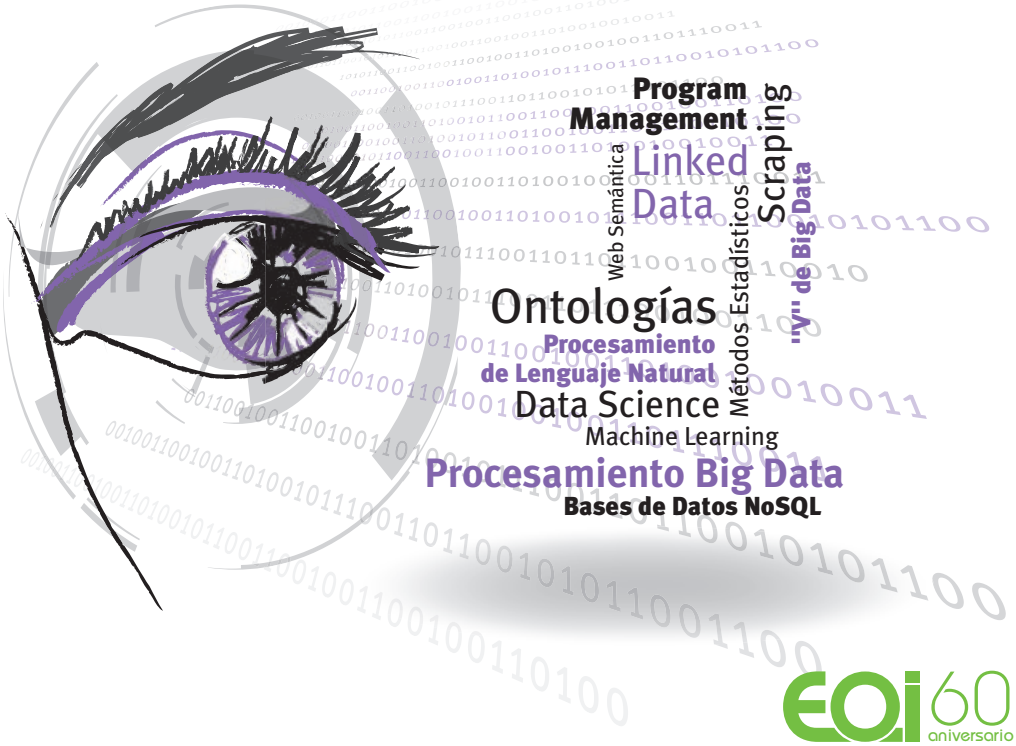
GOBIERNO DE ESPAÑA
MINISTERIO DE INDUSTRIA, ENERGÍA Y TURISMO
EOI Escuela de organización industrial

innovación y creatividad

big intelligence

nuevas capacidades big data para los Sistemas de **Vigilancia Estratégica** e **Inteligencia Competitiva**



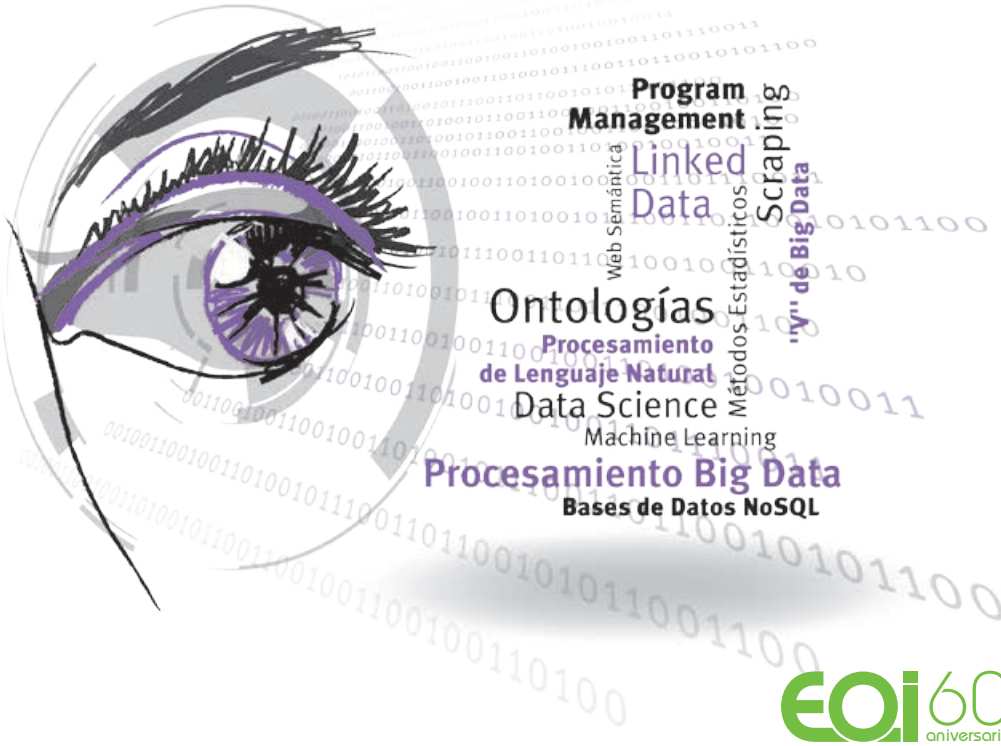
EOI 60 aniversario

GOBIERNO DE ESPAÑA
MINISTERIO DE INDUSTRIA, ENERGÍA Y TURISMO
EOI Escuela de organización industrial

innovación y creatividad

big intelligence

nuevas capacidades **big data** para los Sistemas de **Vigilancia Estratégica** e **Inteligencia Competitiva**



EOI 60 aniversario

CRÉDITOS

DIRECCIÓN DEL PROYECTO

Eduardo Lizarralde
Vicedecano EOI

Juan Jiménez Morillas
Responsable de Proyectos de Investigación EOI

AUTOR

Antonio Miranda Raya
Director de Proyectos en EOI

Libro digital en:
http://a.eoi.es/bigintelligence
Enlace directo en:

ISBN
978-84-15061-61-8

DEPÓSITO LEGAL
M-35029-2015

Madrid, 2015

ÍNDICE

ACERCA DE LOS AUTORES ..... | 5 |

PRÓLOGO ..... | 11 |

Capítulo 1
PRESENTACIÓN ..... | 15 |

Capítulo 2
VIGILANCIA ESTRATÉGICA E INTELIGENCIA COMPETITIVA ..... | 27 |

1. Vigilancia Estratégica, Inteligencia Competitiva y Gestión del conocimiento en el siglo XX ..... | 28 |
2. Vigilancia, Inteligencia, Conocimiento y Prospectiva ..... | 37 |
3. La norma UNE 166.006:2011 "Gestión I+D+i: Sistema de Vigilancia Tecnológica e Inteligencia Competitiva" ..... | 42 |

Capítulo 3
NUEVAS CAPACIDADES BIG DATA ..... | 49 |

1. "V" de Big Data ..... | 50 |
2. Business Bots, Spiders, Scrapers: recuperando información desestructurada de la WEB ..... | 54 |
3. Data Science, Estadística, Inteligencia artificial, Data Mining, Investigación Operativa, Machine Learning, Procesamiento del Lenguaje Natural... el entorno de Big Data ..... | 65 |
4. Machine Learning ..... | 80 |
5. Procesamiento de Lenguaje Natural ..... | 86 |
6. Procesamiento de Lenguaje Natural versus Machine Learning ..... | 94 |
7. Arquitectura Big Data ..... | 100 |
8. Ontologías, Datos Enlazados (Linked Data) y Web Semántica ..... | 112 |
9. Gestionando el Conocimiento y la Veracidad de la información ..... | 131 |
10. Mapeando las tecnologías Big Data y las actividades de Vigilancia Estratégica e Inteligencia Competitiva ..... | 135 |

Capítulo 4
DISEÑANDO SISTEMAS DE VIGILANCIA E INTELIGENCIA CON NUEVAS CAPACIDADES BIG DATA ..... | 139 |

1. Casos de Uso y Necesidades de Vigilancia Estratégica e Inteligencia Competitiva: el estilo de pensar "Big Data" ..... | 141 |
2. Fuentes de Información. Taxonomías ..... | 143 |
3. Integración de Datos ..... | 149 |
4. Modelo de Información: los Módulos de Entidades Estructurales de Información ..... | 154 |
5. NoSQL: las Bases de Datos del Big Data ..... | 165 |
6. Funcionalidades, Implementaciones e Interfaces Big Data para los Sistemas de Vigilancia e Inteligencia ..... | 168 |

Capítulo 5
FORMALIZACIÓN DEL MODELO Y LA METODOLOGÍA ..... | 211 |

1. Elementos del Modelo ..... | 212 |
2. Metodología de Diseño del Sistema de Vigilancia / Inteligencia ..... | 215 |
3. El Modelo organizativo ..... | 226 |
4. Puesta en marcha mediante Program Management ..... | 229 |

Capítulo 6
BIBLIOGRAFÍA Y FUENTES DE DOCUMENTACIÓN ..... | 235 |

ACERCA DE LOS AUTORES

Autor

Antonio Miranda Raya

Cuenta con más de 15 años de experiencia que va desde la Gerencia en el sector de la Consultoría TIC hasta la Dirección de Sistemas de Información y departamentos TIC. Su foco actual es el diseño y dirección de iniciativas que contribuyan a la Innovación y transformación digital de las organizaciones y la sociedad. Actualmente es Director de Proyectos en EOI.

Durante el año 2014 se encargó de la Dirección y Ejecución del proyecto consistente en el diseño de un Sistema Big Data de Vigilancia Estratégica e Inteligencia Competitiva del sector TIC realizado para la Secretaría de Estado de Telecomunicaciones y Sociedad de la Información del Ministerio de Industria y Turismo, proyecto tras el que surge la iniciativa de crear este libro.

Es Licenciado en Informática por la UPM y Executive MBA por el IE. Ha realizado en EOI los Programas de Finanzas para Directivos, Dirección de Proyectos con Metodología PMI, el nuevo rol del Director de Servicios IT y Mapas Estratégicos y Cuadro de Mando Integral entre otros. Asimismo ha cursado programas especializados en Gestión del I+D+i en el CEU y la UPM.

Colaboraciones

Este libro cuenta con las siguientes colaboraciones:

Dr. Asunción Gómez Pérez

Es una experta en Inteligencia Artificial en el área de ontologías y Web semántica. Se licenció en informática por la Universidad Politécnica de Madrid (UPM) en 1991, y alcanzó el doctorado en Ciencias de la Computación e Inteligencia Artificial en la misma universidad en diciembre de 1993. Realizó estudios post-doctorales en el prestigioso Knowledge Systems Laboratory de la Universidad de Stanford, en Palo Alto (California). Habla francés e inglés, y tiene un Máster en Dirección y Administración de Empresas. Ha ganado el premio Ada Byron a la Mujer Tecnóloga en 2015. Ha sido distinguida recientemente como "una de las tres mujeres más reconocidas y con mayor presencia mundial en el amplio campo de investigación de las tecnologías semánticas. Actualmente es Catedrática en la Universidad Politécnica de Madrid (UPM).

"Ontologías, Datos Enlazados (Linked Data) y Web Semántica". Apartado 3.8.

Juan Jiménez Morillas

Es un especialista en prospectiva y vigilancia tecnológica, habiendo desarrollado su labor en el Observatorio de Prospectiva Tecnológica Industrial (OPTI). Actualmente es Director de Proyectos en el Vicedecanato de EOI. Su formación universitaria es de Ingeniero de Caminos, Canales y Puertos por la UPM, en la especialidad de Urbanismo y Ordenación del Territorio. Juan es también Executive MBA por la EOI.

"Vigilancia Estratégica, Inteligencia Competitiva y Gestión del conocimiento en el siglo XX". Apartado 2.2.

María Poveda Villalón

Es estudiante de doctorado en el Ontology Engineering Group, grupo de investigación perteneciente a la Universidad Politécnica de Madrid (UPM). En dicha universidad realizó también los estudios de Ingeniería Superior en Informática (2009) y Máster en Investigación en Inteligencia Artificial (2010). Sus intereses en investigación son principalmente la ingeniería ontológica, la web semántica y datos enlazados. Durante los últimos años ha realizado estancias de investigación en centros extranjeros como University of Liverpool (2011), Free University of Berlin (2012) y en empresas como Mondeca en Paris (2013).

"Ontologías, Datos Enlazados (Linked Data) y Web Semántica". Apartado 3.8.

Antonio Sánchez Valderrábanos

Ha trabajado durante más de 20 años en el campo de la lingüística computacional. Desarrolló su carrera profesional en importantes multinacionales del sector IT como IBM y Novell Corporation, en las que trabajó en el despliegue de tecnologías lingüísticas para sistemas de recuperación de información y entornos de publicación electrónica. Antonio fundó la empresa Bitext en 2008 con el objetivo de proporcionar tecnología semántica multilingüe OEM para diferentes áreas de negocio, como Social Media, búsqueda y análisis de textos. Antonio es licenciado en Filología y doctor en Lingüística por la Universidad Autónoma de Madrid.

"Procesamiento de Lenguaje Natural versus Machine Learning"[1]*. Apartado 3.5.*

Dr. María del Carmen Suárez de Figueroa Baonza

Es profesora ayudante doctor en la Escuela Técnica Superior de Ingenieros Informáticos de la Universidad Politécnica de Madrid (UPM) e investigadora senior del Ontology

[1] Original del inglés "Machine Learning & Deep Linguistic Analysis in Text Analytics". Traducción al castellano de Antonio Miranda Raya.

Engineering Group. Es Doctora en Ciencias de la Computación e Inteligencia Artificial por la UPM desde 2010 y ha recibido el Premio Extraordinario de Tesis Doctoral de la UPM. Sus líneas de investigación se centran en la Ingeniería Ontológica y en Linked Data. Es co-editora del libro "Ontology Engineering in a Networked World" (Springer 2012). Ha co-organizado sesiones, conferencias, workshops y tutoriales en eventos internacionales.

"Ontologías, Datos Enlazados (Linked Data) y Web Semántica". Apartado 3.8.

Revisión

Además de los autores, varios profesionales del sector TIC han contribuido con sus revisiones, comentarios y recomendaciones: Alberto Latorre, Carlos Hernando Carasol, Jerónimo García Loygorri, Mónica Blanco, Pedro Bernad, Elena Salinas y Sergio Jiménez.

Especialmente me gustaría destacar la participación en la revisión del libro a:

***Sergio Montoro Ten** es emprendedor y consultor especializado en tecnologías de la información. También es redactor del blog La Pastilla Roja, dedicado a la Tecnología y sus usos sociales. Cuenta con más de 20 años de experiencia, especialmente como líder técnico y CTO. Su formación universitaria es de Licenciado en Informática por la UPM y cuenta además con un Master en Gestión de Recursos Humanos por ESIC.*

***José Luis Jerez** es un experto en tecnologías de la información, redes y seguridad informática con más de 15 años de experiencia, principalmente como Gerente de Operaciones de Seguridad. Su formación universitaria es de Ingeniero en Informática por la UPM y cuenta con los certificados CISSP, CISM, CISA y CRISC entre otros.*

***José Antonio Leiva** ha trabajado como CTO y líder técnico de la startup Smartvel, como Development Manager en Prosegur e IT Manager de Plettac Electronics. Se ha incorporado recientemente al equipo de Ingeniería de Amazon España. Su formación universitaria es de Ingeniero en Informática por la UPM. José Antonio también es MBA por el Instituto de Empresa de Madrid.*

Sus aportaciones han sido incorporadas en buena medida y han ayudado enormemente a mejorar la calidad del libro.

Nota del Autor

Este libro se escribe tras la ejecución del proyecto "Diseño de un Sistema Big Data de Vigilancia Estratégica para el sector TIC", realizado por EOI para la Secretaría de Estado de Telecomunicaciones y la Sociedad de la Información, al percibir la oportunidad de consolidar el conocimiento adquirido desarrollándolo, ampliándolo y compendiándolo.

Se empezó a escribir a finales de 2014 y ha estado en proceso de redacción durante buena parte del 2015. En este breve lapso de tiempo han surgido novedades y transformaciones en el mundo del Big Data que han podido ser incorporadas, otras no. El abrumador ritmo de cambio en el mundo del Big Data convertirán en obsoletos algunos de los contenidos: esperamos poder ir incorporándolos en próximas ediciones.

Puede ser útil tanto para estudiantes y personas en general interesadas bien en el conocimiento de "Big Data", bien en el conocimiento de la Vigilancia Estratégica y la Inteligencia Competitiva y también lógicamente en ambas áreas de conocimiento. Inicialmente el libro fue concebido como una colección de artículos de múltiples autores, aunque finalmente terminó siendo obra mayoritariamente de un sólo autor, habiendo realizado un esfuerzo en integrar en lo posible todos los apartados. Cada apartado y subapartado pretende mantener, sin embargo, una cierta independencia para que el lector pueda acudir a ellos como unidades de aprendizaje independientes.

Varios profesionales del sector TIC han leído y revisado el libro y sus revisiones, comentarios y recomendaciones han sido incorporados en buena medida, con objeto de reducir al máximo el número de erratas contenidas. Les agradecería que en caso de que encontrase alguna errata o quiera hacer alguna sugerencia, comentario o solicitud para futuras ediciones lo comunicara por e-mail a antoniomiranda@eoi.es

Este libro estará a su disposición para su descarga gratuita desde SAVIA, el repositorio de conocimiento de EOI, accesible en http://www.eoi.es/savia/ En este mismo repositorio pueden encontrar una clase sobre "Big Data y Vigilancia Estratégica" impartida en el Máster de Innovación de EOI, del que este libro es su referencia principal: http://www.eoi.es/savia/video/2554/big-data-y-vigilancia-estrategica

He optado por incluir tanto el término en español como el término en inglés, frente a la opción de incluir únicamente el término en español. Para ello añado al término en español el texto siguiente: (en inglés "*término-en-inglés*"). El idioma inglés es de facto el latín actual del mundo científico y tecnológico. La consulta y lectura de publicaciones en inglés es habitual en el día a día de cualquiera que trabaje en estos ámbitos. Creo que conocer los dos términos facilita el aprendizaje y la toma de referencias a los lectores.

PRÓLOGO


Program Management
Linked Data
Web Semántica
Scraping
Métodos Estadísticos
"v" de Big Data
Ontologías
Procesamiento de Lenguaje Natural
Data Science
Machine Learning
Procesamiento Big Data
Bases de Datos NoSQL

Los datos se han convertido en un activo muy valioso para la sociedad del siglo XXI. Se están desarrollando importantes innovaciones para explotar la enorme cantidad y variedad de datos que se generan y almacenan de forma constante a una velocidad creciente. Asimismo los datos se han convertido en una destacada fuente de nuevos empleos: Big Data o Data Science son sin duda dos de los trending topics más relevantes en los últimos tiempos en cuanto a la empleabilidad, siendo destacable el hecho de que los empleos generados son de alta calidad.

Se abren grandes oportunidades no sólo para los sectores más digitales, sino también para sectores más tradicionales como los sectores de la Salud, el Transporte o la Energía, que pueden generar valor incorporando a sus actividades la explotación de datos provenientes de sensores, satélites, vídeos, señales GPS y también los generados por personas, por ejemplo los generados en las Redes Sociales.

Grandes cosas son las que podemos esperar de la convergencia entre el Big Data y la Vigilancia y la Inteligencia Competitiva: nuevos servicios y productos, transformaciones de negocios, reducciones de costes operativos, servicios más personalizados, empresas e instituciones más sofisticadas que incorporan la investigación y la innovación en sus cadenas de valor o empresas mejor gobernadas, más rentables y más sostenibles en el tiempo.

Europa encauza su impulso de digitalización a través de la Agenda Digital Europea[2], que tiene su reflejo en España a través de la Agenda Digital para España[3]. Uno de sus grandes bloques, el de la Economía Digital, incluye un apartado específico dedicado al Big Data[4]. Se estima que las inversiones en Big Data alcanzarán alrededor de los 2.5 billones de euros[5] entre 2016 y 2020. La Comisión Europea, a través de su programa Horizonte 2020 le ha destinado 500 millones de euros, cantidad que se espera se multiplique por cuatro gracias a la inversión privada, llegando a los 2 billones de euros. Otros mercados relacionados, como el mercado del Procesamiento de Lenguaje Natural (NLP) también están de enhorabuena: se espera que crezca desde los 3787.3 millones de dólares de 2013 hasta casi el billón de dólares para el año 2018[6].

---

[2] EC. Digital Agenda for Europe http://ec.europa.eu/digital-agenda

[3] EC. Agenda Digital para España http://www.agendadigital.gob.es/

[4] EC. Digital Economy: Making Big Data work for Europe http://ec.europa.eu/digital-agenda/en/big-data

[5] Public-Private Partnership (PPP) for Big Data http://europa.eu/rapid/press-release_MEMO-14-583_en.htm

[6] EC. Digital Agenda Web Site Press Releases: https://ec.europa.eu/digital-agenda/en/news/natural-language-processing-nlp-market-worldwide-market-forecast-analysis-2013%E2%80%932018 y Research and Markets. Natural Language Processing (NLP) Market - Worldwide Market Forecast & Analysis (2013–2018) http://www.researchandmarkets.com/research/3tl4zb/natural_language (October 2013)

*La **Vigilancia Estratégica y la Inteligencia Competitiva** también se beneficiarán de su convergencia con el Big Data ya que los negocios que desarrollan procesos de toma de decisiones e integran el conocimiento generado desde el Big Data **están incrementando su productividad entre un 5% y un 6%**.*

Big Data, la Vigilancia Estratégica y la Inteligencia Competitiva constituyen grandes oportunidades para las empresas españolas aunque también, como en todas las situaciones emergentes de mercado, presentan desafíos importantes. España cuenta con varias fortalezas que puede aprovechar:

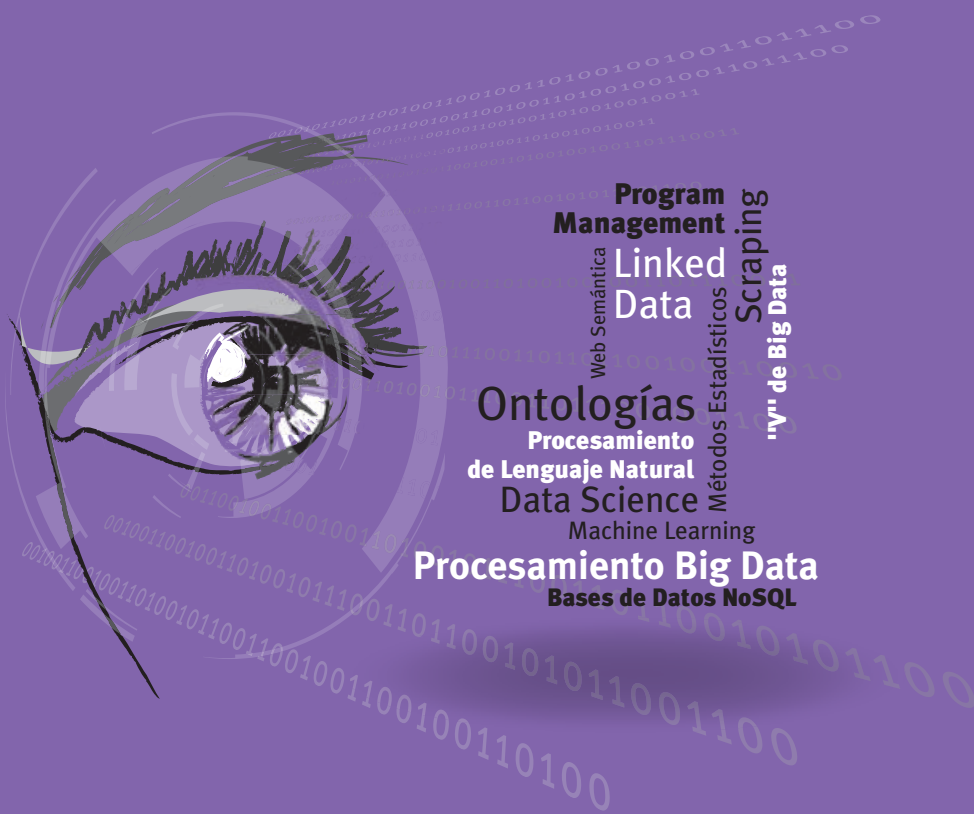
- Una más que notable red de investigadores en Universidades y Centros de Investigación en tecnologías y áreas de conocimiento están siendo impulsadas gracias al Big Data, como pueden ser las Ontologías, el "Machine Learning" o el Procesamiento de Lenguaje Natural (NLP).
- Prestigiosas Escuelas de Negocio que han puesto en marcha los primeros masters en Big Data de Europa.
- Una importante cantidad y calidad de profesionales del sector IT con formación universitaria en las tecnologías y áreas de conocimiento clave en Big Data como son la inteligencia artificial, programación funcional, estadística y matemáticas avanzadas citando las áreas más importantes.
- Grandes empresas en sectores que pueden obtener importantes beneficios del Big Data, como la banca, las telecomunicaciones, el sector público o la distribución.

Big Data puede ser un driver para el crecimiento y la internacionalización de las empresas españolas, que pueden crear una importante cantidad de empleos basados en nuestro país, sostenibles en el tiempo y de alta calidad. Es una industria de futuro, sin duda, pero que ya tiene un presente esperanzador y por el que merece la pena apostar.

EOI está haciendo una fuerte apuesta por el Big Data, destacando lo que fue la creación del primer máster en Big Data en Europa. Hoy dispone de hasta 4 programas Big Data, orientados cada uno a segmentos diferentes de profesionales, lo que la convierte posiblemente en la primera Escuela de Negocios de Big Data de España y uno de los líderes destacados en Europa.

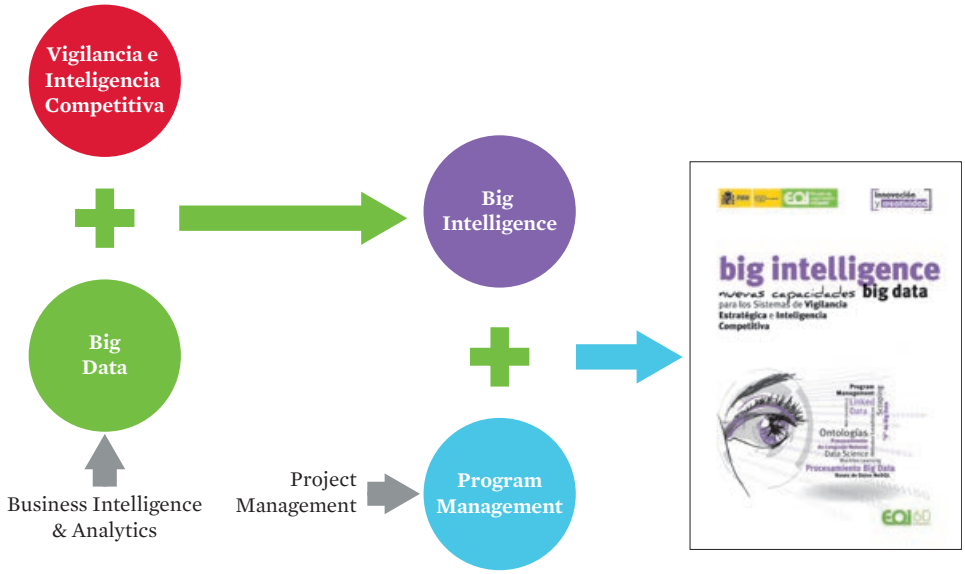
Este libro, que quedará disponible para su descarga desde nuestro repositorio de conocimiento Savia, es la penúltima aportación de EOI a la sociedad sobre Big Data porque sin duda no será la última.

PRESENTACIÓN



En las disciplinas de **Dirección de Programas** (en inglés "Program Management") y Dirección de Proyectos se denomina **Programa** al conjunto de **proyectos** interrelacionados que son gestionados de forma coordinada con el objetivo de obtener **Beneficios** no alcanzables si se gestionan de forma individual. Estos Beneficios proporcionan un conjunto de **Nuevas Capacidades** en la Organización en la que se implantan.

Se postula que las grandes empresas de Internet han creado un **Nuevo Mercado** cuyos productos y servicios son el fundamento de un término paraguas que llamamos **Big Data**, que le da nuevas alas a las actividades y procesos que suelen englobarse en los conceptos de **Vigilancia Estratégica e Inteligencia Competitiva**. A esa fusión de Big Data aplicado a la Vigilancia Estratégica e Inteligencia Competitiva lo hemos venido a llamar en este libro **"Big Intelligence".**



Estas cuatro ideas fuerza **"Big Intelligence", "Nuevas Capacidades"**, "**Big Data"** y "**Vigilancia e Inteligencia Competitiva"** le dan nombre a este libro. **La puesta en marcha de Programas Big Data de Vigilancia e Inteligencia Competitiva en las Empresas e Instituciones les proporcionarán Nuevas Capacidades que hasta muy recientemente la tecnología no ha hecho viable.**

Tras la presentación general, se introduce los conceptos principales relacionados con la Vigilancia e Inteligencia Competitiva. A continuación se presentan las tecnologías, disciplinas y áreas de conocimiento más relevantes que se suelen englobar bajo el término Big Data, proponemos la evolución de los Sistemas de Vigilancia e Inteligencia Competitiva mediante Big Data y finalmente se presenta una metodología para el diseño del Sistema, un modelo funcional y un modelo organizativo que lo soporte.

Un Nuevo Mercado para el siglo XXI

Un **Nuevo Mercado sostenible en el tiempo** se ha generado encabezado por las grandes empresas de **Internet**, como Google, Facebook, Amazon, Yahoo! o Twitter, y a la que han seguido con fuerza las grandes empresas intensivas en el uso de información en su cadena de valor, como las de los sectores de la telecomunicación o la banca. Este nuevo mercado ha llegado para quedarse y para seguir evolucionando. Esta situación nos aporta sostenibilidad y por tanto podemos considerarla no moda ni flor de un día sino tecnologías que han venido para permanecer y para ser incorporadas en procesos de negocio.

Además de las grandes empresas de Internet, las grandes corporaciones del sector IT, como Apple, IBM, Oracle o Microsoft están incorporando tanto nuevas empresas a través de adquisiciones, como personal especializado en las tecnologías y áreas de conocimiento que son base de las **Nuevas Capacidades de los Sistemas de Vigilancia e Inteligencia Competitiva.**

Estas nuevas capacidades las proporcionan toda una serie de tecnologías, disciplinas y áreas de conocimiento que se agrupan bajo el término paraguas **Big Data** y que trataremos en este libro, entre ellas las **ontologías, el machine learning, la inteligencia artificial, el procesamiento de lenguaje natural, el procesamiento distribuido o la estadística**. De hecho varias empresas ya tienen disponibles diversas herramientas con las que pueden implementarse no ya sólo las nuevas capacidades que planteamos sino entornos completos con los que construir un Sistema de Vigilancia Avanzado.

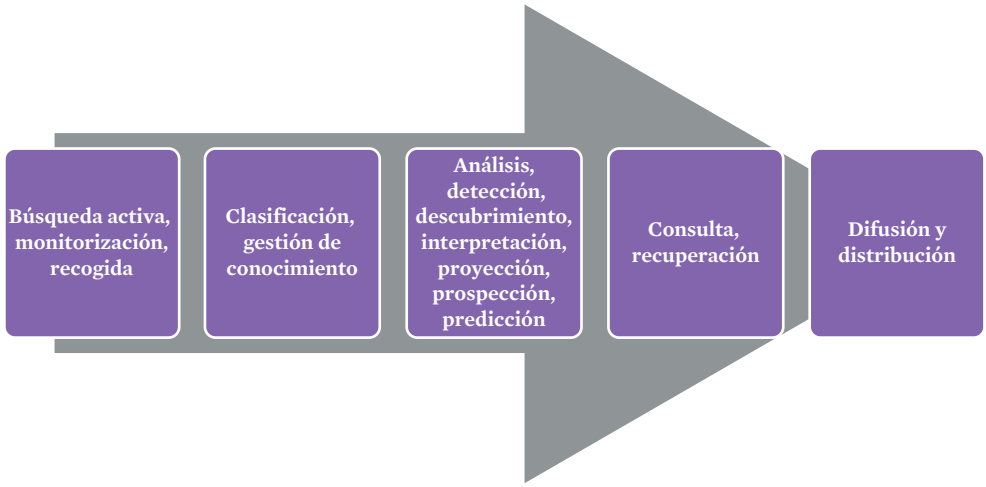
Los Sistemas de Vigilancia e Inteligencia Competitiva

Big Data está caracterizado por tecnologías y paradigmas renovados que explotan grandes repositorios de información y aprovechan la evolución del hardware y el software. Más que una evolución, señalan la viabilidad de realizar **transformaciones significativas en los Sistemas de Vigilancia Estratégica e Inteligencia Competitiva.**

Resulta plausible por tanto, ya que empieza a ser plenamente viable, que se incorpore la función de Vigilancia Estratégica e Inteligencia Competitiva dentro de los sistemas de información de las empresas, bien como una aplicación IT como las actuales, bien incorporando funciones de vigilancia a las grandes aplicaciones IT, los ERPs, los CRMs o los SCMs. El recorrido que tiene este mercado es muy importante.

Toda la cadena de actividades que ejecutan los **Sistemas de Vigilancia Estratégica e Inteligencia Competitiva** queda fuertemente afectada por estos cambios: **la búsqueda activa, monitorización, recogida, clasificación, análisis, detección, descubrimiento, interpretación, proyección, prospección, predicción, gestión de conocimiento, consulta, recuperación, difusión y distribución.** Consecuentemente se modifican al alza las expectativas, los objetivos y las funciones que ahora cumplen los Sistemas de Vigilancia Estratégica e Inteligencia Competitiva.

ILUSTRACIÓN 1
Cadena de Actividades de la Vigilancia e Inteligencia Competitiva



Es posible, y esta posibilidad se ha convertido en una necesidad competitiva, expandir el conocimiento en la cadena de valor de empresas e instituciones públicas incorporando información de vigilancia, conocimiento e inteligencia competitiva sobre **entidades de negocio clave**: tecnologías, productos, procesos, servicios, proveedores, competidores, clientes, empresas, mercados, sectores, áreas de conocimiento, instituciones, empleados, stakeholders, financiación, investigación y formación.

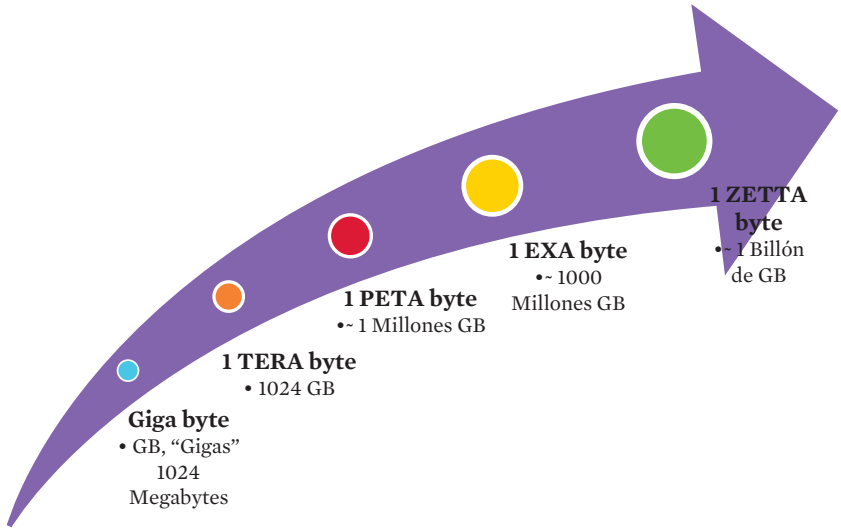
Cada vez más empresas e instituciones comprenden que deben evolucionar para poder competir mejor en un entorno cada vez más globalizado, competitivo y cambiante. Para ello sofistican sus cadenas de valor, introduciendo en la misma la **gestión de la innovación** y explicitando la **gestión estratégica** de sus organizaciones. La misma adopción exitosa por parte de su competencia de estas mejoras, para las que son clave los procesos de vigilancia e inteligencia competitiva, son acicate para que otras inicien su implantación. Identificar sectores, tecnologías, productos y servicios emergentes, determinar su impacto económico y cómo afecta a la competitividad de

nuestras organizaciones; determinar oportunidades susceptibles de ser viables para una estrategia determinada; proporcionar a tiempo información relevante junto con su valor estratégico que permita reducir los riesgos y anticiparse a los cambios en los procesos de toma de decisiones; diseñar políticas vinculadas a la sostenibilidad económica en el medio y largo plazo. Todo esto ahora **ya es clave para la dirección de las empresas y las instituciones**.

Big Data y los fundamentos del Cambio

Una de las herencias más importantes que el siglo XX le dejó al siglo XXI fue el acceso a la información. En los años 90 los "*Peta y Zetta*" rememoraban únicamente una inolvidable golosina. Hoy hemos pasado a ver de cerca estos conceptos y a hablar con naturalidad de Terabytes, *Petabytes,* Exabytes, *y Zettabytes.*

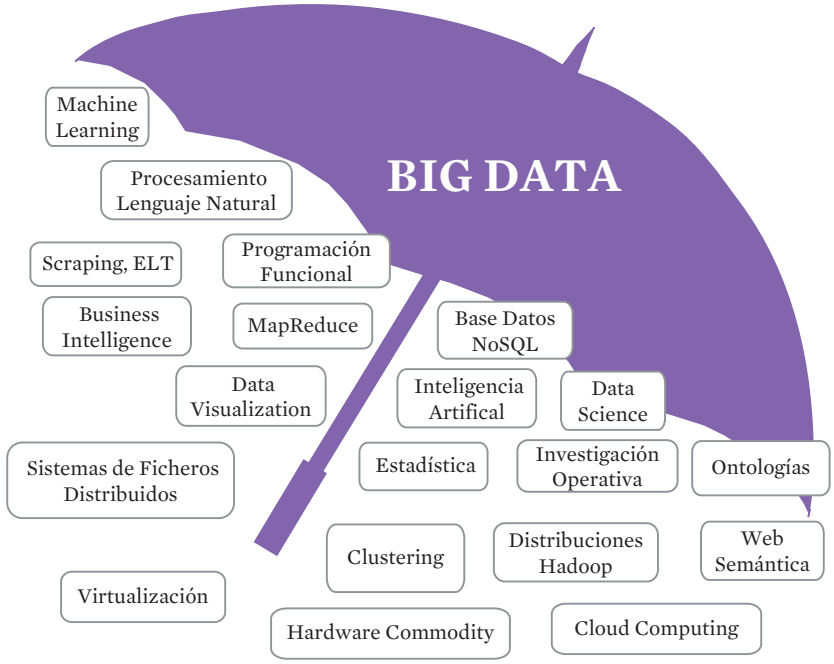
Se calcula que **la información que está disponible hoy a través de la web es de unos 10 Zettabytes (más de 1 billón de gigabytes)**. **Este es el primer fundamento del cambio que estamos viviendo.** Esta estimación se multiplica por 500 si tenemos en cuenta la Internet profunda (en inglés "*Deep Web*"), es decir la información no accesible directamente por buscadores. Textos, sonidos y vídeos de los grandes medios de comunicación, bases de datos abiertas, enormes recursos como Google o la Wikipedia, las páginas web de las empresas, páginas web personales y las grandes Redes Sociales son parte de ese enorme conglomerado de información.



La capacidad de cómputo del hardware y el software crece exponencialmente. Hoy en día tenemos en nuestro bolsillo, concretamente en nuestros modernos teléfonos móviles, más capacidad de cómputo que los ordenadores de la NASA que llevaron al hombre a la luna. Los ordenadores personales de los que disponíamos a finales de los años 90 son hoy tristes antiguallas, apenas útiles más que en exposiciones de juegos retro.

El siglo XXI nos ha traído nuevas técnicas y las nuevas capacidades del hardware y del software **que nos hacen posible usar ahora viejos paradigmas informáticos de altas capacidades** que hasta hace pocos años eran computacionalmente inviables.

Estas nuevas tecnologías pueden habilitar **Nuevas Capacidades para las organizaciones** fundamentadas en el término paraguas **Big Data**, materializadas en servicios, funciones u operaciones nuevas o muy mejoradas. La implementación de estas nuevas capacidades pueden **conseguir como resultado importantes beneficios**.



Viejas promesas de **la inteligencia artificial** como el **Machine Learning** o el **Procesamiento de Lenguaje Natural** (**PLN**) son hoy drivers de desarrollo de soluciones nuevas, asequibles y potentes para problemas cuya solución hasta ahora o bien ha requerido enormes recursos o bien han sido muy pobres. También paradigmas, como el de la **programación funcional,** que sus exigentes necesidades de cómputo relegaban a los laboratorios universitarios y a entornos restringidos o las **Bases de Datos VLDB** (del inglés "*Very Large DataBases*"), son hoy parte constituyente de esta revolución tan

importante que llamamos **Big Data**. El ejemplo más claro de ello es el de las funciones "Map" y-"Reduce", combinadas en el modelo de programación funcional **MapReduce** y popularizadas por su implementación en el proyecto **Apache Hadoop,** posiblemente el proyecto que ha hecho viable la adopción masiva del Big Data.

Big Data como paradigma también nos ha aportado Sistemas de Archivos Distribuidos y escalables y **nuevos sistemas de gestión de bases de datos** preparados para dar respuesta a la necesidad de manejar grandes volúmenes de información de forma distribuida. Ejemplos hoy de rabiosa actualidad son las **Bases de Datos NoSQL**, entre las que destacan las Orientadas a Columnas, las de Clave-Valor, las orientadas a la Gestión de Documentos, Objetos o Grafos.

Un enfoque emergente para el tratamiento de los textos desestructurados de las páginas web es leerlas mediante aplicaciones software deduciendo de este proceso su contenido, su estructura y su semántica. Para ello, además de la técnica de **Scraping**, usamos un campo de la inteligencia artificial que llamamos **Procesamiento de Lenguaje Natural**, para el que se usa tanto el acrónimo en inglés (*NLP, "Natural Language Processing")* como en español (PLN). Sus objetivos son tanto comprender el lenguaje humano como generar respuestas en lenguaje humano coherentes con un contexto dado, como pueda ser una pregunta. Actualmente el uso de PLN es relativamente primitivo. Un ejemplo de ello es el uso que hacemos de un buscador para hacer una búsqueda, para lo cual introducimos nada más que palabras clave con las que tenemos la esperanza de que en la lista de respuestas que nos genere el buscador encontraremos alguna referencia que nos ayude a dar respuesta a nuestras preguntas y necesidades reales. En ningún caso escribimos una pregunta en la caja del buscador ni esperamos que el buscador extraiga de varios documentos un resumen estructurado conteniendo las respuestas más relevantes a la pregunta hecha.

Durante muchos años a PLN le han faltado tanto las necesidades de mercado como un conjunto suficiente de textos con los que trabajar. La explosión de Internet ha ejercido de catálisis que ha habilitado este mercado. Los **análisis de sentimiento o de opinión** en Redes Sociales son buenos ejemplos de aplicaciones, muy populares hoy en día en las que PLN está proporcionando las mejores soluciones. Estamos todavía muy lejos de hacer una comprensión profunda de un texto complejo y por tanto de poder disponer de un software inteligente capaz de realizar pensamientos complejos. Sin embargo el recorrido que veremos en el medio y largo plazo en PLN promete ser muy amplio.

Los otros enfoques emergentes son los del **Aprendizaje Automático**, popularmente conocido por su denominación en inglés, ***"Machine Learning"***, y los **Métodos Probabilísticos y Estadísticos.** Estos dos enfoques, aplicados tanto a textos desestructurados como a datos masivos, proporcionan resultados novedosos aplicados a los procesos analíticos, prospectivos y predictivos.

En ***Machine Learning*** utilizamos conjuntos de información y un algoritmo para entrenar a una aplicación. Una vez entrenada, cada vez que necesitemos analizar una nueva información dicha aplicación clasificará la nueva información a partir del entrenamiento recibido.

En el algoritmo de entrenamiento podemos estar utilizando tanto los métodos probabilísticos y estadísticos mencionados anteriormente como otras técnicas de inteligencia artificial como redes neuronales, árboles de decisión, etc.

Los **métodos probabilísticos y estadísticos** nos van a ofrecer un modelo de referencia para un conjunto de datos, gracias al cual podamos clasificar una nueva información ofreciendo una **predicción** a partir de dicho modelo. Estos modelos se aplican tanto a datos numéricos como a conjuntos de palabras dentro de documentos. Son aplicados actualmente, por ejemplo, por los grandes buscadores de Internet para determinar qué documentos son más relevantes para una búsqueda dada.

Para agrupar todo este conocimiento que se está concentrando en torno al término de Big Data ha emergido el concepto de **Data Science.**

Las implementaciones Big Data serían imposibles sin las nuevas capacidades de los ordenadores actuales, que han evolucionado enormemente tanto en el hardware como en el software. La reducción del coste de hardware ha sido enorme en estos últimos años, llegando a convertir en ***hardware commodity***[7] a sistemas cuyo coste a final de siglo pasado era superior al millón de euros. Además de la capacidad de procesamiento, el **Almacenamiento** es el otro punto en el que el hardware ha evolucionado: el coste de un dispositivo de 1Gb de capacidad ha disminuido de 300.000 € en 1980, a unos 10 € en el año 2000 y a apenas unos céntimos en la actualidad.

En cuanto al **software** las claves están en la evolución y mejora de los sistemas operativos y en la **Virtualización**, encarnada en las **Máquinas Virtuales**, un software capaz de emular a una computadora, pudiendo ejecutarse en un mismo ordenador varias máquinas virtuales.

Ambas evoluciones, de hardware y software, han habilitado una **Paralelización** potente y fiable, haciendo posible poner a funcionar en paralelo cientos o miles de estos ordenadores que, aplicando el viejo lema de Julio César **"divide et vinces",** *divide y vencerás*, separan los problemas en multitud de pequeños problemas fáciles de solucionar y luego integran todas esas pequeñas soluciones en la solución final del

[7] Hardware Commodity: http://www.webopedia.com/TERM/C/commodity_hardware.html

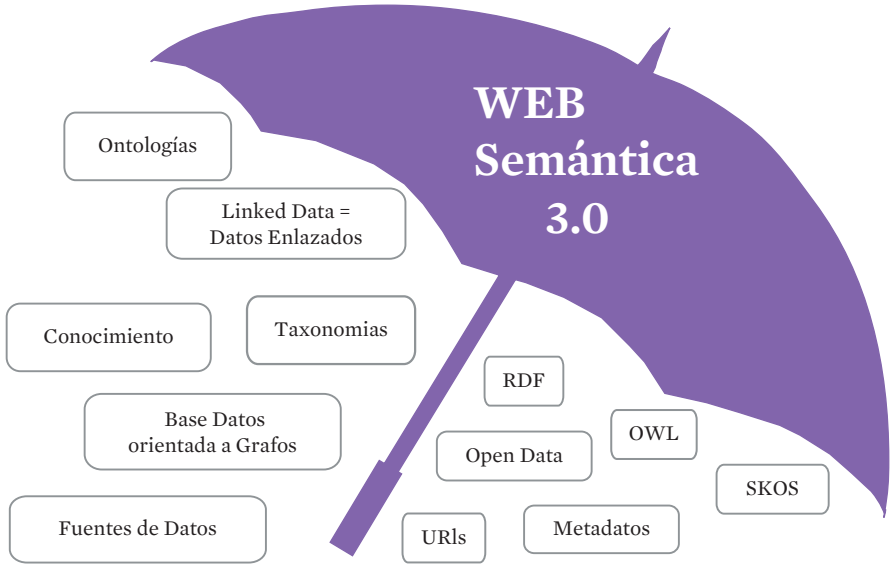
problema planteado, todo ello realizado en un intervalo de tiempo pequeño. A este tipo de sistemas lo llamamos **Sistemas Distribuidos**.

Gracias a todo esto se ha habilitado la posibilidad de que en grandes centros de datos se implementen todas estas nuevas capacidades de cómputo y se le ofrezcan nuevos servicios al mercado. A este otro paradigma lo llamamos "**Cloud Computing**", computación remota, en definitiva.

Por último hay que citar el concepto de **Software Libre**. El uso intensivo de proyectos de software libre, entre los que destaca el **Apache Hadoop**, ha hecho posible esta revolución. Las grandes empresas de internet han promovido y hecho uso masivo de software libre principalmente por su capacidad de adaptación rápida a sus nuevas necesidades, pero también hay que mencionar que el reducido o inexistente coste de licencias del mismo ha posibilitado la viabilidad económica de estas empresas.

La WEB Semántica

Internet, y **la evolución de la World Wide Web** son dos de las ideas que dan sentido a este libro. Como todo sistema emergente, la web actual está adaptada a las necesidades y oportunidades que la hicieron nacer. Sin embargo nuevas necesidades están emergiendo, cada vez con más fuerza. La siguiente evolución que viene es lo que llamamos "**Web 3.0"**, o más conocida como **Web Semántica.**



Solemos reunir bajo el concepto de **Web Semántica** la idea de **añadir el conocimiento a las páginas web** mediante **metadatos semánticos y ontológicos**. Las aplicaciones actuales, y por ende los sistemas dedicados a tal fin, tienen cada vez más funciones de Vigilancia y requieren capacidades más sofisticadas que permitan extraer información y conocimiento, no sólo datos, de las páginas web. En la nueva Web una nueva aplicación software debe poder conectarse a nuestra página web y extraer de forma automatizada información que ahora mismo únicamente es leída de forma totalmente fiable por personas.

Las **Ontologías** son los mecanismos que nos van a proporcionar fiabilidad en cuanto a la semántica de lo expresado, lo que queremos comunicar. Las ontologías son descripciones de conocimiento, esquemas conceptuales en dominios de información concretos. Estos esquemas nos van a permitir clasificar el conocimiento y razonar sobre él de forma automatizada. A la hora de publicar la información en internet vamos a hacer uso de estas ontologías para asegurarnos de que estamos expresando la información de forma unívoca y que va a ser reconocida universalmente como tal. Para ello usaremos los **lenguajes de ontologías** de los que RDF (Resource Description Framework), RDFS (RDF Schema) o OWL (Web Ontology Language) son buenos y relevantes ejemplos.

Actualmente existen numerosas **ontologías públicas** que se han convertido en estándares de facto en diversos dominios y son aceptadas como tal por empresas, instituciones y personas. Dos ejemplos serían **FOAF**[8] (del inglés *"Friend of a Friend"*, acrónimo **foaf:**), que describe actividades y relaciones entre personas y objetos, así como a las personas en sí. Otros ejemplos podrían ser **GoodRelations**[9], (acrónimo **gr**) orientado a la descripción de productos y servicios de una empresa, y **Open Graph**[10] (acrónimo **og:**), pensado para facilitar que una página web tenga propiedades de redes sociales. Utilizando estos estándares nos aseguraremos de estar usando una semántica común globalmente aceptada. La manera más adecuada de responder a nuestras necesidades, en caso de superar la semántica contenida en estas ontologías públicas, será extender la ontología.

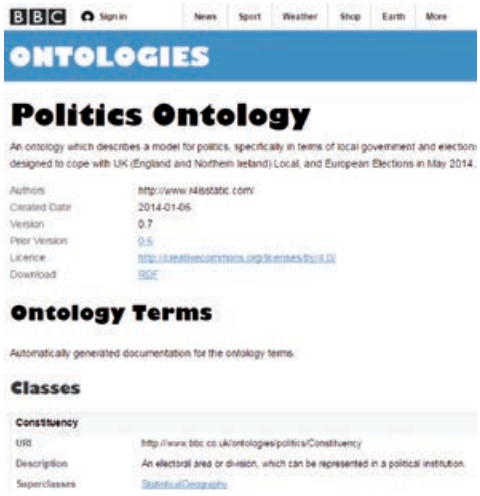
---

8 FOAF: http://xmlns.com/foaf/spec/

9 Good Relations: http://www.heppnetz.de/projects/goodrelations/

10 Open Graph http://ogp.me/

Los lenguajes de ontologías presentan tripletas (sujeto, predicado, objeto). Por ejemplo que "el teléfono de contacto (ficticio) de una persona de nombre Jaime García es el "913495600" se representaría como ("Jaime García", foaf:phone, 913495600) o que "el título del libro con ISBN 978-84-95598-65-3 es las Aventuras de Don Quijote de la Mancha" se podría representar como ("978-84-95598-65-3, og:title, "Las Aventuras de Don Quijote de la Mancha").



Otro pilar relevante de esta nueva Web son los **Linked Data**, término acuñado por Tim Berners-Lee, que suele traducirse como **Datos Enlazados**, un método de publicación de datos estructurados que permite que sean enlazados y accesibles de forma sencilla tanto para personas como automáticamente por programas software. Por ejemplo la **Biblioteca Nacional de España** dispone de su portal de **Linked Data** en http://datos.bne.es/ con información enlazada de Autores, Obras y Temas.

Linked Data se basa en el concepto de URI (Uniform Resource Identificator), similar al bien conocido de URL (Uniform Resource Locator). Un URI identifica un recurso en Internet mediante una dirección http. Existen ya multitud de URIs accesibles: por ejemplo la popular BBC de Londres dispone de multitud de ontologías con URIs para los conceptos que gestiona. Por ejemplo en la Ontología " BBC Sport"[11] para el concepto "Competition", presenta la URI http://www.bbc.co.uk/ontologies/sport/Competition y la descripción *"A competitive sporting event that usually appears as an occurrence of a recurring competition, for example the recurring English Football Premier League has a seasonal competition occurrence during 2012/13".*

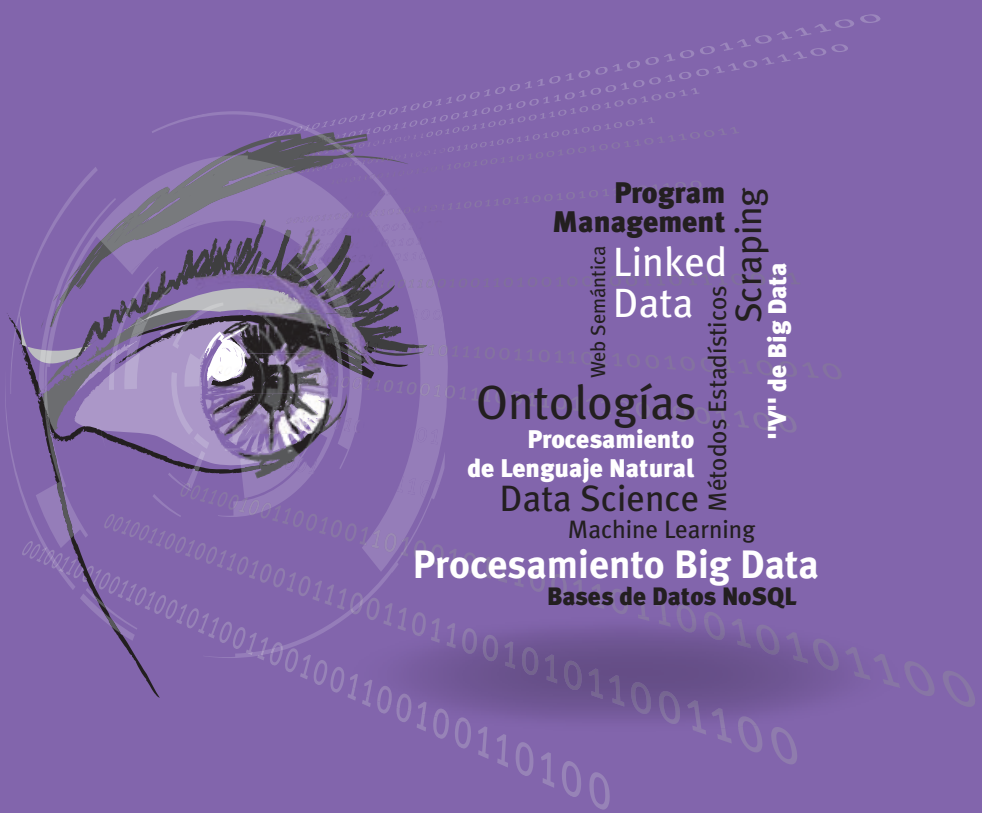
La otra dinámica interesante es la de **Open Data**. En los últimos años se ha promovido la idea de poner datos relevantes a disposición de forma accesible y reutilizable en internet. Instituciones públicas y privadas han abierto sus repositorios por lo que se han multiplicado el número de **Fuentes de Datos** disponibles y por ende la capacidad de ofrecer mejores respuestas a necesidades y casos de uso de los Sistemas de Vigilancia. En unos casos son meros ficheros publicados en bruto; en otros casos nos

[11] BBC – Sport: http://www.bbc.co.uk/ontologies/sport

encontramos en el límite con datos publicados con URIs en RDF e integrados con ontologías públicas y enlazando datos de terceros.

Esta multiplicación de las Fuentes de Información trae asociado otro elemento clave: las **Taxonomías** que clasifican los datos. Tener la misma taxonomía clasificando diferentes fuentes permite asociar los datos entre dichas fuentes, lo cual puede ser de gran utilidad. Sin embargo la explosión del número de fuentes a explorar en un Sistema de Vigilancia puede conllevar también la necesidad de integrar un número creciente de taxonomías, dejando además sin resolver otro de los grandes problemas existentes: la diferente granularidad en la clasificación en diferentes taxonomías o la diferente granularidad existente en la taxonomía frente a la necesidad de vigilancia.

VIGILANCIA ESTRATÉGICA
E INTELIGENCIA COMPETITIVA



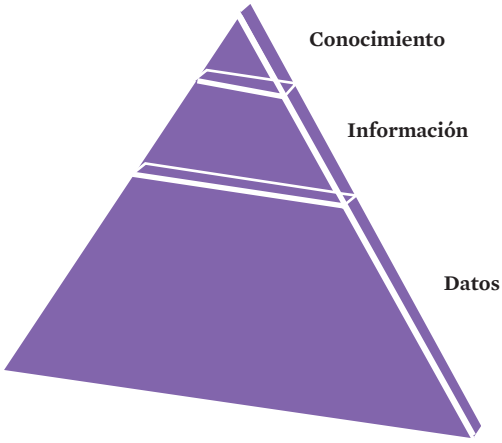
Presentamos en este capítulo el marco conceptual en torno a la Vigilancia Estratégica y la Inteligencia Competitiva, sus objetivos, procesos y la normativa que recoge las mejores prácticas para su puesta en marcha y gestión.

1. Vigilancia Estratégica, Inteligencia Competitiva y Gestión del conocimiento en el siglo XX[12]

1.1. Antecedentes

La construcción de un modelo del mundo que permita tomar decisiones y anticiparse al entorno forma parte del proceso del conocer y entronca con la evolución misma del sistema nervioso.

Muchas teorías sostienen que efectivamente, el sistema nervioso surge evolutivamente como una solución biológica al problema de la relación con el mundo, comenzando con la optimización de las reacciones al medio (movimiento) a partir de la información captada por los sentidos. Así surgió el sistema nervioso primitivo.



A medida que este sistema evolucionaba y se volvía más complejo, comenzaron a emerger nuevas propiedades de esa complejidad y el sistema nervioso fue poco a poco evolucionando hasta convertirse en un verdadero cerebro. A su vez, este cerebro fue también incrementando su nivel de complejidad, dando lugar a la aparición de la capacidad para el razonamiento simbólico y el pensamiento abstracto, dos aspectos fundamentales de la capacidad de creación de modelos del mundo.

Se suele explicar el proceso del conocimiento mediante una pirámide que tiene en su base lo que llamamos datos y que termina en la cúspide en lo que llamamos conocimiento o, yendo un paso más allá, lo que otros autores denominan “sabiduría”, que estaría un nivel por encima del conocimiento.

[12] Texto original de Juan Jiménez Morillas.

Los **datos** representan valores asignables a características propias de los objetos o de la realidad. En sí mismo no tienen ninguna utilidad, por eso algunos autores (Zeleny[13]) dicen de ellos que son "ignorantes", y no son más que medidas, el producto de una observación, y que no tienen ningún valor si no se vinculan con un contexto que permita interpretarlos. Un ejemplo podría ser una medida de temperatura: 24 °C.

La **información** contiene en sí misma un cierto grado de significado, incluso de propósito. Es importante señalar (de ahí la imagen de la pirámide para describir el proceso), que la información se desprende de los datos, y que son elementos externos a los propios datos (el contexto) lo que permite dotarles de un significado y que se conviertan, individualmente o por agregación, en información. En nuestro ejemplo, si decimos que los 24 °C representan la temperatura de la habitación en la que estamos, ya tenemos información. Mientras que los datos son objetivos, **la información representa un estado cognitivo**.

Progresando más hacia la cúspide de la pirámide, **cuando la información se procesa y se organiza de determinada manera, que no es unívoca**, llega a constituirse en **conocimiento**, una de cuyas características es representar un estado totalmente subjetivo. Podría decirse que los datos están en el mundo mientras que el conocimiento está en los cerebros. El conocimiento tiene en cuenta no sólo el contexto, sino la experiencia e incluso los valores de su propietario, ya que emerge a partir de una interiorización de los datos y una asignación subjetiva de valor de la información puesta en relación con otra información. En nuestro ejemplo, los 24°C tendrán un valor muy diferente si la habitación es un dormitorio o si la habitación es una cámara frigorífica. La experiencia nos puede indicar la existencia de un problema serio en el segundo caso, e incluso nos vendrán a la mente casi de inmediato las posibles causas y acciones paliativas a tomar.

Cuando el peso de lo subjetivo es máximo, así como la aplicación de las funciones de la actividad mental que llamamos "juicio", ya estamos hablando de lo que hemos mencionado anteriormente que algunos denominan sabiduría. A los efectos de "gestión del conocimiento" que nos interesan aquí, nos quedaremos con los tres niveles que aparecen en la pirámide. **El nivel de sofisticación que más nos interesa estriba en la capacidad predictiva basada en cierto modelo del mundo. El conocimiento entronca ese nivel superior de sofisticación (la sabiduría)** con otro tipo de cuestiones, como las relacionadas con los valores humanos, las estrategias y las motivaciones, elementos cuyo análisis queda fuera del alcance de esta obra y que en todo caso se relacionan con la motivación y el diseño estratégico de los sistemas de vigilancia.

[13] Milan Zeleny es profesor de de sistemas de gestión en la Universidad de Fordham y autor de Sistemas de apoyo a la gestión: hacia un sistema integrado de gestión del conocimiento.

1.2. La Vigilancia tecnológica como actividad clave para la innovación

Enmarcando la actividad de la vigilancia en el contexto que acabamos de describir, la Vigilancia sería un **proceso proactivo de captura de datos y contextualización que permita la generación de conocimiento**. El producto resultante del proceso permite por tanto a un agente humano con criterio integrar la información y adquirir conocimiento. Así, estrictamente, la vigilancia no produce conocimiento ella misma: detecta cambios en el entorno y los contextualiza, pero la tarea de dotar de valor a esos datos depende de un agente externo al proceso de vigilancia colocado al final de la cadena. Esto requiere de una persona con capacidad para aportar ese valor añadido que salva la frontera entre el conocimiento y la información.

Para nombrar ese rol ha emergido el concepto de "Curador de Contenidos" (traducción literal del inglés "content curator") que tiene sus raíces en actividades habitualmente realizadas por documentalistas y bibliotecarios. Su papel resulta clave para que el producto final sea verdadero conocimiento.

En este punto, cabe hacer una distinción entre el proceso de vigilancia, como entidad propia y autónoma, y la vigilancia tecnológica como un todo. El proceso de vigilancia describe la captación de datos, su conversión en información y su transmisión hasta el intérprete. El intérprete es "el que vigila". Dado que el producto típico de la vigilancia tecnológica es algún tipo de informe, que contiene la interpretación de la información, puede producirse confusión entre el proceso de vigilar y la vigilancia tecnológica propiamente dicha. En este sentido, el proceso de vigilar termina en la información y la vigilancia tecnológica, como actividad, termina en el conocimiento y su difusión. Para simplificar, cuando se habla de vigilancia se suele hablar del proceso completo, como veremos más adelante.

Como veremos, el proceso de Vigilancia Tecnológica tiene mucho en común con el proceso de adquisición de conocimiento desde los datos que se ha ilustrado en la figura de la pirámide. Pero antes, estaría bien poner en valor el proceso de relevancia relacionándolo con una actividad clave: la innovación.

1.2.1. Algunas reflexiones sobre la innovación

Se dice con frecuencia que la innovación no es más que hacer las cosas de una manera diferente. Esta definición es desde luego tan amplia que admite casi cualquier tipo de actividad, desde diseñar un material con propiedades nuevas que poder aprovechar, hasta volver a casa dando un paseo en lugar de utilizando el autobús por un mero impulso. En este sentido, conviene restringir un poco el concepto e introducir al menos unas pinceladas de algo más que podemos conceptualizar como "intención", lo que

nos permite acotar la innovación como una actividad más específica sin dejar por ello de tener una definición amplia. Se hacen las cosas de manera diferente con el fin de mejorar. El fin último es la obtención consciente (o la maximización) de un beneficio, entendido en sentido amplio (mejorar la salud, en el caso de haber elegido el paseo para volver a casa como excusa para hacer ejercicio).

La innovación es, por lo tanto, **un cambio motivado.** Si nos centramos en el nivel de las organizaciones, esta innovación puede referirse tanto a los procesos como a los productos y servicios que ofrecen. Al mismo tiempo, se habla de innovación incremental, que es la más común, y que se basa en la incorporación de pequeñas mejoras dentro de un modelo existente, y de innovación rupturista, que es aquella cuyas consecuencias son, en principio impredecibles, y que cambian el comportamiento de la demanda y de los productores, los modos de vida, los hábitos de consumo, y dan lugar a modelos nuevos.

Cabe decir que dentro de una organización, la capacidad de innovación se relaciona estrechamente con la cultura de gestión de la misma, con sus valores y con la capacidad de asumir riesgos. Parece razonable pensar que la innovación tiene más éxito (como proceso específico) allá donde se dan las condiciones favorables para su desarrollo. Culturas con exceso de control, falta de libertad o donde el coste de asumir riesgos sea demasiado elevado lo que tienden a fomentan es, precisamente, que las cosas se hagan siempre de la misma manera.

Innovación sistematizada: los sistemas nacionales (no sólo)

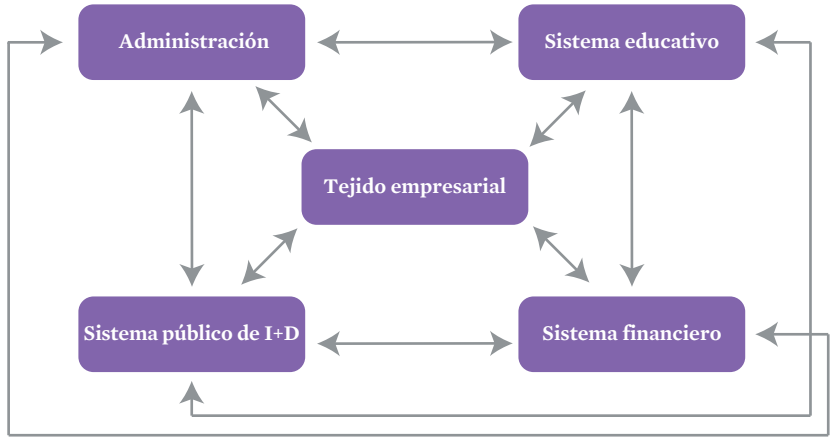
El Sistema Nacional de Innovación es la red de instituciones del sector público y el sector privado cuyas actividades e interacciones contribuyen a lanzar, a importar, a modificar y a difundir nuevas tecnologías. (Freeman, 1987)[14]

Si nos elevamos en la jerarquía de las organizaciones hasta el nivel de los estados o entidades nacionales, estas condiciones internas de las que acabamos de hablar tienen su reflejo en distintos elementos que de manera formal o informal constituyen el sistema nacional de innovación, entendiendo que nos estamos refiriendo ya en concreto a la capacidad y la actividad productora de conocimiento que pueda servir como base para innovar.

