Tecnologías de Información

Tema 1. Introducción y lenguajes de consulta

Bibliografía:
Baeza-Yates y Ribeiro-Neto: “Modern Information Retrieval” 1999 (Capítulo 1).

Dr. Carlos Castillo

UPF – 2005

¿Sinónimos?

Datos

Información

Conocimiento

Introducción

- Datos
  - Hechos representados en forma legible
    - *Bases de datos relacionales*
- Información
  - Datos organizados de forma coherente
    - *Sistemas de recuperación de información*
- Conocimiento
  - Información que puede ser utilizada para un propósito

Temas de esta asignatura

- 1. Lenguajes de consulta

  Formular necesidades de información
- 2. Procesamiento del texto
- 3. Modelos de recuperación

  Transformar contenido en información
- 4. Evaluación de los resultados
- 5. Indexación

  Acelerar búsqueda de información
- 6. Aplicaciones

Temas de esta clase

- Proceso de recuperación de información
- Punto de vista del usuario
  - Lenguajes de consulta
  - Interfaces

Motivación

- Colecciones de datos enormes y potencialmente valiosos
- Cifras
  - Una presentación con algunas fotos = 500 KB
  - 1 minuto de audio comprimido = 1 MB
  - 1 CD = 650 MB
  - Biblioteca grande (congreso US) = 10 Terabytes
  - Todas las páginas Web >> 170 Terabytes
  - Todos los e-mail > 400.000 Terabytes por año

¿Cuánta información?

- Kilobyte KB = 1.000 bytes
  - 100KB = una fotografía a baja resolución
- Megabyte MB = 1.000.000 bytes
  - 5MB = todos los libros de Shakespeare
- Gigabyte GB = 1.000.000.000 bytes
  - 1GB = una camioneta llena de libros
- Terabyte TB = 1.000.000.000.000 bytes
  - 1TB = 50.000 árboles hechos papel e impresos
  - 2TB = una biblioteca grande

Números grandes

- Petabyte PB 1.000.000.000.000.000 bytes
  - 70 PB = toda el cine y tv producido en 1 año
  - 200 PB = todo el material impreso en el mundo
- Exabytes EB 1.000.000.000.000.000.000 b
  - 17 EB = todas las conversaciones telefónicas en un año
  - 5EB = cantidad de información

Mayor acceso a la información

- Múltiples fuentes de información
  - Más colecciones disponibles on-line
    - Directorio de buscadores especializados: http://www.invisible-web.net/
  - A veces desagregadas
- Auto-servicio de información
  - Usuarios no especializados

Algunas preguntas clave

- ¿Cómo representar contenidos?
  - Para búsquedas más efectivas
- ¿Cómo representar necesidades de información?
  - Ayudar al usuario a formular consultas
- ¿Cómo seleccionar lo relevante?
- ¿Cómo desplegar los resultados?

Desarrollos históricos

- Libros => explosión de información
- Índices en tarjetas (más versatilidad)

Sistemas de catalogación temáticos
Ejemplo: código Dewey

Código Decimal

- 000 Generalidades
- 100 Filosofía
- 200 Religión
- 300 Cs. Sociales
- 400 Lenguajes
- 500 Ciencias puras
- 600 Ciencias aplic.
- 700 Artes
- 800 Literatura
- 900 Hist. y geo.

Inventado en 1870 por Melvin Dewey, sistema jerárquico con divisiones de 10 en 10 ...

Ejemplo

- Tema: “Eclipses” en DEWEY
  - 500 Ciencias puras
  - 520 Astronomía
  - 523 Cuerpos celestes específicos
  - 523.7 Sol
  - 523.78 Eclipses

¿Por qué el esquema jerárquico no es suficiente?

- 1.- Desconocimiento de la jerarquía
- 2.- Clasificación manual
- 3.- ¿Todo se puede clasificar así?
- 4.- Múltiples jerarquías
- Posible solución:
  - Asociar palabras clave a documentos o temas
  - Búsqueda por palabras clave

Contenidos en un sentido amplio

- Texto sin etiquetar
  - Ej.: desde OCR
- Texto con campos
  - Ej.: base de datos
- Texto con estructura
  - Ej.: documentos legales, médicos
- Mapas, diagramas
- Fotografías
- Música, Videos

Contenidos multimedia

- Búsqueda por metadatos
  - Se traduce en búsqueda de texto
- Búsqueda por contenido
  - Audio: detección de voz, sonidos, etc.
  - Música: melodía, melodías similares, etc.
  - Imágenes: buscar caras, paisajes, etc.
  - Video: buscar gente, segmentación en escenas

Definición: metadatos

- “Datos acerca de otros datos”
- Ejemplos
  - Título
  - Fecha de creación
  - Autor
  - Tamaño
  - Formato
  - etc..

Proceso de recuperación de información

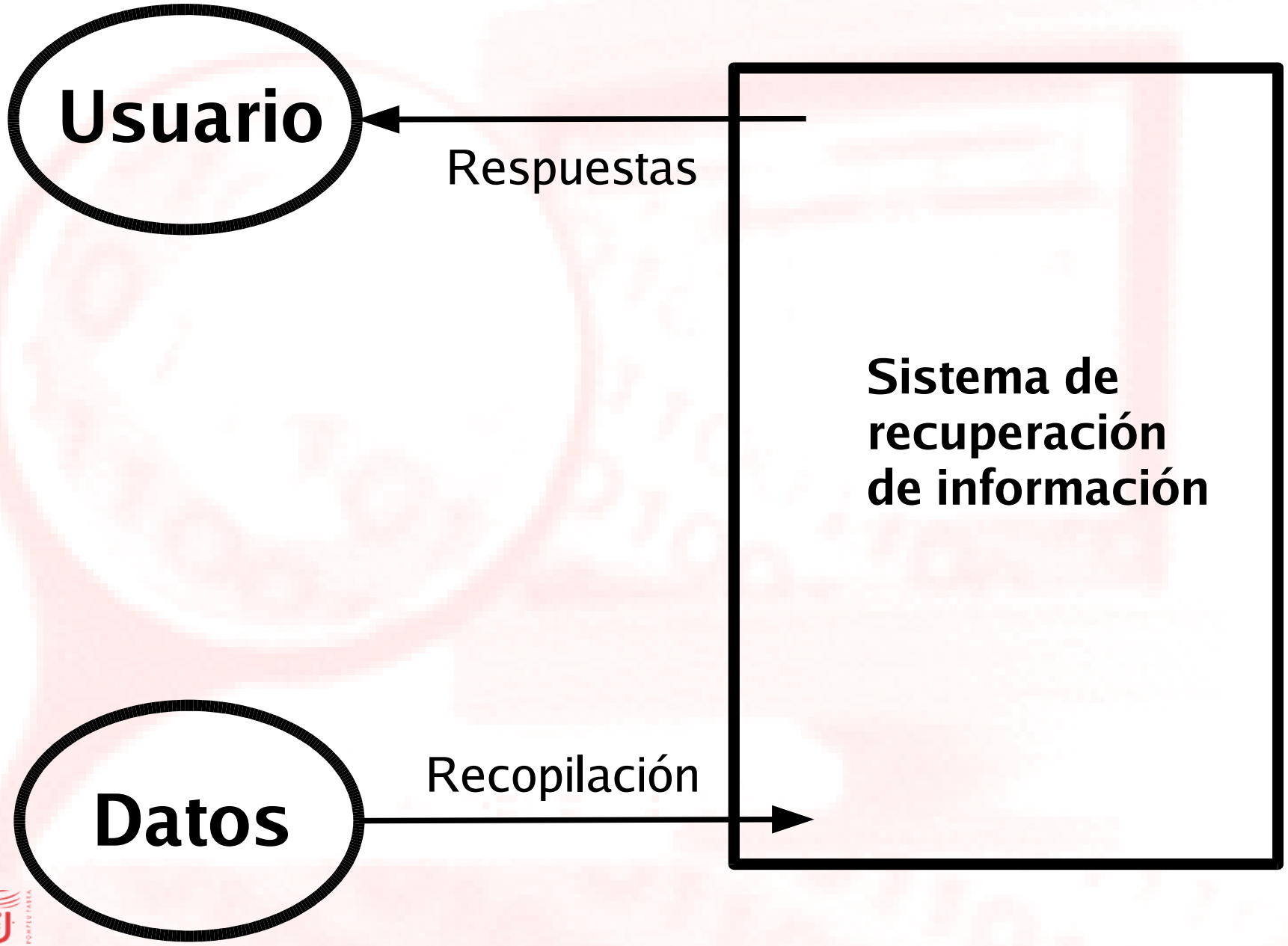
Recuperación de ...

