Extracción automatizada para la Unidad de Vigilancia Tecnológica e Inteligencia Competitiva de la Patagonia

Damián Barry[1], Juan Manuel Cortez[22], Luis Ignacio Aita[12], Romina Stickar[1]

LINVI, Departamento de Informática, Facultad de Ingeniería, UNPSJB, 9120 Puerto Madryn, Argentina,
Sur Software S.H., Patagonia 687, 9120 Puerto Madryn, Argentina,
I.damian_barry@unpata.edu.ar, II.ignacioaita@sursoftware.com.ar,
WWW home page:http://madryn.unp.edu.ar

Resumen El presente trabajo presenta la constitución, desarrollo y actividades realizadas por la Unidad de Vigilancia Tecnológica e Inteligencia Competitiva de la Patagonia. En particular el desarrollo realizado de una herramienta que permite automatizar y homogeneizar las ecuaciones de búsqueda que utilizarán los expertos de la Unidad de Vigilancia.
La Unidad está conformada por los consorcistas que integran el Parque Tecnológico Puerto Madryn (PTPM) y tiene la finalidad de incorporar a la ciudad de Puerto Madryn una plaza más de la red de Antenas de Vigilancia Tecnológica e Inteligencia Competitiva que estratégicamente están planteadas dentro del Plan Argentina Innovadora 2020. En este sentido, se describen a continuación la metodología realizada y resultados obtenidos hasta la fecha.

Keywords: Vigilancia Tecnológica, Inteligencia Competitiva, Gestión del Conocimiento, Antena Tecnológica, Innovación Abierta

1. Introducción

La Unidad de Vigilancia Tecnológica e Inteligencia Competitiva de la Patagonia es una Plataforma que pone a disposición de cámaras y asociaciones empresariales, empresas, entidades gubernamentales, universidades y organismos públicos y privados de investigación, información clave y estratégica sobre distintos sectores.

"La información generativa y la información circulante pueden transformarse la una en la otra, pero la transformación de una información circulante o de señales de información generativa no es posible más que si se encuentra un aparato capaz de registrarla y tratarla."[10]

El contexto que la globalización ha generado, se encuentra caracterizado por una dinámica de presión competitiva que exige la redefinición constante del mapa de ventajas comparativas y por lo tanto, requiere de intervenciones apropiadas difícilmente generalizadas y duraderas.

En este contexto se pueden distinguir elementos tales como el aumento permanente de competidores a nivel global, escenarios sin fronteras físicas, generados principalmente a partir de las tecnologías de la información y las comunicaciones (TIC), disminución permanente de los ciclos tecnológicos y comerciales, internacionalización de las empresas y libre circulación del conocimiento, entre otros.

Por otra parte, se sostiene que la innovación productiva debe originarse asociada a la ciencia y la tecnología, poniendo en valor el conocimiento y los procesos de I+D+i, como argumento estratégico para el impulso y desarrollo de ventajas competitivas.

En este contexto, debido al desarrollo actual de las TIC, surgen y adquieren un rol central nuevas herramientas para la toma de decisiones; entre ellas, la Vigilancia Tecnológica e Inteligencia Competitiva (VTeIC), procurando la mejora de la competitividad de las empresas y sectores económicos, reduciendo los niveles de incertidumbre y riesgo, propios de la complejidad existente en los procesos de innovación.

La detección de información relevante sobre tendencias, novedades de clientes, invenciones y potenciales socios y competidores; a partir de datos codificados y analizados, brindan la posibilidad de planificar y formular estrategias tecnológicas minimizando la incertidumbre del contexto.

La Inteligencia Competitiva se ocupa del análisis, el tratamiento de la información, la evaluación y la gestión de los procesos de decisiones estratégicas dentro de las empresas e instituciones, integrando los sistemas de vigilancia tecnológica, comercial, de competidores y entornos, entre otros.

Tales actividades se convierten en herramientas clave dentro de procesos de innovación y en el fortalecimiento de empresas, por lo cual existe la necesidad de posicionar y lograr un alto nivel de penetración de esta área temática en los distintos actores sociales, logrando la concreción de una práctica sistemática por parte de los mismos.