Si definimos "tecnología" como la aplicación del conocimiento científico a la resolución de problemas prácticos, entenderemos que la tecnología es uno de los elementos de la innovación. Dentro del sistema, se puede acceder a la tecnología por distintos

[14] Christopher Freeman fue un economista británico, teórico de los ciclos económicos.

caminos: **Desarrollo, Compra, Transferencia desde el sistema académico, Adquisición incorporada a equipos, Ingeniería inversa...**



En este marco es donde aparece la **Vigilancia Tecnológica,** como actividad cuyo fin es **identificar señales emergentes de cambio relacionadas con la tecnología.** Los ámbitos a considerar aparecen ya reflejados en el diagrama del sistema nacional de innovación: la actividad pública, la actividad privada (ámbito empresarial) y el mundo académico. Cuando se añade a este proceso la dimensión económica y de negocio, se habla de Inteligencia Competitiva. El proceso tiene en cuenta los productos de conocimiento que emanan de cada uno de estos ámbitos: patentes, publicaciones y productos y servicios en el mercado. La finalidad es conseguir ventajas competitivas gracias al establecimiento de un proceso sistematizado de detección, captación y análisis de la información para generar un conocimiento apto para la toma de decisiones.

1.3. Sistemas de Vigilancia Estratégica e Innovación Competitiva

Hemos dicho que una de las características fundamentales de la Vigilancia es que se trata de un proceso sistematizado, que por lo tanto consta de una serie de etapas que cierran un ciclo, en el que el conocimiento generado sirve a su vez para interpretar la información en las iteraciones posteriores. Por lo tanto, no sólo es un proceso sistematizado, sino también continuo.

1.3.1. Objetivos de los usos de la Vigilancia

Vamos a señalar seis objetivos que engloban la mayor parte de los usos de la Vigilancia, sin ánimo de limitar el alcance, sino con la intención de ilustrar sus posibilidades

y destacar los que consideramos más importantes. La Vigilancia es un instrumento, y está por lo tanto al servicio de una estrategia o de un fin más alto, que puede estar relacionado con la actividad de organizaciones empresariales o con la gobernanza del país; con la necesidad, en definitiva, que tienen las organizaciones de realizar una asignación racional de los recursos con el fin de alcanzar unas determinadas metas consideradas estratégicas.

1. **Identificar cambios en el entorno**, entendiendo por el entorno sectores de actividad económica, conjuntos de tecnologías asociadas a esta actividad, productos y servicios disponibles en el mercado y señales débiles o emergentes de que están a punto de producirse cambios.

2. Estar al tanto y **conocer estos cambios** no sólo en nuestro entorno propio, sino **también en el entorno próximo** (otros países y áreas económicas).

3. **Reducir la incertidumbre y por lo tanto el riesgo en los procesos de toma de decisiones**, identificando dónde queremos (y podemos) posicionarnos estratégicamente.

4. **Dilucidar caminos de evolución del Sistema**, al identificar nuevas necesidades de clientes, usuarios y ciudadanos en general.

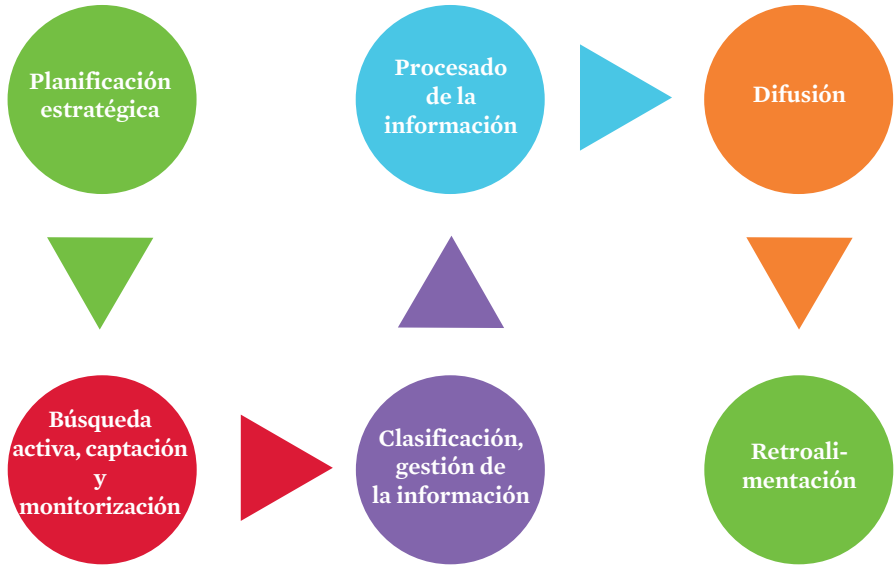
5. **Identificar nuevas tendencias** que permitan realizar innovaciones en los procesos, los productos, la gestión del talento y del capital humano...

6. **Conocer la competencia**, descubrir posibles alianzas con nuevos socios e identificar expertos a los que poder solicitar asesoramiento.

1.3.2. Ámbitos

Los ámbitos de la actividad ya se han esbozado en la descripción del marco conceptual de la Vigilancia: sectores de actividad, productos y servicios ofrecidos en el mercado, tecnologías y avances científicos (patentes, publicaciones) y personas (identificación de expertos).

1.3.3. Etapas del proceso



Planificación estratégica

Sin duda la piedra de toque del procedimiento es determinar qué se quiere vigilar y para qué, porque esto condicionará el resto del proceso: asignación de recursos, fuentes de información, procesado de la misma, etc.

Una definición insuficiente de los fines en esta etapa suele producir problemas en etapas posteriores.

El producto de esta etapa se concreta en una serie de factores críticos de vigilancia, que son los elementos que se considera clave tener controlados y que está relacionados con los ámbitos ya señalados: patentes, productos y servicios, personas, empresas de un sector, otros agentes, etc.

Una vez identificados los factores críticos de vigilancia, es necesario parametrizarlos: recordemos que el proceso del conocimiento se apoya en los datos, y los datos son el fruto de medidas. Estos parámetros debe ser medibles, y su evolución en el tiempo es también un importante elemento de análisis.

Ejemplos de parámetros pueden ser los siguientes:

- Número de patentes anuales realizadas en un determinado campo.

- Número de productos disponibles en el mercado que solucionan determinada necesidad.
- Número de publicaciones académicas relacionadas con cierta tecnología.
- Número de eventos (ferias, congresos, etc.) relacionados con cierta actividad.

Existen parámetros que se construyen mediante indicadores, que agregan otros datos para dar valores agregados tales como:

- Grado de madurez de una tecnología.
- Posición competitiva del país.
- Grado de aceptación de la población de determinada tecnología.

Existe un tercer tipo de información a captar, desestructurada, constituido por eventos y noticias relacionados con los ámbitos de vigilancia, que aportan a los expertos elementos de juicio adicionales o que son del interés de los usuarios del sistema.

Los factores críticos de vigilancia se suelen concretar en una serie de términos o palabras clave sobre las que se realiza la captación de información.

En esta etapa deben determinarse también los productos de información que se van a extraer del sistema, generalmente, informes de algún tipo. Se deben diseñar los productos atendiendo tanto a su contenido como a su difusión.

Búsqueda activa: captación y monitorización

En esta etapa es preciso identificar y clasificar las fuentes de información que se van a emplear para alimentar el sistema. La clasificación se realiza atendiendo a atributos tales como su fiabilidad, la frecuencia de su actualización, su accesibilidad, su completitud, etc.

Dichos atributos condicionan su captación y posterior procesado, así como la capacidad de mantener la información actualizada y detectar los cambios (monitorización).

Una vez implementado el sistema de Vigilancia Tecnológica, la búsqueda activa es la tarea más común y sostenida en el tiempo, ya que, a menos de que se redefinan los objetivos o se designen nuevos factores críticos de vigilancia o palabras clave, la actividad base es la captación de datos.

Clasificación, gestión de la información

Una vez establecidos los mecanismos para la captación de información desde las distintas fuentes, se hace necesario tratar la información obtenida. El primer paso es clasificarla, relacionándola con los distintos aspectos que se han seleccionado como más relevantes y con los factores críticos de vigilancia de manera que quede correctamente identificada.

En el caso de fuentes estructuradas, obtendremos una serie de conjuntos de datos que se van alimentando con cada iteración del sistema, por ejemplo el número de patentes en un determinado campo.

En el caso de fuentes no estructuradas, se clasifican por familias, que deberán ser revisadas por el *curador de contenidos* de manera individual (el ejemplo claro son las noticias tecnológicas: no son almacenes de datos, se clasifican por temas y deben ser leídas por el curador para asignarles un valor, basado en su experiencia y conocimiento previo; otro podrían ser los objetos de las patentes, que ofrecen conocimiento sobre lo que la técnica va haciendo posible y que serán más o menos relevantes en base al conocimiento del curador).

En este punto es interesante reseñar que un recurso muy enriquecedor si se consigue incorporar al sistema, aunque no siempre esté disponible, es la figura del experto. Así, tener identificado un listado de expertos en los diferentes ámbitos de interés y mantener abierto con ellos un canal de comunicación puede enriquecer enormemente la calidad de los productos de vigilancia obtenidos gracias al sistema, pues su criterio está altamente cualificado y siempre aporta valor a la información. En esta etapa del proceso, su función sería valorar la relevancia de una información, en la etapa de procesado, ayudaría a darle sentido.

Procesado de la información

En esta etapa se trabaja sobre la información captada. Corresponde a la cúspide del conocimiento en la figura de la pirámide que hemos empleado como imagen del proceso del conocer. En esta etapa se pone en relación la información con el conocimiento previo del contexto, y con información de otras categorías, se analizan los datos, se buscan señales débiles, cambios de tendencia... En esta fase es cuando la inclusión de una red de expertos en el sistema es capaz de aportar más valor añadido. El resultado de esta etapa se concreta en los productos de información previamente definidos en la fase de planificación, que deben integrar la información relevante identificada y el valor añadido que puedan aportar los agentes humanos involucrados en el sistema.

Difusión

Una vez se han generado los productos de información, se deben poner a disposición de los consumidores finales. Dentro de una organización, existe un trabajo previo (que debería ser un input en la etapa de planificación), en el que se determina qué tiempo de información precisan los diferentes consumidores, de manera que los productos se adapten a sus necesidades concretas. En esta etapa de difusión, se ponen a disposición de esas personas o perfiles de la organización los productos específicos que les corresponden. Puede estar a cargo de un agente humano, responsable de esta distribución, o automatizarse.

Retroalimentación

La información obtenida en cada iteración se incorpora al sistema y lo enriquece, y acrecienta el acervo de conocimiento de los curadores, de manera que lo que es información en la iteración n, se utiliza en la etapa de filtrado y se incorpora a los criterios de selección de información en la iteración (n+1).

1.3.4. Elementos del sistema

Al explicar el procedimiento, se han puesto en evidencia los elementos fundamentales del sistema de vigilancia. Indicaremos algunas de las herramientas que se emplean habitualmente, pero sin profundizar en ellas.

A modo de resumen, presentamos este listado introductorio:

- **Recursos de información (fuentes):** Bases de datos especializadas.
- **Herramientas (TIC):** Buscadores, Spiders, Indexadores, Alertas, Buscadores especializados, Metabuscadores, Marketplaces, Software específico de vigilancia tecnológica, Open analytics.
- **Agentes humanos:** Curador de contenidos, Experto, Consumidor de los productos o usuario del sistema, Responsable de la difusión de los productos de información.

2. Vigilancia, Inteligencia, Conocimiento y Prospectiva

Existen un conjunto de conceptos relacionados con la **Vigilancia** y la **Inteligencia,** los protagonistas de este libro, que se nos presentan siempre acompañándolos, como apellidos de un nombre. La Vigilancia se nos presenta como *vigilancia estratégica,*

tecnológica, comercial, competitiva, jurídica o financiera, entre otros. La Inteligencia la "apellidamos" con los adjetivos de *"competitiva", "de negocios", económica o corporativa*. Para tener una foto razonablemente completa debemos acompañar a la Vigilancia y la Inteligencia con dos conceptos adicionales: la **gestión del conocimiento** y la **prospectiva**, también conocida como futurología.

Las diferencias entre unos y otros conceptos no siempre son claras. A veces incluso aparecen razones de preferencia en uno u otro idioma utilizando con un sentido general términos que son más restringidos. Por ejemplo en el entorno francófono se usa mucho el de Inteligencia Económica, aunque en puridad la Inteligencia económica tiene en cuenta principalmente dimensiones económicas. Lo mismo ocurre en inglés, con el concepto de *Business Intelligence* es decir Inteligencia de Negocio.

Muy posiblemente nos estemos enfrentando a ese hecho tan humano de buscar una semántica, un significado, una manera unívoca de referirnos a conceptos de tal manera que se diferencie suficientemente de otros similares y que capte todas los fundamentos e incluso los matices de la actividad per-se. Empezaremos por los "nombres", seguiremos a continuación con "los apellidos", intentando componer un marco en el que quepan todos los conceptos y sus variaciones.

En general se acepta que la **Vigilancia** es un concepto más "pasivo" que el de Inteligencia, mediante la que se pretende obtener la información más relevante para nuestro entorno

e intereses y suele incluir el análisis de dicha información. Frente a esto, la **Inteligencia** trasciende las actividades que realiza la vigilancia destacando la importancia en la presentación de la información en tiempo y forma adecuada para que los directivos puedan realizar una toma de decisiones correcta, ganándose así el atributo de ser más "activa". Además destaca la necesidad de medir el efecto de la implantación de un sistema de Inteligencia. Es adecuado matizar, de todos modos, que la Vigilancia mantiene un proceso de revisión y mejora continua de los elementos del sistema, por ejemplo la vigencia de las Fuentes utilizadas, manteniendo así los objetivos de vigilancia del sistema, aunque no pone su foco en la entrega activa de información oportuna a los directivos.

Nos encontraremos con **varios tipos** de **Inteligencia**. La primera, la **Competitiva**, interpreta prácticas y movimientos estratégicos o tácticos de los competidores que afecten a la posición competitiva de la empresa. No sólo de competir vive la Inteligencia: otra razón para implementar un sistema de Inteligencia puede ser el seguimiento de acuerdos o prácticas establecidas en el sector, un *statu quo* explícito o implícito. A este tipo de Inteligencia la llamamos **Cooperativa.** Otras razones, cubiertas por lo que llamamos **Inteligencia Neutral**, pueden ser consolidar actividades de la empresa, realizar investigaciones de marketing, realizar seguimiento de escenarios futuros sobre un sector o confeccionar informes de amplio alcance como los que realizan asociaciones sectoriales, think-tanks o centros de investigación. Por último, muy impulsada por la información existente en redes sociales profesionales, de manera reciente ha surgido la **Inteligencia Individual,** orientada a interpretar el entorno y las características de la organización en la que trabaja la persona e integrar ese conocimiento con su carrera profesional e intereses.

El apellido fundamental a explorar es el de "**Estratégica**". Podemos definir la estrategia empresarial como el conjunto de actividades de la empresa puestas en marcha con el objetivo de asegurar la sostenibilidad de la empresa a largo plazo. Existen numerosas escuelas estratégicas[15] que definen su propio enfoque y sus propias herramientas. Por ejemplo la Escuela de Diseño popularizó el análisis **DAFO**, por otra parte la Escuela estratégica del Posicionamiento y Michael Porter[16] popularizaron los análisis de **Cadena de Valor** y **5 Fuerzas**. Actualmente las diferentes versiones del **Business Canvas Model**[17] son imprescindibles en cualquier análisis estratégico moderno.

Los componentes que aparecen en cada uno de estos análisis nos dan pistas sobre los sentidos de diversos apellidos que presentábamos anteriormente. El análisis **DAFO** se compone de un **Análisis Externo**, en el que se incluyen las **Amenazas** y las **Oportunidades** para la organización y un **Análisis Interno**, que incluye las **Fortalezas** y **Debilidades** de la organización. En la **Cadena de Valor**, se incluyen un conjunto de

[15] Henry Mintzberg, Bruce Ahlstrand, Joseph Lampel (1999). "Safari a la Estrategia"

[16] Michel Porter (1980), "Estrategia Competitiva"

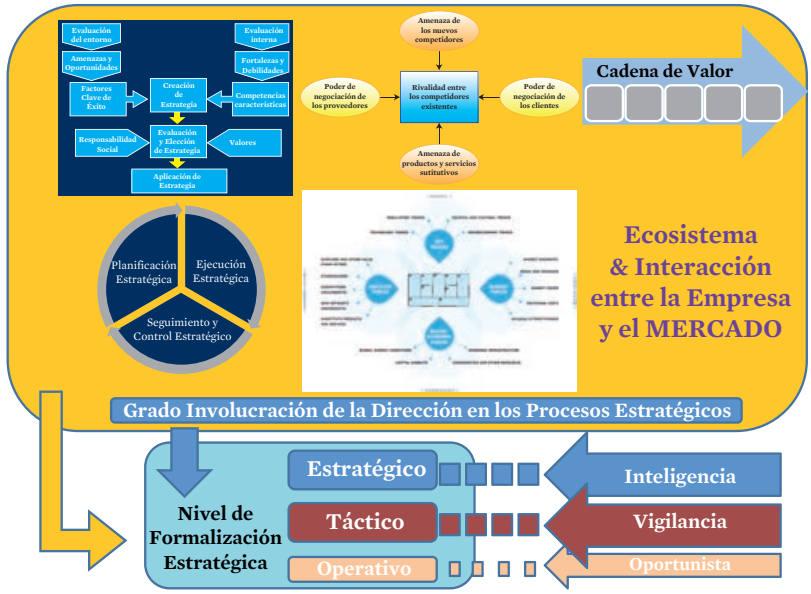
[17] Alexander Osterwalder, Yves Pigneur. (2010). Business Model Generation

actividades principales, entre las que se incluyen las actividades comerciales, de operaciones o logísticas y **actividades de soporte**, que incluyen los sistemas de información, los recursos humanos, las adquisiciones y la financiera. En el análisis de **5 fuerzas** se destaca la Competencia en el Mercado en el que se establecen equilibrios entre los **Proveedores**, los **Clientes** y **nuevos Competidores** que saltan nuestras **barreras de entrada** al mercado con sus Productos y Servicios y nuestra empresa se enfrenta a sus propias **barreras de salida** cuando aparecen en el mercado **Productos y Servicios Sustitutivos** de los que nuestra empresa presta en un momento dado.

Cada empresa mantiene una diferente relación con el Mercado. Unas tienen un foco 100% en el cliente, se les llama empresas orientadas **"hacia fuera".** Otras en cambio tienen una visión **"hacia dentro",** diseñando productos y servicios a partir de su conocimiento, confiando en que el mercado los aceptará.

En las empresas más maduras estratégicamente hablando existen **procesos explícitos de planificación, ejecución y control estratégico**. En el otro extremo, cubierto también por otras escuelas estratégicas, tenemos empresas en las que la estrategia está fundamentalmente "en la cabeza del líder" y por tanto la visión está poco formalizada.

Con estos ingredientes se nos dibujan ya varias de las situaciones que nos dan lugar a los conceptos que presentamos en este apartado. En una empresa con un liderazgo muy personalista es probable que no expliciten sus procesos de gestión estratégica y consecuentemente como mucho sólo llegue a hacer Vigilancia. Por otra parte tendremos empresas maduras estratégicamente con procesos estratégicos explicitados en los que tendrá todo el sentido el disponer de procesos de Inteligencia.



Nos encontraremos con situaciones o aplicaciones orientadas a eslabones de la cadena de valor o en ámbitos concretos, haciendo Vigilancia Comercial, Jurídica, Financiera o Tecnológica y en el otro extremo con organizaciones maduras que expliciten procesos de **Inteligencia Competitiva** para estudiar los mercados, conocer el entorno, analizar la información disponible, agregar valor y tomar decisiones coherentes con el conocimiento adquirido para que la empresa compita de manera sostenible en los mercados.

Frente al foco de la Vigilancia y la Inteligencia en el exterior, la **Gestión de Conocimiento** se enfoca más al Interior, a los resultados del Análisis Interno del DAFO. Parte de los Conocimientos existentes en la empresa, tanto los explícitos de cualquier organización como en mayor medida los conocimientos implícitos que se destilan del conocimiento de los miembros de la empresa y que unas veces permean la empresa y otras son repositorios de valor sin explotar de gran importancia.

Las empresas orientadas "hacia dentro" serán más proclives a disponer de procesos y aplicaciones de Gestión de Conocimiento, frente a las empresas orientadas "hacia fuera" que serán mejores candidatos para poner en marcha procesos de Inteligencia.

El último concepto especialmente relevante es la **Prospectiva,** que en inglés denominan de forma muy ilustrativa: "*future studies*"[18], estudios acerca del **futuro**. Dos son las definiciones más aceptadas:

> ***Ejercicio colectivo*** *de análisis y comunicación entre expertos para identificar las componentes probables de escenarios de futuro: las proyecciones tecnológicas de los mismos, sus* ***efectos sociales y económicos, obstáculos y fuerzas a favor".***

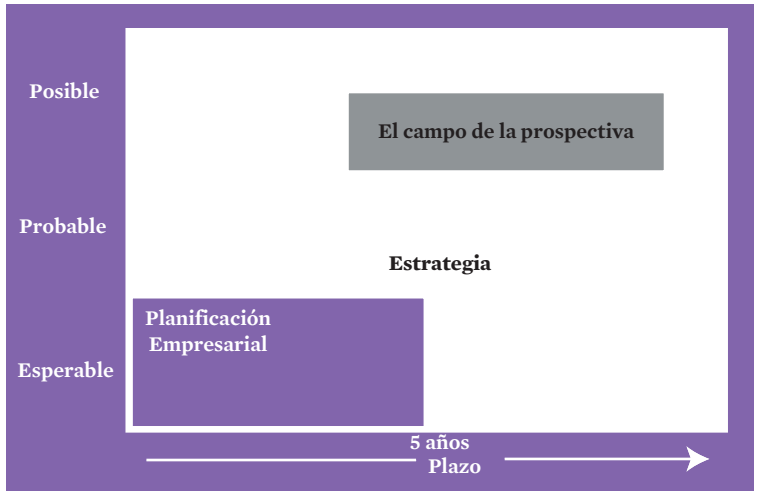
Y también es muy utilizada la de la OCDE:

> *"Tentativas sistemáticas para observar a largo plazo el futuro de la ciencia, la tecnología, la economía y la sociedad con el propósito de identificar las tecnologías emergentes que probablemente produzcan los mayores beneficios económicos y sociales"*

El sentido de la prospectiva es por tanto el estudio del futuro a medio y sobre todo a largo plazo. Frente a esto tanto la vigilancia como la inteligencia y la gestión del conocimiento hacen foco en información del pasado, histórica por tanto, y del presente cercano.

---

[18] Wikipedia. "Future Studies". http://en.wikipedia.org/wiki/Futures_studies.



En la prospectiva se tienen en cuenta tanto escenarios razonablemente continuistas sobre la realidad actual como escenarios digamos revolucionarios en los que se producen cambios mayores sobre cuestiones que ahora consideramos totalmente axiomáticas.

La Prospectiva permite sentar las bases para planificar acciones que influyan en el futuro, evitando los escenarios más negativos y promoviendo los más positivos.

3. La norma UNE 166.006:2011 "Gestión I+D+i: Sistema de Vigilancia Tecnológica e Inteligencia Competitiva"

Los comités de normalización

El comité AEN/CTN 166 de la Asociación Española para la Normalización, AENOR, es uno de los líderes actualmente a nivel mundial en la normalización de la Gestión del I+D+i. Buen ejemplo de ello es la norma UNE 166.006:2011 sobre "Gestión de I+D+i: Sistema de Vigilancia Tecnológica e Inteligencia Competitiva", traducida al inglés como *"Technological Watch and Competitive Intelligence System"*, que presentaremos posteriormente en este apartado. Esta versión 2011 sustituye a la versión 2006 y destaca por la incorporación a la misma del concepto de "Inteligencia Competitiva", ampliando así su ámbito de actuación a la estrategia competitiva.

Su campo de actividad está en la normalización de los aspectos de organización y definición de las actividades de I+D+i en las empresas industriales, incluyendo la definición

y terminologías de las actividades de I+D+i, los requisitos directrices y recomendaciones de los sistemas de gestión y proyectos de I+D+i., las guías de auditoría de los sistemas de gestión de la I+D+i y de los proyectos de I+D+i y la transferencia de tecnología.

Actualmente[19] mantiene como vigentes los siguientes documentos estándares sobre Gestión de I+D+i:

Código	Título	Fecha
UNE 166000:2006	Gestión de la I+D+i: Terminología y definiciones de las actividades de I+D+i.	03/05/2006
UNE 166001:2006	Gestión de la I+D+i: Requisitos de un proyecto de I+D+i.	03/05/2006
UNE 166002:2014	Gestión de la I+D+i: Requisitos del Sistema de Gestión de la I+D+i.	21/05/2014
UNE 166005:2012 IN	Gestión de la I+D+i: Guía de aplicación de la Norma UNE 166002 al sector de bienes de equipo.	25/07/2012
UNE 166006:2011	Gestión de la I+D+i: Sistema de vigilancia tecnológica e inteligencia competitiva.	16/03/2011
UNE 166007:2010 IN	Gestión de la I+D+i: Guía de aplicación de la Norma UNE 166002:2006.	19/05/2010
UNE 166008:2012	Gestión de la I+D+i: Transferencia de tecnología.	25/07/2012
UNE-CEN/TS 16555-1:2013 EX	Gestión de la innovación. Parte 1: Sistema de gestión de la innovación.	10/07/2013

Asimismo es responsable de estos otros dos documentos relacionados, del ámbito de la pyme:

Código	Título	Fecha
EA 0043:2015	Requisitos para la consideración como Joven Empresa Innovadora.	04/02/2015
EA 0047:2015	Requisitos para la consideración como Pequeña o Mediana Empresa Innovadora.	04/02/2015

La norma UNE 166.006:2011, ha sido la referencia para los estándares publicados por el **Comité Técnico 389 (CEN/TC 389)[20] del Comité Europeo para Estandarización (CEN),** dedicado a la Gestión de la Innovación. Los estándares publicados por este organismo son **6 partes de la Gestión de la Innovación**:

[19] http://www.aenor.es/aenor/normas/ctn/fichactn.asp?codigonorm=AEN/CTN%20166&pagina=1

[20] Comité Europeo para la Normalización - Comité Técnico CEN/TC 389 - Gestión de la Innovación: http://standards.cen.eu/dyn/www/f?p=204:7:0::::FSP_ORG_ID:671850&cs=1E977FFA493E636619BDED775DB4E2A76

- Sistema de Gestión de la Innovación: la referencia para esta parte es el documento UNE-CEN/TS 16555-1:2013 EX "Gestión de la innovación. Parte 1: Sistema de gestión de la innovación", del 10 de Julio del pasado año 2013.
- Gestión de la Inteligencia Estratégica.
- Pensamiento Innovador (en inglés "Innovation Thinking").
- Gestión de la Propiedad Intelectual.
- Gestión de la Colaboración.
- Gestión de la Creatividad.

Está pendiente de aprobación una séptima parte dedicada a la "Valoración de la Gestión de la Innovación".

El CEN/TC 389 tiene como alcance de su trabajo la estandarización de herramientas que permitan que las organizaciones, instituciones y empresas mejoren su gestión de la innovación, incluyendo todo tipo de aspecto relacionado con la innovación y las actividades de Investigación y Desarrollo.

A nivel global, la actividad de estandarización de la Gestión de la Innovación es desarrollada por **el Comité Técnico "ISO/TC 279 Innovation Management"[21]. Actualmente no existen estándares ISO en esta materia.** El Business Plan de ISO para Innovation Management[22] señala como referencias a los estándares europeos CEN que hemos mencionado así como a los estándares nacionales de diversos países.

La Norma española UNE 166.006:2011

Posiblemente la novedad más relevante de esta versión 2011 de la norma sea la inclusión de la **Inteligencia Competitiva**. La Inteligencia Competitiva añade a la Vigilancia Tecnológica dos aspectos fundamentales:

- La comunicación en tiempo y forma adecuada de la información de vigilancia así como su análisis a la dirección de la organización y su integración en procesos de gestión de la toma de decisiones.

---

[21] ISO Comité Técnico para la Gestión de la Innovación: http://www.iso.org/iso/iso_technical_committee%3Fcommid%3D4587737

[22] ISO (Diciembre 2014) "Strategic business plan – Innovation Management" Comité ISO/TC 279: http://isotc.iso.org/livelink/livelink/fetch/2000/2122/687806/ISO_TC_279__Innovation_management_.pdf?nodeid=16913333&vernum=-2

- El foco en los aspectos de análisis competitivo de la organización en el mercado, entre los que se pueden encontrar los clientes, los proveedores, los competidores, las barreras de entrada y salida al mercado, los productos sustitutivos, el ecosistema de stakeholders del mercado, etc.

Otra cuestión relevante es que la **Vigilancia Tecnológica** hace foco en la Tecnología. Otros enfoques de Vigilancia, como la Vigilancia Comercial o Jurídica hacen foco en otros aspectos como los de comercial, marketing o legislativo, que también pueden influenciar indirectamente en la Tecnología, cuestión que deberá tenerse en cuenta a la hora de definir el Sistema.

La norma **UNE 166.006:2011** queda encuadrada en dos normas de alcance más amplio: la **UNE 166.002** para la **Gestión del I+D+i**, hará referencia a la norma **UNE 166.000:2006** que recoge **Terminología y Definiciones**, y la norma **ISO 9000** para la **Gestión General de la Organización**. Consecuentemente presentará una estrategia de Mejora Continua similar al ciclo de Deming con las 4 fases bien conocidas: Planificar (Plan), Hacer (Do), Verificar (Check) y Actuar (Act) que guiará **el incremento de la efectividad del sistema**.

La norma ayudará a la implantación y puesta en marcha de procesos de vigilancia tecnológica e inteligencia competitiva adecuados para los objetivos de la empresa o institución así como la organización que la gobierne.

La Inteligencia Competitiva requiere de la participación, compromiso y liderazgo por parte de la Dirección de la Organización en la que se diseña, desarrolla, implanta y mantiene el Sistema. Es por ello que la norma incluye un apartado específico sobre las **Responsabilidades de la Dirección**. La Dirección deberá implicarse activamente en el establecimiento de la Política y Objetivos de Vigilancia Tecnológica y la Inteligencia Competitiva, la Planificación necesaria para el cumplimiento de los Requisitos identificados y la Revisión y Mejora del Sistema.

Un punto clave en la puesta en marcha y operación de un Sistema de Vigilancia o de Inteligencia son las personas, por lo que la norma incluye un apartado sobre **Recursos Humanos**. La formación del personal, las competencias necesarias, los recursos materiales e infraestructura e incluso la motivación necesaria se tratan en este apartado.

Para la realización de la Vigilancia Tecnológica e Inteligencia Competitiva (en adelante VT/IC) la norma presenta un conjunto de **procesos**:

- Identificación de necesidades, fuentes y medios de acceso a la información.
- Búsqueda, tratamiento y validación de la información.

- Puesta en Valor de la información.
- Productos de la VT/IC.
- Resultados de la VT/IC.

ILUSTRACIÓN 2

Modelo inspirado en el ciclo de Deming (Plan-Do-Check-Act) y los grupos de procesos marco utilizados en Dirección de Proyectos

Los **Requisitos** deberán especificarse de acuerdo a la parte del Objeto posible del Sistema que se considere viable y se decida poner en marcha.

La norma establece, entre otros, los siguientes requisitos[23]:

- Documentación y registro de los procedimientos, hallazgos y otros aspectos relevantes para la norma.
- Responsabilidad de la Dirección en el proceso de vigilancia tecnológica.
- Disponibilidad de recursos suficientes y adecuados.
- Tomar Acciones en relación a los resultados obtenidos.
- Medición, análisis y mejora del proceso.

De forma coherente con la realidad de las empresas, organismos e instituciones, que muchas carecen del tamaño, volumen de negocio, estructura o capacidad de gestión

[23] Wikipedia "UNE 166006" http://es.wikipedia.org/wiki/UNE_166006

de los riesgos estratégicos, la norma incluye la posibilidad de **Externalización** de los Servicios de Vigilancia e Inteligencia Competitiva a terceros que se encarguen de dicho servicio. Todo se detalla en un apartado titulado *"Contratación de Servicios en los Sistemas de Vigilancia*".

Por último decir que la norma puede adquirirse[24] a través de la web de AENOR a un precio muy asequible, lo cual recomendamos a todos los interesados en estos sistemas.

[24] http://www.aenor.es/aenor/normas/normas/fichanorma.asp?tipo=N&codigo=N0046930&PDF=Si#.VZ-QCvmU05w

3

NUEVAS CAPACIDADES BIG DATA

"Big Data" son dos palabras que comunican muy bien y todo el mundo entiende: "*muchos datos*". Es un concepto pulido por los departamentos de marketing y acordado por la industria con el objeto de ser fácilmente reconocido y aceptado en el mercado. Otra situación similar ocurrió por ejemplo a finales del siglo XX con la palabra "*Portal*" para referirse al sitio web de una empresa.

Sin embargo "Big Data" es mucho más que "muchos datos", incluso **constituye toda una manera de pensar**. Bajo su paraguas encontramos un grupo de tecnologías y áreas de conocimiento; una parte de ellas son nuevas, a otras Big Data les ha dado nueva vida y a otras Big Data les ha dado la oportunidad de salir de laboratorios universitarios y ámbitos restringidos de trabajo y solucionar necesidades de mercado. Todas ellas configuran las bases de nuevas soluciones, totalmente sinérgicas con la Vigilancia Estratégica y la Inteligencia Competitiva.

Resulta imposible ser totalmente exhaustivo con las tecnologías, ideas, procesos y áreas de conocimiento que merecidamente deberían incluirse en un apartado sobre "Big Data". Se pretende presentar en este capítulo 3 una visión horizontal de "Big Data", aceptando que quien mucho abarca poco aprieta y por tanto algunos conceptos o tecnologías apenas se enunciarán y nos remitiremos a otros libros, tratados y páginas web que profundizan sobre los mismos. Se le dedican apartados a aquellos que nos resultan especialmente relevantes para la Vigilancia Estratégica y la Inteligencia Competitiva y que son clave para aportarles nuevas capacidades.

Como colofón, en el último apartado de este punto se mapean el Modelo Big Data que se presenta a continuación con la Cadena de Actividades de la Vigilancia Estratégica y la Vigilancia Competitiva.

1. "V" de Big Data

¿Qué es y qué no es realmente un proyecto "Big Data"? ¿Está mi competencia haciendo proyectos "Big Data"? ¿Debemos emprender proyectos "Big Data"? ¿Estamos en un entorno "Big Data" y por tanto podemos realmente emprender proyectos "Big Data"? Con frecuencia me encuentro incluso con grandes profesionales del sector IT haciendo este tipo de reflexiones. La primera respuesta que doy a estas preguntas proviene de una propuesta: la **"V" de Big Data.**

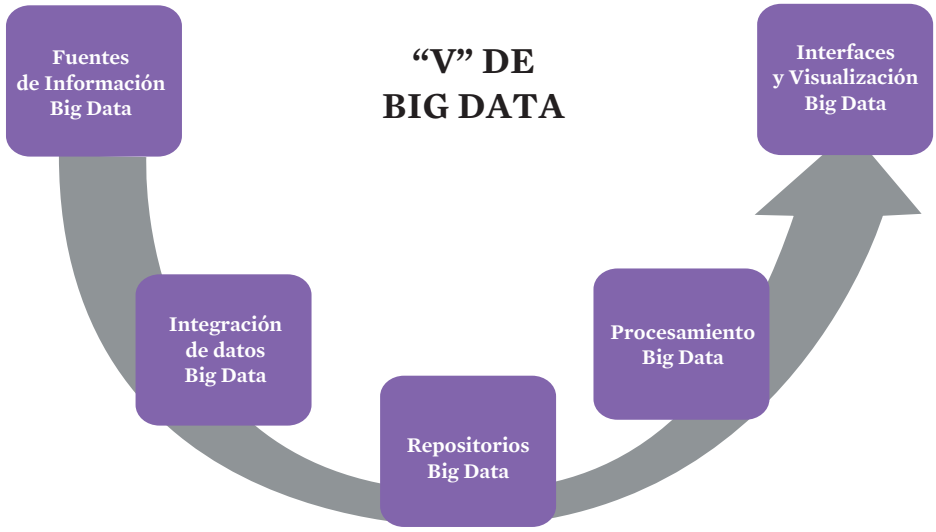
La popularización de Big Data ha venido explicada inicialmente por 3 Vs: el procesamiento de grandes **Volúmenes** de datos que llegan a grandes **Velocidades** y con una **Variedad** de fuentes de información nunca vista hasta ahora. Pensemos por ejemplo

en Google, Facebook o Twitter recogiendo peticiones simultáneas de servicio por usuarios de todo el mundo, procesándolas y generando resultados a las mismas.

La "V" es una letra mágica en informática, que enraíza en sus orígenes en el último cuarto del siglo XX con el "Modelo en V"[25], que se refería a la metodología de desarrollo de nuevas aplicaciones recogiendo las fases incluidas en un proyecto IT, desde la especificación de requisitos, los diferentes análisis, el desarrollo, las pruebas de unidad y de sistema y la puesta en producción.

Seguramente todas estas uves han servido de inconsciente inspiración para que el Modelo que propongo para "Big Data" sea también una "V", **la "V" de Big Data**.

ILUSTRACIÓN 3
Modelo "V" de Big Data, inspirado en el Modelo de Desarrollo en V.
Original del autor



En el modelo en V de Big Data se proponen 5 grupos de procesos:

- **Fuentes de Información Big Data:** enriquecemos nuestras fuentes de datos con nuevas fuentes disponibles de forma abierta en internet. Toda esta Variedad de fuentes de información genera grandes Volúmenes de datos que llegan a gran Velocidad. Las taxonomías que clasifican esas fuentes son relevantes.

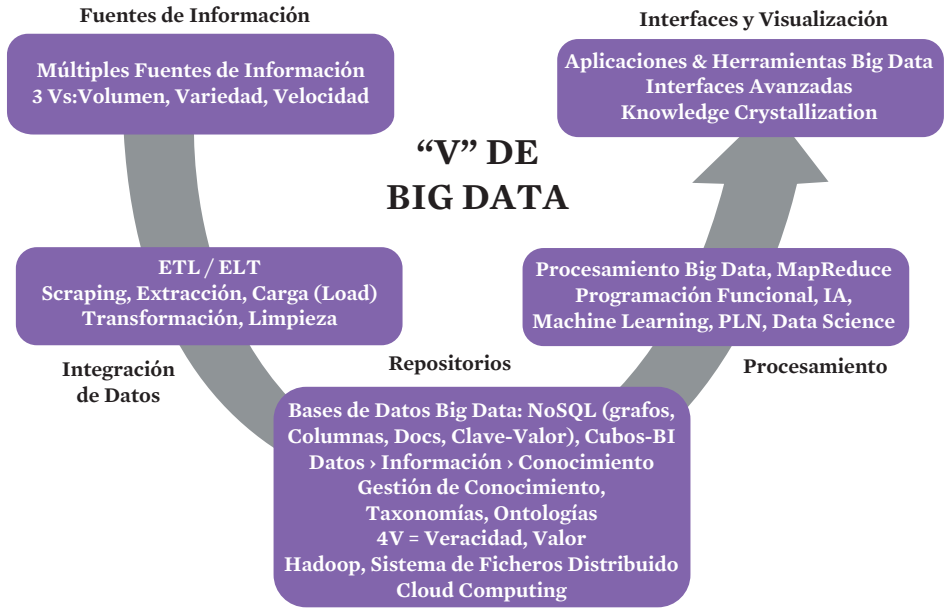
[25] Método /Modelo en V: http://es.wikipedia.org/wiki/Método_en_V

- **Integración de datos Big Data:** extraemos los datos y los cargamos en Repositorios de Información especialmente diseñados para tratar Big Data. Frente a la posibilidad de transformar y limpiar los datos antes de cargarlos la tendencia es cargar todos los datos para poder explotarlos a posteriori para otros fines. Cobra asimismo importancia el proceso de Scraping de información, de lectura de datos directamente de la web mediantes aplicaciones software que llamamos Bots.

- **Sistema y Repositorios Big Data:** nuevos tipos de Bases de Datos, que llamamos NoSQL son los nuevos contenedores de información, especialmente preparados para los tipos de procesamiento necesarios. Además de datos e información gestionamos el conocimiento en Ontologías, que son reflejo de una 4ª V, la Veracidad. El Sistema de Ficheros Distribuido y el Cloud Computing son la base de este Sistema Big Data.

ILUSTRACIÓN 4

Modelo "V" de Big Data ampliado



- **Procesamiento Big Data:** tecnologías tradicionales como la programación funcional, el machine learning, el procesamiento de lenguaje natural, y un grupo de áreas de conocimiento que agrupamos bajo los paraguas de la "Data Science" y la Inteligencia Artificial se aprovechan de nuevas capacidades de procesamiento distribuido y masivo de datos para ser el 4º eslabón de la "V" de Big Data. En torno a esta grupo de procesos aparece para algunas empresas una 5ª "V", la Viscosidad, referenciando con ese concepto la mayor o menor facilidad para correlacionar los datos.

- **Interfaces y Visualización Big Data**: los usuarios necesitan nuevos sistemas de visualización, interacción y análisis para interactuar con el Big Data, diferentes a los tradicionales provenientes del mundo del Business Intelligence. Aparecen situaciones en las que, por ejemplo, una misma pregunta cristaliza en diferentes respuestas para diferentes usuarios según su contexto.

¿Todos estos elementos son necesarios, entonces, para que un proyecto sea Big Data? ¿Qué elementos son totalmente necesarios y cuáles en cambio son coyunturales o innecesarios? ¿Si tenemos Volumen y Velocidad de datos pero no Variedad estamos en un contexto Big Data? ¿Es sensato plantearse hacer una taxonomía que nos ayude en esta clasificación? ¿Los proyectos que anteriormente decíamos que eran de Business Intelligence o de Análisis Estadístico son ahora de Big Data?

La necesidad de responder a estas preguntas viene de nuestro propio carácter como seres humanos. De nuestro yo más interior surge la necesidad de clasificar, de saber lo que es comestible y lo que no, lo que es un peligro o un aliado, cómo gestionar cada situación, persona o cosa. "Big Data" es, sin embargo, un concepto artificial, un término de marketing, un paraguas bajo el que se recoge una nueva realidad que todavía está definiéndose y evolucionando.

La respuesta, más que en "el todo", debemos buscarla "en las partes". Nos encontraremos con pocos proyectos "Big Data puros", con todos sus elementos mencionados en el modelo propuesto.

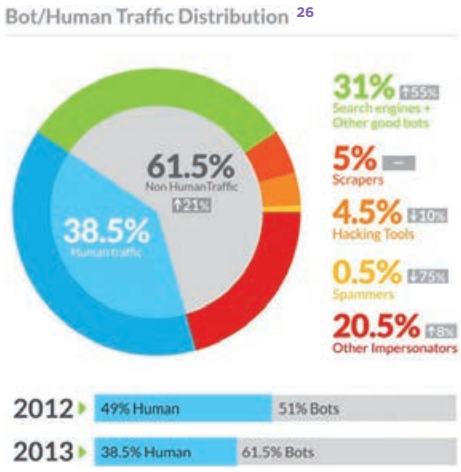
Nos encontraremos más proyectos que usen varios de estos elementos y poco a poco nos encontraremos más que utilicen algún elemento o alguna tecnología relacionada. Sin duda también aparecerán elementos que deberán ser considerados, con buen criterio, parte de proyectos Big Data.

La otra respuesta la podemos encontrar en "la forma de pensar Big Data": múltiples fuentes de información que enriquecen nuestro conjunto de datos, Scraping de páginas web para extraer información no preparada para su tratamiento, integración de la información en repositorios especializados para el tipo de información y conocimiento que necesitamos, herramientas y técnicas especializadas para el tratamiento de la información y la generación de soluciones y finalmente la aplicación de técnicas de entrega, visualización y análisis avanzado para la presentación de la información.

Finalmente lo más importante será, sin duda, que el proyecto proporcione nuevas soluciones a casos de uso que hasta ahora no nos planteábamos solucionar por considerar que era imposible abordarlos.

2. Business Bots, Spiders, Scrapers: recuperando información desestructurada de la WEB

En los proyectos Big Data es necesario habitualmente recopilar datos de diversas fuentes, bien por ser parte intrínseca y necesaria del proyecto, bien con el objeto de enriquecer dichos datos y obtener consecuentemente soluciones a los casos de uso y necesidades de negocio de más calidad.



Para ello utilizamos el **Scraping,** un conjunto de técnicas que tienen como objetivo la extracción de información bien de páginas web normalmente simulando la navegación que las personas hacemos a través de un navegador, bien de otras fuentes, habitualmente colecciones de documentos. Ni la documentación ni las páginas web están pensadas para ser leídas a través de una aplicación software sino para ser vistas por personas a través de aplicaciones que facilitan su lectura página a página. Concretamente las páginas web son vistas a través de las aplicaciones software que llamamos "Navegadores".

Para ello se desarrollan aplicaciones específicas, que llamamos Webbots o sencillamente Bots, que automatizan la interacción con el sitio web en cuya información estamos interesados. Los Bots realizan diversas funciones, destacando la función de navegación por la página web y la de lectura de los contenidos. A la primera labor le llamamos habitualmente Crawling, y a los bots que realizan esa función Crawlers, Spiders o WebSpiders. A la segunda le llamamos Scraping y a los bots Scrapers o WebScrapers. A estos Bots se les incorpora adicionalmente todo tipo de funciones, por ejemplo de automatización de tareas o de integración con otras aplicaciones y sistemas.

El Bot más conocido es el **GoogleBot**, que recorre la World Wide Web, recogiendo información en su base de datos para su motor de búsqueda, aunque son miles, posiblemente millones los Bots activos en internet. De hecho, se calcula que el tráfico en internet atribuible a Bots supera ya el 60%. Concretamente el tráfico atribuible a WebScrapers estaría en torno al 5% del total del tráfico de internet.

---

[26] Bot Traffic Report 2013, reparto de tráfico en internet entre tráfico de Bots y navegación de personas https://www.incapsula.com/blog/bot-traffic-report-2013.html

Oportunidades de Negocio

En la actualidad hay una fuerte demanda de desarrollo de este tipo de Bots[27], por las enormes oportunidades que hay como resultado de incorporar capacidades de toma de decisiones, integración y automatización a sus páginas web corporativas. La gestión de la experiencia de usuario, la gestión de cambios, la inteligencia competitiva o la integración de reglas de negocio son áreas de negocio en las que los Bots se están aplicando. Este tipo de aplicaciones constituyen un cambio sustancial en nuestra manera de interactuar con Internet, contribuyendo en que a medio y largo plazo se transforme completamente.

Se presentan a continuación algunos ejemplos de Bots aplicados al negocio:

- **Análisis de precios y compras automáticas**. Se aplican en compras y pujas por eBay y otros *marketplaces*. Este tipo de Bots se denominan *Snipers*. Incorporan reglas de negocio que manejan situaciones y excepciones. Este tipo de Bots se han llevado al límite fuera de la web, en el mercado financiero, en el que se hacen compras automáticas a altísima velocidad, lo que denominamos **HFT** "***high frecuency trading"***[28].
- **Búsqueda de ficheros**, aplicado al control de pirateo de contenidos.
- **Verificación de calidad de contenidos**, por ejemplo URLs mal construidas, mejora de la calidad de los contenidos, cálculo de rankings.
- **Agregadores de contenidos**, por ejemplo son muy conocidos los agregadores RSS. Empiezan a popularizarse otros, como los agregadores de ofertas de empleo.

La utilización masiva de los Bots debe considerarse una tecnología todavía emergente, tanto por su difusión limitada como por la falta de madurez de la Word Wide Web. La mentalidad de interaccionar con los servidores web únicamente mediante navegadores de forma individual por las personas es algo que se mantendrá todavía durante muchos años.

---

[27] "Webbots, Spiders and Screen Scrapers: A guide to Developing Internet Agents with PHP/CURL (2nd Edition), Michael Schrenk, No Starch Press, 2012

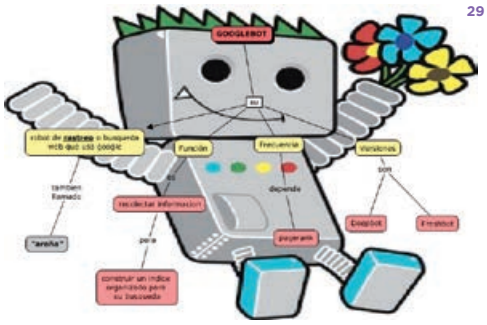
[28] Jacob Loveless, Sasha Stoikov, Rolf Waeber - Communications of the ACM Vol. 56 No. 10, Pages 50-56 – "Online Algorithms in High-Frequency Trading: http://cacm.acm.org/magazines/2013/10/168184-online-algorithms-in-high-frequency-trading/abstract

Con frecuencia el uso de Bots está también asociado a situaciones de negocio de **Investigación e Inteligencia Competitiva**, lícita, pero que se desea no hacer pública para precisamente evitar darle directa o indirectamente pistas a la competencia.

A medida que se profundiza en el tránsito hacia una vida digital este tipo de agentes inteligentes cobran más protagonismo. Cada vez es más necesario hacer las webs accesibles a Bots y aplicaciones en general.

Funcionamiento de los Bots, Spiders, Scrapers…

La primera actividad que suelen realizar los Bots es la **Descarga de Páginas**. Esta es una labor realizada por un tipo de Bots que llamamos **Arañas**, (o también en inglés, "**Spiders")**, Crawlers, Web Crawlers o Web Walkers. Las Arañas, descargan páginas web se gún los objetivos marcados en la aplicación. Una vez descargada una página buscan los enlaces contenidos dentro de ella y siguen dichos enlaces para descargar la página enlazada. Como esta podría ser una labor infinita, se establece un límite de profundidad que llamamos nivel de penetración.



[29]

La tendencia actual está en el almacenamiento masivo de los datos para utilizaciones de los mismos, comprimiéndolos en lo posible. Tradicionalmente se ha hecho en bases de datos relacionales, actualmente se está pasando a utilizar bases de datos tipo Big Data, al que le dedicamos un apartado en este libro. Una de las razones para realizar almacenamiento masivo es el poder realizar estudios históricos, proyecciones a largo plazo basadas en la información histórica y finalmente el no perder información que pudiera ser utilizada en el futuro aplicando técnicas todavía no inventadas o solucionando necesidades de negocio todavía no expresadas. Es importante realizar también el almacenamiento de meta-datos que permitan integrar los datos con los objetivos de negocio implementados en la aplicación.

Tras la descarga de las páginas viene la etapa de **Análisis Sintáctico** (en inglés, "**parsing"**), cuyo objetivo es separar de los textos lo que es útil, lo que está orientado a los objetivos de la aplicación, de lo que no, construyendo para ello una **estructura de datos ad-hoc para cada página**. Es frecuente, de todos modos, que el análisis sintáctico

[29] Imagen de Google Bot incluyendo un mapa conceptual mediante la herramienta IHMC Cmap tools: http://cmapspublic.ihmc.us/rid=1K03VVV5X-1R1G2XN-1G1J/googlebot.ryna.cmap

se realice durante la descarga para reducir la cantidad de información almacenada, pero perdiendo la capacidad de volver a analizar los datos y de aplicar técnicas de proyección y predicción. La técnica del parsing data de los orígenes de la informática. Se usa por ejemplo en los compiladores, los programas que convierten un texto escrito en un lenguaje de programación en un programa ejecutable en un ordenador.

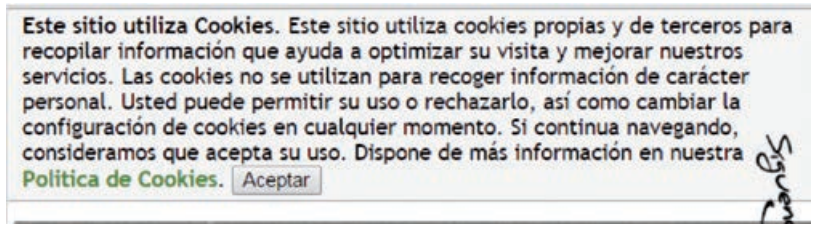
Por ejemplo el **GoogleBot** busca **imágenes** que mostrar en Google Images, **ficheros** que mostrar cuando usamos la palabra clave "filetype:" en una búsqueda o **enlaces y contenidos** con los que alimentar al algoritmo que decide qué resultados mostrar cuando hacemos una búsqueda sobre unas palabras clave concretas.

Una dificultad adicional del análisis sintáctico es la calidad del texto de la página web. Pueden ocurrir diferentes circunstancias, desde código HTML de baja calidad como mezclas de diferentes contenidos, por ejemplo publicidad, que dificultan el proceso de averiguar en qué consiste el texto que está siendo leído, cuál es su sentido, qué es lo que se quiere comunicar. Para solucionar o al menos paliar esta situación se aplican funcionalidades de limpieza de textos.

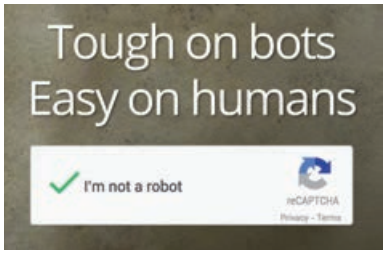
Lo que no es sencillo de analizar es el lenguaje humano, que técnicamente llamamos Lenguaje Natural, por la complejidad y ambigüedad del mismo. Sin embargo es parte habitual de los proyectos Big Data, por lo que también le dedicamos un apartado al Procesamiento de Lenguaje Natural en este libro. Frecuentemente nos vamos a encontrar con que no nos va a ser posible cumplir los objetivos de negocio especificados debido a la incapacidad tecnológica y científica actual de analizar el lenguaje natural. Sin embargo Internet y la WWW han hecho posible que estas tecnologías empiecen a despegar y nos estén proporcionando aplicaciones de alto valor añadido que hasta ahora eran impensables. Esta situación la tratamos en los apartados dedicados al "Procesamiento de Lenguaje Natural" en los capítulos 3 y 4.

La siguiente situación más común que tienen que gestionar los Bots son los **Formularios**, destacando en particular el formulario de **Autenticación Básica,** es decir, cuando se accede a una página web a través de usuario y password. Entender el formulario y completarlo emulando como lo entiende un usuario no es trivial. Suponiendo que el formulario sea entendido y completado, entregarlo al Servidor Web de forma correcta y completa es un proceso muy proclive a errores.

Otra situación habitual con la que se tienen que enfrentar los Bots es el establecimiento de sesiones con el servidor web. Cuando accedemos a una página web, el Servidor Web proporciona un identificador (*session value*) con el objetivo de otorgar una identificación a la persona que navega y proporcionarle diferentes características que tengan en cuenta esa navegación. A este proceso lo denominamos **Autenticación de Sesión.**



La autenticación de sesión más habitual es la autenticación por **Cookies**. Las cookies son ficheros que, provenientes del Servidor Web se guardan en nuestro ordenador. Ayudan a que el servidor recuerde preferencias y hábitos de navegación de los usuarios y para que los identifique manteniendo la autenticación de la sesión. Por ejemplo guardamos en una cookie el carrito de la compra con los productos que vamos seleccionando al hacer una compra online. Un navegador no puede acceder a los ficheros de nuestro ordenador, salvo que explícita y voluntariamente lo hagamos, por ejemplo al cargar un fichero en nuestro webmail. Las cookies son una excepción necesaria.



Cada vez que se interacciona con el Servidor Web se le ha de enviar la cookie. Hay dos tipos de cookies: *Temporales*, que desaparecen al cerrar el navegador y *Permanentes*, que persisten en el disco duro hasta que llega su fecha de expiración, que es un valor indicado por el servidor web. El servidor modifica los valores incluidos en la cookie, no pudiendo hacerlo el navegador nunca. Sin embargo los Bots no tienen esa limitación, pudiendo hacerlo a voluntad.

A la hora de programar el Bot es importante tener en cuenta que las cookies pueden afectar a los formularios ya que contienen variables de sesión.

Un formulario muy particular, que se completa de forma previa a la autenticación, es el de los ***códigos Captcha*** *(Computer Automated Public Turing test to tell Computers and Humans Apart),* en el que se inserta un texto dentro de una imagen con el objetivo de solicitar explícitamente que sean sólo las personas y no los Bots los que accedan a una determinada página web.

El otro sistema de autenticación de sesión más utilizado es usar la propia URL, usando cadenas de consulta (query string). Las URLs deberían cumplir la arquitectura REST (Representational State Transfer), caracterizada porque cada petición HTTP contiene toda la información necesaria para comprender la petición, incluida la *sessión value.*

El proceso de Scraping puede ejecutarse mediante varios hilos de ejecución en paralelo, lo cual resulta ideal para el **procesamiento en sistemas distribuidos** y por tanto en sistemas **Big Data**. En ocasiones esto resulta totalmente necesario: si un servicio web detecta que es repetida, continua y sistemáticamente visitado desde una misma máquina puede interpretar que se trata del ataque de un hacker y rechazar las visitas desde la dirección IP de la máquina que realiza el Scraping. Un Sistema Distribuido facilita la realización de peticiones desde diferentes direcciones IP, evitando así esta circunstancia. Existen, de todos modos, otras técnicas que no necesitan el sistema distribuido como la de usar máquinas proxies, que consigue que las conexiones se realicen desde diferentes ubicaciones que nos convengan.

En caso de ejecutar la aplicación en un sistema distribuido necesitaremos también un planificador que decida a qué dominios y subdominios y cada cuánto tiempo hacer peticiones.

El Servidor Web

Cuando accedemos mediante un navegador a un Sitio Web, estamos interaccionando con un **Servidor Web**, la aplicación encargada de gestionar, confeccionar y servir las páginas web que se le demandan principalmente a través del protocolo HTTP. **Los Bots tienen que interactuar con el Servidor Web y enfrentarse a la misma problemática que los navegadores**. La página web que le sirve al navegador constituye un entorno atractivo y bien conformado con el que las personas interactuamos, pero no está pensado para que interactúe un Bot. En este apartado veremos los aspectos más relevantes que deberá solucionar un Bot en su interacción con el Servidor Web para poder cumplir sus funciones.

Las páginas web que nos presenta el navegador están escritas en **HTML**, el lenguaje de la World Wide Web. Para confeccionar las páginas web que el Servidor Web le envía al navegador, éste tiene que acceder a diversos **repositorios** que contienen textos, diversos tipos de ficheros de los que cabe destacar las imágenes, vídeos y otros tipos de recursos multimedia.

En sus inicios, las páginas web eran sencillas y visualizando el HTML eran relativamente fáciles de interpretar: estaba claro lo que era un título, qué era un encabezado o que algo era más importante porque estaba en negrita. Las enormes posibilidades que ofrece internet hicieron evolucionar las páginas web hasta convertirse en lo que hoy son las modernas Aplicaciones WEB. Las páginas web con las que los Bots tienen que interactuar actualmente son complejas. Se lista a continuación algunas de las características ahora existentes:

- La **separación de contenidos y estilos** en hojas de estilo **CSS.**
- La integración en las páginas HTML de pequeños programas denominados **scripts**, que aportan **funcionalidad dinámica**. El lenguaje más popular actualmente es javascript.
- **Formularios**, para **recoger información** de los usuarios.
- Sistemas de **Autenticación**, para identificar a los usuarios.
- **Cookies**, pequeños ficheros que se almacenan en el ordenador en los que se recoge información de los usuarios y que habilitan que el Servidor Web y los usuarios interaccionen **manteniendo conversaciones coherentes que llamamos Sesiones**.
- **Protocolos Seguros,** como la evolución del HTTP, el **HTTPS**, que hicieron posible el **comercio electrónico y las interacciones seguras**, certificando que las páginas web que presenta el navegador vienen realmente del servidor con el que queremos interactuar.
- Tecnología **FLASH**, que utiliza plugins de los navegadores y **protocolos cerrados.**
- **AJAX** (Asynchronous javascript and XML), estandarizada en el año 2006, un Sistema que posibilita la **consulta asíncrona de la página web con el servidor**, sin necesidad de recargar la página.
- Diversos protocolos, como **SOAP, RMI, RPC, CORBA y especialmente REST**, que utilizan los servidores para interactuar entre sí.
- **Aplicaciones integradas en las páginas web**, denominadas **Widgets**. Las primeras fueron las applets, desarrolladas en el lenguaje de programación Java.

Cuestiones que a las personas nos resultan viables, incluso fáciles o triviales no lo son tales para la inteligencia artificial que hoy somos capaces de programar en un Bot. Buenos ejemplos de ello son los siguientes: diferenciar el tema principal de una página web, lo que es importante y lo que no, lo que es publicidad frente a lo que es el contenido, seleccionar un objeto, una imagen, una fecha en un calendario, capturar la información que va cambiando según la página web se lo solicita dinámicamente al servidor mediante AJAX o simular de forma convincente que es un humano y no un programa quien está manteniendo una sesión con el Servidor Web.

Interaccionando con el Servidor Web y los Administradores de Sistemas

Los Servidores Web tienen que lidiar continuamente con los Bots. La primera línea de interacción es el fichero **"Robots.txt".** En este fichero se le indica a los Bots en qué páginas pueden y no pueden entrar. De todos modos esta es una medida "volunta-

ria", ya que el Servidor Web no tiene una manera de saber a priori qué tráfico es de un bot y qué tráfico es de un navegador tras del cual hay un ser humano. Incluir una página en Robots.txt también tiene sus contraindicaciones: es una señal para un Bot malintencionado de que dicha página contiene información relevante.

La única manera que tienen los **Administradores de Sistemas** que gestionan el Servidor Web de saber lo que ocurre o ha ocurrido en el Sistema es a través de la **Analítica Web**, es decir el análisis de la información que se genera cada vez que llega una petición al Servidor Web. A esta información se le denomina "**Log**". Se consideran varios tipos de logs: de acceso, de error, propios de una aplicación, de kernel, de depuración, etc. Analizar los logs es la manera más relevante que tienen los administradores de determinar diversos problemas del Servidor Web, entre ellas los que puedan ser causadas por los Bots.

Incluso la actividad de, llamémosle un "**Bot de Negocio**" o "**Business Bot**", puede provocar inintencionadamente diversos problemas en los Servidores Web. Si un Bot hace muchas peticiones a un Servidor Web puede desde consumir ancho de banda *(en inglés "bandwidth stealing")* que deberá pagarse al proveedor de comunicaciones que le esté dando el servicio hasta incluso llegar a colapsarlo o al menos reducir su calidad de servicio. Un uso abusivo puede provocar la necesidad por parte del proveedor del Servidor Web de escalar la capacidad, o sea aumentar el número de máquinas para dar el servicio, lo cual conlleva un daño económico. Los administradores de sistemas evitan de facto por ejemplo que una página web aloje referencias a imágenes almacenadas en su servidor ya que cuando se cargue dicha página web le solicitará a su servidor la imagen, consumiendo el ancho de banda correspondiente.

Es importante también evitar que un Bot lance descontroladamente peticiones al servidor ya que puede hacer parecer que se está efectuando un tipo de ataque al que los administradores de sistemas están muy atentos: el ataque de Denegación de Servicio, también conocido por sus siglas en inglés DoS (*Denial of Service*). En este ataque un conjunto de Bots hacen peticiones al servidor web de forma simultánea, agotando su capacidad de responder a peticiones y por tanto colapsándolo. Las técnicas que utilizan los criminales informáticos pueden resultar similares a las utilizadas en Scraping por lo que se debe ser especialmente cuidadoso al utilizar estas técnicas.

Al interaccionar mediante un Bot con un servidor web hay que esforzarse en simular que se trata de actividad humana, evitando por ejemplo el tráfico nocturno, realizar peticiones siempre a la misma hora, o en fines de semana, días de vacaciones, o sin intervalos entre cada una de las peticiones al servidor. Una aparente buena idea, como dejar un Bot pobremente programado funcionando una noche o un fin de semana esperando empezar a trabajar a partir de resultados en el siguiente día laboral, puede acabar en una desagradable sorpresa al descubrir que la actividad del Bot ha sido rechazada por el servidor por considerarla sospechosa y que no tenemos nada con lo que empezar a trabajar. Acceder desde direcciones IP públicas asociadas a centros de datos bien conocidos también puedes ser un problema ya que, lógicamente, se considera que nadie navega desde máquinas ejecutándose en esos centros de datos.

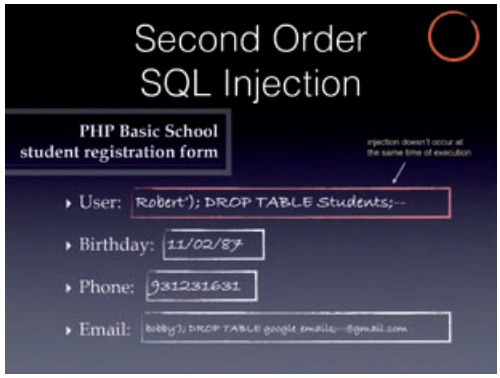
También hay que tener cuidado con la información que enviamos hacia el servidor web. Si ésta no es información estándar puede ser identificada en primera instancia por los Cortafuegos de aplicación (en inglés "*Application Firewalls*") y en segunda instancia en los logs del servidor por los administradores de sistema como actividad inusual y por tanto sospechosa, teniendo como consecuencia el bloqueo preventivo del Bot.

Otra de las cuestiones que vigila el Servidor Web es **desde qué países se reciben las peticiones**. Por ejemplo, una situación legítima es la siguiente: si desde España accedemos a www.nba.com la petición se redirige a http://baloncesto.as.com, en virtud del acuerdo existente entre la NBA y el Diario AS. La recepción de tráfico inusual desde un país concreto en unas circunstancias concretas puede conllevar el tomar medidas preventivas contra el generador de dicho tráfico. Medidas similares se toman por ejemplo en plataformas de comercio electrónico (en inglés *"e-commerce"*), que rechazan pagos realizados por tarjetas de crédito provenientes de países poco habituales por ser una de las técnicas usadas en pagos fraudulentos.

Una técnica para engañar al servidor web es la utilización de un "**Proxy**", un ordenador que intermedia en las peticiones, identificándose a sí mismo como el generador del tráfico, anonimizando así el acceso, hacen pensar que están en otra ubicación. Dentro de este abanico de opciones estarían algunos más:

- ***Open Proxies***, un conjunto de proxies que están disponibles de forma abierta en internet.

- ***TOR***, un servicio que encamina el tráfico a través de diferentes proxies haciendo muy difícil rastrear el origen del tráfico.
- ***Proxies comerciales*** disponibles en el mercado o finalmente tener *un servicio proxy propio*.
- **Crawlera**, un servicio que realiza peticiones a través de un pool de direcciones IP, aplicando diversas técnicas para gestionar los problemas en el Scraping, como el baneado de direcciones IP.



La interacción con **formularios** también ha de ser cuidadosa, en particular la autenticación básica de un usuario, ya que de nuevo la actividad es similar a la de *una* de las técnicas más conocidas de hacking se denomina *Inyección SQL (*en inglés, *"SQL Injection"),* que consigue interaccionar con la base de datos del servidor usando un formulario. Por tanto un error al interaccionar con un formulario va a hacer saltar todas las alarmas del administrador de sistemas al confundir de nuevo la actividad de Scraping con un ataque de un criminal informático.

Las páginas web pueden sufrir cambios con frecuencia, tanto por la propia dinámica de la organización propietaria de la misma como por ser una técnica utilizada cuando la organización no desea que los Bots lean la información en general o alguna en particular. Es por ello que los Bots tienen que ser tolerantes a cambios y a fallos. Además el Bot ha de adaptarse a cambios en la gestión de cookies, congestión de red o problemas de ejecución en el servidor. Hay que tener en cuenta que no sólo el desarrollo del Bot puede ser una tarea costosa sino que la cantidad de gastos operativos de la explotación y mantenimiento del Bot también lo serán.

Otra técnica consiste en mantener una sucesión iterativa de páginas que ha de seguirse para llegar a un contenido en concreto. Por ejemplo esto se aplica cuando ofrecemos a un usuario que se descargue un documento pero previamente queremos que haya dejado sus datos de contacto y además no queremos que se lo descarguen los Bots.

También los administradores de sistemas ponen trampas a los Bots en el código html: una manera es poner un enlace invisible al ojo humano, por ejemplo dentro de una imagen de tamaño 1x1. Si llega tráfico hasta la página enlazada significa que es un Bot ya que ningún humano llegaría de forma natural allí.

Desde otro punto de vista, la empresa está muy interesada en ponerle todas las facilidades posibles a los Bots que indexan y establecen un ranking para las páginas web. Por esa razón surgió una disciplina, el SEO (*Search Engine Optimization*) cuyo objetivo es optimizar las páginas web y facilitar la labor de las Arañas de los Buscadores. Se establece por tanto un equilibrio de fuerzas, un yin y un yang, entre el deseo de obtener la mayor relevancia posible de cara a los buscadores frente a la necesidad de mantener la seguridad y la privacidad del Sitio Web de la empresa.

