



## OPTATIVA

# Recuperación **Avanzada** de Información

---

Dr. J. Federico Medrano  
*@jfedemedrano*

Unidad N° 1 – Parte 2

# Temas a desarrollar

---

- ~~Organización de la asignatura~~
- ~~Introducción.~~
- ~~Documentos electrónicos.~~
- ~~Modelos de recuperación de información.~~
- ~~Algoritmos y estructuras básicas.~~
- ~~La recuperación de información en Internet.~~
- La recuperación multilingüe.
- Sistemas hipertextuales y recuperación documental.
- La recuperación de documentos multimedia o no textuales.
- La recuperación basada en la citación.
- Sistemas de filtrado y recomendación

# La Recuperación de Información Multilingüe

---

# RI multilingüe

---

- Un motor de búsqueda ideal recuperaría todos los documentos relevantes (lo que implica una **cobertura** completa) y sólo aquellos documentos que son relevantes (**precisión** perfecta).
- Este modelo tradicional lleva consigo muchas restricciones implícitas; entre ellas, la suposición de que la consulta y el documento están escritos en el mismo idioma.
- La recuperación de información translingüe, que trata el problema de encontrar documentos que están escritos en idiomas distintos al de la consulta.

# RI multilingüe

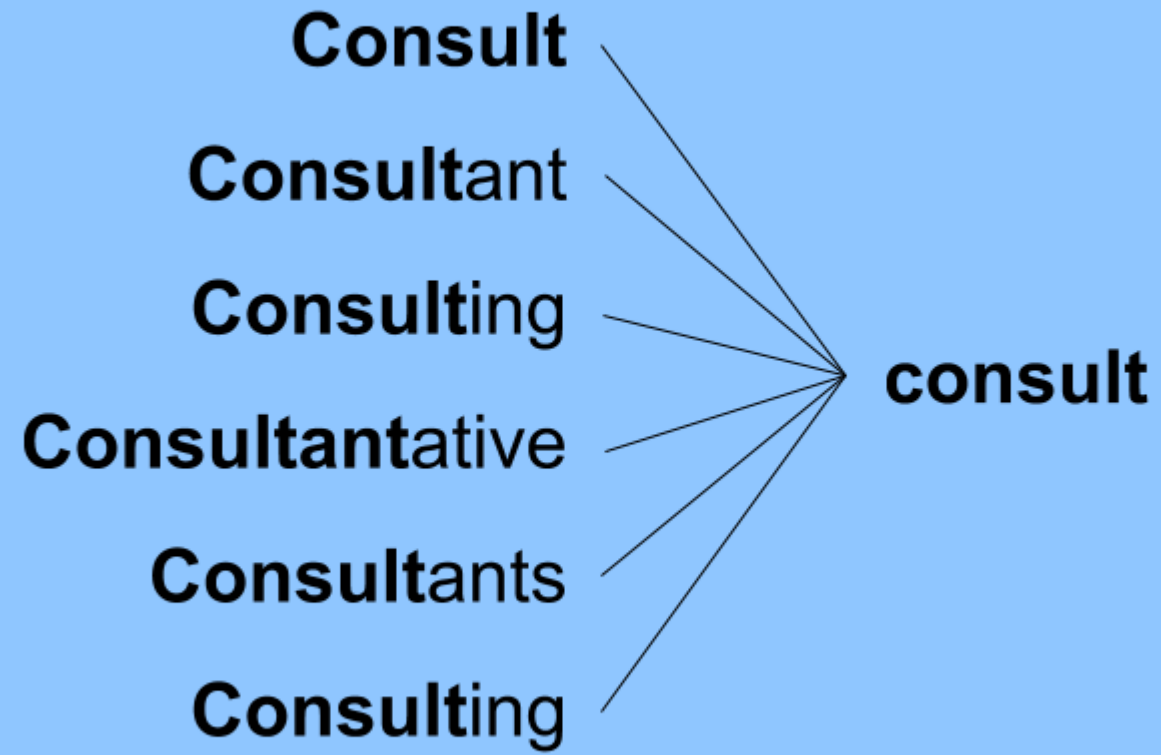
---

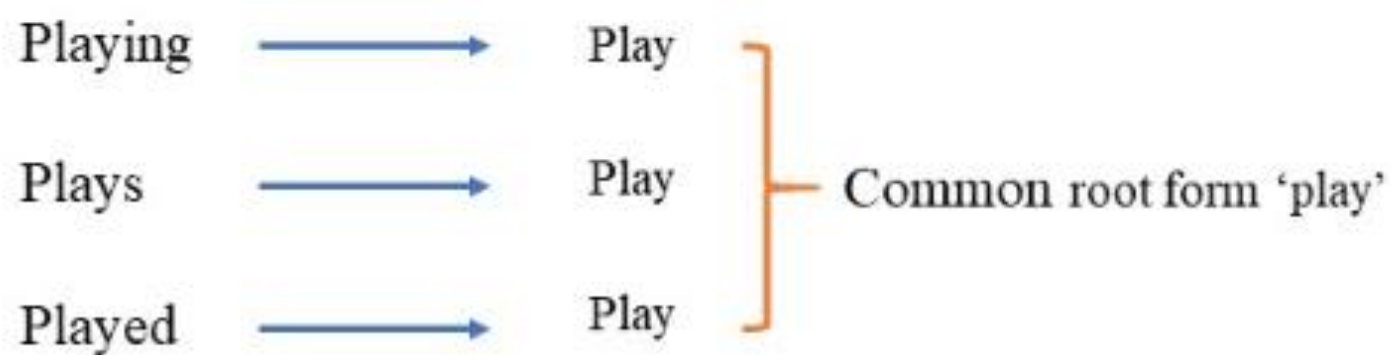
- En primer lugar es necesario estudiar las características propias de cada idioma a la hora de efectuar la recuperación monolingüe de documentos.
- En segundo lugar, hablaremos de búsqueda bilingüe cuando la consulta esté en un idioma origen y los documentos en un único idioma destino.
- Finalmente, hablaremos de búsqueda multilingüe cuando la consulta esté en un idioma origen y los documentos distribuidos en varias colecciones de idiomas diferentes.
- En este caso, el problema consiste en devolver un único ranking de documentos relevantes escritos en todos los idiomas considerados.