2. Contexto

2.1. Antecedentes

Los territorios deben enfrentar nuevos desafíos para el diseño de estrategias de desarrollo dentro de un contexto de mayor complejidad, incertidumbre y velocidad de cambios, adquirir mayores competencias, adaptarse a las exigencias del mercado y avanzar hacia el desarrollo del territorio.

Así, la utilización de las potencialidades endógenas se presenta como la estrategia para lograrlo. Teniendo en cuenta que la difusión de las innovaciones y el conocimiento entre las empresas y la organización de los sistemas productivos en formas más flexibles, son dos pilares fundamentales para el proceso que mejoran las economías internas de las firmas y favorecen el posicionamiento competitivo de las ciudades y territorios.

Puerto Madryn presenta espacios institucionales activos en los que participan representantes de diferentes sectores que avanzan en experiencias concretas de

articulación y cuyo aprendizaje aportan las bases para la creación de una oficina de VTeIC. A la vez, esta unidad permitirá enriquecer tales espacios interinstitucionales.

Agencia de Desarrollo Productivo de Puerto Madryn (ADP) Es una asociación civil sin fines de lucro, creada en septiembre de 2003 por la Municipalidad de Puerto Madryn, en conjunto con la Cámara de Industria, Comercio, Producción y Turismo (CAMAD), que nuclea instituciones del mundo productivo, de la tecnología y la innovación productiva, tales como Centro Nacional Patagónico (CENPAT-CONICET) , Universidad Tecnológica Nacional Facultad Regional Chubut (UTN-FRCh), Universidad Nacional de la Patagonia San Juan Bosco (UNPSJB) e Instituto Nacional de Tecnología Industrial (INTI), con el fin de apoyar el desarrollo productivo de la ciudad.

Desde el año 2012, se ha re-impulsado su tarea a través de la designación de nuevos representantes municipales, la incorporación en estatutos de un Gerente Operativo y la invitación a integrantes del mundo productivo local a conformar el Consejo Asesor de la misma.

Incubadora Madrynense de Emprendimientos Tecnológicos (IMET) En el 2010 la Secretaría de Ciencia, Tecnología e Innovación Productiva de la provincia del Chubut (SCTeIP) impulsó la formación de una incubadora de empresas de base tecnológica, la que se conformó en el año 2011 , integrada por la Universidad Tecnológica Nacional Facultad Regional Chubut, la Universidad Nacional de la Patagonia San Juan Bosco, el Centro Nacional Patagónico (CENPAT – CONICET) y la Cámara de Industria, Comercio, Producción y Turismo de Puerto Madryn (CAMAD).

IMET es una institución conformada por entidades públicas y privadas, promovidas por el estado, que brindan servicios y asistencia a emprendedores para que desarrollen su proyecto de negocio.

IMET propone como visión ser la entidad líder que promueva emprendimientos sustentables e innovadores que potencien la dinamización y diversificación de la actividad productiva, colaborando con el Sistema de Innovación Local y de esta forma contribuir al desarrollo económico y social de la región.

IMET promueve el desarrollo de actividades emprendedoras e innovadoras de base tecnológica, en sectores tales como: alimentación, metalmecánico, construcción, software, energías alternativas, gestión de recursos del mar, tecnologías del ambiente, biotecnología y producción en general.

Parque Tecnológico Puerto Madryn (PTPM) La finalización del proyecto de viabilidad, el otorgamiento del predio donde se ubicará y el reciente proyecto de infraestructura para la instalación del Parque Tecnológico de Puerto Madryn como nodo central, conectando y vinculando a los centros productivos y de generación y transferencia de conocimiento, permitirá dar impulso al desarrollo de la región, teniendo en cuenta que existen la intención, vocación y compromiso

de las instituciones que conforman el Sistema de Innovación Local, de crear, mantener y dar contenido a este espacio.

El PTPM propone como visión, ser el agente dinamizador que genere la articulación entre los actores del Sistema de Innovación, promoviendo las actividades de I+D+i entre los sectores académicos, científicos, tecnológicos y productivos, siendo protagonista de la nueva economía del conocimiento y contribuyendo al desarrollo local.