En toda web se debe poner un apartado legal, incluyendo un apartado de Acuerdo de Servicio (Terms of Service Agreement), indicando la política de uso aceptada en la web, con un apartado a la interacción con Bots. Antes de interaccionar con una web, el administrador del Bot debería leer dicho apartado y adecuar la configuración y programación del Bot para cumplirlo.

HTML5, el nuevo estándar

En octubre de 2014 se publicó HTML5, la quinta versión del estándar HTML. HTML5 incluye novedades, constituyen importantes aportaciones que se han de tener en cuenta a la hora de hacer Scraping[30]:

- **Elementos semánticos**, como <header>, <footer>, <article>, and <section>.
- **Elementos multimedia**, como <video> y <audio>.
- **Elementos gráficos**, como <svg> y <canvas>.
- **Elementos para gestionar conjuntos de datos**, como <datagrid>, <details>, <menu> y <command>.
- Mejoras en **formularios.**

La evolución del HTML en un futuro va ligada a las funciones que hoy en día sólo tiene el sistema operativo de cada dispositivo. Sería deseable que se pudiera acceder a cámaras web, micrófonos, puertos USB e incluso a la CPU y la RAM pero las implicaciones en seguridad informática de estas funciones no son precisamente triviales y acarrean graves riesgos. Posiblemente sea esta la razón por la que no estas posibilidades no han sido implementadas.

[30] W3schools.com "What is New in HTML5": http://www.w3schools.com/html/html5_intro.asp

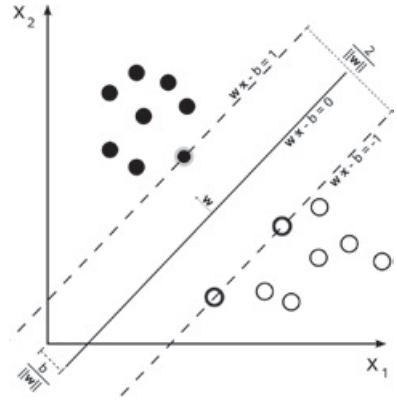
3. Data Science, Estadística, Inteligencia artificial, Data Mining, Investigación Operativa, Machine Learning, Procesamiento del Lenguaje Natural... el entorno de Big Data

Introducimos en un primer apartado algunos términos que nos encontramos habitualmente en torno a Big Data. Todos ellos son áreas de conocimiento a las que Big Data les ha dado alas y cuya combinación e integración están liderando y dando sentido a buena parte de los proyectos Big Data. En el segundo apartado destacamos algunos algoritmos y técnicas que frecuentemente aplicamos en dichas áreas de conocimiento y por ende en Big Data.

Este apartado ha quedado ilustrado con muchos ejemplos y referencias en el apartado 4.6 "Funcionalidades, Implementaciones e Interfaces Big Data para los Sistemas de Vigilancia e Inteligencia", especialmente el 4.6.3 "Implementaciones de las Funcionalidades de Vigilancia, el 4.6.4 "Aplicaciones de Machine Learning y técnicas de Data Science" y 4.6.7 "Integrando Información en la Interfaz de Usuario".

3.1. Algunas áreas de conocimiento utilizadas en proyectos Big Data

ILUSTRACIÓN 5
Máxima separación entre hiperplanos con margen. Técnica utilizada en Machine Learning



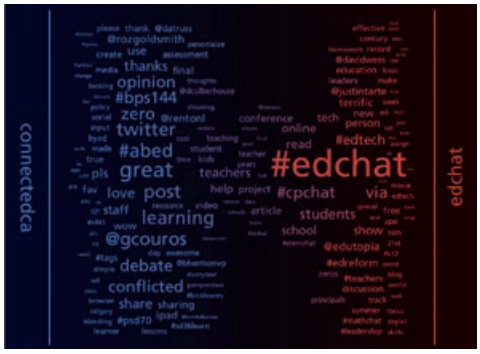
Fuente: wikimedia Commons.

Machine Learning es una **técnica** mediante la que **el ordenador aprende de los ejemplos y el aprendizaje se aplica para resolver los nuevos casos que surjan.** El término **Machine Learning** está de moda en la actualidad, es un término *sexy*, fácil de compren-

der o más bien digamos que todos somos capaces de otorgarle un significado, es un buen término desde el punto de vista de marketing. No se trata únicamente una moda sino que hay un estallido de startups que dedican sus esfuerzos a esta área y que están aportando resultados muy interesantes a empresas e instituciones. Es necesario, sin embargo, ponerla en contexto para entender su rol, su realidad, su proyección y sus relaciones con otras áreas de conocimiento.

En primer lugar, Machine Learning se encuadra, al igual que el Procesamiento de Lenguaje Natural, en la **Inteligencia Artificial**, una ciencia dedicada al diseño y la creación de entidades capaces de resolver cuestiones por sí mismas. Normalmente se implementa en un software y se ejecuta en un ordenador o una máquina robotizada. Actualmente se utiliza como referencia la inteligencia humana. Además de las mencionadas, dentro de la inteligencia artificial se encuadran un gran número de áreas de conocimiento, como son los sistemas expertos, la visión artificial, los algoritmos genéticos, las redes neuronales, las redes bayesianas, las redes semánticas, la lógica difusa, la realidad virtual, los agentes artificiales, el razonamiento automático, la representación del conocimiento o el análisis de decisiones. A Machine Learning le dedicamos el siguiente apartado, 3.4.

La **Estadística** es el área de las Matemáticas que trata del análisis y la obtención de conclusiones a partir de los datos por un lado y la recolección y la organización de los datos por otro. La estadística incluye un conjunto de técnicas, metodologías y modelos que son usados extensivamente por las diferentes áreas de la Inteligencia Artificial, y en particular por Machine Learning.



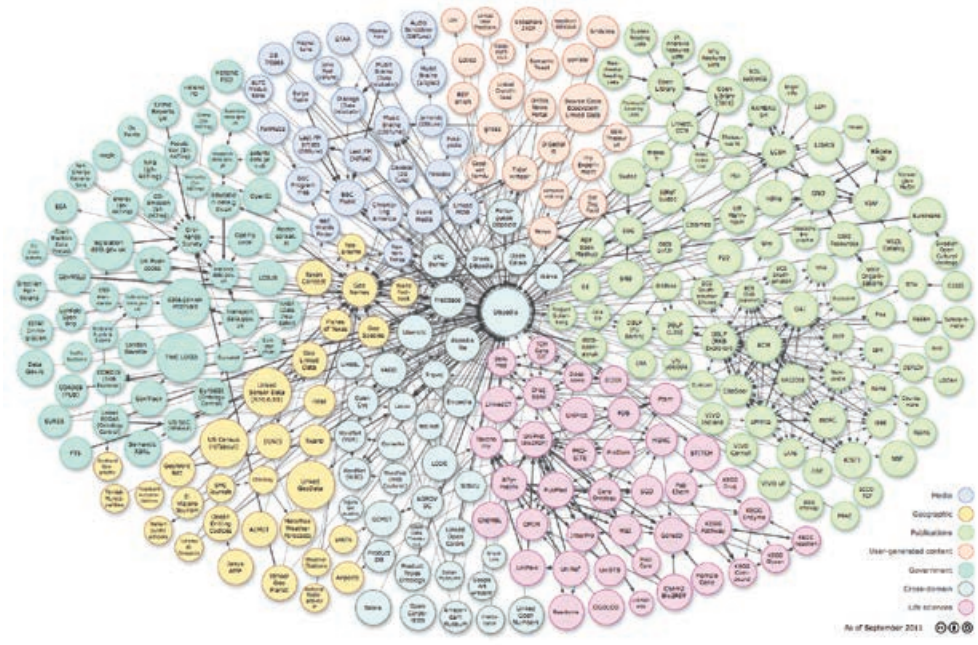
Data Mining consiste en la aplicación de diversos algoritmos a conjuntos subyacentes de datos con el objetivo de resolver problemas relacionados con dichos datos. Varios de dichos algoritmos se utilizan en Machine Learning, por lo que a su vez está usando la Estadística y la Inteligencia Artificial. Sin embargo se suele encuadrar dentro del área de Base de Datos, debido a que los datos están almacenados y organizados en Bases de Datos y suelen ser las personas formadas en Bases de Datos y en Business Intelligence los encargados de tareas relacionadas con Data Mining.

Data Science es el término más novedoso, aparecido entre el florecimiento del término Big Data. Ha resultado ser un término paraguas, al igual que puede ser la palabra

"Informática": no es casualidad que en inglés se traduzca como **"Computer Science".** Su punto de partida es el mismo que Big Data: ingentes cantidades de datos que no pueden tratarse con las técnicas convencionales de tratamiento de datos. Encuadra conocimientos, sobre fuentes de datos y su organización, técnicas de procesamiento masivo de datos, técnicas de tratamiento de datos, repositorios especializados de almacenamiento de Big Data, técnicas y metodologías de análisis de datos y de visualización de dichos análisis. Asimismo incluye una visión horizontal y complementaria sobre conocimientos sobre el hardware, software, sistemas operativos, middleware, frameworks y aplicaciones en general para Data Science. Finalmente también se incluye bajo el epígrafe de Data Science los nuevos modelos de negocio con fundamentos en Data Science y diversos análisis resultado de aplicar Data Science a modelos de negocio, sectores y empresas actuales. Como puede verse, Data Mining, Machine Learning y Procesamiento de Lenguaje Natural quedan albergados para el epígrafe de Data Science.

ILUSTRACIÓN 6

Linked Open Data Cloud Diagram (Wikimedia Commons)[31]



La **Investigación Operativa** es una disciplina que pone el foco en determinar el máximo o el mínimo posible para un problema determinado. Este conocimiento es

[31] https://commons.wikimedia.org/wiki/File:LOD_Cloud_Diagram_as_of_September_2011.png

determinante en numerosas situaciones a las que se enfrenta, entre otras áreas de conocimiento, el Machine Learning. El entorno Big Data permite aplicar la investigación operativa a problemas modelizados con miles de variables y restricciones. Uno de los algoritmos de investigación operativa más populares es el **Simplex**, que está implementado por ejemplo en la popular aplicación **Solver**, de Microsoft Excel. Simplex se aplica a sistemas de programación lineal. La Investigación operativa se aplica también en otras técnicas y metodologías como los procesos estocásticos o las cadenas de Markov.

Al **Procesamiento de Lenguaje Natural** también le dedicamos un apartado en este libro, concretamente el 3.5 y lo comparamos con Machine Learning en el 3.6. Encuadrado en la Lingüística Computacional, está ayudando a solucionar nuevas necesidades del mercado derivadas de las enormes cantidades de texto que sobre todo las Redes Sociales han traído a Internet. Se ocupa tanto de la comprensión como de la generación de textos escritos en cualquiera de los lenguajes humanos. El procesamiento de frases sencillas, complejas, párrafos o documentos nos marcan niveles de complejidad en sus tareas. Lo que aporta ahora mismo es muy básico: el día que aporte valor añadido de verdad será la señal de que la inteligencia artificial está llegando "de verdad" a las máquinas.

3.2. Algunas técnicas útiles para Data Science

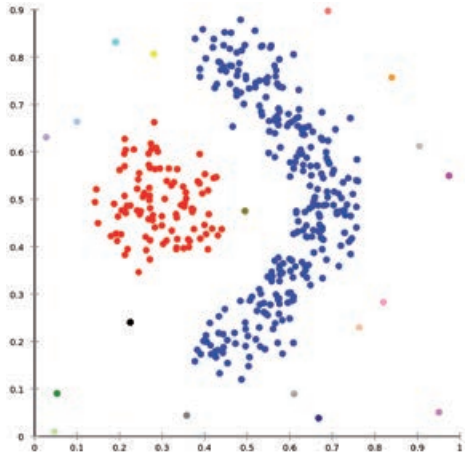
Provenientes de los mundos conexos de la Estadística, la Inteligencia Artificial y la Investigación Operativa, presentamos varios algoritmos y técnicas frecuentemente utilizados tanto en Machine Learning como en Data Mining y proyectos Big Data en general.

Clustering

El **Clustering**, que puede traducirse como **Agrupamiento**, es uno de los análisis más utilizados. Consiste en clasificar los datos en diversos grupos (*clusters*), de tal manera que los datos de cada grupo compartan similitudes, propiedades comunes. En la imagen[32], tomada de la Wikipedia, se pueden observar a simple vista dos agrupamientos de los datos. En este caso puede visualizarse con facilidad, ya que presentamos los valores de dos variables de los datos, cuestión que no ocurriría si estuviéramos representando 4 o más variables.

[32] Image By Chire (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons: http://commons.wikimedia.org/wiki/File:SLINK-density-data.svg

ILUSTRACIÓN 7
Análisis tipo Clustering aplicado a un conjunto de datos



La cuestión más compleja es que el algoritmo determine tres cuestiones:

- El número de agrupamientos (clusters).
- Qué datos se integran en qué agrupamiento.
- La definición de distancia a utilizar.

En la figura[33] se perciben con claridad dos agrupamientos, pintados en rojo y en azul. La cuestión adicional es qué hacer con los puntos que en la figura se identifican con otros colores ¿son ruido o son parte integrante de uno de los dos agrupamientos? Estos puntos podrían ser situaciones excepcionales que merezcan un análisis especial. Por ejemplo, si fuera una representación de un estudio prospectivo de fraude por parte de la Agencia Tributaria constituirían casos de empresas que merecerían atención y análisis especial.

En los algoritmos de clustering es relevante el concepto de **Centroide** del cluster, que es la media de los valores de los datos que pertenecen a un agrupamiento. Cuando queremos clasificar un dato nuevo, medimos la distancia con todos los centroides como primer criterio para incorporarlo a uno u otro *cluster*.

A veces sí que se dispone de información o conocimiento que determina cuantos agrupamientos hay o queremos tener, lo cual te da un punto de partida para ir agrupando

[33] Image By Chire (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons: http://commons.wikimedia.org/wiki/File:SLINK-density-data.svg

datos en torno al centroide. En sucesivas iteraciones se va refinando la clasificación hasta determinar los centroides y las agrupaciones de datos. En otras ocasiones no es así, por lo que se va segregando agrupaciones, a partir de una matriz de distancias entre todos los datos, y hasta que se determina el número de agrupaciones óptimo y los datos de cada agrupación.

En cuanto a la definición de distancia a utilizar dos son las más utilizadas: el **teorema de Euclides**, que todos estudiamos de pequeños para calcular la hipotenusa de un triángulo y la **distancia de Manhattan**, que evoca a la ciudad de los rascacielos en la que para ir de un punto a otro no se pueden atravesar los edificios sino recorrer las calles, lo que en la práctica significa sumar los valores de los dos catetos componiendo un triángulo rectángulo entre dos puntos.

Análisis de Grafos

El **Análisis de Grafos** puede verse como un tipo de Clustering. Un grafo consiste en un conjunto de nodos y un conjunto de relaciones entre dichos nodos. En los nodos representamos cualquier tipo de Entidad y en los relaciones representamos cualquier tipo de transacción, ocurrencia y relación en general entre los nodos. Podemos por ejemplo representar personas y su pertenencia a grupos de interés, relaciones de cliente o proveedor entre empresas, áreas de conocimiento y las universidades que las imparten...

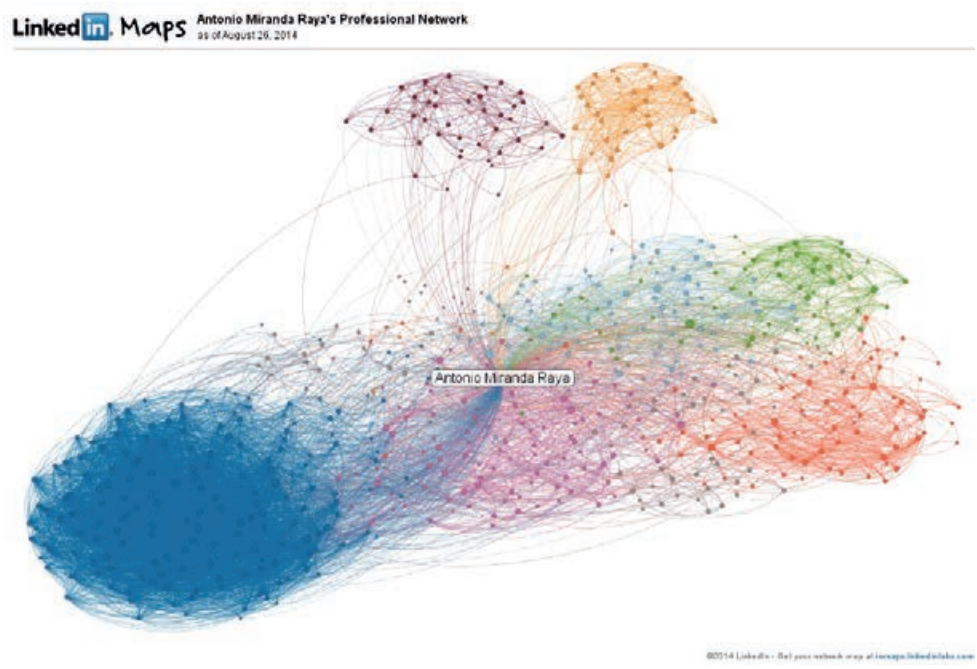
Mediante el análisis de los grafos resultantes pueden descubrirse agrupaciones de entidades, entidades que hacen de puente entre agrupaciones, influenciadores, etc. Este tipo de análisis puede por ejemplo facilitar la realización de campañas de marketing ad-hoc a un grupo de personas, el diseño de productos que respondan a las necesidades concretas de grupos de interés, la búsqueda de líderes de opinión...

La pujanza de las redes sociales ha provocado un creciente interés en este tipo de análisis. Por ejemplo en Twitter podemos encontrar muy significativa la relación de seguidores entre personas y cuentas o también los retweets o los hashtags y palabras clave de tweets. Las relaciones pueden tener dirección: siguiendo el ejemplo de Twitter, ser seguidor de una persona no implica que dicha persona sea seguidor de la primera.

La red social LinkedIn recientemente discontinuó la aplicación INmaps, que podemos ver en la imagen. La herramienta permitía visualizar las relaciones entre los contactos de LinkedIN de una persona. El proceso de creación de los clusters, presentados en diferentes colores, empieza con el estudio de similaridades entre todas las personas de la red, la agrupación de las personas con más similaridades entre si y diferencias con el resto de los miembros, proceso que coincidiría con el proceso de determinación de Centroide que veíamos en el apartado de Clustering y la visualización de los grupos

y las conexiones. Puede visualizarse además la posible relevancia de determinadas personas por ser nexos entre diferentes grupos (puentes).

ILUSTRACIÓN 8
Grafo resultante de aplicar LinkedIN Maps a los contactos en LinkedIN del autor



También se ha hecho muy popular la idea de que todas las personas del mundo están conectadas a través de un máximo de 6 saltos a través de las relaciones entre las mismas. A ese número de saltos lo llamaremos Distancia entre dos nodos. El otro concepto relevante es el de grado o valencia que referencia al número de relaciones que tiene un nodo.

Modelización probabilística de tópicos

La modelización probabilística de tópicos (en inglés "*probabilistic topic models*") se aplica muy frecuentemente a grandes colecciones de documentos. Su objetivo es descubrir y anotar de forma automatizada los **temas** subyacentes en cada documento mediante el análisis de las palabras contenidas en los mismos. Asimismo se identifica la mayor o menor proporción de cada tema en cada documento y cómo cada tema está conectado con otros temas.

Cada tema, denominado **Tópico**, consiste en una distribución sobre un **conjunto de palabras**, es decir que **cada palabra tiene asociada una probabilidad** de aparecer en dicho tópico.

ILUSTRACIÓN 9

Aplicación de LDA con 4 tópicos a un corpus documental

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Un tipo de esta modelización es la LDA[34] (Latent Dirichlet Allocation), donde se asume la existencia de un número de tópicos finito, definido a priori, contenidos en cada documento. Cada documento tendrá asociado cada tópico en una determinada proporción. Una mayor proporción determinará una mayor adscripción de un documento a un tópico, lo que en la práctica nos determina el tema principal y los temas secundarios de un documento.

Consecuentemente se consigue funcionalidad útil para:

- Relacionar documentos entre sí.

[34] "Latent Dirichlet Allocation", David M. Blei; Andrew Y. Ng, Michael I. Jordan, Journal of Machine Learning Research 3 (2003) 993-1022

- Proporcionar una manera de organizar y resumir grandes conjuntos de documentación.
- Recuperación de información.
- Exploración de corpus de conocimiento.
- Manejar, organizar y anotar grandes colecciones de documentos.

En Modelización Probabilística de tópicos asimismo se estudia **cómo los temas evolucionan en la dimensión tiempo y cómo están relacionados diferentes temas entre sí**. Además de la dimensión tiempo, podrían definirse otras dimensiones para el conjunto de documentos y estudiar sus variaciones. Por ejemplo podría ser útil la dimensión "origen" del documento, para estudiar para un mismo tema las diferencias de los documentos según la fuente y el tipo de fuente.

Resulta habitual que en una misma colección aparezcan documentos dispares en el tiempo, y que esa cuestión deba tenerse en cuenta ya que los temas emergen, evolucionan y llegan a un declive. Puede ser de muy alto interés **detectar un tópico emergente** pues puede conducir a diversas necesidades prospectivas, como detectar nuevos productos, nuevas necesidades de negocio, tecnologías emergentes, nuevas empresas competidoras en el mercado, etc.

Existen diferentes variaciones sobre este método, que tienen en cuenta otras cuestiones, como el orden de las palabras en los documentos, el orden de los documentos, no asumir que el número de tópicos es sabido y fijo sino que los nuevos documentos pueden proporcionar nuevos tópicos, valorar la correlación entre tópicos, prohibir palabras en tópicos, incorporar estructuras y modelos en las distribuciones de tópicos, incorporar metadatos, valorar enlaces entre documentos, las distancias entre los corpus de las palabras y los NER (Named Entity Recognition).

Métodos Bayesianos. "Naíve Bayes": el Clasificador Bayesiano ingenuo

Los métodos bayesianos son procesos basados en las reglas de la inferencia Bayesiana, que actualiza las probabilidades a medida que se adquieren nuevas evidencias. En Machine Learning se usa específicamente un conjunto de clasificadores denominados **Bayesianos simples** o "**ingenuo**" (*Naive Bayes classifier*), que asumen una relación independiente entre las variables. También se aplica Bayes en la **Predicción Basada en Modelos**: asumimos que los datos siguen un modelo estadístico concreto y se usa el teorema de Bayes para identificar los clasificadores más adecuados.

Precisamente una de las técnicas más utilizadas para clasificar documentación es el Clasificador Bayesiano Ingenuo (en inglés "Naïve Bayes"). Se basa en una simplificación del teorema de Bayes. Puede aplicarse a otras situaciones, pero es muy utilizada para clasificar documentación.

La idea es la siguiente: tenemos un conjunto de documentos que queremos clasificar por su tema. Una manera sencilla es estudiar las palabras que compone cada documento: según sea el tipo de documentos contendrá unas palabras u otras según el tema del que se trate. El Clasificador Bayesiano ingenuo parte de la premisa, evidentemente falsa, de que la probabilidad de que una palabra aparezca en un documento es independiente del resto de palabras lo cual, aunque no tenga ningún sentido, consigue resultados bastante buenos y es fácil de implementar en un algoritmo y se ejecuta con un alto rendimiento en un ordenador.

En la práctica, tendremos que calcular la **probabilidad de que cada palabra aparezca en uno de los temas** y posteriormente y almacenar dichas probabilidades. Cuando llegue un nuevo documento, extraeremos sus palabras y sus probabilidades de estar en un documento del tema en cuestión, las multiplicaremos entre sí y por la **probabilidad de que haya un documento del tema en cuestión**. Así:

Probabilidad (tema) X Probabilidad (palabra1 | tema) X... X Probabilidad (ultima-palabra |tema)

Lo mismo haremos con otro tema:

Probabilidad (otro-tema) X Probabilidad (palabra1 | otro-tema) X...X Probabilidad (ultima-palabra |otro-tema)

La comparación entre el resultado de las multiplicaciones nos dará el tema ganador al que será asignado el documento.

Lógicamente **esta técnica necesitará de un conjunto de datos de prueba, suficientemente significativo en cantidad y calidad**, que nos proporcione las probabilidades de cada palabra para cada tema.

Regresión

Es el proceso que tiene el objetivo de obtener una función que predice los valores de un conjunto de variables dependientes en función de la variación de los valores de otro conjunto de variables independientes.

La regresión más básica es la **Regresión Lineal**. Se caracteriza por que se expresa como una función matemática lineal:

> **Variable a explicar** = *parámetro-0 + parámetro-1* ***X*** *Variable-independiente-1 + ... + parámetro-p* ***X*** *Variable-independiente-p*

En las regresiones lineales se usan varios indicadores para determinar y mejorar la calidad de la regresión, entre ellos el **coeficiente[35] de determinación R^2**, el resultado de la prueba **F de Fisher[36]**, y de la distribución **T de Student[37]**.

En algunas ocasiones, podemos simplificar el modelo dando menos peso o eliminando los coeficientes de algunas variables compensando con otras. A esto lo llamamos **Regresión Regularizada** (*Regularized Regression*).

La **Regresión no lineal** se basa en funciones no lineales por ejemplo **polinomiales, exponenciales o logarítmicas**. Finalmente mencionaremos una técnica, la **Regresión Segmentada**, por la que se dividen los valores de la variable independiente en intervalos y a cada uno de ellos se le aplica una línea o curva diferente, según el tipo de regresión aplicado.

Árboles de clasificación y regresión

Es una técnica consistente en segregar los datos en grupos según valores de variables, evaluar cada grupo y repetir el proceso en cada grupo creado hasta que el proceso de evaluación determine que no es necesario segregar más grupos. Cuando usamos más de un árbol hablamos de Bosque (Forest). Dos son las técnicas más conocidas: **Random Forest** y **Bagging** (también Agregación de *Bootstrap*)[38]. En Random Forest cada árbol depende de los valores de un vector probado aleatoriamente. Bagging no es estrictamente una técnica de árboles pero se usa muy frecuentemente con árboles. En Bagging se construyen un conjunto de árboles no correlacionados y luego se promedian. Cada árbol proporciona un predictor para un tramo de los valores de una variable.

---

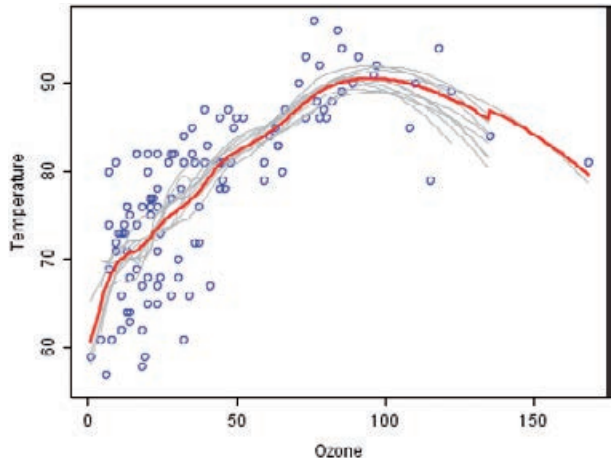
[35] Coeficiente de determinación R2: http://es.wikipedia.org/wiki/Coeficiente_de_determinación

[36] Prueba F de Fisher: http://es.wikipedia.org/wiki/Prueba_F_de_Fisher

[37] T de Student: http://es.wikipedia.org/wiki/Distribución_t_de_Student

[38] Bagging (Agregación de Bootstrap): http://es.wikipedia.org/wiki/Agregación_de_bootstrap y http://en.wikipedia.org/wiki/Bootstrap_aggregating#/media/File:Ozone.png

ILUSTRACIÓN 10
Gráfico de datos de ozono aplicando la técnica de Agregación de Bootstrap



Fuente: (Wikimedia Commons).

Otra técnica habitualmente consiste en combinar varios predictores consiguiendo un predictor más adecuado. A esta técnica se le llama **Boosting**.

Pronóstico ("forecasting")

Con frecuencia las empresas tienen la necesidad de saber con la mayor certeza posible determinadas cuestiones sobre circunstancias futuras. Por ejemplo unos grandes almacenes desearían saber a qué horas y qué días van a venir más clientes por ejemplo para tener suficientes empleados para atenderlos en esas horas en lugar de en horas a las que vendrán menos personas y consecuentemente serán menos necesarios. Sin duda disponen de datos, cada vez más y más sofisticados, sobre días, semanas y años pasados que les pueden ayudar, información sobre las campañas de marketing puestas en marcha, el tiempo meteorológico, circunstancias económicas... pero **no disponen de certeza**, sino que únicamente pueden hacer un **pronóstico**, una **predicción**. Cientos de cuestiones relevantes están en juego para los grandes almacenes: tener stock suficiente de mercancías, ajustar los precios para maximizar el beneficio o para fidelizar a los clientes. ¡La sostenibilidad en el largo plazo del negocio está en juego!

ILUSTRACIÓN 11

Evolución de la empresa Yahoo en bolsa en 2014, capturado de su Web "Yahoo Finance", incluyendo Medias Móviles (superpuesto en verde y rojo) e Indicadores de Volumen y de Fuerza Relativa

Otro ejemplo del día a día: un *trader* de bolsa desearía saber cómo van a evolucionar los mercados, con objeto de realizar las compras y las ventas en los momentos adecuados, con objeto de maximizar las ganancias y minimizar las pérdidas. En la imagen podemos ver la evolución en bolsa de la empresa Yahoo, capturado precisamente en su servicio Yahoo Finance[39].

El pronóstico o predicción estadística, (en inglés "forecasting"), es el proceso de realizar afirmaciones sobre valores futuros proyectando a partir de los valores de los que se dispone.

Si estimamos que la demanda estará alineada con la media de los consumos históricos lo llamaremos **Alisado Exponencial** (en inglés "*exponential smoothing*"). En este tipo de pronóstico el valor pronosticado para un tiempo "t" se compondrá de la media de los valores anteriores sumado a un nivel de error asociado a dicho tiempo "t". En la imagen se puede ver la captura de los valores en bolsa del último año. Habitualmente

[39] Cotización de Yahoo en el servicio Yahoo Finance: http://finance.yahoo.com/q?s=YHOO

se le dará más peso a los datos más recientes frente a los más antiguos, porque resultarán más significativos. También es importante el momento "t" para el que se desea hacer el pronóstico. Si nuestro trader está haciendo compras y ventas intra-día, o sea, que la "t" corresponde al mismo día, le serán relevantes los valores más cercanos en el tiempo. Sin embargo si lo que pretende es hacer una inversión a más largo plazo, por ejemplo a un año vista, sí que puede resultar razonable usar la media de valores del último año.

La idea subyacente consiste en establecer **la premisa de la existencia de un sistema que genera cada uno de los valores y que como consecuencia el sistema será válido para calcular valores futuros desconocidos.** Se trataría entonces de diseñar y construir ese sistema teniendo en cuenta los datos ya existentes. Evidentemente los sistemas van a ser imperfectos y van a tener un nivel de error en cada valor, por eso hablamos de pronóstico, de predicción y no de cálculo matemático. Lo que sí van a hacer estos sistemas es descubrir tendencias y patrones en los datos que pueden ser interpretados como un pronóstico, una predicción.

El más sencillo es el **Alisado Exponencial Simple**. Este sistema tiene dos claves: el **nivel** y una constante denominada **Alpha**.

Valor-pronosticado-T+1 = Nivel-T

En la práctica significa que si tenemos datos del día 1 al día 100 y queremos pronosticar que valor vamos a tener en el día 101, el sistema dice que corresponde al Nivel de dicho día 100. ¿Y cómo se calcula el Nivel de T? Con la siguiente fórmula.

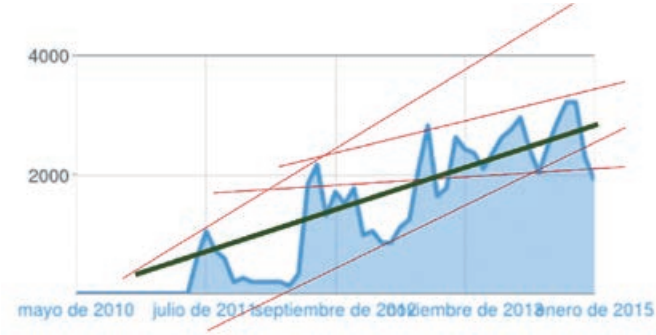
Nivel-T = Nivel-T-1 + **alpha** x (valor-T – nivel-T-1)

Teniendo en cuenta que a Nivel-0 le daremos el valor de la media de los valores de los que disponemos ya que no podemos calcular el Nivel de T menos uno.

¿Y cómo se calcula el alpha? Le damos un valor inicial entre 0 y 1, calculamos todos los niveles, calculamos los errores entre los niveles y los valores calculados y posteriormente buscamos con el algoritmo simplex el alpha que minimice los errores.

ILUSTRACIÓN 12

Aplicación de técnicas de Pronóstico a las visitas a un blog según estadísticas proporcionadas por Google Blogger



El sistema de Alisado Exponencial Simple nos proporciona un único valor, lo cual es bastante pobre. Observando gráficos de ventas, de bolsa, de visitas a una página web, frecuentemente tenemos la impresión de que siguen patrones, que existen correlaciones entre máximos y mínimos, que existe una o varias tendencias que podrían ofrecer información valiosa para la predicción fiable de valores futuros. En la imagen de a continuación se presenta una gráfica de número de visitas a un blog a lo largo de cinco años. He conectado varios máximos y mínimos ilustrando posibles tendencias.

El modelo más popular para establecer una tendencia son los **modelos de Holt-Winters**[40]. En el más sencillo se establece una sencilla ecuación lineal de las que enseñan en primaria a los niños *(y = a*x + b).* En la imagen anterior hemos marcado en verde oliva esta ecuación.

Al parámetro Alpha del modelo anterior se le añade un parámetro de tendencia, denominado **Gamma** de tal manera que tendremos una ecuación:

Valor-Pronosticado-T = **Alpha** X Tiempo-T + **Gamma**

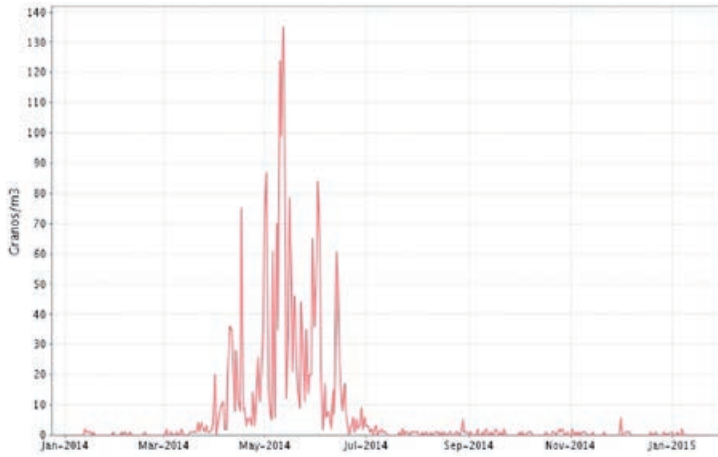
En un modelo más completo de **Holt-Winters**, denominado **Triple Exponencial**[41], se tienen en cuenta la **Estacionalidad** (en inglés "seasonality"), es decir, las tendencias que se repiten a lo largo del tiempo. En este modelo se añade un tercer parámetro vinculado a la estacionalidad, denominado **Delta**.

---

[40] Método de Alisado Exponencial en el que se incluye el Modelo de Holt-Winters: https://es.wikipedia.org/wiki/Alisado_exponencial

[41] Triple Exponential Smoothing: https://en.wikipedia.org/wiki/Exponential_smoothing#Triple_exponential_smoothing

ILUSTRACIÓN 13
Recuento de polenes en Madrid, según www.polenes.com



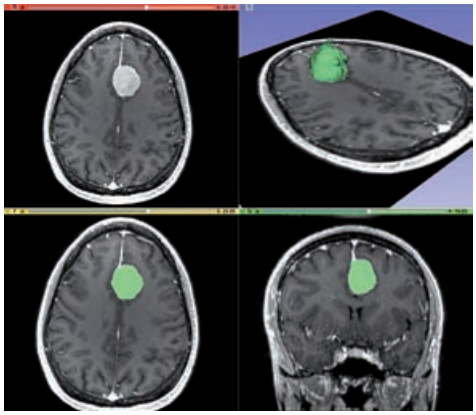
Para entender la necesidad de tener en cuenta la estacionalidad pensemos por ejemplo en el **sector del turismo,** concretamente un negocio de Casas Rurales: nos aparecerán picos de visitas los fines de semana, si lo miramos a escala semanal y en navidades, Semana Santa y verano si lo miramos a escala anual. La ecuación anterior nos marcará una tendencia en el largo plazo pero nos sería mucho más útil tener un modelo que nos ofreciera información sobre si podríamos o no a tener muchos o pocos visitantes en un fin de semana o en una temporada concreta.

Otro ejemplo interesante es el de los **índices de pólenes para los alérgicos**, en los que la estacionalidad es precisamente la clave: en la imagen vemos un gráfico de recuento de pólenes de pino en la ciudad de Madrid, que marca una fuertísima estacionalidad en los meses de primavera. El pronóstico puede ayudar a los alérgicos, al personal médico y a los gestores públicos a reducir el impacto de esta enfermedad en la población.

Esto es lo que nos proporciona el modelo de Triple Exponencial de Holt-Winters. Por último, dado que lo que estamos haciendo es una predicción, se acompaña por intervalos de predicción superior e inferior sobre la predicción efectuada.

4. Machine Learning

Machine Learning, también conocido en español como *"**Aprendizaje Automático**"*, es una disciplina científica habitualmente encuadrada dentro del área de la Inteligencia Artificial, aunque también es posible verla encuadrada en el área de la Estadística.



Se utiliza en aplicaciones de muy diferentes ámbitos: **detección de enfermedades** a partir de datos clínicos, **visión artificial**, **clasificación de documentación** en general (por ejemplo detección de spam), **detección de fraude** en el sector bancario, **segmentación de clientes** (por ejemplo para determinar productos de su interés o gusto), **predicciones en Bolsa**, **inversión de capitales, detección de anomalías en general en los datos** (por ejemplo fallos de red), **predicción de precios, traducción automática o establecimiento de rankings** (por ejemplo en respuestas a búsquedas de información), **predicción de comportamiento humano, detección de ataques informáticos, identificación de sonidos de razas animales en grabaciones, interfaces hombre-máquina en neurociencias o en juegos** y un largo etcétera.[42]

Como podemos ver, **muchos de estos casos tienen que ver con Prospectiva y Predicción**, que son parte de los objetivos de la Vigilancia Estratégica e Inteligencia Competitiva.

De forma coloquial, se suele decir que **Machine Learning** es una **técnica** mediante la que **el ordenador aprende de los ejemplos y el aprendizaje se aplica para resolver los nuevos casos que surjan**. Resulta así sencillo de expresar, de hecho resolver un caso simple puede ser relativamente fácil. Sin embargo la realidad es más compleja puesto que las aplicaciones lo son.

De una forma más estricta, bajo el concepto de *Machine Learning* se encuadra un proceso por el que dado un conjunto de datos ejemplos disponibles resultado de un caso de uso, objetivo o tarea a realizar, se diseña un algoritmo que, generalizando a partir de características de los datos ejemplos es capaz de resolver dicho caso de uso, objetivo o tarea tanto para los datos disponibles como para otros datos disponibles a posteriori a los que se le aplique el algoritmo diseñado, dentro de un margen de error considerado como aceptable.

[42] "MeningiomaMRISegmentation" by Rkikinis at English Wikipedia. Licensed under CC BY-SA 3.0 via Wikimedia Commons: http://commons.wikimedia.org/wiki/File:MeningiomaMRISegmentation.png#/media/File:MeningiomaMRISegmentation.png

Metodología

Una vez que tenemos un objetivo para el que queremos probar Machine Learning, el primer paso consiste en estudiar qué **datos** tenemos, cuántos son y qué **características** tenemos de esos datos. Lo ideal sería que tengamos un conjunto de datos grande, o al menos mediano. Nos podemos encontrar, sin embargo, que sólo dispongamos de un conjunto relativamente pequeño. En ese caso los resultados que demos deberán venir acompañados de una advertencia, ya que cuanto más pequeño sea el conjunto de datos sobre el que saquemos conclusiones, más posible es que el modelo obtenido no sea válido.

La **elección de características** debe hacerse en función del conocimiento experto del problema. Si podemos elegir, los datos deben reflejar la estructura del problema, reflejando la máxima diversidad posible y balancear entre características. Los datos van a tener siempre **ruido**, un conjunto de datos de comportamiento anómalo. Al resto de los datos los llamamos "**señal**": lo ideal es **encontrar un conjunto de datos que reflejen de forma correcta y completa el conjunto de datos señal**. Estas tareas de elección de datos y características son clave, parece una tarea trivial pero es una tarea delicada y complicada: los datos constituirán el alimento del modelo y se cumple el principio de *"Garbage in, garbage out",* es decir, *"si metes basura, consigues basura"*.

Por ejemplo, si estamos intentando identificar un tipo de tumor a partir de imágenes tomadas por un aparato médico nos interesará tener imágenes de todo tipo de pacientes, hombres, mujeres, niños, de diferentes edades, sanos y enfermos, con diferentes enfermedades, con el tumor en diversas etapas de crecimiento y sin el tumor, etc.

Puede ser adecuado enriquecer estos datos con otros datos. Estos datos deben tener únicamente características relevantes siempre íntimamente relacionadas, de lo contrario pueden llevar a conclusiones erróneas.

El siguiente paso consiste en **organizar los datos**. El proceso de Machine Learning conlleva la separación de los datos disponibles en dos grupos: **datos de entrenamiento** (training set) **y datos de prueba** (test set), con una proporción recomendada de 60%-40%. Si hay disponible un tamaño grande de datos podemos considerar un tercer grupo, que llamaremos de **Validación**, separando en dos el conjunto de datos de prueba. Usaremos en el proceso el conjunto de datos de entrenamiento y a posteriori probaremos el resultado con el conjunto de datos de prueba. De esta manera estaremos simulando con los datos de prueba un escenario real en el que llegarán datos nuevos que previamente no estarán disponibles y que deberán ser clasificados por el sistema de machine learning.

A la hora de tratar el conjunto de datos de entrenamiento usaremos también una técnica denominada **Validación Cruzada** (en inglés *"Cross-validation"*). Consiste en dividir el conjunto de datos de entrenamiento a su vez en varios conjuntos de datos de entrenamiento y de pruebas que usaremos durante los siguientes pasos.

A continuación tenemos que **diseñar el predictor** en función de las características disponibles de los datos, **es decir el algoritmo, la función de predicción**, que aplicaremos al conjunto de entrenamiento. Al diseñar este algoritmo tendremos que balancear la exactitud de sus resultados contra otras características necesarias: que sea interpretable, o sea razonablemente de interpretar sus resultados, que sea sencillo, rápido de entrenar y probar y finalmente que sea escalable, es decir que una vez diseñado sea viable el ser ejecutado en un sistema en tiempo real y datos reales.

Si un predictor es muy complicado de entender, usar o poner en funcionamiento, no será mantenible ni integrable en los sistemas, ni la empresa podrá crecer con él, siendo un buen ejemplo de que **"lo mejor es enemigo de lo bueno".** Para ilustrar esto es bien conocido el caso del premio Netflix[43] en la plataforma Kaggle[44], dedicada a competiciones sobre predicción en datos. Netflix, la plataforma de cine online, planteó un premio de un millón de dólares para el equipo que mejorara en un 10% las predicciones que aportaba su sistema de recomendaciones sobre las películas en las que un cliente estaría interesado. Sin embargo, una vez realizado el concurso, no implantaron el predictor que ganó el concurso, sino una versión simplificada del mismo ya que no resultaba viable técnicamente ponerlo en marcha.

El trabajo de diseño del predictor es iterativo. Por una parte tenemos los datos, que ya tenemos organizados en datos de entrenamiento y datos de prueba. Por otra parte tenemos la función de predicción, para la que usaremos algunos algoritmos de base, que explicaremos después. Aplicaremos la función de predicción a los datos de entrenamiento, siguiendo la técnica de Validación Cruzada, que mencionábamos antes. Iremos refinando la función de predicción hasta que nos resulte un modelo con un error aceptable. El resultado final también puede ser la combinación de varios predictores previamente ensayados.

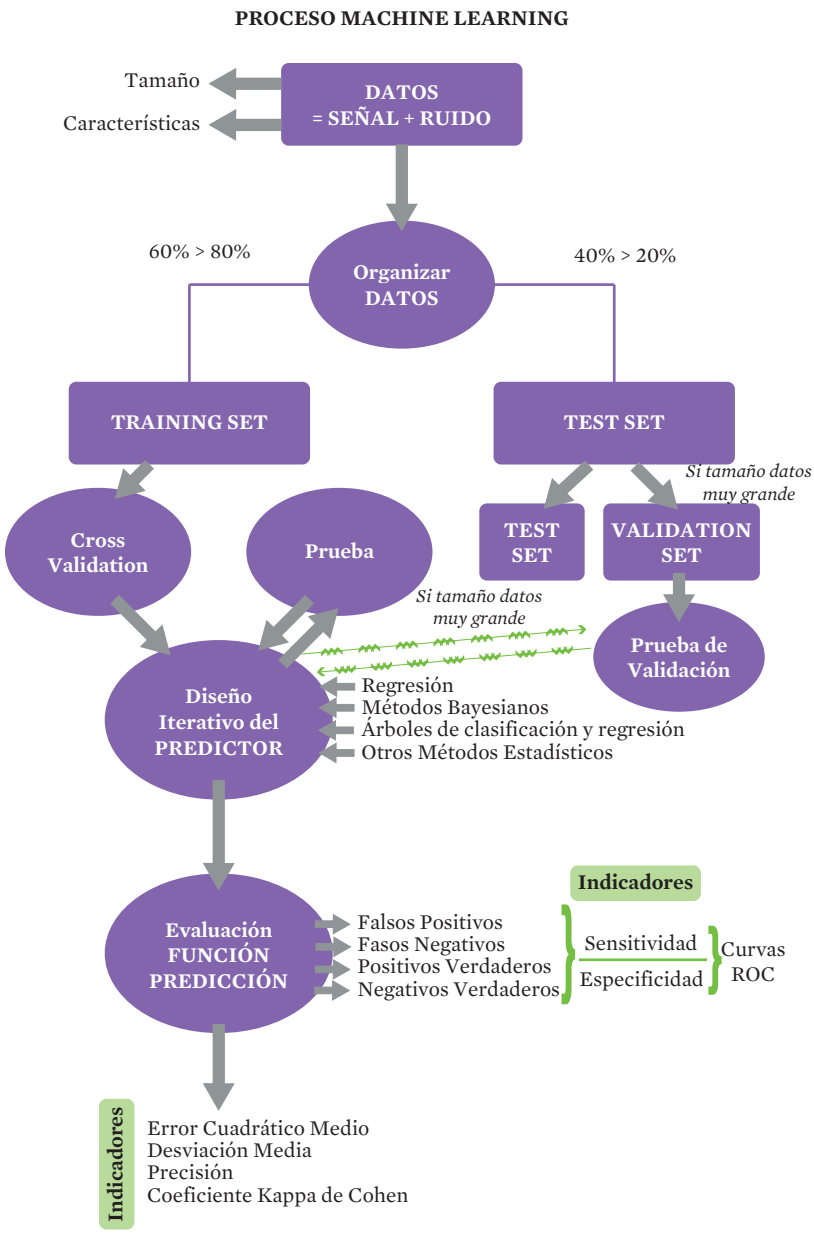
Una vez que lo tengamos **pasaremos a evaluar la precisión de la función de predicción.** Para ello aplicamos la función de predicción a los datos de prueba y mediremos el error resultante. Si el error supera el umbral considerado como aceptable volveremos a empezar a refinar la función de predicción de nuevo.

---

[43] Netflix Prize: http://www.netflixprize.com/

[44] Kaggle, the Home of Data Science http://www.kaggle.com

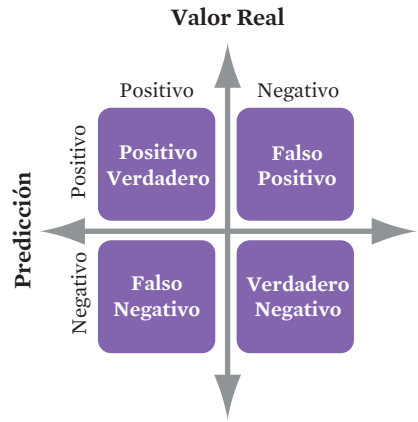
ILUSTRACIÓN 14
Proceso de Machine Learning (original del autor)



Los resultados los clasificaremos de la siguiente manera:

- **Falsos positivos:** nº de resultados identificados como correctos pero que son falsos.

- **Falsos negativos:** nº de resultados identificados como incorrectos pero que son verdaderos.
- **Positivos Verdaderos:** nº de resultados positivos bien identificados.
- **Negativos Verdaderos:** nº de resultados negativos bien identificados.



A partir de esta clasificación establecemos dos **indicadores** importantes:

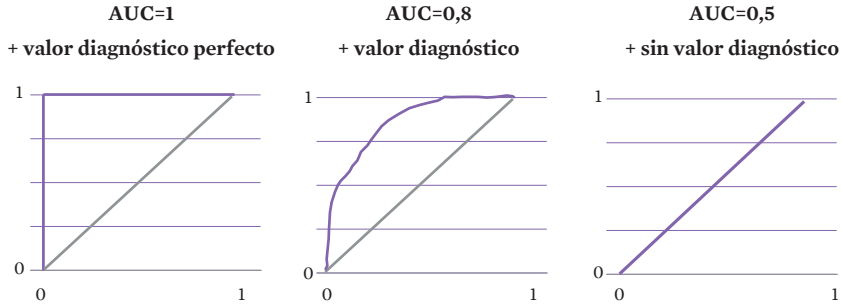
- **Sensibilidad**: (en inglés "sensitivity" o frecuentemente "***recall***") cociente entre positivos verdaderos y todos los positivos reales. Si queremos **maximizar el número de positivos verdaderos** bien clasificados buscaremos un modelo que maximice este indicador.
- **Especificidad**: cociente entre los negativos verdaderos y todos los negativos reales. Si queremos **minimizar el número de falsos negativos** buscaremos un modelo que maximice este indicador.

Estos dos indicadores son usados para determinar la calidad de un predictor, a través de las **Curvas ROC** (Receiver Operating Characteristic), que relacionan la sensibilidad y la especificidad. Construye una curva relacionando la sensibilidad y (1-especificidad). Si el área debajo de la curva (AUC = área under curve) se acerca a uno es que es un predictor muy bueno; a medida que se acerca a 0,5, el predictor pierde valor. En caso de disponer de varios posibles modelos los valores del indicador AUC de cada uno nos ayudará a elegir entre ellos.

A la hora de diseñar el predictor tenemos un riesgo conocido, denominado "**Sobreajuste**" (en inglés "O*verfitting"*), que se basa en el concepto de señal y ruido en los datos. Una función de predicción puede ser muy buena clasificando los datos de entrenamiento, debido a que termina siendo demasiado específica para poder incluir el ruido y a la hora de probarla en los datos de prueba (o en datos de validación, si disponemos de ellos) nos encontramos con que versiones de iteraciones anteriores de la función de predicción que está siendo diseñada muestran mejor comportamiento de predicción. El indicador[45] AUC[46] nos puede resultar muy útil para comparar las diferentes versiones de la función de predicción.

[45] Curva ROC http://es.wikipedia.org/wiki/Curva_ROC

[46] Imagen Curvas.png publicada en http://commons.wikimedia.org/wiki/File:Curvas.png



Otros indicadores usados habitualmente son:

- **Error cuadrático medio**, suma de las diferencias entre los valores válidos y los erróneos elevadas al cuadrado.
- **Desviación media**, media de la suma de las desviaciones absolutas, es decir, el valor absoluto de la diferencia entre cada valor y la media.
- **Precisión:** fracción de los datos clasificados correctamente, es decir, el cociente entre los positivos verdaderos y la suma de positivos verdaderos y falsos positivos.
- **Coeficiente kappa de Cohen (k):** coeficiente que relación la concordancia observada y la concordancia por puro azar.

Para diseñar la función de predicción se utilizan una serie de algoritmos base, provenientes en su mayoría de la disciplina de la Inteligencia Artificial, que pasamos a nombrar a continuación y que son descritos en el apartado de 3.2 "Data Science, Estadística, Inteligencia Artificial...":

- **Regresión.**
- **Métodos Bayesianos.**
- **Árboles de clasificación y regresión.**

5. Procesamiento de Lenguaje Natural [47]

La profusión de textos en redes sociales, especialmente en Facebook y Twitter y la participación de los usuarios en Webs sectoriales especializadas, posiblemente desta-

[47] Textos tomados del Proyecto Fin de Carrera del autor (A. Miranda), "Sistema de Consulta a una Ontología", ETS Ingenieros Informáticos (anteriormente "Facultad de Informática) – UPM.

quen aquí las del sector turismo, han sido los catalizadores para que las aplicaciones de Procesamiento de Lenguaje Natural hayan obtenido la confianza del mercado con el objeto de solucionar la necesidad de gestionar y sacarle partido a todos esos textos. Presentamos en este apartado las bases de esta disciplina, de la que veremos un recorrido enorme en los próximos años.

5.1. Lingüística Computacional y PLN

La **lingüística computacional** es una disciplina que se ocupa de las cuestiones relacionadas con los sistemas informáticos cuyo propósito es el de intentar comprender, analizar o generar textos en lenguaje natural.

Cada sistema informático que contenga un sistema de generación de lenguaje natural tiene sus propios objetivos: generación de informes, generación de cartas, análisis de los contenidos de páginas web, clasificación de artículos dependiendo de sus contenidos, etc. Para ello utilizará un sistema que utilice técnicas de lingüística computacional.

En puridad deberíamos considerar que la Lingüística Computacional es un superconjunto que incluye al PLN, y que incluye otras áreas de conocimiento como la traducción automática, la minería de textos, la ingeniería ontológica o la representación de conocimiento. La realidad es que ambos términos se están usando en el mercado de forma indistinta, de hecho tanto la traducción automática como la minería de textos se consideran como parte de NLP por empresas y grupos de investigación.

Un posible enfoque para diferenciarlos puede ser el foco de NLP en la creación de herramientas que resuelvan problemas relacionados con el tratamiento lingüístico de textos y el foco de la Lingüística Computacional en el estudio de la Lengua utilizando entre otras las herramientas que NLP crea.

Podemos diferenciar dos áreas: **comprensión de lenguaje natural y generación de lenguaje natural.**

La **comprensión del lenguaje natural** puede verse como la traducción de los textos en un conjunto de representaciones propias, denominadas formas lógicas, que están interrelacionadas de forma significativa para el programa a través de una gramática. Hay que enmarcarla en los objetivos con que se plantee el sistema informático. Una etapa de comprensión de lenguaje natural se habrá completado cuando el sistema haya obtenido la información necesaria y suficiente para los objetivos que se plantea resolver.

La **generación de lenguaje natural** es otra área de la lingüística computacional. Se ocupa de construir sistemas que generen texto en lenguaje natural. Los sistemas informáticos de generación de lenguaje natural pretenden construir textos que sean sintáctica y semánticamente correctos, pero además que sean lo más parecidos posibles a los textos que escribiría un ser humano. Por ejemplo no tienen la misma calidad los dos siguientes textos:

- El hierro un metal. El hierro tiene electronegatividad positiva. El hierro es abundante en la Tierra.
- El hierro es un metal que tiene electronegatividad positiva y es abundante en la Tierra.

Tanto para la comprensión del lenguaje natural como para la generación se utiliza una gramática en la que se contempla el lenguaje que es capaz de comprender y/o generar el sistema. Cuando se utiliza la misma gramática para la etapa de comprensión de lenguaje natural que para la generación de lenguaje natural se dice que se cumple la propiedad de la **reversibilidad**[48]**.**

Podemos hablar también de dos tipos de procesamiento, ya sea para el objetivo de **generación** o el de **comprensión**, según la cantidad de texto que tengamos:

- **Procesamiento Táctico:** generación o comprensión de una sola frase.
- **Procesamiento Estratégica o Planificación del texto**: generación o comprensión de uno o más párrafos.

Un aspecto mucho más avanzado y ambicioso en sus objetivos es el **Registro**, definido en la lingüística. Provee un marco de trabajo completo para el estudio del lenguaje. Facilita que se realice todo el trabajo de la generación de textos. Incluye aspectos abstractos y psicológicos.

El registro incumbe a la habilidad del generador de expresar variaciones adecuadas de locución dependiendo de los tres siguientes aspectos de la comunicación:

- **Ámbito**: el asunto o tema del que se está tratando.
- **Tono**: los roles y relaciones interpersonales del interlocutor.
- **Modo**: la situación y significado de la comunicación.

[48] G. VanNoord – Survey of the State of the Art in Human Language Technology. Cambridge University Press, 1996

Podríamos decir que el problema clave del procesamiento de lenguaje natural estriba en que el ser humano no ha terminado todavía de entender cómo estos fenómenos se realizan dentro de nuestro cerebro. Por ello intentamos adecuar tareas que nos parecen lógicas, son algorítmicamente realizables y son adecuadas a los objetivos del sistema diseñado.

Con estas etapas engarzamos un modelo de objetos lógicos cada vez más complejos que arrojan luz sobre el intrincado mundo de la lingüística computacional.

El procesamiento del lenguaje natural se enfrenta a diversas cuestiones de la comunicación, claves para poder realizar tanto la comprensión como la generación adecuada de los textos.

5.2. Procesamiento de frases o un solo párrafo

Muchas veces se ha pensado que la **oración** es la unidad lingüística máxima porque por encima de ella no se ha definido una organización posible pero esto significa reducir el discurso a un conjunto de oraciones unidas al azar lo cual no es lógico. Por ejemplo Alcaraz Varó[49] señala que existe un concepto de la organización discursiva llamado **enunciado**. Se compone de oraciones contextualizadas que actúan como eslabones articuladores del discurso que introducen nueva información y que contribuye a la progresividad textual. Relacionados con el enunciado están la **proposición** y la **oración**. La proposición predica algo de una persona, animal, cosa o acontecimiento. Por su parte, la oración analiza la manifestación externa de la proposición.

Otra cuestión clave a la hora de procesar lenguaje natural es **la información dada y la nueva** adoptan distintos nombres según distintos autores: **tema** y **rema**[50] (en inglés *"theme and rheme"*), **asunto** y **comentario** (en inglés *"topic and comment"*), **dado** y **nuevo** (en inglés *"given and new"*), y **asunto** y **enfoque** (en inglés *"topic and focus"*). Entre ellos, en cierto sentido, pueden establecerse paralelismos.

Se suele considerar al tema como la información previamente mencionada en el contexto lingüístico. Otra visión de los conceptos de tema y rema se basa en considerar al tema como la parte de la información compartida por emisor y receptor dentro de un discurso. Veamos un par de ejemplos:

1. ¿Qué ha hecho **Laura**? — **Laura** ha regado las macetas
2. ¿Quién ha regado **las macetas**? — **Las macetas** las ha regado Laura

[49] Enrique Alcaraz Varó, Paradigmas de la investigación lingüística, Ed. Marfil 1990

[50] M.A.K. Halliday - An Introduction to Functional

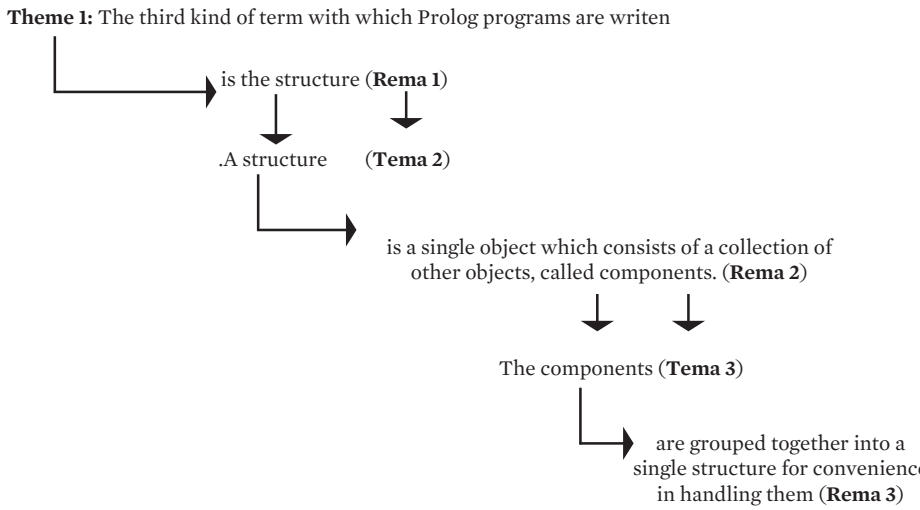
En el primer ejemplo, "**Laura**" constituye el tema de la oración mientras que en el segundo el tema de la oración es "Las macetas".

Otra cuestión a identificar en la comunicación es la **tematización,** es decir, la estrategia comunicativa mediante la cual pasan a posición temática o de arranque del enunciado algunos constituyentes que no suelen estar en posición inicial.

De forma asociada nos aparece el fenómeno de la **progresión temática** dentro del texto. En el siguiente ejemplo vemos una progresión lineal, reconocible por la progresión sin interrupciones de tema a rema:

> ***The third kind of term with which Prolog programs are written is the structure. A structure is a single object which consists of a collection of other objects, called components. The components are grouped together into a single structure for convenience in handling them.***

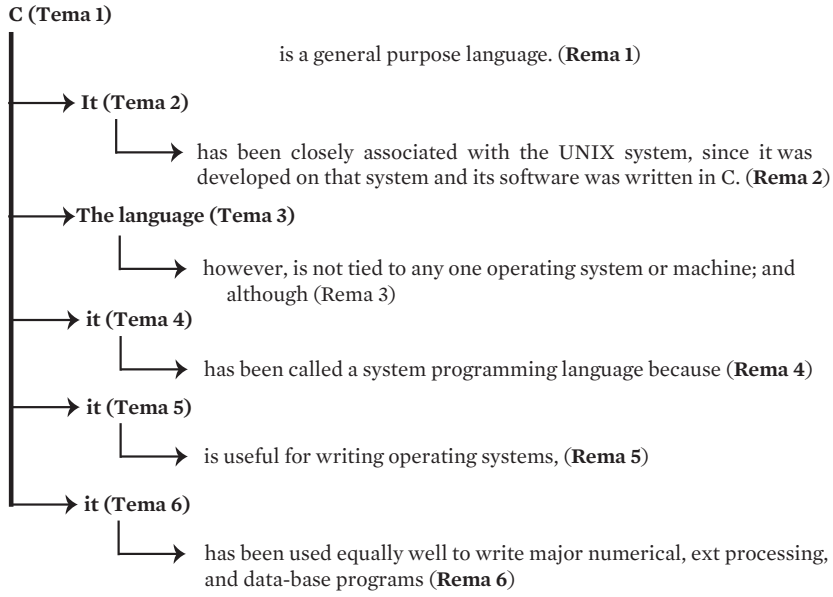
puede esquematizarse de la siguiente manera:



Este otro en cambio se denomina **progresión con tema constante**:

> ***C is a general purpose language. It has been closely associated with the UNIX system, since it was developed on that system and its software was written in C. The language however, is not tied to any one operating system or machine; and although it has been called a system programming language because it is useful for writing operating systems, it has been used equally well to write major numerical, text processing, and database programs.***

Que se puede esquematizar de la siguiente manera:



Existen otros patrones, como el de **progresión con tema derivado**, en el que no se explicita y debe inferirse, normalmente a partir de afirmaciones precedentes o el **desarrollo de un rema bifurcado**, cuando un único tema es usado como introducción de dos remas.

5.3. Procesamiento de varios párrafos

Los conceptos que hemos presentado hasta ahora son útiles a nivel de frase o de un solo párrafo. Lo que viene a continuación está relacionado con **textos de varios párrafos**. La necesidad que subyace al procesamiento de varios párrafos está en identificar un conjunto de patrones que señalen la coherencia de los párrafos y en otro desarrollar un método para procesarlos dinámicamente formando o comprendiendo los párrafos deseados.

Una teoría clásica, que usaremos para explicar la problemática del procesamiento de textos con varios párrafos es la **RST**[51] (Rhetorical Structure Theory) [Mann y Thompson 1988]. En RST **se representan con 25 relaciones las relaciones subyacentes dentro de**

[51] W.C. Mann y S. A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization

los textos que normalmente se encuentran en el inglés. Este número está en principio abierto a nuevas relaciones aunque se presume un ritmo de aparición lento.

Suponen, según esta teoría, que un párrafo es únicamente coherente si todas sus partes pueden finalmente ser construidas para encajar bajo una relación que constituye una **arquitectura de relaciones**. Cada párrafo coherente puede ser descrito por una estructura arborescente que captura las dependencias retóricas entre sintagmas adyacentes y bloques de sintagmas. Ejemplos de relaciones RST son la **Secuencia** (then, next...), **Propósito** (in order to), **Elaboración** y **Concesión**.

RST asume una serie de principios, entre los que destacan las siguientes:

- **Organización**: un texto está compuesto por partes funcionalmente significativas. Dichas partes son elementos de patrones en los que las partes se combinan para crear otras partes más grandes y textos completos.

- **Jerarquía y homogeneidad de la jerarquía**: Un texto está organizado de tal modo que partes más elementales componen partes más grandes que a su vez contribuyen a formar partes aún más grandes. existe un conjunto de patrones estructurales que puede organizar un texto a cualquier escala desde la más grande hasta la más pequeña. A esta serie de patrones se les llama esquemas.

- **Composición Relacional**: el patrón más importante en procesamiento de párrafos es el **patrón relacional**. Se usa un pequeño número de relaciones altamente recurrentes mantenidas entre pares de partes los textos conjunto de para enlazar partes que conformen partes más grandes.

- **Asimetría de las Relaciones**: el tipo más común de relaciones de estructuración de textos son una clase asimétrica llamada relación **núcleo - satélite**. Es asimétrica porque un miembro del par de texto es más central (el núcleo) y otro más periférico (el satélite). Además una parte de texto que es el núcleo va a tener funcionalidades similares a otros núcleos. Una relación consistirá en dos campos:

  - **Restricciones**: comprenden un conjunto de restricciones del núcleo, un conjunto de restricciones del satélite y un conjunto de restricciones de la combinación de núcleo y satélite.

  - **Efectos**: incluye los efectos que posiblemente el escritor intentaba producir en el lector y el "locus" de efecto, identificado tanto como el núcleo sólo o la combinación de núcleo y satélite.

Presentamos a continuación algunos ejemplos explicativos de las 25 relaciones definidas, incluyendo las condiciones para que se den y el efecto que tienen. Además del núcleo (N) y satélite (S), se usan los conceptos de Receptor (R) y Productor (P):

Nombre	Condiciones	Efecto
Prueba	• R no podía creer en N de forma satisfactoria para P. • R considera creíble S. • La comprensión de S por R aumenta la creencia de R en N.	Aumenta la creencia de R en N.
Justificación	• La comprensión de S por R aumenta la disposición de R a aceptar el derecho de P a presentar N.	Aumenta la disposición de R a aceptar el derecho de P a presentar N.
Antítesis	• P considera positivamente la situación presentada en N.	Aumenta la consideración positiva de N por R.
Concesión	• P considera positivamente la situación presentada en N. • P no afirma que no se produzca la situación presentada en S. • P reconoce incompatibilidad entre las situaciones de S y N aunque no las ve incompatibles. • Al reconocer la compatibilidad entre ambas situaciones aumenta la consideración positiva de la situación de N por R.	Aumenta la consideración positiva de R hacia la situación presentada en N.
Circunstacia	• S ofrece un marco en el cual se desea que R interprete la situación presentada en N.	R reconoce que la situación presentada en S proporciona el marco para la interpretación de N.
Solución	• S presenta un problema. • La situación presentada en N es una solución para el problema presentado en S.	R reconoce la situación presentada en N como la solución al problema presentado en S.

6. Procesamiento de Lenguaje Natural versus Machine Learning[52]

Dos son los enfoques más habituales en el análisis de textos. Por un lado tenemos "Machine Learning", que está basado en métodos probabilísticos, que es el enfoque más común. Por otro lado tenemos los enfoques lingüísticos, basados en el conocimiento y estructura del lenguaje, que son menos utilizados

Frecuentemente se perciben estos dos enfoques como alternativas, que compiten entre sí, especialmente en los proyectos de análisis Big Data. Esta percepción constituye un verdadero obstáculo para el progreso de la industria del Big Data. Los dos enfoques deben ser vistos como complementarios. Cuando combinamos los dos enfoques de forma cooperativa obtenemos la forma más efectiva de extraer análisis[53] de la más alta calidad del Big Data.