# Mejorando la búsqueda monolingual

---

- ***Stemming***, consiste en la obtención de la raíz de las palabras, de forma que el proceso de indexación se lleve a cabo sobre ellas en lugar de sobre las palabras originales. Asumiendo que dos palabras que tengan la misma raíz representan el mismo concepto, esta técnica permite a un sistema de recuperación de información **relacionar términos** presentes en la consulta y en los documentos que pueden aparecer bajo diferentes variantes morfológicas. Además, reduce apreciablemente el espacio de indexación.
- Existen varios algoritmos que se basan en un conjunto sencillo de reglas que truncan las palabras hasta obtener una raíz común.





am, are, is → be

Car cars, car's, cars' → car

Using above mapping a sentence could be normalized as follows:

the boy's cars are different colors → the boy car be differ color

# Mejorando la búsqueda monolingual

---

- **Lematización**, a diferencia del *Stemming*, reduce las palabras flexionadas adecuadamente asegurando que la palabra raíz pertenece al idioma. En la lematización, la palabra raíz se llama *Lema*. Un lema es la forma canónica, la forma del diccionario o la forma de cita de un conjunto de palabras.
- Por ejemplo: *runs, running, ran* son todas formas de la palabra *run*, por lo tanto ***run*** es el lemma de todas estas palabras.
- En español, por ejemplo, sabemos que *canto, cantas, canta, cantamos, cantáis, cantan* son distintas formas (conjugaciones) de un mismo verbo (*cantar*). Y que *niña, niño, niñita, niños, niñotes*, y otras más, son distintas formas del vocablo *niño*.

# Mejorando la búsqueda monolingual

---

- La lematización: relaciona una palabra flexionada o derivada con su *forma canónica* o *lema*.
- Como el proceso de lematización toma en consideración la probable *clase de palabra* (adjetivo, verbo, sustantivo...) — también llamados POS (Parts of Speech) — podemos usar dicha información para filtrar nuestra lista de lemas.

# Mejorando la búsqueda monolingual

---

- La lematización: relaciona una palabra flexionada o derivada con su *forma canónica* o *lema*.
- Como el proceso de lematización toma en consideración la probable *clase de palabra* (adjetivo, verbo, sustantivo...) — también llamados POS (Parts of Speech) — podemos usar dicha información para filtrar nuestra lista de lemas.
- <https://www.datacamp.com/community/tutorials/stemming-lemmatization-python>

|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |          |          |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------|----------|
| <pre> import nltk from nltk.stem import WordNetLemmatizer wordnet_lemmatizer = WordNetLemmatizer()  sentence = "He was running and eating at same time. He has bad habit of swimming after playing long hours in the Sun." punctuations="?!.,;" sentence_words = nltk.word_tokenize(sentence) for word in sentence_words:     if word in punctuations:         sentence_words.remove(word)  sentence_words print("{0:20}{1:20}".format("Word", "Lemma")) for word in sentence_words:     print ("{0:20}{1:20}".format(word, wordnet_lemmatizer.lemmatize(word))) </pre> |          |          |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Word     | Lemma    |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | He       | He       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | was      | wa       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | running  | running  |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | and      | and      |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | eating   | eating   |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | at       | at       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | same     | same     |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | time     | time     |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | He       | He       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | has      | ha       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | bad      | bad      |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | habit    | habit    |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | of       | of       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | swimming | swimming |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | after    | after    |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | playing  | playing  |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | long     | long     |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | hours    | hour     |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | in       | in       |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | the      | the      |
|                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | Sun      | Sun      |

```

sentence_words
print("{0:20}{1:20}".format("Word", "Lemma"))
for word in sentence_words:
    print("{0:20}{1:20}".format(word, wordnet_lemmatizer.lemmatize(word, pos="v")))

```

| Word     | Lemma    |  | Word     | Lemma |
|----------|----------|--|----------|-------|
| He       | He       |  | He       | He    |
| was      | wa       |  | was      | be    |
| running  | running  |  | running  | run   |
| and      | and      |  | and      | and   |
| eating   | eating   |  | eating   | eat   |
| at       | at       |  | at       | at    |
| same     | same     |  | same     | same  |
| time     | time     |  | time     | time  |
| He       | He       |  | He       | He    |
| has      | ha       |  | has      | have  |
| bad      | bad      |  | bad      | bad   |
| habit    | habit    |  | habit    | habit |
| of       | of       |  | of       | of    |
| swimming | swimming |  | swimming | swim  |
| after    | after    |  | after    | after |
| playing  | playing  |  | playing  | play  |
| long     | long     |  | long     | long  |
| hours    | hour     |  | hours    | hours |
| in       | in       |  | in       | in    |
| the      | the      |  | the      | the   |
| Sun      | Sun      |  | Sun      | Sun   |

# Mejorando la búsqueda monolingual

- **Segmentación de compuestos**, en los idiomas aglutinativos, como alemán y holandés, se unen palabras para formar otras más largas. La descomposición de estas palabras en lemas individuales produce una significativa mejora en las búsquedas en este tipo de idiomas al considerar cada elemento de la palabra compuesta como un término.

