar grupos de documentos. Y buscar anuncios apropiados para agregar en los resultados de búsqueda.

3.3.6. Ranking

3.3.6.A Scoring

Calcula la puntuación para los documentos utilizando el algoritmo de ranking, basado en el modelo de recuperación. Las características y pesos utilizados en los algoritmos de clasificación que pueden haber sido derivados empíricamente (a través de pruebas y evaluaciones) tiene que estar relacionado al tema y a la relevancia del usuario, o el motor de búsqueda no funcionará. El cálculo básico para determinar la puntuación utilizado por muchos modelos puede ser así:

$$\sum_i q_i.d_i$$

La sumatoria es sobre todos los términos del vocabulario de la colección, qi es el peso del término de la consulta del término i, y di es el peso del documento. El peso del término depende del modelo particular de recuperación utilizado, pero generalmente es similar al de TFxIDF.

La puntuación de los documentos debe ser calculada muy rápidamente y comparada para determinar el orden de los documentos que son enviados al componente de salida.

3.4. ANATOMÍA DE LAS BÚSQUEDAS

Una búsqueda parece bastante sencilla, se escribe una búsqueda en una casilla, se envía la consulta, se muestra una pequeña oración mientras se buscan los resultados, se muestran unos resultados y uno puede seguir con su vida.

Existen diferentes algoritmos que procesan la consulta en algo que el software pueda entender y para determinar una priorización de los resultados. Esta priorización tiene diversas aristas.

3.4.1. Análisis de documentos y conectividad semántica

En el análisis de un documento, los motores de búsqueda tienen en cuenta los términos más importantes de un documento, el título, los metadatos, las etiquetas del título y el cuerpo del contenido. También tienden a medir automáticamente la calidad de un documento basado en un análisis.

Es necesario para los buscadores de hoy en día chequear también la conectividad semántica de los documentos indexados. La conectividad semántica refiere a conjuntos de palabras o frases que son comúnmente asociadas a otras; la palabra “aloha”, será asociada con Hawái y no con Florida sin ser ni siquiera un sinónimo. Los motores de búsqueda construyen sus propios tesauros y diccionarios para determinar qué términos y tópicos están relacionados. Escaneando sus propias bases de contenidos en la web y aplicando ciertas ecuaciones pueden conectar términos y empezar a “entender” las páginas web y los sitios de una forma más similar a como lo hacen los humanos.

Los motores de búsqueda utilizan la teoría fuzzy, una rama de la lógica creada por el DR Lotfi Zadeh, para descubrir conexiones semánticas entre dos palabras. Aunque este proceso pueda parecer complejo, las fundaciones son muy simples. Los motores de búsqueda deben

confiar en la lógica de una máquina (verdadero/falso, sí/no, etc.), la cual tiene muchas ventajas sobre los humanos, pero no posee la forma de pensar de los humanos, conceptos que son simples para los humanos son complejos para que una máquina los entienda. Por ejemplo, la banana y la naranja son frutas, pero las naranjas y las bananas no son ambas redondas, esto para un humano es intuitivo.

Para que una máquina entienda este concepto y pueda responder cosas similares, las conexiones semánticas son la clave. El conocimiento masivo que existe en la web puede ser analizado para artificialmente crear las conexiones que los humanos han realizado. Una máquina podría saber que una manzana es redonda y una banana no, por el simple hecho de analizar miles de resultados donde la palabra manzana y redonda tiene muchas ocurrencias mientras que banana y redonda no.

4. DEFINIENDO LA INFORMACIÓN SOBRE LA INFORMACIÓN

Un ambiente interactivo de información -como lo es un sitio web- es una colección de sistemas interconectados que tienen dependencias complejas. Un simple enlace puede determinar en simultáneo el uso de etiquetado, organización, navegación y hasta el uso de un sistema de búsqueda. Es muy útil estudiar cada uno de estos puntos de manera independiente, pero es crucial su interacción.

Los metadatos y los vocabularios controlados nos muestran claramente la red que relaciona los sistemas. En muchos productos de metadatos, los vocabularios controlados se han transformado en el pegamento que mantiene a los productos juntos.

4.1. METADATOS

Dijimos anteriormente que los metadatos son los datos que proveen información sobre uno o más aspectos del dato, por ejemplo:

- Método de creación del dato.
- Propósito del dato.
- Tiempo y fecha de creación.
- Creador o autor del dato.
- Locación de la red donde el dato fue creado.
- Estándares usados por el dato.

A modo de explicación, una imagen digital puede incluir metadatos de autoría de la imagen, fecha la resolución e incluso datos sobre la exposición y velocidad al momento de sacar la foto.

Las etiquetas de metadatos son utilizadas para describir documentos, páginas, imágenes, software, videos y archivos de sonido con el fin de mejorar la navegación y la recuperación. Los atributos <meta> del lenguaje HTML, utilizados por varios sitios web, proveen un ejemplo simple y concreto. Los autores pueden describir libremente palabras que describen el contenido, palabras que no son mostradas en la web, pero están disponibles para los motores de búsqueda.

```
    <meta name=”keywords” content=”information architecture, content
management”>
```

Se pueden utilizar los metadatos de formas más sofisticadas, si utilizamos manejadores de contenidos y vocabularios controlados. Estas herramientas tienen un modelo basado en

metadatos, los cuales soportan autorías de contenidos de forma distribuidas y navegaciones complejas. Este modelo implica un cambio en la forma de pensar los sistemas de información. En vez de preguntarse dónde pondremos el contenido nos preguntamos cómo describimos este contenido y la herramienta es la que se encarga del resto.

Los modelos de metadatos que vamos a tratar en esta sección son los más influyentes y populares que hoy en día existen en el mercado. La principal característica de los modelos elegidos es que fueron ampliamente adoptados por la comunidad en general; desde CMS, sistemas informáticos de todo tipo y hasta buscadores de datos en línea.

4.2. DUBLIN CORE

Es un modelo de datos elaborado y auspiciado por la Dublin Core Metadata Initiative, una organización dedicada a fomentar y a promover el desarrollo de los vocabularios especializados de metadatos para describir recursos. Es el ejemplo más significativo y conocido dentro de los campos de la documentación y bibliografía. Dublin Core es producto de un esfuerzo internacional e interdisciplinario con una vida muy intensa y el más influyente en relación con el desarrollo de la teoría del uso de los metadatos.

Muchas instituciones y personajes han participado en el desarrollo de este formato y su desarrollo se dio en el mismo momento que XML y DRF. Dublin Core tiene como objetivo, definir un conjunto básico de atributos que sirvan para describir todos los recursos existentes en la red ayudando así a la recuperación de datos de la web.

4.2.1 Elementos del Dublin Core

Para cumplir con su objetivo, el Dublin Core define un conjunto de quince elementos los cuales pueden modificarse y ampliarse permitiendo a los autores de las páginas Web codificar sus documentos en el momento de generarlos. Estos se agrupan en tres categorías: contenido, propiedad intelectual e instanciación.

4.2.1.A Contenido

- **Título:** el nombre dado a un recurso, habitualmente por el autor.
    - Etiqueta: DC.Title
- **Claves:** los temas del recurso. Típicamente, Subject expresará las claves o frases que describen el título o el contenido del recurso. Se fomentará el uso de vocabularios controlados y de sistemas de clasificación formales.
    - Etiqueta: DC.Subject

- **Descripción:** una descripción textual del recurso. Puede ser un resumen en el caso de un documento o una descripción del contenido en el caso de un documento visual.
    - Etiqueta: DC.Description
- **Fuente:** secuencia de caracteres usados para identificar unívocamente un trabajo a partir del cual proviene el recurso actual.
    - Etiqueta: DC.Source
- **Tipo del Recurso:** la categoría del recurso. Por ejemplo, página personal, romance, poema, diccionario, etc.
    - Etiqueta: DC.Type
- **Relación:** es un identificador de un segundo recurso y su relación con el recurso actual. Este elemento permite enlazar los recursos relacionados y las descripciones de los recursos.
    - Etiqueta: DC.Relation
- **Cobertura:** es la característica de cobertura espacial y/o temporal del contenido intelectual del recurso.
    - La cobertura espacial se refiere a una región física, utilizando por ejemplo coordenadas.
    - La cobertura temporal se refiere al contenido del recurso, no a cuándo fue creado (que ya lo encontramos en el elemento Date).
    - Etiqueta: DC.Coverage

4.2.1.B Propiedad Intelectual

- **Autor o Creador:** la persona u organización responsable de la creación del contenido intelectual del recurso. Por ejemplo, los autores en el caso de documentos escritos; artistas, fotógrafos e ilustradores en el caso de recursos visuales.

Etiqueta: DC.Creator

- **Editor:** la entidad responsable de hacer que el recurso se encuentre disponible en la red en su formato actual.
    - Etiqueta: DC.Publisher
- **Otros Colaboradores:** una persona u organización que haya tenido una contribución intelectual significativa, pero que esta sea secundaria en comparación con las de las personas u organizaciones especificadas en el elemento Creator. (por ejemplo: editor, ilustrador y traductor).
    - Etiqueta: DC.Contributor

- **Derechos:** son una referencia (por ejemplo, una URL) para una nota sobre derechos de autor, para un servicio de gestión de derechos o para un servicio que dará información sobre términos y condiciones de acceso a un recurso.
    - Etiqueta: DC.Rights

4.2.1.C Instanciación

- **Fecha:** una fecha en la cual el recurso se puso a disposición del usuario en su forma actual. Esta fecha no se tiene que confundir con la que pertenece al elemento Coverage, que estaría asociada con el recurso en la medida que el contenido intelectual está de alguna manera relacionado con aquella fecha.
    - Etiqueta: DC.Date
- **Formato:** es el formato de datos de un recurso, usado para identificar el software y, posiblemente, el hardware que se necesitaría para mostrar el recurso.
    - Etiqueta: DC.Format
- **Identificador del Recurso:** secuencia de caracteres utilizados para identificar unívocamente un recurso. Ejemplos para recursos en línea pueden ser URLs y URNs. Para otros recursos pueden ser usados otros formatos de identificadores, como por ejemplo ISBN ("International Standard Book Number").
    - Etiqueta: DC.Identifier
- **Lengua:** lengua/s del contenido intelectual del recurso.
    - Etiqueta: DC.Language

4.1.1.D Ejemplo

Los elementos Dublin Core puede ser usado como atributo de un HTML tag dentro de la cabecera "<head>".

En lo siguiente daremos un ejemplo de uso de Dublin Core dentro de la cabecera de un documento HTML. Esta es una de las formas en las que un crawler podría llegar a las etiquetas semánticas del contenido.

```
<html>
 <head>
   <title> A Dirge </title>
   <link rel="schema.DC" href="http://purl.org/DC/elements/1.0/">
   <meta name="DC.Title" content="A Dirge">
   <meta name="DC.Creator" content="Shelley, Percy Bysshe">
   <meta name="DC.Type" content="poem">
   <meta name="DC.Date" content="1820">
   <meta name="DC.Format" content="text/html">
   <meta name="DC.Language" content="en">
 </head>
```

```
  [...]
</html>
```

4.3. SCHEMA.ORG

El sitio de schema.org define el modelo como “una actividad colaborativa en comunidad con la misión de crear, mantener y promover esquemas para los datos estructurados en la internet, páginas web, correos electrónicos y más”.

El vocabulario Schema.org puede ser usado con diferentes codificaciones usando formas más modernas como RDFa, Microdata o incluso JSON-LD. Schema.org cubre entidades, relaciones entre entidades y acciones que pueden ser fácilmente extensibles a través de un modelo de extensión bien documentado. Más de 10 millones de sitios lo usan y los buscadores más grandes lo soportan, entre ellos, Google, Bing, Pinteres, Yandex, etc.

Schema.org está sponsoreado por Google, Microsoft, Yahoo y Yandex y sus vocabularios están desarrollados por un proceso abierto de la comunidad.

4.3.1. Esquemas y herencia de atributos

Schema.org define más de 171 tipos y más de 830 propiedades. Los tipos pueden usar el concepto de “Tipo padre” / “Tipo hijo” para representar la herencia de atributos. Un ejemplo de esto es que el tipo “Book” desciende directamente del tipo “CreativeWork”, así como también lo hacen “Movie”, “TVSerie” y “Recipe”. Esta herencia funciona heredando los atributos del padre, si el tipo “CreativeWork” define sus atributos como “about”, “author”, “comment” y “creator”, el tipo “Book” también tendrá estos atributos y le agregará, por ejemplo, “isbn”, “ilustrator” y “bookFormat”.

4.3.2. Contenido explícitamente vinculado

Si, por ejemplo, tenemos un sitio institucional como el INTA en donde existen personas que publican documentos, al entrar a la página de una persona veremos lo siguiente:

http://inta.gob.ar/personas/villanova.ingrid

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Person",
  "name": "Ingrid VILLANOVA",
  "content:encoded": "",
  "givenName": "Ingrid",
  "url": "http://inta.gob.ar/personas/villanova.ingrid",
  "familyName": "VILLANOVA",
  "description": "",
  "image": "http://inta.gob.ar/sites/default/files/villanova.ingrid.jpg"
}
```

```
</script>
```

De ahora en adelante, un buscador cuando entre a la página de la persona "Ingrid Villanova" sabrá que está página **es** en definitiva una persona. Si luego quisiéramos entrar a un trabajo publicado por esta persona vamos a ver algo parecido a lo siguiente:

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "CreativeWork",
  "title": "Relevamiento del comercio minorista de la floricultura en la
Argentina
",
  "content:encoded": "",
  "url":"http://inta.gob.ar/documentos/ relevamiento-del-comercio-minorista-de-la-
floricultura-en-la-argentina",
  "author": ["inta.gob.ar/personas/villanova.ingrid",
"http://inta.gob.ar/personas/pizarro.maria",
"http://inta.gob.ar/personas/barrionuevo.nestor",
"http://inta.gob.ar/personas/morisigue.daniel"  ]
}
</script>
```

De ahora en adelante, "`Relevamiento del comercio minorista de la floricultura en la Argentina`", será identificado por un buscador como un trabajo creativo cuyo autor, en vez de ser un nombre como lo sería usando Dublin Core, es una URI a un tipo "Persona", en este caso son varios autores, entre los que está Villanova, en donde se definen completamente sus datos semánticos construyendo así un conjunto de "nodos" en un grafo de conocimiento interconectado explícitamente. A medida que el sitio adopte el uso de schema.org y valores de atributos como URI (Schema.org también podría aceptar cadenas de caracteres como "Ingrid Villanova") el grafo de conocimiento crece, tanto en nodos como en vínculos ya que, hasta las palabras clave, tienen una clase que las representa en Schema.org.

4.3.3. Visión de Google con el esquema de datos estructurados de Schema.org

Al usar un esquema de datos estructurados como schema.org los buscadores tienen otra visión del sitio; saben qué es cada página del sitio y, además, entienden las propiedades esenciales de cada una de ellas. Una vez incorporado schema.org, el crawler de Google recorre el sitio, analiza la información y sabe qué es cada cosa. Desde la consola de búsquedas de Google, ahora sabe que en inta.gob.ar hay "personas", "trabajos de creación intelectual" (Creative Work), "libros", "cosas" (clase genérica que engloba páginas no descritas específicamente)

Data Type	Source	Pages	Items
Person	Markup: schema.org	4,286	4,324
Document	Markup: xmlns.com	4,405	29,416
Item	Markup: rdfs.org	4,405	29,416
Image	Markup: xmlns.com	24,569	51,195
CreativeWork	Markup: schema.org	13,486	25,981
BreadcrumbList	Markup: schema.org	4,284	4,284
Book	Markup: schema.org	486	486
Thing	Markup: schema.org	12	12

Además, también, el crawler halló un listado de "Breadcrumb Lists" que describen el acceso por niveles, de forma jerárquica, a páginas del sitio. Un ejemplo de esto se puede ver, ahora, al buscar una persona del INTA:

Vemos que debajo del título "Diego Javier CELDRAN…", se incluye el índice jerárquico "inta.gob.ar > Personas", accesible en todos sus niveles, que lleva primero a "inta.gob.ar" y luego a un listado de personas. Antes de agregado el breadcrumb list en la sección de personas del sitio, solamente aparecía la url de la página. El código que describe el breadcrumb list para Diego Javier Celdrán es de la siguiente forma:

```
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "BreadcrumbList",
  "itemListElement": [{
    "@type": "ListItem",
    "position": 1,
    "item": {
      "@id": "http://inta.gob.ar",
      "name": "INTA"
    }
  },{
    "@type": "ListItem",
    "position": 2,
    "item": {
      "@id": "http://inta.gob.ar/busqueda/tipo-de-contenido/persona/p/sort_by/search_api_aggregation_1
/sort_order/ASC?buscar=",
      "name": "Personas"
    }
  },{
    "@type": "ListItem",
    "position": 3,
    "item": {
      "@id": "http://inta.gob.ar/personas/celdran.diego",
      "name": "Diego Javier CELDRAN"
    }
  }]
}
</script>
```

4.4 VOCABULARIO CONTROLADO

El vocabulario controlado aparece en diferentes formas y tamaños, definido de forma simple, un vocabulario controlado es un subconjunto definido de lenguaje natural.

El objetivo de un vocabulario controlado es lograr la consistencia en la descripción de contenidos de los documentos y facilitar su recuperación.

Para esto se tienen cinco finalidades:

1. Traducción. Proporcionar medios para convertir el lenguaje natural de los autores, indizadores y usuarios en vocabularios que pueda ser utilizado para indizar y recuperar.
2. Consistencia. Promover uniformidad en la forma y asignación de los términos.
3. Indicación de relaciones. Indicar relaciones semánticas entre los términos.
4. Etiquetado y visualización. Proporcionar jerarquías claras y consistentes en los sistemas de navegación para ayudar a los usuarios a localizar objetos de contenido deseados.
5. Recuperación. Servir de ayuda en la búsqueda para localizar objetos de contenido.

Se distinguen cuatro tipos de vocabularios controlados, determinados por su estructura compleja creciente: listas, anillos de sinónimos, taxonomías, y tesauros.

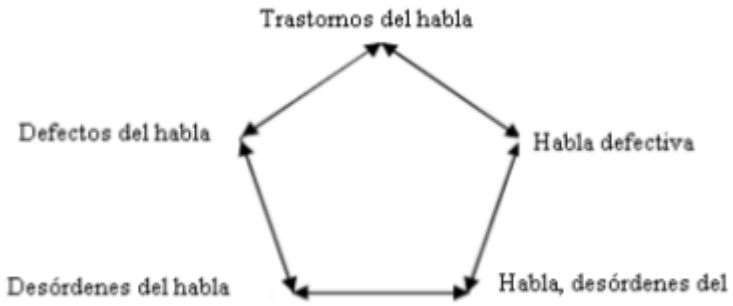
4.4.1 Lista

Es un grupo simple de términos preferentes sin estructura y suelen presentarse en orden alfabético u otra secuencia lógica.

4.4.2 Anillo de sinónimos

Conecta un conjunto de palabras que son definidas como equivalentes al propósito de su recuperación. Estas palabras no siempre son sinónimos reales. Es un tipo de vocabulario controlado que no puede ser utilizado en el proceso de indexación, solo en el de recuperación. El uso del anillo de sinónimos asegura que un concepto pueda ser descrito por múltiples sinónimos o términos equivalentes que serán utilizados si uno de los términos es usado en la búsqueda. Se utilizan generalmente en las interfaces de sistemas de información, y proporcionan acceso al contenido que representa en el lenguaje natural, no controlado.

Ejemplo:



4.4.3 Taxonomía

Se define como una lista de términos preferentes con estructura jerárquica. Estos pueden tomar diferentes formas y servir a diferentes propósitos:

- Interfaz, una jerarquía de navegación que es visible y parte integral de la interfaz del usuario.
- Backend, como herramienta utilizada para que los autores y los indexadores organicen y etiqueten contenidos.

Por ejemplo, netflix utiliza un esquema de clasificación sofisticado, para que sus usuarios puedan encontrar películas que puedan llegar a disfrutar, además de los obvios (Drama, comedia, etc.). Las películas son categorizadas en cientos de micro-géneros incluyendo

algunos como “basado en hechos reales”, “Protagonizada por una mujer” y algunos más específicos aún como “Drama gangster de suspenso”.

4.5 TESAURO

La palabra tesauro, proviene del griego tesauros, el cual significa tesoro, almacén de algo valioso.

La UNESCO, en 1976 propone dos definiciones distintas a partir de dos puntos de vista.

Según su función, “son un instrumento de control terminológico, usado para trasladar, desde un lenguaje natural de los documentos, los descriptores, a un sistema lingüístico”. Y según su estructura, “Son vocabularios controlados y dinámicos de términos relacionados semántica y genéricamente, que cubren un dominio específico del conocimiento”.

Podemos definirlos como vocabularios controlados, de estructura combinatoria, definidos a priori, es decir, fijados con anterioridad, compuestos por términos que reflejan conceptos que se relacionan entre sí semántica, jerárquica y asociativamente, que tienen como finalidad el control de sinónimos, y se utilizan para describir de manera unívoca, el lenguaje natural al documental, el contenido de los documentos, para su posterior recuperación en un sistema documental dado, con el fin de satisfacer las necesidades de información,

4.5.1 Funciones del tesauro

1. Normalizar el vocabulario. Es decir, controlar los accidentes que puede generar el vocabulario, que son la sinonimia, la polisemia, así como el género y el número de las expresiones.
2. Inducir al usuario hacia posibles alternativas, a términos en los que el usuario no habia pensado, gracias a una serie de referencias cruzadas que indican las relaciones asociativas, jerárquicas o preferenciales que se pueden dar entre la terminología que lo compone.
3. Representar los conceptos presentes en los documentos. Esta función es compartida con el resto de lenguajes documentales.

4.5.2. Objetivos del tesauro

1. Facilitar la representación consistente de las materias por parte de indizadores y usuarios que recuperan, evitando la dispersión de los elementos relacionados. Esto se consigue con el control (agrupación) de los sinónimos y cuasi sinónimos y la distinción de los homógrafos.

2. Facilitar la realización de una búsqueda amplia sobre una materia enlazando los términos con relaciones paradigmáticas y sintagmáticas.

4.5.3 Tipos de tesauros

Si se decide construir un tesauro para un sitio, se debe elegir entre tres tipos de tesauros clásico, de indexación o de búsqueda. la decisión tiene que estar basada en cómo se pretende utilizar el tesauro lo cual tendrá mucha implicación en el diseño.

