

# ONTOLOGY

El PLN (Procesamiento del  
Lenguaje Natural)

aplicado a la

**Clasificación Automática  
de  
Documentos**

en vanguardia de la  
Innovación Tecnológica

# Ontology

## Clasificación automática de documentos

### Introducción

Ontology es una aplicación de servidor desarrollada por THERA, Centre de Llenguatges i Computació, S.L. para la clasificación automática de documentos.

Se trata de una aplicación de servidor de fácil administración, transparente para el usuario y de solvencia técnica que basa su funcionamiento en automatismos que permiten obtener documentos de diferentes fuentes de datos y procesarlos de acuerdo con planes de proceso llamados "tratamientos".

El procesamiento de los documentos incluye un completo análisis del lenguaje natural que permite la lematización lingüística de los textos, así como un proceso adaptativo de entrenamiento que aumenta la calidad de la clasificación.

Este documento contiene una explicación técnica de Ontology.

### Funcionamiento

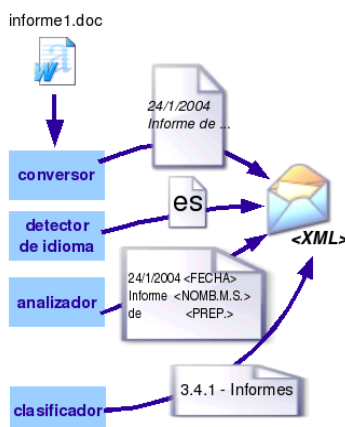
La finalidad de Ontology es obtener documentos de diferentes fuentes, procesarlos de acuerdo con un plan de proceso (tratamiento) y clasificarlos en uno o varios temas de una ontología (estructura jerárquica de temas).

De la obtención de documentos para el proceso se encargan las **tareas**, mientras que del proceso de estos documentos se encargan los **tratamientos**.

#### Tratamientos

Un tratamiento es la planificación tanto de la fuente de la cual se obtienen los documentos, como de los procesos que han de aplicarse a ése conjunto de documentos. Estos planes de proceso tienen como objetivo transformar los documentos originales en un resultado de clasificación, es decir, en una asignación temática para el documento.

Los tratamientos se construyen mediante la conexión de varios módulos de proceso, cada uno de ellos encargado de una parte puntual del proceso, por ejemplo:



- El módulo **conversor** es el encargado de extraer el texto del documento original y convertirlo en un documento XML.
- El módulo **detector de idioma** es capaz de asignar un código de idioma al documento. Para ello necesita tener acceso al texto del documento, y por tanto necesita que el conversor haya actuado antes.
- El módulo de **análisis** realiza un análisis morfológico del texto del documento: segmenta el texto en palabras, las analiza morfológicamente y determina el lema de cada una de ellas (por ejemplo, el lema de "vendimos" y "venderemos" es "vender"). Para poderlo hacer necesita tener acceso al texto del documento (conversor) y al idioma del documento (detector de idioma).

Hay muchos más módulos. Y todos ellos funcionan generando una salida a partir de la transformación de sus entradas, incorporando información al documento.

En ejecución, cada módulo se convierte en un proceso diferenciado, gestionado por un proceso permanente (planificador) que se encarga de arrancar o parar los módulos dependiendo de las necesidades del sistema.

El intercambio de información entre los procesos que componen un tratamiento se hace mediante un documento XML (uno por cada documento original), al que cada módulo añade el resultado de su proceso. Cuando este documento XML llega al último módulo de su proceso, se guarda en el sistema para su posterior recuperación.



Los módulos entrenables son un caso particular de módulo que necesitan ser entrenados para su correcto funcionamiento; el **clasificador** es uno de ellos.

## Tareas

Los procesos encargados de la obtención de los documentos para el proceso reciben el nombre de tareas.

Las tareas son las encargadas de conectarse a una fuente de documentación mediante el protocolo correspondiente y de obtener los documentos susceptibles de ser procesados por el tratamiento que tiene asociado.

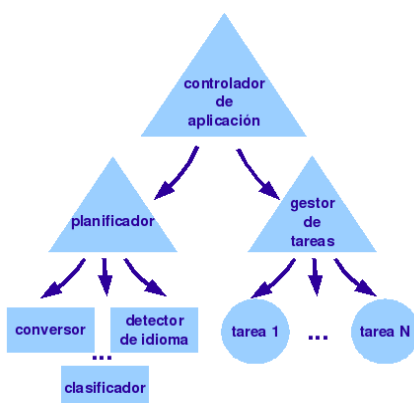
Ontology dispone actualmente de tareas que son capaces de obtener documentos de directorios locales del servidor, de recursos compartidos de Windows y de páginas Web. Y próximamente estarán disponibles otros protocolos.