Es importante destacar que del mismo participan las áreas centrales planteadas en el triángulo de Sábato[12], donde intervienen el sector privado desde la perspectiva de la producción de bienes y servicios, el conocimiento desde sus instituciones académicas y científicas en su rol de transferencia y asistencia científica y tecnológica, y el estado en la articulación y promoción de las actividades de desarrollo.

2.2. Visión de la Unidad de VTeIC Patagónica

Ser líderes en detectar, analizar y sintetizar información y ponerla a disposición para la toma de decisiones estratégicas que permitan a las organizaciones de la región generar desarrollos innovativos, establecer mecanismos de innovación abierta e intercambio de saberes y conocimientos.

2.3. Objetivos

Objetivo General Crear en el seno del PTPM, una Unidad de VTeIC orientada a brindar información relevante para la industria de la región como del sector público local y provincial; facilitando la creación de ventajas competitivas de las empresas y la mayor eficiencia en la toma de decisiones como resultado de la reducción de incertidumbre y riesgo asociados a procesos innovativos.

Objetivos Específicos Entre los objetivos específicos se destacan los siguientes:

- Identificar información relevante a corto y mediano plazo que permita establecer las tendencias de la ciencia, la tecnología y la economía, mediante la exploración periódica y sistémica de las mismas, detectando los emergentes que puedan impactar económica y socialmente las estructuras productivas de la región; a través de la construcción, desarrollo e implementación de herramientas de monitoreo.
- Generar un ámbito de identificación, análisis y resolución de problemas en las estructuras productivas del sector por medio de:
  - Identificación de líneas potenciales de I+D+i
  - Análisis de información técnico-científica.
  - Identificación de tecnologías existentes en los mercados de interés.
  - Análisis de aspectos específicos y generales de los mercados nacionales e internacionales.

- Promover un espacio colaborativo e interdisciplinario, que favorezca la generación de propuestas innovadoras en materia productiva que profundizar el nivel de desarrollo tecnológico de la industria.
- Vincular ofertas y demandas de los sectores productivos, académicos/científicos y gubernamentales mediante el desarrollo de mapas tecnológicos basados en bibliografía, publicaciones y patentes.
- Combinar el conocimiento interno con el conocimiento externo para sacar adelante los proyectos de estrategia y de I+D+i, acelerando el proceso de innovación interna y expandiendo los mercados para la aplicación de la innovación. El modelo de "Innovación Abierta" permite que el flujo de conocimiento circule desde dentro y desde fuera, considerando tres diferentes niveles de intercambio: interdepartamental, entre empresas y con el entorno institucional (instituciones gubernamentales, universidades, etc.).
- Identificar potenciales socios regionales, nacionales e internacionales de proyectos tecnológicos en las áreas de interés definidas para la Unidad.
- Brindar asesoramiento y capacitación a las instituciones tractoras de la región en técnicas de VTeIC.

2.4. Áreas tecnológicas de interés para la Unidad de VTeIC Patagónica

Las actividades de vigilancia se realizarán en referencia a las siguientes áreas tecnológicas y sus derivados de interés de los miembros del PTPM:

- Pesca y acuicultura
- Aluminio y sus derivados
- Minería
- Energía eólica
- Logística y distribución

3. Desarrollo de la Unidad de VTeIC Patagónica

3.1. Introducción

Un Sistema de Vigilancia e Inteligencia consta de 7 (siete) fases preliminares: planificación, identificación de necesidades, búsqueda de información y herramientas, monitoreo y validación, tratamiento y análisis, difusión y protección y evaluación y seguimiento conformando lo que se llama el Ciclo de VTeIC.

- Fase 0: Estructura Organizativa, planificación de actividades de VT, proyección de productos y servicios de VTeIE, recursos físicos y humanos, etc.
- Fase 1: Identificación de necesidades e interpretación del sector - árbol tecnológico, fuentes de información, definición de palabras claves y términos técnicos, recopilación documental, distribución geográfica.
- Fase 2: Búsqueda de información, herramientas y generación de ecuaciones de búsqueda.