La creciente importancia del Big Data

El Big Data es sin duda un negocio floreciente. El análisis Big Data está demostrando ser una herramienta efectiva para ser soporte de los procesos de toma de decisiones. Los proyectos Big Data extraen conocimientos y análisis clave mediante la explotación de la estructura de los datos y la captura de las relaciones entre los elementos asociados a los datos, es decir, objetos, conceptos y acciones. Buenos ejemplos de este conocimiento extraído son las preferencias de los consumidores, clientes potenciales, actitudes existentes, las características de los productos, puntos de venta más adecuados, estrategias legales, la satisfacción de los empleados, etc.

Los datos en forma de texto son una parte muy significativa del Big Data. La naturaleza de los datos en los proyectos Big Data son de una variedad extrema, aunque pueden clasificarse en dos grandes grupos: datos numéricos y textos. En cuanto a los datos numéricos destacan las hojas de cálculo y los registros de bases de datos; en cuanto a los textos tenemos tanto los textos generados directamente por personas (noticias, blogs, comentarios, e-mails, redes sociales, etc) y los generados desde programas de ordenador que denominamos **logs**. Hay un montón de información valiosa en estos datos, pero los enfoques de Big Data basados en clasificaciones (opiniones

[52] Texto original en inglés de Antonio Sánchez Valderrábanos, CEO de Bitext. Traducción al español de Antonio Miranda Raya.

[53] A estos análisis y conocimientos en inglés suelen denominárseles "insights". También se pueden traducir como puntos de vista, perspectivas, introspecciones o incluso percepciones, según el contexto. El término se refiere al proceso de extraer conocimiento de los datos y la información.

positivas y negativas, categorías, etc) extraen únicamente información superficial, fundamentalmente porque no abordan el entendimiento de la estructura del lenguaje.

Machine Learning, la técnica más popular actualmente, **lidera este tipo de enfoques Big Data** basados en clasificaciones. Las técnicas de Machine Learning están basadas en marcos conceptuales provenientes de las matemáticas y la estadística por lo que puede parecer sorprendente que se haya convertido en las herramientas más habituales para minería (en inglés "text mining") y análisis de textos.

No debemos considerar enfoques competidores a la Linguïstica y al Machine Learning, a pesar de que esta visión es la que se ha extendido de forma generalizada. Este malentendido viene de algunos proyectos Big Data en los que se mezclaron varios tipos de enfoques de Machine Learning, que pueden en principio integrarse indistintamente uno antes que otro y viceversa frente a la integración de Machine Learning, Métodos Probabilísticos y Aplicaciones Lingüísticas, que no hacerse adecuadamente pueden resultar incompatibles. Esta idea equivocada no está basada en bases sólidas: los enfoques de Machine Learning y la Lingüística Computacional pueden trabajar de forma conjunta. Los enfoques lingüísticos son los más adecuados para entender el lenguaje y proporcionar la estructura que Machine Learning necesita para extraer conclusiones precisas de los textos.

¿Por qué usamos Machine Learning para Análisis de Textos?

Los enfoques estadísticos son una opción de rápida implementación pero limitada. Son dos las razones por las que probablemente son la tendencia más popular actualmente: en primer lugar, de forma primordial, existen herramientas software que implementan estos enfoques matemáticos que son las herramientas naturales y bien conocidas por muchos ingenieros y científicos de datos, que los hace la opción más natural y con la que se sienten más cómodos a la hora de enfrentarse a un problema. Al convertir un caso de uso en un problema de clasificación, *clustering* o modelización, cualquier ingeniero sin conocimientos lingüísticos es capaz de obtener resultados inmediatos. Esta visión está cimentada por el relativo éxito de los enfoques estadísticos a la hora de solucionar problemas de reconocimiento del habla y traducción automática, incluso cuando los enfoques lingüísticos ya han madurado y son más eficientes.

Los perfiles con formación en enfoques lingüísticos no se encuentran con facilidad. Además requieren tener conocimiento extenso tanto en lingüística como en ingeniería informática. Desafortunadamente las personas formadas en esta disciplina, que llamamos Lingüística Computacional, son escasas.

Algunos inconvenientes de Machine Learning

Machine Learning ignora la estructura de la frase. Los productos y soluciones comerciales de análisis de textos fundamentadas en Machine Learning no tienen en cuenta la estructura de la frase. En su lugar, utilizan el enfoque de "bolsa de ideas", que ignora las relaciones entre las palabras y por ende el conocimiento que se deriva de las mismas.

Detectar una similaridad entre "*Seguridad Social en la Red*" y "*Seguridad de la Red Social*" ilustra uno de los errores que cometerán los paquetes de análisis de textos basados en machine learning. Otro ejemplo sería la detección de un sentimiento positivo en un tweet al leer la expresión *"bien perecedero",* confundiendo el sustantivo "bien" por el adjetivo, en principio de tono positivo. De la misma manera no se tienen en cuenta el efecto que cierto tipo de palabras tiene en otras. El ejemplo más claro de este tipo de fenómenos habituales en el lenguaje es el de la negación: "*No me gusta este teléfono*" constituye claramente una opinión negativa, aunque no aparezca ninguna palabra similar a *disgustar, odiar* o *despreciar".* Otra situación habitual es de las frases en condicional: "Recomendaría este teléfono si la pantalla fuera mejor" es claramente una opinión negativa que muy probablemente sería identificada como positiva por contener la palabra "recomendaría", que al procesarla por *stemming* quedaría reducida a "Recomendar". Tampoco son capaces de gestionar la granularidad del lenguaje: una frase como "*La pantalla es maravillosa pero odio el teclado en pantalla*", que contiene dos opiniones distintas que deben ser evaluadas de forma separada.

Machine Learning necesita datos de entrenamiento. Entrenar un programa que implementa Machine Learning es una cuestión no menor. Todo el proceso de elegir el conjunto de datos de entrenamiento y todo el proceso de entrenamiento no es una cuestión trivial. Fenómenos que ocurren en Machine Learning, como el **sobreajuste** (en inglés "overfitting") puede etiquetar como positivos o negativos conceptos en principio neutrales como "Harvard" o "Stanford".

Otro fenómeno a tener en cuenta es el del diferente Registro de las fuentes de datos. Big Data es sinónimo de Variedad de datos y por tanto de variedad de textos, que puede ir desde el lenguaje formal de noticias de cualquier medio de comunicación al lenguaje informal de los correos electrónicos, transcripciones de conversaciones en un call-center, respuestas a encuestas, comentarios en redes sociales, etc. Cada tipo de textos requerirá de diferentes necesidades de entrenamiento, lo que constituye un gran desafío. Otra situación similar es la que nos proporcionan los objetivos de negocio, que pueden marcar los textos generados. La mejora del servicio al cliente, prevenir la fuga de clientes, generar clientes potenciales, prevenir impagos... todos ellos son objetivos de negocio que van a provocar importantes particularidades en los textos generados.

La necesidad del proceso de entrenamiento es una limitación para el Machine Learning. Por definición, Machine Learning resuelve los problemas para los que ha sido entrenado por lo que la variedad de tipos de datos de entrada y la variedad de objetivos de negocio, constituyen un desafío para los principios en los que se basa ya que para cada caso y tipos de datos se ha de crear datos de entrenamiento, lo cual es un proceso ejecutado a mano por lo que es costoso y sensible a errores. Por tanto la necesidad de entrenamiento es un obstáculo mayor para el éxito de forma sostenible de Machine Learning como una manera de extraer análisis y conclusiones de los textos.

El tipo de resultado que nos proporciona Machine Learning es lo que se denomina en los entornos IT "una caja negra", en la que si el conjunto de datos de entrenamiento origina que el sistema clasifique incorrectamente una frase, visto desde un punto de vista sintáctico, o detecta un sentimiento equivocado, visto desde un punto de vista de Caso de Uso, no existe una forma sencilla de ajustar el sistema para corregir el error sino que debemos proceder a un entrenamiento adicional o re-entrenamiento del sistema. En cualquiera de los dos casos resulta un proceso costoso, derivado de la necesidad de recoger nuevos datos e incorporarlos al sistema.

¿Cómo aborda el Análisis Lingüístico todas estas cuestiones?

La Lingüística comprende la estructura de las frases. El Análisis Lingüístico utiliza el conocimiento sobre el lenguaje (gramáticas, ontologías y diccionarios) lo que permite tratar con la estructura del lenguaje a todos los niveles: morfológico, sintáctico y semántico.

El Análisis Lingüístico puede tratar con exactitud fenómenos complejos como la negación o los condicionales, gracias a que tiene en cuenta la estructura del lenguaje, especialmente en casos complejos donde para diferentes significados el conjunto de palabras utilizado es similar. Un buen ejemplo serían las frases "No voy a comprar este producto" y "Si no compro este producto hoy podría hacerlo mañana que DLA trataría correctamente.

El Análisis Lingüístico aporta un gran valor añadido en la granularidad de las frases, es decir encontrando diferentes significados en las frases gracias al correcto entendimiento de la estructura de las frases. Por ejemplo, en la frase "la pantalla es maravillosa pero no me gusta el teclado en pantalla", detectamos dos partes en la frase, una con opinión positiva y otra con opinión negativa que podría resultar una opinión neutral. También es reseñable que el Análisis Lingüístico nos permite identificar los temas y conceptos que están siendo discutidos, pudiendo consecuentemente asociar la opinión positiva a la pantalla y la negativa al "teclado en pantalla".

El Análisis Lingüístico **ya salió de los laboratorios de investigación**, **ya tiene nivel de madurez suficiente para los usos empresariales**, desde los 140 caracteres de un tweet de Twitter hasta largos documentos del ámbito legal. Las gramáticas computacionales, las ontologías y los diccionarios describen eficientemente la estructura y contenido del lenguaje y consecuentemente puede ser aplicado a diferentes tipos de textos.

Frente al enfoque de "caja negra" de Machine Learning, el Análisis Lingüístico usa el enfoque denominado de "caja blanca" (en inglés "glass box"), en el que las reglas y el código son programadas explícitamente y cualquier mejora puede implementarse fácilmente bien añadiendo nuevas reglas, bien modificando las existentes, en ambos casos con resultados predecibles.

Un motor de Análisis Lingüístico puede ser configurado para analizar una amplia variedad de textos, basados en las semejanzas que comparten las expresiones del lenguaje humano. Por ejemplo, es factible definir léxicos y gramáticas capaces de analizar diferentes tipos de noticias o textos de redes sociales. Asimismo el Análisis Lingüístico es adecuado para ser adaptado a abordar todo tipo de aplicaciones de negocio: generar clientes potenciales, prevenir la pérdida de clientes (en inglés "customer churn").

Buenas noticias: la Lingüística y Machine Learning son compatibles

El Análisis Lingüístico proporciona dos ventajas fundamentales:

- Estructura diferentes tipos de texto y diferentes tipos de propósitos
- Convierte textos desestructurados en textos estructurados.

Con Machine Learning podemos analizar y extraer conclusiones de los textos estructurados que nos proporciona el Análisis Lingüístico por lo que podemos concluir que la combinación del Análisis Lingüístico y Machine Learning pueden proporcionar aplicaciones con una fiabilidad y fiabilidad muy alta. Este es el tipo de resultados que el mercado está demandando hoy en día.

La propuesta, por tanto, consiste en diseñar y construir un proceso Big Data con dos fases: una primera de extracción de una estructura lingüística de los textos y generando una representación exacta y enriquecida de los mismos y una segunda en la que aprovechamos las capacidades de Machine Learning para extraer análisis y conclusiones de esta representación de los textos. Los resultados serán presumiblemente mucho más adecuados y exactos.

Dos Casos de Estudio

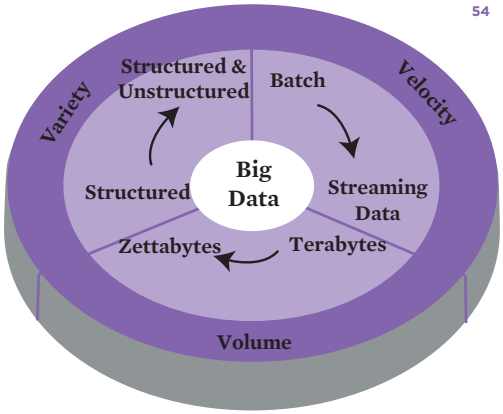
Los siguientes Casos de Estudio son Buenos ejemplos de cómo la el Análisis Lingüístico ayuda a solucionar objetivos de negocio específicos, como por ejemplo identificar qué aspectos de mis servicios son valorados positivamente por mis clientes. Estos objetivos pueden ser logrados sencillamente analizando el contenido de las fuentes de datos que habitualmente están disponibles para cualquier Empresa o Institución y alineando estos análisis con objetivos de negocio específicos.

Identificación de fortalezas y debilidades en Enterprise Feedback Management (EFM). Una empresa proveedora de EFM proporciona servicios a una cadena de hoteles, monitorizando opiniones registradas en páginas web por usuarios de los hoteles de la cadena. Las opiniones se recogen de fuentes públicas como TripAdvisor o Expedia, recogiendo millones de opiniones por día en diferentes idiomas. El Análisis Lingüístico puede proporcionar no sólo si las opiniones son positivas o negativas, sino conocimiento significativo como qué aspectos de la experiencia de cliente (la habitación, los empleados, el servicio al cliente, etc) son percibidos como positivos o negativos y porqué: la habitación era demasiado pequeña, los empleados no fueron profesionales.... También resulta muy valioso el poder categorizar la información disponible según los objetivos de negocio. De esta manera podemos organizar todo el conjunto de opiniones priorizando aquellas relativas por ejemplo a la atención del cliente o la percepción de la marca y presentando los aspectos específicos de estos objetivos de negocio que necesitan ser mejorados o los aspectos positivos que pueden ser explotados, por ejemplo en una campaña de marketing.

Generación de listas de clientes potenciales en un CRM. Un banco estaba interesado en descubrir nuevas oportunidades de negocio para su línea de préstamos para empresas. El banco quería monitorizar fuentes públicas de noticias, buscando historias de las que se pudieran deducir necesidades financieras, como el lanzamiento de nuevos productos, nuevas instalaciones de producción, nuevas líneas de inversión o fusiones y adquisiciones. Los sistemas de análisis de textos tradicionales clasificarían las noticias en categorías y temas pero no serían capaces de extraer la información crítica para que el sistema sea efectivo: qué empresa está lanzando un nuevo productos, qué compañía está siendo adquirida y por quién y cuándo ocurrirán estas situaciones. El Análisis Lingüístico utiliza e incluye en su sistema el conocimiento necesario para implementar los objetivos de negocio siendo de esta manera capaz de detectar para el banco clientes potenciales mucho antes que sus competidores, anticipándose de esta manera también a las necesidades de sus clientes.

7. Arquitectura Big Data

Las Arquitecturas Big Data surgen ante el nuevo requisito de gestionar el crecimiento exponencial del volumen de datos en los Sistemas, la velocidad a la que estos datos están siendo generados y la variedad de los mismos, es decir, los diferentes tipos de datos que existen, surgen y consecuentemente han de gestionarse. En torno a estos tres conceptos se ha hecho famosa la idea de "Las 3 Vs de Big Data: **Volumen, Variedad y Velocidad** (en inglés *"Volume, Variety and Velocity")*.



[54]

A esta V también se le ha añadido una 4 "V": la Veracidad (*"Veracity"),* la fiabilidad de los datos por su aplicación en los procesos de toma de decisiones. Este concepto lo desarrollamos más en el apartado "Gestionando el Conocimiento y la Veracidad de la información". Otras versiones hablan de la "V" de "Valor" ("Value"), con un sentido similar y de la "V" de "Viscosidad", haciendo referencia a la mayor o menor facilidad para correlar los datos. Es realmente tentador esto de encontrar conceptos que empiecen con "V" y relacionados con Big Data, no hay ninguna duda.

La solución a estas necesidades aportará asimismo varias oportunidades:

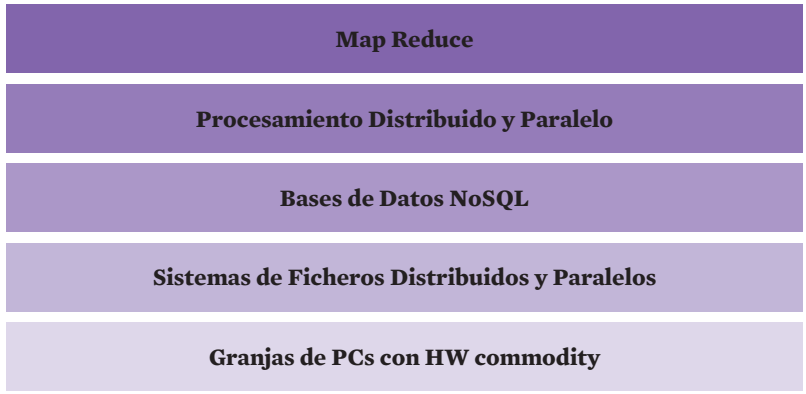
- Incrementar la cantidad de datos que se gestionan en los Sistemas de Vigilancia.
- Incluir en los sistemas el tratamiento masivo e inteligente de datos no estructurados.
- Generar nuevos indicadores, reflejo de nuevas capacidades generadas.
- Tratamiento analítico de datos en tiempo real.
- Generación de información predictiva y prospectiva que puede ser integrada en otros sistemas.

Con estas arquitecturas se multiplican las posibilidades de la vigilancia estratégica y la inteligencia competitiva y lo que es más importante, crece enormemente el valor añadido y por tanto las razones para integrar información y capacidades de vigilancia

[54] Caracterización de las 3Vs del Big Data (Volumen, Velocidad, Variedad), según IBM.

e inteligencia competitiva en las grandes aplicaciones IT y los Sistemas de Información de las empresas.

ILUSTRACIÓN 15
Capas de una Arquitectura Big Data



7.1. Distribución y Paralelismo en el Sistema de Ficheros

En los nuevos sistemas Big Data es necesario el uso de sistemas distribuidos para la ejecución distribuida de aplicaciones y el acceso paralelo a los datos.

Los Sistemas de Ficheros Distribuidos permiten incorporar miles de ordenadores independientes que se convierten en nodos independientes de una misma red. Gestionan distintos dispositivos en diferentes nodos de forma transparente a usuarios y aplicaciones ofreciendo servicios con las mismas prestaciones que si fuera un sistema de ficheros centralizado. En caso de que un nodo falle, el sistema de ficheros gestiona automáticamente la situación. Los Sistemas Distribuidos destacan, por tanto, por su escalabilidad.

Los Sistemas de Ficheros Paralelos son un grupo específico dentro de los sistemas distribuidos que se caracterizan la distribución de datos entre múltiples dispositivos de almacenamiento y el acceso paralelo a los mismos. Se popularizan ante las siguientes necesidades:

- Transportar una gran cantidad de archivos sobre la red puede causar bajas prestaciones debido a la alta latencia, cuellos de botella en la red, alta escalabilidad y sobrecargas.
- La necesidad creciente de las aplicaciones de manejar repositorios masivos de datos.

- La falta de crecimiento del ancho de banda y la latencia a los discos frente al crecimiento enorme de su capacidad.

En ambas situaciones se cuenta con que los fallos en el hardware y en el software son norma, y no excepción y que se ha de gestionar la situación para que el sistema sea tolerante a dichos fallos.

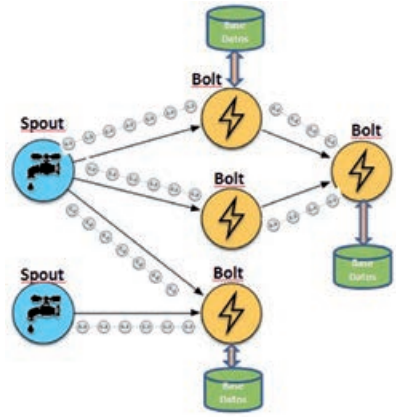
Google File System y sobre todo **HDFS** (*Hadoop Distributed File System*) son los dos Sistemas de Ficheros más populares en la actualidad que cumplen estas características.

7.2. Sistemas de Procesamiento Big Data

Se presentan a continuación los Sistemas de Procesamiento Big Data, que **basculan entre renovados Sistemas de Procesamiento por Lotes hasta los potentes Sistemas de Stream Computing y algunas soluciones mixtas.** Dependiendo de los requisitos de latencia, rendimiento y tolerancia a fallos se elegirá uno u otro sistema de procesamiento.

Será especialmente relevante determinar si la Fuente de Datos puede, en caso de fallo del sistema de procesamiento, disponer de nuevo o no de un mensaje previamente recibido y si los mensajes pueden volverse a encontrar dado un criterio de búsqueda, para lo cual será necesario estar respaldada por una arquitectura con sistema de ficheros distribuido.

ILUSTRACIÓN 16
Spouts & Bolts en "Real time Big Data, Apache Storm Arquitecture and Integration" en Hadoop Summit Europe 2014

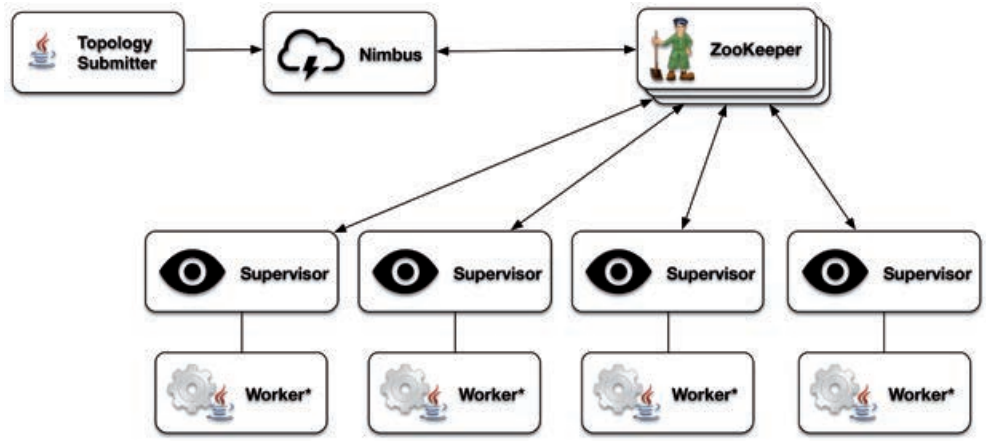


La inclusión de fuentes de datos que generen información masiva en tiempo real en nuestro Sistema de Vigilancia era una tarea técnicamente complicada y frecuentemente no viable tanto técnica como económicamente. Ante esta necesidad nacieron los Sistemas de **Stream Computing,** que permiten leer y analizar datos en tiempo real. Son sistemas escalables, formados por **redes de nodos**, que procesan miles de mensajes por segundo, tolerantes a fallos y confiables (en inglés, "reliable"), que garantizan la entrega del mensaje. Sus modelos son sencillos, basados en topologías, con pocos tipos de nodos que ejecutan tipos de tareas sencillas.

Uno de los sistemas más populares es **Apache Storm**, que cuenta con dos tipos de nodos: nodos "**Spouts**" y nodos "**Bolts**"[55]. Los spouts convierten flujos de datos en tiempo real en flujos de **tuplas *clave-valor*** y los emiten hacia nodos bolts que ejecutan tareas sencillas, como la lectura o escritura de una base de datos o un procesamiento simple de la tupla. Opcionalmente vuelven a emitir la tupla hacia otro nodo bolt. Cada spout y bolt es ejecutado en paralelo en múltiples ordenadores.

ILUSTRACIÓN 17

Nimbus, Zookeeper, nodos Supervisores y Worker en "Real time Big Data, Apache Storm Arquitecture and Integration" en Hadoop Summit Europe 2014



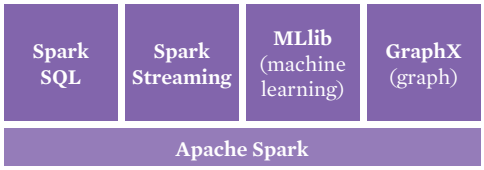
Storm agrupa las tareas asegurando que todas las tuplas con los mismos valores son enrutados hacia la misma tarea. Cuenta con dos aplicaciones de soporte, **Nimbus** y **Zookeeper**. Nimbus recibe la topología diseñada, calcula las asignaciones y se las envía a Zookeeper. Zookeeper envía a nodos supervisores las asignaciones, que lanzan nodos trabajadores

[55] Taylor Goetz, Apache Storm Committer – Hortonworks http://www.slideshare.net/ptgoetz/storm-hadoop-summit2014

para ejecutar la topología. Para asegurar la tolerancia a fallos, los nodos trabajadores informan periódicamente a Zookeeper de que siguen activos y este a Nimbus. Si no llega la notificación, el Supervisor reinicializa el nodo trabajador. Si falla repetidamente el nodo, sea trabajador o supervisor, Nimbus reasigna el trabajo a otros nodos.

Los nodos Bolt emiten señales de ACK o de FAIL para notificar que la tarea ha sido o no ejecutada, haciendo así al sistema confiable. Estos envíos se hacen a través de nodos Bolt especializados únicamente en esta tarea.

A caballo entre los dos sistemas está el **Micro-batching**, que es una técnica que permite empaquetar flujos (stream) de datos entrantes en paquetes para su tratamiento por un sistema de procesamiento por lotes. Un ejemplo es **Trident**, una abstracción de alto nivel basada en Apache Storm. Trident divide los lotes en particiones, cada una orientada a ser ejecutada por un nodo Bolt.



También otro de los referentes en Big Data, el motor de procesamiento de datos a gran escala **Apache Spark**[56] hace *micro-batching*[57] a través de su extensión **Spark Streaming**, un sistema de computación en clusters, de propósito general caracterizado por su alta velocidad. Apache Spark[58] se basa en un módulo core que proporciona funcionalidad básica para planificación y gestión de tareas y de entrada y salida de datos. Define un concepto especialmente relevante, denominado *RDD* (en inglés "*Resilient Distributed Datasets*") que constituyen colecciones lógicas de datos distribuídas entre varias máquinas. Spark permite referenciar a las RDDs a través de APIs. Su arquitectura está orientada al procesamiento con la memoria RAM (en inglés "*in-memory*") en lugar con estar orientado a trabajar con la memoria en disco duro, como hace Hadoop. Esto permite un rendimiento muy superior en algunos tipos de procesamiento, por ejemplo los que se utilizan en Machine Learning.

Cuenta con 4 librerías principales:

- **Spark SQL,** que habilita las consultas mediante el lenguaje SQL a una abstracción denominada SchemaRDD, que da soporte a datos estructurados y semi-estructurados.

[56] Proyecto Spark en Apache http://spark.apache.org/

[57] Taylor Goetz, Apache Storm Committer – Hortonworks http://www.slideshare.net/ptgoetz/apache-storm-vs-spark-streaming

[58] https://en.wikipedia.org/wiki/Apache_Spark

- La librería de **Machine Learning Mlib**, que explota las especiales capacidades de Spark para mejorar enormemente el rendimiento frente a otras aplicaciones.
- **GraphX**, componente dedicado al procesamiento gráfico distribuido.
- **Spark streaming**[59]**,** permite el procesamiento de flujos continuos de datos (en inglés "*live data streams*"). A los datos se les aplican funciones de alto nivel, como las conocidas Map y Reduce o funcionalidades de procesamiento de gráficos o machine learning con MLib o GraphX y su resultado almacenado en bases de datos o sistemas de ficheros distribuidos o publicados en sistemas de visualización Big Data.

Seguirán vigentes aquellas aplicaciones que tradicionalmente hemos llamado de **Procesamiento por Lotes**, (en inglés "*Batch Processing*") con objeto de tratar aquellas fuentes cuya incorporación al sistema se haga de forma puntual o con periodicidades medias o altas, por ejemplo una fuente de una institución estadística que genera sus datos anualmente.

Por último, otra tendencia es la **Arquitectura Lambda** que intenta sacar lo mejor de los métodos de stream computing y batch processing, Parte de un fuente de datos base, en la que se almacena toda la información disponible. Dispone de tres capas: una capa "batch", para grandes cantidades de datos, una capa denominada "speed" (stream computing), para flujos de datos en tiempo real con objetivo de reducir la latencia y entregando los datos en una base de datos NoSQL, y una capa de servidor que recoge las salidas de las otras dos capas y que responde a querys al sistema generando vistas de los datos.

7.3. Procesamiento MapReduce

MapReduce es un modelo de procesamiento de datos, proveniente de un paradigma de la programación denominado "paradigma funcional". Este tipo de soluciones, dise-

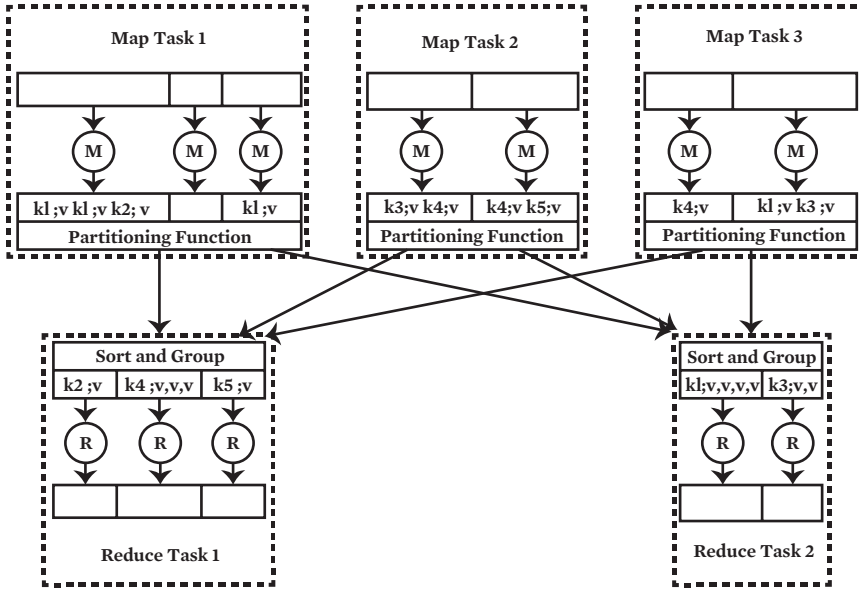
[59] Spark Streaming Programming Guide http://spark.apache.org/docs/latest/streaming-programming-guide.html

ñadas hace decenios y resueltas por los lenguajes de programación funcionales, son hoy especialmente relevantes ya que **existen muchos problemas del mundo real que pueden ser solucionadas aplicando este modelo y específicamente muchos relacionados con Big Data**.

Hasta hace pocos años no se ha dispuesto de una infraestructura de hardware y software que hiciera viable desde un punto de vista técnico y económico el aplicar este tipo de técnicas a cantidades masivas de datos. El gran tiempo de computación necesario hacía inviable la aplicación del paradigma funcional al tratamiento masivo en tiempo real de la información. Los nuevos sistemas distribuidos si permiten este tipo de soluciones mediante paralelización en clusters de ordenadores estándar de precio reducido. Cada fragmento de trabajo en los que es dividida cada aplicación es ejecutado en un nodo del sistema distribuido.

MapReduce consiste en la unión de **dos funciones** de alto nivel: "**Map**" y "**Reduce**". Cada una de estas funciones de alto nivel toman como entrada una lista de pares clave-valor y su propia función, que llamaremos función-map y función-reduce.

Vamos a explicarlo con un ejemplo que usamos todos los días: las búsquedas en Google. Cuando hacemos una búsqueda en su sistema, Google nos presenta los resultados en un orden concreto. Para ello recorre con sus Bots la internet buscando y contabilizando los enlaces que cada página hace a otras páginas, decidiendo así qué página es más relevante y por tanto debe ser presentada antes que otras en las páginas de resultados.



En la figura[60] vemos que se han creado tres nodos Map y dos nodos Reduce. La función de Map consiste en recorrer cada página web, extraer los enlaces (en la figura los llama k1, k2, k3, k4 y k5) y les da un valor (en la figura lo llama "v", digamos para simplificar que es 1). Este sistema es muy habitual en Big Data: se le llama pares de clave-valor (en inglés "*Key-Value*"), de ahí el uso de "K" y de "V".

Se han generado asimismo dos nodos Reduce, cada uno de ellos recogiendo el conteo de diferentes palabras, k2-k4-k5 el de la izquierda y k1-k3 el de la derecha. La función-Reduce consistirá en coger de cada nodo Map las ocurrencias de las palabras que está contando y contabilizarlas y ordenarlas. El resultado final será una lista ordenada como la siguiente:

(k4,3), (k1,3), (k3,2), (k5,1). (k2,1)

Otro ejemplo muy popular para explicar MapReduce es el de la **ordenación de cartas**[61]. En una primera pasada, la función Map reparte en 4 montones las cartas de cada palo. A continuación la función Reduce recoge las cartas de cada palo y las ordena.

En resumen: Map recoge la lista de claves-valor y genera una lista de claves y valores intermedios aplicando la función-definida como función-map. Reduce recoge esta lista de claves y valores intermedios y los combina aplicando la función definida como función-reduce generando una lista de valores definitiva. Tanto la ejecución de la función de alto nivel Map como la de Reduce se realiza a través de una red de nodos trabajando en paralelo en un sistema distribuido Big Data. MapReduce genera una red de nodos, que llamaremos topología, distribuye el trabajo en la red de nodos y recoge el resultado final.

Otros ejemplos de utilización de MapReduce por parte de Google podría ser la generación de una lista de páginas web junto con la lista de URLs que enlazan a cada página, o la determinación de las palabras más relevantes de un conjunto de documentos.

7.4. NoSQL, las Bases de Datos del Big Data

La historia de la informática va unida a los **Sistemas de gestión de Bases de Datos Relacionales (SGBDR)** que posibilitaron la implantación de forma segura de todas las

[60] MapReduce: Simplified Data Processing on Large Clusters, Jeffrey Dean and Sanjay Ghemawat http://research.google.com/archive/mapreduce.html

[61] "Learn MapReduce with Playing Cards" https://www.youtube.com/watch?v=bcjSe0xCHbE Jesse Anderson.

transacciones que hoy en día se gestionan en los sistemas de información, cualquier compra, cualquier venta, cualquier anotación en un banco, altas, bajas, borrados de registros, modificaciones... Sin embargo **resultaron ser demasiado lentos y costosos para las necesidades de los Sistemas Big Data.** Las bases de datos relacionales no se pueden distribuir de forma sencilla sobre varias máquinas y tienen graves problemas de escalabilidad, cuestión no admisible para un sistema Big Data.

El fundamento teórico lo ofrece el **teorema CAP** o teorema de Brewer, que dice que **en sistemas distribuidos no es posible garantizar a la vez:**

- La **Consistencia** (en inglés "*Consistency*"): se recibe la misma información independientemente del nodo que procese la petición.
- La **Disponibilidad** (en inglés "*Availability*"), que todos los clientes puedan leer y escribir aunque algún nodo falle).
- La **Tolerancia a las Particiones** (en inglés "*Partition tolerance*")**:** el sistema funciona aunque falle una partición.

Las bases de datos relacionales cumplen bien las dos primeras, pero **las bases de datos del Big Data necesitan cumplir prioritariamente la Tolerancia a las Particiones.**

Esto empujó a Google, Amazon y otras empresas a construir sus propios sistemas o promover el desarrollo de sistemas open-source que poder utilizar sin pagar un coste de licencias que hacía muy costoso y casi inviable el sistema distribuido.

El almacenamiento distribuido no relacional, o sea, que no sigue el modelo relacional, es uno de los fundamentos del Big Data. Los sistemas de almacenamiento Big Data manejan de forma distribuida diariamente una cantidad de datos del orden de petabytes, en clusters de miles de servidores commodity baratos, con un rendimiento muy alto, cumpliendo el **principio CAE, o sea eficiente en costes con alta disponibilidad y ampliable elásticamente** (en inglés "*Cost-Efficiency*", "*High Availability*", "*Elasticity*").

A las bases de datos Big Data se les agrupa bajo en concepto de Bases de Datos **NoSQL**, que suele traducirse como "Not Only SQL". Como otros paradigmas, el concepto no es nuevo, tiene sus raíces en las bases de datos en red y jerárquicas del último cuarto del siglo XX.

Hay numerosas clasificaciones para las bases de datos NoSQL[62], siendo las más importantes las siguientes:

[62] Aprovechamos la referencia de http://nosql-database.org/ que clasifica más de 150 BD No SQL.

- **Clave - Valor:** a partir de una clave se recupera un objeto binario. DynamoDB, Redis, BerkeleyDB, GenieDB, Voldemort, Oracle NoSQL y Windows Azure NoSQL serían las más destacadas.
- **Big Table/ Columnares:** sistemas que reparten las filas y las columnas de una tabla en diferentes servidores. De este tipo son las muy conocidas Hadoop HBase, Cassandra, Hypertable o Amazon SimpleBD.
- **Documentos**: manejan conjuntos de datos identificados por etiquetas, usado habitualmente para almacenar información de formularios completados por los usuarios. Ejemplos de este tipo serían MongoDB, Elasticsearch, CouchBase o CouchDB. Dentro de este grupo se separan las de **Documentos XML,** bases de datos especializadas en manejar documentos en formato XML como BerkeleyDB XML, EMC Documentum o BaseX.
- **Grafos**: representación de información estructurada en forma de red. Neo4J, InfiniteGraph y OpenLink Virtuoso serían los productos más significativos.

También se puede considerar un quinto tipo de Base de Datos NoSQL, los **Índices "full-text"**, que son textos no estructurados y son la base de los conocidos Apache Lucene y Apache SolR. En el apartado "4.5.1 Tipos de Bases de Datos NoSQL" se abunda y detalla sobre estos cuatro tipos de Bases de Datos NoSQL.

Existen otros grupos, como las bases de datos orientadas a objetos, soluciones grid & cloud, bases de datos multidimensionales, orientadas a eventos o a redes, además de otros muchos tipos orientadas a propósitos específicos.

Como fundamento las bases de datos NoSQL cumplen las propiedades **BASE** que son las siglas de "**Basic Availability",** "**Soft-state**" y "**Eventual consistency**". Estas propiedades significan que después de cada transacción, las bases de datos NoSQL estará en un estado consistente, que se pueden entregar datos obsoletos y que se permiten dar respuestas aproximadas.

Las bases de datos relacionales, en cambio, cumplen las propiedades **ACID** (**Atomicidad, consistencia - integridad, aislamiento y durabilidad - persistencia**), que aseguran que las transacciones no se quedan a medias, que sólo se ejecutan aquellas transacciones que no vayan a afectar la integridad de la base de datos, que en la realización de dos transacciones sobre una misma información una no afecta a las otras y que una transacción realizada no se podrá deshacer.

En la práctica, a la hora de programar interaccionando con bases de datos NoSQL nos vamos a encontrar con un entorno muy rápido y ágil, pero que a veces nos va a generar errores. La cuestión es si ese error es o no relevante. Por ejemplo, puede

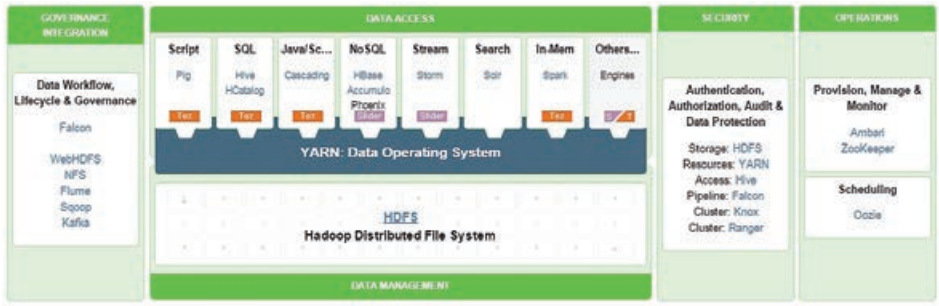
no ser relevante si Facebook nos presenta o no todas las entradas publicadas por un usuario o si Twitter nos presenta todos los tweets de un trending topic concreto o si el número de visitas a un vídeo de Youtube son 14.567 o 15.000 pero sin embargo sí lo sería si estuviéramos hablando de las transacciones realizadas contra una plataforma de e-commerce o en una aplicación de banca online, para lo que usaremos si o si una base de datos relacional.

También es muy relevante que con NoSQL no se pueden hacer JOINs de dos tablas, es decir, que no podemos cruzar dos tablas buscando los denominadores comunes entre ellas. Si nos es necesario, tendremos que usar bases de datos relacionales. Además, los lenguajes de consulta a bases de datos NoSQL son recientes, mientras que SQL es un lenguaje altamente consolidado.

7.5. Apache Hadoop

Hadoop[63] es el grupo de proyectos que ha catalizado el despegue del Big Data en el mundo. Es usado de una manera u otra por grandes empresas que tienen en su core técnico el Big Data y muchas funcionalidades del mundo de la Vigilancia. Ejemplos destacados son Twitter, Facebook, LinkedIN, Reddit o Amazon.

Su origen es el GFS (Google File System), del que toma mucha de sus ideas para construir el primero de sus proyectos, el HDFS, un Sistema de Ficheros Distribuido.



64

Se organiza en 4 módulos:

- **Common**, que da soporte al resto de módulos.

63 Proyecto Hadoop en Apache https://hadoop.apache.org/

64 Arquitectura Hadoop de HortonWorks

- **YARN**, para planificación de tareas y gestión de cluster.
- **MapReduce**, para el procesamiento en paralelo de trabajos que implican grandes conjuntos de datos.
- **HDFS**: como decíamos, el Sistema de Ficheros Distribuido.

Actualmente cuenta con un conjunto de proyectos vinculados, entre los que destacamos los siguientes:

- **Spark y Storm**: motores avanzados de procesamiento de datos.
- **HBase**: la base de datos distribuida de Hadoop.
- **Mahout**: para hacer Machine Learning.
- **Pig**: es un plataforma para crear programas MapReduce en plataformas Hadoop. En esta plataforma se utiliza un lenguaje denominada Pig Latin. Desarrollado originalmente por Yahoo.
- **Hive**: proporciona a Hadoop la infraestructura equivalente a la de un dataware-house, habilitando la consulta y el análisis de los datos. Desarrollado originalmente por Facebook.
- **Sqoop**: aplicación para transferir datos entre bases de datos relacionales y Hadoop.
- **Zookeeper**: proporciona servicio de configuración centralizada, sincronización y registro de nombres.
- **Flume**: servicio para recolectar, agregar y mover grandes conjuntos de datos hacia un entorno Hadoop.
- **Kafka**: sistema de mensajería distribuido.

Otros paquetes, que no suelen incluirse en Hadoop pero que si se están usando en proyectos Big Data son:

- **OpenNLP**: para procesamiento de lenguaje natural, basado en machine learning.
- **UIMA:** utilizado por IBM Watson, es un framework para integrar aplicaciones NLP.
- **Apache SolR:** buscador avanzado, basado en Apache Lucene.

Las Distribuciones de Apache Hadoop

De la misma manera que de Linux surgieron diferentes distribuciones[65] hasta el punto de que hoy en día millones de personas poseen un ordenador Linux, concretamente todos los que poseen un teléfono móvil con la distribución de Linux "Android", una situación similar ha ocurrido con Hadoop. Estas distribuciones incluyen muchos de las aplicaciones que mencionamos en el punto anterior.

En la página web de Apache podemos encontrar una lista completa de productos[66] que incluyen Apache Hadoop, aplicaciones derivadas y soporte comercial.

Entre ellas destacaremos las siguientes:

- **Amazon**: la muy popular Amazon Elastic MapReduce
- **Cloudera**: denominada CDH y complementada con productos propios.
- **Hortonworks**: ofrecen una plataforma 100% open-source y basan la sostenibilidad de su empresa en los servicios. Ofrecen sus servicios como partners de otras empresas destacadas del sector, como SAS.
- **IBM**: denominada IBM InfoSphere BigInsights
- **MapR Technologies**: distribución orientada al alto rendimiento, han realizado reingeniería de algunos de sus componentes.
- **Pivotal HD**: ofrece una distribución que es ofrecida por otras empresas relevantes del sector, como EMC, adquirida en 2015 por DELL.

Otros proyectos y empresas relevantes del entorno Big Data como BigTop, Mahout, o Nutch también son mencionados en esta lista, que presumiblemente irán creciendo y desarrollando en un futuro cercano.

8. Ontologías, Datos Enlazados (Linked Data) y Web Semántica

Este apartado presenta una descripción de los conceptos Ontología y Web Semántica. En primer lugar, en el apartado de "Ontologías", se proporciona la definición de onto-

[65] Wikipedia, "Distribuciones Linux" http://es.wikipedia.org/wiki/Anexo:Distribuciones_Linux

[66] Distribuciones de Apache Hadoop en la página web de Apache http://wiki.apache.org/hadoop/Distributions_and_Commercial_Support

logía más ampliamente utilizada, los tipos de ontologías existentes, los componentes básicos, y se describen algunas ontologías de ejemplo. A continuación, en el apartado "Datos Enlazados y la Web Semántica", se presenta la noción de Datos Enlazados y Web Semántica. Seguidamente, en el apartado "Lenguajes de representación de ontologías y tecnologías", se describen los lenguajes de representación de ontologías más utilizados en el ámbito de la Web Semántica. Finalmente, el apartado "Sistemas de Organización del Conocimiento (KOS) y SKOS" introduce los sistemas de organización de conocimiento y su relación con la Web Semántica.

8.1. Ontologías

El término "Ontología"[67] se deriva del griego: *ontos* (ser) y *logos* (hablar de). Por consiguiente, en el campo de la filosofía, Ontología es la ciencia que tiene al ser como su objeto de trabajo. Es decir, es la rama de la metafísica que trata de proporcionar una explicación sistemática de la existencia. Sin embargo, en áreas de ingeniería, el término "ontología" ha sido adoptado para describir parcelas de conocimiento que pueden ser (computacionalmente) representadas en un programa[68]. Por consiguiente, una ontología se considera una entidad computacional, es decir un recurso artificial que se desarrolla[69].

Una **ontología** es una especificación formal y explicita de una conceptualización compartida, donde la semántica de la información se representa mediante objetos, relaciones y propiedades que los caracterizan, en un lenguaje que sea comprensible para los ordenadores, es decir, un lenguaje formal. Por tanto, una ontología es un ***modelo de conocimientos consensuado*** en un determinado dominio y que es reutilizable en diferentes aplicaciones.

Es importante mencionar que las ontologías definen conocimiento consensuado y compartido sobre un dominio, para que dicho conocimiento pueda comunicarse entre personas y sistemas computacionales. Es decir, las ontologías permiten el intercambio y la reutilización de conocimiento de forma computacional.

Resumiendo, (a) las ontologías son un formalismo de representación de algún tipo de conocimiento, que debe estar consensuado, de un dominio o ámbito, y (b) las ontologías deben especificarse mediante un lenguaje formal. Para definir el vocabulario de un dominio se utilizan los conceptos y las relaciones entre dichos conceptos, así

---

67 Munn y Smith (2008) Sobre el término "Ontología", Diccionario Oxford

68 Studer et alter (1988) "Knowledge Engineering: Principles and Methods."

69 Mahesh (1996) Ontology Development for Machine Translation: Ideology and Methodology

como los axiomas y las reglas que combinan conceptos y relaciones que amplían las definiciones dadas en el vocabulario.

Tradicionalmente, en la comunidad ontológica se han distinguido dos tipos de ontologías: (a) **ontologías ligeras** (*lightweight*) que son principalmente **taxonomías** y (b) **ontologías pesadas** (*heavyweight*) que son una extensión de las ontologías ligeras, a las que **se les añaden axiomas y restricciones**. Sin embargo, como consecuencia del desarrollo colaborativo y distribuido de ontologías se ha evolucionado a una nueva clasificación de las ontologías basada en las relaciones que se pueden establecer entre ellas. De acuerdo a este nuevo enfoque se distinguen los siguientes tipos de ontologías[70]:

- Una **ontología individual** es una ontología que no tiene ningún tipo de relación (dependiente o independiente del dominio) con otras ontologías.
- Un **conjunto de ontologías individuales interconectadas** incluye un conjunto de ontologías que tienen algún tipo de relación ad-hoc dependiente del dominio entre ellas.
- Una **red de ontologías** es una colección de ontologías individuales interconectadas relacionadas mediante una variedad de meta-relaciones[71]. Algunos ejemplos de estas meta-relaciones son: (a) *hasPriorVersion* (si la ontología es una nueva versión de otra ontología ya existente), (b) *useImports* (si la ontología importa otra ontología), (c) *isExtension* (si la ontología extiende otra ya existente), (d) *containsModules* (si la ontología se compone de varios módulos), y (e) *hasMapping* (si alguno de los componentes de la ontología tiene relación de correspondencia con otras ontologías existentes).

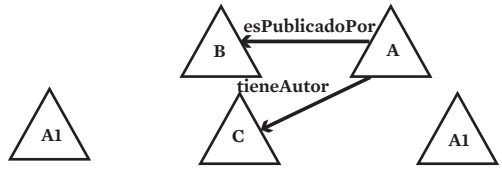
La Figura 1 (a) muestra una ontología individual (Ontología A1); la Figura 1 (b) presenta N ontologías individuales relacionadas entre sí mediante relaciones dependientes del dominio entre conceptos incluidos en dichas ontologías, por tanto, esta figura presenta un conjunto de ontologías individuales interconectadas; finalmente la Figura 1 (c) muestra la red de ontologías asociada al conjunto de ontologías individuales presentado en la Figura 1 (b). En esta red de ontologías, las meta-relaciones ("*useImports*" y "*hasPriorVersion*") se han expresado explícitamente entre las distintas ontologías.

[70] Suárez-Figueroa, (2010) NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse

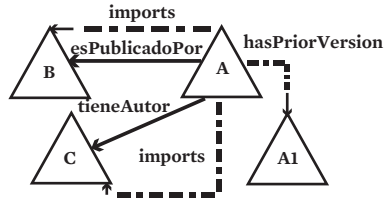
[71] Haase et alter (2007) Networked Ontology Model. NeOn Project

FIGURA 1

Ejemplo de ontologías individuales, ontologías individuales interconectadas y redes de ontologías[72]



(a) Ontología individual (b) Conjunto de ontologías individuales interconectadas



(c) Red de ontologías

Los componentes principales de una ontología son los **conceptos** (p. ej., Hombre, Mujer, Tren, Edificio, Universidad, etc.), normalmente organizadas en **taxonomías** o **jerarquías** (p.ej., Mujer es un tipo de Persona, Tren es un tipo de Vehículo) y las **propiedades** (p.ej., las personas se casan, una persona vive en un edificio, una persona tiene un nombre, una universidad tiene edificios, etc.).

Para que las ontologías sean algo más que una taxonomía codificada en un lenguaje formal, deben incluir los llamados **Axiomas**. Éstos consisten en **especificaciones**, siempre en base lógica, **de las propiedades y de las relaciones entre los componentes de la ontología**. Por ejemplo, un país sólo puede tener una capital, no se puede viajar en tren de Europa a EEUU, o dos personas que tienen los mismos padres son hermanos.

Las ontologías se utilizan en aplicaciones informáticas que van a ser utilizadas por seres humanos y también en aplicaciones que permiten el intercambio de datos máquina-máquina sin intervención humana. Aunque últimamente se conoce e identifica a las ontologías por importante papel en la web semántica, no sólo se aplican en el ámbito web sino también en procesos locales de integración de datos de distintas fuentes y en la gestión del conocimiento entre otros.

[72] Suárez-Figueroa, (2010) NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse

Ejemplos de sistemas que utilizan ontologías son: **iniciativas de publicación y enlazado de datos, anotación de contenidos, interoperabilidad de datos entre sistemas conciliando distintos modelos de datos, recuperación de información, búsquedas avanzadas, procesamiento del lenguaje, análisis de sentimientos, vigilancia tecnológica, razonamiento inductivo, clasificación, y diferentes técnicas de resolución de problemas**.

Ejemplos de áreas de aplicación relevantes son: **gobierno abierto, medicina, gestión de conocimientos, recursos humanos, patrimonio cultural, geografía, vigilancia tecnológica**.

La reutilización de conocimiento permite ahorrar tiempo y recursos y proporcionar una forma común para compartir datos. Por ese motivo, el desarrollo de ontologías debe estar basado en la reutilización de ontologías existentes. En este sentido, existen diferentes herramientas (buscadores, registros y repositorios) que facilitan la búsqueda y localización de ontologías. Merece la pena mencionar los buscadores **Watson**[73] y **Swoogle**[74]. Con respecto a los registros, **Protégé Ontology Library**[75]**,** es un registro general que no dispone de funcionalidad de búsqueda, mientras que **BioPortal**[76] es un registro específico dirigido a la comunidad biomédica que proporciona un espacio en el que compartir y descubrir ontologías. Existen otros registros dirigidos a comunidades específicas como es el caso de las ciudades inteligentes para el que existe un catálogo de ontologías[77] útiles para dicho dominio. Finalmente, ***Linked Open Vocabularies***[78] (LOV) es un repositorio de ontologías, considerado como un observatorio del ecosistema de vocabularios (principalmente usados en datos enlazados). LOV incorpora información sobre interconexión y dependencias entre vocabularios, historia de versiones y políticas de mantenimiento. Los resultados de una búsqueda de ontologías en este repositorio aparecen ordenados en función de la popularidad de los términos en los conjuntos de datos enlazados y en el ecosistema de LOV.

A continuación se describen en detalle los principales componentes de las ontologías así como ejemplos de ontologías en diversos dominios (persona, organización, comercio electrónico, etc.).

---

[73] Watson http://watson.kmi.open.ac.uk/WatsonWUI/ A Gateway for the Semantic Web

[74] Swoogle http://swoogle.umbc.edu/)

[75] Protégé Ontology Library http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library

[76] BioPortal http://bioportal.bioontology.org/

[77] Catálogo de ontologías útiles en el contexto de las ciudades inteligentes http://smartcity.linkeddata.es/

[78] Linked Open Vocabularies (LOV) http://lov.okfn.org/

8.1.1. Principales componentes

Los principales componentes de las ontologías son las clases, las relaciones y los axiomas formales. Además, los modelos representados en las ontologías se pueden completar con datos específicos (individuos concretos), que se denominan instancias.

- Las **Clases representan conceptos**, en un sentido amplio. Por ejemplo, en el ámbito de la eficiencia energética en edificios, los conceptos pueden ser 'Edificio', 'Puerta', 'Ventana', 'Dispositivo', 'Sensor', etc.; y en el ámbito universitario se pueden tener clases como 'Universidad', 'Curso', 'Asignatura', 'Profesor', etc. Las clases de la ontología **se organizan generalmente en taxonomías o jerarquías** a través de las cuales se pueden aplicar mecanismos de herencia. Siguiendo con los ejemplos anteriores, en el dominio de la eficiencia energética en edificios, se puede representar una taxonomía de sensores ('Sensor de Barrido', 'Sensor Óptico', 'Sensor Táctil', etc.) o de diferentes tipos de puertas en los edificios ('Puerta Interior', 'Puerta Exterior', 'Puerta Corredera', 'Puerta Giratoria', etc.).
- Las **Relaciones representan un tipo de asociación entre conceptos del dominio.** Las **relaciones más habituales son relaciones binarias** (es decir, relaciones que involucran dos elementos). Por ejemplo, en el dominio de eficiencia energética en edificios, algunas relaciones podrían ser 'estar localizado en' o 'contiene sensor'; y en el ámbito universitario se pueden tener relaciones como 'imparte asignatura', 'asiste a curso' o 'examina'. En las relaciones binarias se suele hablar de **dominio y rango de la relación** (siendo el primer elemento de la relación el dominio, y el segundo elemento de la relación el rango). Por ejemplo, la relación binaria 'estar localizado en' puede tener como dominio el concepto 'Edificio' y como rango el concepto 'Lugar'; también podría tener como dominio el concepto 'Dispositivo'. Cuando las relaciones binarias tienen como rango un tipo de datos (p. ej., cadena de caracteres, números, etc.), éstas se denominan **atributos**. Por ejemplo, una asignatura tiene un número de créditos, los españoles tienen un DNI, o los cursos tienen un nombre.

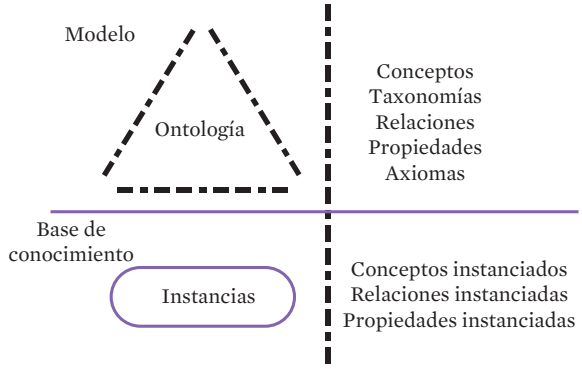
  Es importante mencionar que los ejemplos de relaciones dados se refieren a las llamadas relaciones ad-hoc (o de dominio). Sin embargo, también existen otro tipo de relaciones, que permiten la creación de jerarquías de conceptos (relación 'subclaseDe', que básicamente indica una clasificación). De esta manera, si dos conceptos están relacionados mediante una relación 'subclaseDe' se establece que los individuos de la clase hija son a su vez individuos de la clase padre (también se puede decir que la clase hija es un tipo de la clase padre). Por ejemplo, si se establece que 'sistema de detección de intrusos basado en red' es subclase de (o es un tipo de) 'sistema de detección de intrusos' se puede inferir que todos los individuos declarados como 'sistema de detección de intrusos basado en red' son instancias de 'sistema

de detección de intrusos'. Es importante no confundir este tipo de relaciones con las relaciones '**parteDe'**, que permiten representar la relación de **meronimia**. Por ejemplo, la clase 'Sensor' se podría relacionar con la clase 'Sistema de detección de intrusos' mediante una relación 'parteDe' (un sensor es uno de los elementos que componen los sistemas de detección de intrusos).

- Los **axiomas formales**[79] sirven para **modelar sentencias que son siempre verdad**. Normalmente se utilizan para representar el conocimiento que no se puede definir formalmente usando el resto de componentes. Los axiomas formales se utilizan para verificar la consistencia de la ontología o la consistencia del conocimiento almacenado en una base de conocimientos. Además, se suelen utilizar para inferir nuevo conocimiento. Un axioma en el dominio de eficiencia energética en edificios puede ser "*no es posible construir un edificio público sin una puerta contra incendios*" (basado en cuestiones legales).
- Las **instancias** se utilizan para **representar elementos o individuos particulares en una ontología**. Por ejemplo, en el dominio de eficiencia energética en edificios, algunas instancias de Edificio podrían ser Edificio Picasso o Edificio España; y en el ámbito universitario la Universidad Politécnica de Madrid y la Universidad Carlos III serían instancias del concepto Universidad.

Cuando se habla de ontologías, es habitual realizar la división entre el modelo, que representa el conocimiento del dominio en general, y la base de conocimiento, que incluye los individuos particulares que siguen el modelo. La Figura 2 muestra dicha división así como los componentes que conforman cada parte.

FIGURA 2
Modelo (TBox) y base de conocimiento (ABox)



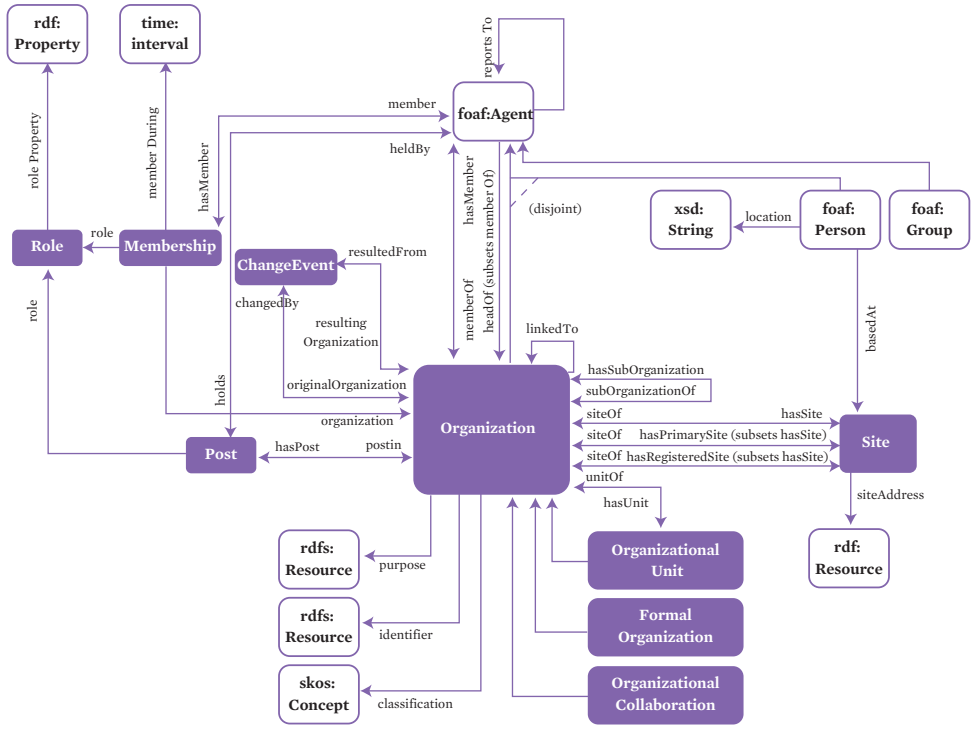
[79] Gruber (1993) A translation approach to portable ontology specifications

Por último merece la pena comentar que si las ontologías se desarrollan usando el formalismo de Lógica Descriptiva, la distinción mostrada en la Figura 2 se realiza entre la llamada TBox y la ABox. La TBox contiene la terminología referente al conocimiento del dominio, mientras que la ABox contiene el conocimiento extendido, es decir las definiciones de los individuos.

8.1.2. Ontologías públicas

Las ontologías no sólo permiten anotar semánticamente datos aportándoles significado si no también compartir conceptualizaciones en un dominio dado, por ejemplo con el objetivo de reconciliar distintas fuentes de datos. En este sentido las ontologías deben ser compartidas entre varias partes e idealmente deberían estar disponibles en la web bajo licencias abiertas. En este escenario, una ontología debe ser recuperable en la web a través de su URI ofreciendo mecanismo para acceder a su código fuente, en formato legible para las máquinas y posiblemente en distintos formatos, y a la documentación asociada a la ontología en formato legible para los humanos. A continuación se muestran algunos ejemplos de ontologías disponibles en la web organizados por dominios:

- **Personas y Organizaciones:**
  - **Friend of a friend Ontology** (prefijo: foaf; URI: http://xmlns.com/foaf/0.1/): es una ontología ampliamente aceptada por la comunidad de la web semántica que permite describir personas y grupos sociales incluyendo información básica sobre los mismos así como sus relaciones con otras personas y proyectos.
  - **vCard Ontology** (prefijo: vcard; URI: http://www.w3.org/2006/vcard/ns#): es una ontología que permite describir personas y organizaciones usando la información que normalmente se detalla en las tarjetas de visita.
  - **Organization Ontology** (prefijo: org; URI: http://www.w3.org/ns/org#) es una recomendación del W3C desde enero de 2014 que permite la publicación de información sobre organizaciones y estructuras organizacionales como son las gubernamentales. Esta ontología proporciona un vocabulario genérico y reutilizable que puede ser extendido o especializado para situaciones particulares.



- **Documentos:**
  - **Dublin Core** (prefijo: dc; URI: http://purl.org/dc/terms/): es un ontología para describir recursos mediante metadatos simples y generales (DC *terms*).
  - **Bibliographic Ontology** (prefijo: bibo; URI: http://purl.org/ontology/bibo/): es una ontología para describir datos bibliográficos en la Web. Se puede usar como ontología de citas y para clasificación de documentos.
  - **Metadata Authority Description Schema** (prefijo: mads; URI: http://www.loc.gov/mads/rdf/v1) es un vocabulario para describir datos utilizados en el contexto de las bibliotecas y la comunidad de la sociedad de la información, como museos, archivos y otras instituciones culturales. Este vocabulario permite la descripción de personas, corporaciones, geografía, etc

- **Información geográfica:**
  - **WGS84 Geo Positioning** (prefijo: geo; URI: http://www.w3.org/2003/01/geo/wgs84_pos) es una ontología para representar puntos geográficos (longitud y latitud).

- **Estadísticas:**
  - **RDF Data Cube Vocabulary** (prefijo: qb; URI: http://purl.org/linked-data/cube#; http://www.w3.org/TR/vocab-data-cube/): es una ontología para la publicación de datos multidimensionales (como estadísticas) en la web.

- **Otros vocabularios:**
  - **Semantic Sensor Network Ontology** (prefijo: ssn URI: http://purl.oclc.org/NET/ssnx/ssn) ha sido desarrollada en el contexto del W3C. Esta ontología se ha convertido en el estándar de facto para representar redes de sensores, observaciones y conceptos relacionados como dispositivos, plataforma, capacidades de medidas como precisión y etc.
  - **PROV Ontology** (prefijo: prov; URI: http://www.w3.org/ns/prov#): es una ontología que permite representar e intercambiar información sobre procedencia (*provenance*) generada por diversos sistemas y en contextos diferentes.
  - **Vocabulary of Interlinked Datasets** (prefijo: void; URI: http://rdfs.org/ns/void#; http://www.w3.org/TR/void/): es una ontología para describir conjuntos de datos RDF. Facilita el descubrimiento y la utilización de conjuntos de datos RDF por parte de usuarios potenciales (agregadores, indexadores, desarrolladores de aplicaciones).
  - **Data Catalog Vocabulary** (prefijo: dcat; URI: http://www.w3.org/ns/dcat#; http://www.w3.org/TR/vocab-dcat/): es una ontología para facilitar la interoperabilidad de catálogos de datos publicados en la Web. Facilita el descubrimiento de catálogos de datos y el consumo de metadatos procedentes de distintos catálogos.
  - **Asset Description Metadata Schema** (prefijo: adms; URI: http://www.w3.org/ns/adms): es una ontología que describe conceptos relacionados con la interoperabilidad semántica de recursos.
  - **Public Contracts Ontology** (prefijo: pc; URI: http://purl.org/procurement/public-contracts#): es una ontología que sirve para expresar en RDF datos estructurados sobre los contratos públicos. Permite el intercambio fluido de datos anotados semánticamente a través de fuentes de datos distribuidas.
  - **GoodRelations** (prefijo: gr; URI: http://purl.org/goodrelations/v1): es una ontología que proporciona el vocabulario para anotar ofertas de comercio electrónico. En este vocabulario se definen y relacionan conceptos como producto, precio, oferta, venta, alquiler, método de pago, licencia, entre otros. Esta ontología ha sido ampliamente aceptada convirtiéndose en el estándar de facto para publicar ofertas de comercio electrónico.

- **Ontology for Media Resources** 1.0 (prefijo: ma; URI http://www.w3.org/ns/ma-ont#): es una ontología que permite reducir las diferentes descripciones de recursos de medios de comunicación (*media resources*) y que proporciona un conjunto básico de propiedades descriptivas. Esta ontología define las asignaciones entre su conjunto de propiedades y elementos de los formatos de metadatos más utilizados para describir recursos multimedia.

8.1.3. Herramientas de desarrollo de ontologías

En este apartado se describen brevemente las características de las herramientas de libre distribución de desarrollo de ontologías más importantes.

La herramienta de desarrollo de ontologías **Protégé** ha sido desarrollada en el Stanford Medical Informatics (SMI) de la Universidad de Stanford. Esta aplicación autónoma de código abierto consiste en una arquitectura cuyo núcleo es el editor se puede ampliar mediante las extensiones disponibles que le aportan mayor funcionalidad al entorno de desarrollo. Las extensiones o plug-ins disponibles aportan funcionales de importación/exportación de ontologías, visualización, razonamiento, etc. La plataforma de Protégé da soporte estable principalmente a dos formas/paradigmas de modelado de ontologías:

- El editor **Protége-Frames** da soporte al desarrollo y población de ontologías construidas mediante el paradigma de representación basado en marcos de acuerdo con el Open Knowledge Base Connectivity protocol (OKBC[80]). En las últimas versiones de este editor se han incorporado funcionalidades para dar soporte a los lenguajes OWL 1.0 y RDF(S). A parte de la conexión con distintos razonadores para la realización de inferencias, en Protégé-Frames se da soporte a la realización de consultas en SPARQL, ejecución de reglas en SWRL. Además, esta herramienta permite la edición multiusuario a través de una versión cliente-servidor.
- El editor **Protégé-OWL** da soporte al desarrollo de ontología para la Web Semántica, concretamente en OWL 2, el lenguaje de implementación de ontologías recomendado por el W3C[81]. Actualmente esta herramienta no permite realizar consultas en SPARQL pero si permite mediante un editor básico ejecutar reglas en SWRL. Además da soporte a tareas de razonamiento mediante distintos razonadores como FaCT++, Pellet entre otros.