- Recuperación de datos
  - ¿Qué documentos contienen un conjunto de palabras?
  - ¿Qué documentos responden a una cierta restricción estructural?
  - Cualquier diferencia => omitir documento
- Recuperación de información
  - Información sobre un tema o tópico
  - Semántica más relajada
  - Se toleran errores

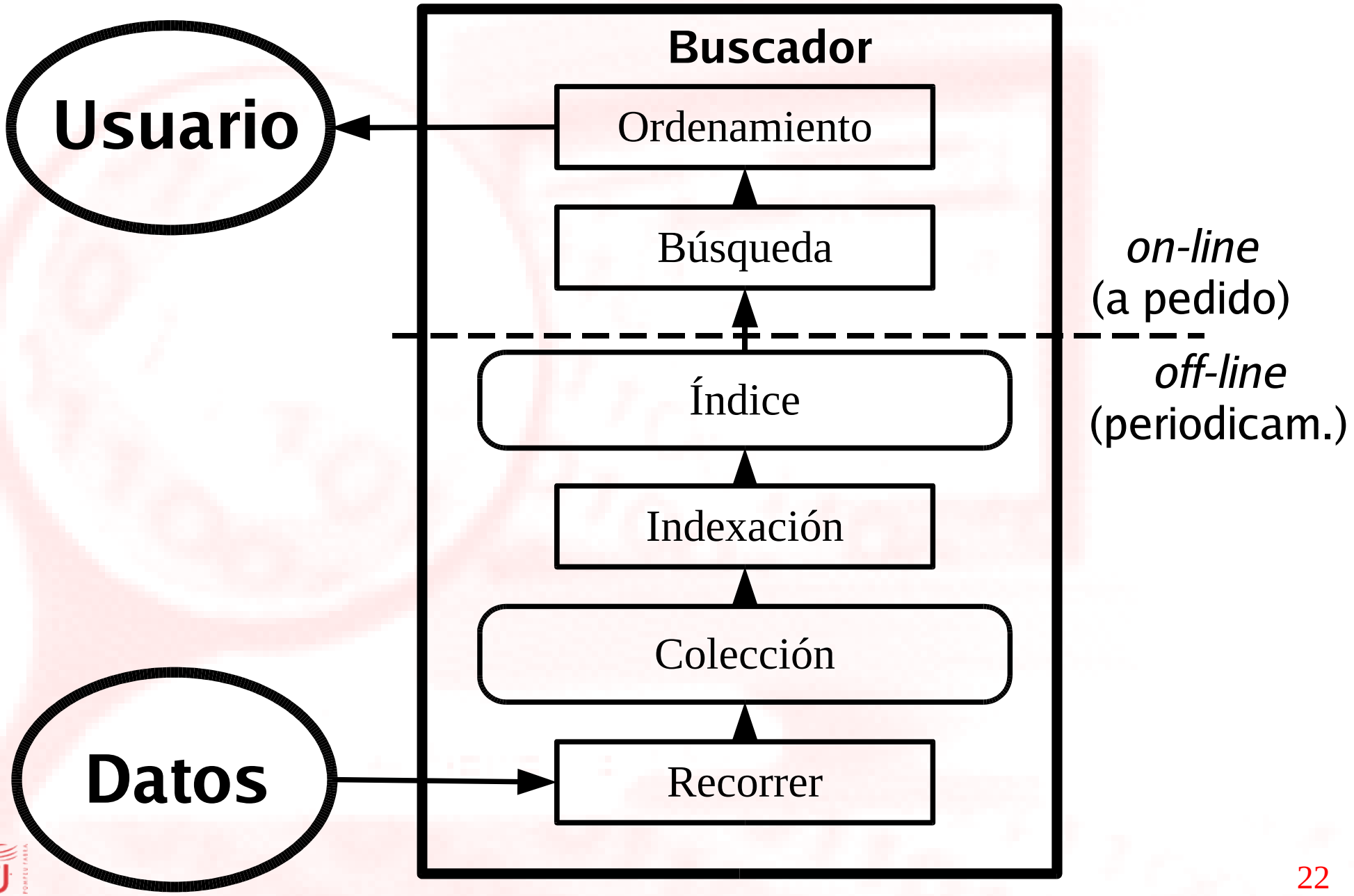
Recuperación de ...

	Datos	Información
Buscado	Registros	Documentos
Correspondencia	Exacta	Parcial
Pregunta	Completa	Difusa
Respuesta	Adecuada	Relevante
Modelo	Determinista	Probabilista
Lenguaje de consulta	Artificial	Natural
Clasificación	Un eje	Múltiples ejes

Proceso


Usuario
Respuestas
Sistema de recuperación de información
Datos
Recopilación

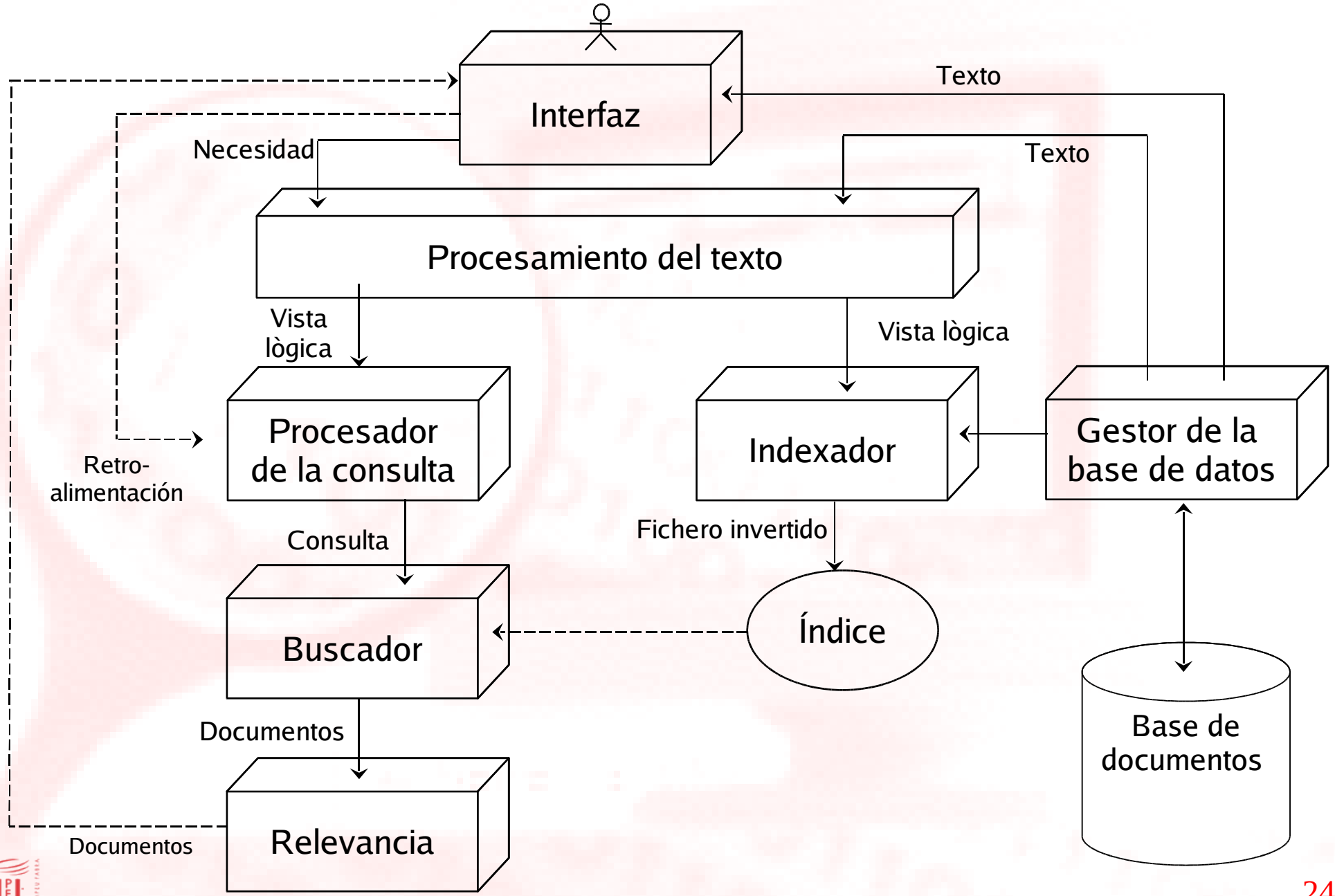
Proceso


Usuario
Buscador
Ordenamiento
Búsqueda
on-line
(a pedido)
off-line
(periodicam.)
Índice
Indexación
Colección
Datos
Recorrer

Proceso

- Acceso a los documentos
- Procesamiento de los documentos
- Indexación de los documentos
- Necesidad de información del usuario
- Procesamiento de la petición del usuario
- Búsqueda de la respuesta
- Ordenación de los resultados
- Presentación de los resultados
- Retroalimentación