- Fase 3: Monitoreo y validación de la información.
- Fase 4: Tratamiento y análisis de información.
- Fase 5: Difusión y protección de la información.
- Fase 6: Evaluación, seguimiento y actualización del proceso de VeIE y presentación de Informe Final

3.2. Bigdata como estrategia de la Unidad de VTeIC

Contexto BigData A partir del creciente uso de Internet y de la popularidad en el uso de las Redes Sociales las cuales, según Edgar Morín[10], generan "ruido" en el ciber-espacio, o lo que ahora se lo denomina como ïnfoxicación", es que se crea la necesidad de poder procesar y transformar la gran cantidad de datos en información útil para las personas o para las organizaciones que desean entender sucesos desde la perspectiva social.

Por otra parte la producción y obtención de información ha pasado a ser uno de los grandes activos de las organizaciones, ya sean públicas, mixtas o privadas. En este sentido el desarrollo y estudio de la generación, administración, explotación, interpretación y clasificación de información se ha convertido en un desafío tecnológico y científico a nivel mundial. Para poder abordarlo, no sólo se requiere del soporte de científicos y tecnólogos en el área de la informática sino además de la integración con investigadores y expertos de distintas áreas vinculadas con las actividades que se desean analizar y comprender, donde a través de la conformación de equipos multidisciplinarios generen verdadero valor a la información circundante.

En este sentido Edgar Morin[9] en su libro "El Método" dice que "La idea de cibernética–arte-ciencia del gobierno puede integrarse y transformarse en cocibernética – arte-ciencia de pilotear conjuntamente, donde la comunicación ya no es útil del mando, sino una forma simbólica compleja de organización."

"Así la información sólo puede nacer a partir de una interacción entre una organización generativa y una perturbación aleatoria al ruido. Ergo la información no puede desarrollarse más que a partir del ruido. Y desde luego, en el nacimiento de una información, siempre se necesita una actitud organizacional de carácter neguentrópico que se supere a sí misma transformando el evento en novedad."

Claramente los meta-buscadores utilizados actualmente en las unidades de Vigilancia Tecnológica, como ACM Digital Library[5], Scopus[3], SciELO[11] o ScienceDirect[2] entre otras tantas, han solucionado en parte el problema, pero han introducido otro que no es menor: la supuesta información recuperada es sesgada al criterio de clasificación y ordenamiento de cada buscador y no precisamente resuelve la problemática de encontrar lo que se necesita, más aún de generar la información correcta como objetivo a la resolución de problemas concretos.

El carácter público o mediante una membresía de gran parte de la información generada en estas bibliotecas o meta-buscadores y gracias al acceso de la misma, mediante interfaces públicas, permite acceder a la misma con técnicas de Information Search and Retrieval (ISR).

Por otra parte además en la actualidad existe gran cantidad de Corpus Documentales, artículos, sitios web, librerías digitales, que procesan grandes volúmenes de información, la cual es necesario manejar eficientemente. En suma se ha pasado de hablar de gigabyte de información a hablar con total normalidad del orden de los petabytes [7,1,14].

Esta situación ha generado el desafío de mejorar las herramientas de búsqueda en lo que se denomina "Information Search and Retrieval" utilizando para ello diversas técnicas que necesitan ser evaluadas, re-formuladas y si es posible mejoradas.

Asociado a esta problemática, se suma la necesidad de escalabilidad, disponibilidad y desempeño en el manejo de grandes volúmenes de información, situación que requiere de técnicas de sistemas distribuidos.

El volumen de información hace impensable utilizar mecanismos manuales supervisados para su clasificación, ordenamiento y uso, especialmente como aportes al análisis científico de datos.

Es por esto que técnicas de Inteligencia Artificial asociadas a lo que denominamos Data Mining son necesarias a la hora de analizar grandes volúmenes de información.

A modo de ejemplo, podemos mencionar soluciones similares como la adoptada por la Casa Blanca que ha utilizado la combinación de Drupal y Apache Solr en el portal de contenidos documentales[13] para su recuperación e indagación del material de forma eficaz.

