# Estado actual del proyecto GDA (Gestión Documental Automatizada): planteamiento teórico y descripción práctica

**José A. Moreiro González**
**Juan Lloréns Morillo**
**Manuel Velasco de Diego**
Universidad Carlos III de Madrid

## 0.1 Resumen

El Grupo de Tecnología de la Información, viendo las necesidades existentes en el campo de gestión documental, ha desarrollado un sistema que permite gestionar, en algunos casos de forma automática, todo el conjunto de información perteneciente a una organización. Para ello se ha definido una estructura de repositorio basada en los Tesauros de descriptores de las Ciencias de la Documentación. Contra esta estructura se han implementado con éxito una serie de procesos tales como un sistema de gestión automática, un sistema de consulta inteligente, un sistema de indización avanzada. Se han trabajado en las técnicas modernas de Análisis de Dominios con vistas a la generación automática del repositorio. (Autor)

**Palabras clave:** Clasificación automática. Entornos multimedia. Gestión de información, Indización automática. Recuperación de información. Reutilización. Tesauros de descriptores.

## 0.2 Abstract

Due to existing needs in the Information Management area the Information Technology Group has developed a system to manage, in some cases in an automatic way, the whole set of information related to an organization. A repository structure (Software Thesauri) has been defined. This structure is based on Thesauri which belongs to Information Science. A set of processes has been successfully performed to manage this repository. Some examples are: an automatic management system; an intelligent retrieval system; and an advanced automatic indexing system. Domain analysis modern techniques have been applied in order to construct the repository. (Author)

**Keywords:** Automatic Classifying. Automatic Indexing. Information Management. Information Retrieval. Multimedia. Reuse. Thesauri.

## 1. Origen, composición y objetivos del GTI

El Grupo de Tecnologías de la Información (GTI) surgió a comienzos de 1993, buscando alcanzar un equipo estable de trabajo multiinstitucional y, dentro de la Universidad Carlos III de Madrid, multidepartamental, incluyendo miembros tanto del profesorado como del alumnado. Para conseguirlo se precisaba obtener financiación que soportase las investigaciones tanto desde organismos nacionales como internacionales.

Fueron miembros desde los inicios: por parte del Departamento de Ingeniería de la Universidad Carlos III de Madrid, los profesores Antonio Amescua, Juan Lloréns y Manuel Velasco; por el Departamento de Biblioteconomía y Documentación de la misma Universidad, José Antonio Moreiro, y por el CINDOC los investigadores Jorge Páez, Antonio Valle y Alfredo Del Rey. Esta colaboración ha producido ya dentro del grupo una tesis doctoral. Se están realizando otras dos, así como se ha completado una treintena de trabajos de fin de carrera en la titulación de Informática Técnica de Gestión

El grupo enfocó su trabajo hacia la gestión documental automatizada mediante la utilización de tecnología informática y metodologías avanzadas de ingeniería de software. Las áreas de trabajo tratadas desde entonces han sido las siguientes :

- Diseño de software avanzado en entorno multimedia.
- Sistemas de indización automática de información: aplicación a la construcción automática de tesauros.
- Integración de herramientas ofimáticas en sistemas de gestión de información.
- Sistemas avanzados de clasificación de información: redes neuronales y metodologías estadísticas avanzadas.

### 1.1. Objetivos y naturaleza

El principal objetivo del proyecto consiste en diseñar y construir una plataforma de gestión de repositorios (Tesauro de Software) capaz de almacenar, procesar, gestionar y recuperar cualquier tipo de documento, sin importar su presentación, soporte y forma de acceso, creando un sistema autogenerable (que el tesauro crezca por sí mismo desde su nacimiento o desde algún punto de desarrollo).

Procesar y gestionar implica todo el trabajo interno del sistema, el cual permite que la información sea analizada vocablo por vocablo, determinando su naturaleza, el número de incidencias, las relaciones entre sí, etc.; para de ésta manera alcanzar el análisis textual y documental. Se pretende ayudar de manera eficiente y relevante en la recuperación de la información de acuerdo a las necesidades de los usuarios. Por otra parte, si se logra que el sistema sea autogenerable se podrá analizar toda la cantidad de información que se genera día a día en las "autopistas de la información", facilitando la labor de documentalistas e informáticos. El proyecto busca desarrollar una estructura específica orientada a la recuperación documental automatizada. Dicha estructura de información se define como Tesauro Autogenerable y está basada en la teoría documental de los Tesauros de Descriptores en Documentación (según la norma ISO 2788).

