

# GeoCrawler y gvSIG: un Tándem para la Generación Automática de Metadatos

Arturo Beltran, Joaquín Huerta, Laura Díaz, Carlos Granell



# Contenidos

- ▶ **Introducción**
  - Motivación
  - Qué son los Metadatos?
  - Caso de uso
- ▶ **Metodología propuesta para la generación automática de metadatos**
- ▶ **GeoCrawler**
  - Descripción
  - Arquitectura general
  - Una primera versión
- ▶ **Trabajo Futuro**
- ▶ **Conclusiones**
- ▶ **Preguntas**

# Introducción

- ▶ La información juega un papel fundamental en la sociedad donde vivimos
  - Deseo de información a nivel global
  - Fácilmente accesible en un entorno colaborativo
  - Esencial organizar, publicitar y facilitar el acceso
  - Gran cantidad de sistemas de información: Bibliotecas digitales, directorios y buscadores de internet
- ▶ En qué situación nos encontramos en el contexto de los SIG/IDEs ?

# Motivación

- ▶ Actualmente
  - Creación y publicación de recursos en paralelo
  - SIG e IDEs navegación poco transparente
  - Difícil descubrir y acceder a los recursos
- ▶ Metadatos: pieza clave en una IDE
  - Descripción de recursos
    - Organizar, publicitar y facilitar el acceso
  - Base del resto de componentes de una IDE
- ▶ No solo deseable, **requerido** por iniciativas como INSPIRE (Directiva 2007/2/EC)
  - IG armonizada, de calidad y disponible
    - Creación y mantenimiento de MD y servicios de descubrimiento

# Qué son los Metadatos ?


- ▶ *“Datos estructurados acerca de los datos”*
- ▶ *“Datos que describen los atributos de un recurso”*
- ▶ *“Información acerca de los datos”*
  
- ▶ Beneficios de la creación de metadatos
  - Organizar y mantener colecciones de datos: reusabilidad
  - Publicitar la existencia en catálogos: poder ser encontrados
  - Facilitar acceso, adquisición y utilización: interoperabilidad

# Caso de Uso

## ▶ Contexto general actual

- Multitud de datos no accesibles al público
- Deseo o imposición de publicar esos recursos

## ▶ Contexto SIG

- Necesidad de compartir grandes conjuntos de datos
- Por directivas como INSPIRE no solo deseable, sino **requerido !**
- Hacer esos datos accesibles y fácilmente descubribles  **PROBLEMA**
  - Recursos sin documentar

## ▶ Aproximaciones

- Poner personas a documentar y publicar los recursos
  - Altamente ineficiente, costoso y propenso a errores
- Solución que nos permita documentar y publicitar nuestros recursos de una forma fácil y transparente al usuario

# Metodología para la generación automática de metadatos

- ▶ **Objetivo:** describir los recursos de forma completa y veraz
- ▶ Técnicas para la generación de metadatos
  - Introducción manual por teclado
  - Extracción de metadatos del propio dato
  - Extracción de metadatos a partir del contenido
  - Recolección en el proceso de creación de los datos
  - Aprovechamiento del contexto
  - Búsqueda (look-up) desde una tabla de referencia
  - Medición del valor
  - Computación del metadato
  - Inferencia del metadato
- ▶ **Metodología propuesta:** combinación de todas ellas orquestadas de forma eficiente

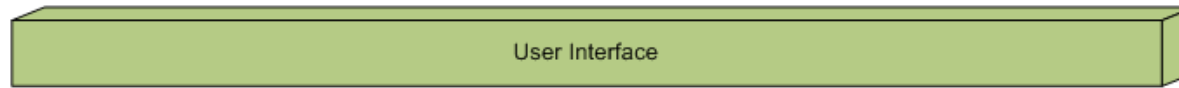
# Qué es GeoCrawler ?