---

[80] Open Knowledge Base Connectivity protocol (OKBC): A Programmatic Foundation for Knowledge Base Interoperability. Chaudri, Farquhar, Fikes, Karp, Rice. Proceedings of AAAI-98, July 26-30, Madison, WI.

[81] W3Consortium http://www.w3.org

Además existe una versión de Protégé llamada **WebProtege**[82] que consiste en un editor de ontologías de código abierto cuyo principal objetivo es dar soporte al desarrollo colaborativo de ontologías en entornos web.

TopBraid Composer es una herramienta desarrollada por la compañía TopQuadrant. Esta herramienta está basada en el lenguaje OWL e incorpora funcionalidades[83] como crear y ejecutar consultas en SPARQL. Además, permite definir reglas y restricciones usando SPIN[84] (SPARQL Inference Notation).

NeOn toolkit[85] es un entorno de desarrollo multi-plataforma de libre distribución que proporciona un soporte integral para el desarrollo de redes de ontologías a lo largo de su ciclo de vida. NeOn toolkit se basa en la plataforma de desarrollo Eclipse y está formado por un núcleo, que contiene funcionalidades básicas en el desarrollo de ontologías, y más de treinta plug-ins que extienden dichas funcionalidades básicas. Estos plug-ins dan soporte a numerosas actividades relacionadas con el desarrollo de redes de ontologías como anotación, documentación, adquisición del conocimiento, reutilización de recursos ontológicos, gestión, modularización, evaluación, matching y razonamiento entre otras.

8.2. Datos Enlazados y la Web Semántica

8.2.1. Que son los Datos Enlazados

El objetivo de la iniciativa de **Datos Enlazados** es crear una Web de datos con relaciones explícitas y semánticas entre los mismos utilizando el lenguaje RDF para representar tanto los datos como las conexiones entre ellos. En esta nueva concepción de la Web, se pasa de una Web basada en documentos, en la que el usuario es el destinatario de la información publicada, a una Web de datos enlazados en la que programas software pueden publicar, navegar, interpretar, visualizar y utilizar estos datos enlazados de forma automatizada mediante ontologías.

El mecanismo básico de los datos enlazados es el siguiente:

1. Asignar un identificador único (**identificador o URI**) en la Web para cada uno los recursos (organizaciones, personas, productos, edificios, etc.). Estos identificadores

---

[82] WebProtege http://webprotege.stanford.edu/

[83] TopBraid Composer – funcionalidades http://www.topquadrant.com/tools/modeling-topbraid-composer-standard-edition

[84] SPIN (SPARQL Inference Notation). http://spinrdf.org/

[85] NeOn toolkit http://neon-toolkit.org/

únicos reciben el nombre de URI[86] (*Uniform Resource Identifier*) se refieren a los identificadores de los conceptos, relaciones o individuos que se generen o reutilicen durante el trabajo realizado.

2. Modelar los datos usando una **ontología**.
3. Utilizar el lenguaje **RDF** para describir los **datos.**
4. Potencialmente, **enlazar los recursos** con otros recursos en la Web utilizando los identificadores del punto 1. Dichos enlaces permiten enriquecer los recursos con información contextual (por ejemplo de otras clasificaciones de productos transformadas, de plataformas de contratación, etc.).

Esta generación y publicación de datos enlazados ha de realizarse de acuerdo a los principios de publicación descritos en[87] es decir[88]:

1. Utilizar URIs para nombrar cosas.
2. Utilizar URIs HTTP para que se puedan consultar/buscar los nombres asignados.
3. Cuando alguien acceda a una URI, proporcionar información útil utilizando tecnologías estándar (RDF, SPARQL).
4. Incluir enlaces a otros URIs permitiendo que se descubra nueva información.

A modo de ejemplo, podemos imaginar varios conjuntos de datos relacionados que proporcionan información sobre Cervantes sin definir, mediante ontologías, a lo que el término se refiere a Cervantes. De ahí Cervantes puede ser un bar de tapas, una autoridad en una biblioteca, el nombre de una calle, el nombre de un municipio, etc. Si estábamos usando URIs para referirnos a todos estos significados de Cervantes y realizamos una búsqueda simple, posiblemente basada en palabra clave en lugar de por tipo de entidad, en la que se pregunte por las propiedades de Cervantes, se podría tener como respuesta uno o varios números de teléfono, coordenadas GPS, la longitud de la calle, una fecha de nacimiento, el número de habitantes de una ciudad, o el famoso libro Don Quijote de la Mancha, entre otros. Mediante la anotación semántica se ayudaría a evitar esta confusión, definiendo el bar de tapas como un bar, el autor como escritor español, la calle como parte de la infraestructura de una ciudad, y el municipio como unidad territorial.

---

[86] URIs http://www.ietf.org/rfc/rfc1630.txt

[87] Berners Lee (2006) http://www.w3.org/DesignIssues/LinkedData.html

[88] Traducción de http://www.w3.org/DesignIssues/LinkedData.html

8.2.2. La Web Semántica

La **Web semántica** es una extensión de la Web en la que el significado (semántica) de la información y de los servicios está definido de acuerdo a modelos de datos consensuados que reciben el nombre de **ontologías.** Las ontologías se implementan en lenguajes específicos propuestos por el Consorcio de la World Wide Web, siendo los más importantes RDF para representar datos, RDF (S) y OWL para representar ontologías, y SPARQL para realizar consultas a los datos y las ontologías.

La cantidad de datos semánticos publicados en la Web ha experimentado un enorme crecimiento en los últimos años, principalmente impulsado por iniciativa conocida como ***Linked Data***. El concepto de datos enlazados nació con la idea de usar la Web para "conectar datos" y está transformando la Web en una "base de datos global" en la cual **los datos están conectados con otros datos y descritos mediante ontologías**.

Como se observa en los puntos anteriores hay una cuestión fundamental: por un lado está el **modelo de datos** (ontología) y por otro **los datos** (recursos, metadatos, etc.) que se describen con la ontología y se publican en la Web como datos enlazados. Para poder beneficiarse de los ventajas ofrecidas por los datos enlazados, la publicación de datos debe hacerse y ser dirigida en base a una ontología validada, correcta y robusta, desarrollada siguiendo un proceso metodológico.

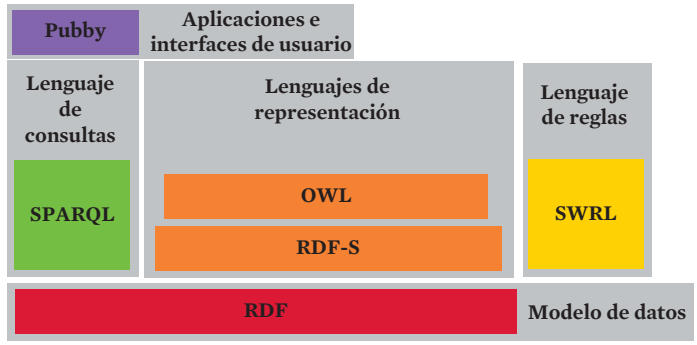
En los últimos años han sido numerosas las instituciones que se han sumado a esta tendencia. Gobiernos de numerosos países e instituciones europeas, medios de comunicación, en el ámbito académico, medio ambiente, salud, bibliotecas, museos, etc., han comenzado a publicar datos y recursos en la Web utilizando ontologías y los estándares mencionados arriba, para facilitar la **integración de datos**, la **interconexión** y **enriquecimiento** de datos con fuentes diversas y el **desarrollo de aplicaciones innovadoras.** La lista de recursos ya disponibles como datos enlazados es cada vez más numerosa y las aplicaciones que se construyen sobre ellos también.

8.3. Lenguajes de representación de ontologías y tecnologías

Debido a la diversidad de paradigmas de representación del conocimiento existe una variedad de lenguajes de implementación de ontologías con distintos niveles de expresividad y de mecanismos de inferencia. En esta sección nos centramos en los lenguajes de ontologías enfocados al desarrollo de la web semántica, basados en lógicas descriptivas.

También se describen a continuación, las distintas tecnologías utilizadas para el desarrollo y publicación de datos enlazados. Estas tecnologías se han desarrollado en el contexto del "*W3C Data Activity*"[89]. Como se muestra en la Figura 3 la tecnología base sobre la que se asienta es el modelo de datos de RDF. Sobre dicho modelo de datos se apoyan los lenguajes de representación RDF-S y OWL. El lenguaje de consultas utilizado es SPARQL. Finalmente Pubby ofrece capacidades de visualización de los datos para navegadores web.

FIGURA 3
Pila de tecnologías



RDF[90] **(Resource Description Framework)** es un modelo de datos estándar para la descripción de recursos en la web. La estructura básica de información se denomina "tripleta" y se compone de tres elementos "sujeto", "predicado" y "objeto". Este modelo de datos es la base de los desarrollos e datos enlazados, es decir la tecnología utilizada para representar los datos, las clasificaciones y las relaciones entre las mismas. De tal manera la publicación de los datos se realiza en RDF a través de un SPARQL endpoint.

OWL[91] **(Web Ontology Language)** es un lenguaje de implementación de ontologías diseñado para representar y compartir ontologías en la Web basado en RDF. Este lenguaje extiende las capacidades de otros lenguajes como RDF y **RDF-S** incorporando un vocabulario adicional junto con una semántica formal que permite realizar inferencia. OWL permite además de definir clases, jerarquías y relaciones, añadir características a las relaciones como transitividad o simetría, crear axiomas utilizando los operadores universal y existencial, añadir cardinalidades, uniones, etc.

89 W3C Data Activity http://www.w3.org/2013/data/

90 RDF (Resource Description Framework) http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/

91 OWL – Web Ontology Language http://www.w3.org/TR/owl-ref/

SWRL[92] **(Semantic Web Rule Language)** es un lenguaje de reglas basado en la combinación de los sublenguajes OWL DL y OWL Lite con sublenguajes de Unary/Binary Datalog RuleML y el lenguaje de marcado de reglas[93]. Las reglas siguen la forma de implicaciones, de manera que consisten en un antecedente y un consecuente. Este lenguaje permite desarrollar reglas entendidas de la siguiente manera: **en caso de que las condiciones especificadas en el antecedente se cumplan, entonces las condiciones especificadas en el consecuente se deben cumplir también**.

SPARQL[94] **(SPARQL Protocol and RDF Query Language)** es un lenguaje de consultas para fuentes de datos almacenados como RDF establecido como recomendación oficial del W3C. Este lenguaje ofrece la opción de realizar encaje de patrones obligatorios u opcionales así como sus conjunciones y disyunciones. Además, SPARQL permite la aplicación de restricciones del ámbito de las consultas indicando los grafos sobre los que se opera. Los resultados de las consultas SPARQL pueden ser conjuntos de resultados o grafos RDF.

Pubby[95] se puede definir como una interfaz web para SPARQL endpoints cuyas principal característica es proporcionar un interfaz HTML para navegadores convencionales que permite navegar sobre datos enlazados representados en RDF. Así mismo Pubby se encarga de la resolución de los URIs de los recursos almacenados en los endpoints y permite soporte para más de un endpoint en una misma instalación. Además, permite la posibilidad de añadir metadatos a los datos proporcionados.

8.4. Sistemas de Organización del Conocimiento (KOS) y SKOS

El vocabulario **Simple Knowledge Organization System** (prefijo: skos; URI: http://www.w3.org/2004/02/skos/core) es una recomendación del W3C desde agosto del 2009 ampliamente aceptado por la comunidad de la web semántica que define un modelo para compartir y enlazar sistemas de organización de conocimiento. Este vocabulario está formalizado en OWL y define principalmente esquemas de conceptos y conceptos. Debido a la importancia de este vocabulario como pieza clave del desarrollo de este trabajo a continuación se detallan sus características.

Se entiende por **sistemas de organización del conocimiento** (KOS del inglés *Knowledge Organization Systems*) el conjunto de esquemas para organizar la información y

[92] SWRL – Semantic Web Rule Language http://www.w3.org/Submission/SWRL/

[93] Lenguaje RuleML http://wiki.ruleml.org/index.php/RuleML_Home

[94] SPARQL protocol and RDF Query Language http://www.w3.org/TR/rdf-sparql-query/

[95] Pubby, Interfaz web para SPARQL http://wifo5-03.informatik.uni-mannheim.de/pubby/

facilitar la gestión del conocimiento. Este tipo de esquemas incluye clasificaciones, tesauros, glosarios, diccionarios, esquemas de clasificación, etc. Con el objetivo final de contribuir a la web semántica y traducir sistemas KOS a un lenguaje formal procesable por máquinas se creó el vocabulario SKOS (del inglés *Simple Knowledge Organization System*). SKOS[96] es en esencia una ontología desarrollada en OWL que proporciona un modelo para la representación de la estructura básica y el contenido de sistemas KOS. Un uso básico de SKOS permite identificar los recursos conceptuales (conceptos) mediante URIs, etiquetarlos con literales de uno o varios idiomas, documentarlos con diversos tipos de notas, relacionarlos entre sí mediante estructuras jerárquicas informales o redes asociativas, y agregarlos a esquemas de conceptos. Al tratarse de una aplicación de RDF, SKOS permite la creación y publicación de conceptos en la Web, así como vincularlos con datos en este mismo medio e incluso integrarlos en otros esquemas de conceptos.

En la Figura 4 se muestran los principales conceptos y relaciones definidas en el vocabulario SKOS. Como se puede observar las principales clases son **skos:ConceptSchema** que sirve para representar los sistemas KOS, como entidad, en RDF y la clase **skos:Concept**, que se utiliza para representar los conceptos recogidos en el sistema KOS en cuestión. Es importante mencionar que para representar un sistema KOS es necesario crear un URI que identifique dicho sistema de manera única. Por ejemplo, si quisiéramos transformar el tesauro UNESCO, el individuo que representa el tesauro UNESCO en sí que sería instancia de la clase skos:ConceptSchema. Además, se ha de crear un URI por cada uno de los conceptos recogidos por dicho tesauro que se definen como instancias de la clase skos:Concept.

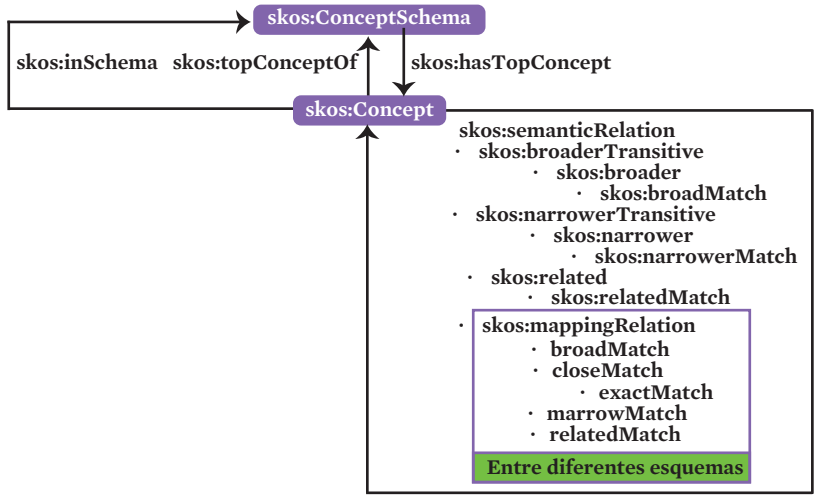
Mediante la relación skos:inSchema se establece la relación entre cada concepto perteneciente al sistema KOS y el sistema KOS al que pertenece. De esta manera, dado un concepto se puede saber a qué sistema de conocimiento pertenece. Además mediante la relación **skos:hasTopConcept** y su inversa **skos:topConceptOf** se identifican los conceptos del primer nivel de la clasificación o sistema KOS.

Entre instancias de la clase skos:Concept se pueden dar distintos tipos de relaciones como se muestra en la figura 4. Las principales relaciones son **skos:narrower** y su inversa **skos:broader** que permiten establecer relaciones de **hiponímia** (indicar qué conceptos son más específicos) e **hiperonimia** (indicar qué conceptos son más generales) entre conceptos respectivamente.

[96] SKOS – Simple Knowledge Organization System http://www.w3.org/TR/skos-primer/

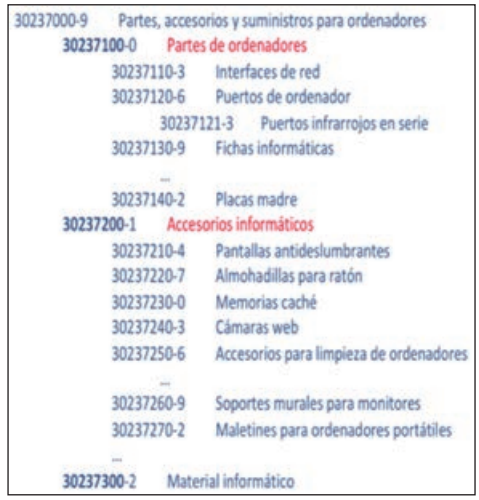
Las relaciones entre conceptos de distintas clasificaciones dadas por las siguientes propiedades: **skos:mappingRelation, skos:broadMatch, skos:closeMatch, skos:exactMatch, skos:narrowMatch y skos:relatedMatch** mostradas también en la figura 4.

FIGURA 4
Resumen entidades y relaciones principales de SKOS



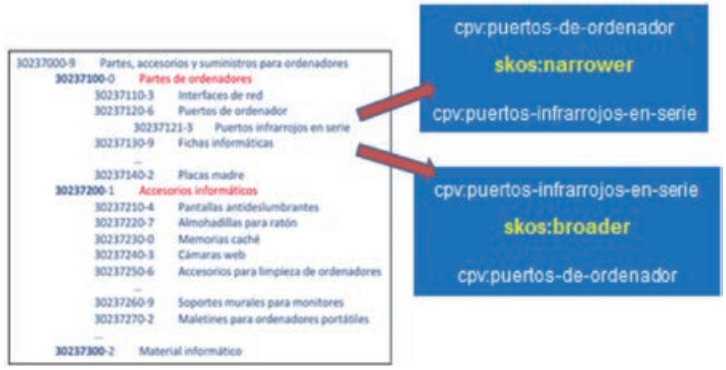
Es importante mencionar cierta característica de los sistemas de organización del conocimiento en contraposición con las estructuras ontológicas. Es decir, para ciertos recursos, como los tesauros o algunas clasificaciones, no se puede asegurar que entre un concepto de un recurso y los conceptos superiores de dicho recurso se satisfaga la semántica de la relación **"subclase de"** para todos los casos. Además, en algunos nodos se mezclan conceptos que tienen tanto relaciones jerárquicas como de **meronimia** (parte de) que deberían ser modeladas en distintas clases perdiendo por tanto la estructura original de la clasificación. Para ilustrar este problema nos basaremos en el ejemplo extraído de la clasificación CPV mostrado en la Figura 5. Como se puede observar dentro de un misma concepto se mezclan partes de ordenadores como accesorios. Si bien es cierto que cada uno de los código inferiores de la clasificación es subclase del concepto superior, es un error de modelado definir una clase como el conjunto de conceptos de distinta naturaleza (partes, accesorios y suministros), por lo que esta clase debería estar dividida en tres conceptos perdiendo así como se ha comentado antes, la estructura original de la clasificación. Además, siguiendo ese tipo de modelado no se podría transformar las clasificaciones automáticamente pues habría que identificar y revisar manualmente que conceptos tienen relaciones jerárquicas entre ellos y cuáles de meronimia.

FIGURA 5
Ejemplo de códigos CPV



8.4.1. Aplicando SKOS a una taxonomía

Vamos a aplicar a continuación SKOS a unos elementos de una taxonomía, traduciéndola a RDF. En la parte izquierda de la imagen podemos ver un conjunto de códigos correspondientes una taxonomía denominada CPV utilizada habitualmente en entornos de Contratación Pública.



Para dicha clasificación CPV se generan tantos individuos como categorías contenga la clasificación como instancias de skos:Concept y un individuo que representa la clasificación en sí como instancia de skos:ConceptSchema.

Cada uno de estos individuos se identificará de forma única mediante su URI. Dichos individuos se relacionarán entre sí mediante las propiedades mostradas en la formando la estructura definida en la clasificación dada.

En la imagen se puede observar que el concepto "puertos de ordenador" tiene como concepto más específico (equivalente a "*narrower" en SKOS*) "puertos infrarrojos en serie" y viceversa, "puertos infrarrojos en serie" tiene como concepto más general (equivalente a "*broader" en SKOS*) el concepto "puertos de ordenador". Consecuentemente Este ejemplo se formalizaría mediante las tripletas que están a la derecha de la imagen.

9. Gestionando el Conocimiento y la Veracidad de la información

Muy posiblemente la 4ª V del Big Data sea la **Veracidad**. La tentación de acumular sin más datos, información en un gran repositorio del cual nos creamos que mágicamente vamos a extraer conocimiento que no existía previamente gracias al Big Data, se convierte en una quimera si no tenemos en cuenta que podemos estar acumulando datos buenos, datos regulares y datos malos. "*Garbage in, garbage out*" se convierte de nuevo en la maldición bíblica del Big Data.

Esta gestión de la veracidad se apoya en un proceso denominado *knowledge crystallization* o *knowledge fusion*, que utilizan grandes Bases de Conocimiento.

El objetivo de un proceso de **Knowledge Crystallization** es obtener una descripción lo más compacta posible sobre un concepto a partir de un conjunto de datos, dejando a un lado la información superflua, irrelevante, repetida o falsa y destacando la información más relevante.

El conocimiento está evolucionando de forma constante. Incluso el que se considera ya "cristalizado" puede recibir interacciones por parte de la comunidad que lo mejoren. Se va construyendo de forma evolutiva, incremental y distribuida[97]. Evoluciona hacia un estado estructurado y refinado de dos posibles maneras, que frecuentemente son también complementarias

- Por medio de interacciones de comunidades de usuario.
- Por medios automatizados, extrayendo hechos de forma automática de Internet.

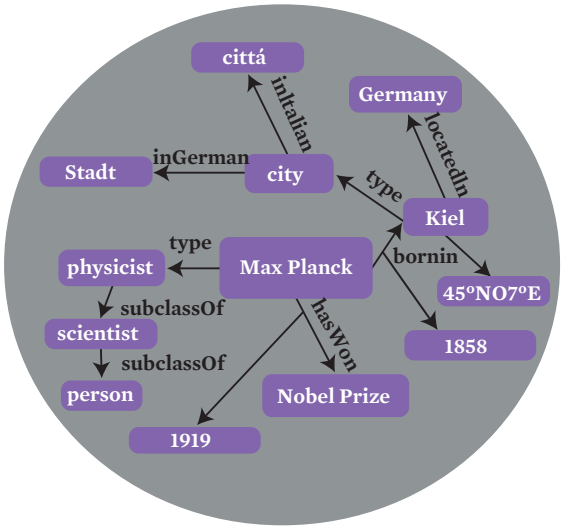
[97] http://cdn.intechopen.com/pdfs-wm/29147.pdf Ruth Cobos, UA Madrid

9.1. Knowledge Crystallization basado en Comunidades de usuario

En el caso de las comunidades de usuario el criterio que se utiliza, siempre dentro de un proceso de evaluación, es el de aceptación por parte de la comunidad; por otra parte el que no obtiene suficiente aceptación es borrado. Al principio un Comité Director de Conocimiento necesita estar encargado de la evaluación del conocimiento, debido a la falta de una masa crítica de conocimiento y de interacción. Cuando se llega a una masa crítica la evaluación cambia a estar basada en comunidades virtuales de expertos.

Se consideran expertos a aquellos usuarios que haya añadido conocimiento que haya cristalizado y cuyo trabajo es reconocido por la comunidad virtual. Las comunidades virtuales de expertos se construyen en base a la suma de comunidades especializadas en sub-áreas específicas de conocimiento. Cada comunidad hace su propio proceso de evaluación, de manera similar al mecanismo de evaluación por pares (en inglés "*peer review*").

Durante este último decenio se han desarrollado varias grandes **Bases de Conocimiento** (en inglés "knowledge bases"), construidas a partir de la Wikipedia y otras fuentes como **YAGO**, **DBpedia** o **FreeBase.**



YAGO[98] (Yet Another Great Ontology) es una base semántica de conocimiento construida por el Max Planck Institut a partir de la Wikipedia, WordNet y GeoNames. Cuenta con más de 10 millones de Entidades, incluyendo personas, organizaciones, ciudades, etc y más de 120 millones de hechos sobre dichas Entidades. Destaca por contener una exactitud de sus hechos superior al 95%.

DBpedia[99] es una versión en forma de Web Semántica de la Wikipedia, por lo que utilizan RDF y admite consultas a través de SPARQL. Es realizado por la Universidad de Leipzig

[98] YAGO http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

[99] DBPedia http://wiki.dbpedia.org/

y la empresa OpenLink, que destaca también por Virtuoso[100], su Base de Datos Big Data orientada a grafos. Está conectada a otras fuentes de datos como GeoNames, MusicBrainz, CIA World Factbook, Proyecto Gutenberg y Eurostat. Cuenta con cerca de 4 millones de Entidades.

FreeBase[101] es el otro repositorio ontológico de referencia. Fue desarrollada por la empresa Metaweb, que fue adquirida por Google en el año 2010. Recoge datos de fuentes como Wikipedia, ChefMoz, NNDB y MusicBrainz y aportaciones individuales de usuario voluntarios. Metaweb es también una base de datos orientada a grafos. Cuenta con su propio lenguaje de acceso, denominado MQL y genera representaciones en RDF.

Estos repositorios almacenan millones de hechos, la unidad de conocimiento, pero todavía están muy lejos de poder considerarse "completas". Además la fuente de mantenimiento y evolución de su contenido es mediante personas, generalmente voluntarios, ya que el mantenimiento de sus fuentes es así (por ejemplo los bibliotecarios voluntarios de Wikipedia).

9.2. Knowledge Crystallization basado en la Extracción automática de la Web

El otro enfoque es aplicar **técnicas de extracción de la web de información abierta**, **sin esquemas previos** ni ontologías de referencia, como hacen Reverb, OLLIE y PRISMATIC **o extrayendo información de la web pero apoyándose en una ontología** como hacen NELL, ReadTheWeb, PROSPERA, Elementary / Deep Dive. En el entorno de las buscadores existen notables esfuerzos en estas líneas.

Son conocidos los esfuerzos que está realizando actualmente Google, entre ellos los publicados en el paper "*Knowledge Vault: a Web-Scale Approach to Probabilistic Knowledge Fusion*"[102], escrito por varios miembros destacados de dicha empresa.

Es razonable pensar que en torno a este u otros productos aparezcan nuevos servicios y que se estructuren y evolucionen los servicios existentes. Alguna de las últimas contrataciones que ha realizado Google también apunta en esta línea[103]. De hecho

[100] Virtuoso http://virtuoso.openlinksw.com/

[101] Freebase http://www.freebase.com/

[102] Xin Luna Dong , Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy , Thomas Strohmann, Shaohua Sun, Wei Zhang. Google. "Knowledge Vault: a Web-Scale Approach to Probabilistic Knowledge Fusion https://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf

[103] Natasha Noy https://www.linkedin.com/in/natashafnoy

se espera que incluso el algoritmo por el que Google presenta uno u otro resultado evolucione en torno a esta Base de Conocimiento.

TABLA 1

Comparison of knowledge bases. KV, DeepDive, NELL, and PROSPERA rely solely on extraction, Freebase and KG rely on human curation and structured sources, and YAGO2 uses both strategies. Condent facts means with a probability of being true at or above 0.9

Name	# Entity types	# Entity instances	# Relation types	# Condent facts (relation instances)
Knowledge Vault (KV)	1100	45M	4469	271M
DeepDive [32]	4	2.7M	34	7M[a]
NELL [8]	271	5.19M	306	0.435M[b]
PROSPERA [30]	11	N/A	14	0.1M
YAGO2 [19]	350,000	9.8M	100	4M[c]
Freebase [4]	1,500	40M	35,000	637M[d]
Knowledge Graph (KG)	1,500	570M	35,000	18,000M[e]

Esta Knowledge Vault combina conocimiento extraído de Bases de Conocimiento, especialmente de **FreeBase**, con conocimiento extraído con procesos automáticos de la web mediante técnicas de **Scraping**, como las que explicamos en el apartado "3.1 Business Bots, Spiders, Scrapers".

Destaca[104] por ser mucho más grande que las otras bases de conocimiento que presentamos en el apartado anterior y por tener un conjunto de hechos muy grande con niveles de confianza superiores al 70% y al 90%.

Cuenta con tres componentes principales:

- **Extractores**, que extraen hechos, codificados en tripletas RDF (sujeto, predicado, objeto) de un enorme número de fuentes de información, añadiendo en cada extracción un **nivel de confianza y certidumbre** a cada tripleta extraída.

[104] Xin Luna Dong , Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy , Thomas Strohmann, Shaohua Sun, Wei Zhang. Google. "Knowledge Vault: a Web-Scale Approach to Probabilistic Knowledge Fusion https://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf https://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf

Por ejemplo a la tripleta ("Miguel de Cervantes y Saavedra, /people/person/place_of_birth, "/España/Madrid/Alcalá de Henares") se le asignaría un nivel de confianza muy alto, pero existiría una equivalente con el pueblo de "/España/Zamora/Sanabria" o "España/Ciudad Real/Alcázar de San Juan" con nivel de confianza muy bajo, ya que también hay páginas web que afirman que dichas ciudades fueron su lugar de nacimiento.

- **Valoraciones a priori** *("Graph-based priors"):* aprenden la probabilidad que tiene a priori una nueva tripleta a partir de la información disponibles de otras tripletas ya almacenadas previamente.

  Por ejemplo si disponemos de dos tripletas, una que conecta a un hijo con un padre y otra que conecta dicho hijo con su madre, otra tercera tripleta en la que dicho padre y dicha madre están casados tendrá una valoración a priori muy alta.

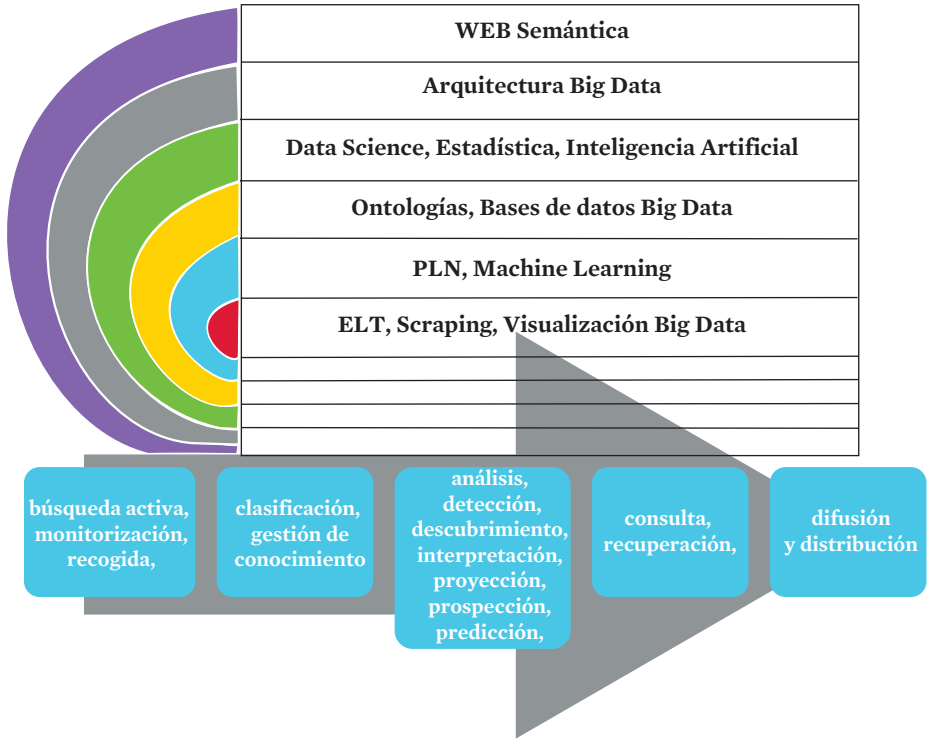
- **Fusión de Conocimiento** *("Knowledge Fusion"):* evalúa la probabilidad de que una tripleta sea cierta, en base a los a la Valoración a Priori y el nivel de confianza que proporcionan los Extractores.

  Por ejemplo un hecho que esté disponible en Freebase apoyará un nivel de probabilidad alto o muy alto frente a otros de los que sólo se disponga información extraída de un blog con pocas visitas, por ejemplo.

10. Mapeando las tecnologías Big Data y las actividades de Vigilancia Estratégica e Inteligencia Competitiva

Este apartado pretender mapear los puntos anteriores contra los grupos de actividades que se realizan en los proyectos de Vigilancia Estratégica e Inteligencia competitiva, a partir de la cadena de valor que usamos en la presentación del libro.

Para el primer grupo de actividades, dedicado a la **Búsqueda Activa, la Monitorización y la Recogida de Información** destacamos como clave el **Scraping** de información, que veíamos en el punto 3.1. Para la recogida de información se utiliza frecuentemente las aplicaciones de ETL ("Extract, Transform and Load"), que se encuadran en el Business Intelligence. Big Data cambia el orden de ETL a **ELT**, realizando la carga masiva de datos para poder explotarla en las actividades del resto de grupos. Abundamos sobre este tema en el apartado 4.4 de Integración de Datos.

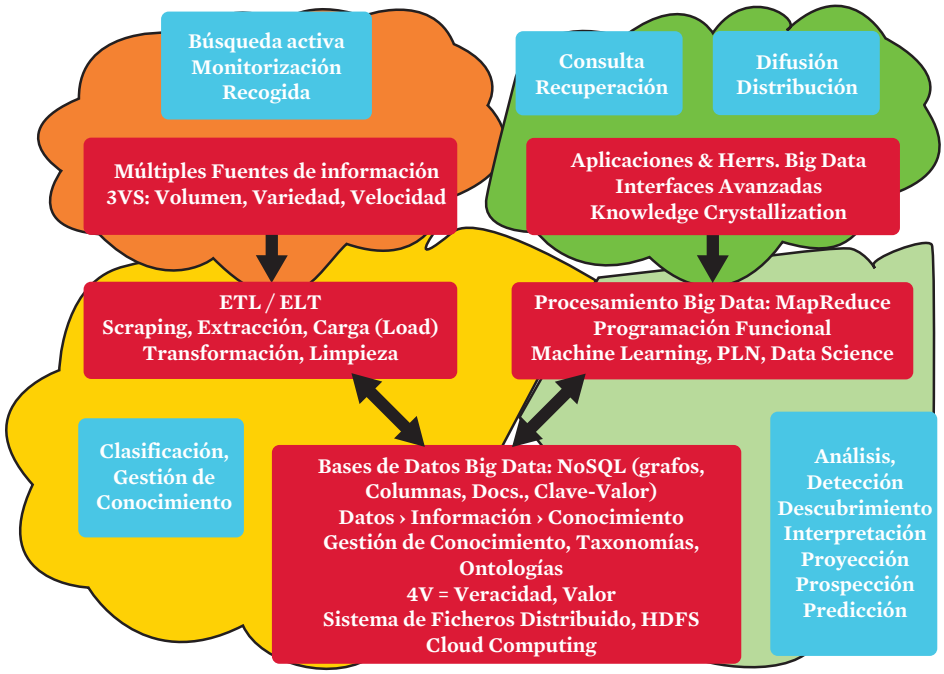


En el segundo grupo de actividades, dedicado a la **Clasificación y Gestión del Conocimiento**, son clave las **Ontologías**, como repositorios de referencia de conocimiento. También son importantes las distintas **bases de datos NoSQL**, que pueden facilitarnos la organización de conocimiento. Para la clasificación los **Sistemas de Procesamiento Big Data y MapReduce** son herramienta imprescindible en caso de que sea necesario el procesamiento en masivo y continuo de los datos. **Machine Learning y PLN** también pueden proporcionar mucho valor en la Clasificación de la Información.

El tercer grupo de actividades está relacionado con la explotación de la información y el conocimiento. Incluye actividades relacionadas con el **análisis, la detección, el descubrimiento, la interpretación, la proyección, prospección y predicción**. Todo lo que presentamos en los apartados de **"Data Science", MapReduce**, **"Machine Learning"** y **"Procesamiento de Lenguaje Natural"** tienen la capacidad de evolucionar y hacer transformar estos procesos para proporcionar nuevas capacidades. Las herramientas de **Análisis y Visualización de Datos** que presentaremos en el punto 4.7 dedicado a "Soluciones e Interfaces Big Data" también serán de mucha actividad en las actividades de **Análisis y Detección**.

ILUSTRACIÓN 18

Propuesta de mapeo entre Big Data y Vigilancia Estratégica e Inteligencia Competitiva. Original del autor

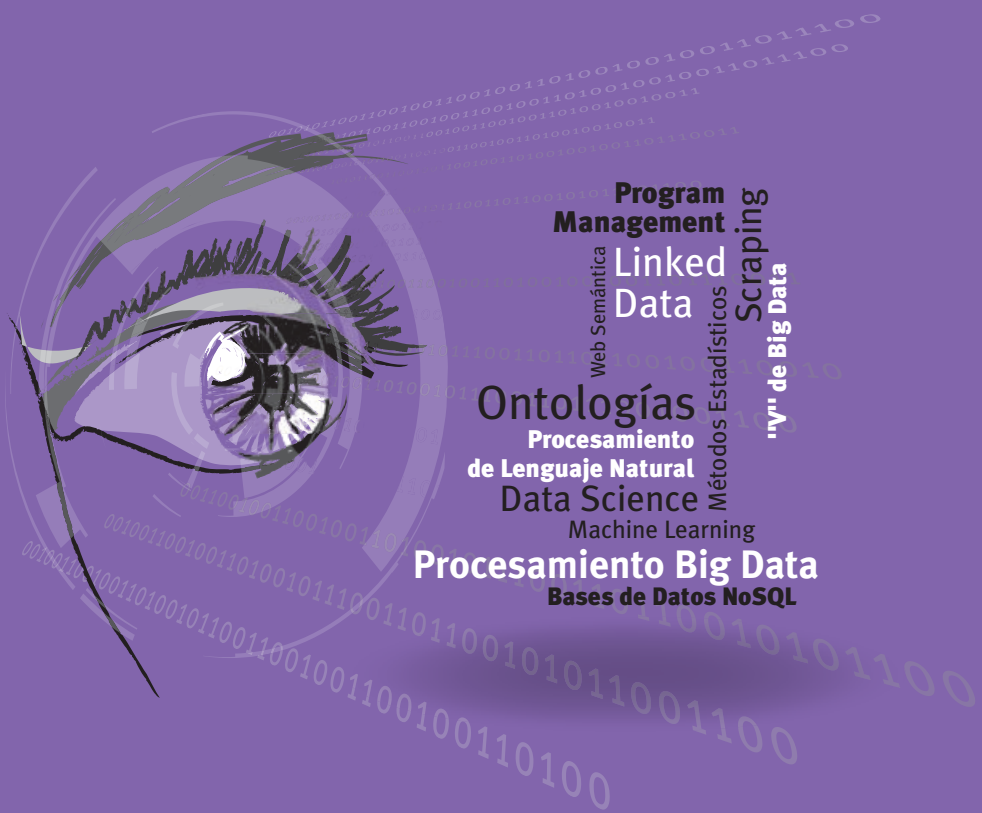


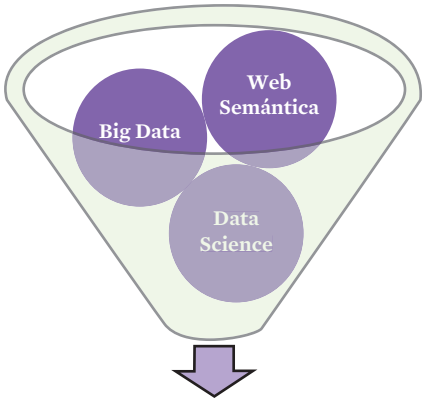
Para la **Consulta y Recuperación de Información**, englobada en el cuarto punto, es importante el apartado 4.7 que mencionábamos anteriormente, especialmente todo lo relacionado con Análisis y Visualización de la información. Los **lenguajes de consulta a Bases de Datos Big Data** también son la herramienta básica para acceder a la información almacenada en las **Bases de Datos Big Data**, especialmente las orientadas a grafos, que usamos en las **Ontologías**. El **Procesamiento de Lenguaje Natural** también puede ser una herramienta que aporte gran potencia, al poder hacer búsquedas más inteligentes que la mera búsqueda de palabras clave.

Las actividades de **Difusión y Distribución** se ven claramente afectadas por la **Web Semántica**. Cualquier información puesta a disposición va a tener su semántica mediatizada por las Ontologías actualmente publicadas en Internet e impulsadas por organismos tan importantes como el W3C. Las **Herramientas de Visualización Big Data** nos van a proporcionar nuevos formatos e interfaces con los que comunicar mejor la complejidad Big Data. La Distribución de la información también se ve afectada por otras tecnologías y procesos que no vemos en este libro, como la integración con otras grandes aplicaciones o con tecnologías de gestión de procesos de negocio (BPM, "Business Process Management").

4

DISEÑANDO SISTEMAS DE VIGILANCIA E INTELIGENCIA CON NUEVAS CAPACIDADES BIG DATA

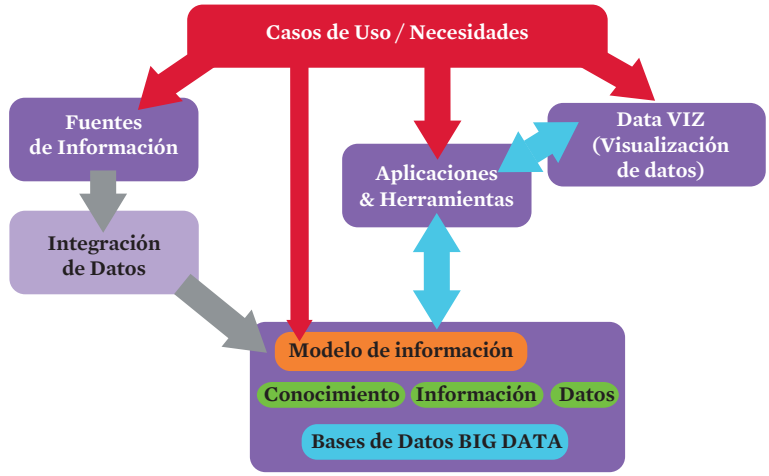




Una vez introducidas las nuevas capacidades que nos traen la **Web Semántica, Big Data y Data Science**, y realizado en el apartado anterior su mapeo con la Vigilancia Estratégica y la Inteligencia Competitiva, en este apartado nos planteamos hacer evolucionar la Vigilancia Estratégica y la Inteligencia Competitiva con estas nuevas capacidades que hemos presentado.

El concepto de "**Capacidad**", que veremos en el apartado 5 dedicado a la "Formalización del Modelo y la Metodología" tampoco es baladí: está extraído de otra área de conocimiento, el **Program Management**, en la que también vamos a encuadrar el Diseño que proponemos para nuestro Sistema de Vigilancia.

En la imagen que presentamos a continuación se refleja un esquema con los grandes bloques en los que se estructurará la Arquitectura Funcional del Sistema de Vigilancia/Inteligencia y que nos sirve también como guía para estructurar este capítulo 4.



Se trata en el primer apartado los **Casos de Uso y Necesidades** (4.1) más habituales en Vigilancia Estratégica e Inteligencia Competitiva, seguiremos con un apartado dedicado al efecto de la explosión exponencial de contenidos en internet, materializado en nuevas **Fuentes de Información y Taxonomías** (4.2), trataremos la **Integración de Datos** (4.3) de las mismas en las **Bases de Datos NoSQL, las Bases de Datos del Big Data** (4.5)y el **Modelo de Información** (4.4.) necesario para dar soporte al Sistema. Finalmente se presenta una propuesta de **Aplicaciones, Herramientas y Sistemas de Visualización Big Data** en el apartado 4.6.

Las actividades a realizar se formalizan en el apartado 5 “Modelo y Metodología”. En este apartado se presenta conceptualmente cada bloque funcional, se explica su relación con la Vigilancia Estratégica y la Inteligencia Competitiva y se destacan las características particulares que nos aporta Big Data.

1. Casos de Uso y Necesidades de Vigilancia Estratégica e Inteligencia Competitiva: el estilo de pensar "Big Data"

Resumiendo lo que decíamos en el capítulo 2, los casos de uso más habituales de la Vigilancia Estratégica y la Inteligencia Competitiva responden a:

- La identificación de cambios en el entorno.
- La identificación de tendencias y nuevas necesidades.
- El conocimiento de la competencia.
- La identificación del posicionamiento estratégico para la reducción de la incertidumbre y el riesgo y para dilucidar caminos de evolución del Sistema.

Nos encontraremos por supuesto con otras necesidades de negocio expresadas para el sistema que no respondan a estas líneas estratégicas, fundamentalmente relacionadas con la obtención de información base sobre las entidades de negocio que sustentan el posicionamiento estratégico.

Para cada necesidad y caso de uso realizaremos las siguientes actividades:

- Examinaremos el grado de viabilidad de aportar una solución al caso de uso expresado.
- Determinaremos las fuentes de información necesarias.
- Diseñaremos la integración de datos necesaria para obtener todos sus datos.
- Revisaremos el modelo de información del sistema para asegurarnos de que cuenta con todas las Entidades de Negocio necesarias.
- Elegiremos las herramientas útiles y diseñaremos con ellas la aplicación adecuada para solucionar el caso de uso.
- Diseñaremos el sistema de visualización y distribución de los datos generados por las aplicaciones de tal manera que optimicen los objetivos del caso de uso.

Un ejemplo de caso de uso con cierto nivel de complejidad podría ser por ejemplo el de una empresa de IT en expansión que desea realizar ***"la detección de los mejores candidatos expertos en Big Data susceptibles de ser contratados por la empresa"***. Como hemos visto, Big Data refleja una realidad novedosa, con escasas referencias y poco claras. ¿Cómo lo resolveríamos?

En primer lugar tenemos que simplificar el caso de uso, para lo cual lo vamos a segregar en requisitos más sencillos, de una granularidad inferior, con el criterio de que les podamos aplicar las actividades señaladas anteriormente. Para ello el caso lo podemos **segregar en tres requisitos más sencillos**:

1. Descripción Perfiles de Empleo tipo Big Data.
2. Conceptos relacionados con Big Data.
3. Personas expertas en Big Data.

Vamos a aplicar las actividades mencionadas al primer requisito, "**Descripción Perfiles de Empleo tipo Big Data**", aprovechando también a la vez para introducir

- **Fuentes de Información:** lo inmediato y necesario es usar la lista interna de **Perfiles de Empleo** pero podemos hacer mucho más. Si buscamos en redes sociales profesionales, como **LinkedIn** o **InfoJobs**, posiciones relacionadas con "Big Data" podremos hacernos con un conjunto más rico y completo de Perfiles de Empleo. También podemos usar **Twitter**, **webs de universidades** con titulaciones relacionadas con Big Data, centros de investigación relevantes... *Las fuentes susceptibles de ser útiles crecen*. Buscaremos **taxonomías** relacionadas con las Fuentes, con el concepto Empleo y con Big Data.
- **Integración de Datos:** para la integración de datos programaremos un **Bot** que utilizará las técnicas de **scraping** vistas en el apartado de "Business Bots, Spiders y Scrapers".
- **Modelo de Información:** se relacionan dos Entidades de Negocio importantes: **Empleo** (los perfiles) y los P**roductos, Soluciones, Servicios y Tecnologías** relacionados con Big Data.
- **Base de Datos Big Data:** vamos a realizar una descarga masiva de información documental, para lo que puede parecer oportuno el uso de una base de datos documental. Los conceptos en torno a Big Data, puro conocimiento, parece adecuado que sean organizadas en una base de datos orientada a grafos.
- **Herramientas y Aplicaciones:** parece clara la oportunidad de **usar Procesamiento de Lenguaje Natural** para detectar las Entidades de Negocio por un lado y puede ser interesante explorar el uso de la técnica **LDA** (Latent Direchlet Allocation)

que explicábamos en el apartado de "Algunas técnicas útiles para Data Science" y posiblemente también **Machine Learning**. El uso de un **Buscador** tipo Apache SolR, también puede aportar valor.

- **Visualización y Distribución de Datos:** diseñaremos un sistema de visualización que permita integrar el buscador, extraer las Entidades de Negocio, relacionar y categorizar las Fuentes de Información, las Entidades de Negocio más importantes, los diferentes perfiles y los tipos de bases de datos que hemos elegido.

Aplicaríamos el mismo proceso a los otros dos requisitos, obteniendo de la integración de las soluciones la solución completa.

Resulta interesante también estudiar el conjunto de información disponible y buscar otras utilidades para la misma. Por ejemplo, estudiando qué empresas publican Perfiles de Empleo Big Data podemos obtener información interesante sobre la posible competencia. Asimismo dichas empresas son susceptibles de tener empleados con el perfil que necesitamos, por lo que debemos priorizar la búsqueda en redes sociales profesionales de información vinculada con esas empresas.

> Este es, en mi humilde opinión, el **"estilo de pensar Big Data"**: la **"Big Data way of thinking".**

2. Fuentes de Información. Taxonomías

La explosión exponencial de contenidos en Internet se ha materializado en miles de fuentes de información, unas especializadas, otras genéricas, pero en todo caso susceptibles de ser incorporadas dentro de un Sistema de Vigilancia Estratégica.

Seleccionar y mantener las Fuentes de información que serán parte del Sistema de Vigilancia es una de las tareas de nivel más estratégico. En este apartado diferenciamos y tratamos los dos grandes tipos de fuentes, las estructuradas y las no estructuradas, e introducimos el concepto de taxonomía.

2.1. Estructura de Catalogación de Fuentes

Todas las fuentes incluidas en un Sistema de Vigilancia deben ser estudiadas y catalogadas para que puedan responder a los objetivos y necesidades a las que responde el sistema. Debe considerarse recoger la siguiente información:

- **Información sobre el Organismo, Entidad o Empresa** que la elabora y mantiene. Es relevante saber también si la información está o no siendo actualizada y mantenida en la actualidad o si ha sido descontinuada. Debe valorarse también la calidad y fiabilidad del Organismo en cuestión, por ejemplo a partir de los medios disponibles o el histórico de Fuentes generadas y mantenidas.
- **Origen de la información:** es relevante saber si la información es original o si está elaborada a partir de otras fuentes y de qué fuentes se trata. Asimismo debe recogerse información sobre el método de recopilación de la misma, que nos determinará la calidad y fiabilidad que le podemos asignar. Es importante saber si ya estamos incorporando la información por estar incluidas las fuentes originales en nuestro sistema o, visto desde otra perspectiva, si nos resulta valiosa la elaboración de la información que realiza el Organismo en cuestión.
- **Información Temporal:** debemos conocer el origen y el final temporal de la información, la fecha de última actualización y el periodo de actualización, es decir, cada cuanto tiempo se actualiza la información.
- **Formato de la información:** información relevante sobre el tipo de archivo en el que se encuentra la información, si es accesible a través de internet, si es de forma abierta o mediante usuario y contraseña, si es de acceso libre o de pago. El formato de la información puede conllevar diferentes tipos de campos, por ejemplo un "cubo" de Business Intelligence requerirá campos específicos, diferentes a una simple hoja Excel.
- **Contenidos:** se recogerá información sobre los campos contenidos en la fuente y consecuentemente las tecnologías necesarias para la explotación de la misma. Debe tenerse en cuenta que diferentes campos pueden requerir diferente tipo de tecnología. Asimismo se recogerá información sobre el idioma y el tema sobre el que se recoge información. También debe tomarse como referencia lo que llamamos Entidades Principales (ver punto 4.3.1) y estudiarse si es relevante para los objetivos del sistema recoger información específica sobre los mismos.

La integración técnica de las fuentes puede ser más o menos complicada, dependiendo de las diferentes circunstancias que tratamos en el apartado de Scraping. Generalmente, las herramientas de ETL (Extracción, Transformación y Carga, en inglés "*Load*") de los paquetes de software de Business Intelligence son suficientes para realizar la integración fiable de una fuente de información estructurada.

2.2. Fuentes Estructuradas, semiestructuradas y no estructuradas

Llamaremos Fuentes Estructuradas a aquellas que estén organizadas en forma de tabla y sobre la que podemos deducir el tipo de información que contiene. En el otro extremo nos encontramos con Fuentes no Estructuradas, es decir, aquellas que carecen de una estructura de la que deducir la información que contiene, como ocurre en páginas web o colecciones de documentos por ejemplo.

En un término medio nos encontramos con fuentes semi-estructuradas, es decir fuentes que cuenta con dicha estructura de referencia de la que deducimos el tipo de información que contiene pero que el contenido de uno o varios campos es no estructurado. Un campo no estructurado de una fuente estructurada puede ser por ejemplo el resumen de un libro: si no disponemos de otra fuente podemos necesitar deducir de él otra información, como el área de conocimiento al que pertenece o los temas que trata.

2.3. Tratamiento de Fuentes no estructuradas

Tanto de Páginas Web como de Documentos, información de Redes Sociales y resto de fuentes no estructuradas podemos extraer diferentes propiedades que llamamos Metadatos, y que nos pueden ser útiles para estudiar y clasificar la documentación. Los metadatos podemos encontrarlos por ejemplo accediendo a las Propiedades de un documento Office (Archivo > Información > Propiedades) o en las etiquetas <META> de las páginas web que podemos visualizar desde cualquier navegador haciendo uso de la opción “Ver código fuente”.

Para la descripción de metadatos existe el estándar *Dublín Core* (DC)[105], que define varios campos habitualmente utilizados como tal. Por ejemplo si nos encontramos en una página web el siguiente campo:

<META name=”DC.description” content=”descripción de esta página web”>

significa que estamos utilizando el campo estándar *Dublín Core* “DC.description” para incluir una breve descripción del contenido de la página web.

Es relevante destacar que aunque todavía siguen siendo utilizadas, la tendencia en curso es que las etiquetas <META> sean sustituidas por el uso de otros formatos, entre

[105] Dublin Core Metadata Iniative http://dublincore.org/

los que destacan Microformats[106] y Schema.org[107]. Estos formatos permiten representar conceptos habitualmente utilizados en internet como personas, productos[108], eventos[109], etc, con información semántica asociadas. Los principales buscadores, como Google, Bing o Yahoo indexan y dan soporte a estos formatos. En el ejemplo de la imagen[110], se presenta un comentario emitiendo una opinión sobre un restaurante:

```
<div class="hreview">
    <span class="item">
        <span class="fn">L'Amourita Pizza</span>
    </span>
    <span class="rating">3.5</span>
    <span class="reviewer">Ulysses Grant</span>
    <span class="dtreviewed">2009-01-06</span>
    <span class="summary">"Deliciuous, tasty pizza in Eastlake."</span>
</div>
```

Para el tratamiento de fuentes no estructuradas se están utilizando diferentes técnicas, disciplinas y metodologías, que presentamos en dentro del capítulo de "Nuevas Capacidades Big Data para los Sistemas de Vigilancia":

- Scraping.
- Linked Data
- Procesamiento de Lenguaje Natural.
- Machine Learning.
- Inteligencia Artificial, Estadística, Investigación Operativa y Data Science en general.

Debe asimismo recogerse toda la información posible, según el esquema presentado en el apartado de "Fuentes Estructuradas".

2.4. Taxonomías

La **Taxonomía** es la ciencia de la clasificación y asimismo se utiliza como un sinónimo de **Clasificación**. Originalmente se utiliza en el mundo de la biología, sin embargo su uso

[106] Que són los Microformatos http://microformats.org/wiki/what-are-microformats (en inglés) y https://es.wikipedia.org/wiki/Microformato

[107] Schema.org http://schema.org/

[108] Representación de Productos con Microformats http://microformats.org/wiki/hProduct

[109] Representación de eventos en Calendarios con Microformats http://microformats.org/wiki/hcalendar

[110] Google Rich Snippets Examples: http://microformats.org/wiki/google-rich-snippets-examples

se ha extendido al resto de ámbitos del conocimiento. Ante la evidente explosión de Fuentes de Información en Internet se ha hecho especialmente relevante la clasificación de sus campos, con el objetivo de armonizar todo lo posible los contenidos y poder así cruzar y unir diferentes fuentes de información gracias a la estandarización que proporciona el uso de taxonomías. Nos van a interesar especialmente las taxonomías que ofrezcan clasificaciones sobre las Entidades Principales que mencionaremos en el apartado 4.4 "Modelo de Información", que también se presentan en el punto de introducción sobre "Vigilancia Estratégica e Inteligencia Competitiva".

Las taxonomías se enfrentan actualmente a la vorágine de nuevos conocimientos emergentes, de cambios y de matices en los existentes y en general en la dificultad de seguirle el ritmo a la realidad. Asimismo se enfrentan a un relativamente pobre punto de partida. Sin embargo las taxonomías son claramente una pieza clave en la arquitectura de los Sistemas de Vigilancia.

Una de las situaciones más comunes que tengan un nivel de granularidad inferior al necesario, es decir que se queden en un nivel de abstracción superior al necesario. Por ejemplo si queremos clasificar "*Neo4j*" como una *base de datos NoSQL orientada a grafos*, como mencionaremos en el apartado dedicado a Big Data, va a ser frecuente que la taxonomía se quede, como mucho, en Base de Datos, sin discernir si es relacional o NoSQL ni el tipo de base de datos NoSQL del que se trata. Consecuentemente a la fuente en la que se encuentre este tipo de información **no voy a poder hacer preguntas del tipo *"¿qué empresas puedo contratar porque tienen experiencia haciendo proyectos sobre las revolucionarias tecnologías Big Data?"*** o ***"¿qué personas son especialistas en bases de datos NoSQL para ofrecerles un puesto de trabajo remunerado con un sueldo un 30% por encima de mercado?".*** El freno, en este caso, a la innovación, es evidente.

Existen numerosas taxonomías utilizadas en la actualidad, que son promovidas y mantenidas por diferentes Organismos nacionales e Internacionales. Pasamos a continuación a nombrar algunos de los más relevantes:

- Sobre **Actividades Económicas:** es muy relevante la taxonomía **NACE** de la CE, que tiene en el **CNAE** su réplica en España. EEUU, Canadá y México usan la **NAICS** (North American Industrial Classification System). En Australia y Nueva Zelanda disponen de la **ANZSIC** (Australian and New Zealand Standard Industrial Classification). Los países africanos utilizan la taxonomía **NAEMA**. Por último es necesario mencionar también ISIC (International Standard Industrial Classification), la clasificación de las Naciones Unidas.
- Sobre **Contratos** destaca la CPV (Common Procurement Vocabulary). Sobre **Personas y Empleo** destacamos la International Standard Classification of Education

(ISCED), la International Standard Classification of Occupations (ISCO), gestionado por la Organización Internacional del Trabajo y la CNO (Clasificación Nacional de Ocupaciones), del INE español.

- Sobre **Productos**, destaca las taxonomías de la UE CPC (Central Product Classification) y CPA (Clasificación de Productos por Actividad). También son relevantes: la HS (Harmonized Commodity Description and Coding System), UNSPSC (United Nations Standard Products and Services Code) y NC (Nomenclatura combinada).
- Sobre **Propiedad Industrial** y **Patentes** destaca la IPC (International Patent Classification), la USPCS (U.S. Patent Classification System), la Clasificación de Niza y la Clasificación Armonizada de la OAMI y la OMPI, que extiende la de Niza.

Para documentar una taxonomía debemos recopilar información similar a la de una fuente y adicionalmente deberá documentarse la **estructura de la clasificación**.

Para los casos en los que nos encontremos Fuentes clasificadas con diferentes taxonomías existen pasarelas que establecen correspondencias entre las mismas. Habrá que documentarse sobre la existencia previa de pasarelas entre las taxonomías que nos interese. En caso contrario deberá diseñarse un mecanismo de traducción entre las taxonomías. Las dos opciones habituales son: realizar la pasarela en sí e implementar una ontología que integre las dos taxonomías.

2.5. Mapas de Información

Es recomendable también realizar mapeos entre las diferentes taxonomías, entre las Entidades de Negocio y las taxonomías y entre las Entidades de Negocio y las Fuentes. Recomendamos asimismo realizar este mapeo de forma visual. Existen diversas técnicas y metodologías en ingeniería informática que pueden inspirar estos diseños, por ejemplo los diagramas utilizados en UML (Unified Modeling Language) o los de modelado relacional de bases de datos.

2.6. Tipos de Datos

En las fuentes estructuradas más habituales son las siguientes: Bases de Datos estructuradas, Cubos de información, Ficheros con tablas en diferentes formatos como CSV, XLS o de ancho fijo, Ficheros con información geoespacial y Ficheros en diversos formatos estándar como XML, RSS, JSON o LDAP.

El tratamiento de fuentes no estructuradas es una de las áreas en las que la Vigilancia y el Big Data más se están desarrollando en la actualidad. Páginas web de cualquier tipo (medios de comunicación, blogs,...) o documentos de cualquier tipo (docs, e-mails...) y formato (Word, PDF...) son las fuentes más abundantes de fuentes no estructuradas.

El mundo Big Data nos ha traído otras fuentes estructuradas y no estructuradas, como el audio, el vídeo, información de todo tipo de sensores, ficheros de logs e información proveniente de teléfonos y otros dispositivos inteligentes, información de Redes Sociales, que podrían ser introducidos en Sistemas de Vigilancia Estratégica. Destaca la información de Redes Sociales: Twitter, Facebook y LinkedIN se han convertido en fuentes muy populares.

3. Integración de Datos

En este apartado vamos a abundar un poco sobre el estilo de integración de datos en los proyectos Big Data, específicamente en aquello que difiere de las integraciones clásicas que se realizan, sobre todo, en proyectos de Business Intelligence. Podemos considerar al Business Intelligence como "el padre" del Big Data, de hecho muchas empresas que tradicionalmente se han considerado de Business Intelligence ahora se publicitan como empresas Big Data.

3.1. ETL / ELT y Federación de Datos

La primera etapa de los proyectos de Business Intelligence es tradicionalmente la ETL, es decir, la Extracción de Datos de fuentes, la Transformación de dichos datos y la carga (en inglés "*Load*", de ahí la "L") de los datos transformados en una Base de Datos, un Repositorio de Datos.

La tendencia Big Data ha convertido las ETLs en ELTs[111], es decir, tras la extracción de los datos de la fuente no se hace ningún tipo de transformación sino que se cargan todos los datos en el Sistema, para posteriormente poder hacer todos los análisis y transformaciones que sean pertinentes. En definitiva, hasta la llegada de Big Data los datos se quedaban en las Fuentes de Información, integrándose sólo la información necesaria en los sistemas. **Con Big Data, la cadena *Dato > Información > Conocimiento* se encuentra de forma completa en el Sistema**. La razón fundamental es habilitar que posteriormente sea viable realizar tanto predicciones basada en datos históricos como responder a casos de uso que serán definidos en el futuro.

[111] "Big Data", Bill Schmarzo. Ed. Wiley 2013

Las nuevas capacidades Big Data de los nuevos ordenadores, con gran capacidad de cómputo en paralelo y grandes capacidades de almacenamiento de datos es parte del nuevo "Big Data way of thinking". Ahora podemos descargar grandes cantidades de datos y procesarlas con las técnicas avanzadas que se presentan en el apartado de "Nuevas capacidades Big Data".

Hay dos procesos, que no suelen explicitarse en el concepto de ETL pero que son clave para cualquier sistema que incorpore datos externos a sus bases de datos: la **Limpieza de Datos** (en inglés "*data cleansing*"), y el **Enriquecimiento de Datos** (en inglés "*data enrichment*"). En un Sistema de Vigilancia Big Data como el que estamos diseñando estos procesos siguen siendo clave.

Con el proceso de **Limpieza de Datos** intentamos conjurar el viejo dicho de la calidad de datos en los Sistemas de Información "*Garbage in, garbage out*". Da igual que diseñemos e implementemos un Sistema de Vigilancia fantástico combinando Machine Learning, Procesamiento de Lenguaje Natural, Ontologías, fantásticos repositorios Big Data, y aplicaciones y sistemas de visualización de datos súper-avanzados: si nuestro sistema gestiona datos erróneos, incompletos, duplicados o poco veraces, los resultados que se generarán contendrán contradicciones, anomalías, problemas de consistencia e irregularidades que alimentarán al Sistema de Vigilancia y llevarán a los usuarios a tomar decisiones incorrectas ya que estarán basadas en datos con una calidad inferior a la necesaria. Varias de las técnicas explicadas en el apartado de "Data Science" son utilizadas para este proceso.

El **Enriquecimiento de los Datos** se ha convertido en un proceso clave en los proyectos Big Data. Frecuentemente los datos disponibles o con los que se ha trabajado habitualmente en la organización no son suficientes para dar solución a los casos de uso y objetivos de negocio planteados. Una solución consiste en acudir a la gran cantidad de nuevos datos disponibles en Internet gracias al movimiento Open Data o directamente a empresas que se han especializado en obtener, mantener y poner datos a disposición de terceros.

Es posible obtener por múltiples fuentes información sobre Personas, Empresas, Productos, Tecnologías, Instituciones y en general para todas las Entidades de Negocio a las que nos referiremos en el punto 4.4 "Modelo de Información". Esta información puede consistir en datos sobre campos sobre los que ya disponemos información o datos sobre campos sobre los que no tengamos información. Será aconsejable categorizar también las Fuentes de Información por estas entidades de negocio.

Por último, deberemos tener en cuenta otra tendencia, la **Federación de Datos**, (en inglés "***Data Federation***"). Una **Base de Datos Federada** es un sistema que intermedia

con otras bases de datos de forma transparente al usuario, ofreciéndole una única interfaz de acceso a todos los datos. No será necesario por tanto trasladar a un único repositorio todos los datos sino que cada base de datos permanece autónoma y se programa en la base de datos federada cómo acceder a los datos de cada una de las otras bases de datos.

3.2. Fabricando un Bot para hacer Scraping

Todo proyecto que implica la agregación de información dispersa a lo largo de un conjunto de websites incluye la necesidad de implementar Bots, programas específicos dedicados a realizar la tarea de extracción de un modo lo más automatizado posible. Este tipo de programas suelen constituir las bases de recolección de información que alimenta el resto del sistema. A la hora de desarrollarlos existen diferentes acercamientos o enfoques que deberán ser cuidadosamente seleccionados en función del problema concreto a resolver, la capacidad técnica del equipo encargado de ello y la infraestructura disponible. Podemos agrupar los enfoques para la fabricación de un Bot en cinco grandes bloques:

- Desarrollo a medida completo.
- Utilizar aplicaciones de testing.
- Ejecutar macros y scripts dentro de un navegador.
- Desarrollo sobre frameworks.
- Utilizar servicios en la nube.

El primer enfoque sería la **programación completa del mismo mediante un desarrollo a medida**. Los lenguajes de programación más populares hoy en día, entre ellos Java, .NET, Python y PHP, proporcionan librerías y funciones especializadas que facilitan en gran medida la programación a medida de la funcionalidad habitualmente ejecutada por los Bots.

El segundo enfoque viene a partir de **aplicaciones diseñadas originalmente para hacer testing de aplicaciones web**. Probar una aplicación incluye programar un conjunto, potencialmente enorme de pruebas, ejecutar las pruebas, recoger los resultados y evaluar los mismos. Con la aparición de las aplicaciones web fue necesario construir aplicaciones capaces de ejecutar dichas aplicaciones, simulando ser navegadores que acceden a los servidores web permitiendo emular el comportamiento del usuario especialmente en entornos en los que hay mucha lógica de negocio. La funcionalidad de estas aplicaciones es muy similar a la necesaria para hacer Scraping, por lo que han

evolucionado para permitir realizar esta actividad. Al fin y al cabo estas herramientas han sido construidas con el objetivo de descargar y ejecutar el código javascript de los sites para asegurar su correcto comportamiento. Como hándicap, con este enfoque se pierde flexibilidad a la hora de manipular la información obtenida o tomar decisiones sobre cómo avanzar en detalle en las páginas web. Debe tenerse en cuenta que el comportamiento del navegador y su interacción con el servidor web es realmente difícil de imitar.