Como características generales del proyecto podemos señalar las siguientes:

- Su funcionamiento mediante un tesauro de descriptores global (utilizando la superestructura de los tesauros).
- Autogenerabilidad: que permite que el propio sistema se actualice conforme vaya almacenando, mediante la construcción automática de relaciones-de aplicación multilingüe.
- Con realización de la indización mediante analizadores sintácticos, semánticos y morfológicos.
- Basado en la reutilización de software.
- Con aplicación a la indización de todo tipo de documentos (textos, imágenes, etc.).
- Permitirá la gestión y recuperación de la información.

Además de los parámetros anteriores, el proyecto busca diseñar y ejecutar herramientas que permitan una gestión documental mucho más ágil:

- Trabajar en un entorno multiusuario: el proyecto trabaja en un ambiente que permite su acceso a varios usuarios de manera simultánea (sistema de multiproceso).
- Permitir la gestión del repositorio que almacena la información desde fuera por especialistas.
- Presentar la información de la manera más clara y sencilla.
- El programa es modular, lo cual permite su desarrollo y perfeccionamiento por unidades, como también facilita su ampliación, cobertura y especialización.
- Permitir gran amplitud en las relaciones semánticas de los términos: es ilimitada la cantidad de específicos que se puede asignar a cada término,

igual sucede con los términos relacionados, tanto los circunstanciales como los permanentes.

- Contener un sistema de gestión bibliográfica (registro bibliográfico con información secundaria de los documentos).
- Se conseguirá un sistema de indización y recuperación automática (búsquedas inteligentes).
- Supone la integración de diversos tipos de aplicaciones, con la ayuda de protocolos OLE-DDE (protocolos de interconexión de aplicaciones que permiten el traspaso de información entre aplicaciones).
- Realiza clasificación temática de los descriptores.
- Sistema de importación y exportación de la Ficha Bibliográfica del documento propiamente dicho.

### 1.2. Fundamentos teóricos

Los fundamentos teóricos del proyecto, sin contar con el aporte informático, se pueden descomponer así: Se parte del concepto de información y de los diversos soportes que pueden contener dicha información, para adoptar la teoría general de Tesauros de descriptores como fundamento para la gestión y recuperación de dicha información.

Partiendo, pues, del concepto general de información (todo mensaje que se trasmite, sin importar su canal o soporte) se recalca la importancia del contenido: todo aquello que ofrece datos y elementos para la construcción misma del saber, para la toma de decisiones o para la comunicación en sí. Por lo tanto, la información toma diversas formas o vías de acceso: encontramos información textual, gráfica, en imágenes, numérica, audiovisual, información en códigos (lenguajes de programación), de señales, etc (Moreiro, 1993. Capítulo 1). Para que dicha información cumpla su función social y democrática de uso, se requiere que se administre de manera correcta para su posterior recuperación. Por lo tanto, para su gestión normalizada, la unidad mínima de trabajo es el documento, en cuanto conjunto de información con cohesión y sentido propios. El documento se puede presentar como monografías, informes técnicos, tesis doctorales, artículos, diagramas de flujos de datos, hojas de cálculo, vídeos, fotografías, diseños gráficos, programas de ordenador, lenguajes de programación, etc.

Toda esta variedad de documentos tiene sentido en la medida en que se puedan recuperar de la manera más rápida, efectiva y eficaz, para lo cual este proyecto se basa en la teoría documental de la recuperación de la información mediante los lenguajes controlados. ¿Y por qué lenguajes controlados? Pues, porque toda la información fluye en un lenguaje natural, el lenguaje del usuario,

y para su gestión se hace necesario crear un lenguaje que permita la normalización de dicho lenguaje natural.

Para afrontar esta labor seguimos los presupuestos teóricos planteados por Van Slype (1991) por los que se concibe al tesauro de descriptores como una lista controlada de términos de un área del conocimiento estructurada semánticamente. No brinda significado gramatical de la palabra, pero sí establece relaciones semánticas entre los términos. El tesauro de descriptores es pues una herramienta para la "representación de conceptos" expuestos en los documentos y cuya finalidad es la recuperación de la información contenida en ellos (Slype, 1991).

Un tesauro está constituido básicamente por dos elementos: unidades léxicas y relaciones semánticas. Las unidades léxicas se dividen en cuatro categorías:

- *Campo semántico* o grupo de familia de términos.
- *Descriptor o término preferente*, que designa un concepto y que sirve para representar el contenido de un documento y realizar consultas.
- *No descriptor o término no preferente*, originados en los sinónimos o cuasi-sinónimos. No sirven para indizar, pero reenvían la indización o la consulta hacia los descriptores. Su misión es servir de inferentes hacia los descriptores.
- *Descriptores auxiliares*: por sí solos no aportan ningún concepto, pero sumados a los descriptores forman conceptos o descriptores compuestos.