4.5.3.A. Tesauro clásico

Es utilizado en el punto de la indexación y en el punto de la búsqueda. Los indexadores lo utilizan para mapear los términos variantes a los preferidos, cuando hacen la indexación del documento. Los buscadores usan el tesauro para recuperar. Las consultas son emparejadas contra el vocabulario rico de los tesauros, habilitando sinónimos, navegación de jerarquías y links asociados.

4.5.3.B. Tesauros de indexación

Se utiliza el vocabulario controlado para la indexación, pero no está habilitado para la búsqueda, lo cual es una gran debilidad. Hay razones para mencionar que es mejor un tesauro indexado que nada:

- Estructura el proceso de indexación, promoviendo consistencia y eficiencia.
- Permite construir índices navegables de términos preferidos, habilitando a los usuarios a encontrar todos los documentos de un tipo particular, a través de un punto de acceso.

4.5.3.C. Tesauros de búsqueda

A veces un tesauro clásico no es práctico, quizás se trata con contenidos que no son propios o novedades que cambian todo el tiempo. Cuando un usuario ingresa un término en el motor de búsqueda, el tesauro de búsqueda puede mapear el término antes de ejecutar la búsqueda full-text.

4.5.4. Estándares y normas

El primer especialista que propuso normas para la construcción de encabezamientos alfabéticos fue el norteamericano Charles Ammi Cutter en 1876: Rules for a Dictionary Catalogue.

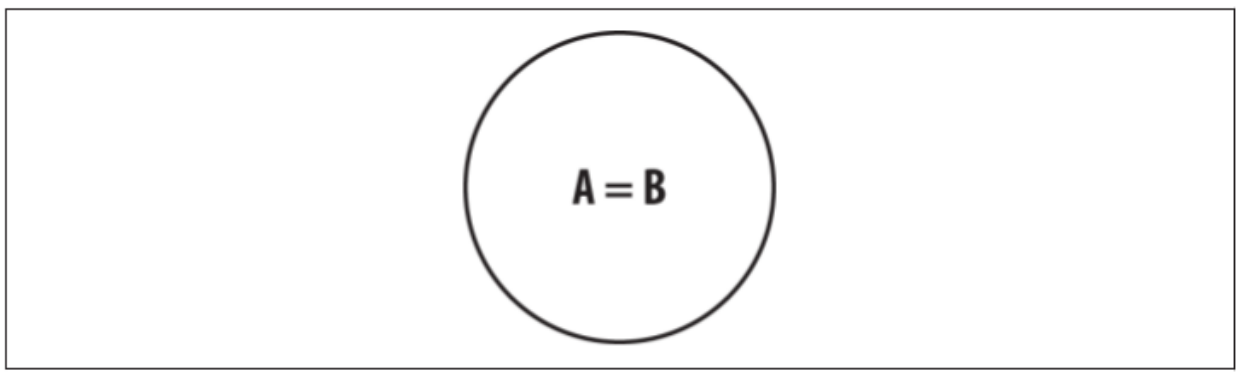
Las normas nacionales americanas para la construcción de tesauros, la ANSI Z39.19 (National Information Standards Organization, 2005) publicadas por el American National Standards Institute aparecieron por primera vez en 1974. Estas normas fueron revisadas a finales de los años 70, y en 1980 apareció la segunda edición. Con la posterior edición de 1993 se amplió de modo significativo el ámbito de aplicación de la norma con la indicación format and management. Ésta ofrece más detalles respecto al control terminológico y las presentaciones en pantalla. Esta tercera edición es análoga al estándar internacional ISO 2788 y la norma británica BS 5723. Una nueva revisión fue aprobada en 1998, a pesar de que fue muy criticada por no recoger la evolución de los sistemas informáticos (Franchini, 2005). Por ejemplo, Oracle dice del estándar: "El título estándar de tesauro es engañoso, en la industria informática se considera un estándar como una especificación de un comportamiento o una interface. Estos estándares no especifican nada. Si estás buscando una interfaz de un tesauro, o un formato de un archivo tesauro, no lo vas a encontrar aquí. En vez de eso encontraras una guía para los compiladores de tesauros, siendo un compilador un humano y no un programa." Estamos en el medio de una transición de un tesauro tradicional a un nuevo paradigma que incluya los contenidos informáticos.

4.5.4.A. Relaciones semánticas

Lo que diferencia al tesauro de los vocabularios controlados es su riqueza en relaciones semánticas. Exploremos las diferentes posibilidades.

4.5.4.B. Equivalencia

La relación de equivalencia es utilizada para conectar los términos preferidos y sus variantes. Si bien se puede confundir con el manejo de sinónimos, es importante entender que equivalencia es más amplio que el sinónimo.



Relación de equivalencia.

El objetivo es el de agrupar términos definidos como “equivalentes al propósito del retorno”, esto incluye sinónimos, cuasi-sinónimos, acrónimos, abreviaciones, variantes léxicas y faltas de ortografía.

La relación de equivalencia entre términos preferentes y no preferentes se indica mediante los símbolos siguientes:

- "UF" (used for = usado para), colocado delante del término no preferente (no descriptor); Ejemplo: parque tecnológico UF parque científico
- "USE" (= utilícese), colocado delante del término preferente (descriptor). Ejemplo: parque científico USE parque tecnológico

Cuando varios términos representan el mismo concepto, la relación de equivalencia indica el término de indización que debe utilizarse. En este ejemplo, "parque tecnológico" (perteneciente al micro tesauro "6806 Política y Estructuras Industriales") es el término preferente que debe utilizarse para indizar conceptos tales como "parque científico".

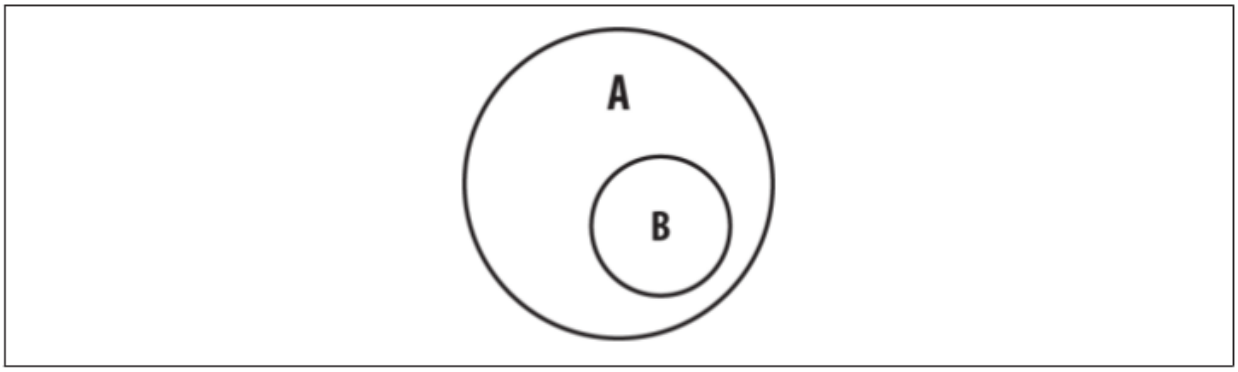
En EuroVoc, la reciprocidad entre términos preferentes y no preferentes se aplica a los:

- Sinónimos, ejemplo: tercera edad USE persona de edad avanzada
- Cuasi sinónimos o términos considerados diferentes en el uso corriente pero tratados como sinónimos a efectos de indización. Ejemplo: sacrificio de animales USE matanza de ganado.
- Antónimos o palabras de significado opuesto. Ejemplo: independencia tecnológica USE dependencia tecnológica.

En el caso de los términos marginales a los campos temáticos del tesauro se remite al usuario al término genérico. Los nombres de una clase y de sus miembros se tratan como equivalentes. El término genérico de la clase se considera término preferente. Ejemplo: higo, membrillo, pera USE fruto de pepita.

4.5.4.C. Jerárquica

Esta relación está basada en niveles jerárquicos de superioridad o subordinación entre conceptos. El concepto superior constituye una clase, mientras que los conceptos subordinados representan elementos o partes de esa clase.



Relación Jerárquica

Esta relación se indica mediante los siguientes símbolos:

- BT (broader term = término genérico), situado entre un concepto específico y un concepto genérico y acompañado de una cifra que indica el número de niveles jerárquicos que hay entre el término específico y cada uno de los términos genéricos que le corresponden. *Ejemplo:*
    - norma
        - BT1 normalización
            - BT2 reglamentación técnica.

    Los conceptos que carecen de término genérico se denominan **términos cabecera** (**top terms**).
- NT (narrower term = término específico), situado entre un concepto genérico y un concepto específico y acompañado de una cifra que indica el número de niveles jerárquicos que hay entre el término genérico y cada uno de los términos específicos que le corresponden. *Ejemplo:*
    - normalización

- NT1 armonización de normas
- NT1 homologación
    - NT2 certificación comunitaria
- NT1 marca de conformidad CE
- NT1 norma
    - NT2 norma de calidad
    - NT2 norma de producción
    - NT2 norma de seguridad
    - NT2 norma técnica
- NT1 norma internacional
    - NT2 norma europea

La relación jerárquica permite al usuario de un sistema documental adaptar el nivel de especificidad del mismo navegando por la jerarquía. Puede ampliar o precisar su pregunta seleccionando conceptos que tengan un sentido más amplio (por ejemplo, "normalización", o "reglamentación técnica") o más estricto ("norma de producción", "norma de calidad" o "norma técnica").

La relación jerárquica precisa el sentido del concepto en su contexto. La acepción del concepto prensa, por ejemplo, queda determinada por su subordinación al concepto medio de comunicación de masas.

En EuroVoc la relación jerárquica se establece a partir de situaciones lógicas.

a) Relación genérica, Identifica el vínculo entre una clase (término genérico o término cabecera) y sus elementos (términos específicos).

Ejemplo:

- área protegida

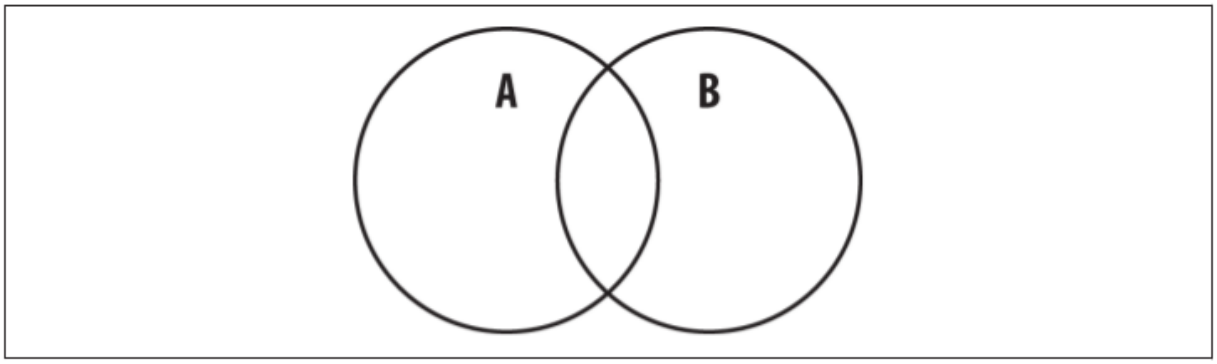
- NT1 parque nacional
- NT1 reserva natural

b) Relación partitiva, corresponde a algunas situaciones en las cuales el nombre de la parte indica el nombre del todo con independencia del contexto. El nombre del todo representa el término genérico y el nombre de la parte representa el término o términos específicos del concepto. En EuroVoc se aplica principalmente a las siguientes clases de términos:

- *Lugares geográficos, Ejemplo:*
  - océano
    - NT1 Océano Antártico
    - NT1 Océano Ártico
    - NT1 Océano Atlántico
    - NT1 Océano Índico
    - NT1 Océano Pacífico
- *Disciplinas o ámbitos de conocimiento, Ejemplo:*
  - química
    - NT1 bioquímica
    - NT1 electroquímica
    - NT1 fotoquímica
    - NT1 química analítica
- *Estructuras sociales jerarquizadas, Ejemplo:*
  - Mesa del Parlamento
    - NT1 cuestor
    - NT1 presidente del Parlamento
    - NT1 vicepresidente del Parlamento

4.5.4.D. Asociativa

La relación asociativa es una relación entre dos conceptos que no pertenecen a la misma estructura jerárquica, aunque sean semántica o contextualmente similares. Esta relación debe explicitar, ya que sugiere al indizador la utilización de otros términos de indización de significado similar o próximo y que pueden servir para la indización o la búsqueda.



La relación asociativa se indica mediante el símbolo **RT (related term = término relacionado**), situado entre dos conceptos asociados, y es recíproca.
Existen diferentes subtipos de relaciones asociativas.
Ejemplos:

Subtipo de relación	Ejemplo
Campo de estudio o Objeto de estudio	Cardiología RT Corazón
Proceso y su agente	Control de termitas RT Pesticidas
Conceptos y sus propiedades	Veneno RT Toxicidad
Acción y el producto de una acción	Comer RT Indigestión
Conceptos relacionados por una dependencia causal	Fiesta RT Fin de año

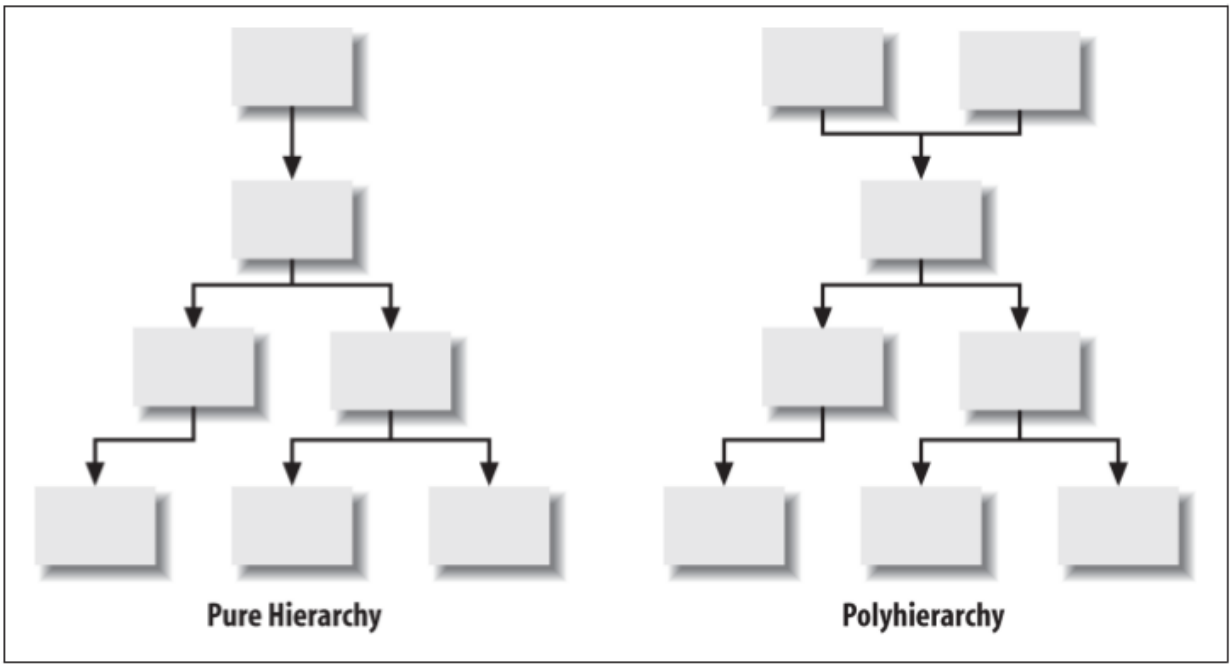
En el comercio online, la relación asociativa provee una excelente forma para relacionar productos y servicios para los usuarios. La relación asociativa permite lo que el marketing llama “venta cruzada”, permitiendo que un sitio de ventas muestre un anuncio que diga “Lindas zapatillas, irían bien con esta remera” . Si estas relaciones están bien contempladas, no sólo mejoran la experiencia del usuario, sino que mejoran las ventas.

4.5.4.E. Jerarquía múltiple

En una jerarquía estricta, cada término aparece en un lugar y solamente en uno. Esta fue la idea original de la taxonomía biológica. Cada especie debería encajar solamente en una rama del árbol de la vida.

Sin embargo, las cosas no fueron de acuerdo al plan. De hecho los biólogos estuvieron discutiendo por décadas el lugar correcto de cada especie. Algunas especies tienen la particularidad de mostrar características de múltiples categorías.

Si uno es purista puede defender la idea de una jerarquía estricta en tu sitio web pero, si uno es pragmático, puede agregar algunos niveles de multi-jerarquía, permitiendo que algunos términos estén listados en múltiples categorías.



4.6 ONTOLOGÍAS

Si los metadatos sirven para la estructuración del contenido, tanto los tesauros como las ontologías, hacen posible una semántica para construirlos. Una ontología es una especificación de una conceptualización, esto es, un marco común o una estructura conceptual sistematizada y de consenso no sólo para almacenar la información, sino también para poder buscarla y recuperarla. Una ontología define los términos y las relaciones básicas para la compresión de un área del conocimiento, así como las reglas para poder combinar los términos para definir las extensiones de este tipo de vocabulario controlados.

Se trata de convertir la información en conocimiento mediante unas estructuras de conocimiento formalizadas (las ontologías) que referencian los datos, por medio metadatos,

bajo un esquema común normalizado sobre algún dominio del conocimiento. Los metadatos no sólo especificarán el esquema de datos que debe aparecer en cada instancia, sino que también podrán contener información adicional de cómo hacer deducciones sobre ellos, es decir, cómo establecer axiomas que podrán, a su vez, aplicarse en los diferentes dominios que trate el conocimiento almacenado. De esta forma, los buscadores podrán obtener información al compartir los mismos esquemas de anotaciones web y los agentes de software no sólo encontrarán la información precisa, sino que podrán realizar inferencias de forma automática buscando información relacionada con la que se encuentra situada en las páginas web y con los requerimientos de las consultas realizadas por los usuarios. Además, los productores de páginas y servicios web podrán intercambiar sus datos siguiendo estos esquemas comunes consensuados e, incluso, podrán re-utilizarlos.

El término ontología se ha empleado desde hace muchos siglos en el campo de la filosofía y del conocimiento y hace ya varias décadas cobró especial relevancia en el campo de la biblioteconomía y la documentación. Hoy ha sufrido un nuevo impulso debido al desarrollo de la Web Semántica donde prima la idea de transformar la red no sólo en un espacio de información, sino también en un espacio de conocimiento.

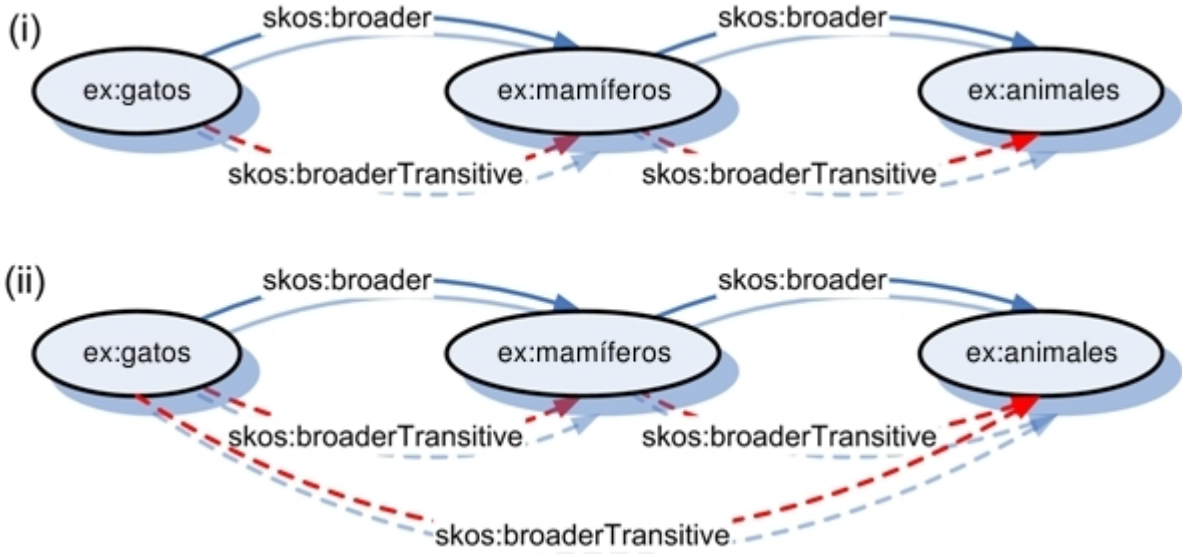
Tanto los tesauros como las ontologías son herramientas que sirven para estructurar conceptualmente determinados ámbitos del conocimiento por medio de vocabularios controlados. La diferencia entre los tesauros y las ontologías radica en la complejidad ya que estas últimas introducen un mayor nivel de profundización semántica y proporcionan una descripción lógica y formal que puede ser interpretada tanto por las personas, como por las máquinas, mientras que los tesauros sólo pueden ser interpretados por humanos. Las ontologías permiten, además, la interoperabilidad entre sistemas distintos.

4.7 SKOS

SKOS (siglas de Simple Knowledge Organization System) es una iniciativa del W3C en forma de aplicación de RDF que proporciona un modelo para representar la estructura básica y el contenido de esquemas conceptuales como listas encabezamientos de materia, taxonomías, esquemas de clasificación, tesauros y cualquier tipo de vocabulario controlado.

En SKOS los conceptos se identifican con referencias URI. Estos conceptos pueden etiquetarse en cadenas de texto en uno o varios idiomas, documentarse y estructurarse a través de relaciones semánticas de diversa tipología. Este modelo permite mapear conceptos de diferentes esquemas, así como definir colecciones ordenadas y agrupaciones de conceptos.

También permite establecer relaciones entre las etiquetas asociadas a los conceptos. El uso de RDF en el desarrollo de SKOS permite obtener documentos en un formato que permita su lectura por parte de aplicaciones informáticas, así como su intercambio y su publicación en la Web. SKOS se ha diseñado para crear nuevos sistemas de organización o migrar los ya existentes adaptándose a su uso en la Web Semántica de forma fácil y rápida. Proporciona un vocabulario muy sencillo y un modelo intuitivo que puede ser utilizado conjuntamente con OWL o de forma independiente. Por todo ello, SKOS se considera como un paso intermedio, un puente entre el caos resultante del bajo nivel de estructuración de la Web actual y el riguroso formalismo descriptivo de las ontologías definidas con OWL. (SKOS, s.f.)