¿Big Data? ¿Qué significa? Desde que el MGI (McKinsey Global Institute) acuñó el concepto a mediados del 2011, han surgido numerosas definiciones de Big Data, la mayoría de ellas como intentos de acotar el concepto a un área determinada.

Una de las más completas es la realizada por Gartner [http://www.gartner.com/it-glossary/big-data/] "Son activos de información caracterizados por su alto volumen, velocidad y variedad, que demandan soluciones innovadoras y eficientes de procesado para la mejora del conocimiento y toma de decisiones en las organizaciones."

Un estudio realizado por el IBM Institute for Business Value entre más de mil profesionales en IT y expertos en la materia de casi 100 países muestra que el término para algunos es tan abarcativo como "Nuevos tipos de datos y análisis" mientras que otros lo vinculan solamente a las redes sociales.

Para este trabajo consideraremos el concepto Big Data simplemente como Análisis y Tratamiento de grandes volúmenes de datos.

Más allá de la definición lo importante del Big Data es la tendencia de avance de las tecnologías que abren puertas para nuevos enfoques de entendimiento y toma de decisiones, ampliando el alcance de los sistemas de bases de datos relacionales, permitiendo almacenar y procesar grandes cantidades de datos de todos los tipos posibles (estructurados, semi-estructurados y sin estructurar)

3.3. Herramienta de Vigilancia automatizada

Como hemos expresado entendemos que no es posible pensar implementar un tratamiento manual para la gestión de conocimiento realizado por expertos para la implementación de las fases 2, 3 y 4 planteadas en la sección 3.1, es que la Unidad de VTeIC de la patagonia ha diseñado y desarrollado una herramienta integral que permita realizar la extracción de información tanto científica, de patentes como comercial y de mercado de forma automatizada y homogénea para su tratamiento.

Para realizar este producto se establecieron 3 etapas:

1. Extracción y pre-clasificación de información.
2. Clasificación avanzada de información y construcción de espacios semánticos y taxonómicos mediante técnicas de inteligencia artificial.
3. Explotación de la información clasificada. Automatización de los informes y reportes dela Unidad de Vigilancia Tecnológica e Inteligencia Competitiva.

El presente trabajo es alcanzado solamente por la etapa 1 del proyecto.

4. Motor de extracción de información

Para la extracción de contenido público de bibliotecas digitales y portales de patentes se desarrollo un motor de extracción de contenido público de internet denominado ÇrawlingExtractor.[el] mismo está basado en un trabajo anterior para extraer noticias de orden público georreferenciadas según su contenido denominado "ZCrawler"[4], el mismo se compone de un lenguaje de extracción que le permite a los usuarios mediante una especificación formal comenzar a extraer información siguiendo múltiples criterios y permitiendo la clasificación de esta información según a quien se está evaluando.

Para ello se desarrollaron servicios de ejemplo para las API de Springer[8], Scopus[3] y para medios digitales que contaran con RSS estándar.

Adicionalmente se adaptaron extractores ya que sus definiciones de RSS no accedían a información complementaria de los artículos publicados como por ejemplo el perfil de los autores y las citas bibliográficas.

4.1. Proceso de Extracción

El proceso de extracción de la información para almacenar los documentos extraídos en el índice se divide en dos etapas claramente diferenciadas: configuración y extracción.

Como se puede apreciar el proceso consta de 3 fases bien definidas:

1. Definición de ecuaciones de extracción.
2. Ejecución de las extracciones de acuerdo a sus parámetros de planificación.
3. Clasificación e Indexación de las noticias extraídas. A su vez esta fase se divide en las siguientes tareas:
   a) Extraer artículos por fuente.

b) Expandir la información no provista por la fuente.
c) Clasificar la información.
d) Indexar y almacenar la información en la base de datos Apache Solr[6].