“wereld” (mundo),  
“bevolking” (población) y  
“conferentie” (conferencia), y  
se traduce como  
“Conferencia sobre la  
población mundial”



# Mejorando la búsqueda monolingual

---

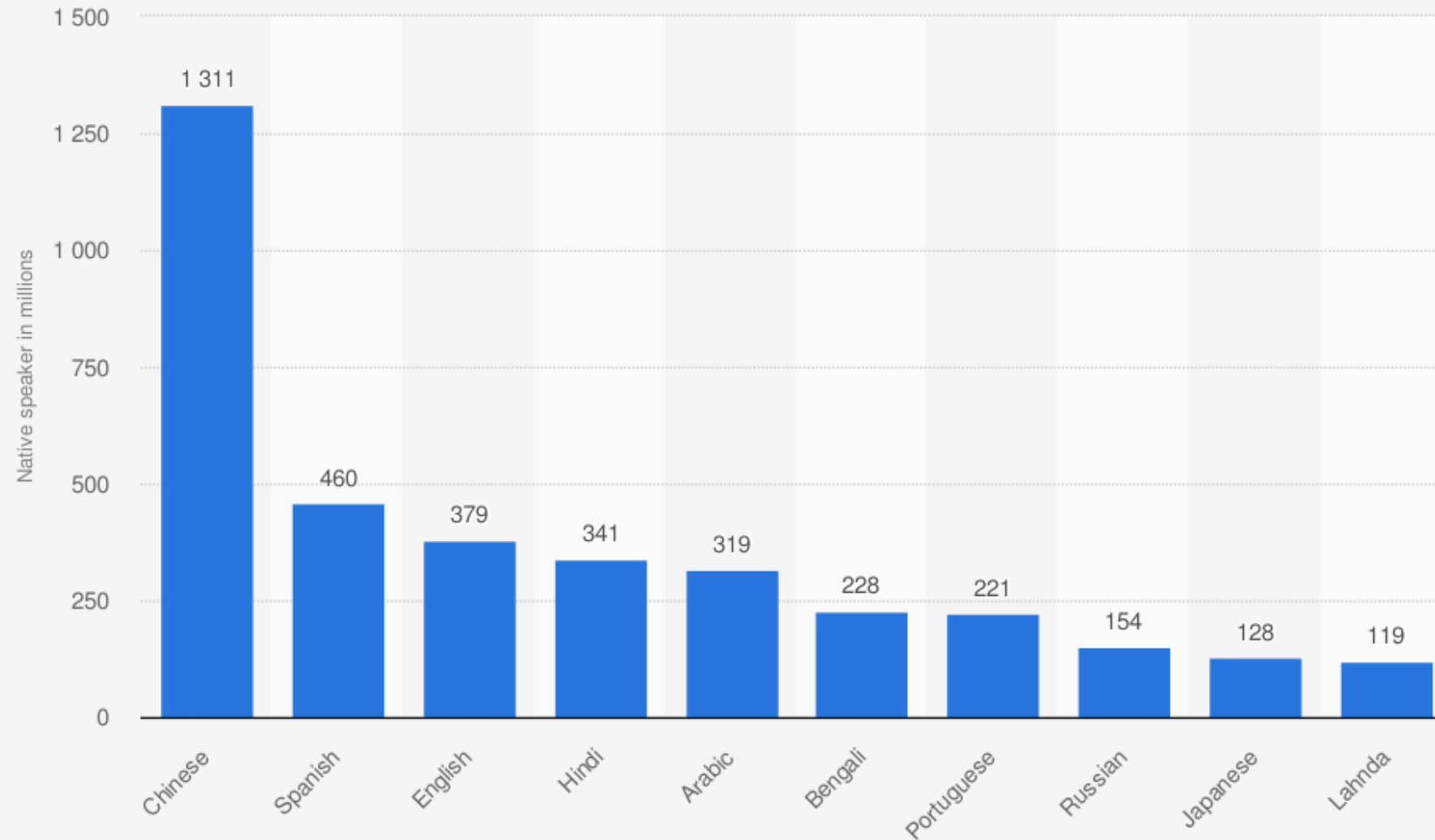
- ***Segmentación de palabras***, en los idiomas asiáticos, como japonés, coreano y chino, los límites de las palabras no se marcan de manera explícita.
- Indexación basada en texto segmentado: que incluye la indexación de palabras y/o de sintagmas.
- Indexación de caracteres: basada en **n-gramas**