El modelo de datos SKOS es en realidad una ontología definida con OWL Full. Obviamente, al estar basado en RDF, SKOS estructura los datos en forma de tripletas que pueden ser codificadas en cualquier sintaxis válida para RDF. SKOS puede ser utilizado conjuntamente con OWL para expresar formalmente estructuras de conocimiento sobre un dominio concreto ya que SKOS no puede realizar esta función al no tratarse de un lenguaje para la representación de conocimiento formal. El conocimiento descrito de manera explícita como una ontología formal se expresa como un conjunto de axiomas y hechos. Pero un tesauro o cualquier tipo de esquema de clasificación no incluye este tipo de afirmaciones, sino que identifica y describe (con el lenguaje natural o expresiones no formales) ideas o significados a los que nos referimos como conceptos. Estos conceptos pueden organizarse en estructuras que carecen de una semántica formal y que no pueden considerarse como axiomas o hechos. Es decir, un tesauro únicamente proporciona un mapa intuitivo de cómo están organizados los temas dentro de procesos de clasificación y búsqueda de objetos (generalmente documentos) relevantes a un dominio específico. Para convertir un tesauro o esquema de clasificación en

conocimiento formal, debe transformarse en una ontología, un proceso que resulta muy costoso.

En efecto, transformar la estructura de un tesauro en una ontología OWL conlleva un gran esfuerzo ya que una ontología no proporciona un modelo de datos que se pueda aplicar fácilmente. Esto sucede porque los tesauros se han desarrollado sin una semántica formal, fundamentalmente como herramientas que ayudan en la navegación o en la recuperación de información. No obstante, resulta factible aplicar OWL para construir un modelo de datos (en este caso concreto SKOS) que sea apropiado al nivel de formalización exigido por un tesauro. De esta forma, los conceptos de un tesauro se modelan como entidades en el modelo de datos SKOS y las relaciones entre conceptos como hechos sobre dichas entidades. Los elementos del modelo SKOS son esencialmente clases y propiedades. La estructura e integridad del modelo de datos están definidas por las características lógicas y por las relaciones entre dichas clases y propiedades. Para SKOS, un sistema de organización del conocimiento se expresa en términos de conceptos que se estructuran en relaciones para conformar esquemas de conceptos.

Tanto los conceptos como los esquemas de conceptos se identifican mediante URIs. Los conceptos pueden ser etiquetados en cualquier idioma. Un concepto puede tener asociadas múltiples etiquetas, pero sólo una de ellas por cada idioma puede asociarse como etiqueta preferente. El resto de etiquetas asociadas al concepto se denominan etiquetas alternativas. También pueden definirse etiquetas ocultas con la finalidad de asignar a un concepto etiquetas que solo serían aplicables en los procesos de búsqueda e indización sin que sean visibles para los usuarios. Es posible asignar a los conceptos códigos de clasificación o de identificación dentro de un esquema conceptual determinado. Estas notaciones no están expresadas en lenguaje natural sino en forma de códigos nemotécnicos o similares. Los conceptos también pueden ser documentados con notas de diferente naturaleza como definiciones, notas de alcance o notas de edición entre otras. El modelo SKOS contempla el establecimiento de enlaces entre conceptos denominados relaciones semánticas. Estas relaciones pueden ser jerárquicas o asociativas, contemplándose la posibilidad de ampliar la tipología de relaciones. Los conceptos también pueden agruparse en colecciones que a su vez pueden etiquetarse y ordenarse. SKOS se complementa con la posibilidad de que conceptos de diferentes esquemas se pueden mapear entre sí empleando relaciones jerárquicas, asociativas o de equivalencia exacta.

4.8. TESAUROS AL ATAQUE

Como explicamos previamente, un tesauro es un vocabulario controlado y estructurado formalmente, formado por términos que guardan entre sí relaciones semánticas y genéricas: de equivalencia, jerárquicas y asociativas. Se trata de un instrumento de control terminológico que permite convertir el lenguaje natural de los documentos en un lenguaje controlado, ya que representa, de manera unívoca, el contenido de estos, con el fin de servir tanto para la indización, como para la recuperación de los documentos.

Frente a los lenguajes clasificatorios cuya función es describir el tema de un documento, los términos contenidos en un tesauro responden al análisis del texto o materia. Un tesauro recoge todos los conceptos y no sólo los que corresponden al título o el texto. Un único tema (aquello de lo que trata el documento) suele desarrollarse mediante una serie de ideas o conceptos que se pueden describir por medio de una serie de términos o descriptores. El tesauro incorporará todos esos términos en una base de datos y cada uno de ellos se convertirá en un punto de acceso para la recuperación del documento. La potencia de un tesauro radica además, en la posibilidad de combinar todos esos términos o descriptores, lo que le convierte en un lenguaje combinatorio mucho más rico que las tradicionales encabezamientos de materias. Un tesauro es pues, una herramienta de control terminológico muy útil para el análisis, descripción y recuperación automatizados.

4.8.1 Tesauros en línea

La siguiente es una lista de recursos de información de apoyo al análisis documental que se encuentran disponibles para consulta gratuita hasta el momento de este trabajo. La lista es extensa, pero vamos a hacer un resumen de los más conocidos, para luego centrarnos en AGROVOC, el cual es el tesauro elegido para este trabajo.

4.8.2 Generales, multidisciplinarios

4.8.2.A. OECD Macrhothesaurus Chapter Headings

http://bibliotecavirtual.clacso.org.ar/ar/oecd-macroth/es/index.htm

Versión en línea del Macrothesaurus para el procesamiento de la información relativa al desarrollo económico y social publicado por las Naciones Unidas en 1985. Disponible en español, inglés y francés.

4.8.2.B. **Tesauro de la UNESCO**

http://vocabularies.unesco.org/browser/thesaurus/es/

El Tesauro de la UNESCO es una lista controlada y estructurada de términos para el análisis temático y la búsqueda de documentos y publicaciones en los campos de la educación, cultura, ciencias naturales, ciencias sociales y humanas, comunicación e información. Continuamente ampliada y actualizada, su terminología multidisciplinaria refleja la evolución de los programas y actividades de la UNESCO.

4.8.2.C. **Tesauro Plurilingüe de Tierras**

ftp://ftp.fao.org/docrep/fao/005/X2038S/X2038S00.pdf

Tesauro de la Organización de las Naciones Unidas para la Agricultura y la Alimentación (FAO) creado en 2003, a partir de la versión en francés publicada en 1999. Esta obra persigue los siguientes objetivos: contribuir al análisis y a la proyección de la temática de tierras a partir de un uso más riguroso de conceptos y nociones; adecuar el empleo de la terminología al espacio de habla hispana (y a otros espacios lingüísticos); explicitar la polisemia de estos términos y expresiones, y contextualizar los diferentes usos para poder preservar su diversidad, sobre todo, cultural.

4.8.2.D. **Tesauro spines**

http://thes.cindoc.csic.es/index_SPIN_esp.php

Tesauro de la Unión Europea para políticas de ciencia, tecnología y desarrollo. La versión en línea fue creada en 2004. Se trata de un tesauro estructurado que contiene 10,865 términos, 42,752 relaciones entre términos, 3,000 términos equivalentes y 679 notas de alcance.

4.8.2.E. **Glosario de Drogas**

http://thes.cindoc.csic.es/index_GLODRO_esp.php

Recoge más de 1,000 términos relativos a drogas y sus traducciones, preparado por personal del grupo TermEsp del Instituto de Estudios Documentales sobre Ciencia y Tecnología (iedcyt), del Consejo Superior de Investigaciones Científicas de España (csic).

4.8.3. Agrociencias

4.8.3.A. Cab

Thesaurus http://www.cabi.org/cabthesaurus/

Creado en 1983 para la base de datos “Cab Abstracts”. Contiene términos en español, inglés y portugués. Se actualiza regularmente, la última actualización data de febrero de 2010. Especializado en agro ciencias, cubre además temas de ciencias puras, aplicada, de la vida, tecnología y ciencias sociales. Incluye aproximadamente 98,000 términos (66,000 preferidos y 32,000 no preferidos), así como 62,000 nombres de plantas, animales y microorganismos.

4.6.3.B. Tesauro y Glosario Agrícola

http://agclass.nal.usda.gov/agt_es.shtml

Iniciado en 2002 es producto de la cooperación entre la Biblioteca Nacional de Agricultura de los Estados Unidos y el Instituto Interamericano de Cooperación para la Agricultura (IICA). El tesauro y el glosario son herramientas de vocabulario especializado para la agricultura disponible en inglés y español. Se actualiza anualmente e incluye más de 80,000 términos y 29,000 referencias cruzadas. Tiene una cobertura amplia de temas sobre agricultura y biología. El glosario incluye definiciones de terminología técnica, así como términos locales y regionales de países latinoamericanos. El contenido puede consultarse por medio de un buscador que, además de dar al usuario la oportunidad de seleccionar la forma de visualizar los resultados, permite interrogar por medio de cadenas de caracteres o truncando los términos por su parte final. La consulta también se puede realizar de modo manual en un índice jerárquico expandible con 17 áreas temáticas o en un índice alfabético de términos, ambos en formato HTML.

4.8.3.C. Agrovoc

http://www.fao.org/aims/ag_intro.htm

AGROVOC es un vocabulario controlado que abarca todos los ámbitos de interés de la Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO), entre ellos la alimentación, la nutrición, la agricultura, la pesca, las ciencias forestales y el medio ambiente. Lo publica la FAO y una comunidad de expertos se encarga de su edición.

La Organización de las Naciones Unidas para la Alimentación y la Agricultura (FAO) publicó por primera vez el tesauro AGROVOC a principios del decenio de 1980 en inglés, español y francés. Su finalidad es ofrecer un vocabulario controlado para la indexación de publicaciones en el ámbito de la ciencia y tecnología agrícola.

En el año 2000 AGROVOC dejó de imprimirse en papel y pasó a formato digital, con un almacenamiento de datos administrado por una base de datos relacional. En 2004 se

realizaron pruebas para la conversión a OWL, y en 2009 AGROVOC se transformó en un recurso SKOS.

En la actualidad, AGROVOC está disponible en 23 idiomas en forma de esquema conceptual SKOS-XL. También se publica como conjunto de datos de acceso abierto vinculados (LOD por sus siglas en inglés), alineado con 13 conjuntos de datos relacionados con la agricultura.

La ventaja de publicar un tesauro como AGROVOC en forma de datos abiertos vinculados consiste en que una vez vinculados los tesauros los recursos indizados por estos también quedan vinculados. Un buen ejemplo de ello es AGRIS, una aplicación web híbrida que vincula el depósito bibliográfico de AGRIS (indizado con AGROVOC) con los recursos web relacionados (indizados con vocabularios vinculados a AGROVOC).

AGROVOC puede usarse para encontrar el nombre popular de una planta en una lengua que no se domina, o para encontrar las relaciones entre un producto y los cultivos del cual procede.

A la fecha, AGROVOC es utilizado por investigadores, bibliotecarios y gestores de información para para la indización, recuperación y organización de datos en sistemas de información y páginas web sobre la agricultura. Actualmente AGROVOC es un esquema conceptual SKOS-XL, un conjunto de datos de abiertos vinculados (LOD por su sigla en inglés) que está alineado con otros 16 sistemas multilingües de organización del conocimiento relacionados con la agricultura.

AGROVOC se edita a través de VocBench, una plataforma de código abierto basada en la Web que permite la edición de tesauros multilingües y recursos RDF-SKOS en colaboración.

4.8.3.D. USDA

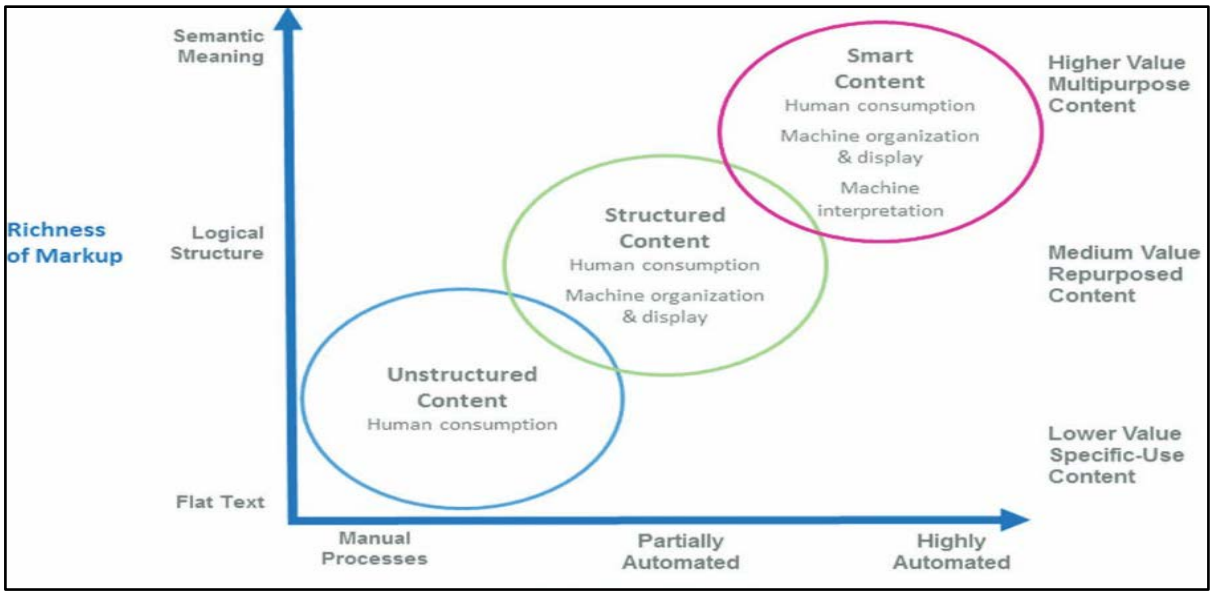
El tesauro y glosario son herramientas de vocabulario especializado para la agricultura disponibles en inglés y en español. Este producto es el resultado del esfuerzo cooperativo entre la Biblioteca Nacional de Agricultura de EE.UU. (NAL por sus siglas en inglés), perteneciente al Departamento de Agricultura de los EE.UU. , y el Instituto Interamericano de Cooperación para la Agricultura (IICA), así como de otras Instituciones Agropecuarias de América Latina pertenecientes al Servicio de Información y Documentación Agrícola de las Américas (SIDALC).

- Actualización anual cada Enero desde el 2002
- Versiones bilingües español/inglés paralelas
- Disponible como Linked Open Data
- Cobertura exhaustiva de temas como agricultura, biología y temas afines

- Contiene más de 105.432 términos, incluyendo 38.138 referencias cruzadas
- Glosario de definiciones para terminología técnica
- Términos locales/regionales de países Latinoamericanos

5. ENRIQUECIMIENTO SEMÁNTICO Y CONTENIDO INTELIGENTE

El enriquecimiento semántico es el proceso mediante el cual se le agrega a un contenido una capa de metadatos y tópicos de manera tal que las máquinas puedan entender un contenido y construir conexiones entre otros similares. Una vez que un contenido es semánticamente enriquecido es muchas veces llamado "Contenido inteligente" porque contiene dentro de sí todo lo que un agente informático necesita para entenderlo estructural y temáticamente. La práctica de agregar información semántica como la hablada anteriormente -metadatos, clasificación en distintos tipos de vocabularios controlados- es usualmente llamado "semantic tagging" (etiquetado semántico). Dado un artículo a ser etiquetado (proceso de tagging), el mismo puede ser realizado usando archivos XML que hagan referencia a este contenido. Esta información puede ser puesto en la cabecera del documento de una forma que el lector humano no lo vea (pero sí los agentes de software que indexan o realizan el crawling del contenido), en un bloque de texto o hasta nivel de sentencia o palabra usando, por ejemplo, microdata.



5.1 ¿QUIÉN (O QUÉ) HACE EL "SEMANTIC TAGGING"?

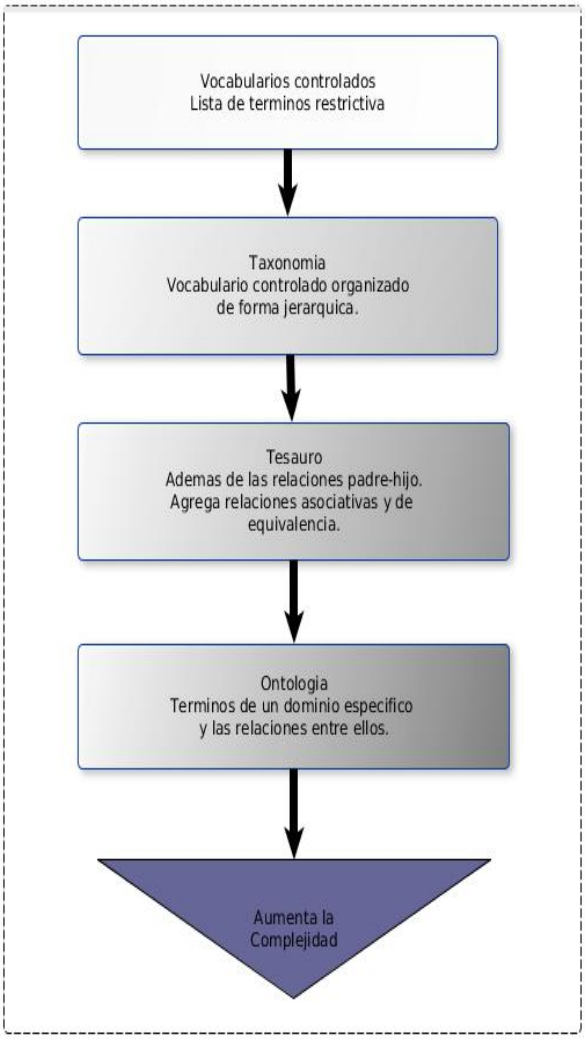
5.1.1 Etiquetado manual y etiquetado automático

Principalmente hay dos aproximaciones: manual y automático. El etiquetado manual se refiere a una persona, usualmente un experto, quien lee los contenidos y aplica las etiquetas. Esto es necesario cuando se requiere un alto grado de precisión, pero no es escalable. En

cambio, el etiquetado automático, está hecho por un agente de software que agrega contenidos basado en patrones estadísticos, macheado de patrones y análisis lingüístico. Generalmente estos agentes de software pasan por una etapa de entrenamiento en donde, usualmente, varios expertos en etiquetas toman un conjunto de contenidos, los etiquetan con una alta precisión y se los proveen al agente de software para que este mejore su precisión en el campo y lenguaje requerido. El etiquetado automático es bueno para grandes volúmenes de datos, pero puede dar falsos positivos y así etiquetar erróneamente contenidos. Lo mejor es usar una aproximación híbrida donde, por ejemplo, al publicador de un contenido se le sugieren las etiquetas que devuelve el agente de software el cual a su vez se puede basar en reglas específicas al conjunto de datos o incluso a un tesauro para que, el publicado, pueda luego agregar y/o quitar las etiquetas provistas automáticamente.

5.1.2. Sistemas de organización de conocimiento

Es importante decidir hasta dónde se llegará con el nivel de complejidad de enriquecimiento del contenido. El siguiente gráfico muestra los distintos niveles de complejidad, ordenados de ascendentemente.



Como vemos en el gráfico, los niveles de complejidad de los sistemas de organización de conocimiento van ascendiendo desde una lista de términos predefinidos administrados manualmente – “vocabularios controlados”,

siguiendo por una lista de términos con relaciones de padre-hijo – "taxonomías", tesauros hasta las especificaciones ontológicas.

5.1.3. Ventajas del enriquecimiento semántico

- Agrupado temático y relaciones de jerarquía: en un "semantic tagging" gobernado por taxonomías permite agrupar los contenidos por tópico de forma jerárquica permitiéndole al usuario buscar jerárquicamente por un listado de tópicos.
- *Contenido relacionado por vínculos*: ofrece al usuario contenidos relacionados a partir de las etiquetas proporcionadas. Un usuario puede comenzar viendo un artículo específico y llegar a otros que contengan etiquetas relacionados entre sí.
- *Interconexión*: haciendo uso de tesauros los contenidos de un sitio se pueden conectar con los contenidos de otros que usen los mismos tesauros y así ofrecer fuentes externas de datos a través del tagging.

Conectar contenidos a los usuarios: a través del análisis de las etiquetas de los contenidos que un usuario visita, se le puede ofrecer a este usuario contenido con etiquetas similares o incluso publicidades.

6. HACIA LA GESTIÓN DEL CONOCIMIENTO

La web es un acervo virtual de información científico técnica en todo tipo de temas y en un espectro de formatos, en gran parte textual. Sin embargo, al hacer uso de este repositorio de información textual, debe ser organizado y estructurado de tal manera que sean significativos para las personas y procesable por los dispositivos que diariamente usan la web. Una de las visiones de la Web Semántica ha sido enriquecer la información en la Web con la organización y la estructura. Sin embargo, dado que el texto está en un lenguaje natural (por ejemplo, español, inglés, alemán, etc.), lo que la gente puede entender, máquinas no. Es necesario aplicar entonces, algún tipo de procesamiento automatizado del propio texto. En este capítulo, se discute una aproximación a la información técnica generada por las organizaciones de ciencia y técnica en la World Wide Web, la Web Semántica y Procesamiento del Lenguaje Natural (NLP). Cada uno de estos son grandes temas de investigación en si mismos, por lo que en esta sección intentaremos dar un enfoque particular y parcial.

6.1 INTRODUCCIÓN

Para el público y los investigadores, la Web ofrece una oportunidad sin precedentes para buscar, encontrar, y exponer la información generada, como papers, artículos, libros o material relacionado con el dominio de conocimiento.Dominios de conocimiento: Son núcleos de aprendizaje esenciales de la ciencia que conforma cada área curricular; tienen un sentido abarcador e intentan dar cuenta de todos los aspectos principales del área. al cual pertenece. Con una herramienta de búsqueda como Google u otras búsquedas indexadas, el público puede dar palabras claves de entrada en una búsqueda y obtener a cambio una lista de los documentos que contienen o se indexan por esas palabras clave.

El problema subyacente es que la mayoría de los documentos resultantes de la producción intelectual están expresados en lenguaje natural, donde los conceptos son más importantes que lo términos. Entonces, ¿cómo podemos aportar a mejorar la inyección de metadatos para mejorar las búsquedas?

Generalmente en la publicación del documento, se adopta un esquema de metadatos para su descripción. La gestión de metadatos nos brinda los elementos necesarios para contextualizar todo el ciclo de vida de los documentos; es por ello, que se requiere de la creación, diseño y mantenimiento del esquema de metadatos que permita asegurar la buena gestión de los documentos y su relación con los sistemas de información de la entidad u organización.

Independiente de la estrategia elegida para gestionarlos, son una parte importante al momento de estructurar las búsquedas, pero estos metadatos dependen de factores externos al propio contenido; como de la "expertise" del documentalista que los trata y el método elegido para hacerlo.