Naturalmente, una tarea no leerá los documentos que previamente ya haya procesado, evitando así un tráfico inútil en la red.

Ontology gestiona las tareas mediante un proceso permanente (gestor de tareas) que se encarga de lanzar las tareas cuando están programadas. Una tarea puede programarse para ejecutarse periódicamente (con cadencia mínima de 1 minuto) o no periódicamente, en cuyo caso será el administrador del tratamiento correspondiente el encargado de lanzarla cuando sea preciso.

## Arquitectura del sistema

Ontology es una aplicación de servidor para sistemas Linux que hace un fuerte uso de las características de paralelismo del sistema operativo. El sistema se controla mediante tres procesos permanentes:



- El **gestor de tareas** es el proceso encargado de arrancar las tareas (encargadas de obtener nuevos documentos) cuando están programadas y de reprogramarlas en caso de que sean periódicas.
- El **planificador** es el proceso encargado de arrancar y parar los módulos de proceso que componen los tratamientos en función de la carga de trabajo que tengan cada uno de ellos.
- El **controlador de aplicación** es un proceso genérico encargado de asegurarse que los otros dos estén siempre en funcionamiento y de rearrancarlos en caso de algún fallo.

## Seguridad

Tanto las tareas como los módulos son procesos UNIX independientes, lo cual aumenta la seguridad de la aplicación impidiendo interferencias.

De hecho, el sistema ha sido programado con la seguridad como primer argumento. Los procesos controladores, que son los encargados de crear nuevos procesos, realizan un **control activo** de los procesos hijos. Las tareas y los módulos están obligados a informar de su situación en intervalos regulares a su controlador; esto permite al controlador detectar cuándo uno de sus procesos controlados ha caído inesperadamente.

Los módulos aplican, a su vez, un sistema que impide que un documento defectuoso paralice el sistema. Esto se consigue intentando el reproceso de un documento sólo un número determinado de veces.

## Base de datos relacional

Ontology hace uso extensivo de bases de datos relacionales. Se ha puesto un especial cuidado en la interfaz de base de datos. Como consecuencia de ello, Ontology es:

- 100% compatible con SQL-92.
- Independiente del sistema de gestión de base de datos (SGBD).

La independencia de base de datos se consigue mediante un nivel de abstracción intermedio entre la aplicación y las librerías de acceso particulares de cada SGBD.

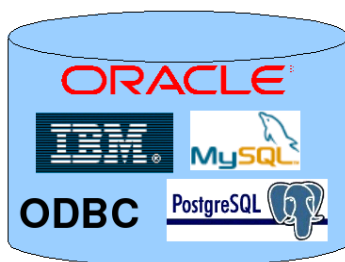
Disponemos de drivers para ORACLE<sup>(c)</sup>, DB2<sup>(c)</sup>, PostgreSQL, MySQL, Interbase<sup>(c)</sup>, Informix<sup>(c)</sup> y muchos otros.

## Interfaz de programación

Ontology dispone de una API en Perl, lo cual permite la implementación de nuevos algoritmos, en tiempo récord, para añadir nuevas características.

## Interfaz de administración

La interfaz de administración de Ontology está basada en tecnología Web, facilitando en gran medida la configuración y administración de una herramienta tan compleja.



## Procesamiento del lenguaje natural

Los módulos de Ontology hacen uso extensivo de las herramientas y recursos para el procesamiento del lenguaje natural de que dispone THERA, resultado de muchos años de investigación.

### ¿Por qué PLN?

La inmensa mayoría de las herramientas de procesamiento y clasificación de documentos parten de algoritmos pensados para el inglés. Algoritmos como el *stemming* pierden efectividad cuando se enfrentan a idiomas ricos en matices morfológicos como los peninsulares.

El procesamiento del lenguaje natural (PLN) es un paso adelante en el análisis de textos en español, catalán, gallego y euskera, porque permite analizar el texto a un nivel muy afinado.

El PLN permite identificar la palabra "tardamos" como la primera persona plural del presente del indicativo y del pretérito indefinido del indicativo del verbo "tardar".

### **Detección de idioma**

El algoritmo de detección de idioma de THERA tiene una precisión del 98% con textos de entre 5 y 10 palabras, superando este porcentaje hasta el 99% con textos de más de 10 palabras.