# Enfoques basados en la traducción de la consulta

---

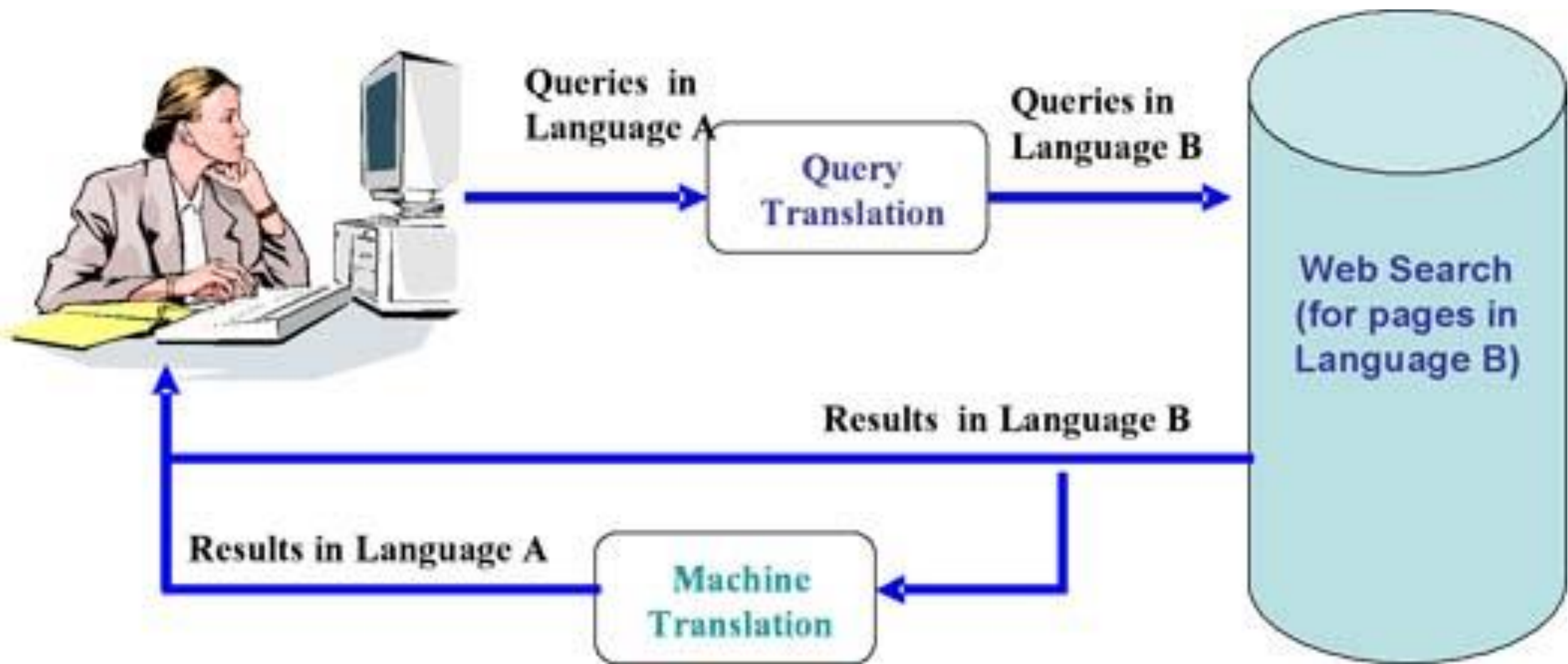
- A la hora de realizar una búsqueda translingüe de información, nos enfrentamos a la siguiente situación: la consulta y los documentos no están escritos en el mismo idioma.
- Es, por tanto, necesario efectuar alguna forma de traducción para poder realizar una búsqueda en la que tanto consulta como documentos se encuentren en el mismo idioma.
- La traducción de la consulta es la opción más frecuente