Los tipos de relaciones semánticas entre los términos pueden ser:

- *Equivalencias interlinguísticas*, conceptos iguales o equivalentes en diferentes lenguas o equivalencias semánticas intralinguísticas, dentro de una misma lengua (de un descriptor a un no-descriptor). Dentro de esta relación encontramos sinonimía verdadera, sinonimia por variante ortográfica, siglas, variantes de escritura, extranjerismos, antonimias, lenguaje usual versus lenguaje científico.
- *Relaciones de jerarquía*: basadas en niveles de super o subordinación, en que un término superordenado representa un todo o clase y los términos subordinados corresponden a los miembros o partes del término superior.
- *Relaciones de asociación*: se muestra así la necesidad de establecer asociaciones que sugieran desde un concepto otros con los que esté relacionado, sin que entre ellos exista dependencia jerárquica. Como tipos de relaciones asociativas podemos nombrar las que se generan en la causalidad, la instrumentación, la sucesión en el espacio y en el tiempo, la concomitancia, la similaridad, la antonimia, las propiedades de los objetos y hechos, la loca-

lización, y los objetos de acciones, procesos o disciplinas (Aitchinson, 1987).

## 2. El proyecto Gestión Documental Automatizada (GDA) (Automated Information Management —AIM—)

De las diversas actividades realizadas por el Grupo de Tecnología de la Información destaca el proyecto de Gestión Documental Automatizada (GDA), que intenta crear una herramienta capaz de indizar automáticamente un conjunto de documentos para facilitar la posterior recuperación de su información. Para desarrollar los objetivos propuestos se trabaja en el diseño y desarrollo de un conjunto de herramientas capaces de indizar, clasificar y gestionar automáticamente documentos de cualquier procedencia y tipo.

### 2.1. Memoria técnica

El Proyecto GDA ha cumplido durante el periodo 93/97 las siguientes fases aplicadas a su diseño y desarrollo:

*1ª fase: 1993-94*

- Diseño de un tesauro global:
  - Tesauro ISO 2788.
  - Árbol de áreas temáticas (AAT), para gestionar la clasificación decimal enumerativa (Dewey, 1979).
- Desarrollo informático del tesauro global (Módulo de gestión):
  - ABMC tablas.
  - Módulo de gestión de tesauros con tratamiento de términos inclasificados.
  - Desarrollo de un módulo de indización básico.
  - Desarrollo de un módulo de consulta.
  - Desarrollo del árbol de áreas temáticas (AAT).
- Puesta en funcionamiento de un sistema de administración del GDA:
  - Seguridad.
  - Módulos de impresión.
  - Módulo de configuración.
  - Trabajo en *batch.*
- Integración GDA - Procesamiento de textos (Winword).

*2ª fase: 1994-96*

- Puesta en marcha de la indización morfológica:
  - Tratamiento de palabras compuestas complejas.
  - Tratamiento de palabras compuestas no yuxtapuestas.
  - Tratamiento de homonimia.
- Indización estadística mediante el método IDF:
  - Aplicación de los métodos estadísticos sobre frecuencias de aparición en documentos.
- Desarrollo de un servidor distribuido de indización.
- Utilización de las estructuras especiales del tesauro global para la realización de una recuperación optimizada:
  - Recuperación mediante el AAT (localización y clasificación de la búsqueda por proximidad al AAT => Recuento de incidencias de términos asociados a cada nodo del AAT ).
  - Creación de la metodología para la interconexión de Tesauros.
- Búsqueda de documentos por texto completo:
  - Localización de documentos utilizando como claves de acceso información textual.
- Desarrollo del Interfaz de Usuario para el sistema de gestión del Tesauro Global:
  - Modificación de estructuras jerárquica.
  - Integrar la consulta al tesauro dentro del módulo de gestión.
  - Permitir la integración de términos genéricos, específicos etc. sobre un inclasificado directamente, sin tener que recogerlo de una tabla.
  - Módulo inteligente de ayuda.
  - Módulo de configuración inteligente del GDA.
  - Módulo de gestión de temas y usuarios asociados a un documento:
    - Asignar tema a cada documento (tesis, etc.).
- Importación/ exportación de tesauros entre formatos ISO-GDA (SQL).
- Integración ofimática del GDA:
  - Desarrollo del Kernel OLE-DDE para el sistema GDA:
    - Estandarización del protocolo DDE para procesadores de textos.
    - Integración de la Información Excel y formato de datos de entrada.
    - Indización de tablas Excel a tablas SQL.
    - Gráficos, etc.