Proceso

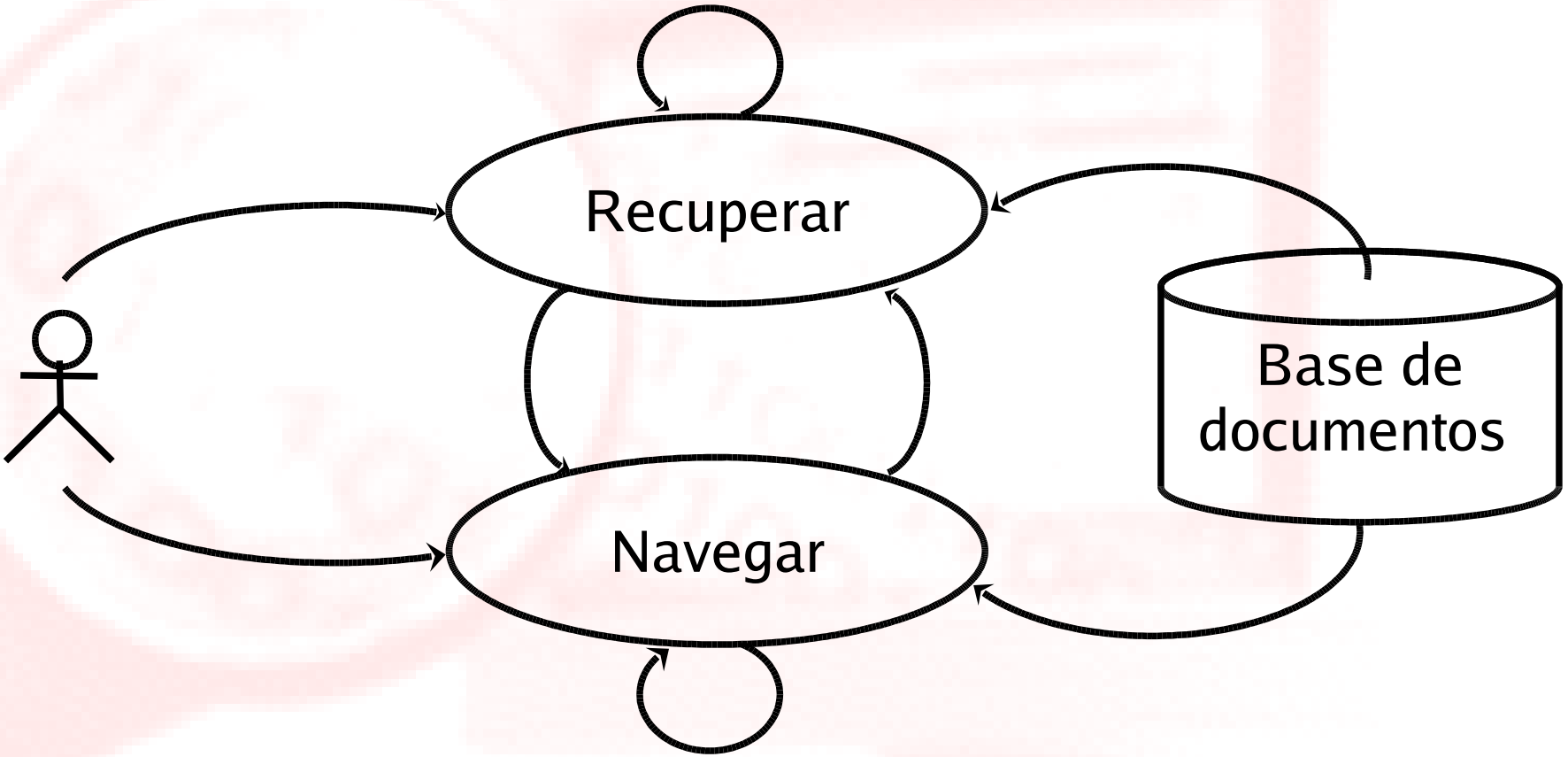

Interfaz
Texto
Necesidad
Texto
Procesamiento del texto
Vista lògica
Vista lògica
Procesador de la consulta
Indexador
Gestor de la base de datos
Retro-alimentación
Consulta
Fichero invertido
Buscador
Índice
Base de documentos
Documentos
Relevancia
Documentos

Sistema de recuperación de información

- Interpreta el contenido de los elementos de información
- Genera clasificaciones según relevancia de acuerdo a ciertos parámetros
- Si hay pocos elementos a buscar
  - Basta con retornar aquellos que son apropiados
- Si hay muchos elementos a buscar
  - La parte “fácil” es encontrar cuáles son apropiados
  - La parte “difícil” es seleccionar unos pocos

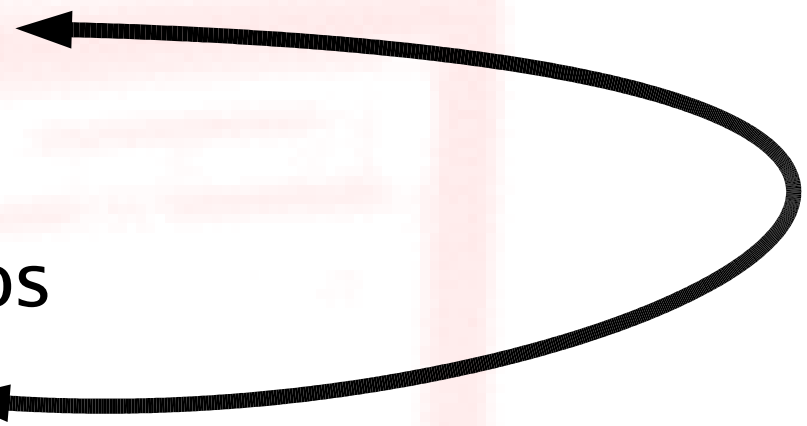
Punto de vista del usuario

Función del usuario



Función del usuario

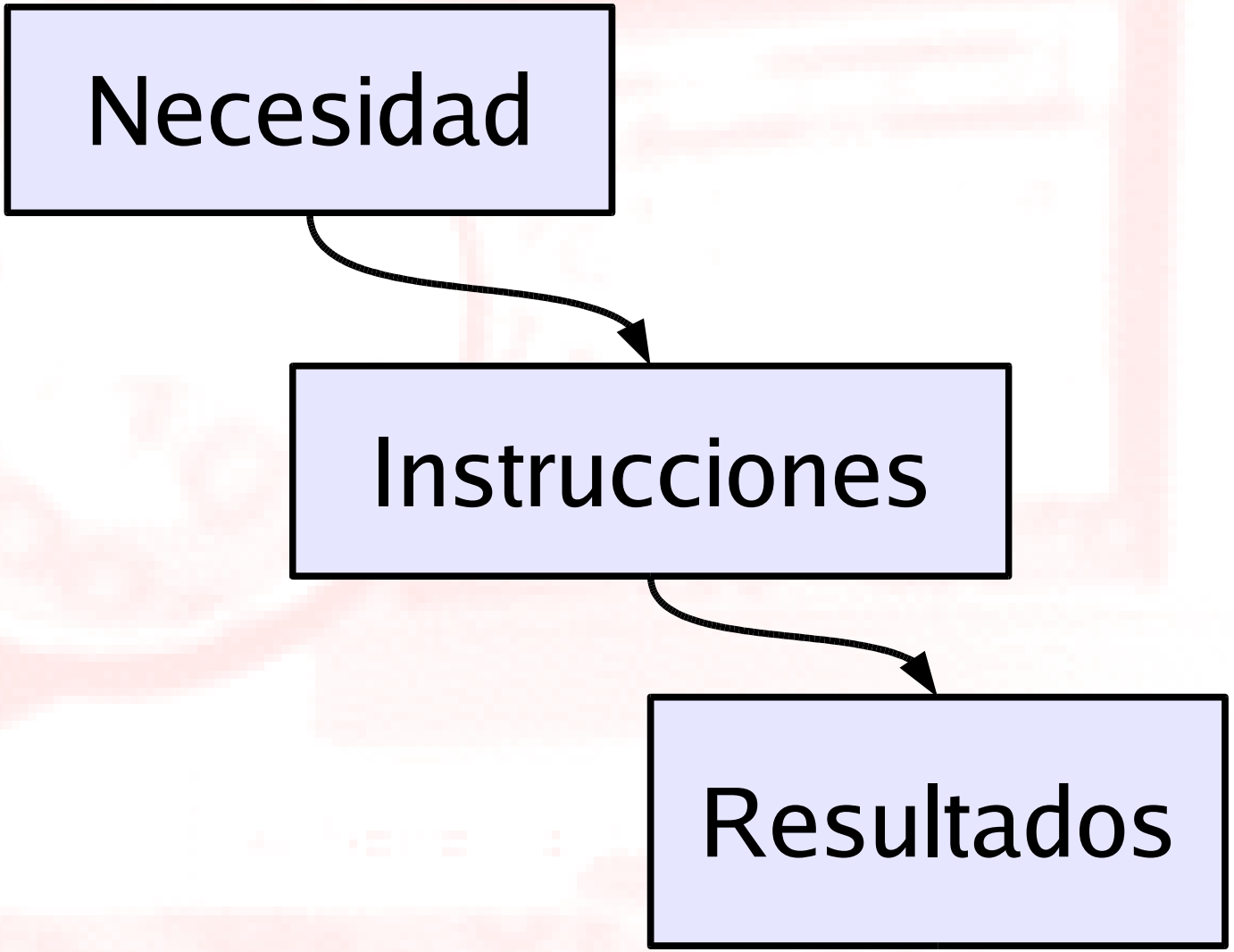
- Recuperación (retrieval)
  - Ingresar palabras clave
  - Revisar listas de resultados
- Navegación (browsing)
  - Navegar por sitios