## The most spoken languages worldwide (native speakers in millions)

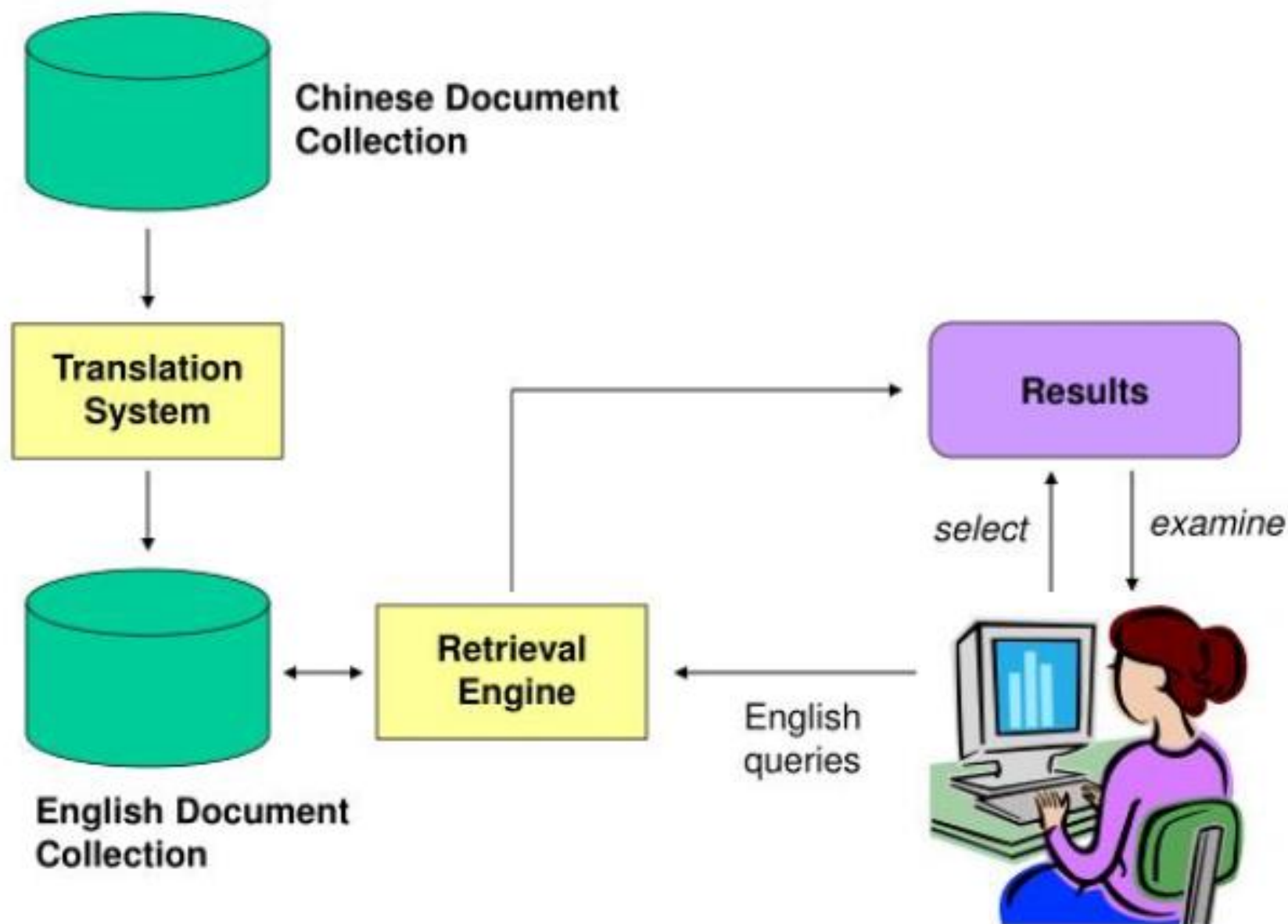


Source  
Ethnologue  
© Statista 2019

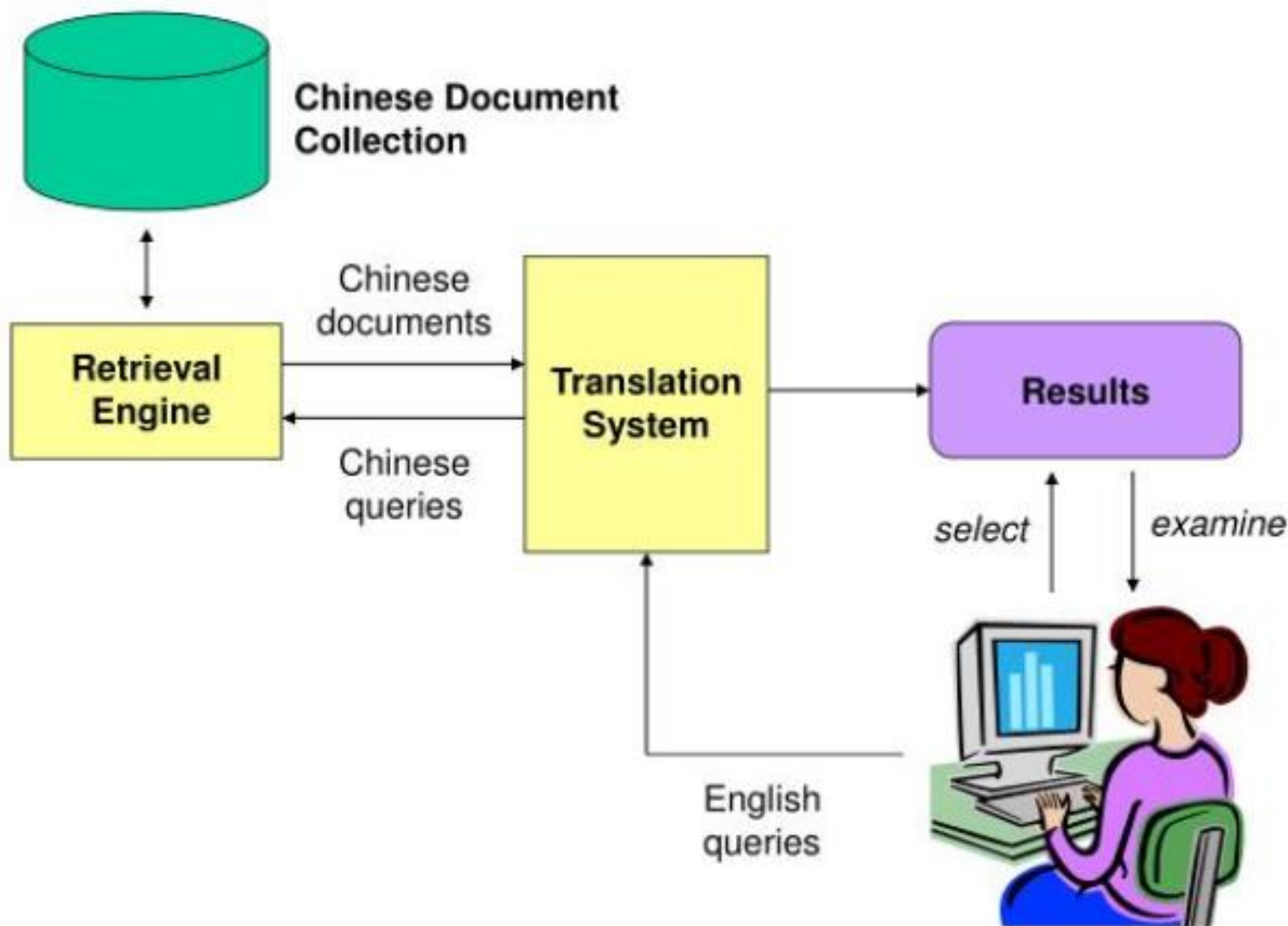
Additional Information:  
Worldwide; 2019



# Query-Language IR



# Document-Language IR



# Diccionarios

---

- La utilización de versiones electrónicas de diccionarios bilingües como recurso de traducción palabra por palabra, ha sido ampliamente estudiada en la literatura. **Problemas:**
- La cobertura del diccionario puede no ser completa, por lo que algunos términos no son traducidos
- No contemplan todas las posibles variantes morfológicas de una palabra.
- En ocasiones es necesario traducir los nombres propios de personas.
- Para cada contexto, sólo algunas traducciones son apropiadas.
- La traducción errónea de los términos es particularmente perjudicial en los conceptos representados por expresiones multipalabra.