El tercer enfoque es **incluir y ejecutar pequeños programas dentro de un navegador**. Tendríamos dos posibilidades: macros y scripts. Las **macros** son registros de instrucciones que se ejecutan secuencialmente. Habitualmente se generan grabando de forma manual un conjunto de actividades realizadas en una primera ocasión, generando una lista de instrucciones que pueda ser invocada posteriormente a voluntad. Es una técnica bien conocida, usada en otros ámbitos, siendo especialmente conocidas las macros de Microsoft Office. Las macros tienen la desventaja de ser monolíticas y no permitir variaciones en función de la interacción con los resultados de la navegación. Los **scripts** son pequeños programas y por lo tanto permiten una ejecución condicionada a la interacción que se esté realizando con el servidor web. Será especialmente útil en la interacción con AJAX y aplicaciones que intercambien información con el servidor y en la extracción de datos compleja.

El cuarto enfoque consiste en el **desarrollo a medida utilizando frameworks** con funcionalidades orientadas a la creación y operación de un Bot. La utilización de frameworks de scraping permite recopilar información de un modo muy efectivo dejando que el propio framework sea el encargado de la gestión de los flujos de datos. Estos sistemas permiten la implementación de estructuras completas incluyendo scrapers, spiders y crawlers en los que el usuario sólo tiene que preocuparse por las particularidades del fuentes de datos definidas como objetivo. Como contrapartida, estos frameworks, al ser fruto de procesos de reutilización y abstracción hasta obtener los elementos comunes útiles en cualquier escenario, suelen perder flexibilidad y coartan comportamientos específicos que pueden ser necesarios en algunos casos particulares. De todos modos, la alternativa suele ser implementar todo el sistema de scraping desde cero, por lo que siempre merece la pena abordar su implementación en primera instancia

Finalmente el quinto enfoque consiste en utilizar un **Servicio en la Nube**. Este enfoque, además, resulta muy útil para realizar una primera aproximación exploratoria en este tipo de proyectos. El proceso de implementación de un scraper suele ser iterativo, de modo que en cada paso aumenta el conocimiento del sistema de extracción que se está construyendo y de la información extraída en sí misma. Sin embargo, en muchas ocasiones, la información que está siendo extraída no es la adecuada, o aparece una fuente adicional que no había sido contemplada adicionalmente. Muchas veces, si no

se ha prestado la atención debida el sistema de scraping desarrollado está totalmente acoplado con las fuentes y multiplica el trabajo. Estos servicios, pese a ofrecer mucha menos flexibilidad, nos permiten implementar pequeños scrapers que ayudan a profundizar en las fuentes de información de un modo muy rápido. Es posible, incluso, implementar sistemas de scraping parciales en los que este servicio es el encargado de recopilar la información mientras que el resto de los sistemas se conectan para obtener los datos en claro y procesarlos

Algunos productos y servicios existentes en el mercado

Entre los **productos** más populares están los siguientes:

- **Selenium**: suite de productos orientada a la automatización de tareas realizadas con navegadores. Incluye un IDE que permite crear scripts y programas orientados, entre otras funcionalidades, a crawling y scraping. Puede asimismo ser invocado desde otros programas, a través de varios lenguajes de programación. www.seleniumhq.org
- **GreaseMonkey**: extensión para navegadores http://www.greasespot.net/
- **iMacros**: es un producto con funcionalidad para automatización de navegadores, pruebas de páginas web y scraping de datos. http://imacros.net/
- **cURL**: Proyecto que incluye la librería libcurl y una herramienta orientada a transferir datos con la sintaxis URL. Es muy utilizado desde el lenguaje PHP, usando el módulo PHP/CURL. http://curl.haxx.se/
- **HTMLUNIT**: un navegador web sin interfaz (“GUI-less”) para programas desarrollados en lenguaje **Java**. htmlunit.sourceforge.net/
- **Scrapy:** framework para hacer Scraping con el lenguaje de programación Python http://scrapy.org/
- **Frontera:** es un framework que se utiliza para gestionar la lógica y las políticas a seguir al leer websites por medio de un bot: orden, prioridades, frecuencias de visitas, comportamientos en páginas concretas. Desarrollado por http://scrapinghub.com/ para proyectos con Scrapy aunque su planteamiento es tecnológicamente agnóstico.
- **CasperJS:** es una aplicación para automatizar tareas y realizar pruebas que se integra con el navegador PhantomJS[112], orientado a la navegación automática. http://caspersjs.org

[112] Webkit PhantomJS http://phantomjs.org/

Existen en el mercado varios **servicios**, entre ellos destacamos los siguientes:

- **DiffBot**: extrae contenido estructurado de artículos, páginas de productos de sites de comercio electrónico, imágenes, vídeos. También analiza páginas completas determinando qué tipo de páginas son y qué contenido tienen. http://www.diffbot.com/
- **Kimono**: se instala como extensión de un navegador, permite interactuar con la página de la que se quiere extraer los datos y convierte esa interacción en una interfaz (API) que alojan en su propio cloud y ponen a disposición de los usuarios para que sean ejecutadas por ellos mismos o para que planifiquen su ejecución para que Kimono realice el scraping en el momento deseado. https://www.kimonolabs.com/
- **CrowdFlower**: plataforma de crowdsourcing que permite a gran escala extraer, recoger, categorizar y mejorar datos. Asimismo permite crear y moderar el contenido así como realizar búsquedas relevantes y análisis de sentimiento en los contenidos. http://www.crowdflower.com/
- **Connotate**: Connotate presenta una visión compatible con una plataforma de Vigilancia Tecnológica con las capacidades avanzadas que presentamos en este libro. Destaca la funcionalidad de Web Scraping por su escalabilidad, la aplicación de inteligencia artificial para facilitar la comprensión de cómo están diseñados las páginas web. http://www.connotate.com
- **WorkFusion**: este producto reúne capacidades de recolección, limpieza y enriquecimiento de datos, machine learning, workflow y ejecución de tareas. http://workfusion.com/
- **Import.io:** Permite leer páginas web y convertirlas con facilidad en tablas de datos que exportar a otros formatos. Dispone de un API para integrar su servicio con tu propia aplicación y dispone de integraciones con terceros. Es uno de los más populares en la actualidad. https://import.io/

4. Modelo de Información: los Módulos de Entidades Estructurales de Información

En los Sistemas de Información Empresariales nos encontramos con grandes aplicaciones especializadas, como pueden ser por ejemplo el ERP *(Enterprise Resource Management),* el CRM *(Customer Relationship Management)* o el SCM *(Supply Chain Management).* Cada uno de ellos implementa fundamentalmente un conjunto diferenciado de procesos de negocio. El CRM se enfoca principalmente a los procesos comerciales, de marketing y de servicio post-venta. El SCM se enfoca a los procesos logísticos, de gestión de la cadena de suministro. En el ERP nos encontramos tanto un

enfoque a los procesos financieros como otro enfoque más generalista, dependiendo del fabricante, intentando cubrir todos los procesos de la empresa o institución.

Grandes fabricantes de software, como SAP, Oracle o Microsoft tienen diferentes implementaciones de cada una de estas piezas. Sin embargo, si descendemos en el análisis de las entidades de negocio que se reflejan en cualquiera de estas implementaciones nos encontramos con que tienen **un conjunto de Entidades de Negocio comunes** en cada aplicación.

Aplicando esta misma idea a los Sistemas de Vigilancia e Inteligencia nos encontramos con **una serie de Entidades de Negocio en torno a las cuales se pueden caracterizar dichos conceptos de Vigilancia e Inteligencia**, en general de forma independiente del sector a vigilar o las necesidades y casos de uso que determinarán los objetivos del sistema. Asimismo nos encontramos con un conjunto de entidades de negocio que tienen más ocurrencias y más relaciones que el resto, y que denominaremos **Entidades Principales**, frente al resto que denominaremos **Entidades Secundarias**. Asimismo, del resultado del estudio de las **relaciones entre las Entidades** también nos aparece una posible clasificación de las mismas.

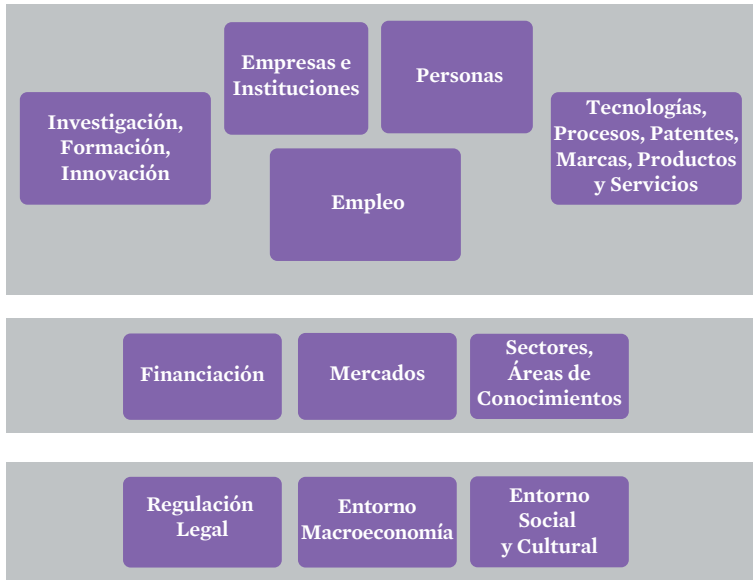
Esta idea resulta relevante por varias razones:

- Nos permite obtener, a partir del superconjunto de Entidades y Relaciones, una **estructura a priori del Sistema de Vigilancia** con la que generar herramientas orientadas a obtener de forma productiva una especificación del Sistema de Vigilancia correcta y completa. Por ejemplo se propone construir una plantilla a priori de casos de uso que presentar como “casos de uso candidatos” a los entrevistados.
- Nos permite **estructurar de forma coherente y sostenible en el tiempo la interfaz de usuario.** El conjunto de Entidades de Negocio principales puede constituir una estructura que puede ser la base para módulos funcionales, menús, cintas o pestañas, dentro de la cual se acceda a la información resultado de implementar los Casos de Uso.
- Nos permite **estructurar a priori las tareas de análisis** de fuentes de datos, taxonomías y casos de uso a incorporar al sistema.
- Nos permite **definir un conjunto de funcionalidad a priori que podría necesitar ser resuelta en el sistema**, posiblemente mediante las aplicaciones de nueva generación que se postulan en este libro (procesamiento de lenguaje natural, machine learning, etc).
- Nos permite visualizar desde el principio el posible recorrido a largo plazo del Sistema de Vigilancia.

4.1. Las Entidades Principales

Las Entidades Principales en torno a las que se estructuraría el Sistema de Vigilancia son las siguientes:

- Empresas e Instituciones.
- Tecnologías, Procesos, Patentes, Marcas, Productos y Servicios.
- Mercados.
- Sectores, Áreas de Conocimiento.
- Personas.
- Empleo.
- Financiación.
- Investigación, Formación e Innovación.
- Regulación legal.
- Entorno Macroeconómico.
- Entorno Social y Cultural.



Podemos encontrar referencia de las mismas en el apartado de 2. Vigilancia Estratégica e Inteligencia Competitiva. En los siguientes apartados trataremos y justificaremos las razones para realizar las agrupaciones que estamos haciendo.

Empresas e Instituciones

Posiblemente la entidad de negocio que está implicada en más casos de uso sea la entidad "**Empresas**". Datos de empresas, productos de empresas, trabajadores, clientes o proveedores de una empresa son algunos ejemplos de información que se requiere sobre esta entidad de negocio.

Un caso especial es la entidad "**Instituciones**", como una manera de diferenciar las instituciones del estado, autonómicas o locales de las Empresas. Asimismo hay que tener en cuenta que dentro del sector público también existen diversos tipos de organismos como empresas públicas, fundaciones, agencias, etc.

Dependiendo del sector a vigilar y los objetivos definidos para el sistema se deberán tratar o no estas entidades como la misma entidad con campos para diferenciarlos o como diferentes entidades.

Tecnologías, procesos, patentes, marcas, productos y servicios

Presentamos de forma agrupada estas 6 entidades de negocio por un lado por su íntima relación y por otro porque en algunos sectores es frecuente encontrarse con confusiones, indefiniciones o mezclas de términos entre dichas entidades. Los contenidos de las páginas web o los folletos de marketing van a estar repletos de información que va a ser objeto de vigilancia pero que no necesariamente va a ser estricta en el uso conceptual de las entidades de negocio que estamos tratando en este apartado. Posiblemente sea todo lo contrario. En el Sistema de Vigilancia Avanzado que proponemos esta información va a ser objeto de tratamiento detallado con el objetivo de extraer información estructurada.

El primer concepto base es el de **tecnología**. Extendiendo la definición que dábamos en el apartado 2. Entenderemos como tal a un conjunto de conocimientos interrelacionados pudiendo incluir procedimientos de construcción, funciones, características y diferentes componentes específicos, que pueden ser a su vez tecnologías, productos, tener marcas... En definitiva pueden aparecer diferentes relaciones entre los componentes de la misma que dificulten la clasificación del concepto tecnológico en sí.

Los **procesos** deben ser entendidos como actividades y procedimientos novedosos y aplicables a la industria. Se encuentran asociados a productos y a servicios.

La **patente** es uno de las entidades de negocio que más a menudo es objeto de vigilancia. Se refiere al conjunto de derechos que se otorgan a una persona o a una

empresa para la explotación comercial de una tecnología o un producto desarrollado o inventado por ellos y sobre el que se tiene propiedad industrial e intelectual. Tanto las tecnologías como los procesos pueden ser objeto de patente. Sin embargo nos podemos encontrar con situaciones de mercado en los que los productos y servicios no puedan ser objeto de patente, no exista tradición de ser patentados o incluso que pueda ser patentado en unos países y no en otros, como ocurre con el software.

Relacionado con los conceptos de patente y producto nos encontramos con las **Licencias**, el permiso a un tercero para utilizar a cambio de una contrapartida un producto sobre el que se tienen unos derechos de autor (en inglés "copyright"), es decir los derechos que asisten a los autores de una obra

Un tipo particular de licencias son las **Licencias Creative Commons.** Presentan 4 características que un autor puede solicitar para una obra: reconocimiento de la autoría, limitación o no a usos no comerciales, la autorización o no para hacer obras derivadas, y la obligación o no de mantener la misma licencia en caso de hacer obras derivadas.

Una vez que dispongamos de la patente, el siguiente paso natural será definir un **Producto**, que podría utilizar la tecnología patentada y consecuentemente la patente en cuestión. En torno a la definición de los productos nos podemos encontrar con varias casuísticas: usar varias patentes, no usar ninguna, usar varias tecnologías o solo una, los diferentes tipos de licencias creative commons, etc.

Asimismo, comúnmente van a ser necesarios una serie de **Servicios** asociados a los productos, que pueden estar asociados al producto de forma unívoca o ser servicios más genéricos que estén relacionados con varios productos a la vez.

Será especialmente común mezclar los conceptos de tecnología con producto y proceso con servicio.

En todo caso lo que sí es habitual es realizar una solicitud de **Marca**. Con esa marca se puede estar denominando indistintamente al producto, al servicio, al proceso o a la tecnología, especialmente una vez que los departamentos de marketing y comerciales entran en los procesos de promoción y venta. Es significativo también el hecho de que lo normal va a ser que el nombre de la patente y el nombre de la marca sean diferentes.

Deberemos por tanto aclarar al inicio del proyecto de definición de un Sistema de Vigilancia la situación específica del sector o sectores a vigilar en cuanto al uso de las entidades de negocio de este apartado, qué tipo de expectativa podemos tener de los textos que vamos a tratar.

Deberá investigarse si se dispone de taxonomías y ontologías para clasificar la información de la entidad de negocio así como si disponen o no del nivel de desagregación adecuado a las necesidades, casos de uso y objetivos de vigilancia. Caso de existir aparece una oportunidad importante en integrarlas y en extenderlas al nivel de granularidad necesario, consiguiendo información importante del cruce de información entre las fuentes y repositorios de información una vez integrados.

Mercados

El concepto **"Mercado"** responderá a un conjunto de **transacciones económicas**, las empresas e individuos que participan en dicho conjunto de transacciones y los conceptos económicos que se utilizan para estudiar y definir dicho mercado.

Se incluirán transacciones económicas realizadas y transacciones económicas de las que se tenga constancia que se van a realizar o podrían realizar. Por ejemplo quedarían incluidas las transacciones anunciadas en una Plataforma de Contratación, pública o privada o las fusiones o adquisiciones de empresas anunciadas en un periódico económico.

Será especialmente importante la detección de información relativa a conceptos importantes a vigilar casi siempre: **Clientes, Proveedores y Competidores**. Normalmente quedarán clasificados bajo la entidad de negocio "Empresa", aunque también se detectarán indirectamente, a través de marcas o productos por ejemplo.

Sectores y Áreas de Conocimiento

El concepto "**Sector**" responderá a un subconjunto de actividad económica y los criterios mediante los que se define dicho subconjunto, frecuentemente un "**Área de Conocimiento**". Es por ello que se propone la creación de una ontología que reúna los conceptos de Sector, los criterios de definición y Área de Conocimiento referida.

Se dispondrá tanto de información de sectores de los que se disponga de algún tipo de organización oficial como de aquellos sectores que no estén oficialmente reconocidos como tal pero de los que se disponga de información.

Personas

Las personas están asociadas a todas las entidades de negocio: empleados de la empresa, profesor de un centro de formación, especialista certificado en una tecnología o en un producto, profesionales del sector, usuarios de un producto, business angels inversores en startups, etc.

Será necesario realizar un estudio detallado para entender e implementar cómo las personas participan en el sector objeto de vigilancia. Se propone para este análisis la utilización del concepto de stakeholder, es decir, todos los interesados, que pueden influenciar, ser influenciados o ser considerados parte del sector.

Actualmente la información disponible en Redes Sociales como Twitter, LinkedIN, Facebook o Infojobs, puede ser una fuente muy valiosa para la obtención de información estratégica de las personas relevante para nuestro Sistema de Vigilancia.

Empleo

La Entidad de Negocio Empleo es un hub que conecta múltiples entidades de negocio. Debe valorarse si, para los objetivos del sistema, es una entidad de negocio estructural o si se incluye como asociada a otras, como puede ser “Personas” o “Empresas”. Es una entidad muy importante de cara a los agentes del sector encargados de definir y poner en marcha políticas públicas y en sectores emergentes o muy dinámicos.

Recoge información de ofertas de empleo, plataformas y agregadores de ofertas de empleo y otras como perfiles demandados y perfiles existentes en el sector o formaciones demandadas en ofertas de empleo.

Financiación

Bajo la entidad de negocio “Financiación” se propone recoger toda la información disponible sobre:

- **Financiación pública** otorgada a la empresa, incluyendo ayudas públicas, programas de financiación de proyectos de investigación o innovación... en general cualquier préstamo o subvención realizada por cualquier administración pública nacional o internacional, siempre que sea relevante para el objeto de vigilancia.
- **Financiación privada.** En este punto se recoge la información sobre inversiones e inversores privados tanto de empresas como de personas (“Business Angels”).

Se propone recoger aquí también información sobre fusiones y adquisiciones. La obtención de crédito y los diferentes tipos de capital privado existente (capital semilla, capital riesgo, etc) también quedaría referenciada aquí.

Formación, Investigación e Innovación

La razón principal de unir estas entidades de negocio es que es muy habitual en numerosos sectores que la investigación sea realizada por profesores en centros de formación universitarios, vinculados a un grupo de investigación.

En sectores más maduros es habitual encontrar Institutos de Investigación y Centros de Innovación tanto públicos como privados independientes de la universidad, aunque también suelen contar con colaboraciones, convenios y partenariados explícitos con la Universidad.

Asimismo también es cada vez más común encontrar en la empresa privada con departamentos de I+D y grupos de innovación, cuya información también proponemos organizarla bajo este epígrafe.

Son notables las relaciones con varias Entidades Principales, como “Financiación”, por ejemplo los programas de financiación de I+D, y “Sectores y Áreas de Conocimiento”, por ejemplo las Áreas de Conocimiento sobre los que investiga un Centro de Investigación o sobre los que forma un departamento universitario.

Entorno Macroeconómico

Explícitamente apenas vamos a encontrar referencias a la palabra macroeconomía. Sin embargo sí que vamos a encontrar información sobre el entorno macroeconómico que sea relevante para los objetivos de nuestro Sistema de Vigilancia.

Nos interesará la información sobre el estado del mercado a nivel global, el mercado de capitales, los precios de los recursos relevantes a nivel mundial y la infraestructura económica del mercado en el sector objeto de vigilancia.

Esto se concreta en la fase en la que la economía se encuentre, el sentimiento general de mercado, el nivel de desempleo, el producto interior bruto, el estado del mercado de capitales, la facilidad de encontrar financiación en el sector y el coste que tiene disponer de él, el estado de los mercados en materias primas, commodities y recursos básicos necesarios para el sector, la facilidad y el precio de obtenerlos y finalmente en

los servicios públicos que estén disponibles en el mercado: enseñanza, transportes, libertad y calidad de acceso a clientes y proveedores, nivel de impuestos y calidad de vida en general.

Deberá analizarse a nivel estratégico cómo el entorno macroeconómico impacta en nuestros objetos de vigilancia.

Una buena referencia para éste apartado y el siguiente es el libro "Business Model Generation", de Alexander Osterwalder & Yves Pigneur, en su apartado "Business Model Environment: Context, Design Drivers and Constraints".

Entorno Socieconómico y Cultural

Cada sector tiene un conjunto de tendencias sociales, económicas y culturales que le afectan. Incluye desde las tendencias demográficas, la distribución de la riqueza, las rentas disponibles, los patrones de gasto hasta el porcentaje de población urbana frente a población rural. Deberán determinarse en los procesos de entrevistas estructuradas el conjunto de tendencias a vigilar y cómo influyen en las Entidades de Negocio y en el entorno de Vigilancia en general.

Regulación Legal

Cada sector está regido por una regulación legal, un conjunto de leyes y reglamentos cuya estabilidad, cambio o tendencia puede influir en nuestros objetivos de Vigilancia.

En este apartado recogeremos información sobre tendencias legales, novedades, esfuerzos por parte de lobbies que pueda representar un cambio en el sector, y cómo esos cambios pueden beneficiar, perjudicar o sencillamente influir en nuestro sector y su entorno de Vigilancia.

4.2. Las Relaciones Estructurales

Entre las Entidades Principales existen una serie de relaciones habituales entre ellas y que responden directamente a requisitos del Sistema. Concretamente estarían las siguientes:

- **Relación Reflexiva con la misma Entidad de Negocio**. Por ejemplo, una empresa cliente o proveedora de otra.

- **Información detallada sobre la Entidad de Negocio**. Mostrar uno o varios de los campos de la Entidad.
- **Búsqueda sobre la Entidad de Negocio.** Filtrar la Entidad de negocio según determinados criterios.
- **Relaciona una Entidad de Negocio con otra.** Por ejemplo Productos y Servicios que ofrece una Empresa.
- **Búsqueda de Entidades de Negocio relacionadas.** Muestra las Entidades de Negocio relacionadas con otra y la relación en sí.
- **Evolución temporal de una Entidad de Negocio.** Presenta cómo ha ido evolucionando una Entidad de Negocio a lo largo del tiempo.
- **Evolución dimensional de una Entidad de Negocio.** Además de por la dimensión tiempo, una Entidad de Negocio puede variar o evolucionar a lo largo de otra dimensión.
- **Escenarios:** evaluar la proyección de una Entidad de Negocio ante un escenario definido.

Tener en cuenta estas relaciones como parte del sistema permite incluir como predeterminados procedimientos de creación, consulta, inserción y borrado de instancias de las entidades.

4.3. Alineando Entidades Principales, Relaciones Estructurales y Ontologías

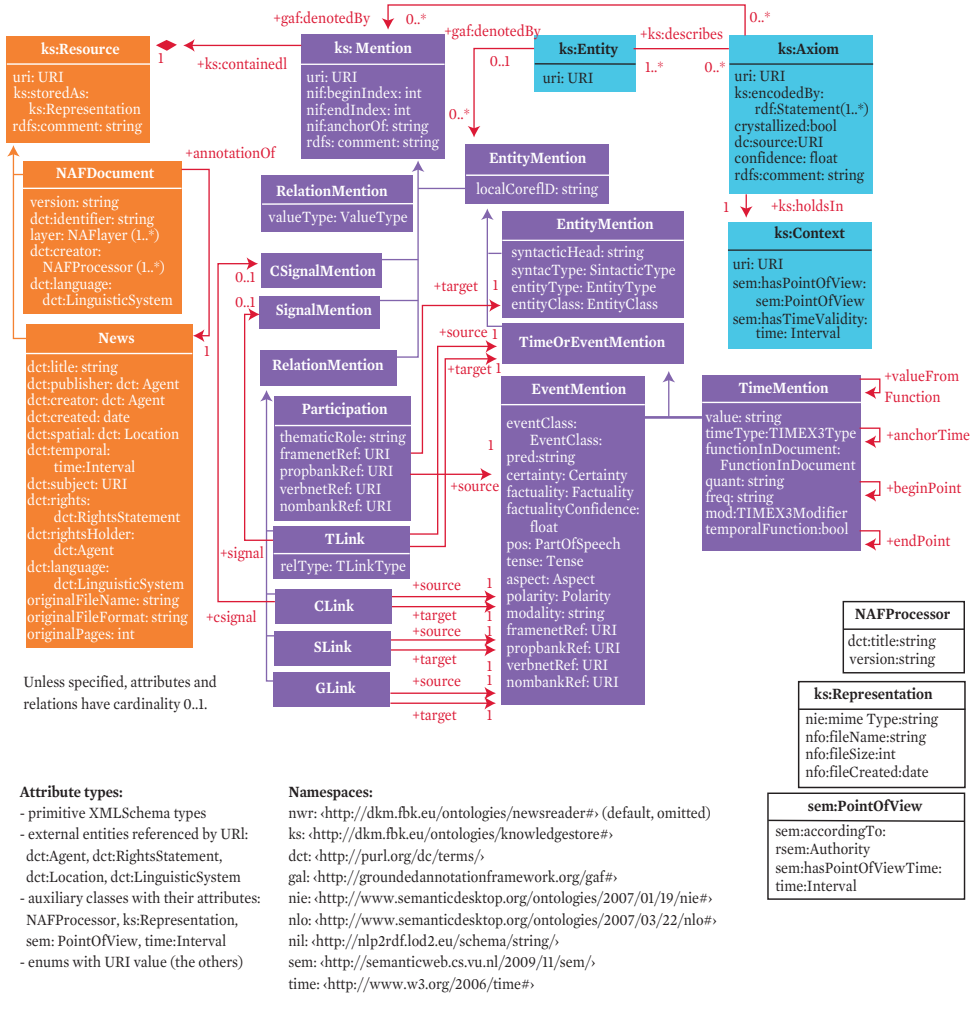
Una línea interesante a explorar es integrar conceptualmente los **objetos**, sus **relaciones** y las **propiedades** de las Ontologías públicas con las **Entidades Principales y sus propiedades y las Relaciones Estructurales.**

El **Modelo de Datos** se convierte así en un **Conjunto de Ontologías interconectadas**, como veíamos en el apartado de "Ontologías, Datos Enlazados y Web Semántica". En las conexiones tendríamos en cuenta el concepto de **Relaciones Estructurales** que presentamos en el punto anterior y los de Dominio y Rango que también presentamos en el apartado de Ontologías.

Para tener en cuenta la evolución en el tiempo de las Entidades de Negocio y la Veracidad de la información estaríamos incluso hablando de una **Red de Ontologías**.

De esta manera podemos relacionar por ejemplo la ontología FOAF con **Personas**, Good Relations con **Productos** o la Organization Ontology org con **Empresas e Instituciones** y consecuentemente usar sus relaciones estandarizadas, al igual que las de SKOS.

La imagen que sigue a continuación del documento "Knowledge store versión 2" del proyecto europeo NewsReader[113] ilustra a la perfección la idea. Se refiere al modelo de datos del Repositorio principal Big Data. Puede verse en las propiedades de los objetos la utilización de numerosos namespaces referenciando Ontologías públicas.



Attribute types:
- primitive XMLSchema types
- external entities referenced by URl: dct:Agent, dct:RightsStatement, dct:Location, dct:LinguisticSystem
- auxiliary classes with their attributes: NAFProcessor, ks:Representation, sem: PointOfView, time:Interval
- enums with URI value (the others)

Namespaces:
nwr: ‹http://dkm.fbk.eu/ontologies/newsreader#› (default, omitted)
ks: ‹http://dkm.fbk.eu/ontologies/knowledgestore#›
dct: ‹http://purl.org/dc/terms/›
gal: ‹http://groundedannotationframework.org/gaf#›
nie: ‹http://www.semanticdesktop.org/ontologies/2007/01/19/nie#›
nlo: ‹http://www.semanticdesktop.org/ontologies/2007/03/22/nlo#›
nil: ‹http://nlp2rdf.lod2.eu/schema/string/›
sem: ‹http://semanticweb.cs.vu.nl/2009/11/sem/›
time: ‹http://www.w3.org/2006/time#›

[113] Proyecto NewsReader http://www.newsreader-project.eu/

5. NoSQL: las Bases de Datos del Big Data

Big Data le ha dado relevancia a varios tipos de bases de datos diferentes a los tradicionales Sistemas de Gestión de Bases de Datos Relacionales. Una de las características más destacadas y conocidas de este tipo de bases de datos es su lenguaje de consulta, denominado SQL (siglas del inglés *"Structured Query Language"*). A las bases de datos del Big Data se les ha agrupado bajo el nombre de "*NoSQL*", que significa "*Not only SQL",* (no sólo SQL), siendo esto una manera de comunicar y destacar que existen tipos de datos diferentes a las Relacionales.

Lo primero que tenemos que tener en cuenta es que utilizar una base de datos NoSQL en nuestro proyecto no es una decisión banal. Existe un gran número de profesionales formados en el uso de bases de datos relacionales y el lenguaje SQL pero no es así con las bases de datos NoSQL, teniendo en cuenta además que no se trata de un grupo homogéneo y que poco van a tener que ver entre sí una base de datos orientada a documentos y una base de datos orientada a grafos. Son tecnologías emergentes, con componentes relativamente nuevos, sujetos a un ritmo importante de evolución y por tanto de cambio.

Para muchos proyectos una base de datos relacional puede ser una solución suficientemente buena, en particular es siempre la más adecuada cuando tenemos que realizar transacciones, con datos tabulares, como en entornos financieros, en la gestión de compras y ventas.

Todos estos elementos constituyen lo que en Dirección de Proyectos denominamos **Riesgos de Proyecto** y en proyectos IT este tipo de riesgos pueden conllevar Planes de Gestión de Riesgos muy costosos por el alto impacto que la ocurrencia de los mismos puede tener en el éxito y la sostenibilidad del proyecto.

Algunas de las razones que pueden aconsejar la elección de una base de datos NoSQL en lugar de una Base de Datos Relacional tradicional son:

- Facilidad a la hora de gestionar la **Variedad de los datos** de la que disponemos, por ejemplo cuando en cada inserción de datos la información a almacenar tiene campos distintos.
- Necesidad de gestionar y mantener **grandes Volúmenes de datos** (terabytes a petabytes) especialmente en picos de uso del sistema.
- Problemas con la **Velocidad** al usar bases de datos relacionales: cuando tenemos picos de uso del sistema que nos provocan **problemas operativos**, por ejemplo cuando los datos llegan a una velocidad superior a la que podemos gestionar

- Cuando la base de datos relacional nos presenta **problemas de escalabilidad** tanto técnica como económica (costes de licencias, replicación en diferentes centros de datos, cloud computing, etc).

- La **complejidad** de las consultas es superior a la que podemos gestionar con una Base de Datos Relacional. Esta situación se mejora con un tipo de bases de datos relacionales que funcionan en paralelo pero una BD NoSQL puede ser mejor solución.

- Hay **alta concurrencia** en las consultas a la base de datos o son **muy intensivas en el uso de la CPU**.

- Cuando a alguna de estas situaciones le acompaña que el tipo de datos que se gestiona **coincide con las especialidades NoSQL** que hemos presentado: grafos, documentos o columnas.

5.1. Tipos de Bases de Datos NoSQL

Como se adelanta en el apartado 3.7.4 "NoSQL, las Bases de Datos del Big Data", cuatro son los principales grupos de bases de datos NoSQL que se destacan en el mercado: "Clave-Valor", orientadas a Columnas, orientadas a Documentos y orientadas a Grafos. Pasamos a continuación a presentarlos, cada una con sus pros y sus contras:

Bases de Datos "Clave-Valor" (Key-Value)

Este tipo de bases de datos destacan por su alta escalabilidad. Abarcan bien proyectos con **textos estructurados y semiestructurados, datos de redes sociales, logs de servidores web y la mayoría de los datos orientados a negocio**, por lo que este tipo de bases de datos son de las más utilizadas en proyectos Big Data. También son utilizadas cuando se realizan grandes volúmenes de escrituras en múltiples nodos o cuando se realizan analíticas a gran escala en grandes clusters.

Encajan muy bien en proyectos en los que **la escritura se realiza una única vez y se realizan muchas lecturas**. Consecuentemente necesitan almacenar la información y recuperarla a alta velocidad.

Son utilizadas por ejemplo en situaciones como la de **gestión de las sesiones de usuario**, caracterizadas por acceso rápido a lecturas y escrituras y por no necesitar durabilidad de los datos. Otro ejemplo muy claro es la **participación en una red social,** (Facebook, Twitter...) que es escrita una única vez por el usuario autor y es leída a continuación por sus seguidores, amigos o el concepto que esté implementado en la red social.

Bases de Datos Columnares u Orientadas a Columna

Son muy utilizadas en **entornos analíticos como el OLAP** (On Line Analytic Processing, los famosos "*cubos*" tan utilizados en entornos financieros y de marketing), tradicionales en el mundo del Business Intelligence, desde el que ha evolucionado el Big Data. Múltiples situaciones de Vigilancia requieren tareas de análisis de información.

Este tipo de Bases de Datos están orientadas a las columnas de datos, en lugar de a los registros como las bases de datos relacionales. Cuando escribimos el registro de una transacción, estamos escribiendo los valores de todos los campos: nombre del comprador, importe de la transacción, nombre del vendedor, el producto vendido, el número de unidades, etc. Las Bases de Datos Columnares estarían orientadas a gestionar de golpe todos los valores de los Compradores, que estarían en la misma columna, de ahí el nombre. De hecho el origen de algunas Bases de Datos Columnares es la necesidad de almacenamiento de columnas de datos.

Son muy utilizadas cuando se ejecutan trabajos tipo **MapReduce**, cuando hay que **actualizar y almacenar registros únicos**, como por ejemplo todo el histórico de relación con un cliente. También son buenas ejecutando el cálculo de métricas de una columna o un conjunto de columnas. En cambio, si han de analizar o escribir filas, es decir, registros, su rendimiento no es bueno.

Uno de sus precursores más relevantes es la base de datos Google Big Table, de ahí que a este tipo de bases de datos también se les llama Big Table, además de "Orientadas a Columnas".

Bases de datos Documentales

En las Bases de datos Documentales cada documento es tratado como un único registro. Gestionan muy bien **texto no estructurado** y particularmente bien **texto semiestructurado**, es decir, texto codificado según un esquema conocido, como XML, JSON, YAML, PDF, e-mail o incluso documentos ofimáticos.

Son muy buenas por tanto en recuperación de conocimiento o temas incluidos en grandes conjuntos de informes y documentación o búsqueda de e-mails. La búsqueda puede ser facilitada añadiendo metadatos, claves y lenguajes específicos dependientes del modelo de base de datos utilizado.

Concretamente son muy utilizadas para búsqueda de patentes, búsqueda de precedentes legales, búsqueda de papers científicos y datos experimentales. Asimismo son

muy buenas para integrar diferentes fuentes de datos que pueden residir en tipos de bases de datos incompatibles.

Los programadores las consideran bases de datos amigables, que permiten un modelado de datos natural y desarrollo rápido. Además encajan muy bien con el paradigma programación orientada a objetos, posiblemente el paradigma dominante en la actualidad.

Se considera que tienen su primera referencia en la aplicación Lotus Notes, creada a final de los años 80. Lotus fue adquirida por IBM en 1995.

Bases de datos orientadas a Grafos

Este tipo de bases de datos están inspiradas por los trabajos de Leonhard Euler y la teoría de grafos. Son especialmente útiles **cuando los datos están muy interconectados y no son tabulares**, en cuyo caso ofrecen un gran rendimiento. Se utilizan en todo tipo de aplicaciones relacionadas con la **Web Semántica**, con **Ontologías**, y son también muy utilizados en almacenamiento de imágenes y cuando en los datos están implicados **algoritmos sustentados en la teoría de grafos**. Enlazan rápidamente **personas, productos, compras y calificaciones**, por ejemplo.

Deben permitir ejecutar como transacciones únicas cualquier consulta que explote las relaciones entre entidades. En una base de datos relacional, para buscar relaciones entre Entidades de Negocio relacionadas, tendríamos que ir paso a paso, relación a relación, ejecutando búsquedas. En una Base de Datos orientada a Grafos ésto se ejecutaría en una única transacción. Se utilizan lenguajes especialmente diseñados, como SPARQL.

Se pueden usar como una base de datos de propósito general pero requiere un cambio de paradigma a la hora de diseñar las relaciones entre los datos ya que sólo ofrecen buen rendimiento con datos muy interconectados.

6. Funcionalidades, Implementaciones e Interfaces Big Data para los Sistemas de Vigilancia e Inteligencia

En este apartado nos centraremos en primer lugar en un conjunto de Funcionalidades que nos hemos encontrado que son denominador común para la solución de muchos casos de uso de Vigilancia e Inteligencia Competitiva.

Presentamos a continuación algunas implementaciones e interfaces de dichas funcionalidades así como algunas de las empresas implementadoras.

No es objetivo presentar aquí plataformas, empresas o productos estrictamente de Vigilancia. Ponemos el foco en la accesibilidad del lector a los servicios y la información y lo adecuado que consideramos la interfaz de usuario, y lo significativos que son los productos de información que ofrecen, en comunicar e ilustrar al lector.

6.1. Funcionalidades de Vigilancia

Al analizar y diseñar las soluciones a los Casos de Uso de los proyectos de Vigilancia e Inteligencia Competitiva nos encontramos con un conjunto de **Funcionalidades** que son denominador común a muchos de ellos. Estas Funcionalidades comunes van a ser el pilar en torno a los que estructuramos este apartado. Hay que reseñar que estas Funcionalidades aparecen a nivel de requisito, es decir tras realizar el proceso que proponemos de convertir cada Caso de Uso en una colección de Requisitos, como en el ejemplo que planteamos en el punto 4.2 "Casos de uso y necesidades de Vigilancia".

En una implementación de este sistema este conjunto de funcionalidades debería convertirse en un *API* (*Application Program Interface*), una **librería de aplicaciones con funciones** que puedan ser llamadas desde un lenguaje de programación o integradas en entornos de desarrollo.

Se presentan a continuación el conjunto de Funcionalidades resultado del análisis, en los siguientes puntos ilustramos cada una de ellas con propuestas de interfaces tomadas de varios implementadores:

1. **Relacionados**: busca Entidades de negocio relacionadas con otra dada, por ejemplo "sectores, áreas de conocimiento, productos o servicios" en los que está trabajando una empresa dada" o también entre sí, por ejemplo empresas partners de negocio de una empresa dada. Otro ejemplo relevante: se aplica a la búsqueda de los stakeholders relacionados con una Entidad de Negocio dada.
2. **Principales**: similar a "Relacionados" pero destacando cuales son las Entidades más importantes. Será importante determinar qué significa en cada caso el concepto "más importante", tanto poder aplicar criterios de clasificación como tener criterios predeterminados que aplicar, como los que aplica el buscador de Google.
3. **Histórico**: presenta información sobre como una Entidad de Negocio ha evolucionado según la dimensión tiempo. Por ejemplo se puede aplicar para estudiar el

impacto en el tiempo de un evento concreto en una empresa, por ejemplo ganar un concurso o la salida a mercado de un producto.

4. **Escenarios**: responde a la necesidad de conocer el posible impacto de un evento o un cambio de circunstancias en una entidad de negocio. Por ejemplo se puede aplicar para estudiar escenarios de fusiones y adquisiciones, de escasez de un componente de un producto, de subida de precios en el petróleo para el sector transporte, para determinar si una empresa es capaz o no de aprovechar o no una subvención.

5. **Tendencias**: estudia los atributos y entidades de negocio relacionadas con una Entidad de Negocio dada para determinar qué cambios en el entorno son determinantes, especialmente los más recientes. Un ejemplo sería determinar qué tendencias se aprecian en la Financiación de las empresas.

6. **Buscador avanzado**: permite realizar búsquedas inteligentes, filtrando y priorizando según las entidades de negocio disponibles en las Bases de Datos.

7. **Navegador Inteligente**: permite navegar a través del conocimiento albergado en bases de conocimiento Big Data. El uso de ontologías y específicamente de los enlaces entre las diferentes ocurrencias de las Entidades de Negocio permiten esta funcionalidad y mediante ella realizar varios tipos de análisis, descubrimiento de resultados y obtención de conclusiones.

8. **Correlación de Eventos / Alarmas:** establece la generación de algún tipo de aviso ante la aparición de algún evento cuya condición es posible programar o configurar previamente. Esta funcionalidad es habitual en cualquier sistema de inteligencia de negocio.

9. **Bases de Datos de Entidades de Negocio:** es frecuente la solicitud de obtener "toda" la información disponible sobre una entidad de negocio dada, por ejemplo papers de investigación, patentes o información legal asociada a un producto o servicio

10. **Comparador:** busca información de entidades de información de diferentes fuentes y presenta su comparación. Por ejemplo se utiliza para buscar información sobre empresas de otros países y dar como salida información comparativa con las empresas españolas

11. **Análisis de Veracidad**: contrastar un conjunto de información disponible sobre una Entidad de Negocio, por ejemplo la presentada por una empresa al solicitar una ayuda pública., o los supuestos en los que se basa un proyecto o la solvencia de un grupo de empresas.

6.2. Algunos implementadores

Destacamos a continuación algunas empresas que implementan funcionalidades e interfaces Big Data. Mostramos tres de las que consideramos más reales y prometedoras junto a un representante del enorme conjunto de startups que están emprendiendo actividades en este sector.

Google

Google, como empresa, se puede considerar uno de los primeros sistemas Big Data. Varias de las innovaciones que implementaron son hoy la base de los conceptos Big Data que presentamos en el "Nuevas Capacidades Big Data". Además implementa per se muchas de las funcionalidades de Vigilancia y disponen de varias herramientas públicas que son estupendos ejemplos para ilustrar las funcionalidades, herramientas e interfaces que queremos presentar en este apartado. Al estar disponible online de forma accesible al gran público lo considero una excelente referencia.

Cuando tengo que explicar en qué consisten los Sistemas de Vigilancia con capacidades Big Data a personas ajenas al mundo TIC y al mundo de la Vigilancia, suelo usar la metáfora de que "es como un Google 2.0", es decir, un buscador al que añadimos las capacidades avanzadas e infraestructura avanzadas del capítulo 3. Google cuenta con una implementación similar pero de momento sin las funcionalidades y herramientas que nos interesan. Se trata del Google Search Appliance, una máquina con software de Google que se integra en los Centros de Proceso de Datos de las empresas y que indexa su documentación.

Wolfram Language

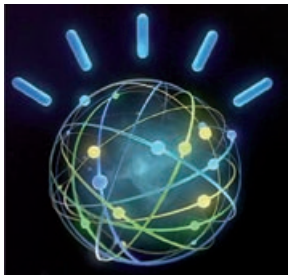
[114]Publicado en junio de 2013, Wolfram es el lenguaje de programación utilizado para el desarrollo de los productos de la empresa Wolfram Research. Es un lenguaje multi paradigma aunque lo presentan como un lenguaje basado en el Conocimiento.

[114] Todas las referencias de Wolfram son parte del Wolfram Language & System Documentation Center http://reference.wolfram.com/language/

Se usará en próximos apartados para la ilustración de varias funcionalidades y aplicaciones, destacando especialmente las relacionadas con Machine Learning y Data Science. Es muy ilustrativo, incluso espectacular.

Cuenta en su web con información y documentación online y cuenta asimismo con un acceso de prueba, un sandbox de aprendizaje que ilustran su espectacular potencial.

IBM Watson



Watson es la gran apuesta de IBM para liderar un grupo de sistemas que ha denominado Sistemas Cognitivos, en los que encajan totalmente el tipo de sistemas que estamos describiendo en este libro. Cuenta con dos líneas de productos que podrían converger en un futuro: Watson Analytics y Watson Content Analytics.

Cuenta con capacidades de procesamiento de lenguaje natural, tanto de comprensión como de generación, Machine Learning, y otras capacidades avanzadas como el aprendizaje automático, generación de hipótesis, representación de conocimiento y razonamiento automático. Internamente usa otras cuestiones que tratamos en este libro, como Fuentes de Información externas, taxonomías y ontologías públicas.

Se ha hecho muy popular en Estados Unidos por el Caso de Uso de Jeopardy!, un famoso concurso de televisión de preguntas y respuestas en el que ha participado y ganado.

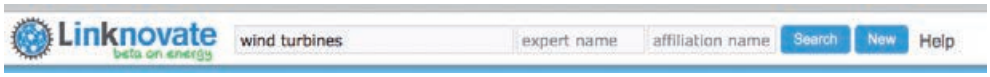
IBM ha hecho en estos últimos años adquisiciones en el ámbito del Business Intelligence, como Cognos y SPSS, que también aportarán valor añadido para configurar soluciones avanzadas de Vigilancia e Inteligencia Competitiva.

Linknovate

Linknovate[115] es una startup que nos proporciona una plataforma enfocada al sector de la energía. Recogen información de múltiples fuentes pero dan más valor a las

[115] Linknovate http://www.linknovate.com/

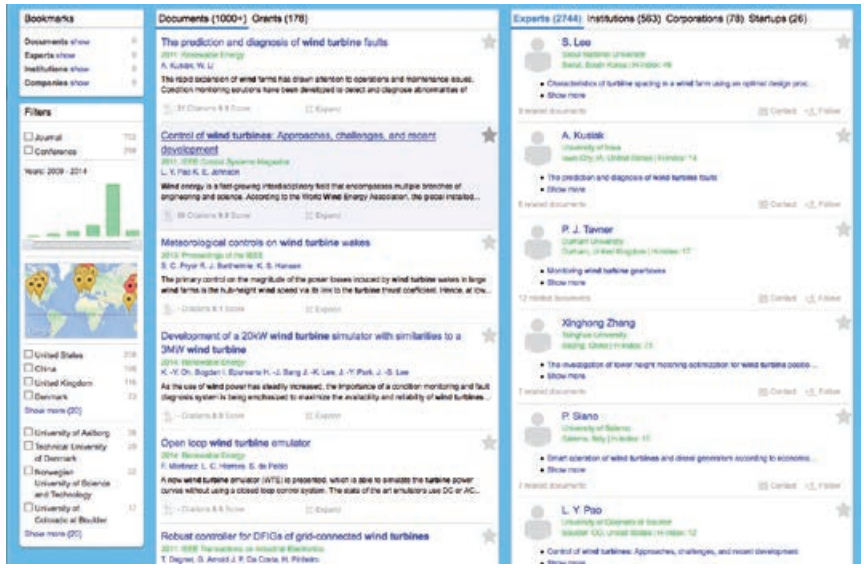
fuentes más recientes: conferencias, información de startups, publicaciones recientes, subvenciones.



Disponen de un buscador en el que localizar conceptos, fuentes, personas, empresas y tipos de empresas. Los resultados se presentan con una columna para filtros, otra para Documentos y Subvenciones y una tercera para Expertos, Instituciones, Empresas y Startups.

Nos permite filtrar según diversas dimensiones, como la fuente, el país o la universidad y llegar a la información de los documentos, subvenciones, expertos, instituciones y empresas.

De las Empresas consolidadas y los Expertos nos ofrece información adicional con palabras clave, expertos, documentos, publicaciones, conferencias y búsquedas relacionadas.



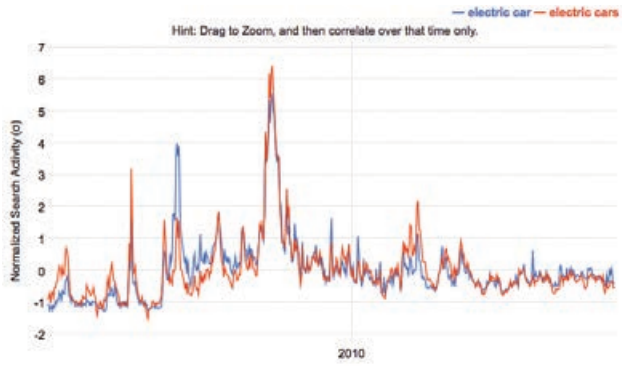
6.3. Implementaciones de las Funcionalidades de Vigilancia

Presentamos a continuación implementaciones de algunas de las Funcionalidades de Vigilancia que hemos identificado. Una imagen vale más que mil palabras...

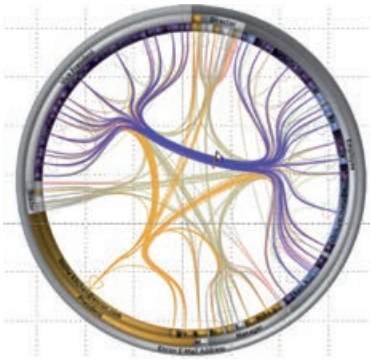
6.3.1. Relacionados / Principales

Las Funciones "**Relacionados**" y "**Principales**" la vamos a ilustrar con la herramienta **Google Correlate**, disponible en Internet en http://www.google.com/trends/correlate/ Esta herramienta permite realizar búsquedas de términos relacionados con un término dado. El criterio de si un término está más o menos relacionado está basado en la correlación entre las búsquedas. También es posible introducir tu propia serie de datos.

El algoritmo está documentado en http://www.google.com/trends/correlate/nnsearch.pdf El resultado es una lista ordenada de textos cuya frecuencia de búsqueda siga un patrón similar con el término que introducimos junto con el nivel de correlación.



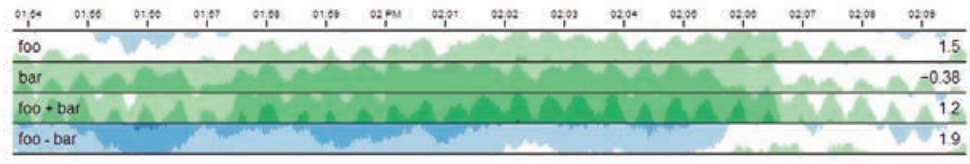
En las imágenes se presentan los resultados de la búsqueda del término "*electric cars*" y los términos Relacionados, de los que podemos ver los Principales en los primeros lugares de la lista. Se presentan dos búsquedas, en la segunda se excluyen en los resultados el término "*electric cars*". En los resultados se descubre una fuerte correlación con términos relacionados con la conversión de coches en coches eléctricos, como "*electric car conversión*", "*electric conversion*" o "*car conversión*", **tendencia que podría señalar incluso una interesante oportunidad de negocio**.



Este otro gráfico, tomado del software **Synerscope**, presenta el estudio agregado de correos electrónicos enviados entre varias personas, categorizado asimismo la posición jerárquica de los que envían y reciben los e-mails.

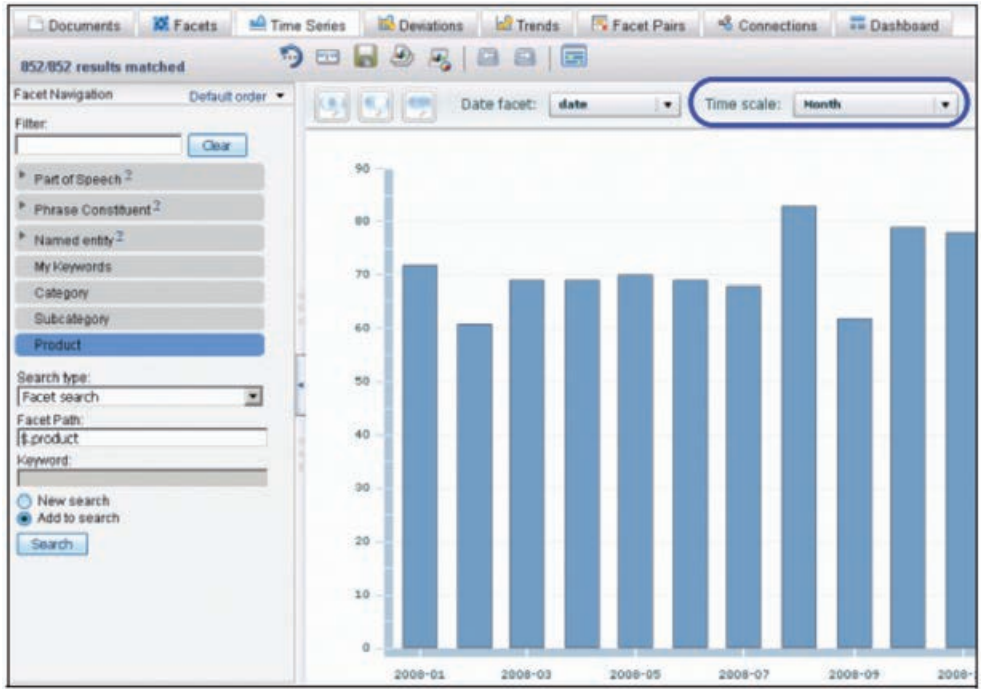
Esta visualización podría ser muy útil por ejemplo para estudiar el flujo de información dentro de la organización, cómo se genera el conocimiento y cumplir multitud de casos de uso en el área de Recursos Humanos. Otro ejemplo evidente serían los proyectos de Informática Forense, buscando relaciones entre personas e incluso aplicando técnicas de procesamiento de lenguaje natural, relaciones entre los conceptos e ideas intercambiadas.

6.3.2. Explotación de Históricos



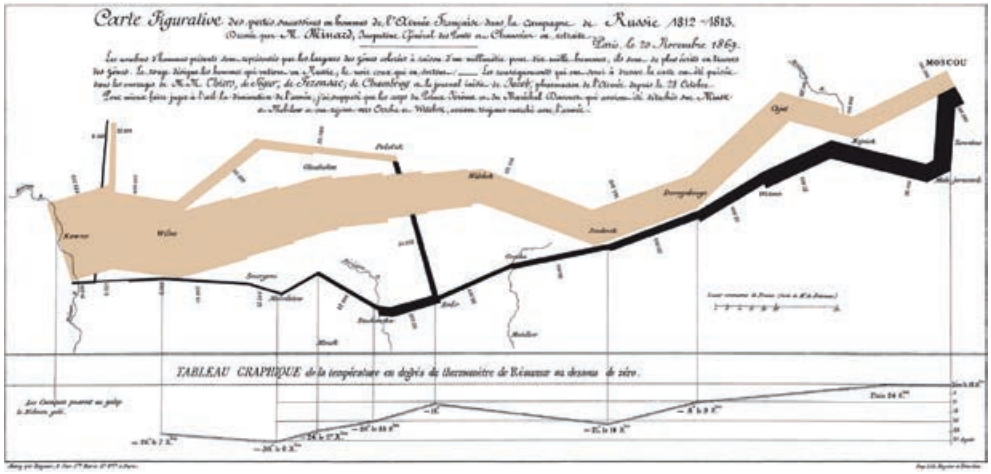
Uno de los casos de uso que más se repiten, junto con la detección de tendencias, es la necesidad de **estudiar la información disponible a lo largo del tiempo** sobre las Entidades de Información del Sistema de Vigilancia. Esta necesidad conlleva una toma de decisiones estratégica muy vinculada al largo plazo. Para que dentro de 10 años sea posible realizar un estudio histórico de la información es necesario tomar hoy la decisión de almacenar esos datos. Todo tipo de casos de uso que pretendan aprender del pasado por ejemplo relaciones causa-efecto necesitan de esta Función, que estará soportada muy frecuentemente por métodos probabilísticos y de data science en general.

Las ilustraciones son de **D3js.org,** concretamente "cubism.js" (arriba) y de **IBM Watson Content Analytics** (en página siguiente).



Deberán estudiarse las visualizaciones más adecuadas para cada Entidad de Negocio y caso de uso, cada una de ellas puede requerir una visualización diferente para que el análisis y extracción de conclusiones sea efectivo.

Otro ejemplo interesante, tomado de **Synerscope**, presenta la evolución del tamaño del ejército de Napoleón durante una temporada de contiendas en Rusia, incluyendo además del tiempo también la temperatura.

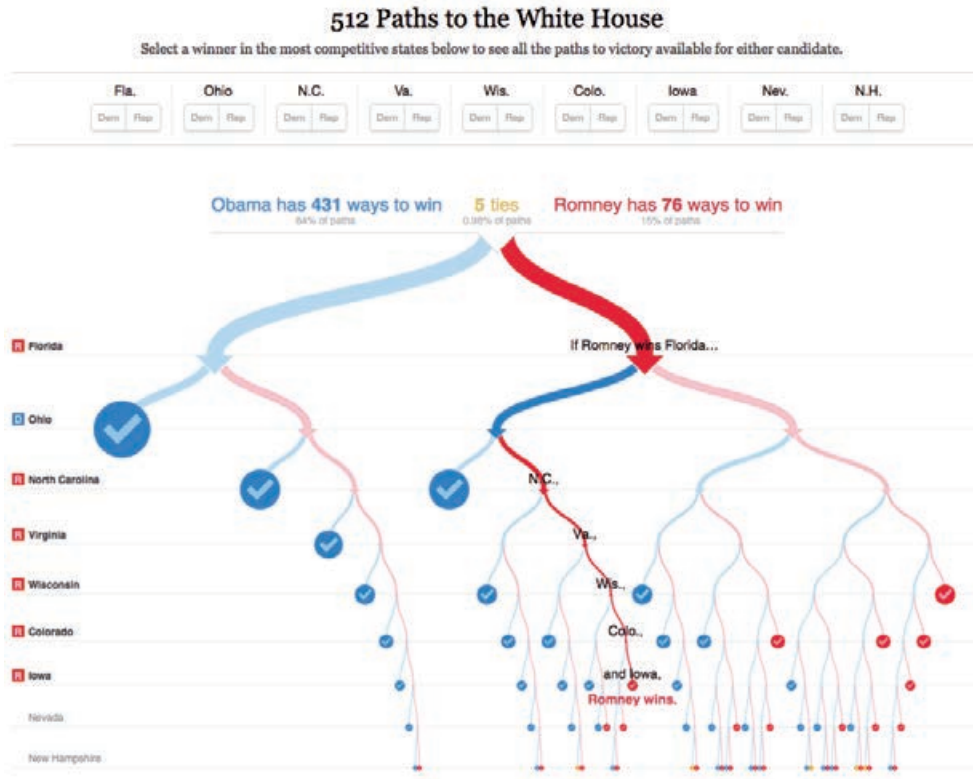


6.3.3. Sistemas de Escenarios

Otra funcionalidad demandada en varios casos de uso es **poder aplicar diferentes escenarios a una Entidad de Negocio para evaluar la plausibilidad de dicho escenario**. Eso implica no sólo la disposición de datos sino de todo un conjunto de reglas de negocio a aplicar a la misma.

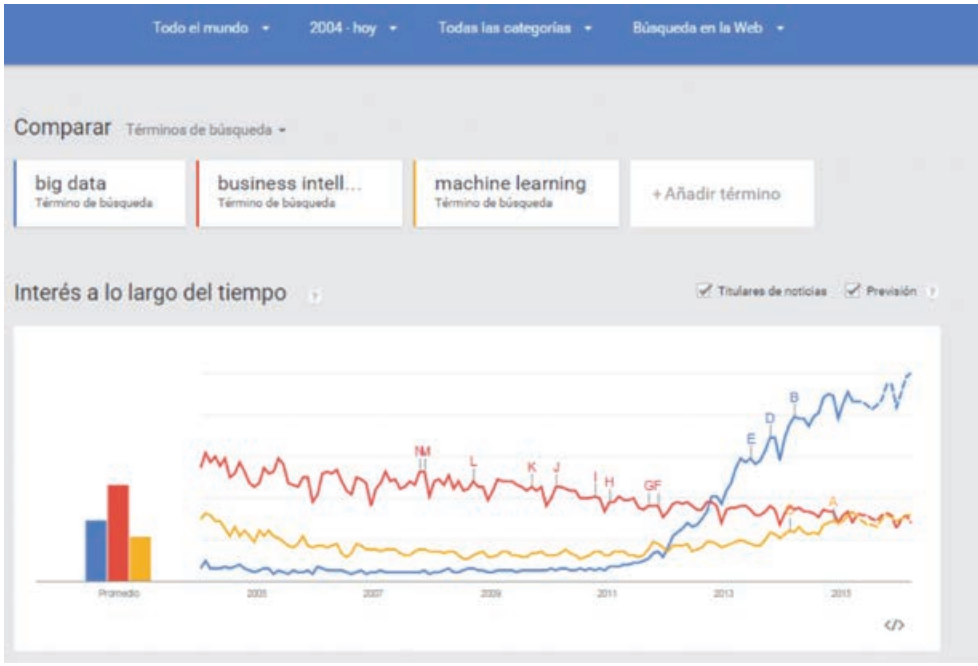
Este tipo de aplicación sería útil para estudiar si una situación es o no viable, por ejemplo si una empresa o grupo de empresas pueden realizar o no un conjunto de proyectos y por tanto si deben otorgársele ayudas públicas, estudiar el resultado de una fusión o una adquisición, el efecto de una abundancia o una escasez de un determinado producto o servicio.

En la imagen presentamos un estudio realizado por el New York Times de diferentes tipos de escenarios posibles en las elecciones americanas entre Obama y Romney. La aplicación permite estudiar varios escenarios posibles mediante la selección de victorias o derrotas en diferentes estados.



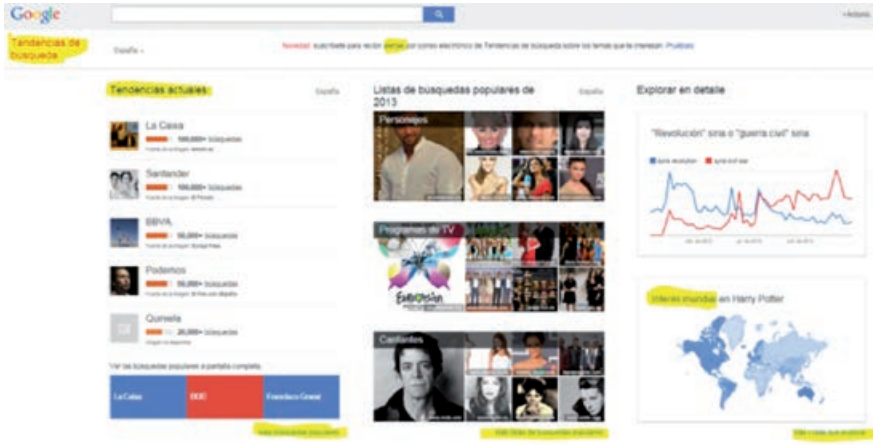
6.3.4. Tendencias

Ilustramos con **Google Trends** la **materialización de la Funcionalidad "Tendencias".** Comparamos la realización de búsquedas de tres términos: "Big Data", "Business Intelligence" y "Machine Learning". Hemos marcado la opción de "Previsión", para que nos aparezca una proyección hacia futuro de los términos. La analítica se puede particularizar por las dimensiones marcadas en la barra azul: ubicación física, tiempo, categorías, y las diferentes aplicaciones en las que se han hecho las búsquedas.



Se aprecia el creciente interés desde el año 2012 en "Big Data" y también un crecimiento notable en "Machine Learning", frente a un decaimiento del interés en "Business Intelligence". Esto posiblemente se deba a que muchos fabricantes de aplicaciones software han pasado de denominarse empresas de Business Intelligence a autoproclamarse empresas de Big Data.

La herramienta presenta una estimación asimismo de qué porcentaje de las búsquedas debe atribuirse a un concepto u otro. Por ejemplo para Big Data, diferencia entre la industria de Big Data y un grupo musical de idéntico nombre.



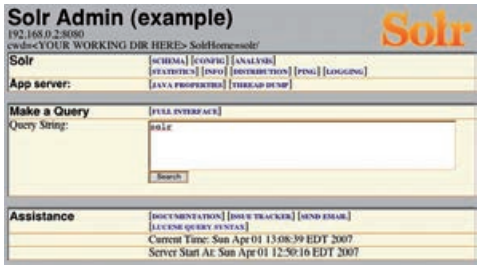
Google Trends presenta información sobre Tendencias actuales generales (por ejemplo en el año) y búsquedas recientes. Concretamente la captura de pantalla está realizada tras un accidente de avioneta de Harrison Ford, el anuncio de un concierto de Alejandro Sanz y una campaña de marketing de la empresa de telecomunicaciones Yoigo. Sin embargo en ese momento sólo "Harrison Ford" era Trending Topic en Twitter, otra de las aplicaciones de Tendencias más populares.

Por último presentamos también la interfaz del IBM Watson Content Analytics, en el que presenta un indicador de crecimiento y un contador de frecuencia sobre los datos que estudia.

6.3.5. Buscadores Avanzados

Una de las piezas clave en torno a la que han ido creciendo las aplicaciones Big Data es el Buscador. Se consideran varios tipos de buscadores, que localizan información dentro de Repositorios de Conocimiento:

Lucene, SolR, ElasticSearch... y Nutch



Lucene y **SolR** son dos proyectos **Apache** íntimamente relacionados. Lucene constituye un buscador y SolR es una aplicación que recubre Lucene con el objetivo de facilitar la usabilidad y la integración de Lucene con otras aplicaciones. Otro proyecto similar a SolR es ElasticSearch, usado por ejemplo por Linknovate, que mencionamos en el apartado de "Implementadores" o también Wikimedia o el CERN.

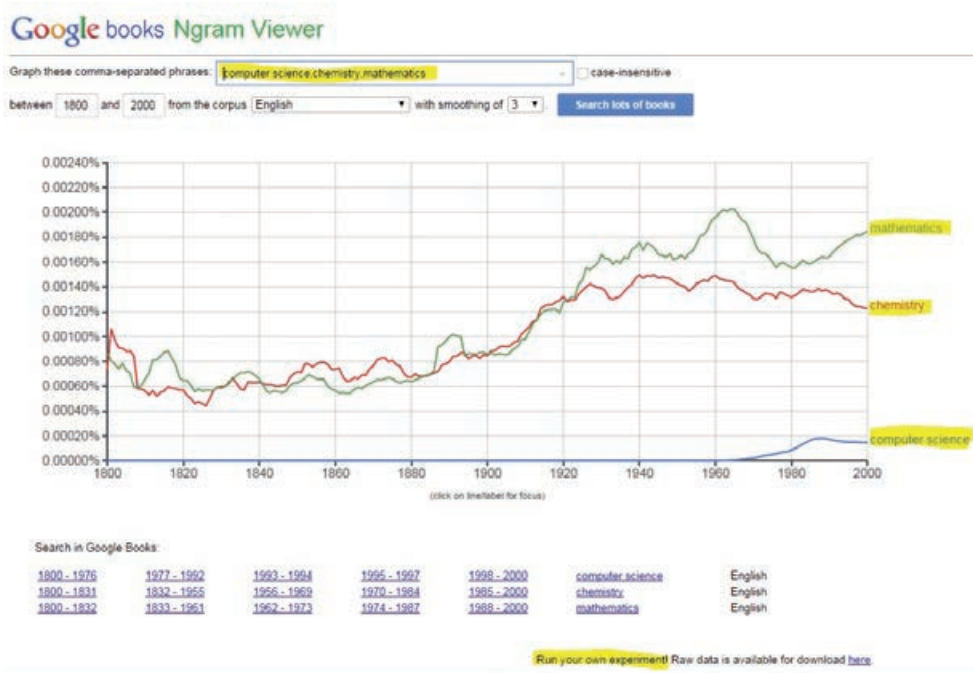
Nutch es otro proyecto de Apache que a menudo funciona de forma integrada con SolR. Es un Bot, un WebCrawler, que ya tratamos en el punto 3.2 de "Business Bots, Spiders, Scrapers". Para construir un motor de búsqueda tan solo tenemos que integrar Nutch y SolR.

Google Public Data Explorer

Google Public Data Explorer[116] se enfoca a la **búsqueda de datos públicos** que ponen a su disposición diversas instituciones públicas y privadas. Asimismo permite subir tus propios datos.