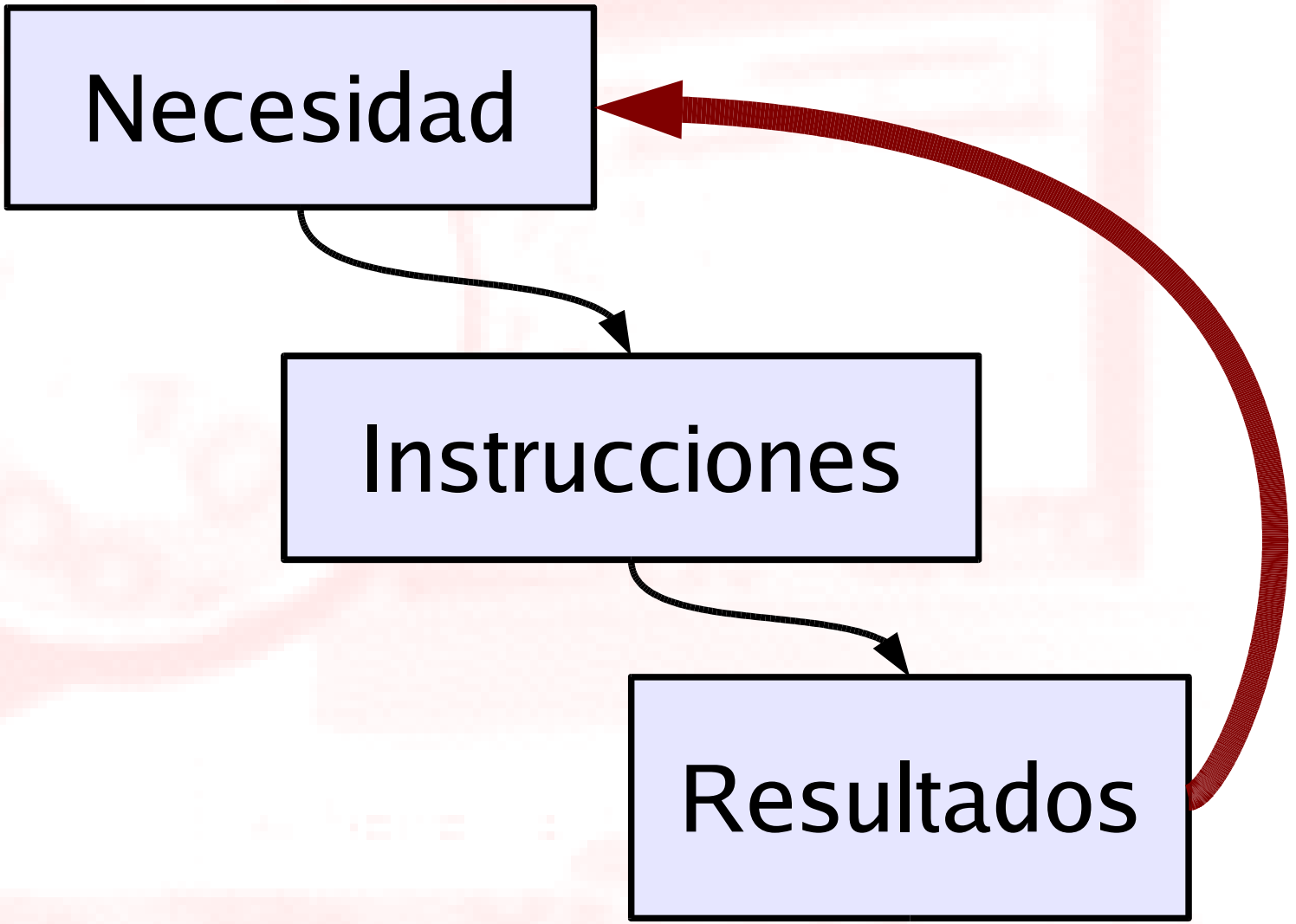
No es sólo filtrado (filtering)

- Filtering
  - Los usuarios deciden cuáles son sus preferencias, y cada vez que reciben un documento, deciden si les interesa o no
  - Las preferencias del usuario no cambian durante la sesión
- No es cierto, porque ...
  - Los usuarios no tienen claras sus preferencias a priori
  - Las preferencias van cambiando conforme se aprende durante la búsqueda

¿Comportamiento del usuario?



Comportamiento del usuario



Búsquedas típicas (zeitgeist)

Spain

Popular Sports Queries 2004	Popular Queries 2004	Popular Google News Queries - 2004	Popular Travel Queries 2004
1. fernando alonso	1. el mundo	1. real madrid	1. renfe
2. ronaldinho	2. renfe	2. alejandro sanz	2. spanair
3. valentino rossi	3. fernando alonso	3. boda real	3. iberia
4. real madrid	4. ronaldinho	4. irak	4. vuelos baratos
5. barcelona	5. valentino rossi	5. eta	5. air europa
	6. fernando torres	6. ronaldinho	
	7. hilary duff	7. atentado en madrid	
	8. melendi	8. la casa de tu vida	
	9. spanair	9. rajoy	
	10. los serrano	10. princesa letizia	

Necesidades del usuario

- Por tipo
  - Específica
    - Acción
  - No específica
- Por ocurrencia
  - Frecuente
  - Infrecuente
- Por formato buscado
  - Artículo largo, artículo breve, código fuente, sólo definición, lista de referencias, imágenes, etc.

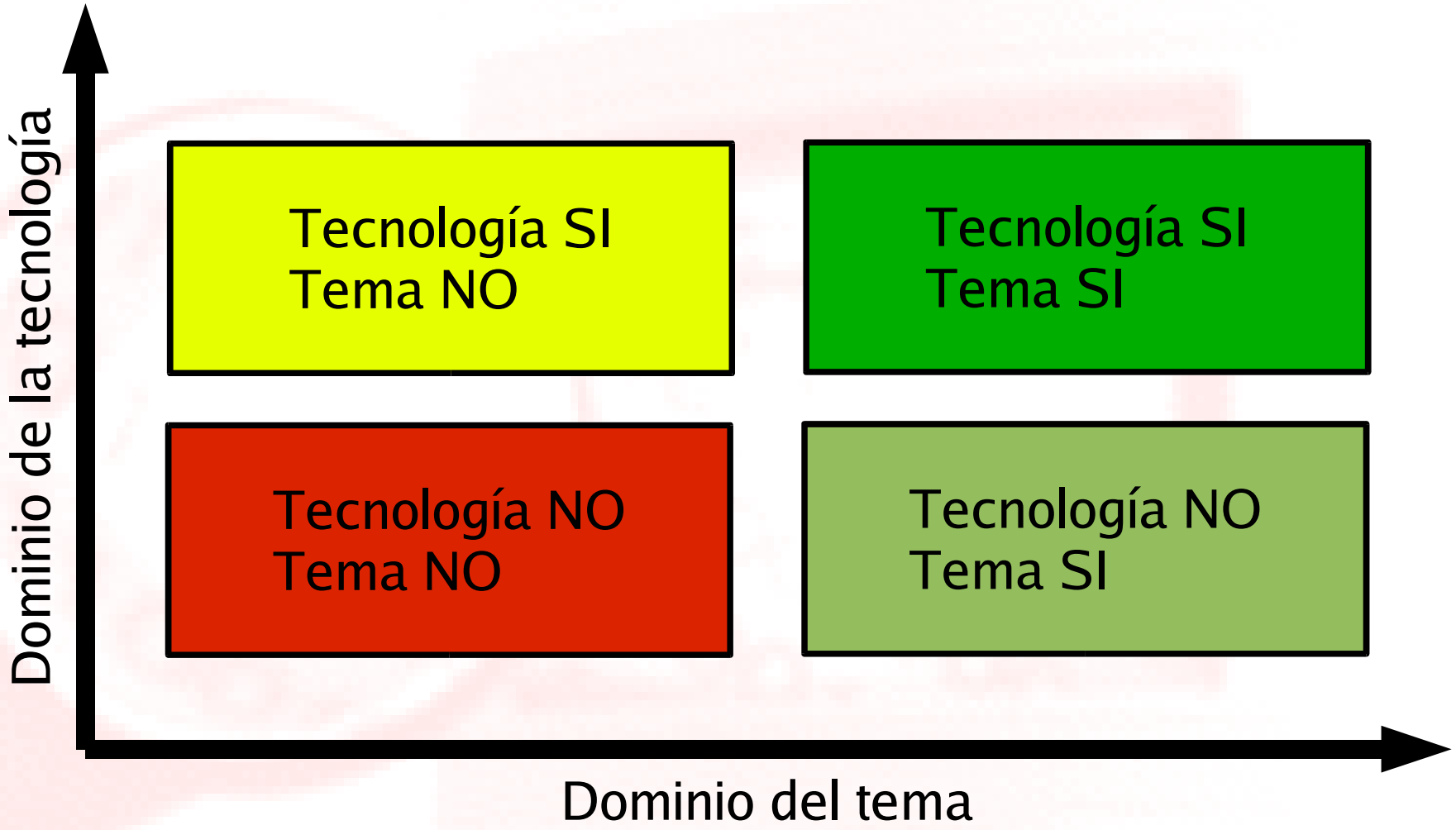
Evaluar un sistema de RI desde un usuario

- Tiempo de aprendizaje
- Sobrecarga cognitiva
- Expresividad
  - ¿Qué consultas se pueden hacer?
- Tiempo hasta primer resultado correcto
- Cantidad de *backtracking*

Usuarios

- Dominio tecnológico
  - Experto
  - No “alfabetizado” en términos informáticos => ¿informatizado?
- Dominio del tema consultado
  - Profundo
  - Superficial

Usuarios (cont...)