# Programas de traducción automática

---

- Este es otro campo de la RI.
- Programas comerciales de traducción automática, siempre que exista uno disponible para el par de idiomas considerados. La traducción de frases cortas es similar a la traducción por palabras, mientras que la traducción de oraciones aportar mejoras notables.
- Si las consultas están formadas por frases, los sistemas de traducción consiguen una traducción mejor que si la consulta está formada por términos independientes sin estructura.
- La creación de estos traductores es costosa, y por eso sólo existen para los pares de idiomas más demandados por el mercado

# Tesauros

---

- Un **tesauro** es una lista de palabras o términos controlados, empleados para representar conceptos.
- Es utilizado en literatura como thesaurus, thesauri o tesoro para referirse a los diccionarios.
- Un tesauro está formado por la colección de términos o palabras clave que se utilizan para realizar la indexación de los documentos, así como las relaciones semánticas que los unen.
- La utilización de tesauros en el campo de la recuperación de información se centra en el enriquecimiento de la consulta con términos relacionados que aparecen realmente en los documentos

# Tesauros

---

- Proporcionan un vocabulario controlado.
- Permiten dar una mejor estructuración a los resultados
- Su estructuración jerárquica hacen posible su utilización en un entorno de búsqueda interactivo
- Un tesoro multilingüe sobre un dominio determinado permite la traducción de términos específicos de ese dominio que quizá no puedan encontrarse en un diccionario bilingüe.

# Tesoro de la UNESCO

Lengua del contenido

español ▾



Buscar

Alfabéticamente

Jerarquía

Grupos

A Á B C D E É F G H I J K L M  
N O Ó P Q R S T U V W X Y Z

Actividad de los museos

Actividad de ocio → **Actividad de tiempo libre**

Actividad de tiempo libre

Actividad fuera de programa

Actividad juvenil

Actividad paraescolar → Actividad fuera de programa

Actividad religiosa

Actividad sensomotriz

Actividad sociocultural → Actividad cultural

Actividad solar

Actividad sísmica → Sismicidad

Actor

Actualidades cinematográficas

Actualización de los conocimientos

Acuerdo bilateral

Acuerdo contractual → Derecho de los contratos

Acuerdo cultural

Acuerdo internacional → Instrumento internacional

Acuerdos → Derecho de los contratos

Acuerdos sobre mercancías

Acuicultura

Acuicultura marina

Aculturación

Acumulación calorífica → Acumulación de calor

Acumulación de calor

Acumulación térmica → Acumulación de calor

Acupuntura

Acústica

Adaptación al cambio climático

Adaptación animal → Ecología animal

Adaptación biológica

Adaptación de las plantas → Fitoecología

TÉRMINO PREFERIDO

## Actividad de tiempo libre

Búsqueda en UNESDOC

CONCEPTOS ESPECÍFICOS

Club

Deporte

Entretenimiento

Juego

Manifestaciones culturales

CONCEPTOS RELACIONADOS

Actividad cultural

Actividad fuera de programa

Comportamiento cultural

Distribución del tiempo

Instalación recreativa

Ocio

Sociología del tiempo libre

Turismo

ETIQUETA ALTERNATIVA

Actividad de ocio

PERTENECE AL GRUPO

Cultura > Ocio

EN OTRAS LENGUAS

Activité de loisir

francés

Leisure time activities

inglés

Занятия во время досуга

ruso

URI

<http://vocabularies.unesco.org/thesaurus/concept368>

Descargue este concepto:

[RDF/XML](#) [TURTLE](#) [JSON-LD](#)

última modificación 23/5/06

# Sistemas hipertextuales y recuperación documental

---

# Hipertexto

---

- El **hipertexto** es una estructura no secuencial que permite crear, agregar, enlazar y compartir información de diversas fuentes por medio de enlaces asociativos y redes sociales. El hipertexto es texto que contiene enlaces a otros textos. El término fue acuñado por [Ted Nelson](#) alrededor de 1965.
- La forma más habitual de hipertexto en [informática](#) es la de [hipervínculos](#) o referencias cruzadas automáticas que van a otros documentos ([lexías](#)). Si el [usuario](#) selecciona un hipervínculo, el [programa](#) muestra el documento enlazado.
- El hipertexto no está limitado a datos textuales, se pueden encontrar dibujos del elemento especificado o especializado, sonido o vídeo referido al tema. El programa que se usa para leer los documentos de hipertexto se llama [navegador](#), *browser*, visualizador o cliente, y cuando el lector o usuario sigue un enlace, se dice que está navegando por la [web](#).