[116] http://www.google.com/publicdata/directory

Google Ngram Viewer



Google Ngram Viewer[117] permite la búsqueda de palabras y frases cortas dentro de los libros disponibles en Google Books, desde el año 1800 hasta el 2012. Es por tanto una **búsqueda dentro de un corpus documental.**

En la imagen vemos la evolución desde el año 1800 de las referencias a "matemáticas" y "química" y la irrupción a finales de los años 60 de la informática (en inglés, "computer science").

Búsqueda distribuida: What do you love?

Finalmente presentamos **What do you love?** [118], que permite realizar **búsquedas distribuidas en diferentes aplicaciones y por tanto en diferentes repositorios**, agregando los resultados.

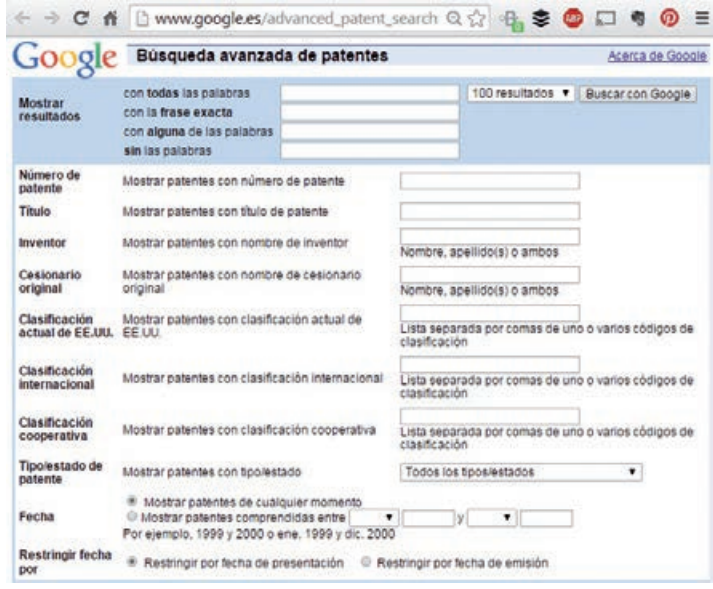
Esta interfaz está muy relacionada con el concepto de **Bases de Datos Federadas** y **Data Federation** que presentamos en el apartado de "Integración de Datos".

---

[117] https://books.google.com/ngrams

[118] http://www.wdyl.com/

Google Patents



Las **Patentes** son uno de las Entidades de Negocio Estructurales que presentábamos en el apartado de "Modelo de Información". **Google Patents**[119] usa como Fuentes de Información las bases de datos de varios organismos globales de patentes y propiedad intelectual, como la USPTO[120], la EPO[121] y la WIPO[122].

6.3.6. Correlación de Eventos y Alarmas

Se debe disponer de funcionalidad que realice una monitorización de datos cuyo resultado dispare una alarma:

- Ante la aparición de información nueva relevante para los usuarios.
- Como resultado de análisis periódicos a realizar en el conocimiento almacenado en los repositorios de información de vigilancia, mediante técnicas analíticas avanzadas (data mining, text mining, etc), como por ejemplo la emergencia de un concepto que empieza a resultar una tendencia relevante.
- Eventos relevantes para los administradores del sistema, como nuevos conceptos o entidades detectadas por el sistema. Como resultas de los mismos deberá tratarse, por ejemplo, la modificación de una ontología.

Como resultado de estas alarmas el sistema generará avisos a través de las diversas interfaces de usuario configuradas en el sistema (e-mails, mensajería, avisos en la interfaz de usuario, etc.). Asimismo es habitual integrarlos con sistemas formales o informales de BPM (Business Process Management).

6.3.7. Navegador

Resulta especialmente relevante la navegación a través de los datos, cuando estos están organizados dentro de una ontología. Para ello se puede disponer de aplicaciones como Pubby[123], que proporcionan interfaces Linked Data para bases de datos NoSQL accesibles mediante SPARQL. El resultado es un conjunto de páginas HTML accesibles mediante cualquier navegador, con enlaces entre las relaciones establecidas entre los datos almacenados.

---

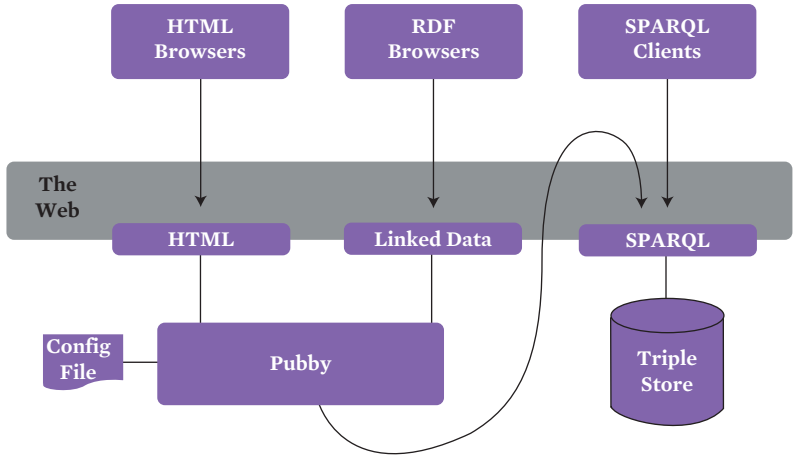
[119] Google Patents http://www.google.es/advanced_patent_search

[120] USPTO United States Patent and Trademark Office

[121] EPO European Patent Office

[122] WIPO World Intellectual Property Organization

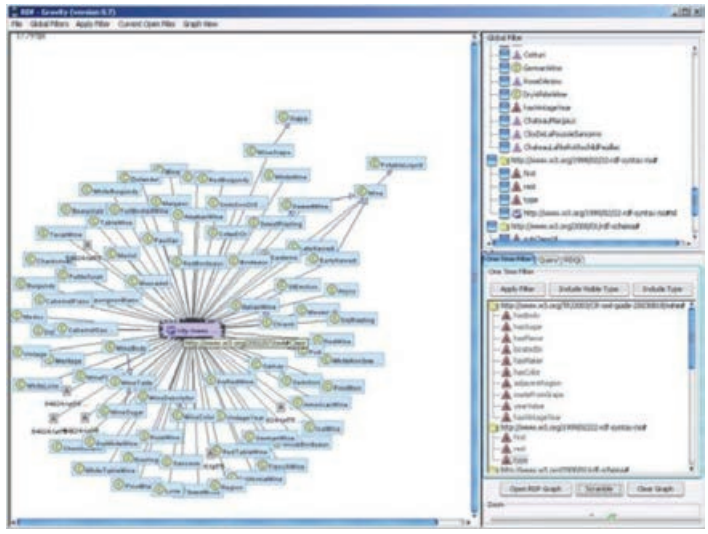
[123] http://wifo5-03.informatik.uni-mannheim.de/pubby/



La DBpedia, que ya mencionamos en el apartado de "Entendiendo la importancia de las Ontologías", es un buen ejemplo de esta navegación.

is http://es.dbpedia.org/property/genero of	▪ dbpedia:Norton_360 ▪ dbpedia:ESET_NOD32_Antivirus ▪ dbpedia:Norton_AntiVirus ▪ dbpedia:BitDefender ▪ dbpedia:ClamAV ▪ dbpedia:ESET_Smart_Security
is http://es.dbpedia.org/property/género of	▪ dbpedia:Windows_Live_OneCare ▪ dbpedia:Avast! ▪ dbpedia:Windows_Defender ▪ dbpedia:Bitfrost ▪ dbpedia:Avira ▪ dbpedia:ClamWin ▪ dbpedia:Microsoft_Security_Essentials ▪ dbpedia:Dr._Web ▪ dbpedia:Herramienta_de_eliminación_de_sc
is http://es.dbpedia.org/property/productos of	▪ dbpedia:Panda_Security ▪ dbpedia:ESET
is skos:subject of	▪ dbpedia:Categoría:Software_antivirus
is foaf:primaryTopic of	▪ http://es.wikipedia.org/wiki/Antivirus

Asimismo se puede disponer de navegaciones visuales con enlaces entre los diferentes conceptos a través de las relaciones existentes entre los mismos.



6.3.8. Consultas a las Bases de Datos de Vigilancia

El Sistema de Vigilancia e Inteligencia debe contener herramientas de consulta a las Bases de Datos que contenga información de Vigilancia, particularizadas para el tipo de repositorio incluido:

- Querys **SQL**: consultas a sistemas relacionales estándar con SQL.
- Querys **MDX**: consultas a datamarts mediante MDX.
- Querys **SPARQL**: consultas a BD NoSQL orientadas a grafos, en las que se almacenan ontologías y estructuras de conocimiento.
- Querys **NoSQL**: querys realizadas en otras bases de datos NoSQL.
- Querys **CRUD**: querys generalistas al RIV que permitirán consultar cualquier tipo de repositorio de información.

En estas consultas deben considerarse las relaciones que estructuran la funcionalidad de las Entidades Estructurales:

- Presenta información y características del EE.
- Relaciona los EE entre si.
- Relaciona cada EE con otros EEs.
- Permite buscar EE con determinados criterios.
- Permite buscar y presenta información de los agentes relacionados con los EE.
- Permite estudiar la evolución histórica de los EE.

Asimismo debe tenerse en cuenta los lenguajes de consulta que presentamos a continuación en el apartado de “Lenguajes destacados para Big Data”.

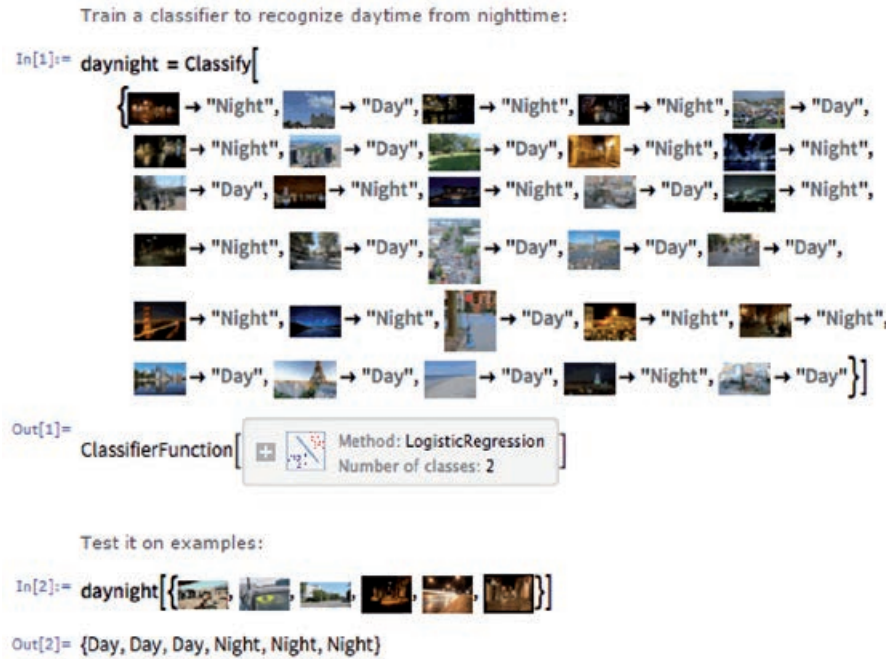
6.4. Aplicaciones Machine Learning y técnicas de Data Science

El Wolfram Language[124] es una manera muy visual y potente de ilustrar la funcionalidad de Machine Learning y las técnicas de Data Science que se presentan en el apartado 3.3.2 "Algunas técnicas útiles para Data Science".

Wolfram dispone de capacidades de machine-learning en el lenguaje incluyendo **aprendizaje supervisado, métodos de aprendizaje sin supervisión y de preparación y filtrado de los datos**. Los datos pueden ser numéricos, textos, imágenes, etc. [125]

ILUSTRACIÓN 19

Implementación de Machine Learning en Wolfram Language para la clasificación de imágenes día/noche



En las imágenes que presentamos a continuación una única función del lenguaje de programación Wolfram es capaz de buscar tumores en un cerebro utilizando técnicas

[124] Wolfram Language & System Documentation Center – Machine Learning http://reference.wolfram.com/language/guide/MachineLearning.html

[125] Las referencias a Wolfram que siguen a continuación en Wolfram Language & System Documentation Center en http://reference.wolfram.com/language/

de Clustering, encontrar grupos relacionados en grupos y grafos, que por ejemplo sería utilizado para buscar comunidades de personas en una red social o clasificar imágenes para determinar si reflejan la noche o el día.

ILUSTRACIÓN 20

Aplicación de la técnica de Clustering
(ver 3.3.1 - Algunas técnicas útiles para Data Science) al análisis de imágenes

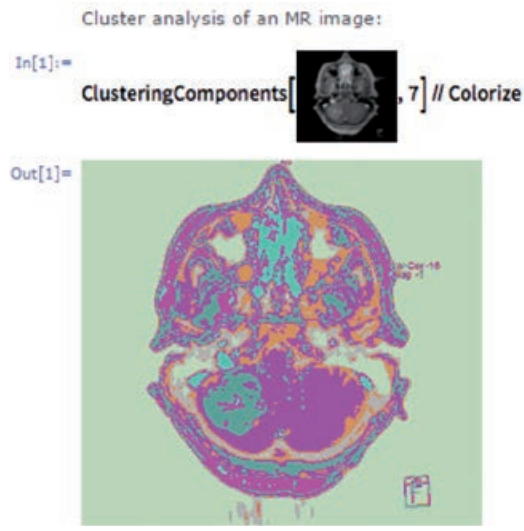
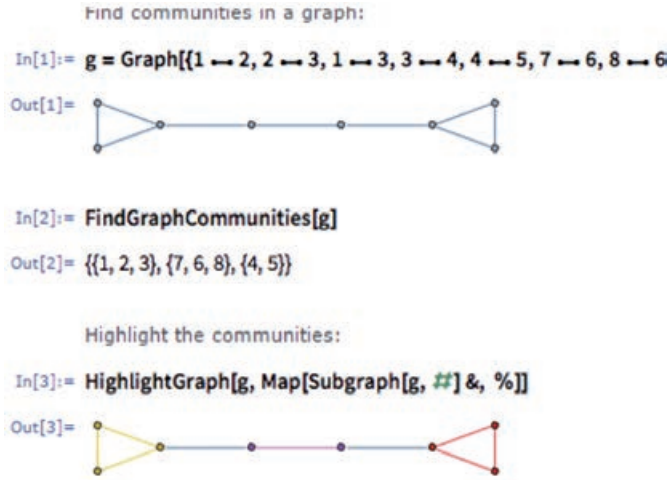


ILUSTRACIÓN 21

Aplicación de la técnica de Data Science "Análisis de Grafos" (ver 3.3.1).
Muy útil para el estudio de Redes Sociales



```
Label two clusters of values in a list:
In[1]:= ClusteringComponents[{1, 2, 3, 7, 8}, 2]
Out[1]= {1, 1, 1, 2, 2}
```

```
In[2]:= CommunityGraphPlot[%, FindGraphCommunities[%]]
Out[2]=
```

Estas capacidades permiten **clasificar datos en diferentes categorías y predecir valores a partir de los datos ya existentes**. Asimismo permiten la búsqueda de patrones o agrupaciones de datos (clusters), y la búsqueda de valores más cercanos a otros cumpliendo criterios previamente definidos.

ILUSTRACIÓN 22

Aplicación de la técnica de Data Science "Clustering" en la clasificación de genomas

```
Cluster genomic sequences based on the number of element-wise differences:
In[1]:= gdata = {"GTCTT", "AAGCT", "GGTAA", "AGGCT", "GTCAT",
        "CGGCC", "GGGAG", "GTTAT", "GTCAT", "AGGCT", "GTCAG", "AGGAT"};
In[2]:= FindClusters[gdata, DistanceFunction → HammingDistance]
Out[2]= {{GTCTT, GTCAT, GTTAT, GTCAT, GTCAG}, {AAGCT, AGGCT, CGGCC, AGGCT, AGGAT}, {GGTAA, GGGAG}}
```

De esta manera se responde de forma directa a **las necesidades de los usuarios de realizar proyecciones a futuro (predicciones) y realizar clasificaciones de los datos** a partir de la información disponible, incluida la información histórica almacenada.

Otra aplicación habitual es la clasificación de documentación existente obteniendo patrones de clasificación y consecuentemente usar dichos patrones en nuevos documentos a ser tratados. Por ejemplo puede utilizarse en la búsqueda de tipos de contenidos en páginas web, como pueda ser las páginas de productos y servicios de una tienda online o de contacto en la web corporativa de una empresa.

Otra de las necesidades de los usuarios es diseñar modelos de los datos y deducir consecuencias a partir de dichos modelos, pudiendo obtener incluso capacidades predictivas.

ILUSTRACIÓN 23

Predicción ("forecasting") de la calidad de un vino

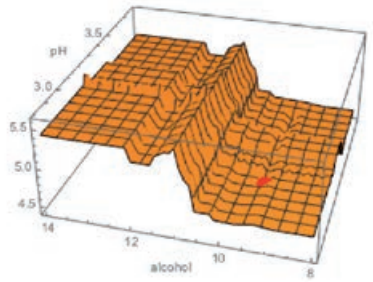
Create a function that predicts the quality of the unknown wine as a function of its pH and alcohol level:

```
In[8]:= quality[pH_, alcohol_] :=
          p[{7.6`, 0.48`, 0.31`, 9.4`, 0.046`, 6.`, 194.`, 0.99714`, pH, 0.61`, alcohol}];
```

Plot this function to have a hint on how to improve this wine:

```
In[9]:= Show[Plot3D[quality[pH, alcohol], {pH, 2.8, 3.8}, {alcohol, 8, 14}, AxesLabel → Automatic],
          ListPointPlot3D[{{3.07`, 9.4`, p[unknownwine]}}, PlotStyle → {Red, PointSize[.05]}]]
```

Out[9]=



Estas funcionalidades incluyen la aplicación a conjuntos de datos a medidas estadísticas (media, varianza, percentiles, etc), suavizado de datos (data smoothing), herramientas de visualización y análisis estadístico y de modelos estadísticos, pruebas de hipótesis y aproximación de funciones.

ILUSTRACIÓN 24

Búsqueda de máximos locales en la cotización de Microsoft en Bolsa. Utilizado frecuentemente para el pronóstico de precios adecuados para la venta de acciones

Find peaks in the stock price of Microsoft in one year:

```
In[1]:= data = FinancialData["MSFT", {{2012, 1, 1}, {2013, 1, 1}, "Day"}][[All, 2]];

In[2]:= peaks = FindPeaks[data, 3]

Out[2]= {{51, 30.98}, {58, 30.73}, {79, 30.37}, {118, 29.36}, {127, 29.2},
          {138, 29.11}, {159, 29.53}, {182, 30.05}, {213, 28.53}, {225, 26.69}, {244, 26.67}}
```

Highlight peaks on the plot of the data:

```
In[3]:= ListLinePlot[data, Epilog → {Red, Point[peaks]}]
```

Out[3]=

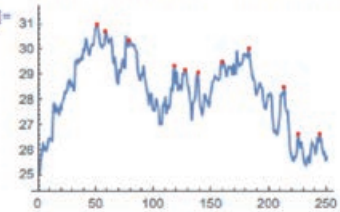


ILUSTRACIÓN 25

Aplicación de los modelos de alisado exponencial de Holt Winters
[ver 3.3.1 Algunas técnicas útiles para Data Science > Pronóstico ("forecasting")]

Determine whether snowfall accumulations in Buffalo are normally distributed:

```
In[1]:= snow = ExampleData[{"Statistics", "BuffaloSnow"}];

In[2]:= ℋ = DistributionFitTest[snow, Automatic, "HypothesisTestData"];
```

Use the Jarque–Bera ALM test and Shapiro–Wilk tests to assess normality:

```
In[3]:= ℋ["TestDataTable", {"JarqueBeraALM", "ShapiroWilk"}]
```

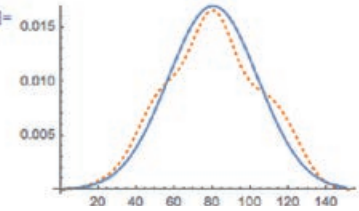
Out[3]=

	Statistic	P-Value
Jarque-Bera ALM	0.896776	0.58867
Shapiro-Wilk	0.983489	0.559265

The SmoothHistogram agrees with the test results:

```
In[4]:= Show[SmoothHistogram[snow, PlotStyle → {Dotted, Orange}],
          Plot[PDF[ℋ["FittedDistribution"], x], {x, 0, 150}]]
```

Out[4]=



6.5. Lenguajes destacados para Big Data

Destacamos en este apartado algunos lenguajes, tanto de programación como de consulta a bases de datos que están surgiendo para el entorno de Big Data.

Lenguajes de Programación

Además de los ya tradicionales lenguajes de programación Java, PHP, Python y C++, varias novedades han surgido o se han consolidado al albor del Big Data. Queremos destacar aquí los siguientes:

Scala

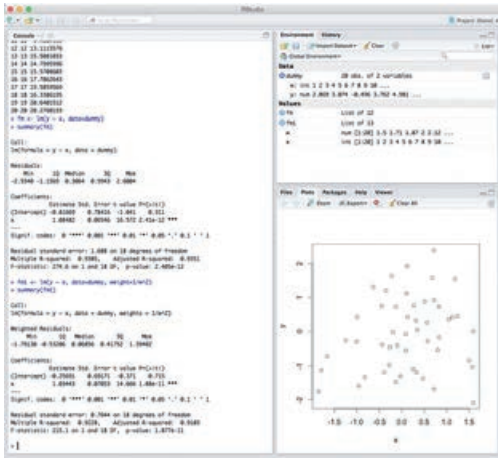
```
def streamRange(lo: Int, hi: Int): Stream[Int] = {
  print(lo+" ")
  if (lo >= hi) Stream.empty
  else Stream.cons(lo, streamRange(lo + 1, hi))
}
```

Scala[126] es sin uno de los lenguajes de programación de moda para los entornos Big Data. Destaca por ser un lenguaje multiparadigma, en el que se integran características de la tra-

[126] http://www.scala-lang.org/

dicional programación orientada a objetos y de la programación funcional que ya hemos visto que es tan relevante debido a MapReduce y al procesamiento Big Data que conseguimos, por ejemplo con Spark. Es un lenguaje muy similar a Java, de hecho corre sobre la JVM (Java Virtual Machine), por lo que ha logrado un gran impacto en la comunidad de desarrolladores de Java.

R



El otro gran lenguaje que está triunfando en la revolución Big Data es "R"[127]. El encabezado de su página web nos refleja con claridad porqué: "*The R Project for Statistical Computing*". Es el lenguaje preferido por muchos científicos de datos (en inglés "*data scientists")* para los proyectos que utilicen Machine Learning y/o Métodos Estadísticos. Cuenta con una amplísima variedad de funciones estadísticas que abordan las cuestiones que hemos presentado en los apartados de "Machine Learning", "Data Science, Estadística, Inteligencia Artificial" y "Algunas técnicas útiles para Data Science".

Además destaca por sus capacidades gráficas, lo cual le ha hecho también muy popular en proyectos de visualización de datos, que hasta hace poco estaba reservado casi exclusivamente el mundo del Business Intelligence.

Wolfram Language

Anunciado a mediados del año 2014, Wolfram Language[128] es el lenguaje que se usa internamente dentro de la empresa Wolfram famosa por el buscador semántico **Wolfram Alpha**[129] o el software **Mathematica**[130].

[127] http://www.r-project.org/about.html

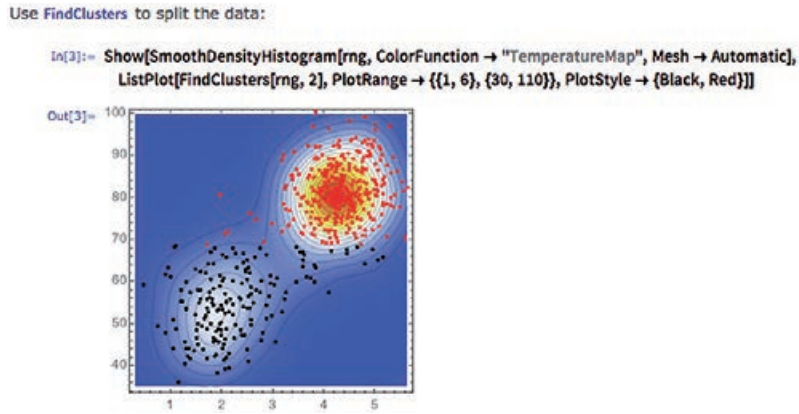
[128] http://www.wolfram.com/language/

[129] http://www.wolframalpha.com/

[130] http://www.wolfram.com/mathematica/

ILUSTRACIÓN 26

Clasificación de datos mediante técnica de Clustering



Es lo que se llama un lenguaje multiparadigma, que permite varios tipos de programación, incluyendo la muy relevante para el Big Data programación funcional. Wolfram le llama a su estilo de programación la "programación basada en el conocimiento".

ILUSTRACIÓN 27

Detección de Género (mas/fem) en imágenes

Cuenta con funciones específicas para procesamiento avanzado de textos, lingüística, redes y grafos, algoritmos de Machine Learning, imágenes, sonido, computación científica, financiera y un largo etcétera. Incluye por tanto un conjunto de capacidades que le capacitan para dar un gran soporte a los casos de uso de Vigilancia y para los de Big Data en general

Lenguajes de consulta a Bases de Datos Big Data

El otro caballo de batalla son los lenguajes de consulta a las Bases de Datos. Nuevos tipos de bases de datos hacen necesarios nuevos lenguajes de consulta que permitan abstraer y ser útiles para especificar el tipo de consultas que se le hacen a cada tipo de base de datos. Seleccionamos aquí los siguientes:

SPARQL

Es el lenguaje de consulta a datos que siguen el estándar **RDF**, uno de los lenguajes más extendidos para la definición de **Ontologías**. Ha sido estandarizado por el **W3C**. Es el lenguaje de referencia para las Bases de Datos orientadas a Grafos. La sintaxis es similar a la de SQL.

Pig Latin

Es el lenguaje de programación para Pig, un proyecto Apache Pig orientado a la creación de programas MapReduce para la plataforma Hadoop.

HiveQL

Lenguaje similar a SQL que convierte consultas a procesos map/reduce. Se utiliza dentro del proyecto Apache Hive para análisis de grandes conjuntos de datos almacenados en HDFS.

CQL

CQL, Cassandra Query Language es el lenguaje de consulta específico de la base de datos del proyecto Apache Cassandra., una de las bases de datos orientada a columnas más popular.

6.6. Aplicaciones Procesamiento de Lenguaje Natural (PLN)

El Sistema de Vigilancia e Inteligencia debería disponer de funcionalidad PLN (Procesamiento de Lenguaje Natural, en inglés *"NLP – Natural Language Processing").* para el tratamiento de frases, párrafos o documentos de los que queramos extraer información lingüística, fundamentalmente Entidades, Conceptos y relaciones semánticas entre los mismos.

A alto nivel el tratamiento consistirá en el indexado y cribado de la información descargada y tras ello al ***Procesamiento de los Textos Relevantes*****,** cuyo resultado será la Información de Vigilancia que se almacenará en el **Repositorio de Información de Vigilancia.**

Para el procesamiento de lenguaje natural el Sistema de Vigilancia debe organizarse por niveles de abstracción, con el objetivo de proporcionar un modelo con fuerte modularidad, encapsulamiento y transparencia:

- **Nivel de Aplicación:** se implementarán los Casos de Uso de los usuarios.
- **Nivel de Análisis:** se realizarán las tareas de procesamiento de lenguaje natural: Análisis Morfológico, Análisis Sintáctico y Análisis Semántico.
- **Nivel de Módulos:** las tareas de los diferentes Análisis serán realizadas por componentes de NLP.

En el módulo NLP se interacciona con los Repositorios de Conocimiento y Bases de Datos Big Data del Sistema de Vigilancia e Inteligencia para obtener información sobre los textos que sean procesados. La interacción entre módulos y los repositorios realizará a través de un formato de anotación, como XML o JSON, cuyos requisitos se detallan a continuación.

Se considerará el procesamiento de información en diferentes idiomas, especialmente el inglés además del español.

Se espera que las aplicaciones que reflejen los casos de uso a implementar puedan necesitar de etiquetados específicos de la información para reflejar información relevante para el caso de uso en cuestión.

6.6.1. Formato de Anotación

Existe en el mercado un número creciente de aplicaciones NLP, sin embargo no existen estándares que aseguren la **interoperabilidad** entre los mismos y la integración en el Sistema de Vigilancia sin comprometer la funcionalidad de cada una de las aplicaciones, por lo que se considera un requisito importante a cumplir.

En el Sistema de Vigilancia e Inteligencia debe definirse y utilizarse un **formato de anotación bien documentado** que responda a las entradas y las salidas de los Componentes. Este formato debe ser **multicapa**, para poder ser ampliable de forma nativa, ya que cada nueva aplicación puede requerir de una manera de anotar el resultado de su procesamiento. El formato debe aunar flexibilidad, eficiencia de procesamiento y reusabilidad.

Cada módulo del sistema recibirá información en el formato definido, lo procesará y generará una salida añadiendo información en la capa correspondiente según su funcionalidad. El último módulo que procese la información la almacenará en los repositorios de información. Ejemplo de capas serían la capa de cabecera, la capa de conceptos, la capa de entidades, la capa de SRL, la capa de opinión, etc.

Cada nueva aplicación del sistema será susceptible de necesitar nuevas capas de información respondiendo a las necesidades de procesamiento, de ahí la importancia de que el formato contemple su ampliación de forma nativa.

El formato deberá incluir la utilización URIs y de RDF para las representaciones lingüísticas, así como la inclusión de enlaces a la procedencia de la información y puntuaciones sobre la confianza en el resultado obtenido.

Actualmente no existe un estándar en este tipo de formatos de anotación. En el diseño del formato deberían tenerse en cuenta las aproximaciones que están teniendo actualmente tanto las empresas como los grupos de investigación.

6.6.2. Consideraciones técnicas

El diseño de la plataforma debe estar orientado al procesamiento de flujos continuos de datos en tiempo real (en inglés *"data streaming"*) con alta disponibilidad, escalabilidad y clusterización, como veíamos en el apartado 3.7.2 "Sistemas de Procesamiento". Debe garantizarse que no se pierden datos en el proceso, que todos son procesados. Este tipo de procesamiento responde al **paradigma denominado *Streaming Computing***.

También debe poder realizarse el **procesamiento por lotes (batch),** que será utilizado en aquellas situaciones en las que pueda postergarse y agruparse el procesamiento de la información.

El procesamiento de la información se realizará a través de la concatenación de módulos dedicados que interaccionarán a través de un formato XML de anotación, cuyos requisitos se tratan a continuación. En lo posible se paralelizará el proceso de cara a maximizar la eficiencia del mismo y reducir los cuellos de botella que puedan surgir debido a dependencias funcionales entre módulos o por tiempos de proceso. Los módulos propuestos para el sistema se presentan en un apartado a continuación.

De cara a la organización del procesamiento se estudiará la utilización de máquinas virtuales que habiliten los requisitos de modularidad y encapsulamiento, ya que la utilización de máquinas virtuales es un estándar de facto en soluciones distribuidas. Se tendrán en cuenta también:

Las dependencias funcionales entre módulos, para no colocarlos en la misma máquina virtual.

- Las dependencias creadas por el tiempo de ejecución de los módulos, que crean caminos críticos de ejecución del pipeline completo.
- Equilibrar el número de máquinas virtuales en el sistema con el TCO del Sistema de Vigilancia, consolidando en lo posible máquinas virtuales en función del uso que se haga de ellas.
- Debe asegurarse la ejecución de varios módulos en paralelo.

Las interacciones con los Repositorios de información de Vigilancia se harán a través del formato XML de anotación a definir. Cada almacenamiento de información deberá contener un enlace a su procedencia.

Deberá tenerse en cuenta que el Sistema de Vigilancia podrá implementarse en Cluster, con un número de nodos grande, con el objetivo de minimizar el tiempo de procesamiento NLP.

6.6.3. Procesamiento NLP de la información

El procesamiento de los textos se realiza mediante la concatenación de módulos que se intercambian información en el formato de anotación presentado en el apartado anterior. El resultado final se almacenará en los Repositorios de Información de Vigilancia, típicamente en la Base de Datos orientada a Columnas.

A continuación se listan los módulos identificados como parte del sistema, cuya funcionalidad es detallada en apartados siguientes:

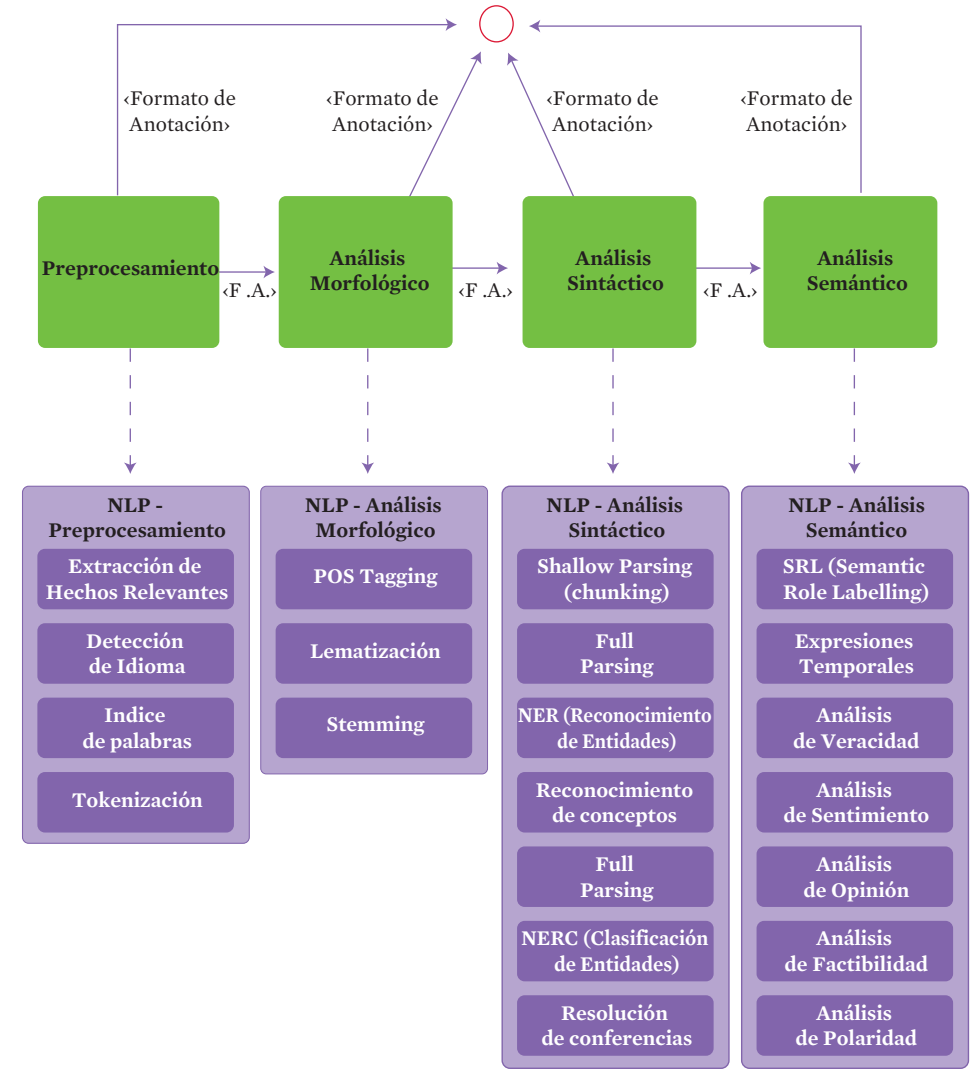
Preprocesamiento	• Detección de idioma. • Índice de palabras. • Tokenización.
Análisis Morfológico	• Análisis Morfológico (POS Tagging). • Lematización. • Stemming.
Análisis Sintáctico	• Reconocimiento de Entidades (NER). • Reconocimiento y Clasificación de Entidades (NERC). • Reconocimiento de Conceptos. • Enriquecimiento de Entidades y Conceptos. • Desambiguación de Entidades (NED). • Resolución de correferencias. • Shallow Parsing. • Full Parsing. • Análisis sintáctico.
Análisis Semántico	• Semantic Role Labelling. • Detección de eventos. • Detección de expresiones temporales. • Análisis de Factibilidad. • Análisis de Sentimiento. • Análisis de Opinión. • Análisis de Polaridad. • Análisis Semántico.

Los módulos no son disjuntos, pueden contener la funcionalidad de otros módulos.

Debe ser posible ampliar el Sistema de Vigilancia con módulos NLP adicionales que cumplan con los requisitos de integración con el sistema. Es condición necesaria que todos los módulos acepten, procesen y generen sus salidas según el formato de anotación a especificar.

A continuación se describen el pipeline de procesos habituales en el procesamiento de lenguaje natural, que constituyen el Pipeline de Procesamiento NLP.

ILUSTRACIÓN 28
Pipeline de Procesos habituales en procesamiento de lenguaje natural



El tratamiento NLP de la información se realizará mediante la concatenación de los módulos listados anteriormente. Existen 4 grupos de Módulos:

- Preprocesamiento.
- Análisis Morfológico.
- Análisis Sintáctico.
- Análisis Semántico.

Preprocesamiento

Dentro de este apartado se consideran los siguientes módulos funcionales:

- Extracción de Textos Relevantes.
- Detección de Idioma.
- Índice de palabras.
- Tokenización.

Una vez descargado un texto el primer paso consiste en la ***extracción de textos relevantes***, donde se ha realizado el cribado de contenidos no relevantes de los documentos y selección de lo relevante, almacenándose en un ***repositorio de textos relevantes***. Este módulo puede considerarse también como parte del proceso de Extracción y Descarga de la información.

El módulo de **Detección de Idioma** señala el idioma del texto descargado para que consecuentemente se pueda configurar los Repositorios de Conocimientos que se utilizarán en el proceso.

Tras el proceso de Extracción de Información se genera y almacena un ***Índice de palabras*****,** una pequeña base de datos que contiene todas las palabras del texto junto con la información necesaria para localizarla.

El proceso de **Tokenización** consistirá en la organización del texto identificando frases y las palabras que las constituyen. Cada palabra quedará incluida en una estructura que llamaremos **Token**.

Análisis Morfológico

El siguiente paso es el **Análisis Morfológico** (POS tagging) de las palabras de cada token. Se genera la forma, clase o categoría gramatical de cada palabra, es decir, el género, el número, si es sustantivo, adjetivo, verbo, adverbio, etc.

Machine_Learning	y	Procesamiento_de_Lenguaje_Natural	son	tecnologías	complementarias
machine_learning	*y*	*procesamiento_de_lenguaje_natural*	*ser*	*tecnología*	*complementario*
NP00SP0	CC	NP00V00	VSIP3P0	NCFP000	AQ0FP0

[131]

Se aplican técnicas que permiten obtener una **forma canónica** de cada palabra que represente a todas sus formas singulares, plurales, variaciones de género o verbales, etc, denominadas "**lematización**" y "**stemming**".

La **Lematización** permite obtener el lema de una palabra, es decir la que se considera representante de un conjunto de palabras. El conjunto estaría constituido por el plural, el masculino, femenino, conjugaciones, etc. El **Stemming** genera la raíz de una palabra.

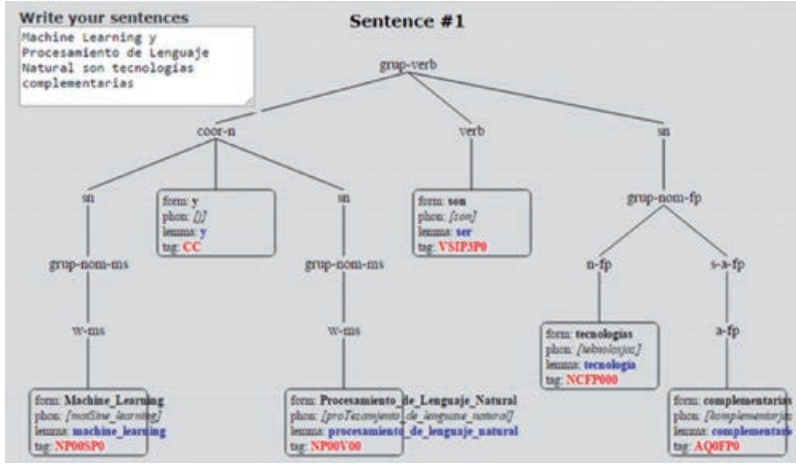
Análisis Sintáctico

A continuación se realiza el **Análisis Sintáctico** cuyo objetivo general es determinar las relaciones de concordancia y jerarquía entre las palabras formando sintagmas, determinando las funciones de las palabras o grupos de palabras dentro de la oración. Se genera una **estructura en árbol de cada frase**, reflejando la sintaxis de la misma.

Hay dos modalidades de análisis sintáctico: **shallow parsing y full parsing**.

ILUSTRACIÓN 29

Freeling Demo, con opción PoS Tagging[132]



[131] Freelimg Demo con opción PoS Tagging: http://nlp.lsi.upc.edu/freeling/demo/demo.php

[132] Demo online de Freeling con opción Full Parsing: http://nlp.lsi.upc.edu/freeling/demo/demo.php
Demo online de Freeling con opción Full Parsing: http://nlp.lsi.upc.edu/freeling/demo/demo.php

Especialmente cuando se hace Full Parsing pero también cuando se hace Shallow Parsing lo primero que se hace es **enriquecer la información** disponible accediendo a Repositorios de Conocimiento (ontologías, taxonomías, diccionarios, etc) etiquetando los sintagmas resultantes del análisis lingüístico. También puede accederse a Repositorios públicos en Internet, como la DBpedia, por ejemplo..

El **Shallow parsing** identifica elementos de una frase pero sin especificar sus estructuras internas. También se le conoce como **chunking.** El sistema realizará los siguientes procesos lingüísticos

- **Reconocimiento de Entidades (NER, Name Entity Resolution)**: reconocerá y clasificará todas las entidades, entendiendo como tal a los nombres de persona, lugares, organizaciones, empresas, y entidades de dominios específicos que estén identificados y organizados como tal en el Sistema de Vigilancia. Se almacenan junto a sus metadatos como Lista de Entidades en el repositorio de información de vigilancia.
- **Reconocimiento de Conceptos**: los conceptos son fragmentos de texto más significativos como los sintagmas nominales pero también pueden ser sintagmas verbales, adjetivales o adverbiales. Los conceptos incluyen a las entidades, que se almacenan junto a los metadatos (url de la página web, nº referencia, oferta, organismo, país, etc) como ***Lista de Conceptos*** en el repositorio de información de vigilancia.

Se incluirá funcionalidad para hacer **clasificación de entidades (NERC), desambiguación de entidades (NED).** Se debe determinar qué tipo de entidad es la palabra, cuál de las posibles acepciones de las palabras tiene sentido en el dominio del texto.

Otro módulo del ámbito sintáctico es el de **Resolución de Coreferencias**, que agrupa y resuelve todas las menciones a la misma entidad en un mismo texto.

El Sistema dispondrá de funcionalidad de **Full Parsing**, que realizará **extracción de hechos, eventos, relaciones entre entidades y conceptos**, del conocimiento lingüísticamente explícito en el texto. Se podrá extraer **relaciones semánticas dinámicas,** como por ejemplo la realización de compras.

La Lista de Conceptos resultado del parsing es normalizada, incluyendo lo que llamaremos **Forma Normalizada del Concepto.**

Los procesos descritos en la introducción, que se detallan a continuación, serán también a su vez herramientas integradas en el sistema de vigilancia y que podrán ser tanto invocadas por un usuario de forma aislada como integradas en una aplicación.

Finalmente, se enviará a un **proceso de mantenimiento** los conceptos que no están en la ontología, de cara a que se evalúe la conveniencia o no de incluirlos en los repositorios de conocimiento y de ser así, que sean incluidos.

Análisis semántico PLN

El proceso más importante contenido en la Plataforma es el **Motor de Análisis Semántico**. Su objetivo es identificar la estructura semántica de los textos que son procesados.

Para ello utiliza el proceso de Análisis Sintáctico y se integra con los Recursos Lingüísticos y de Conocimiento para identificar estructuras en el texto mediante heurísticas. Asimismo contiene las **Gramáticas de Dependencias.**

Dentro de este proceso es necesario para realizar el análisis lingüístico configurar qué elementos, artículos y preposiciones normalmente, son incluidos a la hora de delimitar los sintagmas a obtener.

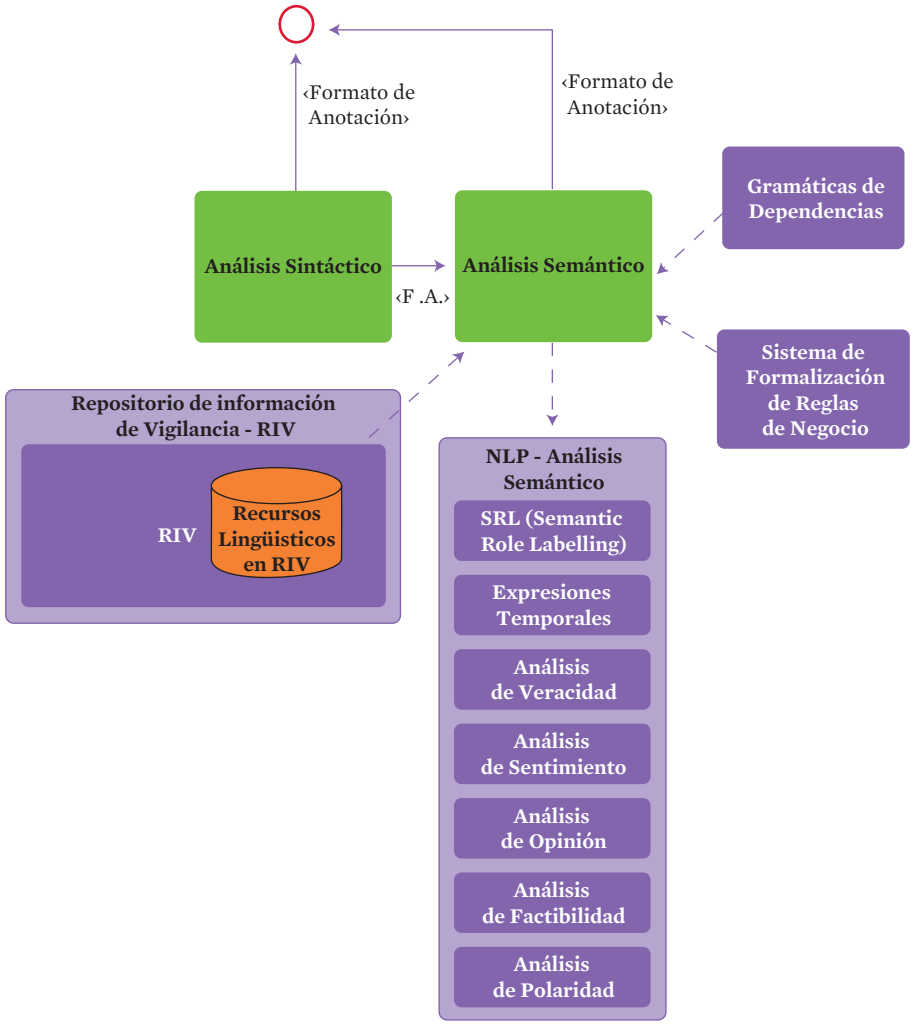
Incluye un **Sistema de formalización de reglas de negocio**, que permite (habitualmente por medio de ficheros de texto estructurados) identificar y definir en la gramática reglas que no son necesarias desde el punto de vista sintáctico pero que definen estructuras relevantes desde el punto de vista de negocio (por ejemplo, "qué es una opinión"). Otras situaciones habituales es cómo se expresan las opiniones positivas y negativas, creencias, actitudes; indicadores de sentimiento, como intensificadores, debilitadores y cambiadores. Esta funcionalidad es determinante para la creación de aplicaciones, como las siguientes:

- **Análisis de Sentimiento.**
- **Análisis de Opinión.**
- **Análisis de Factibilidad.**
- **Análisis de Polaridad.**

A un nivel más básico se encuentran otras aplicaciones de análisis semántico:

- **Semantic Role Labelling (SRL):** detecta argumentos asociados con predicados.
- **Expresiones Temporales:** identifica expresiones temporales mencionadas en el texto.
- **Análisis de Veracidad:** genera una valoración sobre si los hechos referidos en el texto han ocurrido o no o si hay algún nivel de incertidumbre en los mismos.

ILUSTRACIÓN 30
Detalle proceso Análisis Semántico



6.7. Integrando información en la Interfaz de Usuario

6.7.1. Integrando y analizando datos

Google Fusion Tables[133] permite integrar grandes tablas de datos provenientes de datos públicos o datos privados y crear visualizaciones dinámicas, accesibles en internet de

[133] Google Fusion Tables Help https://support.google.com/fusiontables/answer/2571232

forma pública, privada o compartida, que permitan realizar análisis de los datos y obtener conocimiento de dicho análisis a la vez que se protegen los datos de ser modificados.

ILUSTRACIÓN 31

Visualización del presupuesto del condado de Cook

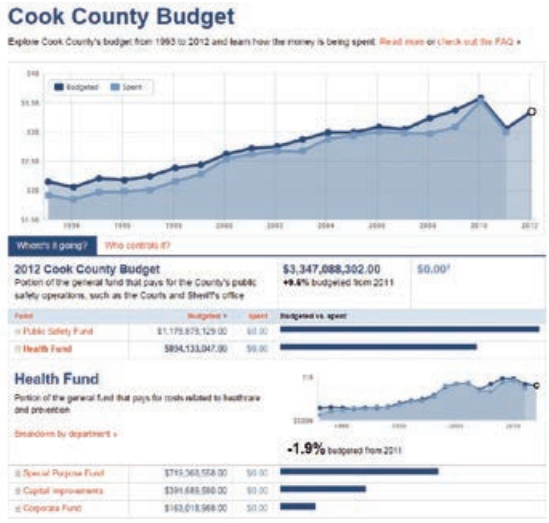
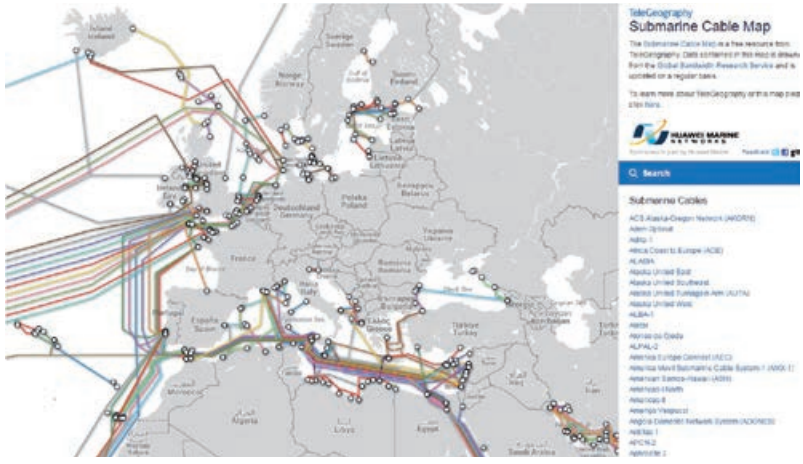


ILUSTRACIÓN 32

Visualización de Mapas de Cables Submarinos de TeleGeography utilizando Google Fusion Tables[134]



[134] TeleGeography Submarine Cable Map http://www.submarinecablemap.com/

6.7.2. Visualización de información clasificada por Repositorios de Conocimiento

Google Knowledge Graph utiliza una base de conocimiento para ofrecer **resultados enriquecidos con información semántica** adicionalmente al resultado del buscador estándar. Utiliza diversas Fuentes de Información, destacando la Wikipedia, Freebase y la CIA World Factbook.

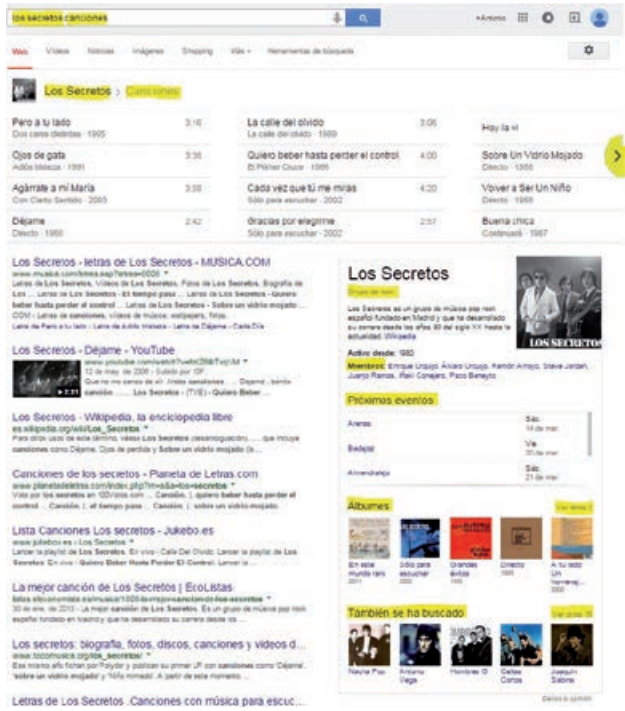
Permite además la interacción y la navegación a través de los resultados semánticos, por lo que frecuentemente en el primer vistazo se consigue resolver el objetivo de la búsqueda realizada.

ILUSTRACIÓN 33

Resultados enriquecidos por Google Knowledge Graph como respuesta a la búsqueda en Google de «comunidad de madrid población»

ILUSTRACIÓN 34

Aplicación de Google Knowledge Graph para la búsqueda «los secretos canciones»



6.7.3. Visualización Avanzada de Datos

Debe proporcionarse funcionalidad que permita crear gráficos que representen los datos de tal manera que se habilite el descubrimiento de conocimiento dentro de los mismos mediante el análisis visual. La diferente funcionalidad habilitará diferentes tipos de representación, cada una de ellas adecuada para un tipo de fin.

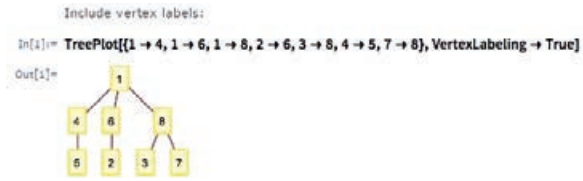
Asimismo debe poder representarse datos de diferente naturaleza, desde números continuos, **números discretos, textos, redes de datos, grafos, datos geoespaciales, etc.**

ILUSTRACIÓN 35

Visualización en mapa de los terremotos en California desde 1980 a 2014[135]

ILUSTRACIÓN 36

Visualización en forma de árbol de datos de enlaces de nodos[136]



6.8. Caso de Uso: aplicación a la Contratación Pública. Entendiendo la importancia de las Ontologías

Conocer la demanda de productos y servicios en un segmento de un sector así como las empresas que lo ofertan y lo demandan es un conocimiento muy interesante para múltiples stakeholders: las empresas que los comercializan, las que compran dichos servicios o para gestores públicos encargados de decidir si se realizan o no ayudas públicas en el sector.

[135] URL de la visualización: http://reference.wolfram.com/language/ref/GeoGraphics.html

[136] URL de la visualización: http://reference.wolfram.com/language/ref/TreePlot.html

Abordar el problema del mercado completo no es viable pero tal vez sí es posible abordar el estudio del sector público y extraer de ahí análisis y conclusiones válidos para el sector completo.

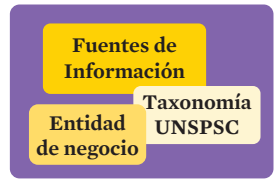
Un enfoque posible es estudiar las plataformas de Contratación Pública. En España tenemos la Plataforma de Contratación del Estado[137], a nivel europeo tenemos la plataforma TED[138] (Tenders Electronic Daily). En cada país vamos a encontrar una o varias plataformas de contratación. La integración de la información de las plataformas puede proporcionar una visión bastante buena de nuestros objetivos.

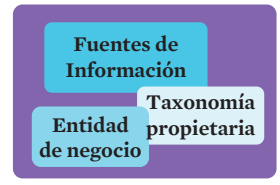
Si integramos toda esa información en un repositorio y procedemos a analizarla podemos conocer la oferta y la demanda de productos y servicios, el mercado de compradores y vendedores, o extraer tendencias y conocimiento a utilizar en múltiples casos de uso, un conjunto importante de funcionalidades que claramente están dentro del ámbito de la Vigilancia y la Inteligencia Competitiva.

Vamos a plantear un caso, incluyendo varios de los problemas que nos encontramos habitualmente, con tres fuentes de información:

- Una plataforma de contratación, que utiliza los códigos **CPV** (Common Procurement Vocabulary).
- Una segunda plataforma, que usa los códigos **UNSPSC** (United Nations Standard Products and Services Code).
- Disponemos adicionalmente de una tercera base de datos con productos, servicios y empresas del sector que queremos estudiar. El nivel de detalle en la clasificación que tenemos de productos y servicios es muy superior al que tienen las otras dos clasificaciones.

Parece que tenemos todo lo que necesitamos. Sin embargo hay que resolver la cuestión de que cada fuente de información utiliza un sistema de clasificación de los contratos diferentes y no nos es posible cruzar la información.





[137] Plataforma Contratación del Estado Español: https://contrataciondelestado.es/wps/portal/plataforma

[138] TED Tenders: http://ted.europa.eu/TED/main/HomePage.do

La ventaja de los códigos internacionales es que a veces existen también traducciones públicas entre las clasificaciones pero además hemos planteado que la tercera base de datos tiene una **clasificación propietaria**, con un nivel de detalle superior al de las otras dos clasificaciones, algo muy habitual ya que las dinámicas del mercado habitualmente superan enormemente a la velocidad de actualización de estas taxonomías.

Deberemos tener en cuenta además que no es infrecuente que haya errores o inexactitudes en la clasificación que se le otorga a cada concurso en las Plataformas de Contratación, es decir, clasificando con un nivel de abstracción superior al necesario. Por ejemplo podría estar clasificado como "Sistemas de Información" cuando se puede concretar mucho más incluyendo por ejemplo "Seguridad" y dentro de "Seguridad" otra clasificación de nivel inferior como "Sistema de detección de intrusos – IDS".

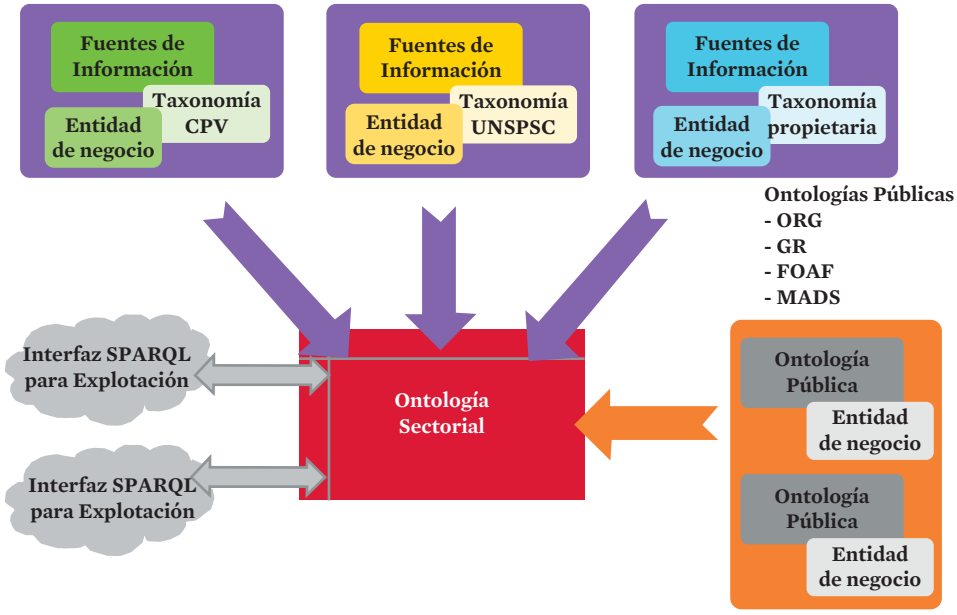
Este es un caso claro que podemos abordar con dos tecnologías: **Ontologías** y **Procesamiento de Lenguaje Natural (PLN)**. La aplicación de PLN sería sobre un campo resumen descriptivo que suele encontrarse en todos los concursos, que suele tener una descripción bastante razonable y que se denomina **"Objeto de Contrato".** Podemos **extraer Conceptos** de este Objeto de Contrato que nos ayuden a clasificar correctamente el Contrato a través de los productos y servicios que se deduzcan de dicho Contrato. No abundaremos más sobre PLN.

¿Qué nos va a proporcionar una Ontología para abordar toda esta problemática?

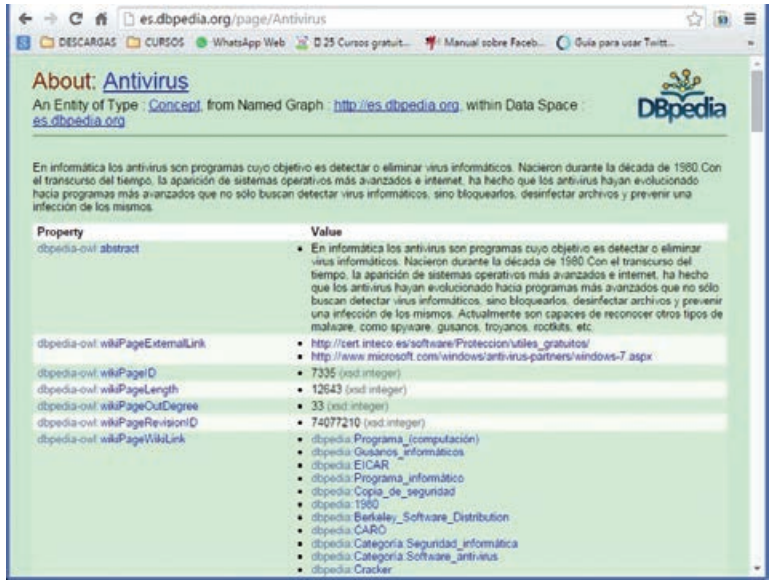
En primer lugar la ontología va a **conectar las tres clasificaciones, las tres Taxonomías**. Gracias a esta conexión no solo las fuentes con las que estamos trabajando sino cualquier fuente que utilice CPV y UNSPSC va a poder ser cruzada con el resto de fuentes, multiplicando así la cantidad de información disponible. Asimismo también va a ser relacionada con nuestra base de datos con toda nuestra información detallada del sector a estudiar. Hay que aclarar que este proceso de conexión ha de hacerse a mano o como mucho con métodos semi-automáticos y que han de participar tanto un experto en ontologías como un experto en la materia en cuestión.

En segundo lugar nuestra aplicación PLN (y cualquier otra aplicación) **consultará la ontología para determinar qué conceptos contenidos en el campo "Objeto de Contrato" son del ámbito del sector a estudiar**. Si extendemos la ontología al sector TIC y a conceptos comunes de productos y servicios ("Consultoría", "Auditoría", "Mantenimiento", etc.) podremos llegar a tener una conceptualización bastante ajustada de la descripción contenida en el campo "Objeto de Contrato".

ILUSTRACIÓN 37
Arquitectura funcional del Caso de Uso



En tercer lugar, la ontología nos ayudará a **relacionar conceptos, tanto dentro de la propia ontología como con otros conceptos usados en otras ontologías**. Para ello integramos nuestra ontología con ontologías públicas (FOAF, ORG, GR, MADS, etc), que ya determinan algunos conceptos y se comparte su significado a nivel global por Internet. Asimismo usamos estándares (SKOS, RDF, OWL, etc) que nos van a permitir decirle a la ontología que si no nos puede dar una respuesta exacta con una búsqueda exacta (en SPARQL se denomina "ExactMatch") o nos pueda dar algo más aproximado (en SPARQL se denomina "NarrowMatch").



En cuarto lugar la ontología permite el **análisis y extracción de conocimiento** mediante navegación a través de la misma, si aplicamos determinados paquetes software, como Pubby, para la navegación. Un buen ejemplo de esta interfaz podemos verla en la DBpedia, que extrae la información disponible de Wikipedia y presentando una estructura clasificada sobre la que podemos navegar.

Las imágenes de este apartado están capturadas de la entrada de "Antivirus" de la DBpedia, cuyo contenido podemos contrastar de la entrada de la Wikipedia de la que captura la información: http://es.wikipedia.org/wiki/Antivirus.

En las imágenes podemos ver referencias a ontologías públicas como SKOS o a FOAF.

5

FORMALIZACIÓN DEL MODELO Y LA METODOLOGÍA

Presentamos en este apartado, de una manera formal:

- Los Elementos del Modelo, referenciándolos a los modelos ya presentados en el libro.
- La Metodología que proponemos para el diseño del Sistema de Vigilancia Estratégica e Inteligencia Competitiva.
- El Modelo Organizativo que se propone.

1. Elementos del Modelo

En el apartado 3.1 "V" de Big Data y en el apartado 3.10 "Mapeando las tecnologías Big Data y las actividades de Vigilancia Estratégica e Inteligencia Competitiva" se ha presentado el modelo de alto nivel que se propone tanto para Big Data como para su alineamiento con los Sistemas de Vigilancia e Inteligencia. También es referencia la Arquitectura Funcional que se presenta en la introducción del apartado 4 "Diseñando Sistemas de Vigilancia e Inteligencia con nuevas Capacidades Big Data".

Se listan a continuación, de forma clasificada, los elementos del modelo. Se incluye una pequeña presentación de cada ítem, a título de introducción. En el siguiente punto, "Metodología de Diseño del Sistema de Vigilancia", se explica la necesidad de cada uno de los ítems.

Infraestructura de negocio: Casos de Uso y Necesidades

- **Biblioteca de Necesidades y Casos de Uso:** necesidades de negocio y Casos de Uso en los que se convierten. Incluye también los Casos de Uso categorizados como no viables.
- **Plantilla de Casos de Uso Base:** plantilla de casos de uso para facilitar las entrevistas con los entrevistados.

- **Biblioteca de Requisitos de Vigilancia:** Requisitos gestionables en que se descomponen los Casos de Uso.
- **Modelo organizativo y estratégico:** propuesta organizativa fundamentada en un modelo de análisis estratégico.

Repositorios de Conocimiento

- **Biblioteca de Fuentes:** conjunto de fuentes de Información estructuradas y no estructuradas, incluyendo tanto fuentes públicas como fuentes privadas a la organización.
- **Biblioteca de Taxonomías:** conjunto de taxonomías asociadas a las Fuentes de Información.
- **Biblioteca de Ontologías:** conjunto de ontologías utilizadas en el sistema, incluyendo ontologías públicas.
- **Diccionarios NLP.** Diferentes diccionarios (sinónimos, antónimos, sentimientos, etc) utilizados para las aplicaciones de procesamiento de lenguaje natural.
- **Bases de Conocimiento.** Otros Repositorios que organicen conocimiento sobre temas concretos.

Modelo de Información

- **Biblioteca de Entidades de Negocio** Principales, Secundarias y de Relaciones estructuradas entre Entidades. Incluyendo su relación con taxonomías y ontologías.
- **Estudio de Alineamiento** de los Requisitos con las Entidades de Negocio y las capacidades de la infraestructura técnica disponible.
- **Modelo de Información** implementable.
- **Diagrama Entidad Relación.** Incluyendo tanto Entidades Principales como Secundarias.

Infraestructura técnica. Bases de Datos, Herramientas y Aplicaciones

- **Bases de Datos:**
  - Bases de Datos Relacionales / Repositorios.
  - Cubos / Datamarts / Datawarehouse: business intelligence tradicional.

- Bases de Datos NoSQL, tipo Big Data.

- **Librería de Herramientas:**
  - Para la **extracción**, **carga**, **limpieza**, **enriquecimiento** y **transformación** de información estructurada.
  - Para el **crawling** y **scraping** de Fuentes de Información no estructurada.
  - Otras herramientas adicionales necesarias.