Para un usuario, es más importante saber qué palabras clave usar que conocer opciones avanzadas del buscador

Lenguajes de consulta e interfaces

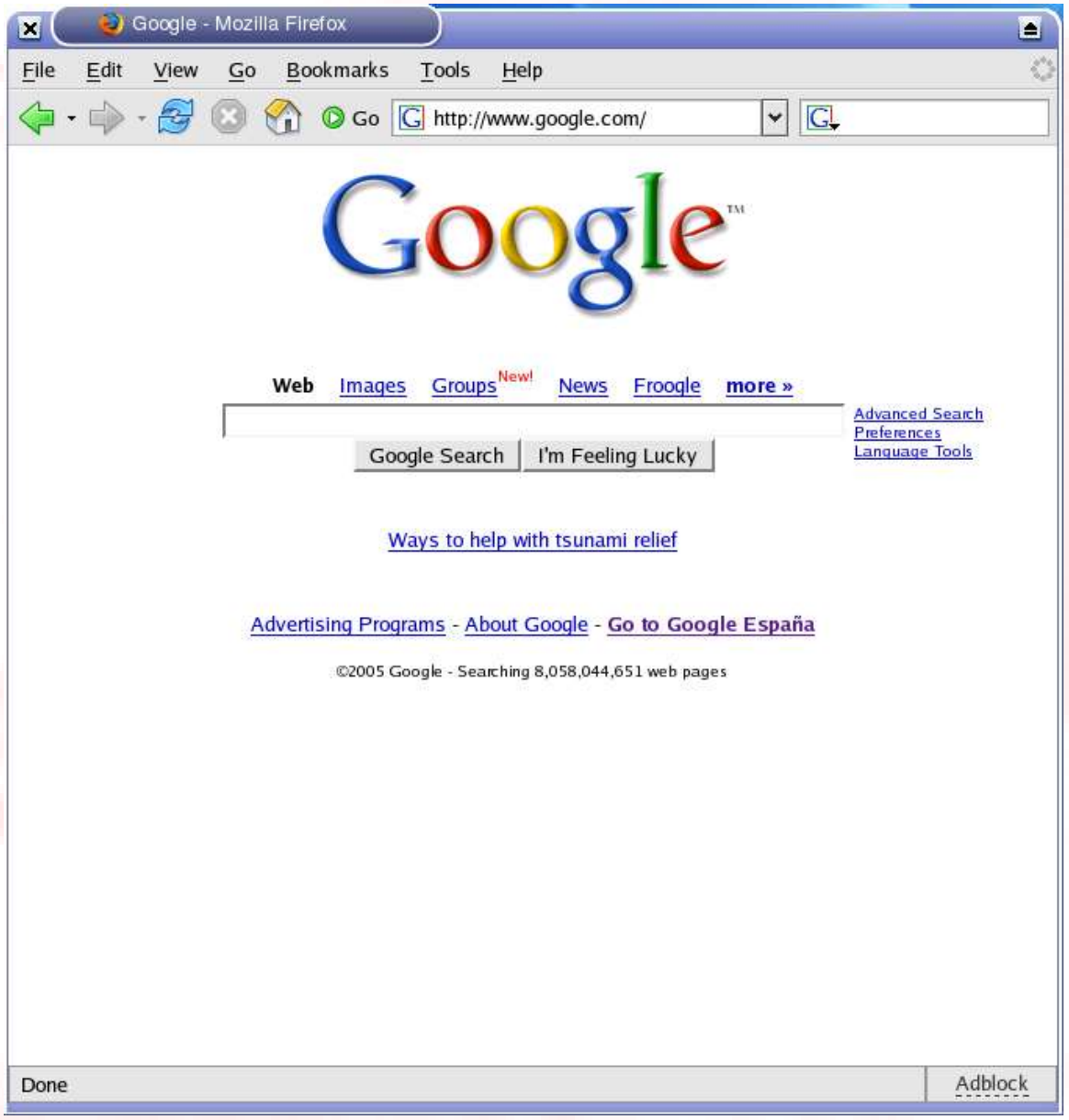
Definiciones

- Lenguaje de consultas
  - Recuperar información: admite ranking
  - Recuperar datos: no admite ranking
- Unidad mínima de recuperación
  - Sitio Web
  - Página
  - Extracto (abstract)

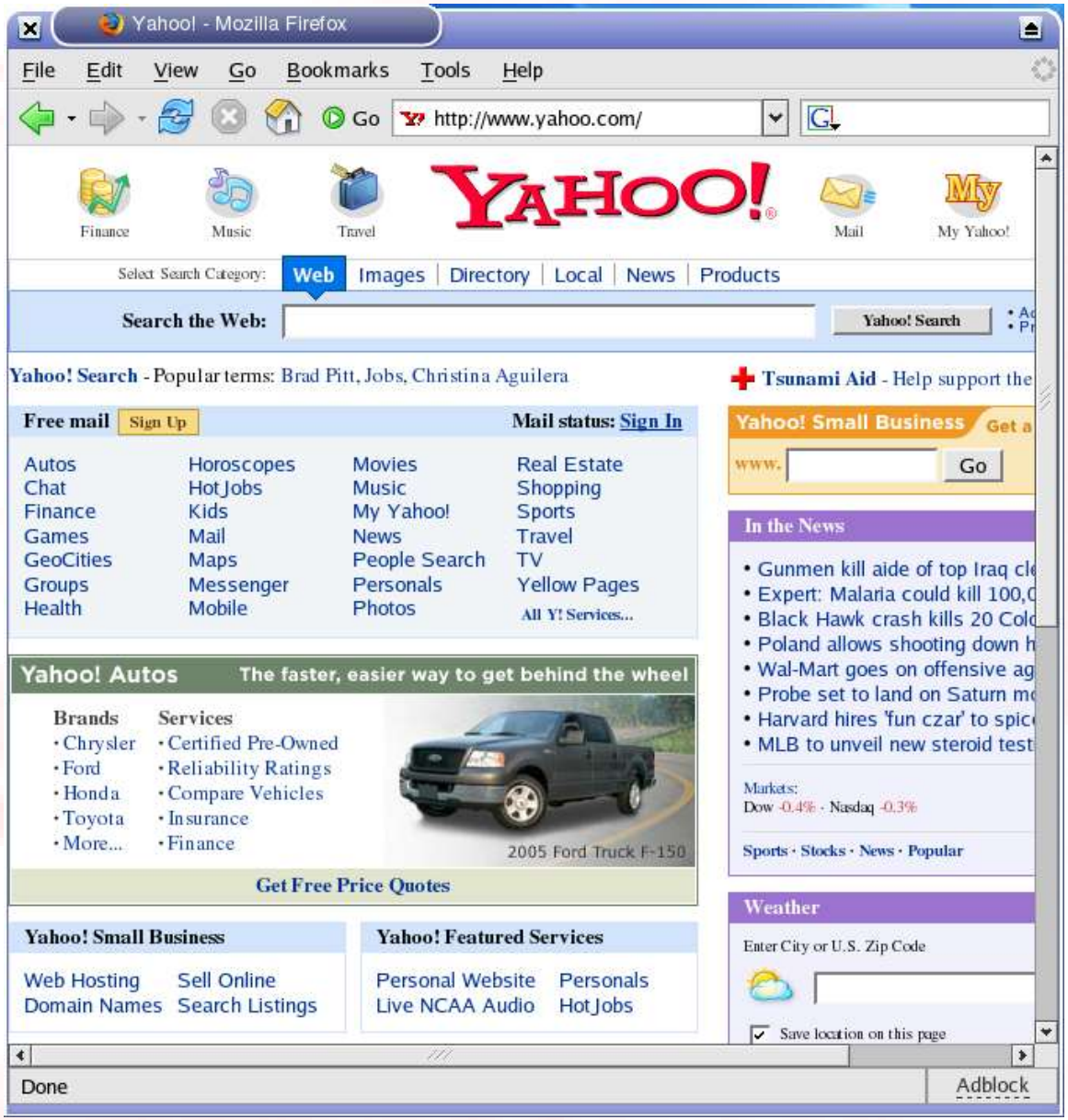
Información disponible

- Base de datos documental
  - Contenido
  - Propiedades
  - Estructura
  - Contexto de cada documento
- Base de datos de usuarios
  - Preferencias
  - Búsquedas anteriores
  - ... perfil