# La Web como un grafo

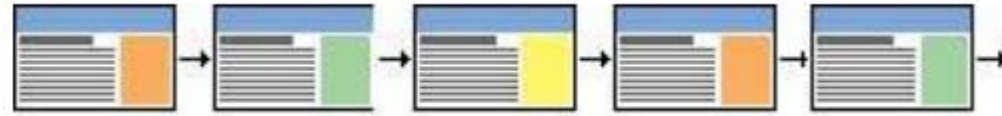
---

- La web puede verse como un enorme grafo dirigido.
- Las porciones pequeñas de la web (un sitio web, un dominio) siguen el mismo principio.
- Cada página web es un nodo, y un link (enlace) entre dos páginas representa una relación entre ambas (relación de entrada o salida)

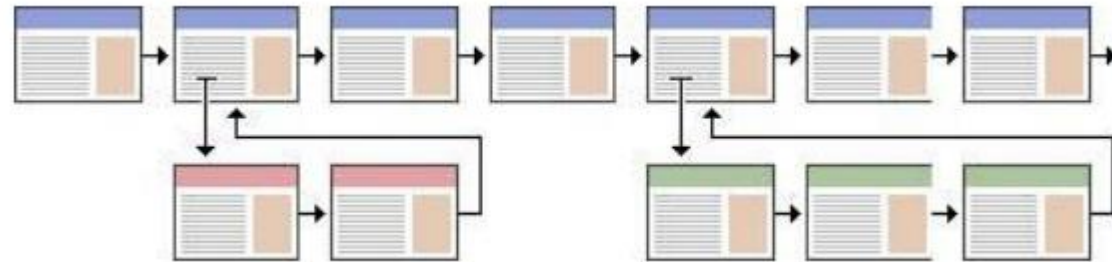
# Estructuras hipertextuales

---

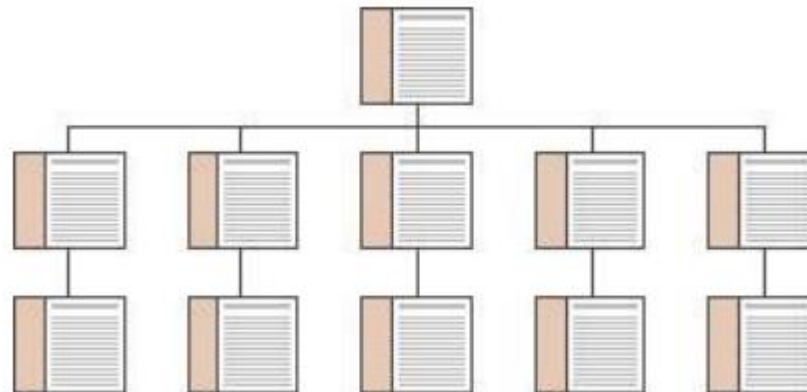
- Estructura secuencial



- Estructura secuencial con desviaciones



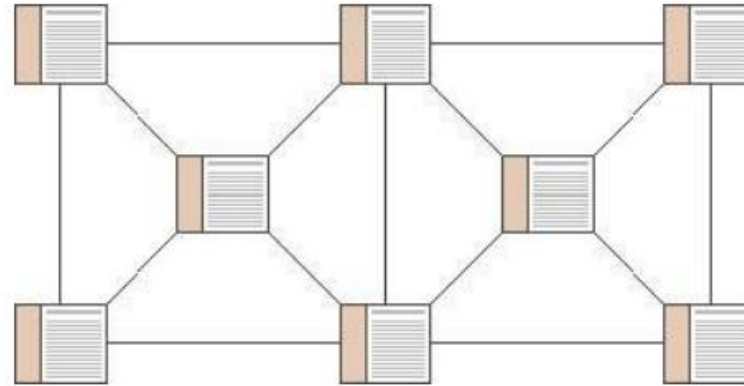
- Estructura jerárquica



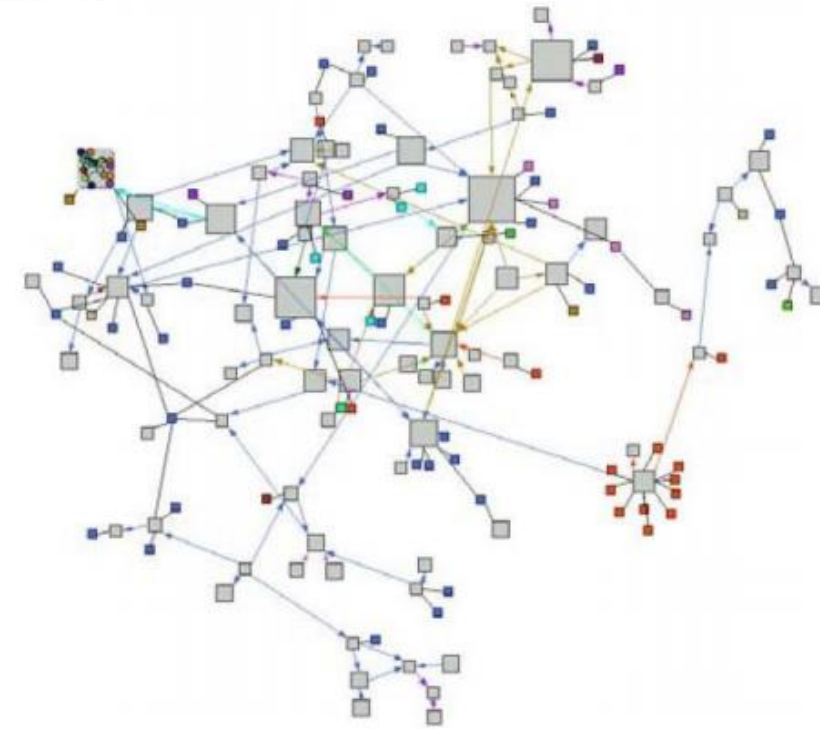
# Estructuras hipertextuales

---

- Estructura en red



- Un método –utilizado habitualmente– consiste en desarrollar un hipertexto con un recorrido “intuitivo” no secuencial. En este tipo de “estructura” tarde o temprano surge el tema de los **laberintos hipertextuales** como el anti-hipertexto didáctico