- ▶ **Objetivo:** Solución que nos permita documentar y publicitar nuestros recursos de una forma fácil y transparente al usuario
- ▶ Aplicación tipo *crawler*
  - Inicialmente de ámbito local
  - Inspecciona el contenido de forma metódica y automatizada
  - Lista de todos los recursos *interesantes* en el sistema
- ▶ En base a la lista
  - Describir los recursos: Metodología propuesta
  - Indexarlos y/o Publicarlos
- ▶ Implementación
  - Java
  - Basado en estándares

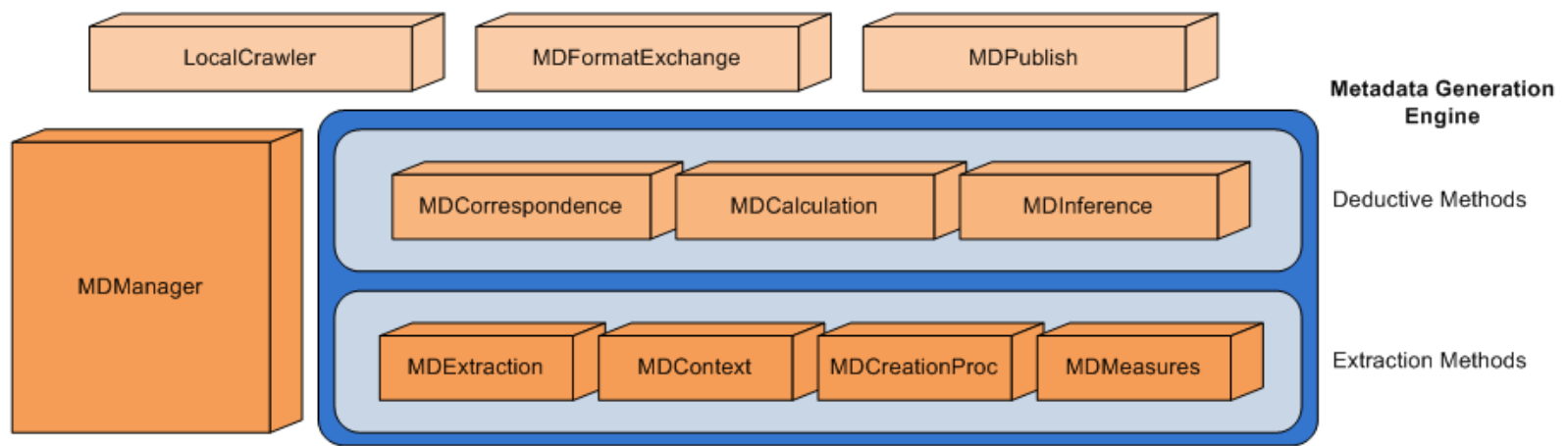


# Arquitectura General

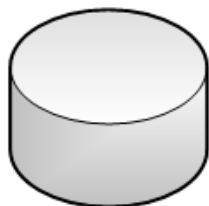
Presentation Tier



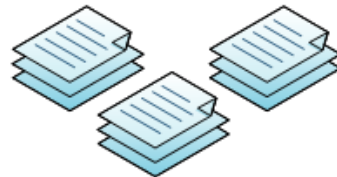
Logic Tier



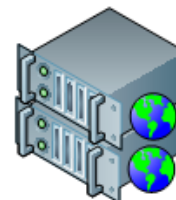
Data Tier



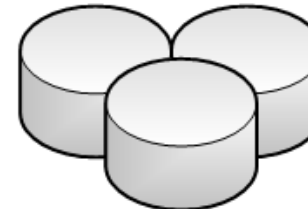
Metadata Database



Data Files



Services