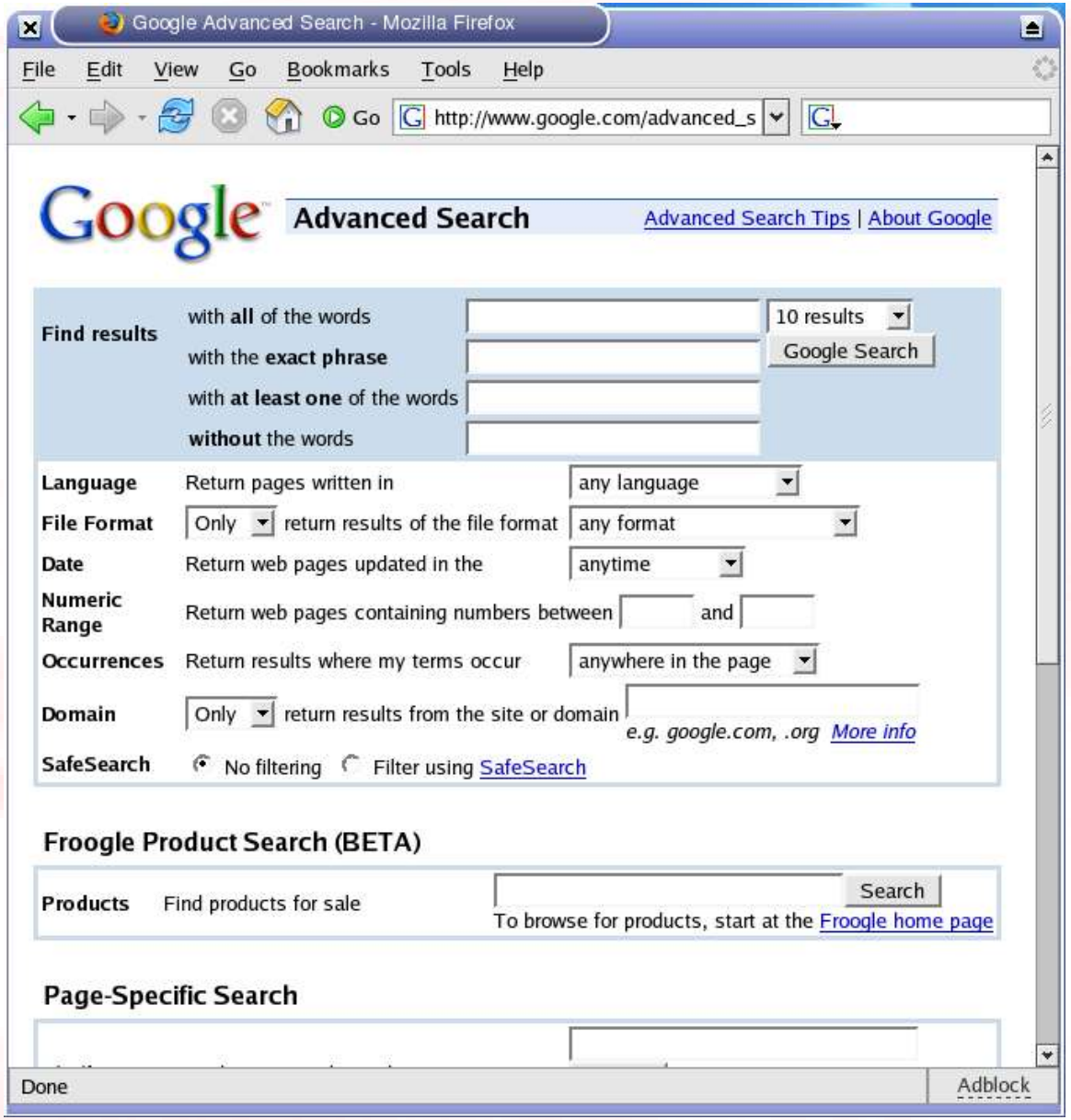
Ejemplo: buscador



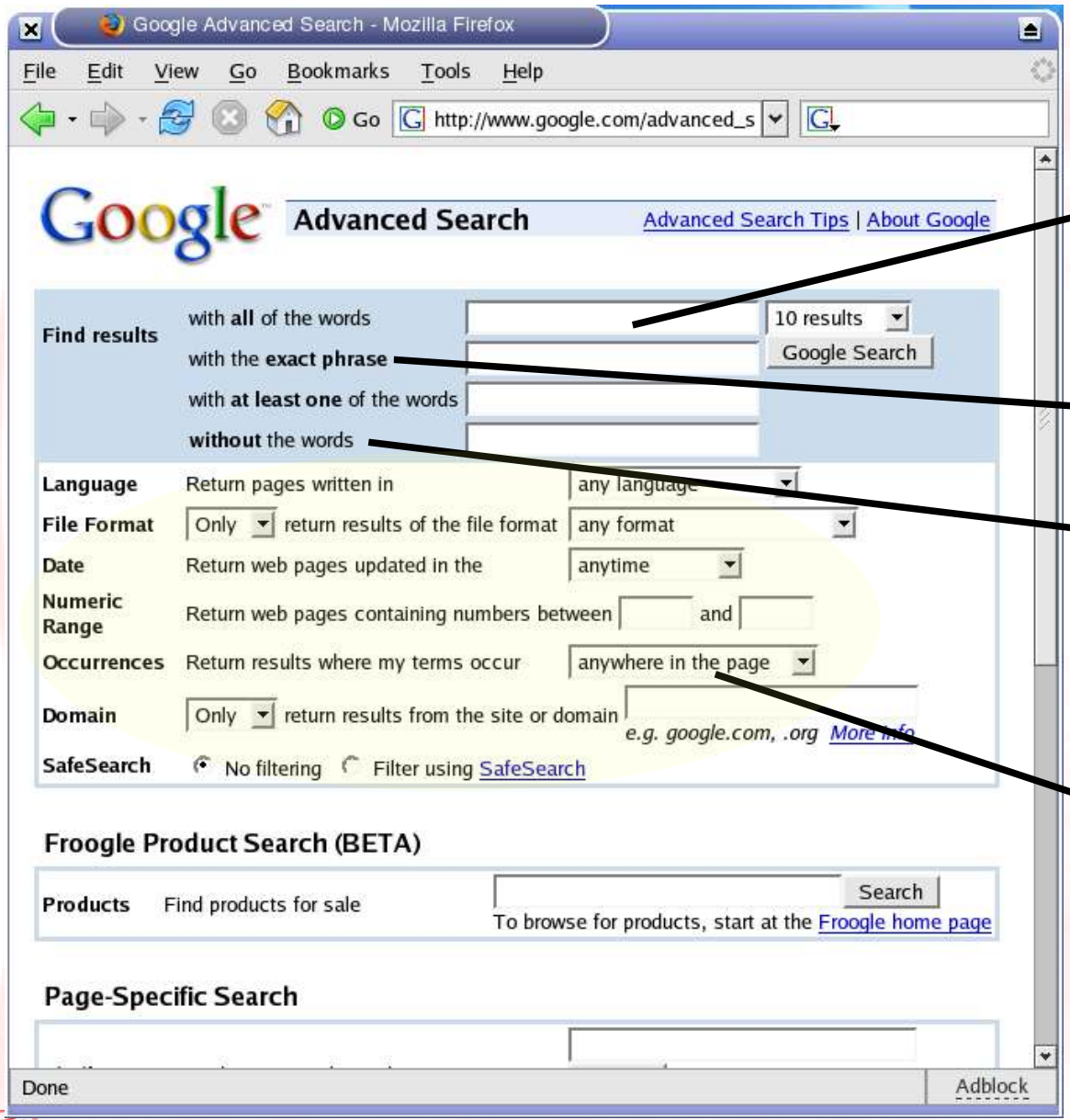
Ejemplo: buscador + portal



Ejemplo: búsqueda avanzada



Elementos



Keywords

Operadores proximidad

Operadores booleanos

Metadatos

Keywords o palabras clave

- Una palabra
  - Definicion de qué es *keyword*
- Palabras en contexto: proximidad
  - Proximidad 1 (frase)
    - Con stopwords
    - Sin stopwords
  - Proximidad arbitraria
  - Proximidad y ordenamiento (más próximas o en mismo elemento estructural, mejor)

Consultas booleanas

- AND, OR, NOT
  - AND (conjunción)
  - OR (disyunción)
- Problema: en el lenguaje no se usa así
  - Ej.: busco pisos grandes y soleados
    - Deben tener ambas características (y=AND)
  - Ej.: mujeres y niños primero
    - Basta con una característica (y=OR)

Lenguaje natural

- En teoría
  - Representación del conocimiento
  - Meta final de la recuperación de información

Ejemplo lenguaje natural

YOU MAY FIND THESE OPTIONS USEFUL:

Other people with your search have also asked:

Ask Where can I find government information for the country Nigeria ?