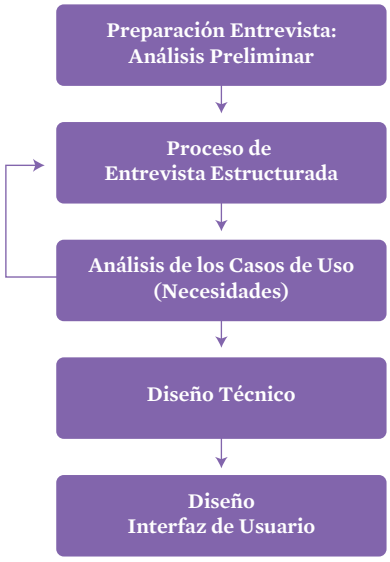
- **Librería de Aplicaciones:**
  - Para el tratamiento inteligente de la información, incluyendo las nuevas capacidades presentadas en el apartado de "Nuevas Capacidades Big Data": procesamiento de lenguaje natural (**PLN**), **machine learning**, aplicaciones estadísticas y aplicaciones para diseñar y desarrollar este tipo de aplicaciones.
  - **Librería de Aplicaciones avanzadas** para la búsqueda e interacción con las Bases de Conocimiento, como Buscadores, Navegadores y Lenguajes de Consulta y Capacidades para diseñar nuevas apps.
  - **Librería de aplicaciones analíticas**, incluyendo tanto las capacidades tradicionales de business intelligence y más novedosas denominadas de business analytics.

- **Biblioteca de Funcionalidades**: recoger conocimiento especializado para abordar tipos de problemas similares.
- Modelo de ejecución de aplicaciones.
- Herramientas para la **Administración y Monitorización del Sistema.**

Interfaz de Usuario. Data Visualization (DataVIZ).

- Modelo de Usuarios.
- Librería de Representaciones Visuales.
- Librería de Querys al Sistema.
- Estrategia de presentación de datos: Data Crystallization.

2. Metodología de Diseño del Sistema de Vigilancia / Inteligencia



Se presenta a continuación una propuesta de metodología basada en 5 procesos: Preparación de Entrevista y Análisis Preliminar, Entrevista Estructurada, Análisis de Casos de Uso, Diseño Técnico y Diseño de la Interfaz de Usuario. A la vez se van introduciendo los Elementos del Modelo, presentados en el punto anterior, según van siendo necesarios. Preparación Entrevistas: Análisis Preliminar.

Tras identificar a los interlocutores que serán entrevistados se propone en primer lugar la realización de una **formación o presentación de las Capacidades Big Data del Sistema de Vigilancia**, con el objetivo de que puedan tomar conciencia de las posibilidades y limitaciones de las mismas. De esta manera se facilitará que surjan nuevos casos de uso en las entrevistas que probablemente no hubieran surgido de no tener conciencia el entrevistado de estas capacidades.

Para la preparación de estas entrevistas se realizará un **estudio previo del sector objeto del diseño del sistema de vigilancia**, con el objeto de tener una base de elementos susceptibles de ser parte del sistema:

- Fuentes de Información y las Taxonomías que las clasifiquen.
- Entidades de Negocio Principales y Secundarias y sus Relaciones Estructurales.
- Ontologías públicas reutilizables sobre las Entidades de Negocio.

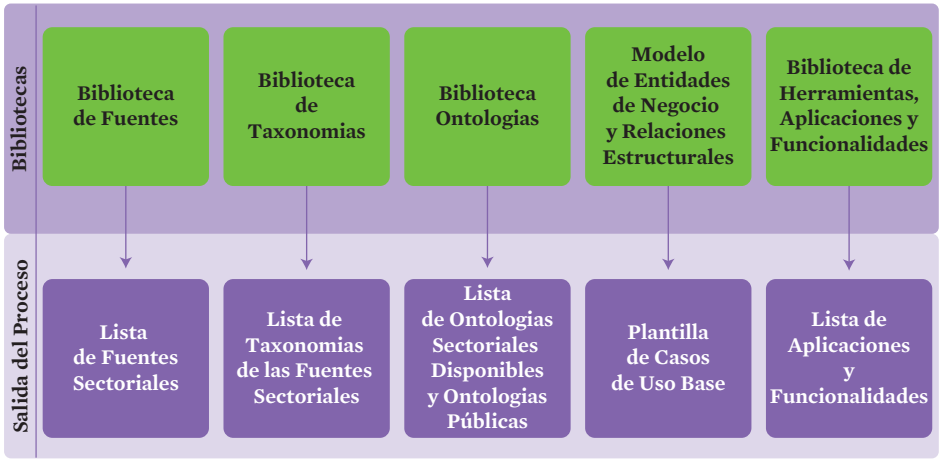
Construiremos una **plantilla de casos de uso base** que tendrán una triple función:

- Encajar en los mismos los casos de uso educidos durante las entrevistas.
- Ser presentados a los entrevistados con el objetivo de contrastar si responden o no a sus necesidades y con qué prioridad e importancia.
- Tener un superconjunto base orientado a cumplir el objetivo de completitud del conjunto de casos de uso.

Para la construcción de los casos de base usaremos dos conjuntos de casos de uso:

- La intersección de las Entidades de Negocio Principales y las Relaciones Estructurales.
- Los casos de uso más habituales de la Vigilancia Estratégica y la Inteligencia Competitiva.

Fase 1. Preparación Entrevistas; Análisis Preliminar



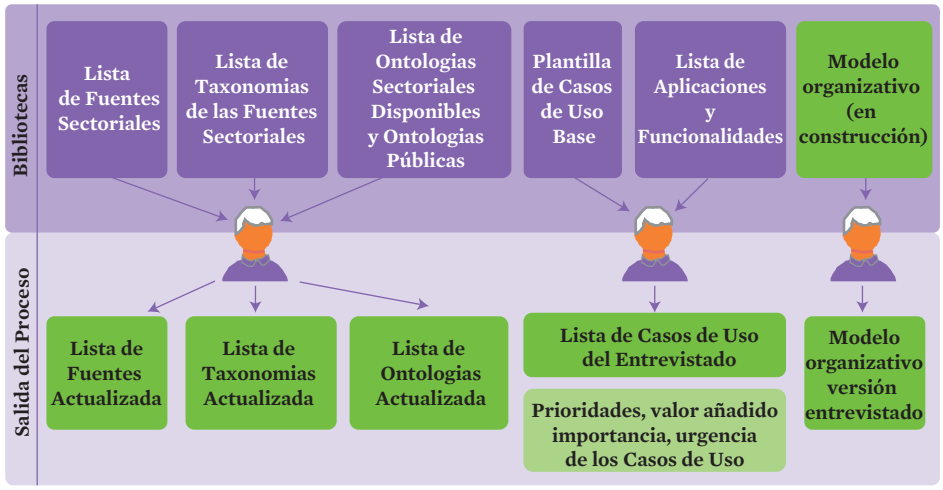
Las otras herramientas que se confeccionarán para el desarrollo de las entrevistas, haciendo una selección a partir de sus respectivas Bibliotecas, serán:

- Lista de Fuentes y Taxonomías.
- Lista de Ontologías.
- Lista de aplicaciones.
- Biblioteca de funcionalidades.

De ambas listas se habrá de disponer de especificaciones y catalogación completa que llamaremos Biblioteca de Fuentes, Biblioteca de Taxonomías, Biblioteca de Herramientas y Aplicaciones y Biblioteca de Funcionalidades con su nivel de viabilidad y calidad.

2.1. Proceso de Entrevistas Estructuradas

Fase 2. Realización Entrevistas Estructurales



En cada entrevista nuestro objetivo principal será obtener el conjunto de necesidades y casos de uso correcto y completo del interlocutor objeto de la entrevista. Para ello:

- Se recogerán de los entrevistados los casos de uso y se intentarán encajar en la plantilla de casos de uso base.
- Se revisará la plantilla de casos de uso con el cliente, **buscando nuevos casos de uso** que el interlocutor no haya expresado.
- Recoger **prioridades** y datos que permitan cuantificar la **importancia** y el **valor añadido** que cada caso de uso proporciona, así como la **urgencia** de disponer de la implementación del mismo.
- Se determinará el **Modelo Organizativo** de los futuros usuarios del sistema. Los Casos de Uso quedarán clasificados según dicho modelo organizativo.

Durante el proceso se recogerán también propuestas concretas que el interlocutor realice para resolver sus casos de uso utilizando las nuevas capacidades de los sistemas de vigilancia.

En segundo lugar se obtendrá del interlocutor el conjunto de **Fuentes que el usuario conozca o resulten relevantes** tanto para sus casos de uso como para su entorno en general, ya que pueden ser útiles para resolver casos de uso de otros usuarios, contrastando con él la Lista de Fuentes y Taxonomías para obtener información experta sobre la **calidad de las Fuentes y Taxonomías incluidas**. También se preguntará de

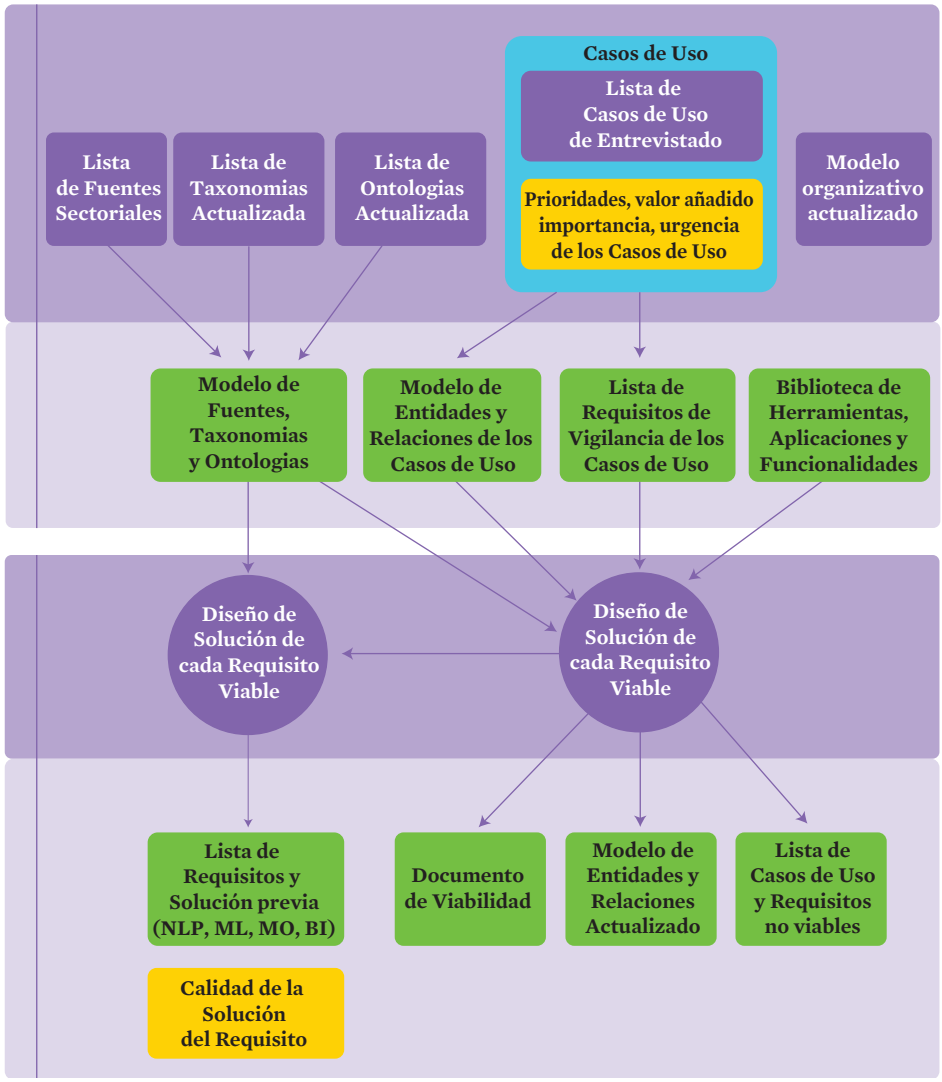
forma abierta por otros componentes del Sistema de Vigilancia, especialmente Taxonomías y Ontologías.

2.2. Análisis de los Casos de Uso

Tras cada entrevista deben analizarse los Casos de Uso con los objetivos de:

- Validar o, en su caso, modificar nuestro modelo: se revisará que las Entidades de Negocio encajan en nuestras clasificaciones de Estructural y Secundaria así como en el conjunto de Relaciones Estructurales. Si no es así, se modificará el modelo.
- **Desglosar los Casos de Uso en Requisitos**, obteniendo una **Lista de Requisitos de Vigilancia** organizada según sus Entidades de Negocio y Relaciones Estructurales. Este paso es clave, porque nos permitirá diseñar y construir una capa con un nivel de abstracción menor que el de los Casos de Uso y aportando sentido funcional completo. A este nivel nos encontraremos con Requisitos que serán denominadores comunes entre varios casos de uso.
  - Los requisitos deberán ser unidades funcionales completas, consistentes, traceables, comprobables, claras (sin ambigüedades y no interpretables) y tener un tamaño manejable para su gestión.
  - Se prepararán los requisitos para poder ser solucionados mediante técnicas de NLP y Machine Learning.
- Se estudiará la **viabilidad de diseñar una solución para cada Requisito** y se cualificará la calidad de la respuesta diseñada (datos exactos disponibles, datos open data, datos obtenidos de fuentes no estructuradas, datos provenientes de soluciones machine learning...).
  - Se diseñará la **solución de cada Requisito** para que pueda ser resueltos mediante la Lista de Aplicaciones y Funcionalidades, la información que pueda estar disponible a partir de las Fuentes existentes y los elementos del Modelo en general.
    - Si no se dispone de la aplicación necesaria se planificará su diseño y se ampliará la Biblioteca de Aplicaciones.
    - Si no se dispone de las Fuentes de Información se realizará una búsqueda prospectiva de los mismos y se ampliará la Biblioteca de Fuentes.
    - Si alguno de los dos puntos anteriores no son viables se anotará en la **Lista de Requisitos y Casos de Uso no viables**, para los que se buscarán alternativas en principio no automatizadas.

Fase 3. Análisis de los Casos de uso de la Entrevista



Tras el análisis de la última entrevista dispondremos de:

- Un conjunto de Necesidades y Casos de Uso clasificados por Interlocutor.
- El desglose de los Casos de Uso en Requisitos.
- El análisis de los Requisitos con la solución diseñada para el mismo, incluyendo las entidades involucradas, las Fuentes de Información origen de los datos a utilizar en las mismas y las aplicaciones involucradas en la misma.

- La calidad y características esperables de cada Caso de uso en función de los de los Requisitos en los que se desglosa.
- El modelo de Entidades de Negocio, Entidades Secundarias y Relaciones Estructurales. Con este modelo construiremos el Modelo de Información del Sistema.
- La lista de Casos de uso y Requisitos no viables.
- El Modelo Organizativo.

2.3. Diseño técnico

El diseño técnico estará orientado a la solución de los Casos de Uso, aunque la metodología propuesta permitirá disponer de la solución al conjunto de la Lista de Requisitos por separado, ya que desglosamos los Casos de Uso en Requisitos.

Partiremos del análisis de viabilidad realizado sobre el Caso de Uso, que nos proporciona una calidad a priori de la solución del mismo. Es importante destacar que es un análisis a priori y que será en la fase de diseño técnico en la que se determinará la calidad real de la misma, normalmente tras realizar pruebas de concepto adecuadas.

Se diseñará la solución técnica de cada Requisito en que se desglosan los Casos de Uso y se diseñará asimismo la integración de cada una de las soluciones para constituir la solución del Caso de Uso.

Para ello se realizarán los siguientes grupos de tareas:

- **Determinar las Fuentes de Información necesarias para obtener los datos.** Puede ser nuevas fuentes de información o también información ya disponible en los repositorios internos del sistema. Se deberá completar la información disponible sobre la Fuente en el Documento de Fuentes, mediante la plantilla disponible en la Biblioteca de Fuentes.

- **Diseño de la integración de los Datos de las Fuentes.** Se estudiará en profundidad la calidad y características de los datos contenidos en las fuentes y se hará el diseño técnico de la aplicación de descarga de información. Típicamente se tratará de una herramienta de ETL tradicional de Business Intelligence para fuentes estructuradas o una herramienta de Crawling & Scraping para fuentes no estructuradas. También nos podemos encontrar con contenidos mixtos, lo que conllevará una solución combinada.

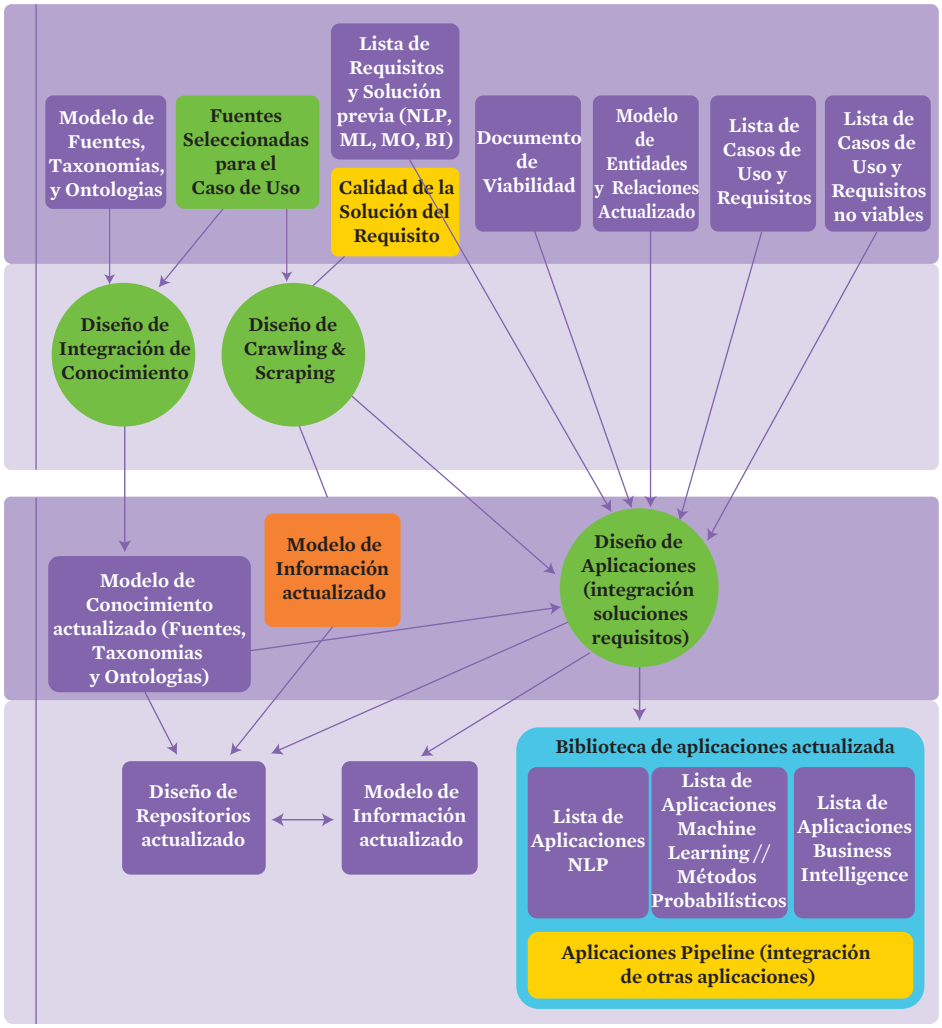
- **Diseño de la integración del Conocimiento de las Fuentes.** Cada fuente de datos puede disponer, por ejemplo, de taxonomías que clasifiquen sus datos o diccionarios que desambigüen los mismos. Asimismo los conceptos recogidos en las fuentes pueden tener matices más o menos importantes sobre el mismo concepto recogido en nuestro DiseñoSistema de Vigilancia.

- **Revisión del Modelo de Información del Sistema.** Se realizará una revisión del Modelo de Información existente, teniendo en cuenta los nuevos campos, la calidad de los datos y las cuestiones técnicas que surjan de los Diseños de Integración de Datos y Conocimiento en el Sistema. Esto puede conllevar por ejemplo la creación de nuevos campos en una Entidad, bien de carácter generalista o particular a la fuente o, en principio debería ser de forma excepcional, la creación de nuevas Entidades de Negocio.

- **Revisión del Modelo de Conocimiento:** en el Sistema de Vigilancia se habrán diseñado y/o integrado diferentes sub-sistemas de gestión del conocimiento sobre los objetos de vigilancia, por ejemplo diccionarios, taxonomías y ontologías. Concretamente en nuestro Sistema se propone la creación de una Ontología que integre estos sistemas de conocimiento, permitiendo así cruzar la información que esté clasificada por diferentes Taxonomías. Asimismo puede ser necesaria la ampliación de la Ontología.

- **Diseño de Aplicaciones.** Las aplicaciones serán desarrolladas teniendo en cuenta tanto el desglose en requisitos de cada caso de uso, generando la solución a cada requisito con objeto de poder reutilizarla a posteriori, como la Integración de las soluciones resultantes montando un pipeline de aplicaciones o incluso una aplicación tipo BPM. Se utilizarán tanto técnicas y herramientas tradicionales de Business Intelligence, Buscadores, Alertas, como técnicas y herramientas punteras basadas en Procesamiento de Lenguaje Natural, Machine Learning, Métodos Probabilísticos, Ontologías, Escenarios, Herramientas avanzadas de explotación y Big Data en general. Cada nueva aplicación será almacenada en la Biblioteca de Aplicaciones para su reutilización.

Fase 4. Diseño técnico



- **Repositorios y Modelo de Información**: cada tipo de aplicación conllevará un tipo concreto de repositorio con sus características específicas.
  - Las aplicaciones de BI pueden necesitar cubos específicos al caso de uso, aunque deberá valorarse la bondad y oportunidad de generar Datamarts más generalistas y reutilizables, orientándolos a las Entidades de Negocio Estructurales.
  - Las aplicaciones que hagan uso intensivo de documentos, como pueden ser las de Métodos Probabilísticos y/o técnicas de Machine Learning, con toda probabilidad necesitarán hacer uso del sistema de ficheros distribuido y la ejecución de procesos MapReduce.

- Las aplicaciones NLP necesitarán bases de datos orientadas a columnas. Finalmente, los repositorios de conocimiento (ontologías, taxonomías, etc) se almacenarán en formato RDF en repositorios Triple Store, accesibles vía SPARQL, en bases de datos NoSQL orientadas a grafos.
- La integración de fuentes clasificadas por diferentes taxonomías conllevará, si aparece una nueva taxonomía, la actualización de la ontología con el objetivo de hacer viable el cruce de datos.
- El Modelo de Información base deberá construirse a partir de las Entidades de Negocio Estructurales, las Secundarias y las Relaciones Estructurales. A partir de la información recogida en la Fase 2 deberá realizarse el diseño técnico definitivo incluyendo las Entidades, Campos y Relaciones que se deduzcan.

2.3.1. Calidad de la solución del diseño técnico

Durante el diseño técnico pueden surgir diferentes cuestiones que resten calidad a la solución del caso de uso. Sin intención de hacer una lista exhaustiva, a continuación se presentan un conjunto de situaciones habituales:

- Series de datos incompletos o sucios en fuentes de información estructuradas.
- Incapacidad del software y/o las técnicas de scraping para proporcionar información estructurada correcta y/o completa.
- Se obtiene información mediante sw y técnicas de scraping pero sin garantías sobre su corrección o completitud.
- Taxonomías de granularidad inferior a la necesaria para clasificar los datos de una fuente.
- Cruces de datos entre fuentes con resultados de granularidad inferior a la deseada debido a taxonomías poco detalladas o no alineadas con los objetivos de negocio.
- Aplicaciones con capacidades NLP o Machine Learning que no proporcionan resultados exactos e incluso pueden no cumplir suficientemente los objetivos de negocio para los que fueron diseñados. Cualquier toma de decisiones debe tener en cuenta esta situación.

Es importante, por tanto, que cada aplicación pueda presentar información completa y relevante sobre la calidad de la información que proporciona.

2.3.2. Fuentes de Información

Cada Fuente de Información candidata a ser integrada en el sistema debe ser estudiada con el objetivo de determinar la calidad de la misma, en función de los parámetros que veíamos en el apartado de "Fuentes de Información y Taxonomías".

2.4. Diseño de la Interfaz de Usuario

El diseño de la Interfaz de Usuario vendrá determinado por:

- El **Modelo Organizacional**, que determinará qué usuarios tienen acceso a qué funcionalidad y datos del sistema, según los casos de uso.
- Los **Casos de Uso** de cada Organización e usuarios de la misma.
- Las **Entidades Estructurales** del Sistema. Una línea de diseño consistirá en tener un Módulo por cada una de las Entidades Estructurales, y organizar en torno a ellas la funcionalidad requerida por los Requisitos.
- Las **herramientas** y **aplicaciones** disponibles en el sistema.
- Otros componentes del Sistema, como son las Fuentes de Información, las Taxonomías, Ontologías y otras Bases de Conocimiento y los Repositorios de Datos.
- La función que ejerza en la organización el usuario del sistema, que determinará las capacidades que tendrá, por ejemplo, desde diseñar nuevas aplicaciones a únicamente a visualizar y analizar datos.

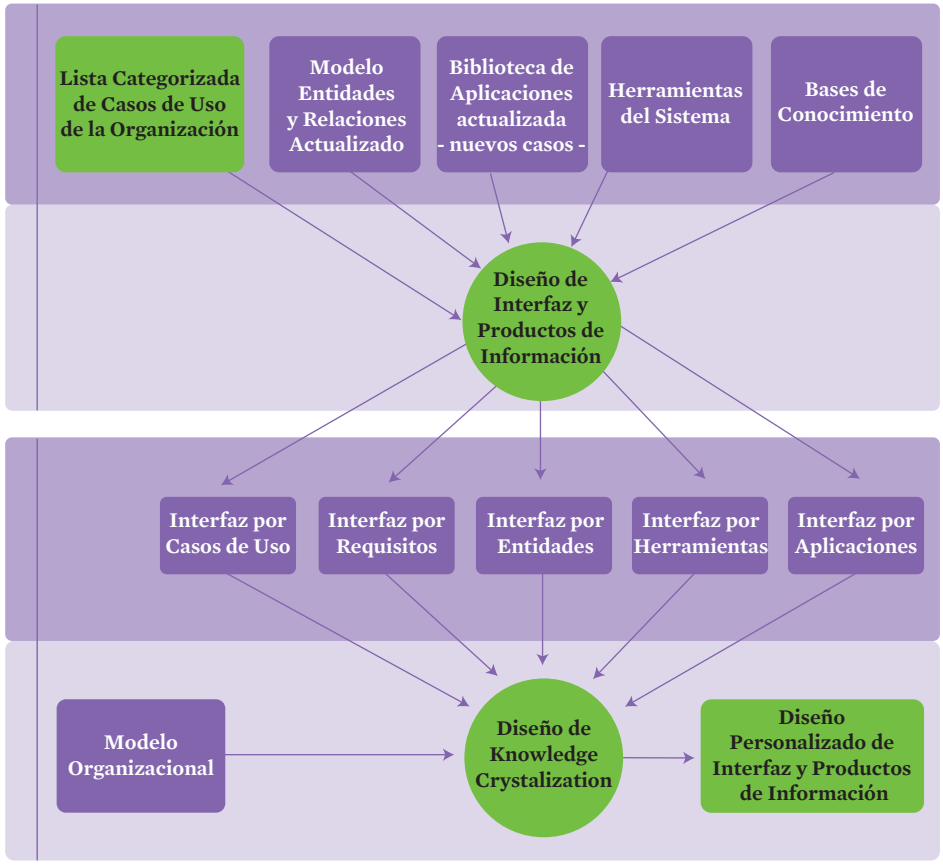
La interfaz de usuario deberá construirse por tanto de tal manera que se permita el **acceso diferenciado a Casos de Uso, Entidades de Negocio y sus Relaciones Estructurales y las Herramientas/Aplicaciones**, por ejemplo a través de menús, cintas o pestañas diferenciadas. Asimismo se propone un **siguiente nivel de desagregación**, de tal manera que desde cada uno de estos elementos se acceda a los elementos relacionados. Por ejemplo, desde los Casos de Uso que se pueda acceder a las Entidades de Negocio relacionadas y consecuentemente a la funcionalidad disponible sobre las mismas.

Típicamente el sistema proporcionará diferentes grupos de datos, que deberán presentarse en **Mashups**, es decir, en aplicaciones **que integran en una única interfaz gráfica** información provenientes de diferentes tipos de fuentes de datos. Asimismo concurre la circunstancia de que la información contenida en cada fuente de datos pueda ser incluso contradictoria entre si o que tenga diferente relevancia según el usuario y su posición en el Modelo Organizacional. Las herramientas y aplicaciones

utilizadas para resolver la solicitud del usuario darán forma al mashup, en función de los tipos de salidas que proporcionen.

Consecuentemente debido a las diferentes naturalezas, calidades y variedades de los datos que están almacenados en el Sistema de Vigilancia será necesario definir una **estrategia de presentación de los datos** orientada a los objetivos y prioridades que hayan quedado determinados en el Modelo Organizacional. A este tipo de estrategias se les llama ***knowledge crystallization***. Se prioriza el ofrecer una descripción lo más compacta posible sin borrar información crítica, priorizando según niveles de confianza que deben otorgársele a cada tipo de datos y perfil o grupo humano de la organización. Este concepto estaría relacionado con la 4ª "V", de Big Data, la relativa a la Veracidad de los datos, que es un concepto que debería aumentar su alcance para incluir los matices que hacen a unos datos más relevantes e importantes que otros para determinados fines, objetivos y personas.

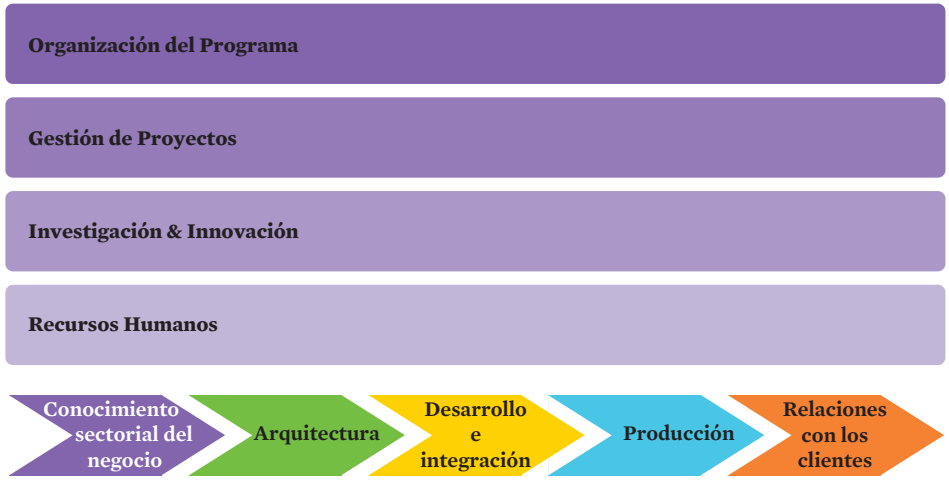
Fase 5. Diseño Interfaz Usuario y Productos de Información



3. El Modelo organizativo

De forma similar al modelo de Cadena de Valor de Porter, estructuramos nuestra propuesta de organización en torno a dos grupos de actividades: un grupo de actividades principales y un grupo de actividades de soporte.

ILUSTRACIÓN 38
Propuesta de Modelo Organizativo inspirado en la Cadena de Valor de Porter



3.1. Actividades Principales

Los proyectos del Sistema de Vigilancia parten de un conjunto de actividades derivadas de obtener un Conocimiento Sectorial y de los objetivos de vigilancia. A partir de dicho conocimiento se diseña una Solución Avanzada, que tenga en cuenta las nuevas capacidades postuladas para el Sistema. A continuación se procede a la Implementación de dicha Solución, y a su puesta en Producción. Se plantean dos grupos de servicio al cliente: el servicio técnico, prestado desde Producción y la gestión de la experiencia de los clientes, prestado desde el grupo de Relaciones con los Clientes.

- **Conocimiento Sectorial del Negocio:** en este grupo de actividades se realiza, desde un punto de vista de negocio el análisis sectorial y de objetivos estratégicos del Sistema de Vigilancia. Entre las actividades de análisis se incluye el análisis de Casos de Uso y Necesidades, el Análisis de Fuentes, Taxonomías y Análisis de Viabilidad a nivel de negocio. Para realizar estas labores a menudo será necesaria la incorporación de expertos sectoriales para lo que será necesario realizar la actividad de gestión de los mismos como proveedores de servicio. Por último también se realizará aquí

la Gestión de Conocimiento del Negocio. Este grupo de actividades se relaciona con el perfil de "Curador de Contenidos" el apartado 2.1.2 "La Vigilancia tecnológica como actividad clave para la innovación".

- **Arquitectura:** en este grupo de actividades se realiza un análisis técnico multinivel, incluyendo un análisis de viabilidad técnica de implementación de los Casos de Uso y Necesidades previamente analizados a nivel de negocio. Su objetivo es, para cada Casos de uso, diseñar de forma trasversal aplicaciones completas, teniendo en cuenta todas las fases de proyecto, tecnologías, componentes y aplicaciones que vayan a intervenir. Asimismo diseñan y mantienen la arquitectura general del Sistema, haciendo viable y compatible la incorporación de nuevas soluciones, aplicaciones y herramientas al Sistema.

- **Desarrollo de Aplicaciones:** se parte del análisis técnico realizado en el eslabón de Arquitectura, se realiza un análisis técnico de detalle que se convertirá en uno o varios proyectos de desarrollo e integración de aplicaciones y de parametrización, configuración e integración de herramientas. Destaca por la variedad y novedad de las tecnologías, lo que hará más compleja tanto la gestión como la implementación de este grupo de actividades. Entre las tecnologías se incluyen los siguientes grupos: Business Intelligence, Big Data, NLP, Machine Learning, Métodos probabilísticos, crawling y scraping.

- **Producción:** el objetivo del grupo de actividades de Producción es asegurar el servicio fiable y sin fallos del Sistema completo en funcionamiento. Incluye las actividades de puesta en producción, explotación y mantenimiento del hardware y todos los componentes software que constituyen el Sistema base así como todas las aplicaciones y herramientas que vayan incorporándose al Sistema de Vigilancia. Esto requiere actividades especializadas aplicadas a diferentes niveles tecnológicos: Sistemas, Redes, Seguridad, Gestión de Red, Gestión de Aplicaciones y Gestión de Bases de Datos. Será necesario atender especialmente a algunas características derivadas del uso de las nuevas capacidades, para la Paralelización y los Sistemas Distribuidos.

  Las referencias marco para la producción serían la ISO 20.000[139], orientada a la **Gestión de Servicios IT**, la ISO 38500, orientada al Gobierno de IT y la librería de buenas prácticas estándar de facto ITIL[140] (en inglés "Information Technology Infrastructure Library") , dedicadas ambas a la **gestión de servicios TI**.

---

[139] ISO/IEC 20000:2011 IT Service Management and ISO/IEC 38500:2015 IT Governance http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=5013818

[140] Sitio web de ITIL.org http://www.itil.org/ También en Wikipedia http://es.wikipedia.org/wiki/Information_Technology_Infrastructure_Library

- **Relaciones con los Clientes:** de este grupo de actividades tenemos dos casos: cuando una organización, institución o empresa le da servicio a otra y el caso en el que un departamento interno le dé servicio a otro. En el primer caso se incluirá las actividades comerciales, de marketing y de servicio postventa al cliente y en el segundo caso sólo las de servicio post-venta, entendiendo que no será necesario comercializar o promocionar los servicios. Sin embargo es cierto que cada vez más, cuando servicios centralizados IT dan servicio a varias empresas de un grupo o a varios países, se están incorporando un cierto tipo de actividades similares a la comercialización y marketing. En cualquier caso se incluirán actividades para controlar y seguir el comportamiento del cliente (o usario), la formación de los usuarios y su interacción con el Sistema, especialmente debido a que varias nuevas capacidades de las aportadas por estos sistemas son realmente novedosas. Se pasará de un enfoque CRM (Customer Relationship Management) a CEM (Customer Experience Management), incluyendo para ello la observación de la experiencia que los clientes y usuarios tienen con el servicio para asegurarse el mejor aprovechamiento del Sistema.

3.2. Actividades de Soporte

- **Dirección y Organización del Programa:** el grupo de actividades de Organización del Programa se desarrolla en el siguiente apartado de "Puesta en macha mediante Program Management". Responde a la necesidad de orientar la organización a la consecución de Beneficios objetivos que trascienden a los entregables que proporciona cada proyecto. Incluye las actividades de Gestión Estratégica (planificación, ejecución y control estratégico). Se recogerá el conjunto de buenas prácticas en Program Management, como las propuestas por el PMI (Project Management Institute).
- **Gestión de Proyectos:** la organización por proyectos va a ser necesaria de forma intensiva en toda la cadena de valor, especialmente en el Desarrollo de Aplicaciones. Es por ello que resulta necesario una labor de soporte de Project Management. Dependiendo del tipo de proyecto, en situaciones más formales se basará en el marcos de mejores prácticas PMBOK o la metodología Prince2, frente a proyectos de cariz más explorador, o con más incertidumbre en el lado del cliente, situaciones para las que las Metodologías Ágiles pueden dar mejores resultados.

  Será necesario tener en cuenta la muy probable necesidad de externalización de proyectos a proveedores especializados, debido a la variedad de tecnologías y enfoques a utilizar. Eso implica una actividad especializada en la formalización de requisitos en forma de pliegos de contratación. A largo plazo, los modelos CMMI-

DEV[141] y CMMI-ACQ[142] pueden servir de referencia. A corto plazo, la exigencia de marcos maduros de procesos IT puede ir en contra de proveedores emergentes e innovadores que entreguen alto valor añadido, lo cual debe valorarse como contraproducente.

- **Investigación & Innovación:** el Sistema va a abordar tanto tecnologías nuevas o relativamente nuevas que además evolucionan a ritmo vertiginoso, junto a funcionalidades y casos de uso de solución novedosa, lo cual conlleva una actividad de investigación e innovación de soporte a toda la cadena de valor. Realizarán la búsqueda de soluciones y nuevos enfoques para los Casos de Uso considerados como "no viables", explorarán enfoques, productos, aplicaciones y servicios innovadores así como los límites de la tecnología.
- **Recursos Humanos:** se prevé un enorme crecimiento en la demanda de personal experto en tecnologías Big Data, una ciencia emergente y compleja, que va a requerir fuertes inversiones en formación del personal y que va a requerir esfuerzos en la fidelización del mismo. Los patrones que podemos encontrar en el medio plazo son similares a los de la burbuja de internet, caracterizados por competencia entre empresas para la captación del talento y sueldos crecientes. RRHH se presenta habitualmente como actividad de soporte en todas las cadenas de valor, en el caso que nos ocupa durante muchos años resultará un área clave.

4. Puesta en marcha mediante Program Management

Llamaremos **Programa**[143] al conjunto de proyectos interrelacionados que son gestionados de forma coordinada con el objetivo de obtener beneficios no alcanzables si se gestionan de forma individual. Los programas, además de **proyectos**, pueden contener otros componentes, como pueden ser **entregables de gestión** o trabajos de **operaciones**. Están vinculados al largo plazo y tienen un nivel de patrocinio alto o muy alto. Los proyectos entregan productos o servicios, sin embargo **los Programas entregan Beneficios**.

Con el conjunto de entregables integrados del programa implantaremos un nuevo conjunto de **Capacidades de la Organización,** con las que explotar **Oportunidades** que se materialicen en **Resultados** y como consecuencia que el Programa entregue **Benefi-**

[141] SEI. – Software Engineering Institute. CMMI-DEV CMMI para Desarrollo, Versión 1.3 http://www.sei.cmu.edu/library/assets/whitepapers/Spanish%20Technical%20Report%20CMMI%20V%201%203.pdf

[142] SEI – Software Engineering Institute. CMMI for Acquisition, Version 1.3 http://www.sei.cmu.edu/reports/10tr032.pdf

[143] The Standard for Program Management, PMI 2013 – Project Management Institute

cios a la Organización, o sea, mejoras medibles como consecuencia de los Resultados alcanzados. Las capacidades pueden ser Servicios, Funciones o incluso Operaciones.

Postulamos la puesta en marcha del Sistema de Vigilancia mediante las fases de gestión de un Programa, que pasamos a presentar.

4.1. Organización del Programa

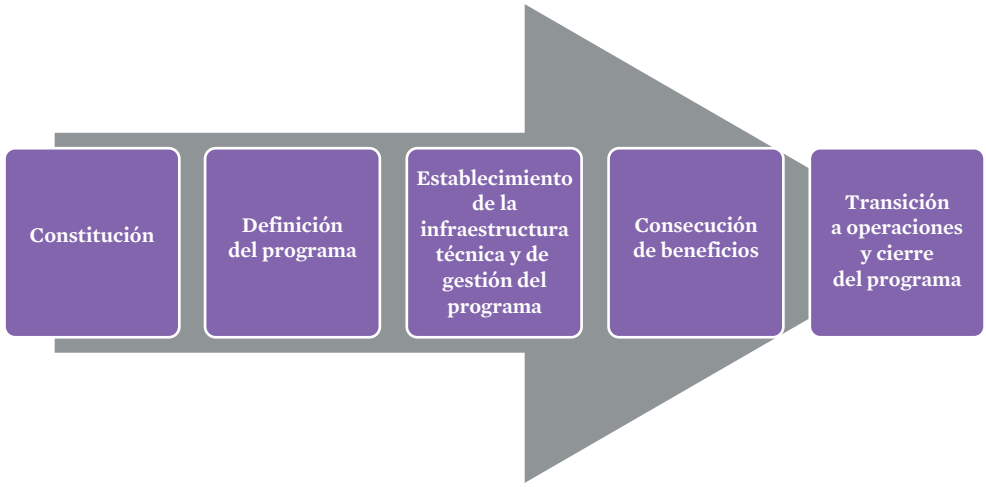
Se establecen tres niveles:

- **Comité de Patrocinio:** compuesto por directivos de la organización. Se encarga del Alineamiento estratégico de las inversiones de la organización. Es liderado por el Patrocinador del Programa.
- **Junta de gobierno del programa:** su objetivo es hacer posible el avance del programa para que proporcione resultados y beneficios. Dirigida por el Patrocinador del programa, contará con una Oficina de Programa, un Jefe de Programa y un BCM, responsable de cambio organizacional.
- **Gestión de proyectos**: asegura la coherencia y la relación entre el programa y los proyectos. El patrocinador del proyecto puede ser el mismo jefe de programa o contar con otro patrocinador que reportaría al jefe de programa.

4.2. Fases de puesta en marcha del Programa

Consta de 5 fases:

- Constitución.
- Definición del Programa.
- Establecimiento de la infraestructura técnica y de gestión del programa.
- Consecución de Beneficios.
- Transición a Operaciones y Cierre del Programa.



Pasamos a explicar a continuación cada una de las fases:

4.2.1. Fase de Constitución

En la fase de Constitución se establecen los objetivos del programa de acuerdo a los objetivos estratégicos de la organización. Se identifican los Beneficios del Programa, los Stakehdolders, el Jefe del Programa y se nombra la Junta del Programa. Asimismo se realiza el **Análisis de Necesidades** del que partimos en la metodología que presentamos.

Como herramientas de Gestión de Programa elaboraremos las siguientes:

- Un **Registro de Beneficios** del programa, perfilado con una descripción, los resultados observables esperados, el propietario y el método de medida.
- Un **Mapa de Beneficios**, en el que relacionaremos entregables, capacidades, resultados y beneficios
- Un **Plan de Beneficios**, a partir del Registro y el Mapa, para tener una visión completa de los beneficios y sus relaciones temporales.

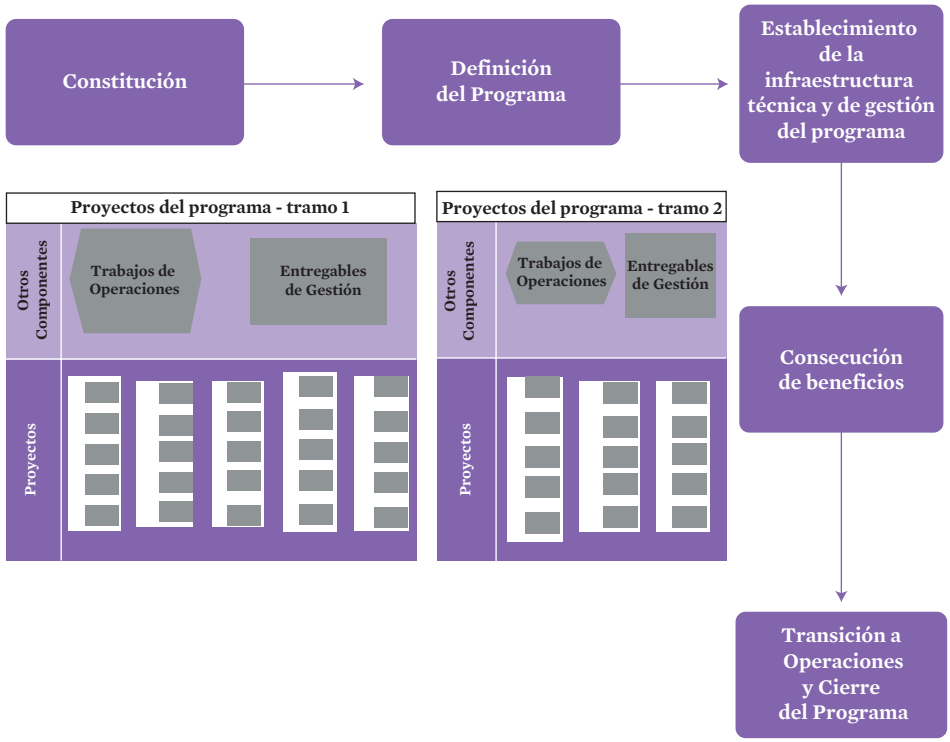
4.2.2. Fase de Definición del programa

Se establece el Plan de Gestión del Programa, incluyendo los Componentes y sus objetivos de tiempo y costo. Se establece la Línea base del programa, incluyendo el alcance, los Riesgos y su Plan de Gestión y se define un mecanismo de Gestión de Cambios en el programa.

También se establecerá un correspondencia entre el Plan de Beneficios del Programa y la Línea Base.

4.2.3. Fase de Establecimiento de la infraestructura técnica y de gestión del programa

Se define el sistema de gobierno del programa, los roles involucrados (sponsor, junta, jefe, oficina, equipo, el Sistema documental de gestión (políticas, procedimientos, plantillas) y se realiza la comunicación de la estructura a los Stakeholders.



4.2.4. Fase de Consecución de Beneficios

En esta fase se ejecutan los proyectos y el resto de componentes del programa. Esta ejecución se divide en **Tramos** (en inglés "*tranches*"). Cada tramo consta de un conjunto de componentes que entregan una o varias nuevas capacidades de forma sucesiva. Consecuentemente se van generando resultados y beneficios de forma incremental. Se deben equilibrar el potencial de generación de nuevas capacidades con los recursos disponibles y su ritmo de recepción.

Además de por la **Ejecución**, esta se caracteriza por sus actividades de control y seguimiento, en forma de **Revisiones**. Se realizan revisiones, conforme a la Línea Base establecida, del estado de los componentes y las interfaces entre los mismos. Asimismo se realiza la revisión de beneficios al acabar cada tramo. Se comprueba el estado de consecución de los beneficios, se supervisa el impacto y el nivel de relevancia para los objetivos del programa, se identifican desviaciones. Finalmente se realiza la comunicación a los interesados.

4.2.5. Fase de Transición a operaciones y cierre del programa

En esta fase se verifican los entregables integrados entregan la capacidad deseada y se comprueba que la capacidad produce los resultados y beneficios esperados. Al final de esta fase la organización receptora de los beneficios habrá cambiado al disponer de nuevas capacidades y a haberse adaptado a las mismas.

En esta fase es relevante la labor del **Business Change Manager (BCM),** que identifica riesgos, coordina y resuelve conflictos entre el programa y las áreas impactadas, receptoras de las capacidades y monitoriza los KPIs de entrega y consecución de beneficios.

Esta fase se estructura en un **Plan de Transición,** que consta de tres sub-fases: pre-transición, gestión de la transición y post-transición.

En la **Pre-transición** se establece la estrategia de implantación cambios, el personal afectado, nuevos métodos de trabajo, la disponibilidad de instalaciones para los BCM, los stakeholders más importantes y su nivel de participación necesario, la integración del plan de transición con línea base del programa y las operaciones y el plan de contingencia o salida.

Durante la **Gestión de la transición** se incorporan los entregables a las operaciones. Se verifica previamente que los entregables están acabados, el personal de operaciones entrenado y dispuesto y organizado, se está ejecutando el plan de gestión de riesgos y se cuenta con la autorización del patrocinador.

En la **Post-transición** se miden los beneficios de acuerdo a las métricas definidas en el perfil de beneficios. Finalmente se extraen las Lecciones aprendidas, se desmantela la infraestructura técnica y de gestión del programa y se formaliza el cierre del programa.

6

BIBLIOGRAFÍA Y FUENTES DE DOCUMENTACIÓN

European Commission (EC). Digital Agenda for Europe http://ec.europa.eu/digital-agenda.

Agenda Digital para España http://www.agendadigital.gob.es.

European Commission (EC). Digital Economy: Making Big Data work for Europe http://ec.europa.eu/digital-agenda/en/big-data.

European Commission (EC). Digital Agenda for Europe. Public-Private Partnership (PPP) for Big Data http://europa.eu/rapid/press-release_MEMO-14-583_en.htm.

European Commission (EC). Digital Agenda Web Site Press Releases: https://ec.europa.eu/digital-agenda/en/news/natural-language-processing-nlp-market-worldwide-market-forecast-analysis-2013%E2%80%932018.

Research and Markets. Natural Language Processing (NLP) Market - Worldwide Market Forecast & Analysis (2013–2018) http://www.researchandmarkets.com/research/3tl4zb/natural_language (October 2013).

Vigilancia Estratégica e Inteligencia Competitiva

Aenor (2011). Norma Española UNE 166006:2011 – Gestión de la I+D+i: Sistema de Vigilancia Tecnológica e Inteligencia Competitiva. http://www.aenor.es.

Aenor (Mayo 2015)– Normas y Publicaciones de I+D+i. Recuperado en: https://www.aenor.es/AENOR/certificacion/innovacion/innovacion_vigilancia_166006.asp.

Aenor, (mayo 2015) Documentos vigentes publicados por el Comité AEN/CTN 166 en su web http://www.aenor.es/aenor/normas/ctn/fichactn.asp?codigonorm=AEN/CTN%20166.

Madrimasd (Junio 2008). Implantación de un Sistema de Vigilancia Tecnológica. La Norma UNE 166.006:2006 – Vigilancia Tecnológica – Resumen publicado en http://www.madrimasd.org/informacionidi/agenda/documentos/Seminario_VT/Seminario_VT_Gerardo_Malvido.pdf.

CEN – Comité Europeo para la Normalización – Comité Técnico CEN/TC 389 Gestión de la Innovación http://standards.cen.eu/dyn/www/f?p=204:7:0::::FSP_ORG_ID:671850&cs=1E977FFA493E636619BDED775DB4E2A76.

ISO- Comité Técnico para la Gestión de la Innovación http://www.iso.org/iso/iso_technical_committee%3Fcommid%3D4587737.

Javier Muñoz, María Marín, José Vallejo. El profesional de la información v15, n6, (nov-dic 2006). La vigilancia tecnológica en la gestión de proyectos de I+D+i: recursos y herramientas. Referenciado desde “Repository teaching materials ISO

http://www.iso.org/iso/fr/home/standards/standards-in-education/education_innovation-list/educational_innovation-detail.htm?emid=3561 redirigido a http://eprints.rclis.org/9400/1/vol15_6.1.pdf.

ISO (Diciembre 2014) "Strategic business plan – Innovation Management" Comité ISO/TC 279 –http://isotc.iso.org/livelink/livelink/fetch/2000/2122/687806/ISO_TC_279__Innovation_management_.pdf?nodeid=16913333&vernum=-2.

Technological Forecasting and Social Change Journal - http://www.sciencedirect.com/science/journal/00401625.

Persistent Forecasting of Disruptive Technologies - Committee on Forecasting Future Disruptive Technologies; National Research Council http://www.nap.edu/catalog/12557/persistent-forecasting-of-disruptive-technologies.

Technological Forecasting – Innovation Portal – Wiley Custom Select http://www.innovation-portal.info/toolkits/technological-forecasting/.

Technological Forecasting – Wiley Online Study Tools http://www.wiley.com/college/dec/meredith298298/resources/addtopics/addtopic_s_02a.html.

Cynertia Consulting –(Septiembre 2010) Presentación online UNE 166.006 – http://issuu.com/bertagar78/docs/une-166006---sistema-de-vigilancia-tecnologica-v2 .

Nuevas Capacidades Big Data

"Big Data: Understanding how data powers big business", Bill Schnarzo (2013) Wiley.

IBM. C. Eaton; D. Deroos; T. Deutsch; G. Lapis; P. Zikopoulos. "Understanding Big Data", http://public.dhe.ibm.com/common/ssi/ecm/en/iml14296usen/IML14296USEN.PDF.

News Reader Project, http://www.newsreader-project.eu/.

News Reader Project Deliverables: http://www.newsreader-project.eu/publications/deliverables/.

News Reader Project. Knowledge Store version 2, Francesco Corcoglioniti, Marco Rospocher, Roldano Cattoni, Marco Amadori, Bernardo Magnini, Mohammed Qwaider, Michele Mostarda, Alessio Palmero Aprosio, Luciano Serafini.

Data Smart, John W. Foreman. Wiley 2014.

Instant Web Scraping with Java, Ryan Mitchell, Packt Publishing, Agosto 2013.

Webbots, Spiders and Screen Scrapers: A guide to Developing Internet Agents with PHP/CURL (2nd Edition), Michael Schrenk, No Starch Press, 2012.

"Probabilistic Topic Models", David M. Blei. Communications of the ACM, April 2012.

Latent Dirichlet Allocation (David M. Blei, Andrew Y.Ng, Michael I. Jordan - Journal of Machine Learning Research 3 (2003) 993-1022.

http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

MapReduce: Simplified Data Processing on Large Clusters, Jeffrey Dean and Sanjay Ghemawat - http://research.google.com/archive/mapreduce.html.

Jacob Loveless, Sasha Stoikov, Rolf Waeber - Communications of the ACM Vol. 56 No. 10, Pages 50-56 – "Online Algorithms in High-Frequency Trading http://cacm.acm.org/magazines/2013/10/168184-online-algorithms-in-high-frequency-trading/abstract.

W3schools.com "What is New in HTML5" http://www.w3schools.com/html/html5_intro.asp.

Clustering Image By Chire (Own work) [CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons http://commons.wikimedia.org/wiki/File:SLINK-density-data.svg.

Taylor Goetz, Apache Storm Committer – Hortonworks - http://www.slideshare.net/ptgoetz/storm-hadoop-summit2014.

MapReduce: Simplified Data Processing on Large Clusters, Jeffrey Dean and Sanjay Ghemawat http://research.google.com/archive/mapreduce.html.

Jesse Anderson: Learn MapReduce with Playing Cards" https://www.youtube.com/watch?v=bcjSe0xCHbE.

Proyecto Hadoop en Apache https://hadoop.apache.org/.

Distribuciones de Apache Hadoop en la página web de Apache http://wiki.apache.org/hadoop/Distributions_and_Commercial_Support.

La Pastilla Roja. "NoSQL para no programadores" Sergio Montoro Ten (2012) en http://lapastillaroja.net/2012/02/nosql-for-non-programmers/.

Versión Cero. "Almacenamiento Distribuido no relacional" Sergio Montoro Ten (2009) http://www.versioncero.com/articulo/596/almacenamiento-distribuido-no-relacional.

NoSQL Databases: http://nosql-database.org/.

Biblioteca Nacional de España. Portal de Datos Enlazados http://Datos.BNE.es (2015) Proyecto realizado por el Ontology Engineering Group (OEG) de la UPM www.oeg-upm.net/.

FOAF (2000 – 2015) Friend of a Friend Linked information system. http://www.foaf-project.org/ y : http://xmlns.com/foaf/spec/.

Good Relations Ontology for Products & Services: http://www.heppnetz.de/projects/goodrelations/.

Open Graph Protocol http://ogp.me/.

BBC Ontologies http://www.bbc.co.uk/ontologies.

BBC – Sport Ontology: http://www.bbc.co.uk/ontologies/sport.

YAGO: A High-Quality Knowledge Base. Max Planck Institut Informatik http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/.

DBPedia. Extracting structured information from Wikipedia. http://wiki.dbpedia.org/.

Virtuoso Universal Data Server http://virtuoso.openlinksw.com/.

Freebase http://www.freebase.com/ Help about Freebase: https://developers.google.com/freebase/index.

Plataforma Contratación del Estado Español: https://contrataciondelestado.es/wps/portal/plataforma.

TED Tenders: http://ted.europa.eu/TED/main/HomePage.do.

Google Patents http://www.google.es/advanced_patent_search.

USPTO United States Patent and Trademark Office.

EPO European Patent Office.

WIPO World Intellectual Property Organization.

Wolfram Language & System Documentation Center – Machine Learning http://reference.wolfram.com/language/guide/MachineLearning.html.

Freeling Demo con opción PoS Tagging: http://nlp.lsi.upc.edu/freeling/demo/demo.php.

Freeling Demo con opción PoS Tagging: http://nlp.lsi.upc.edu/freeling/demo/demo.php.

Google Fusion Tables Help https://support.google.com/fusiontables/answer/2571232.

Ruth Cobos, UA Madrid http://cdn.intechopen.com/pdfs-wm/29147.pdf.

Xin Luna Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, Wei Zhang. Google. "*Knowledge Vault: a Web-Scale Approach to Probabilistic Knowledge Fusion* https://www.cs.cmu.edu/~nlao/publication/2014.kdd.pdf.

Natasha Noy https://www.linkedin.com/in/natashafnoy.

Seth Grimes. Text/Content Analytics 2011. User Perspectives on Solutions and Providers.

Seth Grimes. Text Analytics 2014. User Perspectives on Solutions and Providers.

W3C – Linked Data (Datos Enlazados) http://www.w3.org/standards/semanticweb/data.

G. VanNoord – Survey of the State of the Art in Human Language Technology. Cambridge University Press, 1996.

John Bateman, Eduard Hovy Computers and Text Generation: Principles and Uses. Butler, C.S. Computers and Written Texts, Oxford - Blackwell.

W.C. Mann y Sandra A. Thompson: Rhetorical structure theory: Toward a functional theory of text organization.

M.A.K. Halliday - An Introduction to Functional Grammar, Ed. Arnold 1985.

Enrique Alcaraz Varo, Mª Antonia Martínez Linares. Diccionario de Lingüística Moderna, Editorial Ariel 1997.

Enrique Alcaraz Varó, Paradigmas de la investigación lingüística, Ed. Marfil 1990.

Antonio Miranda Raya - Sistema de Consulta a una Ontología. PFC - ETS Ingenieros en Informática – UPM (2000).

CPV (Common Procurement Vocabulary) REGLAMENTO (CE) No 213/2008 DE LA COMISIÓN de 28 de noviembre de 2007 que modifica el Reglamento (CE) no 2195/2002 del Parlamento Europeo y del Consejo.

UNSPSC (siglas de United Nations Standard Products and Services Code) http://www.unspsc.org/.

OOPS! (OntOlogy Pitfall Scanner!) http://www.oeg-upm.net/oops.

Norma Técnica de Interoperabilidad de Reutilización de Recursos de Información http://datos.gob.es/content/norma-tecnica-de-interoperabiliad-de-reutilizacion-de-recursos-de-informacion.

CPV converted to RDF. Conjunto de datos disponible en publicado por http://opendata.cz.

LODRefine, version de OpenRefine LOD-enabled https://github.com/sparkica/LODRefine.

W3Consortium "R2RML RDB to RDF Mapping Language" http://www.w3.org/TR/r2rml/.

OEG-UPM. Motor RDB2RDF Morph-RDB http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/technologies/315-morph-rdb.

The Compact Edition of the Oxford Dictionary. 1971/1987. Oxford: Oxford University Press.

Munn, K. and B. Smith (eds.). 2008. Applied Ontology: An Introduction. Frankfurt: Ontos Verlag.

Studer, R., V. R. Benjamins and D. Fensel. 1998. "Knowledge Engineering: Principles and Methods." Data & Knowledge Engineering 25(1-2): 161-197.

Suárez-Figueroa, M.C. 2010. NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse. Ph.D. Thesis presented at Universidad Politécnica de Madrid (Spain). http://oa.upm.es/3879/.

Suárez-Figueroa, M.C., A. Gómez-Pérez, E. Motta and A. Gangemi (eds.). 2012. Ontology Engineering in a Networked World. Berlin: Springer.

Haase P., Rudolph S., Wang Y., Brockmans S., Palma R., Euzenat J., d'Aquin M. NeOn Deliverable D1.1.1. Networked Ontology Model. NeOn Project. http://www.neon-project.org/. (November 2006).

Mahesh, Kavi. 1996. Ontology Development for Machine Translation: Ideology and Methodology. Technical Report MCCS 96-292, Computing Research Laboratory, New Mexico State University, Las Cruces, NM.

Watson (http://watson.kmi.open.ac.uk/WatsonWUI/) M. d'Aquin, M. Sabou, M. Dzbor, C. Baldassarre, L. Gridinoc, S. Angeletou, E. Motta. Watson: A Gateway for the Semantic Web. Poster session of the European Semantic Web Conference, ESWC 2007.

(Gruber, 1993), Gruber, T. R. (1993). A translation approach to portable ontology specifications. Knowledge acquisition, 5(2), 199-220.

Open Knowledge Base Connectivity protocol (OKBC) OKBC: A Programmatic Foundation for Knowledge Base Interoperability. Vinay K. Chaudhri, Adam Farquhar, Richard Fikes, Peter D. Karp, James P. Rice. Proceedings of AAAI-98, July 26-30, Madison, WI.

Swoogle http://swoogle.umbc.edu/.

Protégé Ontology Library http://protegewiki.stanford.edu/wiki/Protege_Ontology_Library.

BioPortal http://bioportal.bioontology.org/.

Linked Open Vocabularies (LOV) http://lov.okfn.org/.

Catálogo de Ontologías útiles: http://smartcity.linkeddata.es/.

W3Consortium – Web Ontology Language OWL http://www.w3.org/2004/OWL/.

W3Consortium – Sparql Query Language for RDF http://www.w3.org/TR/rdf-sparql-query/.

W3C Data Activity http://www.w3.org/2013/data/.

W3Consortium – SWRL: A Semantic Web Rule Language Combining OWL and RuleML http://www.w3.org/Submission/SWRL/.

W3Consortium – OWL Working Group http://www.w3.org/2007/OWL/wiki/OWL_Working_Group.

W3Consortium http://www.w3.org/.

WebProtege http://webprotege.stanford.edu/.

TopBraid Composer - funcionalidades http://www.topquadrant.com/tools/modeling-topbraid-composer-standard-edition/.

SPIN (SPARQL Inference Notation). http://spinrdf.org/.

NeOn toolkit http://neon-toolkit.org/.

s/neonglossaryofactivities.pdf Lights and shadows in creating a glossary about ontology engineering. Mari Carmen Suárez-Figueroa; Guadalupe Aguado de Cea and Asunción Gómez-Pérez. Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication, Volume 19, Issue 2, 2013, pages: 202-236.

URIs http://www.ietf.org/rfc/rfc1630.txt.

Berners Lee (2006) http://www.w3.org/DesignIssues/LinkedData.html.

Página del Plan de Impulso de las Tecnologías del Lenguaje en la Agenda Digital española http://www.agendadigital.gob.es/planes-actuaciones/Paginas/plan-impulso-tecnologias-lenguaje.aspx.

Plan de Impulso de las Tecnologías del Lenguaje Humano en la Agenda Digital española http://www.agendadigital.gob.es/planes-actuaciones/Bibliotecaimpulsotecnologiaslenguaje/1.%20Plan/Plan-impulso-Tecnologias-Lenguaje.pdf.

Informe sobre el estado de las tecnologías del lenguaje en España dentro de la Agenda Digital para España http://www.agendadigital.gob.es/planes-actuaciones/Bibliotecaimpulsotecnologiaslenguaje/2.%20Material%20complementario/Informe-Tecnologias-Lenguaje-Espana.pdf.

Presentación, Formalización del Modelo y la Metodología

PMI - Project Management Institute (2014) Project Management Body of Knowledge.

PMI - Project Management Institute (2013) The Standard for Program Management.

Henry Mintzberg, Bruce Ahlstrand, Joseph Lampel (1999). "Safari a la Estrategia", Ediciones Granica.

Alexander Osterwalder, Yves Pigneur. (2010). "Business Model Generation". Ed. Wiley.

Michel Porter (1980), "Estrategia Competitiva".

Michael Porter, (1985), "Ventaja Competitiva".

ISO/IEC 20000:2011 IT Service Management and ISO/IEC 38500:2015 IT Governance http://www.iso.org/iso/home/store/catalogue_tc/catalogue_tc_browse.htm?commid=5013818.

ITIL.org http://www.itil.org/ También en Wikipedia http://es.wikipedia.org/wiki/Information_Technology_Infrastructure_Library.

SEI. – Software Engineering Institute. "CMMI-DEV CMMI para Desarrollo, Versión 1.3" http://www.sei.cmu.edu/library/assets/whitepapers/Spanish%20Technical%20Report%20CMMI%20V%201%203.pdf.

SEI – Software Engineering Institute. "CMMI for Acquisition, Version 1.3" http://www.sei.cmu.edu/reports/10tr032.pdf.

The Standard for Program Management, PMI 2013 – Project Management Institute.

Referencias en Wikipedia

Coeficiente de determinación R^2 http://es.wikipedia.org/wiki/Coeficiente_de_determinación.

Prueba F de Fisher http://es.wikipedia.org/wiki/Prueba_F_de_Fisher.

T de Student http://es.wikipedia.org/wiki/Distribución_t_de_Student.

Bagging (Agregación de Bootstrap) http://es.wikipedia.org/wiki/Agregación_de_bootstrap y http://en.wikipedia.org/wiki/Bootstrap_aggregating#/media/File:Ozone.png.

"MeningiomaMRISegmentation" by Rkikinis at English Wikipedia. Licensed under CC BY-SA 3.0 via Wikimedia Commons – http://commons.wikimedia.org/wiki/File:MeningiomaMRISegmentation.png#/media/File:MeningiomaMRISegmentation.png.

Curva ROC http://es.wikipedia.org/wiki/Curva_ROC.

Imagen Curvas.png publicada en http://commons.wikimedia.org/wiki/File:Curvas.png.

Vigilancia Tecnológica http://es.wikipedia.org/wiki/Vigilancia_tecnológica.

Inteligencia Empresarial http://es.wikipedia.org/wiki/Inteligencia_empresarial.

Technology intelligence http://en.wikipedia.org/wiki/Technology_intelligence.

Technology scouting http://en.wikipedia.org/wiki/Technology_scouting.

Strategic Intelligence http://en.wikipedia.org/wiki/Strategic_intelligence.

Strategic foresight http://en.wikipedia.org/wiki/Strategic_foresight.

Corporate foresight http://en.wikipedia.org/wiki/Corporate_foresight.

Future Studies http://en.wikipedia.org/wiki/Futures_studies.

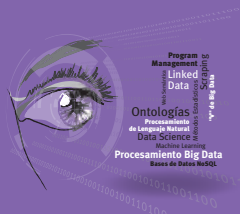
“UNE 166006” http://es.wikipedia.org/wiki/UNE_166006.

“Distribuciones Linux” http://es.wikipedia.org/wiki/Anexo:Distribuciones_Linux.

big intelligence

nuevas capacidades **big data**

para los Sistemas de **vigilancia estratégica** e **Inteligencia Competitiva**



Presentamos bajo el nombre de “Big Intelligence” a la confluencia de “Big Data”, “Vigilancia Estratégica” e “Inteligencia Competitiva”. Estos tres términos junto con el concepto de “Capacidades”, extraído de la disciplina del “Program Management”, son el fundamento de este libro.

La puesta en marcha de Programas Big Data de Vigilancia Estratégica e Inteligencia Competitiva proporcionará a las Empresas e Instituciones nuevas Capacidades que sólo muy recientemente las Tecnologías de la Información han hecho viables.

innovación y creatividad

GOBIERNO DE ESPAÑA · MINISTERIO DE INDUSTRIA, ENERGÍA Y TURISMO · EOI Escuela de organización industrial

www.eoi.es

EOI MADRID
Avda. Gregorio del Amo, 6
Ciudad Universitaria
28040 Madrid
informacion@eoi.es

EOI ANDALUCÍA
Leonardo da Vinci, 12
Isla de la Cartuja
41092 Sevilla
infoandalucia@eoi.es

EOI MEDITERRÁNEO
Presidente Lázaro Cárdenas del Río,
esquina C/Cauce
Polígono Carrús
03206 Elche (Alicante)
eoimediterraneo@eoi.es