### **Análisis morfológico y lematización**

El análisis morfológico identifica la función morfológica de las palabras en las oraciones (desambiguando las posibles interpretaciones múltiples de una palabra) y les asigna un **lema**. Por ejemplo, la frase

*"El informe fue redactado por el Sr. González."*

tiene el siguiente análisis morfológico:

palabra	lema	morfología
El	el	Artículo masculino singular
informe	informe	Nombre masc. sing.
fue	ser	Verbo semiauxiliar 3ª pers. sing. pretérito indefinido indicativo
redactado	redactar	Verbo principal participio masc. sing.
por	por	Preposición
el	el	Artículo masculino singular
Sr. González	Sr. González	Nombre propio
.	.	Fin de oración

El procesamiento del lenguaje natural de Ontology incluye:

- Lematización de las formas comunes (1.200.000 formas del español, sin contar diminutivos, adverbios en -mente, etc. que totalizarían un reconocimiento de más de 5.000.000 de formas flexivas y derivadas).
- Identificación de números, cantidades monetarias, fechas, siglas, etc.
- Identificación de nombres propios.
- Identificación de términos multipalabra (locuciones) como "producto interior bruto", "carnet de identidad", etc.

La lematización, además, reduce el tamaño de los índices necesarios para guardar la información de los documentos y permite una recuperación mejor.

## **Clasificación automática**

Ontology dispone de un módulo de **clasificación automática** de documentos.

### **¿Qué es la clasificación automática?**

La clasificación automática consiste en, dado un conjunto finito de temas, asignar uno o varios de esos temas a un documento.

Es una técnica de inteligencia artificial.

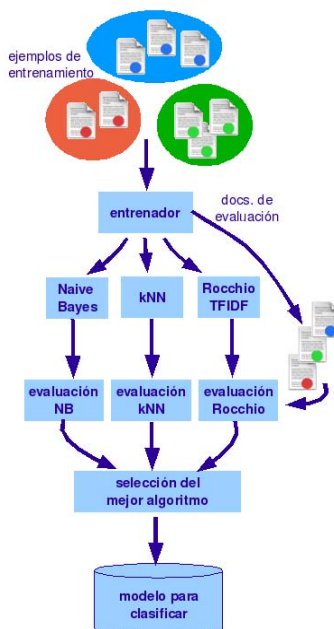
## Ontologies

Para poder clasificar documentos se necesita en primer lugar un conjunto de temas (clases, categorías). En Ontology este conjunto de temas se expresa mediante jerarquías en las cuales un tema contiene a otros.

Ontology permite definir todas las ontologías necesarias y crear tratamientos que clasifiquen con ellas.

## Entrenamiento

Una vez definida una ontología de temas es necesario **entrenar** al clasificador, es decir, proveerle de documentos de ejemplo para cada tema. En esta fase el clasificador **aprende** las características de los documentos de cada tema y extrapola un modelo matemático que le permitirá clasificar los nuevos documentos.



Durante el proceso de entrenamiento se separa un pequeño grupo de documentos que sirve para calcular la bondad del entrenamiento en términos de precisión, exhaustividad, etc.

Ontology, a diferencia del resto de herramientas del mercado, no utiliza un único algoritmo de clasificación sino que dispone de una **batería de algoritmos** que aplica en fase de entrenamiento, seleccionando al final del proceso el que mejor funciona para los datos aportados. Los algoritmos actualmente soportados son:

- KNN (K-Nearest Neighbours)
- Naive Bayes multinomial.
- Rocchio TFIDF.

Cuando el proceso de entrenamiento ha concluido, el administrador del tratamiento valida los datos permitiendo que el tratamiento entre en modo de producción.

## Clasificación

En modo de producción, el módulo de clasificación utiliza el modelo matemático extrapolado en la fase de entrenamiento para decidir los temas que aplica a cada nuevo documento que recibe.

Es decisión del módulo asignar un tema o más de uno a cada documento. La clasificación se enriquece con un valor de confianza para cada tema asignado.

## Mejoras que aporta el PLN a la clasificación

Los algoritmos de clasificación automática se basan principalmente en analizar el vocabulario que se utiliza en los documentos de cada tema, extrayendo un modelo para cada tema.

La lematización produce una mejora en la clasificación muy importante, porque reduce el vocabulario de los documentos y permite que el sistema trate palabras como "redactó", "redactará", "redactado", "redactando", etc. como una única: "redactar".

La detección de nombres propios y de términos multipalabra mejora la calidad de la clasificación. Intuitivamente es evidente: por ejemplo, las palabras "producto", "interior" y "bruto" cobran un especial significado cuando forman un multitérmino: "producto interior bruto".

## **Ontology** ©

Clasificación automática

### **Thera,**

Centre de Llenguatges i Computació, S.L

c/ Adolf Florensa s/n Ed. Florensa  
08028-BARCELONA

Tel. : 93 403 45 58  
Fax 93 403 46 78

**[WWW.thera-clic.com](http://WWW.thera-clic.com)**

**[e-mail:nfo@clic.fil.ub.es](mailto:nfo@clic.fil.ub.es)**