¿Cómo podemos entonces aportar a una mejora para extraer elementos contenidos en el documento para mejorar la clasificación y aportar a las búsquedas para relacionar los conceptos inyectados con otros existentes?

En la siguiente sección propondremos un proceso que aporte dichos elementos para enriquecer las búsquedas.

6.2. EL CONTEXTO INTA

El Instituto Nacional de Tecnología Agropecuaria, INTA, es un organismo nacional cuya misión es "realizar y promover acciones dirigidas a la innovación en el sector agropecuario, agroalimentario y agroindustrial para contribuir integralmente a la competitividad de las cadenas agroindustriales, salud ambiental y sostenibilidad de los sistemas productivos, la equidad social y el desarrollo territorial, mediante la investigación, desarrollo tecnológico y extensión". En este contexto la generación de información juega un rol fundamental tal como se expresa claramente en su Plan Estratégico Institucional (PEI 2005-2015). Tal información se genera a través de las actividades sustantivas: Investigación, Extensión, Vinculación Tecnológica y Relaciones Institucionales y debe ser puesta a disposición de la sociedad a través de su sitio web, su portal de noticias y sus repositorios. La producción intelectual es expresada en gran parte en documentos, lo que dificulta realizar búsquedas globales precisas de conceptos y entidades entre los contenidos distribuidos en la institución.

Lo descrito previamente nos creó la motivación de tomar el problema y desarrollar una solución sustentable para el organismo. Para dar una idea de la producción de conocimiento brindamos algunas estadísticas de su sitio web:

- 3230 archivos de audio
- 132 becas
- 18471 documentos
- 7356 eventos
- 97 elementos del catálogo de maquinarias
- 934 páginas
- 11214 personas

- 1672 servicios
- 376 variedades
- 2691 videos
- 36783 archivos pdf adjuntos (Informes, Artículos de divulgación, Libros, Tesis, etc).

Estos contenidos se encuentran agrupados en taxonomías administradas por los webmasters del INTA y con más de 580 unidades produciendo contenido en el sitio.

6.3. PROPUESTA

Como mencionamos previamente, los documentos son descritos por metadatos antes de su publicación en los repositorios.

Nuestra propuesta es la inclusión de pasos en el proceso previo a la publicación para extraer información adicional a partir del contenido, y en particular establecer como contexto el dominio de conocimiento sobre el cual estamos trabajando.

El primer paso que incluimos el enriquecimiento de Entidades con Nombre. El reconocimiento de entidades con nombre consiste, como su nombre indica, en la detección de elementos con nombre propio dentro de un determinado texto, y su clasificación en categorías predefinidas, como personas, organizaciones, empresas o lugares. La complejidad de este proceso se basa en que muchas entidades equivalentes pueden aparecer escritas de diferentes formas, por lo que es necesario contar con un conjunto de reglas semánticas avanzadas y un corpus (una colección de piezas del idioma que se seleccionan de acuerdo a criterios lingüísticos explícitos con el fin de ser utilizado como una muestra de la lengua para luego cruzar la información para realizar desambiguaciones). En nuestra propuesta seleccionamos las herramientas de código abierto OpenNLP y Stanford NER, ampliamente reconocidas y que proveen soporte al lenguaje español. El resultado de este paso es obtener las entidades expresadas en el documento y que aportan elementos para priorizar las búsquedas.

El siguiente paso, es la detección de frases claves o relevantes condicionadas por un tesauro que nos aproxima a un dominio de conocimiento específico. En el caso de estudio, utilizaremos tres tesauros:

1. Agrovoc (http://aims.fao.org/es/agrovoc), el tesauro de la FAO (FAO - Organización de las Naciones Unidas para la Alimentación y la Agricultura, http://www.fao.com),
2. el tesauro de USDA (Departamento de Agricultura de Estados Unidos) el cual comparte el dominio de conocimiento con Agrovoc y,

3. tuvimos en cuenta también el tesauro de la UNESCO el cual provee términos sobre muchas disciplinas.

El proceso de extracción de palabras claves son un medio importante para el resumen del documento, el agrupamiento, y tópicos de búsqueda. Sólo una pequeña minoría de los documentos tiene frases claves asignadas por el autor, y la asignación manual de frases claves a los documentos existentes es muy laborioso. Por lo tanto es deseable automatizar el proceso de extracción frases claves. Es esperable que la calidad de las palabras clave extraídas mejore significativamente cuando se explota información de un dominio específico.

Las frases claves dan una descripción de alto nivel del contenido de un documento cuando se pretende facilitar a los posibles lectores la relevancia del mismo. Pero tienen otras aplicaciones también: las frases claves al resumir documentos en forma muy concisa se pueden utilizar como una medida de bajo costo de similitud entre documentos, lo que posibilita agrupar documentos mediante la medición de solapamiento entre las palabras claves ya asignadas. Una aplicación relacionada es la búsqueda de temas: al entrar en una frase clave en un motor de búsqueda, todos los documentos con esta frase clave son devueltos al usuario. En resumen, las frases y palabras claves proporcionan un medio poderoso para tamizar a través de un gran número de documentos, centrándose en aquellos que son propensos a ser relevante.

6.3.1 Descripción del proceso

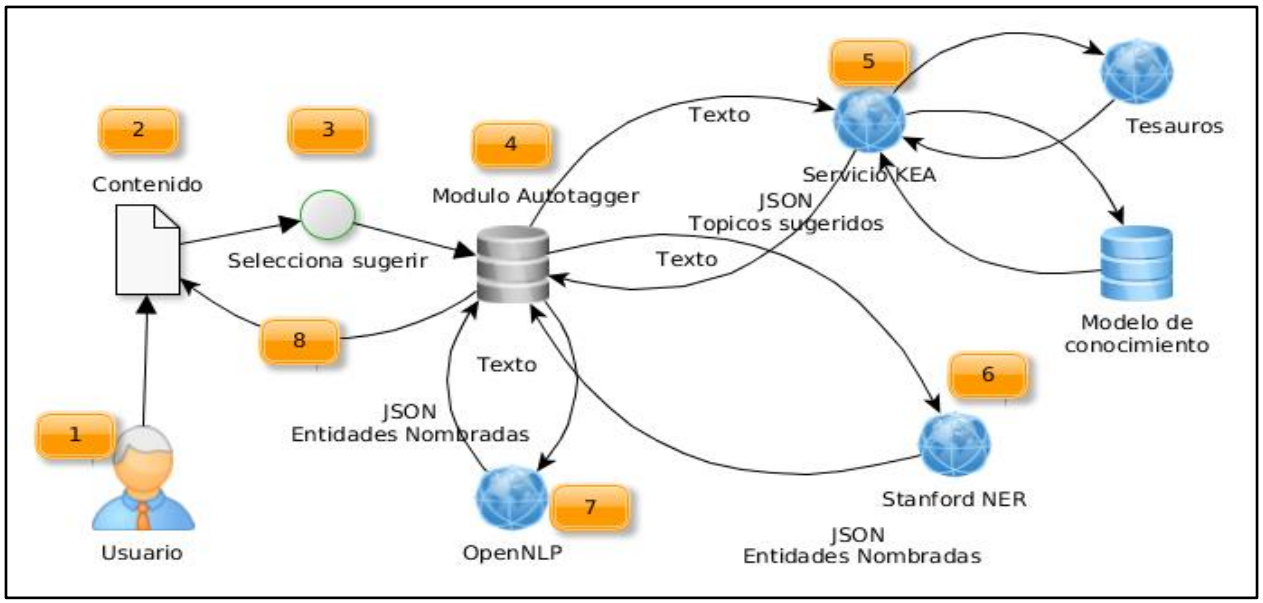
1. El usuario inyecta el documento al repositorio,
2. El documento se describe a través del esquema de metadatos elegido y se almacenan estableciendo la relación Documento->Metadatos,
3. El documento se procesa al efecto de detectar y extraer las entidades con nombre, las que también son almacenadas en relación al documento,
4. El documento se vuelve a procesar extrayendo las palabras claves, priorizando aquellas que estén en el tesauro elegido para el dominio de conocimiento sobre el cual estamos trabajando.

Es decir, al final del proceso tendremos un conjunto de metadatos y etiquetas que describen el documento:

- Metadatos ingresados por un usuario
- Entidades nombradas extraídas automáticamente (Nombre, Organizaciones y Lugares)

- Palabras y frases claves que describen el documento

El uso del tesauro nos posibilita que establezcamos relaciones entre conceptos descritos en el documento y en consecuencia relaciones a partir de ellos entre los documentos que tienen conceptos relacionados. Esquemáticamente:



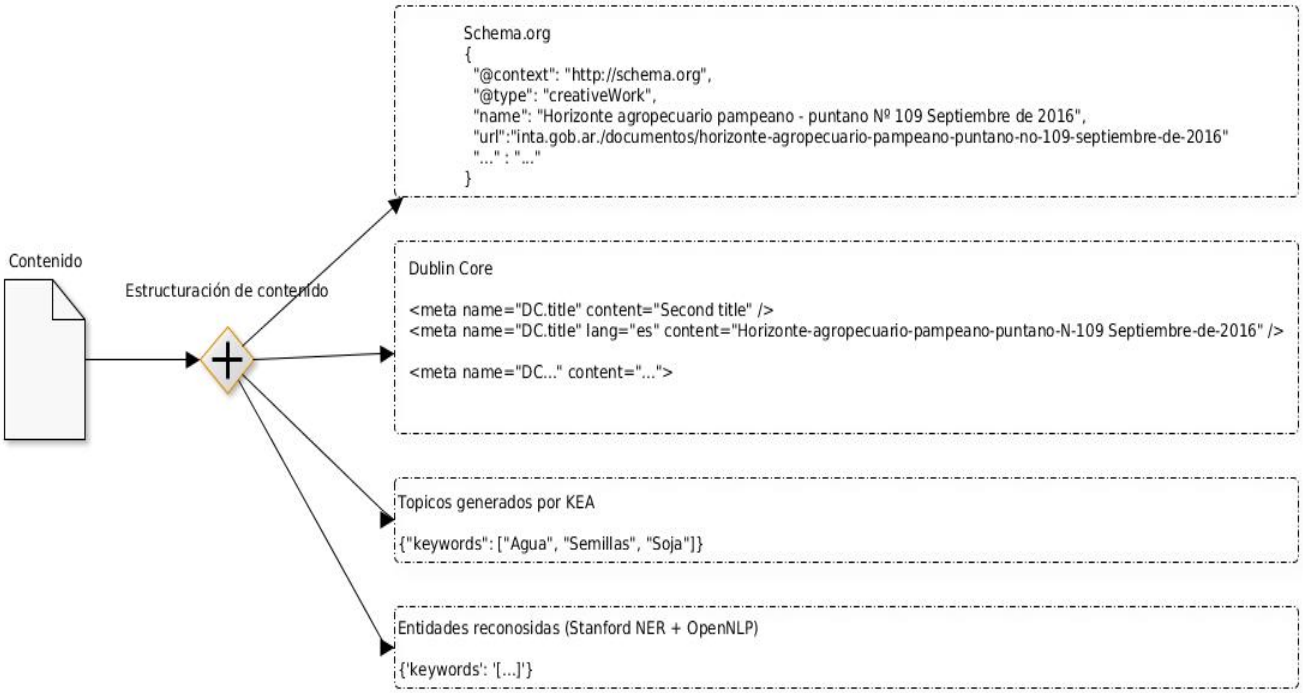
6.3.2. Búsqueda interna y externa

La información de una organización, en nuestro caso los contenidos del INTA, no sólo son accedidos por el buscador interno, sino que también la aplicación de las palabras claves y los metadatos dentro del contenido son importantes para mejorar las búsquedas externas.

Para esto utilizaremos una iniciativa de Google, Bing y Yahoo para mejorar el etiquetado de páginas webs y que puedan interpretar mejor su contenido usando el concepto de datos estructurados que propone schema.org. Este esquema es un conjunto de etiquetas y atributos añadidos al código HTML que le indican a los buscadores qué es cada cosa de nuestro contenido. Por supuesto, sin influir en la maquetación o en la forma de mostrar el contenido.

Las ventajas de usar estos microdatos que propone schema.org son varios: mejorar la indexación de nuestra web, relacionar contenido entre varias páginas por medio de atributos sociales específicos, resaltar nuestro contenido en fragmentos enriquecidos (con fotos y textos) en los buscadores. Además de facilitar la tarea de compartir en redes sociales, como Google + que ya

usa este tipo de datos estructurados para mostrar el contenido en el snippet de su timeline.



6.4. BÚSQUEDAS INTERNAS

6.4.1 Extracción de tópicos del documento en el INTA

En esta sección describiremos la inyección de palabras clave en los contenidos. Éste procedimiento es muy útil para describir los documentos y, como ya dijimos, de gran utilidad para los indexadores, en especial, cuando los contenidos son etiquetados de acuerdo al dominio de conocimiento de la organización (Medelyan, Thesaurus Based Automatic Keyphrase Indexing, 2006).

6.4.1.A Cómo se asignan las palabras claves en la institución

El INTA produce conocimiento a lo largo de todo el país. La tarea de asignación de palabras claves a sus textos es realizada manualmente. Existen alrededor de 15 documentalistas que etiquetan profesionalmente los contenidos en los centros documentales usando términos de un vocabulario controlado. Sin embargo, en el sitio web la realidad es otra; como el proceso de etiquetación por los documentalistas es lento y muchas veces los documentos deben salir a la luz con celeridad, debe ser realizado por los webmasters quienes deben sumar, además de a

sus muchas tareas, la de leer y etiquetar artículos y documentos lo cual es una tarea que lleva un tiempo considerable.

Uno de nuestros objetivos es lograr realizar un etiquetado automático eficiente y veloz con buenas probabilidades de éxito, usando un modelo entrenado en documentos propios del INTA etiquetados por profesionales de la misma organización. Si la aproximación funciona, mejoraría la velocidad y exactitud de los términos dando un respiro a los webmasters en su trabajo y restringiendo a su vez, los términos elegidos a un tesauro bien conocido.

Para lograrlo, usaremos un algoritmo llamado KEA (Keyword extraction algorithm).

6.4.2 KEA

KEA es un algoritmo cuyo objetivo es extraer frases y palabras claves de los documentos. En nuestro caso, lo usaremos junto con uno o varios tesauros al mismo tiempo para mejorar la extracción de palabras claves. (Medelyan, Thesaurus Based Automatic Keyphrase Indexing, 2006)

6.4.2.A Generación de un modelo

Antes de que KEA pueda extraer palabras y frases clave de un texto, se debe realizar un entrenamiento del mismo en donde se proveen al "constructor del modelo" una serie de documentos junto con palabras claves asignadas manualmente para cada documento. A esta colección se le suma un tesauro en formato SKOS para el mejor funcionamiento. El resultado de dicho proceso será un archivo que representará el modelo que KEA necesitará para extraer textos.

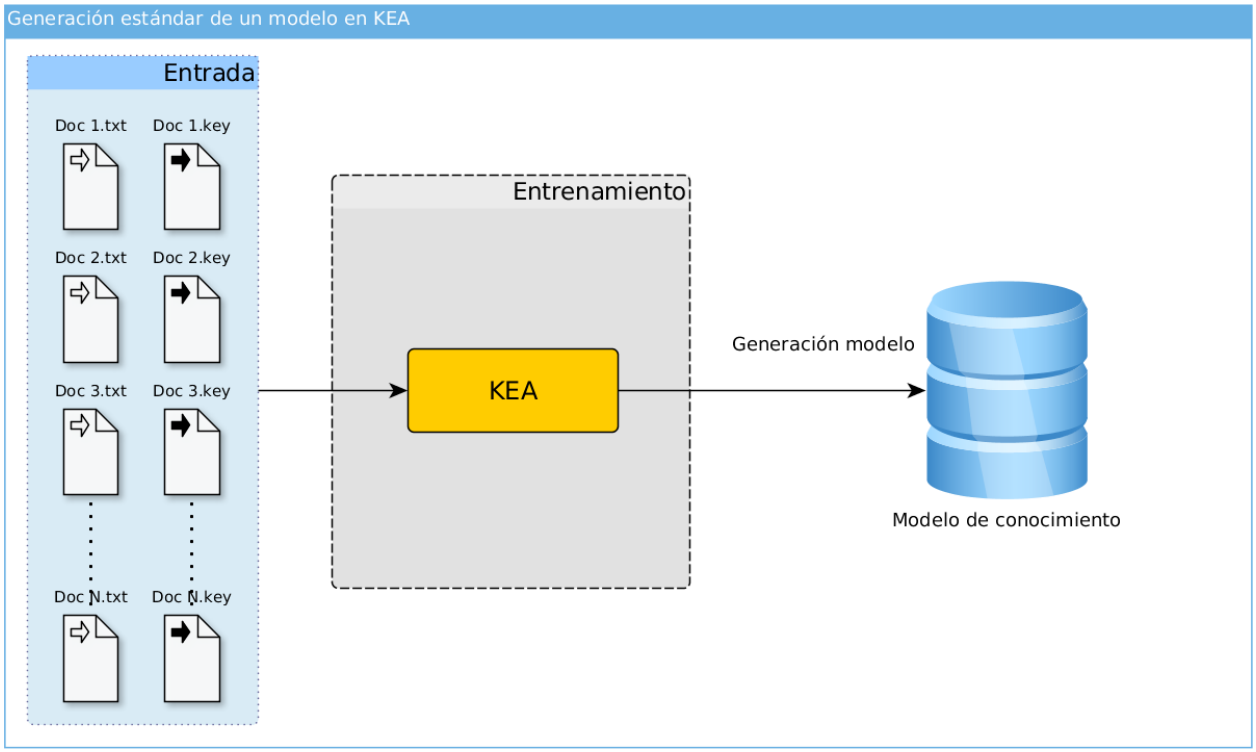
A mayor detalle, para construir el modelo, se usa un conjunto de documentos en donde las frases clave asignadas por el autor son usadas. Para cada documento, los tópicos candidatos son identificados y sus cualidades atractivas calculadas. KEA identifica 4 cualidades:

- TFxIDF: Esta función compara la frecuencia de una frase en un documento en particular con la frecuencia de esa frase en uso general. El uso general es representado por la frecuencia del documento -el número de documentos que contienen la frase en algunos cuerpos-. La frecuencia de un documento de una frase indica lo común que es (y frases más raras son probablemente más indicadas para ser frases clave). Kea construye un documento de frecuencia con un corpus de alrededor de 100 documentos. Frases de candidatas se generan a partir de todos los documentos de este corpus utilizando el método descrito anteriormente. El documento archivo de

frecuencia almacena cada frase y una cuenta de número de documentos en los que aparece.

- La posición de la primera ocurrencia de la frase.
- El largo de la frase en palabras.
- El grado del nodo: el grado del nodo representa el número de relaciones que conectan la frase a otros términos candidatos del tesauro usado. Es decir, si se encuentra un grupo de frases candidatas que estén relacionadas entre sí, este puntaje tendrá gran importancia y será muy probable que sean significativas.

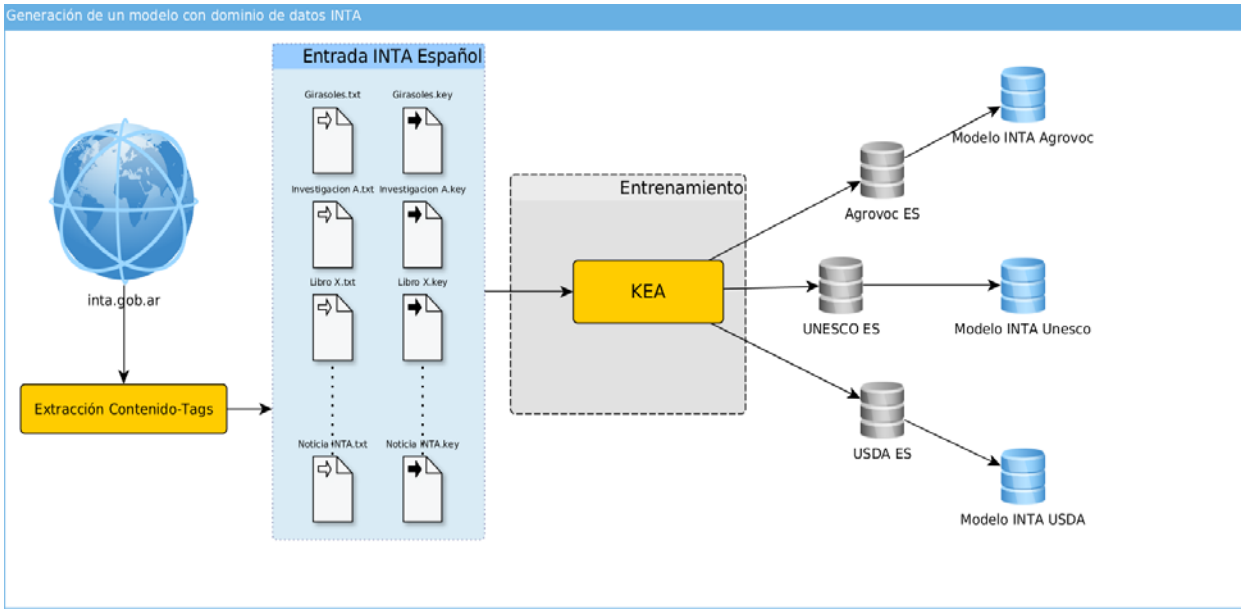
Cada palabra en el conjunto de frases claves asignadas por el autor es marcada como "término de indexado". El algoritmo de KEA tiene la capacidad de predecir términos de indexado en cualquier documento en base a las 4 cualidades descritas anteriormente (Keyword Indexing Algorithm) y lo aprendido en base a las frases clave provista. Cuando se reconoce un nuevo documento, KEA encuentra las palabras importantes usando el indexado de frases clave descrito por las 4 cualidades y, luego, aplica el modelo usado y el tesauro provisto para el entrenamiento para predecir sus probabilidades.



6.4.2.B. Generando el modelo para el INTA

El primer paso para esto, es encontrar un conjunto de documentos y palabras claves para la generación del modelo. Se realizó un desarrollo para obtener archivos de texto con el cuerpo

de contenidos del sitio web y la generación de otro archivo de texto con las frases claves asociadas por los profesionales de la institución de más de 5000 documentos. Con la generación de estos archivos, más la asociación a un Tesauro por parte de KEA, pudimos generar el modelo de conocimiento para automatizar el etiquetado de contenidos.