Ask Where can I find encyclopedic geographical resources on Nigeria ?

Web | Pictures | News | Local NEW! | Products | More »

What time is it in Japan? Search Advanced

Web Search: What time is it in Japan? 1-10 results out of 12,740,000

Email This | About

It is 3:02:33 AM JST in Tokyo, Japan.

Related Topi
› Japa
› Time
› Time Conv

Web Results

Web | Pictures | News | Local NEW! | Products | More »

How old is Madonna? Search Advanced

Web Search: How old is Madonna? 1-10 results out of 876,400

Source

Email This | About

Madonna is 46 years old.

At first a streetwise bubblegum-pop ragamuffin, Madonna used a mixture of talent, cleavage, and relentless self-promotion to become one of the most famous recording artists of the 20th century. She released her self-titled first album in 1983; other albums include Like... More»

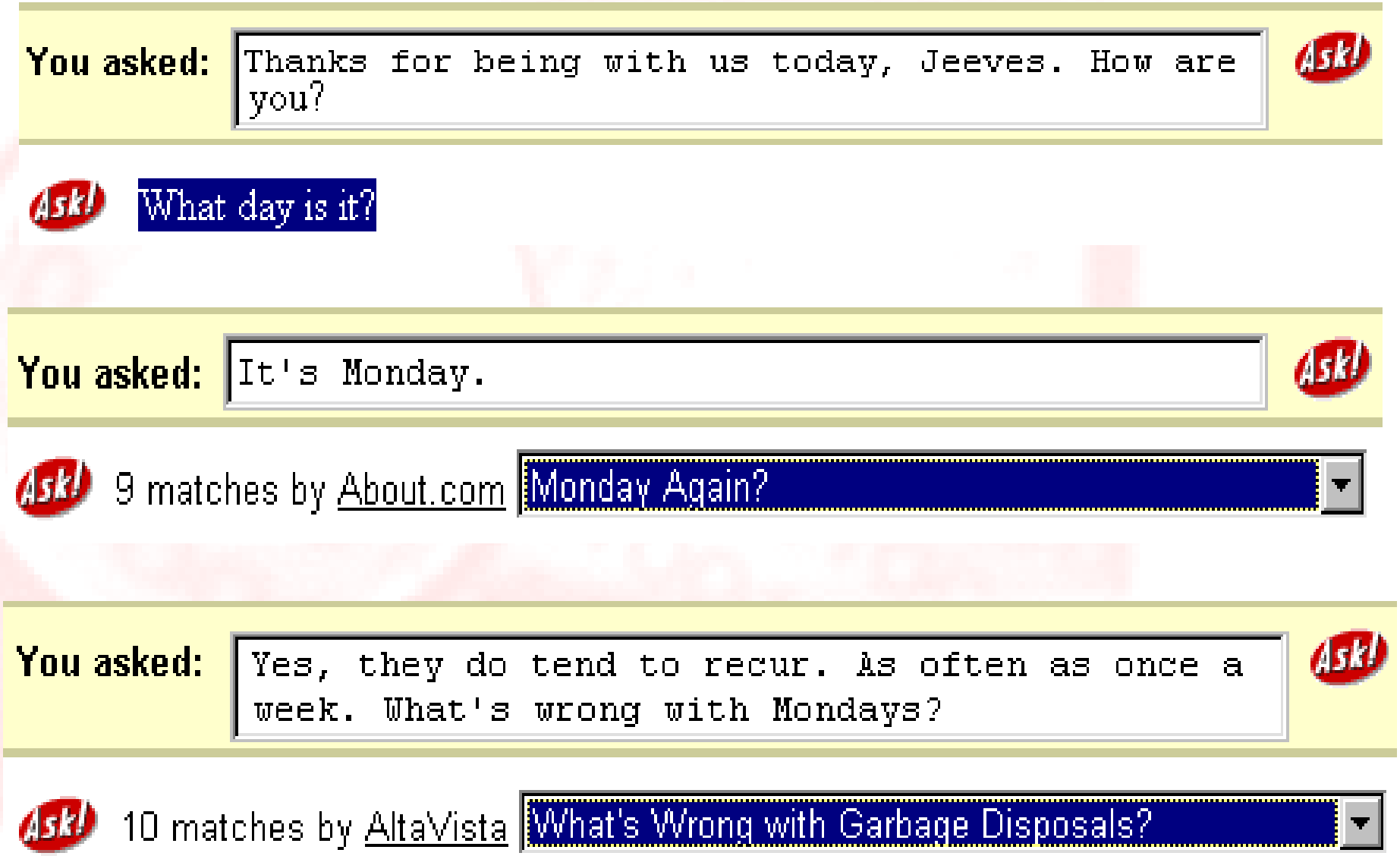
Find: Pictures | News | Products
Go To: Official Web Site | Filmography | Discography

Related Topi
› Biogr Singe
› When Mado
› Mado Birthc
› Full N Mado

Consultas en lenguaje natural

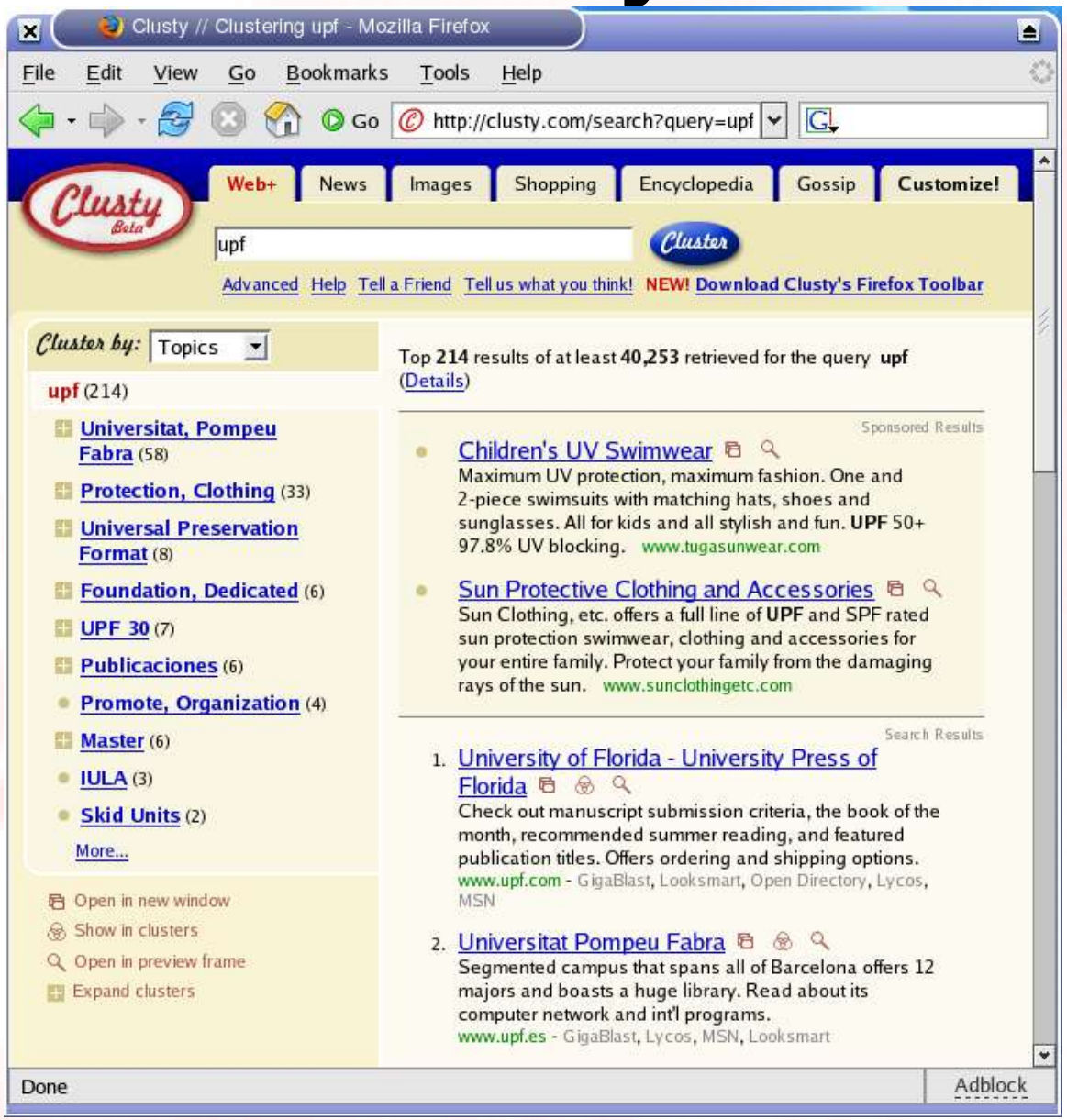
- En la práctica
  - Traducidas a boolean y pesos
  - Uso de diccionario de consultas
    - Keywords y representante
  - Funcionan bien si hay pocas preguntas posibles
- Pueden ser muy demandantes computacionalmente