4.2. Componentes de la Extracción

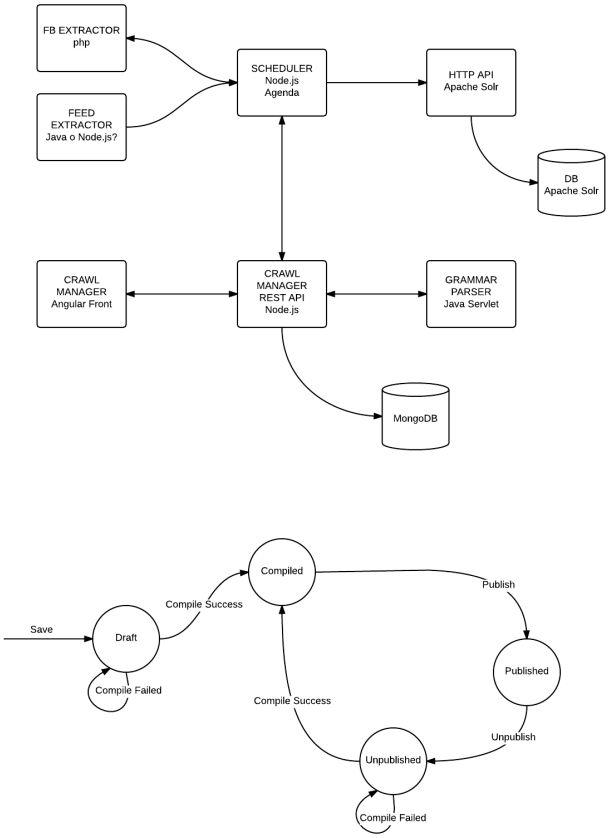

Figura 1. Arquitectura de los componentes de extracción

4.3. Arquitectura de Extractores

El éxito de la herramienta se basa en la posibilidad de estandarizar en un sólo formato y lenguaje todos los sitios y plataformas de extracción de infor-

mación, evitando de esta forma la necesidad que los expertos aprendan muchos lenguajes para elaborar las ecuaciones de búsqueda. Además el esfuerzo de trabajar con formatos heterogéneos de información y estandarizarlos permite luego, independientemente de la fuente unificar la forma de consultar la información.

Para ello es necesario que se programen los distintos formatos de extracción para cada fuente. Si bien el mismo requiere de un esfuerzo por parte de programadores, una vez desarrollado la misma fuente sirve para múltiples consultas.

Componentes de un extractor Los extractores desarrollados para las diversas fuentes de información se alojan en un proyecto Node.js denominado crawl-extractors.

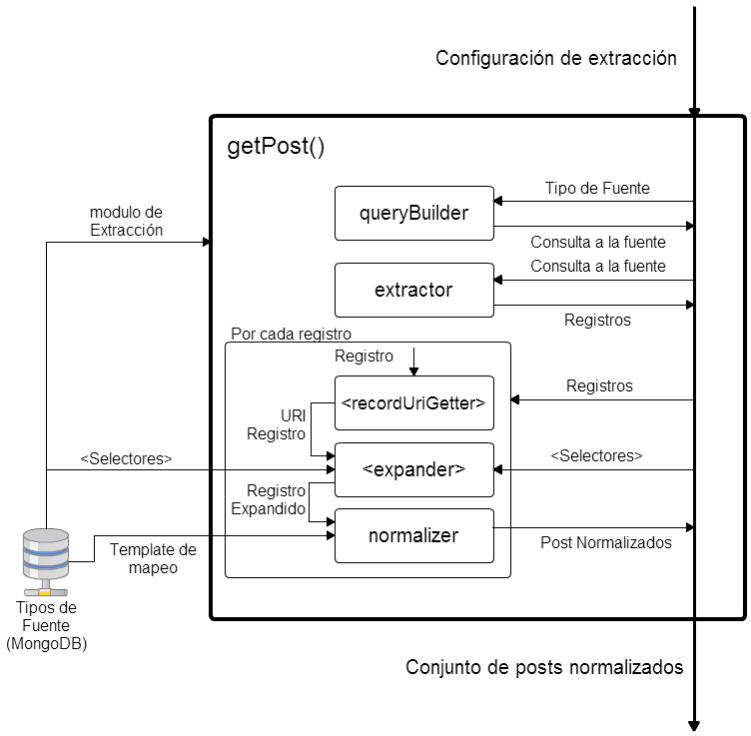

Figura 2. Componentes de un extractor

Todo extractor debe implementar la función getPost() que recibe como parámetro una configuración de extracción y retorna como resultado una colección de posts en formato normalizado.