6.4.2.C. Elección del Tesauro

KEA tiene la capacidad de trabajar sobre cualquier tesauro en formato SKOS. La idea fue usar un tesauro que se acople al dominio de conocimiento de la organización.

En un principio solamente fue tomado en cuenta el tesauro de la FAO (Agrovoc), pero a medida que se hicieron pruebas, comprobamos que el tesauro de la USDA daba muy buenos resultados y que el tesauro de la UNESCO en ocasiones devolvía términos relevantes no incluidos en los otros dos. KEA, al devolver resultados, calcula la probabilidad de éxito de la frase clave entre 0 y 1. Usando los 3 tesauros al mismo tiempo para obtener resultados en combinación con el puntaje, nos dimos cuenta que en los resultados, los términos agropecuarios obtenían un puntaje mayor y, aquellos de la UNESCO, generalmente recibían puntaje bajo. Sin embargo, en algunas excepciones, algunos resultados del tesauro de la UNESCO no incluidos en los demás daban un puntaje alto y se ubicaban en los candidatos correctos, como por ejemplo, “Economía”, “Desarrollo social” y “Cultivos” por lo que elegimos usar los 3 y basarnos en la probabilidad de éxito para incluirlos o no en los candidatos elegidos.

6.4.2.D. Un servicio REST para KEA

KEA al procesar textos toma elementos con extensión ".txt" de una carpeta con el contenido de un solo documento que se desea procesar y luego de su procesamiento, deja en la misma carpeta un archivo de texto con nombre igual al elemento analizado pero con extensión ".key". En nuestro caso este enfoque no era útil; la propuesta es ofrecer etiquetado automático al instante a algo que, probablemente, todavía no fue ni siquiera guardado sin necesidad de tener acceso al sistema de archivos del host donde se encuentra instalado KEA. Además, la implementación de la solución debe ser reutilizable por cualquier pedazo de software sin costo alguno, usando librerías ya provistas en la gran mayoría de los lenguajes proveyendo, además, una respuesta simple y legible al humano.

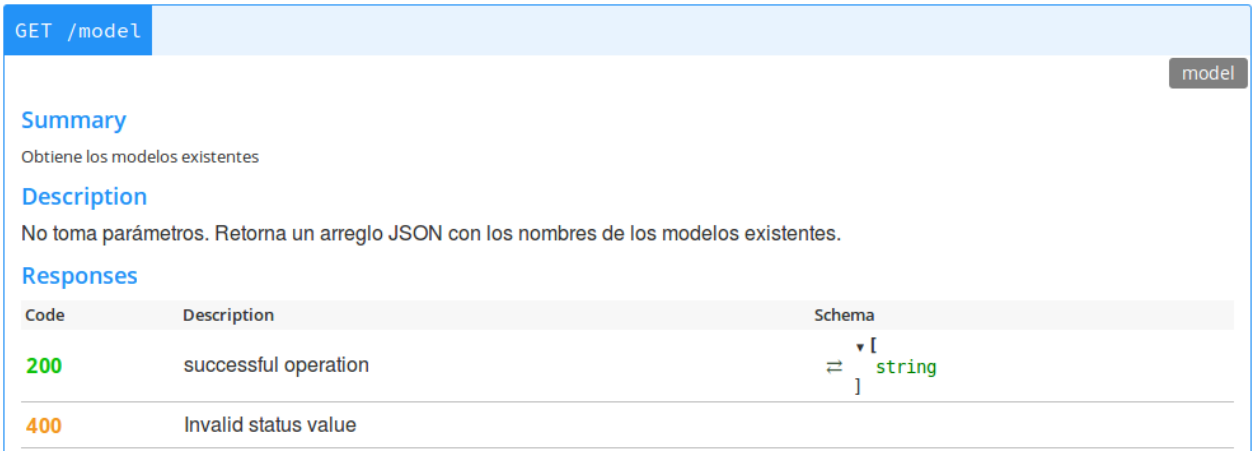
Decidimos implementar un servicio REST, con una API simple, que reciba un texto y, retorne en formato JSON las frases clave del documento.

Para esto tuvimos que, a través de la modificación del código fuente de la librería, adaptarla para que ya no use archivos de texto en el disco y ofrecer una respuesta en JSON con un mayor detalle en la respuesta incluyendo en qué tesauros fue encontrado el término y con qué probabilidad de éxito. También se debió adaptar la librería para que funcione correcta y velozmente con el acceso concurrente ya que, por defecto, no está preparada para este tipo de acceso. Elegimos la última versión de KEA (5.1) que viene preparada para trabajar con cualquier tesauro en formato SKOS y, además, provee stemmers en español.

6.4.2.E. API REST del servicio de KEA

6.4.2.E.i GET /model

Esta operación devuelve los modelos existentes en el servicio para luego poder consultar solo


GET /model

model

Summary

Obtiene los modelos existentes

Description

No toma parámetros. Retorna un arreglo JSON con los nombres de los modelos existentes.

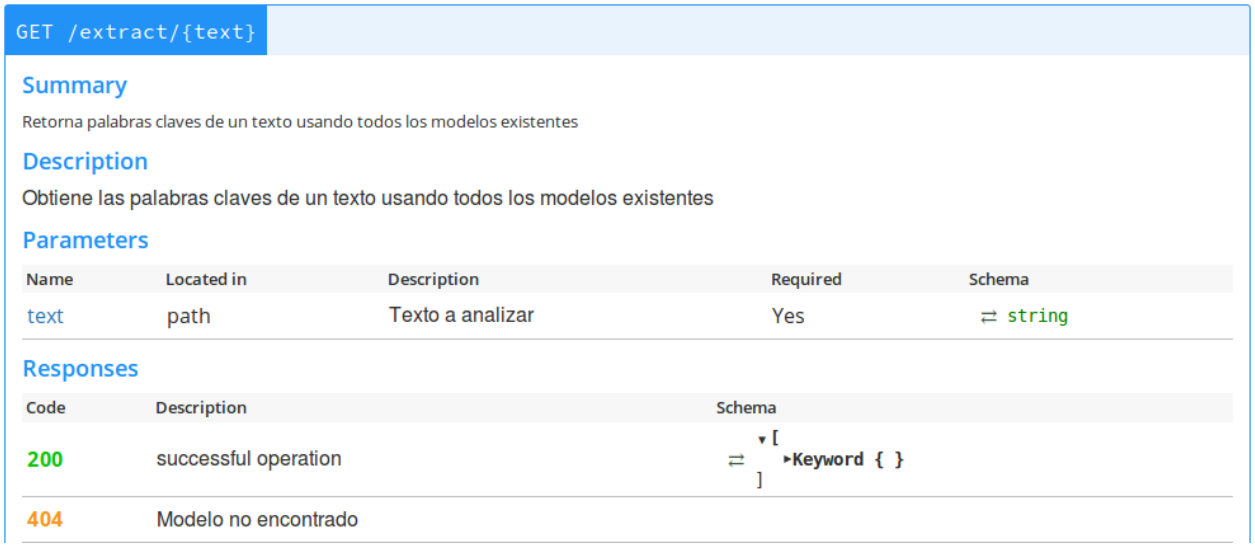
Responses

Code	Description	Schema
200	successful operation	⇄ [string]
400	Invalid status value	

a uno de ellos y no a todos los tesauros. De esta forma se pueden usar modelos en distintos idiomas, con distintos tesauros y con diferentes tipos de documentos de entrenamiento.

6.4.2.E.ii GET /extract/{text}

El método "extract" genérico retorna los tópicos para el texto pasado como parámetro usando


GET /extract/{text}

Summary

Retorna palabras claves de un texto usando todos los modelos existentes

Description

Obtiene las palabras claves de un texto usando todos los modelos existentes

Parameters

Name	Located in	Description	Required	Schema
text	path	Texto a analizar	Yes	⇄ string

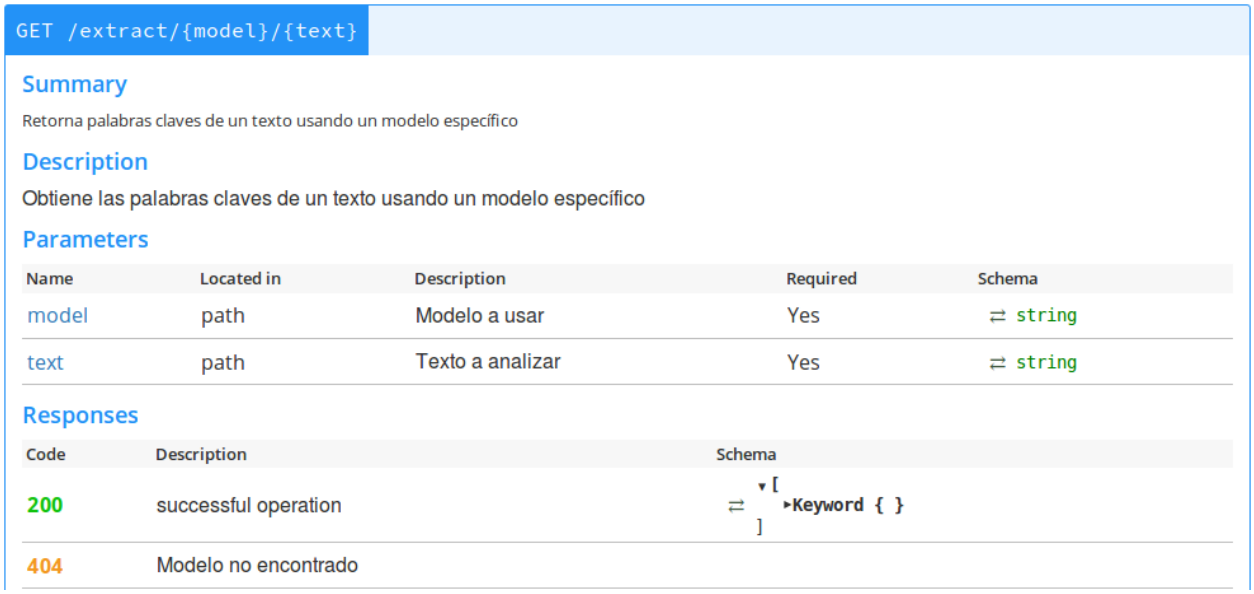
Responses

Code	Description	Schema
200	successful operation	⇄ [▸Keyword { }]
404	Modelo no encontrado	

todos los tesauros.

6.4.2.E.iii GET /extract/{model}/{text}

Esta operación retorna los tópicos del texto pasado como parámetro usando el modelo especificado.



GET /extract/{model}/{text}

Summary

Retorna palabras claves de un texto usando un modelo específico

Description

Obtiene las palabras claves de un texto usando un modelo específico

Parameters

Name	Located in	Description	Required	Schema
model	path	Modelo a usar	Yes	⇄ string
text	path	Texto a analizar	Yes	⇄ string

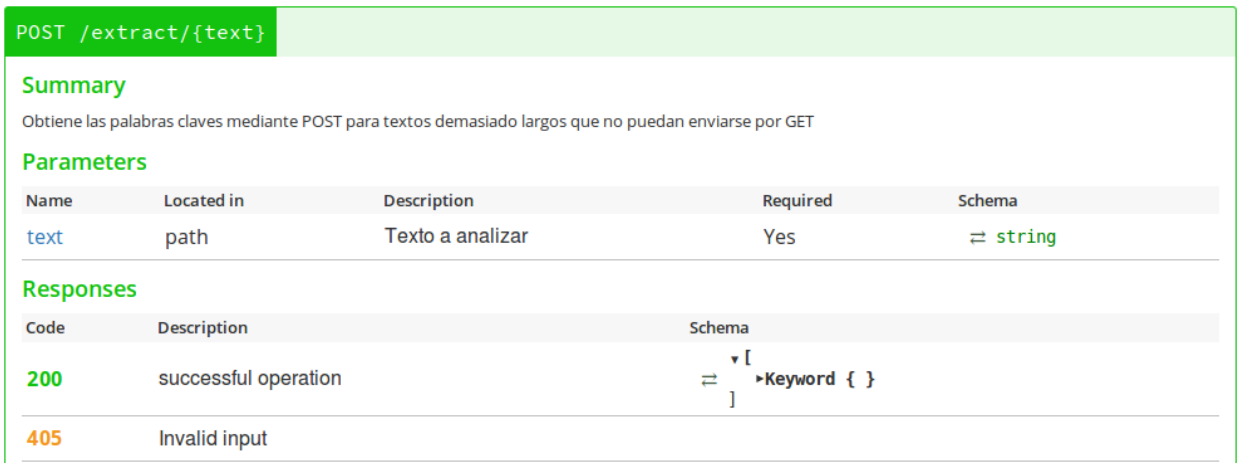
Responses

Code	Description	Schema
200	successful operation	⇄ ▾ [▸Keyword { }]
404	Modelo no encontrado	

6.4.2.F. *Limitaciones de los request GET*

Los métodos que hemos mostrado, todos obtienen datos del servidor sin cambiar nada en el. Se les pasa un parámetro, procesan la respuesta y nos devuelven los tópicos. A misma entrada, misma salida. REST específica que para este tipo de request se debe usar GET. La limitación con la que cuenta GET para este caso, es que, el tamaño del parámetro TEXT puede ser muy grande. Los servicios HTTP pueden tener un límite variable al tamaño de los request GET por lo que, para evitar malfuncionamientos inesperados, hicimos también las alternativas POST que abajo se describen.

6.4.2.F.i. POST /extract/{text}



POST /extract/{text}

Summary

Obtiene las palabras claves mediante POST para textos demasiado largos que no puedan enviarse por GET

Parameters

Name	Located in	Description	Required	Schema
text	path	Texto a analizar	Yes	⇄ string

Responses

Code	Description	Schema
200	successful operation	⇄ ▾ [▸Keyword { }]
405	Invalid input	

6.4.2.F.ii. POST /extract/model/{text}

POST /extract/{model}/{text}

Summary

Retorna palabras claves de un texto usando un modelo específico

Description

Obtiene las palabras claves mediante POST para textos demasiado largos que no puedan enviarse por GET usando un modelo específico

Parameters

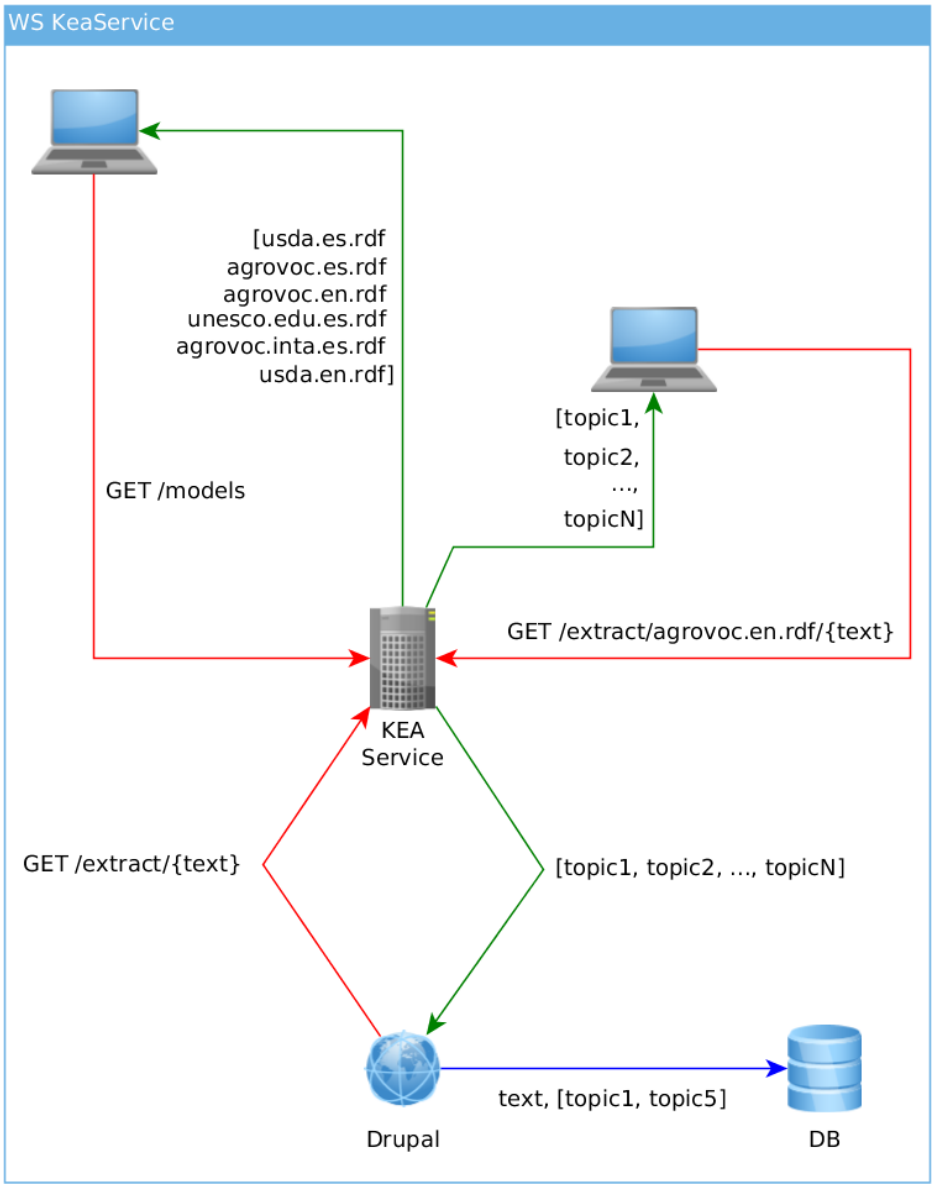
Name	Located in	Description	Required	Schema
model	path	Modelo a usar	Yes	⇄ string
text	path	Texto a analizar	Yes	⇄ string

Responses

Code	Description	Schema
200	successful operation	⇄ ▾ [▸Keyword { }]
404	Modelo no encontrado	

6.4.2.G. Uso de la API

Gracias a la implementación del servicio para KEA, se pudo agregar una capa por delante de las librerías que facilita el uso y acceso a la misma. Usándolo, se pueden tener varios modelos distintos, en varios idiomas generados para distintos tesauros y distintas instituciones.



La respuesta del servicio no es solamente el nombre del tópico o frase clave, sino que devuelve un objeto JSON con los miembros nombre de tópico, lista de tesauros en las que fue encontrado y probabilidad promedio de éxito. Teniendo en cuenta la probabilidad promedio se pueden usar los términos con más probabilidades de ser correctos.

6.4.3. NER Reconocimiento de entidades nombradas

El reconocimiento de entidades con nombre es una tarea cuyo objetivo es identificar y clasificar expresiones de un texto que hacen referencia a personas, organizaciones, lugares, marcas comerciales e incluso fechas, horas y medidas (Felice, 2009). Dado un texto de entrada, el reconocedor de entidades nombradas devolverá un listado de “entidades” y su “tipo”.

Para este trabajo, hemos elegido dos herramientas para esto: Stanford NER y OpenNLP con la intención de analizar los resultados de ambas. Stanford NER es una implementación Java de un reconocedor de entidades nombradas realizado por el Stanford Natural Language Processing Group. OpenNLP es una herramienta de código abierto de la Apache Software Foundation que soporta las tareas de procesamiento de lenguaje natural más comunes. El reconocimiento de entidades nombradas reconoce secuencias de palabras en un texto que son "nombres de cosas" como personas, empresas, genes y proteínas. Incluye modelos en varios idiomas para su funcionamiento y así también el reconocimiento de ubicaciones.
Aquí tomaremos al reconocimiento de entidades nombradas como una de las formas en las que el enriquecimiento semántico del texto ocurrirá.

6.4.3.A. Stanford NER

6.4.3.A.i Clases reconocidas por Stanford NER

Stanford NER viene con modelos entrenados en inglés para 3, 4 y 7 clases en base a conjuntos de datos no públicos. Las clases distinguen la cantidad de "tipos" de entidades que reconocerá Stanford NER al ser ejecutado con un modelo en particular.

Tipo de modelo	Qué reconocerá
3 clases	Personas, organizaciones y lugares.
4 clases	Personas, organizaciones, lugares y "Otros"
7 clases	Personas, organizaciones, lugares, monedas, porcentajes, fechas y horas.

A medida que los modelos crecen en cantidad de clases estos requerirán más memoria y tiempo en el procesamiento. En español, desafortunadamente, Stanford NER ofrece solo un modelo de 4 clases incapaz de reconocer fechas, horas, monedas y porcentajes. Existen, además, modelos que reconocen también animales, elementos químicos e incluso existen también aquellos que soportan jerarquías de tipos, reconociendo más de 150 distintas clases pero estos no funcionan con Stanford NER.

6.4.3.A.ii. Entrenamiento de un nuevo modelo y elección del mismo

Para generar un nuevo modelo, es necesario contar con un conjunto de datos muy extenso. Este conjunto de datos es una planilla con dos columnas: Palabra y Tipo de palabra. Palabra se refiere a una entidad determinada, por ejemplo, "Universidad", "Facultad", "Empresa",

“Torres”, etc, etc. Y tipo de palabra será entonces, la clase de la palabra: “Organización”, “Persona”, “Lugar”, etc, etc.

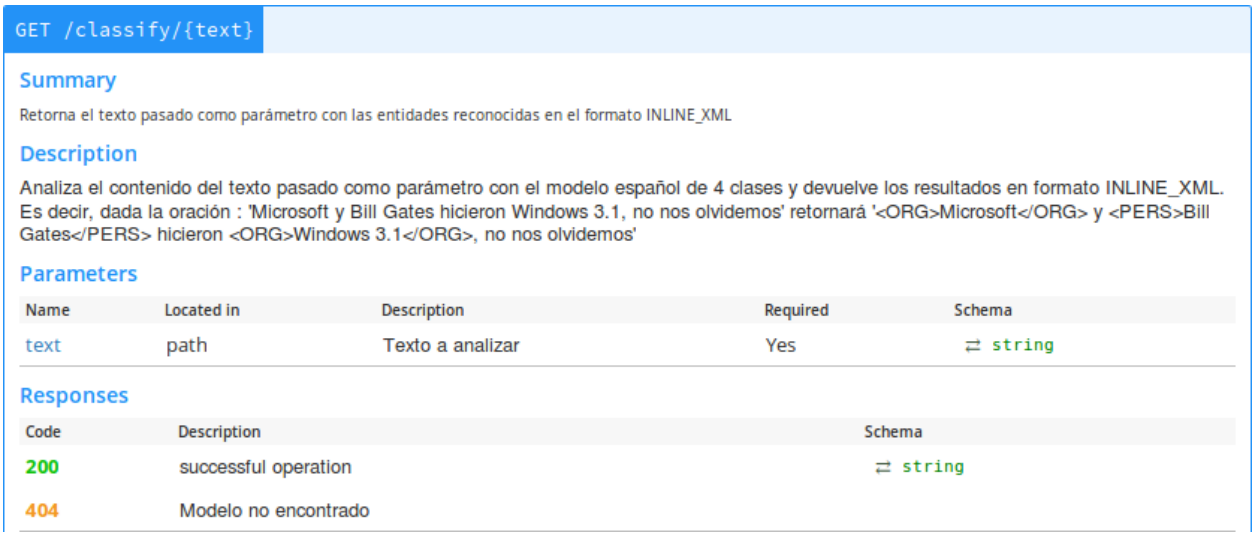
Dado nuestro lenguaje, sería conveniente poder generar un nuevo modelo que cuente con un número mayor de clases pero no contamos con el conjunto de datos necesarios para lograr un modelo mejor que el Stanford NER ofrece en español por lo que, en lo siguiente, usaremos ese modelo.