Satirewire.com (1999)

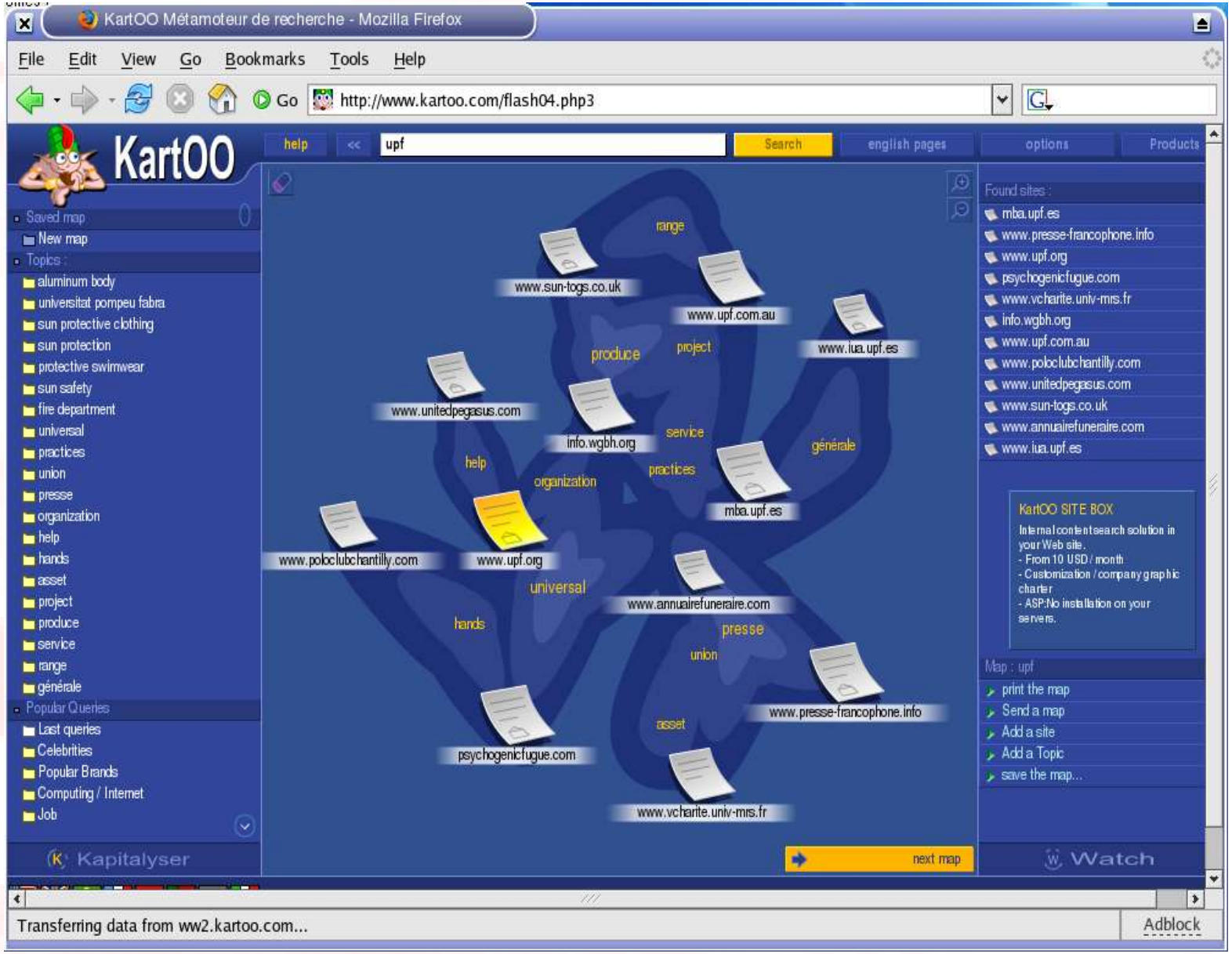


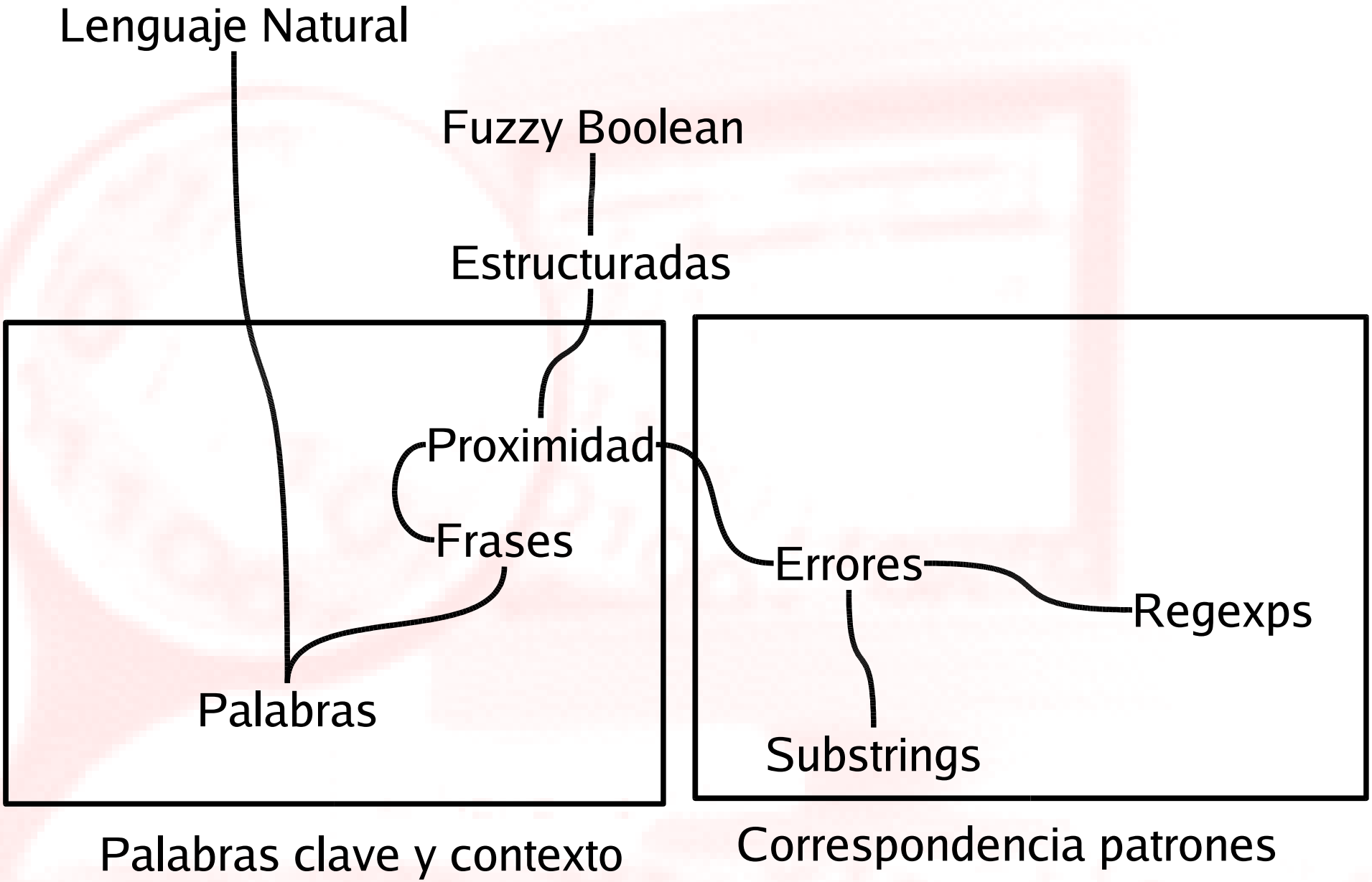
Interfaces con clustering: vivisimo

Interfaces con clustering: clusty



Interfaces con clustering, clustering gráfico: kartoo




Lenguaje Natural
Fuzzy Boolean
Estructuradas
Proximidad
Frases
Palabras
Errores
Regexps
Substrings
Palabras clave y contexto
Correspondencia patrones

Desafíos

- Cambiar la estructura y no las consultas
- Las consultas pueden ser complejas
  - Interfaz
- Lenguaje natural

Resumen

- Datos != Información
- Recuperación de información => relevancia
  - Admite errores
- Necesidad del usuario es cambiante
- Formular esta necesidad puede ser una tarea compleja