*3ª fase: 1996...*

- Implantación de los archivos de indización sintáctica : Parser:
  - control de género.
  - control de número.
  - control de tiempo verbal.
  - control de tipo de palabra (en evaluación) .
- Implantación de la indización multilíngüe:
  - Posibilidad de que el sistema pueda ser indizado en cualquiera de los idiomas para los que ha sido programado.
- Indización mediante redes neuronales : TermNet (en desarrollo):
  - Indización mediante la aplicación de las tecnologías neuronales al análisis gramatical con vistas a la recuperación exacta de los descriptores o sustantivos de una frase.
- Filtrados estadísticos de información mediante tratamiento de cadenas de caracteres
- Clasificación automática de la información (tanto textual como software) (Velasco, 1997):
  - Clasificadores bibliométricos.
  - Clasificadores estadísticos.
  - Clasificadores neuronales.
- Acceso al sistema desde Internet.
- Integración ofimática del GDA:
  - Desarrollo del Interfaz OLE2 para el sistema GDA.
  - Estandarización del protocolo OLE2 para aplicaciones gráficas.
- Diseño de arquitecturas de componentes binarios para la indización y clasificación distribuida.

## 3. Fundamentos Técnicos

El proyecto funciona sobre un entorno Microsoft Windows, con arquitectura cliente-servidor, interactuando con el gestor de base de datos SQL Base Server de Gupta Technologies. La aplicación principal ha sido desarrollada empleando el lenguaje "SQL Windows".

Los soportes técnicos e informáticos del proyecto son:

- Lenguaje SQL (Structured Query Language).

- Servidor SQLBASE: gestor de bases de datos relacional para almacenar la información.
- SQL Windows: lenguaje de cuarta generación sobre windows para desarrollar la aplicación.
- C++: lenguaje de tercera generación de desarrollo para la creación de librerías de enlace dinámico (DLL).
- OCR y scanner: hardware y software adicional para automatizar la entrada de información externa al tesauro.

El proyecto cuenta además con los siguientes apoyos tecnólogicos:

- Nuevos conocimientos de repositorios para almacenamiento y clasificación de software.
- CASE (herramientas de ayuda para desarrollo de software).
- Metodología orientada a objetos.
- Reutilización de códigos (reutilización de software).
- Groupware (trabajo en grupo).
- Prototipado GUI (Graphical User Interface).
- Arquitectura Cliente/Servidor.
- Protocolos OLE y DDE.

## 4. Conclusiones

La concepción modular del proyecto permite su constante expansión y desarrollo. En la actualidad se han conseguido los siguientes hitos:

- Diseño y desarrollo de un Tesauro Global (GT) aplicado a la gestión de software: Tesauro de Software (ST).
- Puesta en marcha de un componente de clasificación e indización de diferentes tipos de documentos (multimedia) con vistas a su referenciación en el ST.
- Integración de Aplicaciones mediante los protocolos OLE/DDE: podemos obtener información de cualquier tipo, y mezclarla de una aplicación a otra
- Desarrollo de las herramientas de gestión del tesauro global con vista a permitir su actualización y modificación de estructura interna de forma automática y gráfica (consulta, gestión, e indización morfológica y sintáctica).
- Definición de los procesos de consulta al tesauro con vistas a poder recuperar de forma inteligente los documentos y la información asociada existente.

- Utilización de la metodología ROM, de reutilización de componentes de software
- Sistema de administración general del repositorio: configuraciones, seguridad, importación y exportación del mismo tesauro u otros
- Acceso vía INTERNET, colocando el sistema a disposición de usuarios remotos

## 5. Anexo: Proyectos de Investigación del Grupo

1. *Proyectos financiados por la Unión Europea:*
   a) IMDEX: Indización distribuida de información multimedia, orientada a la indización de imágenes. (1997-1999).
   b) AUTOSOFT: Indización y clasificación automática de Software. (1998-2000).
2. *Proyectos financiados por la CICYT:*
   a) METRICA 3: Metodología de Desarrollo de Software Métrica Versión 3, utilizada como estándar por la administración pública. (1996-1998).
   b) GATOAC: Generación automática de Tesauros orientada a las arquitecturas distribuidas de componentes binarios. (1997-1999).

## 6. Referencias

Aitchison, J. (1987). Thesaurus Construction : A Practical Manual. ASLIB : 1987.

Dewey, M. (1979). Decimal Classification and Relative Index. Forest Press Inc. :1979.

Moreiro, J. A. (1993). Aplicación de las Ciencias del texto al resumen documental. Madrid : Universidad Carlos III-BOE, 1993.

Van Slype, G. (1991). Les Langages d'indexation: conception, construction et utilisation dans les systèmes documentaires. Paris : Les Editions d'organisation, 1991.

Velasco, M. ; Lloréns, J. ; Martínez Orga, V. (1997). Generación automática de representaciones de dominios. // II Jornadas en Ingeniería de Software. JIS97. San Sebastián. Spain.