6.4.3.B. Un servicio para Stanford NER

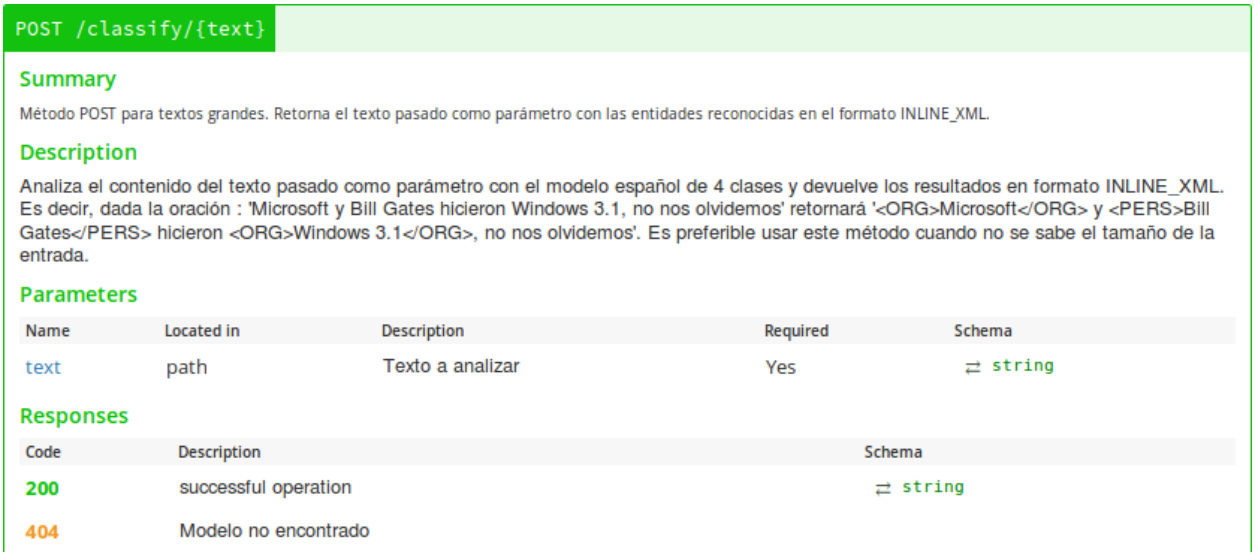
Para ser accedido desde cualquier CMS disponible, en nuestro caso Drupal, elegimos seguir el mismo camino que para KEA al detectar frases claves: un servicio web con una API REST en sus dos versiones GET y POST dado el tamaño de los datos.

La API que generamos es más simple que aquella de KEA pero muy similar.

6.4.3.B.i. GET /classify/{text}



6.4.3.B.ii. POST /classify/{text}



POST /classify/{text}

Summary

Método POST para textos grandes. Retorna el texto pasado como parámetro con las entidades reconocidas en el formato INLINE_XML.

Description

Analiza el contenido del texto pasado como parámetro con el modelo español de 4 clases y devuelve los resultados en formato INLINE_XML. Es decir, dada la oración : 'Microsoft y Bill Gates hicieron Windows 3.1, no nos olvidemos' retornará '<ORG>Microsoft</ORG> y <PERS>Bill Gates</PERS> hicieron <ORG>Windows 3.1</ORG>, no nos olvidemos'. Es preferible usar este método cuando no se sabe el tamaño de la entrada.

Parameters

Name	Located in	Description	Required	Schema
text	path	Texto a analizar	Yes	⇄ string

Responses

Code	Description	Schema
200	successful operation	⇄ string
404	Modelo no encontrado	

6.4.3.C. Open NLP

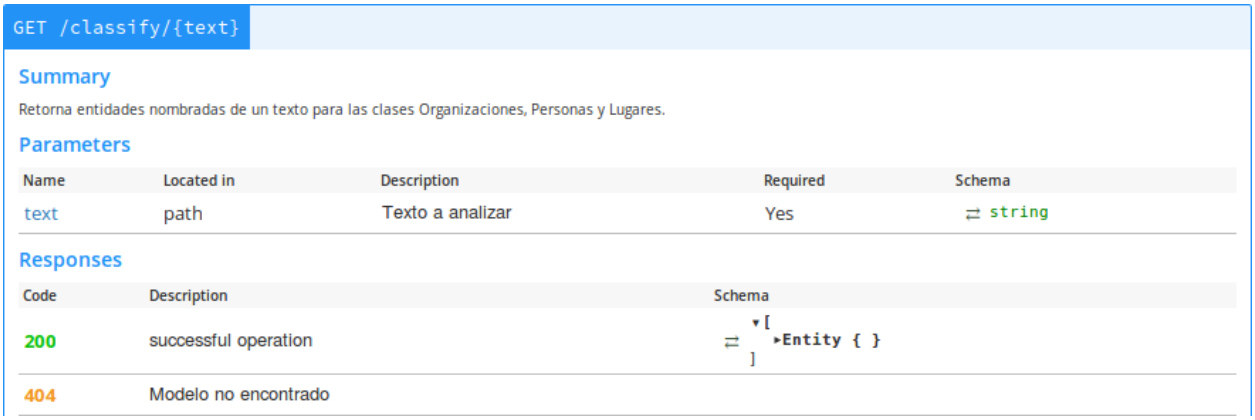
OpenNLP a diferencia de Stanford NER maneja un modelo distinto para cada clase. En español se ofrecen un modelo para personas, otro para localizaciones y otro para organizaciones. Estos modelos pueden ser generados a partir de un entrenamiento similar al de Stanford NER.

Al momento de usar OpenNLP desde la API Java es inmediato el hecho de que la librería está mejor organizada, más documentada y más simple de customizar. OpenNLP no ofrece en sus últimas versiones tokenizadores y detección de oraciones en español por lo que utilizamos el tokenizador y procesador de textos de Stanford NER, más específicamente usamos un fork de la librería que se encuentra en el repositorio de GitHub https://github.com/limves/stanford-ner-spanish que provee versiones en español para los tokenizadores.

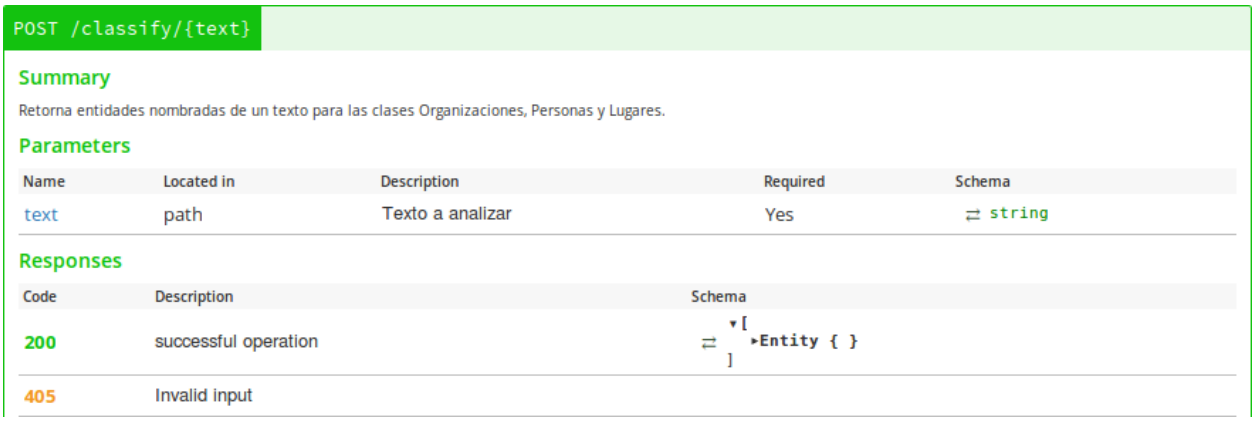
6.4.3.D. Un servicio para OpenNLP

Para el uso de esta librería seguimos el mismo camino que con las demás herramientas. Generamos un servicio REST para OpenNLP. La API se detalla a continuación:

6.4.3.D.i. GET /classify/{text}



6.4.3.D.ii. POST /classify/{text}



6.4.4. Motor de indexación distribuido: SOLR.

Solr es un motor de indexación altamente confiable, escalable y tolerante a las fallas, proporcionando indización distribuida, replicación y equilibrio de carga, automatización de recuperación de fallos y recuperación, configuración centralizada y mucho más. Solr potencia las funciones de búsqueda y navegación de muchos de los sitios de Internet más grandes del mundo.

Provee un motor altamente escalable, proveyendo un conjunto de funcionalidades extenso, sus características principales son:

- Permite realizar peticiones HTTP para indexar o consultar documentos: se podría decir que tiene un API estilo REST, aunque no hace uso de todos los verbos pero sí permite la recuperación de documentos en formato XML y JSON,
- Incluye cachés internas para devolver con mayor rapidez el resultado de las consultas,
- Incluye una administración web que permite: consultar estadísticas de rendimiento, incluyendo el uso de cache, realizar búsquedas mediante un formulario, navegar por

los términos más populares del índice, visualizar un desglose detallado de las matemáticas de puntuación y las fases de análisis de texto.

- Permite la configuración de la indexación y recuperación de documentos mediante ficheros de configuración xml: añade una librería de analizadores textuales a los que provee por defecto Lucene, introduce el concepto de campo tipado, lo que permite introducir fechas y mejorar la ordenación,
- Incluye navegación por facetas en las búsquedas,
- Dispone de un plugin de "spell check" o revisión gramatical, para realizar recomendaciones de búsqueda,
- Permite el manejo de documentos ricos (Word, PDF, …) basándose en el proyecto Apache Tika.
- Se encuentra en un servidor aparte, lo cual beneficia la configuración y la performance.
- Tiene la posibilidad de generar múltiples índices.

Su integración con Drupal es casi inmediata en las versiones 5.x del motor, solamente instalando un conjunto de módulos o "plugins" de drupal se puede extender la funcionalidad de búsqueda para que los contenidos se comienzan a indexar en SOLR proveyendo búsquedas facetadas prácticamente Out-Of-The-Box.

6.4.4.A. Aplicación en INTA

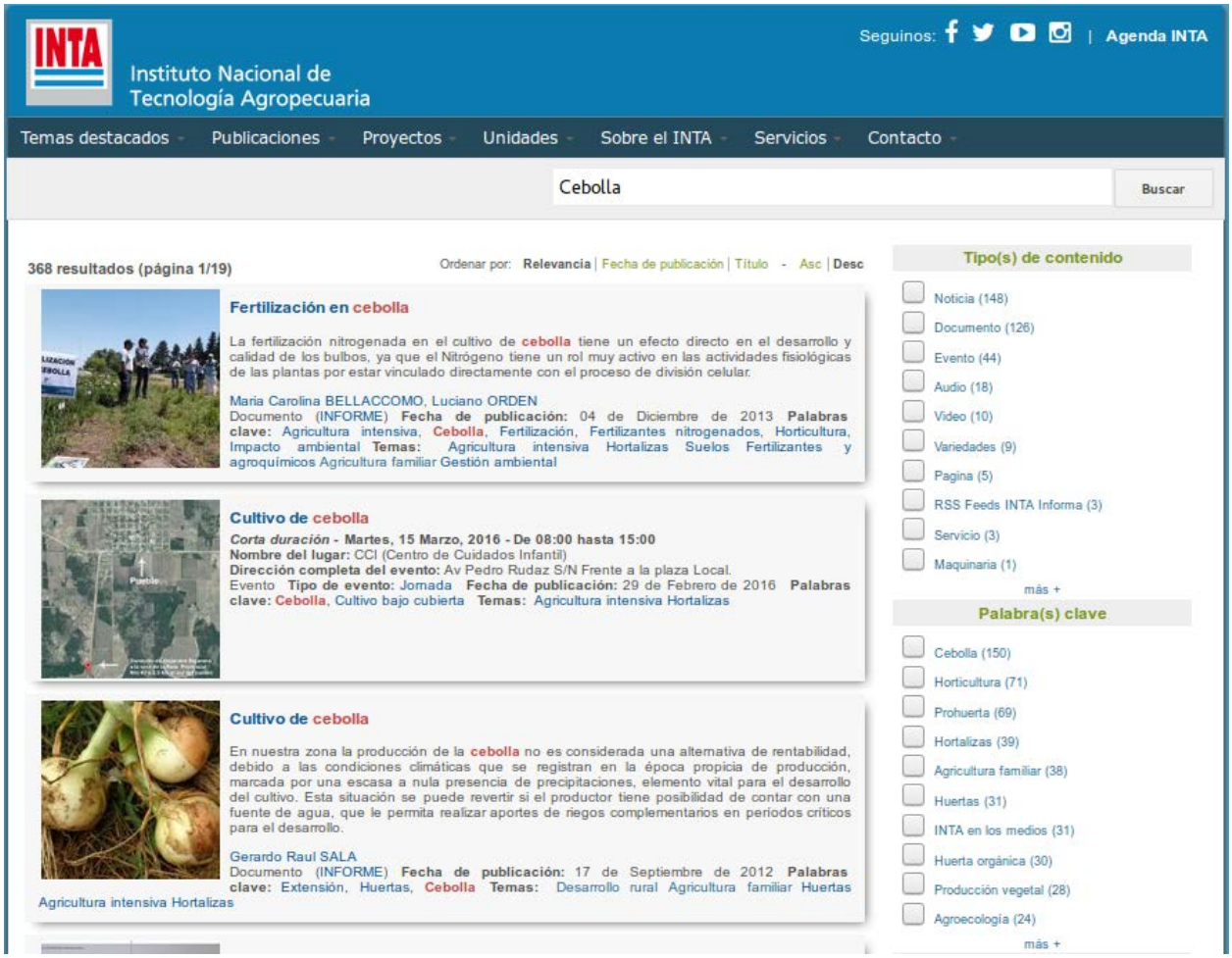
En un servidor separado del CMS, se instaló Apache SOLR y se configuró para poder utilizarlo desde la aplicación web.

En un trabajo conjunto con los webmasters del INTA, se generó un esquema donde definimos la estructura de datos del índice y cómo van a ser analizados los datos en el momento del indexado. Lo primero que se tuvo que definir en nuestra estrategia de indexado fue qué datos se van a almacenar en el índice y con qué objeto. Solr no es un motor de base de datos y no la sustituye, pero puede almacenar determinados datos de cada registro, evitando que se tenga que recuperar información desde la base de datos.

Habitualmente el resultado de la búsqueda se presenta como un listado que muestra una serie de datos del registro en cuestión. Independientemente de si un campo participa o no en el indexado, podemos almacenarlo en el índice y recuperarlo junto con los resultados de nuestras consultas.

6.4.4.A.i. Facetas y filtros

Las facetas responden a campos que nos permiten afinar la búsqueda sesgando el resultado con filtros o distribuyendo los resultados según vocabularios controlados. En otras palabras son filtros de búsqueda dinámicos que responden a la consulta del usuario y a los filtros aplicados actualmente.



En la imagen, se ve una búsqueda por la palabra “Cebolla” en el sitio del inta.gob.ar usando SOLR. Vemos a la derecha las facetas elegidas para ser mostradas, las palabras clave o tópicos detectados anteriormente son parte de las facetas elegidas.

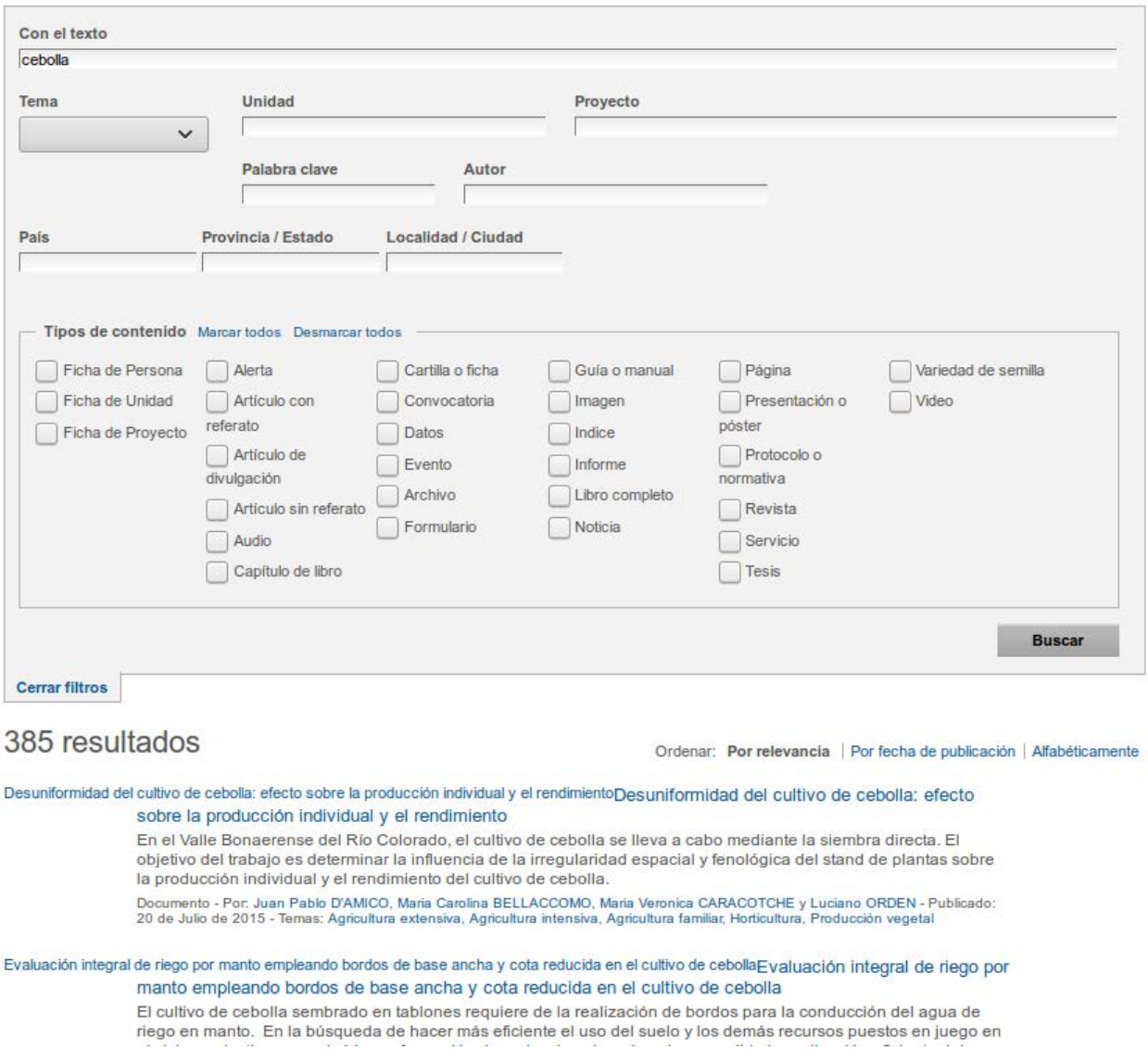
6.4.4.A.ii. Elección de campos a indexar

La indexación de campos en SOLR permite no determina la forma en que los resultados de búsqueda son mostrados. SOLR permite buscar sobre una colección de campos X y mostrar otros campos gracias a la adaptabilidad del módulo de búsqueda con Drupal. La elección de campos a usar fueron discutidos con personal del INTA. Al momento de realizar una búsqueda, los resultados se ordenan por relevancia. Esta relevancia está dada por en qué campos aparece el término buscado. Si el término se encuentra en el título o en el campo

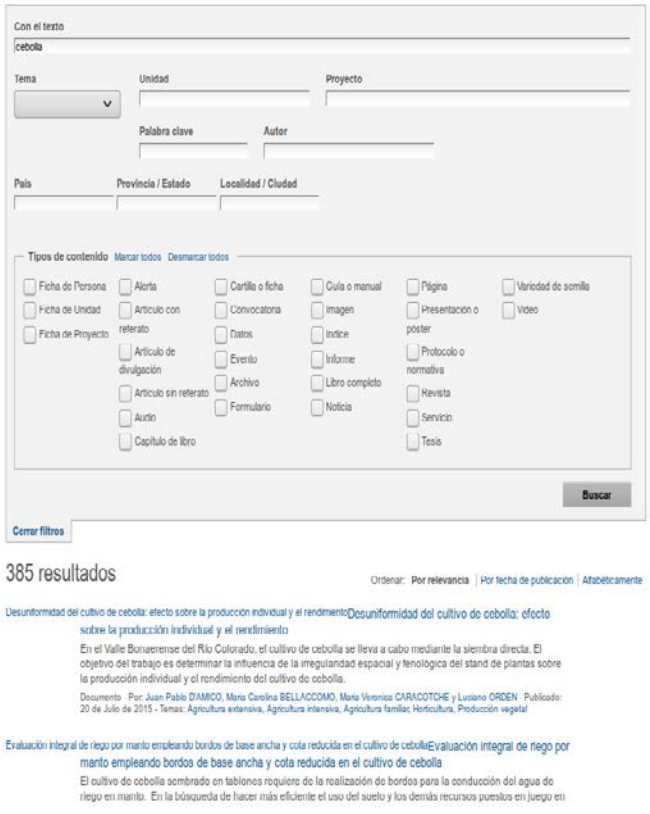
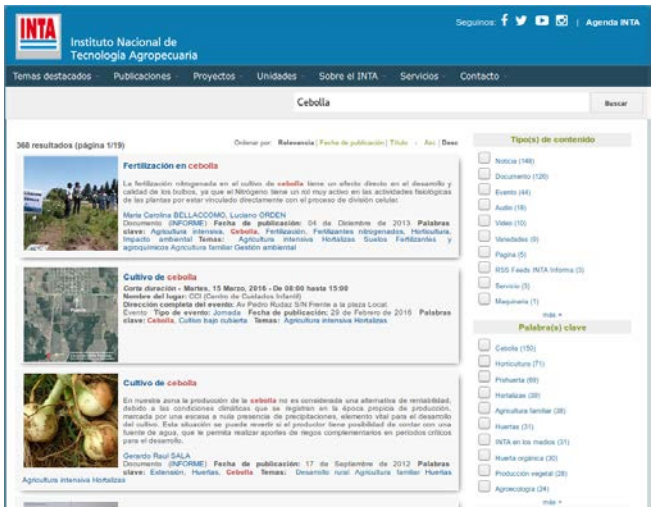
autor de un determinado nodo, este probablemente aparecerá en la cabeza de los resultados ya que tendrá un “ranking” alto de búsqueda.