Existe un extractorFactory que se encarga de retornar el extractor correspondiente de acuerdo al tipo de fuente y la uri configuradas en la extracción.

Para lograr su cometido la función getPosts a su vez ejecuta los 5 pasos que se describen a continuación:

1. **Query Builder** Genera la URI de búsqueda del recurso a partir de la Configuración de Extracción correspondiente, considerando todos los filtros y criterios que puedan incorporarse en la misma.
2. **Extractor** Realiza la extracción de datos desde la fuente, obteniendo los registros correspondientes en el formato original de cada fuente.
3. **Record URI Getter** A partir de un registro en formato original de la fuente, obtiene la URI del recurso original.
4. **Expander** Utilizando un conjunto de selectores (Por defecto selectores CSS), examina el recurso (por defecto una página HTML) y extrae información adicional utilizando los selectores. Retorna un objeto JSON que contiene un conjunto de pares clave-valor. Las claves corresponden con el nombre de cada selector y el valor contiene un arreglo con toda la información expandida.
5. **Normalizer** partir de un registro o un registro expandido, mapea los campos del registro original hacia un registro normalizado VTeIC.

5. Conclusiones

A la presentación del presente trabajo la herramienta cuenta con la etapa 1 "*Extracción y pre-clasificación de información*concluida en etapa de prueba extrayendo información científica y de patentes de acuerdo a las ecuaciones de búsqueda definidas para el proyecto.

En esta etapa se espera depurar el gestor de extracciones y los extractores específicos de cada fuente en función del análisis realizado de información. El planificador de extracciones ya en fase operacional está extrayendo información a la base de datos local para luego poder dar comienzo a las etapas 2 y 3, pues la mismas se basan en la información extraída.

6. Líneas Futuras

Establecido el proyecto completo es fundamental concluir con las etapas 2 y 3 planteadas para el proyecto y contar con una herramienta completa que facilite la tarea analítica de los expertos.

Entendemos que sin esta herramienta que asista a los expertos sería imposible encontrar información útil o valiosa para la Unidad de VTeIC.

Referencias

1. Azza Abouzeid, Kamil Bajda-Pawlikowski, Daniel Abadi, Avi Silberschatz, and Alexander Rasin. Hadoopdb: an architectural hybrid of mapreduce and dbms technologies for analytical workloads. *Proceedings of the VLDB Endowment*, 2(1):922–933, 2009.
2. Elsevier B.V. Sciencedirect.

3. Elsevier B.V. Scopus.
4. Barry Damián, Aita Luis Ignacio, and Cortez Juan Manuel. Zcrawler: Extracción, clasificación y publicación de información pública desde su perspectiva geográfica. *JAIIO 2014*.
5. Association for Computing Machinery. Acm digital library.
6. Apache Fundation. Apache solr reference guide.
7. Cal Henderson. *Building scalable web sites*. .º'Reilly Media, Inc.", 2006.
8. Springer Science+Business Media. Springer.
9. Edgar Morin. El método iii. el conocimiento del conocimiento. *Madrid: Cátedra*, 2:24, 1988.
10. Edgar Morin and Marcelo Pakman. *Introducción al pensamiento complejo*. Gedisa Barcelona, 1994.
11. SciELO Scientific Electronic Library Online. Scielo.
12. Jorge A Sabato. *Ensayos en campera*. Juárez, 1979.
13. Wayan Vota, Rajendra Singh, Siddhartha Raja, Jude Genilo, Shamsul Islam, Marium Akther, and Rohan Samarajiva. Digital government: building a 21st century platform to better serve the american people. 2012.
14. Zhou Wei, Guillaume Pierre, and Chi-Hung Chi. Scalable transactions for web applications in the cloud. In *Euro-Par 2009 Parallel Processing*, pages 442–453. Springer, 2009.