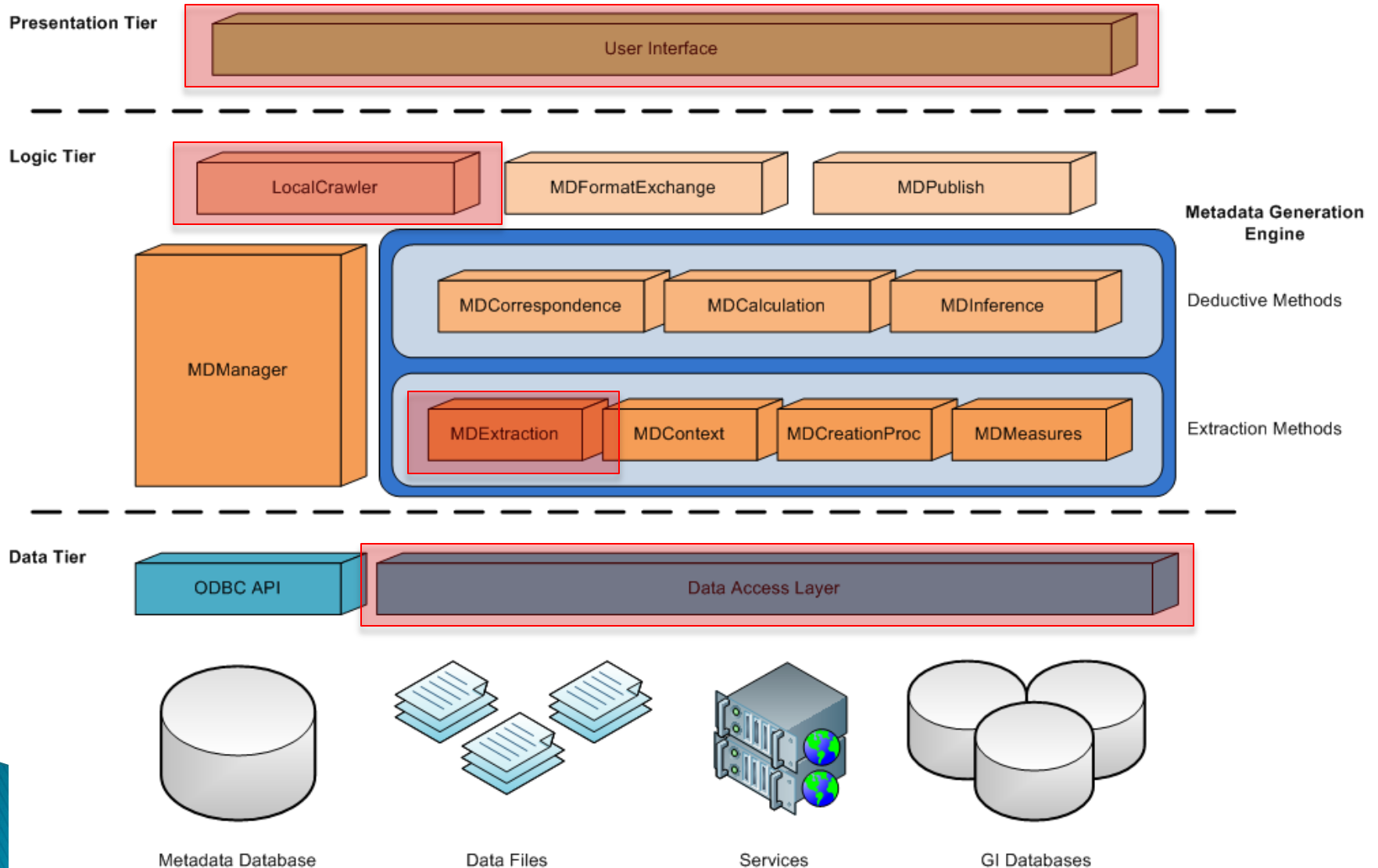


GI Databases

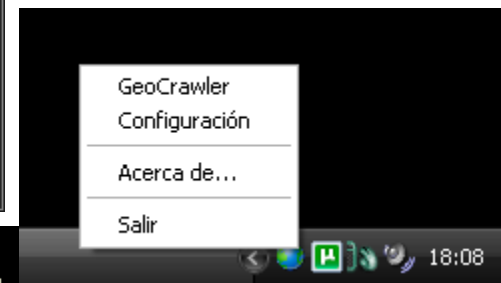
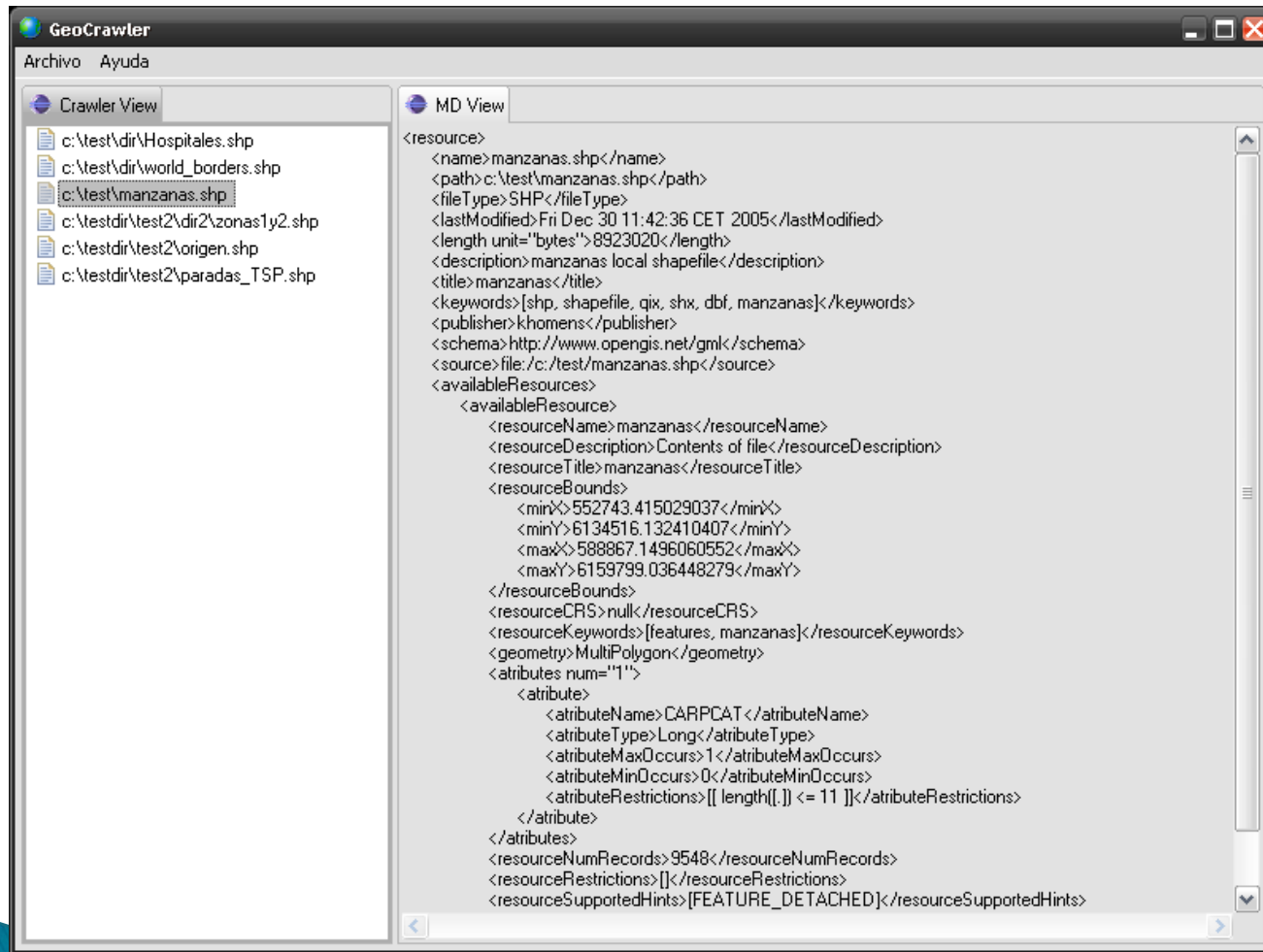
# Primera Versión

- ▶ Funcionalidad muy limitada
  - Aplicación residente
  - Tareas de *crawling*
  - Extracción de metadatos de ficheros de datos vectoriales
    - De los propios recursos: Sistema Operativo
    - Del contenido de los recursos: GeoTools

# Primera Versión



# Primera Versión: Interfaz



# Trabajo Futuro

- ▶ Completar la implementación de GeoCrawler
  - Trabajar en el acceso a los recursos
    - Plataforma que ofrezca información y acceso de forma lo más homogénea posible a recursos heterogéneos
      - Apostamos por la capa de acceso a datos (DAL) de gvSIG.
  - Completar la implementación del motor de generación de metadatos
    - Recolección durante el proceso de creación de los datos
      - Fuente de información “volátil”
      - Información muy importante y raramente tenida en cuenta
      - gvSIG puede jugar un papel muy importante

# Conclusiones

- ▶ Metodología para generación automática de metadatos
- ▶ GeoCrawler
  - **Objetivo:** recopilar, describir, catalogar y publicar recursos
- ▶ Primera versión de GeoCrawler
  - Validar la arquitectura modular diseñada
  - Funcionalidad de *crawler*: lista de recursos disponibles
  - Puesta en práctica metodología
    - Extracción de metadatos explícitos en los propios datos y su contenido
    - Necesidad plataforma común de acceso a recursos: DAL de gvSIG
    - Recopilar información en el proceso de creación de los datos: importante papel de gvSIG

# Preguntas