6.4.4.B. Comparación con las búsquedas en la implementación previa

En lo siguiente veremos una búsqueda por la palabra “cebolla” en la implementación previa:



- Si comparamos con el nuevo esquema, se ve que no hace uso de una búsqueda facetada sino que más bien le permite elegir al usuario, por ejemplo, todas las palabras clave sin importar si tienen relación con el resultado de búsqueda obtenido.
- Los filtros no muestran una visión acotada por cantidad de coincidencias como en la búsqueda facetada.
- La experiencia de un usuario final que no pertenece al INTA es mala ya que debe saber con anterioridad el valor de los filtros para encontrar lo que realmente busca.

Búsqueda en el sitio anterior	*Búsqueda en el sitio nuevo*
	

6.5. BÚSQUEDA EXTERNA

6.5.1 Introducción

Como explicamos anteriormente Schema.org es un esquema de metadatos jerárquico creado en asociación con los principales buscadores para buscar, entre otras cosas, un standard común. Existen otros diccionarios como data-vocabulary.org pero normalmente se suele hablar de schema.org por 2 cosas fundamentalmente, cuenta con un gran número de elementos o entidades para poder marcar y por qué Google, Yahoo y Bing recomiendan el uso de schema.org frente a otros esquemas.

Para aplicar esto, utilizamos JSON-LD (JSON for Linked Data) una alternativa a los microdata, también soportada por schema.org, justamente porque los buscadores recomiendan esta forma de presentar los datos.

El mayor inconveniente a la hora de marcar datos estructurados con microdatos es su implantación ya que la definición del elemento se realiza dentro de las etiquetas HTML, lo que puede ser complicado si debemos tocar muchas partes del código. JSON-Ld soluciona este problema de una forma muy interesante introduciendo una estructura de datos javascript en un solo lugar sin mezclarlo con el código HTML, de una forma "limpia".
El W3C recomienda el formato JSON-LD desde el 16 de enero de 2014 y Google lo adopta como alternativa a los microdatos para poder implementarlo en sitios web.
Para lograrlo generamos un módulo en Drupal, que utilizando la configuración que ya provee el CMS, se puede asociar cada tipo de contenido, y sus campos con atributos de schema.org.
El módulo genera a partir de esta configuración un script en los headers del html indicando estas definiciones.

6.5.2. Aplicación en Drupal

Para lograr esto se implementó un módulo para Drupal, donde utilizando módulos existentes que permiten asociar cada campo de cada tipo de contenido con su respectivo metadato de schema.org. Primero, se configuraron todos los tipos de contenido para que cada campo tenga asociado una propiedad de schema.org. Una vez realizadas todas las asociaciones, el módulo implementado, produce la estructura de datos javascript necesaria en los headers del contenido HTML para conformar una representación JSON-LD conforme a Schema.org del dato que el usuario (o crawler) está visualizando. Con esta implementación al entrar al documento "CONTROL DE INSECTOS-PLAGA EN LA AGRICULTURA UTILIZANDO HONGOS ENTOMOPATOGENOS" en el sitio del INTA se incluye el siguiente esquema de metadatos:

```
<script type="application/ld+json">
{
    "@context": "http://schema.org",
    "@type": "creativeWork",
    "name": "Horizonte agropecuario pampeano - puntano N° 109 Septiembre de 2016",
    "url":            "unlp.tesis.com/documentos/horizonte-agropecuario-pampeano-puntano-no-109-
        septiembre-de-2016",
    "datePublished": "2016-09-30T03:00:00+0000",
    "text": "CONTROL DE INSECTOS-PLAGA EN LA AGRICULTURA UTILIZANDO HONGOS ENTOMOPATOGENOS: …
        ",
    "image": "http://unlp.tesis.com/sites/default/files/tapa_ha_109.jpg",
    "keywords": ["Insecta", "Insecticidas", "Homoptera", "Spodoptera frugiperda (Lepidóptera",
        "México (Estado)", "Departamento", "Raúl Rodríguez-Herrera", "Del Rincón", "Alas
        Marroquín", "Col", "Cultivos", "Plantas", "Plaguicidas", "Daños"],
    "contributor": [],
```

```
    "bookFormat": "http://unlp.tesis.com/tipos-documento/revista",
    "publisher": "Papel",
    "author":["http://unlp.tesis.com/personas/celdran.diego",
        "http://unlp.tesis.com/personas/coeli.marcos",
        "http://unlp.tesis.com/personas/terenti.oscar"],
    "description": "El Horizonte Agropecuario es una publicación de divulgación técnica
        institucional del Centro Regional La Pampa – San Luis del INTA [...]",
    "about":
        ["http://unlp.tesis.com/palabra clave/amaranto","http://unlp.tesis.com/palabra clave/in
        stitucionales", [...]]
}
</script>
```

Podemos ver en el JSON-LD, creado por el módulo implementado, que cada campo del tipo de dato Documento, es definido para el procesamiento de los buscadores, donde se define tipo (“creativeWork”), los autores del documento(“author”), la descripción (“description”), las palabras claves (“`keywords`”). Además, se puede ver que el contenido del campo “autores” no menciona el nombre y apellido del autor, sino más bien, la URL de la página del autor que contendrá el conjunto de metadatos que a él lo representan.

De ahora en adelante, “Horizonte agropecuario pampeano - puntano Nº 109 Septiembre de 2016”, será identificado por un buscador externo como un trabajo trabajo creativo cuyos autores, en vez de ser un nombre como lo sería usando Dublin Core, es una URI a un tipo “Persona”, en donde se definen completamente sus datos semánticos construyendo así un conjunto de “nodos” en un grafo de conocimiento interconectados explícitamente. A medida que el sitio adopte el uso de schema.org y valores de atributos como URI (Schema.org también podría aceptar cadenas de caracteres como “Celdran Diego”) el grafo de conocimiento crece, tanto en nodos como en vínculos ya que, hasta las palabras clave, tienen una clase que las representa en Schema.org.

6.5.2.A. Buscadores externos y Schema.org

El markup de datos estructurados describen “Cosas” en la web y conjunto de propiedades para cada una de ellas. Tomemos el caso de una “Personas” en el INTA. Una persona en el INTA produce contenido en el INTA y por ende en su sitio web. Estas personas son parte de las búsquedas diarias que los usuarios realizan. Brindando esta información estructurada con Schema.org, se da a la posibilidad de que un buscador “entienda” el contenido que está indexando. Sabrá que, por ejemplo, la página http://inta.gob.ar/personas/celdran.diego representa a una persona, sabrá su nombre, su imagen, tendrá una descripción sobre como es

ella e incluso sabrá también como mostrar un breadcrumb list para mostrar en sus resultados de búsqueda.

Google, provee herramientas, documentación y, además, un analizador de datos estructurados para nuestros sitios y así ayudar a la expansión de la web semántica. En la imagen se ve que qué entiende Google por la página personal de Celdrán Diego en el INTA.

6.5.3. Aplicación Dublin Core

Drupal provee esta funcionalidad incluida en módulos en estado estable. Utilizando el módulo "RDF" Drupal permite, a cada campo de cada tipo de contenido en Drupal asignarle su propiedad DC. De esta forma pudimos asignar, por ejemplo, las siguientes asociaciones:

Campo Drupal	**Término DC**
dc:title	Campo título
dc:subject	Palabras y frases clave, tanto aquellas asignadas por el autor como aquellas asignadas por los módulos de tagging automático.
dc:creator	Autor(es) del contenido
dc:date	Fecha de publicación (si existe)
dc:description	Resumen del contenido.

Debemos tener en cuenta que Dublin Core es extensamente usado por muchos productos. Prácticamente todos los motores de búsqueda lo soportan, los repositorios de contenido caracterizan sus documentos con metadatos DC (DSpace es un producto de código abierto que provee herramientas para su uso como repositorio de colecciones digitales) lo que permite una gran interoperabilidad entre repositorios de información.

6.6 MÓDULO DE CONEXIÓN CON LOS WEB SERVICES DE KEA, NER Y OPENNLP Y ETIQUETADO AUTOMÁTICO

6.6.1. Introducción

Teniendo los servicios web implementados para Stanford NER, OpenNLP y KEA, se debe desarrollar un módulo que permita la extracción de entidades y palabras claves desde los formularios de edición y creación de contenido y de, también, los PDF ya cargados en el sistema.

6.6.2. Módulos KEA Client, NER Client y OpenNLP Client

Estos dos módulos desarrollados son una capa de abstracción a los servicios web que permiten realizar las consultas al servicio mediante sólo una llamada encapsulando detalles de implementación y abstrayendo de la comunicación.

6.6.3. Módulo Entity AutoTagger

Es un módulo en PHP para Drupal 7.x que, una vez configurado con los endpoints de los servicios antes nombrados provee dos funcionalidades:

1. Un tipo de control para un campo asociado a una taxonomía de Drupal que le permite al usuario, una vez escrito el cuerpo de un tipo de contenido, obtener un listado de "sugerencias" de palabras claves. Estas palabras clave se muestran organizadas por "Tópicos", "Organizaciones", "Personas", "Lugares" de acuerdo a lo que devuelvan los servicios permitiéndole al usuario elegir las palabras que desee de este listado. Las palabras clave que elija el usuario son guardadas en un campo configurable por el usuario desde las propiedades del módulo.
2. Provee una tarea que permite recorrer todos los contenidos con un archivo PDF, extraer el texto del mismo y guardarlo en un campo asociado al tipo de contenido separado de los demás términos, también configurable.

Como se puede observar en el gráfico, el módulo Autotagger se encarga de utilizar los servicios para reconocimiento de entidades y de sugerencias de frases claves. El procedimiento es el siguiente:

6.6.3.A. Sugerencia de tópicos automática

Tomemos como ejemplo un contenido de la web actual y le aplicamos los servicios generados en este trabajo de tesis para poder comparar los resultados, elegimos el siguiente contenido.

URI: "inta.gob.ar/documentos/sanidad-en-las-colmenas-nosema-varroa-y-virosis"

Cuerpo: La Argentina es el tercer productor de miel del mundo, aunque la apicultura es una actividad económica complementaria para mucho de los apicultores de nuestro país. Por otro lado es el segundo exportador de miel, dado que exporta el 95% del producto cosechado.

Esta última característica hace que esta actividad productiva contribuya, significativamente, a las economías regionales.

Estos datos muestran la importancia de estar atentos a las amenazas que significan los problemas sanitarios, como así también la reducción de la biodiversidad, por el avance de la frontera agrícola.

Dentro de los problemas sanitarios, aún no resueltos, se destacan la infección con el microsporidio *Nosema* sp., infestación con el ácaro *Varroa destructor,* las virosis y sus interacciones.

En este contexto, la Dra. Natalia Bulacio Cagnolo, profesional del grupo de apicultura de la EEA Rafaela y coordinadora del módulo Salud de las Abejas del Proyecto Específico "Estrategias multidisciplinarias para mitigar el efecto del nuevo contexto ambiental y productivo sobre la colmena" del Programa Nacional Apícola del INTA, comentó, "*para tener información que permita dar respuesta a esta problemática, se diseñó un estudio epidemiológico con el objetivo de estimar la prevalencia de las principales patologías en diferentes eco-regiones, e identificar las variables claves asociadas a su presencia y difusión conjunta*".

La profesional explica que durante el otoño 2015, con la colaboración de técnicos de INTA (PROAPI) y del ministerio de Formosa, se tomaron muestras en 385 colmenas, distribuidas en 64 apiarios, de las provincias de Santa Fe, Chaco y Formosa. A estas provincias se las dividió en cinco regiones agroecológicas diferentes: Sur de Santa Fe, Centro de Santa Fe, Chaco Húmedo, Chaco de Transición y Chaco Semi-árido.

La técnica comentó que, en cada apiario se tomaron muestras en el 10% de las colmenas, o un mínimo de seis en aquellos con menos de 60 colmenas. En cada colmena se registró la población de abejas, área con cría, con polen y con miel. También se tomaron muestras para determinar el porcentaje de Varroa forética, el recuento de *Nosema* sp. y la prevalencia de Virus. Los monitoreos se realizaron pre y pos tratamiento para Varroa e inicio de temporada, dependiendo de las condiciones de cada ecorregión.

Por otra parte, simultáneamente, se realizó una encuesta semi-estructurada a los apicultores, recopilándose datos relacionados a número de colmenas, tipo de actividad, práctica de transhumancia, origen y recambio de reinas, suplementos alimenticios aplicados, tratamientos para Varroosis y Nosemosis y tipo de invernada realizada, entre otras preguntas.

Haciendo referencia a los resultados del primer monitoreo, Bulacio comentó "*la media de infestación con Varroa fue 7,12% por colmena. La menor infestación se registró en Chaco Semi-árido (3,01%), mientras que la mayor infestación se concentró en el sur de Santa Fe (10,31%)*".

La profesional manifestó que el recuento promedio de Nosema sp. fue de 341.851 esporos por abeja, y que las regiones chaqueñas presentaron menores recuentos; en comparación con las dos regiones de Santa Fe. Con respecto a los virus, el de mayor prevalencia fue DWV (virus de alas deformes en sus siglas en inglés) (35%), seguido por ABPV (virus de la parálisis aguda) (21,5%), BQCV (virus de las celdas negras reales)(8%), CBPV (virus de la parálisis crónica) (2,2%) y SBV (virus de la cría ensacada)(1,1%).

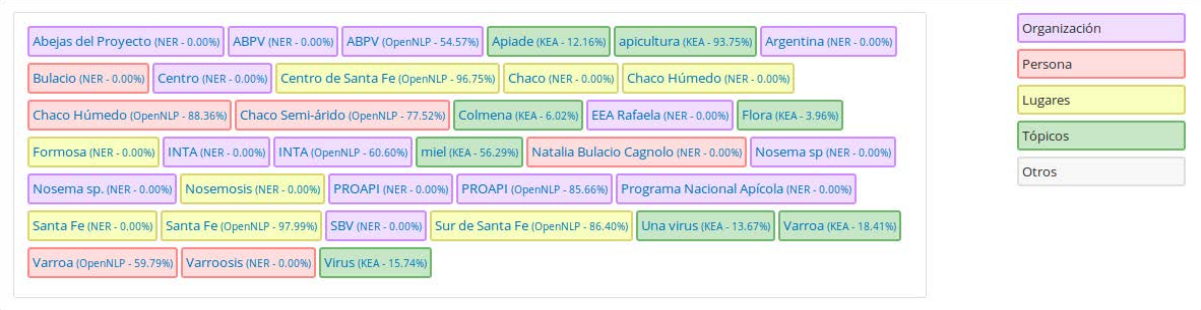
Bulacio resaltó que "*es difícil magnificar la importancia que este trabajo significa, no solo por los recursos humanos involucrados, que sin ellos hubiera sido prácticamente imposible realizar este trabajo, como así también la logística y los recursos involucrados*". Explicó que estos primeros resultados dan la certeza que el manejo adecuado de las colmenas, de acuerdo a las prácticas que se recomiendan en el sendero tecnológico propuesto por el PROAPI, son fundamentales para mantener estas enfermedades por debajo de un umbral de daño económico, he hizo hincapié, en la relevancia de la disponibilidad de flora de importancia apícola para el buen desarrollo de las colonias.

"*La región donde se encuentran ubicados los apiarios es uno de los factores relacionados a la presencia tanto de Nosemosis, Varroosis y ABPV. Dicho factor se encuentra influenciado por la disponibilidad de*

flora de importancia apícola y por las prácticas de manejo utilizadas por los apicultores en cada región. El recambio de reinas, y la frecuencia de recambio, determinan el desempeño general de las colmenas, disminuyendo el desarrollo de las parasitosis. También se verificó la presencia conjunta de Nosemosis y Varroosis en las diferentes regiones, así como la asociación entre Varroosis y los principales virus de importancia apícola" concluyó Bulacio.

Palabras Claves: Abejas, Abeja Reina, Apicultura, Miel, Sanidad Animal.

Cuando el usuario, luego de escribir el texto, hace click en "Sugerir términos", se muestra lo siguiente:



En la imagen se muestran las entidades y tópicos reconocidos por KEA, Stanford NER y OpenNLP. Los términos se encuentran colorizados según el tipo reconocido por los servicios. Con fines evaluativos, dentro de cada término se incluye la librería que arrojó el resultado y el porcentaje de probabilidad de éxito. Además, para aportar a este fin, no se eliminaron duplicados.

6.6.4. Evaluación de los resultados obtenidos por el servicio de KEA

En el ejemplo anterior, el recuadro con términos en color, aparecen representadas las entidades y palabras claves reconocidas. En verde, podemos ver los tópicos sugeridos por KEA. Si revisamos en detalle los resultados arrojados por KEA y las palabras clave asociadas manualmente, comprobaremos que da un excelente resultado: en el contenido existían 5 palabras claves asociadas manualmente por los webmasters del INTA:

- Abejas,
- Abejas Reinas,
- Apicultura,
- Miel,
- Sanidad Animal

El servicio utilizando de KEA con el modelo entrenado por nosotros con documentos del INTA, sugiere los tópicos:

- Apidae (12,16%)
- Apicultura (93,75%)
- Colmena (6,20%)
- Flora (3,96%)
- Miel (56,29%)
- Virus (15,74%)
- Varroa (18,41%)
- Una virus (13,67%).

Coinciden 4 de las 5 palabras claves asignadas manualmente. Esto es así porque tanto "Abejas" como "Abejas Reinas" son representadas por "Apidae" ya que a través del tesauro fueron unificadas en "Apidae" que, según Wikipedia, significa: "Los ápidos (Apidae) son una familia de himenópteros apócritos; constituyen un numeroso grupo de abejas que incluye a la abeja melífera o doméstica (la más conocida), a las abejas sin aguijón, las abejas de las orquídeas, las abejas parásitas, los abejorros y abejorros carpinteros además de otros grupos menos conocidos.". El servicio KEA mejora las asignaciones manuales, ya que agrega como tópicos sugeridos las palabras:

- Colmena
- Flora
- Virus
- Varroa
- Una Virus

Palabras claves aún más específicas sobre el tema tratado en este contenido. Imaginando un escenario, donde un apicultor, ingrese en el sitio de INTA o a través de un buscador como Google se encuentre buscando información sobre el virus Varroa en abejas, encontraría resultados más acordes a lo buscado, porque de no ser por estos tópicos sugeridos, sólo encontraría el contenido por relación con abeja (junto con muchos otros contenidos que traten de este tema) y no con el refinamiento logrado con este etiquetado automático. Este contenido, estará mejor ponderado por ser específico sobre el tema. Dado que KEA utiliza modelos predictivos para encontrar palabras claves también arroja errores al reconocer palabras, en este caso, el tópico "una virus" es uno de ellos. Sin embargo, encontramos que el resultado general de KEA es más que satisfactorio.

6.6.5. Evaluación y comparación de los resultados obtenidos para Stanford NER y OpenNLP

6.6.5.A. Caso de prueba I

Estudio del perfil de producción de toxinas en trigo inoculado artificialmente con Fusarium graminearum

http://inta.gob.ar/documentos/simulacion-de-degradacion-de-plaguicidas-bajo-condiciones-de-almacenamiento-de-cereales-de-invierno

Por Dante Emanuel Rojas, María José Martinez, Gaston PELISSIER, Alba Rocío Castro, Susana Rojas de Maitia y Diego Sebastian Cristos.
En Argentina la producción de cereales y oleaginosos continúa siendo de gran importancia económica frente a la creciente demanda mundial de alimentos y bioenergía. Se espera un aumento de la producción nacional para 2020, alcanzando las 157 millones de toneladas. Entre los cereales de invierno se destaca el trigo, por su producción de hasta 16,5 millones de

Standford NER

Personas

Alba Rocío Castro	CORRECTO
Cromatografía Gaseosa	INCORRECTO
Diego Sebastian Cristos	CORRECTO
Gaston PELISSIER	CORRECTO
Maitia	INCOMPLETO
María José Martinez	CORRECTO
Por Dante Emanuel Rojas	INCORRECTO
Susana Rojas	INCOMPLETO

Organizaciones

Centro	INCORRECTO
Fusarium graminearum	INCORRECTO

toneladas anuales y su aporte nutricional a los hábitos de consumo de la población. El usos de agentes fitosanitarios es una herramienta muy utilizada por el sector agropecuario para mejorar los rendimientos, evitando pérdidas ocasionadas por el ataque de plagas. Sin embargo, el uso inadecuado de agentes fitosanitarios puede provocar efectos negativos en la salud, el medio ambiente y las exportaciones. Los Límites Máximos de Residuos (LMR) son establecidos para cada agente fitosanitario y es la concentración máxima permitida legalmente en alimentos. Los LMR determinados en base a fundamentos científicos, agronómicos y toxicológicos, teniendo como finalidad la preservación de la salud de los consumidores. El objetivo del trabajo fue evaluar la dinámica de disipación de los agentes fitosanitarios aplicados durante el almacenamiento de granos de trigo considerando las recomendaciones de uso en la República Argentina. El trabajo de simulación de almacenamiento post-cosecha, se llevó a cabo en el Laboratorio de Contaminantes Químicos (Instituto Tecnología de los Alimentos dependiente del Centro de Investigación de Agroindustria, INTA) utilizando muestras de grano de trigo tratado con productos comerciales cuyos principios activos fueron pirimifós-metil, diclorvós, clorpirifos-metil; clorpirifos-etil, cialotrina, ciflutrina. Luego del tratamiento químico, las muestras fueron incubadas a temperatura ambiente y a 40 °C. El análisis de los residuos fue realizado por un método de extracción de fase sólida dispersa empleando Cromatografía Gaseosa acoplada a un detector de espectrometría de masas cuadrupolar (GC-MS). La técnica analítica empleada fue evaluada parámetros de linealidad, precisión, exactitud y robustez. Los porcentajes de desaparición de residuos y las tasas de disipación fueron diferentes para cada principio activo, dosis aplicadas y la temperatura de incubación.

Instituto Tecnología	**INCORRECTO**
Laboratorio	**INCORRECTO**
LMR	**INCORRECTO**
Límites Máximos	**INCORRECTO**

Lugares

Argentina	**CORRECTO**
República Argentina	**CORRECTO**

OpenNLP

Personas

Alba Rocío Castro (85,23%)	**CORRECTO**
Diego Sebastian Cristos (96,87%)	**CORRECTO**
María José Martinez (87,09%)	**CORRECTO**

Organizaciones

Centro de Investigación de Agroindustria (90,90%)	**CORRECTO**
Instituto Tecnología de los Alimentos (77,78%)	**CORRECTO**
INTA (71.25%)	**CORRECTO**
LMR (68.13%)	**INCORRECTO**

	Stanford NER	**OpenNLP**	**AVG (% OpenNLP)**
Correctas	6 (37,50%)	6 (85,70%)	84,85%
Incorrectas	8 (50%)	1 (14,30%)	68,13%
Incompletas	2 (12,50%)	0	

6.6.5.B. Caso de prueba II

"Avanzamos en la instalación de un laboratorio de triquinosis"

http://inta.gob.ar/noticias/avanzamos-en-la-instalacion-de-un-laboratorio-de-triquinosis

Lo indicó Juan Manuel Bracci, director de Producción Rural y Actividades Agropecuarias del municipio, en la primera reunión de 2017 de la Mesa de Desarrollo Territorial. Participa INTA AMBA.

Por Federico Gaston GUERRA.

En una jornada con distintos objetivos propuestos para 2017 resaltó la iniciativa conjunta entre las autoridades del municipio de San Vicente, el Ministerio de Agroindustria de la Nación y los representantes de la Mesa de Desarrollo Territorial del municipio, de la que participa INTA AMBA, para la instalación de un laboratorio para análisis de triquinosis.
Juan Manuela Bracci, director de Producción Rural y

Stanford NER

Personas

Bracci	**CORRECTO**
Federico Gaston	**CORRECTO**
Gustavo Tito	**CORRECTO**
Juan Manuela Bracci	**CORRECTO**
Juan Manuel Bracci	**CORRECTO**
Miriam Plana	**CORRECTO**

Organizaciones

Agencia	**INCORRECTO**
Agricultura Familiar del Ministerio	**INCOMPLETO**
Agricultura Familiar del Servicio Nacional	**INCOMPLETO**
Cooperativa	**INCORRECTO**
Cooperativa Apícola San Vicente	**CORRECTO**
CTR	**CORRECTO**
Desarrollo Territorial del Ministerio	**INCOMPLETO**
Dirección	**INCORRECTO**
Escuela Agraria N° 1	**CORRECTO**

Actividades Agropecuarias del municipio de San Vicente, afirmó que este proyecto, que suma aportes de Dirección de Gestión Territorial de la Subsecretaría de Desarrollo Territorial de la Nación, “fue gestado a partir de la necesidad de finalizar un laboratorio que permita realizar el análisis de triquinosis gratuito. Es de destacar que no se cuenta con ningún otro con estas características.”.

Esta iniciativa tiene como objetivo la prevención en salud de la población para minimizar la incidencia de triquinosis, enfermedad zoonótica que se trasmite cuando se ingiere carne de cerdo infectada que no ha sido convenientemente cocida. En suma se busca mejorar la calidad de alimentos que consume la población, destaca el municipio.

Con la puesta en marcha de un laboratorio local se contribuye a disminuir la incidencia del parásito en los chacinados debido a que los productores – elaboradores pueden acceder de manera sencilla y práctica al mismo.

Nuevas metas

En esta reunión, además, se plantearon metas a corto, mediano y largo plazo para planificar los pasos a seguir en 2017. “Fue muy interesante el hecho de compartir las prioridades y posibilidades que tenemos como sector productivo”, dijo Bracci quien

Desde la Mesa se coincidió en analizar otras propuestas vinculadas con mejoras tanto para la Cooperativa apícola como maquinaria rural para el sector de la agricultura familiar del distrito.

Miriam Plana, de la Secretaría de Extensión de la Facultad de Ciencias Agrarias de la UNLZ, indicó que “estos proyectos son muy importantes ya que ayudan y contribuyen al crecimiento productivo y posibilitan generar empleo donde más se lo necesita. Sin dudas son aportes concretos al bien común y general”.

En esta línea Gustavo Tito, director del INTA AMBA, destaca que las Mesas de trabajo son muy importantes para el INTA AMBA. “Por esto participamos en varias en el área metropolitana. Nuestra tarea es asesorar técnicamente siguiendo los lineamientos que propone cada municipio junto a los organismos, pero los protagonistas deben ser los productores”.

Sumar

Esta Mesa está compuesta por organizaciones de productores e instituciones que incluyen el municipio de San Vicente, la Subsecretaria de Desarrollo Territorial del Ministerio de Agroindustria de la Nación, la Oficina de Desarrollo Local INTA AMBA SUR de Lomas de Zamora y Agencia de Extensión de San Vicente del Instituto Nacional de Tecnología Agropecuaria (INTA), la Secretaria de extensión de la Facultad de Ciencias Agrarias de la Universidad Nacional de Lomas de Zamora (FCA-UNLZ), la Coordinación de Agricultura Familiar del Servicio Nacional de Sanidad y Calidad Agroalimentaria (CAF-SENASA), el Instituto Nacional de Asociativismo y Economía Social (Inaes), la Escuela Agraria N° 1 y la Universidad Nacional de La Plata (UNLP).

Son también parte de esta propuesta la Unión de Trabajadores de la Tierra (UTT), Cooperativa de Trabajadores Rurales (CTR) de San Vicente “Bartolina Sisa” Frente Popular Darío Santillán-Corriente Nacional, Sociedad Rural de San Vicente, Cooperativa Apícola San Vicente, Instituto Agrotécnico “San Jose” y la Dirección de Desarrollo Rural y Agricultura

Facultad	INCORRECTO
Frente Popular Darío Santillán-Corriente Nacional	C ORRECTO
Instituto Agrotécnico	INCOMPLETO
Instituto Nacional	INCOMPLETO
INTA AMBA	CORRECTO
Mesa	INCORRECTO
Mesas	INCORRECTO
Ministerio	INCORRECTO
Oficina	INCORRECTO
Secretaria	INCORRECTO
Secretaría	INCORRECTO
Sociedad Rural	INCORRECTO
Subsecretaria	INCORRECTO
Subsecretaría	INCORRECTO
Sumar Esta Mesa	INCORRECTO
Universidad Nacional	INCORRECTO
Unión	INCORRECTO
UNLZ	CORRECTO
Zamora	INCORRECTO

Lugares

Buenos Aires	CORRECTO
San Jose	CORRECTO
San Vicente	CORRECTO

OpenNLP

Personas

Bartolina Sisa (82.91%)	CORRECTO
Bracci (95.13%)	CORRECTO
Calidad Agroalimentaria (93.45%)	INCORRECTO
Gustavo Tito (75.80%)	CORRECTO
Juan Manuela Bracci (97.01%)	CORRECTO
Juan Manuel Bracci (97.45%)	CORRECTO

Organizaciones

Agencia de Extensión de San Vicente (91,66%)	CORRECTO
Cooperativa (62.95%)	CORRECTO
Cooperativa Apícola San Vicente (92.54%)	CORRECTO
CTR (61.83%)	CORRECTO
Escuela Agraria N (94.32%)	CORRECTO
Facultad de Ciencias Agrarias de la Universidad Nacional de Lomas de Zamora (91.79%)	CORRECTO
Facultad de Ciencias Agrarias de la UNLZ (86.16%) CORRECTO	
Frente Popular Darío Santillán-Corriente Nacional (88.53%)	CORRECTO
Instituto Agrotécnico (89.23%)	INCOMPLETO
Instituto Nacional de Asociativismo y Economía Social (91.23%)	CORRECTO
Instituto Nacional de Tecnología Agropecuaria (99.11%)	CORRECTO
INTA (83.21%)	CORRECTO
INTA AMBA (71.17%)	CORRECTO
Mesa (63.91%)	INCORRECTO
Mesa de Desarrollo Territorial (87.97%)	CORRECTO
Ministerio de Agroindustria (94.48%)	CORRECTO
Oficina de Desarrollo Local (90.26%)	CORRECTO
Servicio Nacional de Sanidad (92.10%)	CORRECTO
Sociedad Rural de San Vicente (94.86%)	CORRECTO

Familiar del Ministerio de Agroindustria de la provincia de Buenos Aires.

Universidad Nacional de La Plata (97.28%)	CORRECTO
Universidad Nacional de Lomas de Zamora (97.43%)	CORRECTO
UNLP (84.90%)	CORRECTO
UNLZ (69.01%)	CORRECTO
UTT (56.68%)	CORRECTO

Lugares

Buenos Aires (99.49%)	CORRECTO
San Jose (80.57%)	CORRECTO
San Vicente (91.69%)	CORRECTO
Zamora (85.00%)	INCOMPLETO

	Stanford NER	OpenNLP	AVG (% OpenNLP)
Correctas	15 (40,50%)	30 (88,24%)	86,65%
Incorrectas	17 (45,95%)	2 (5,88%)	78.68%
Incompletas	5 (13,55%)	2 (5,88%)	87.11%

6.6.5.C. Caso de prueba III

Mujeres que hacen pueblo y tienen proyectos
http://inta.gob.ar/noticias/mujeres-que-hacen-pueblo-y-tienen-proyectos

El grupo de mujeres rurales de la localidad de Ochandio que trabaja junto con el grupo de Desarrollo Territorial de la CEI Barrow realizó la primera reunión del año con una buena presencia de asistentes.

Por: Emiliano SOFIA

El grupo “Mujeres que hacen pueblo” de la localidad de Ochandio realizó su primera reunión del año con una muy buena convocatoria que incluyó representantes de la Comisión de Damas del Club Ochandio y el CEPT Nº 34. Además, participaron Daniel Intaschi y Soledad González Ferrín, del área de Desarrollo Territorial de la CEI Barrow.
En el encuentro se trató la posibilidad de financiar la construcción de una cocina comunitaria a través de un proyecto especial de INTA-PROHUERTA. Se instalaría en instalaciones del Club de la localidad, a cargo de la Comisión de Damas, en función a las necesidades de capacitación y de contar con un espacio habilitado a nivel municipal realizar emprendimientos generados por ellas y la comunidad.
A través de esto, se acodó contactar a representantes de Bromotalogía y de aspectos jurídicos de la Municipalidad de San Cayetano para que informen sobre los requerimientos necesarios para su habilitación.
Se consensuó, también, la realización de los avances por etapas, en una primera instancia contar con un lugar azulejado y piletas óptimas. En tanto para una segunda etapa se ideó instalar horno industrial autoclaves y demás artefactos vinculados a la elaboración.
Por otra parte, se acordó diferentes temas para trabajar a lo largo del año que estará relacionados con las capacitaciones en la elaboración de alimentos (a cargo del

Stanford NER

Personas

Daniel Intaschi	CORRECTO
Soledad González Ferrín	CORRECTO

Organizaciones

CEI Barrow	CORRECTO
Centro	INCORRECTO
CEPT	INCOMPLETO
Comisión	INCORRECTO
Desarrollo Territorial	INCORRECTO
INTA-PROHUERTA	CORRECTO

Lugares

Bromotalogía	INCORRECTO
Ochandio	CORRECTO

OpenNLP

Personas

Daniel Intaschi (96.32%)	CORRECTO
Soledad González Ferrín (86.11%)	CORRECTO

Organizaciones

Centro de Formación Profesional (90.34%)	CORRECTO
Desarrollo Territorial de la CEI Barrow (83.18%)	CORRECTO
Mujeres (53.35%)	INCORRECTO

Centro de Formación Profesional), organización interna del grupo, entre otros temas vinculados al desarrollo de la localidad, a cargo del CEPT Nº 34 y de la CEI Barrow.	**Lugares** Ochandio (79.98%) CORRECTO

	Stanford NER	OpenNLP	AVG (% OpenNLP)
Correctas	5 (50%)	5 (82,5%)	87,19%
Incorrectas	4 (40%)	1 (17,5%)	53,35%
Incompletas	1 (10%)	0	

6.6.5.D. Conclusión de las pruebas entre Stanford NER y OpenNLP

De acuerdas a las pruebas desarrolladas en los puntos 6.6.5.B, 6.6.5.C y 6.6.5.D OpenNLP ha demostrado tener un porcentaje promedio de éxito mucho mayor al de Stanford NER: un 85,48% contra un 42,67%. Stanford NER suele arrojar un número de resultados mucho mayor al de OpenNLP pero supera el 50% de entidades reconocidas erróneamente. En cambio OpenNLP, arroja una cantidad menor de resultados, pero casi todos ellos correctos. Esto habilita a OpenNLP a ser usado sin supervisión humana en un etiquetado automático de documentos.

6.6.6 Tarea de reconocimiento de palabras claves y entidades en adjuntos

El mismo módulo provee una funcionalidad extra en forma de una tarea de Drupal donde se sugieren palabras analizando los archivos adjuntos en formato PDF de los contenidos. Por cada contenido, analiza el texto de cada archivo adjunto y sugiere palabras claves y entidades nombradas. Las palabras sugeridas, quedan consistentes en otro campo del contenido que luego es indexado.

Con ambas funcionalidades se logra un análisis automático completo del contenido extrayendo tanto palabras claves como entidades nombradas sobre el cuerpo del contenido y sus archivos adjuntos. El análisis completo del contenido es de gran importancia ya que es muy común en el sitio del INTA que documentos subidos en el sitio web posean archivos adjuntos con investigaciones, tesis, artículos, etc. En estos casos la riqueza del contenido se encuentra en el archivo adjunto y no en el cuerpo HTML del contenido.

Como explicamos anteriormente los webmasters encargados de seleccionar las palabras claves por urgencia y por no poseer experiencia en el tema tratado, no pueden dedicarle el tiempo necesario y, este desarrollo, sería de una gran ayuda a su trabajo diario.

7. CONCLUSIONES

En el contexto de la generación de información, el INTA, juega un rol fundamental tal como se expresa claramente en su Plan Estratégico Institucional (PEI 2005-2015). Tal información se genera a través de las actividades sustantivas: Investigación, Extensión, Vinculación Tecnológica y Relaciones Institucionales y debe ser puesta a disposición de la sociedad a través de su sitio web, su portal de noticias y sus repositorios. Institucionalmente, es muy importante que la información producida por el financiamiento del estado se ponga disponible a los públicos objetivos.

Obviamente la web es el medio más utilizado para acceder a los contenidos y las búsquedas tienen un fuerte impacto en la percepción de usuario de la institución.

Tratando de mejorar la disponibilidad de la información producida constantemente por el organismo se construyó una solución que mejora tanto las búsquedas internas y externas del sitio. Nuestra solución acerca al sitio web institucional del INTA a la web semántica, utilizando datos estructurados para lograr el fin propuesto. Dicha solución plantea una arquitectura escalable y distribuida, que involucra un CMS, Drupal, que puede realizar balanceo de carga, con un frontend que realiza cache y distribuye, Varnish, más un indexador que puede trabajar en cluster, Solr, en su configuración de Solrcloud. Así para las búsquedas internas, se modificó el sitio para que indexe todo el contenido en Solr. Nuestra experiencia con esto fue muy buena; se mejoró el tiempo de respuesta de las búsquedas, ofreciendo facetado y mejores resultados. La arquitectura propuesta de desplegar esta plataforma en un host distinto en donde se encontraba el CMS permitió un mejor rendimiento en general del manejador de contenidos. Además, la vista de los resultados puede ser customizables ofreciéndole al usuario una mejor experiencia. Al usar la extracción de tópicos y entidades nombradas de forma automática, se mejora uno de los escenarios que prevalecía en el INTA, en donde webmasters debían asignar palabras y frases claves a todos los contenidos subidos, siendo esta una tarea que consume mucho tiempo, requiere de “expertise” y que se suma a un gran número de tareas que ya tenían asignadas, resultando en un etiquetado manual “pobre”. A este fin, se investigó sobre un software que permita no solo el reconocimiento de palabras relevantes, sino frases relevantes, en conjunto con el uso de tesauros para acotar el dominio de conocimiento en el cual se etiquetaba. El producto KEA, de la Universidad de Waikato, ofreció muy buenos resultados mejorando, en algunos casos, los términos con que el contenido se hallaba etiquetado previamente. KEA requiere de entrenamiento para mejorar las

sugerencias y en su utilización sugirió términos con una gran exactitud y asociando familias de palabras con el tesauro, por lo que se podría tranquilamente hacer uso desatendido de la herramienta en todos los documentos del INTA. Debemos destacar que KEA es un producto que trabaja Standalone y debió modificarse para que funcione como un servicio REST. En el caso de NER, no cuenta con un modelo en español entrenado correctamente. Stanford NER, parte de la pila de producto de NLP Stanford Group, cuyo objetivo es trabajar con el Procesamiento del Lenguaje Natural. El mismo, en nuestra experiencia, dio un porcentaje de éxito menor al 50% en el lenguaje español con el modelo incluido en el paquete por lo que necesita obligatoriamente el ojo humano al reconocer entidades. En cambio, OpenNLP, reconoció entidades con un porcentaje de éxito mayor al 85%. Para el etiquetado automático de contenidos, nuestra recomendación es usar OpenNLP frente a Stanford NER. Cabe destacar, que usando el módulo de autotagging desarrollado en este trabajo al crear contenido nuevo, se podría llegar prácticamente al 100% de efectividad ya que le permite al usuario elegir entre los términos detectados.

Con respecto a las búsquedas externas no se pudo probar el resultado de la investigación en buscadores debido a que el sitio no se encuentra accesible al público, pero creemos que deberían mejorar sustancialmente ya que, usando Dublin Core y Schema.org, los buscadores ahora entienden qué significa cada página del sitio mejorando el posicionamiento y la relación entre contenidos del sitio y la vista en los resultados de búsqueda.

Finalmente trabajar técnicas aplicables al proceso de indexación para la identificación, extracción y recomendación semiautomática de descriptores susceptibles de ser asignados a documentos web en forma de metadatos, apoyados en tesauros y en datos vinculados nos abre la posibilidad de avanzar sobre la relación de documentos en los repositorios institucionales en una red de contenidos relacionados por metadatos aprovechando las facilidades de los términos equivalentes y relacionados.

BIBLIOGRAFÍA

A. A., & I. G. (s.f.). *Catalogación y metadatos: ventajas y desventajas para lograr una recuperación de información eficienteCatalogación y metadatos: ventajas y desventajas para lograr una recuperación de información eficiente.*

Anderson, P. (2007). What is web 2.0 Ideas, technologies and implications for education.

Brun, R. E. (2003). Topic maps y la indización de recursos electrónicos en la web. *El profesional de la información*, Vol. 12.

Cailliau, R. (2 de 11 de 1995). A Short History of the Web. París, Francia. Obtenido de http://www.netvalley.com/archives/mirrors/robert_cailliau_speech.htm

Clarke, M., & P. H. (2014). How Smart Is Your Content? Using Semantic Enrichment to Improve Your User Experience Enrichment to Improve Your User Experience. *37*.

Comscore. (Abril de 2014). *Comscore*. Obtenido de Comscore: http://www.comscore.com/Insights/Market-Rankings/comScore-Releases-April-2014-US-Search-Engine-Rankings.

D. B., R. G., S. M., P. M., & A. S. (s.f.). *This is a position paper from schema.org for the W3CAuthors.*

E. E., S. S., & J. S. (2012). *The art of SEO.* O´Reilly.

Felice, M. (2009). *Tesis "Enriquecimiento automático de textos"* . Universidad Nacional de Luján, Licenciatura en Sistemas de Información.

Franchini, E. (2005). *Nuove prospettive nell'evoluzione dei thesauri.*

Gate, R. (20 de 02 de 2016). *Research Gate*. Obtenido de https://www.researchgate.net/publication/2618243_The_RBSE_Spider_-_Balancing_Effective_Search_Against_Web_Load
https://www.researchgate.net/publication/2618243_The_RBSE_Spider_-_Balancing_Effective_Search_Against_Web_Load

Graham, P. (Noviembre de 95). *Paul Graham*. Obtenido de http://www.paulgraham.com/web20.html

Grigoris, A., & F. V. (2004). *A semantic web primer.* Cambridge, Massachusetts: Massachusetts Institute of Tecnology.

Hinton, A. (2014). *Understanding Context: Environment, Language, and Information Architecture.* O'Reilly Media.

I. W., G. P., E. F., C. G., & C. N.-M. (s.f.). *Practical Automatic Keyphrase Extraction.*

Internet Live Stats. (s.f.). Recuperado el 8 de Junio de 2016, de Live Stats: http://www.internetlivestats.com/total-number-of-websites

J. P. (2009). *Diseño de un sistema colaborativo para la creación y gestión de tesauros en internet basado en SKOS.* Universidad de Murcia.

J. P., F. M., & J. R.-M. (s.f.). *Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives.*

J. R., & C. M. (s.f.). *Nested Named Entity Recognition.*

K. H., & M. R. (2012). *Content Strategy for the web.* New Riders.

M.Weinberg, G. (2001). *An Introduction to General Systems Thinking.* Dorset House Publishing Company.

Medelyan, O. (2006). *Semantically Enhanced Automatic Keyphrase Indexing.*

Medelyan, O. (2006). *Thesaurus Based Automatic Keyphrase Indexing.*

MIMAS, A. A. (2005). *Guidelines for Encoding Bibliographic Citation Information in Dublin Core Metadata.*

National Information Standards Organization. (2005). *Guidelines for the construction, format and management of monolingual controlled vocabularies.* National Information Standards Organization.

NetCraft. (June de 2016). *NetCraft.* Recuperado el 22 de June de 2016, de http://news.netcraft.com/archives/2016/06/22/june-2016-web-server-survey.html

P. M., J. A., & L. R. (s.f.). *Information architecture for the world wide web.* O´Reilly.

R. J., & P. L. (2016). *Data Recombination for Neural Semantic Parsing.* Association for Computational Linguistics (ACL).

SKOS. (s.f.). *¿Qué es SKOS?* Obtenido de http://skos.um.es/acerca/index.php

W. C., D. M., & T. S. (2015). *Search Engines.* Pearson Education, Inc.

W. M., & T. K. (2012). *Semantic Technologies in Content Management Systems.*

W3C. (2014). *RDF 1.1 Primer.* W3C Working Group.

Wordstream. (s.f.). *Wordstream*. Obtenido de Wordstream: http://www.wordstream.com/articles/internet-search-engines-history