# Recuperación Web

---

- Inmensa cantidad de información: ¿Cómo gestionarla? ¿Cómo asegurar una cobertura total ?
- Dificultad de encontrar lo que se busca
- Imposible acceder directamente a todas las páginas/sitios relacionados
- Es necesario utilizar diferentes herramientas de búsqueda de recursos
- Hay una Internet visible (accesible desde buscadores web y directorios temáticos) pero la mayor parte de la información constituye la Internet invisible (no accesible mediante estas herramientas)
- Las páginas web no se indizan con un vocabulario estándar a diferencia de los catálogos de bibliotecas o de los índices de artículos de revistas
- Escasez de datos de identificación de los documentos: Los autores de los documentos no están identificados, las fechas de publicación o no existen o no son exactas...

# Recuperación Web

---

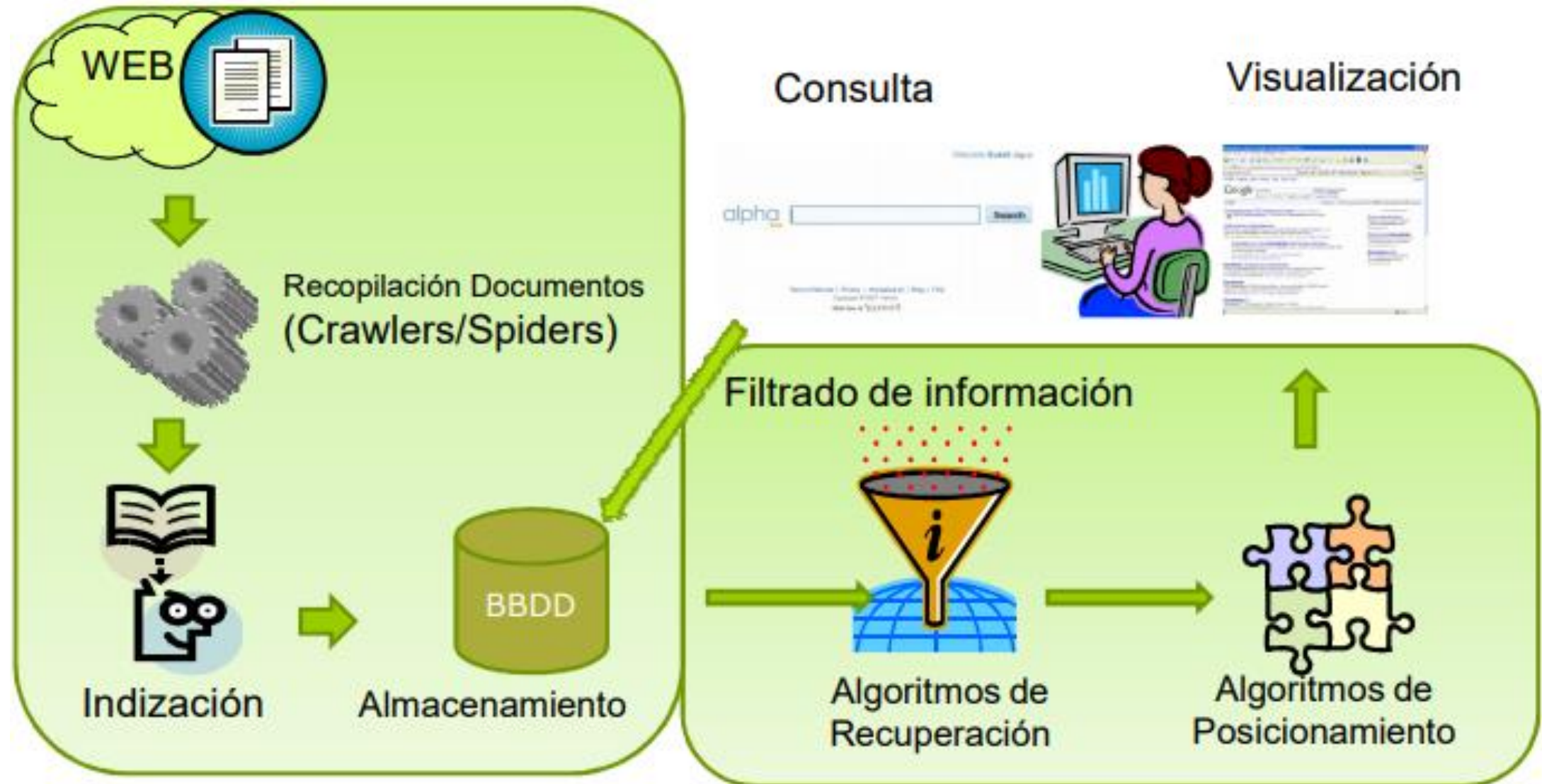
- Información hipermedia: Documentos multimedia e hipertextuales
  - Cualquier persona puede publicar una página web sobre cualquier cosa:
    - La calidad de la información no está garantizada
    - Pocas páginas tienen “crítica editorial” o peer reviews
    - A veces se obtiene información maliciosa o equivocada
  - Información dinámica/inestable: Los sitios web aparecen y desaparecen
- Problemas éticos y legales: censura, propiedad intelectual

# Recuperación Web

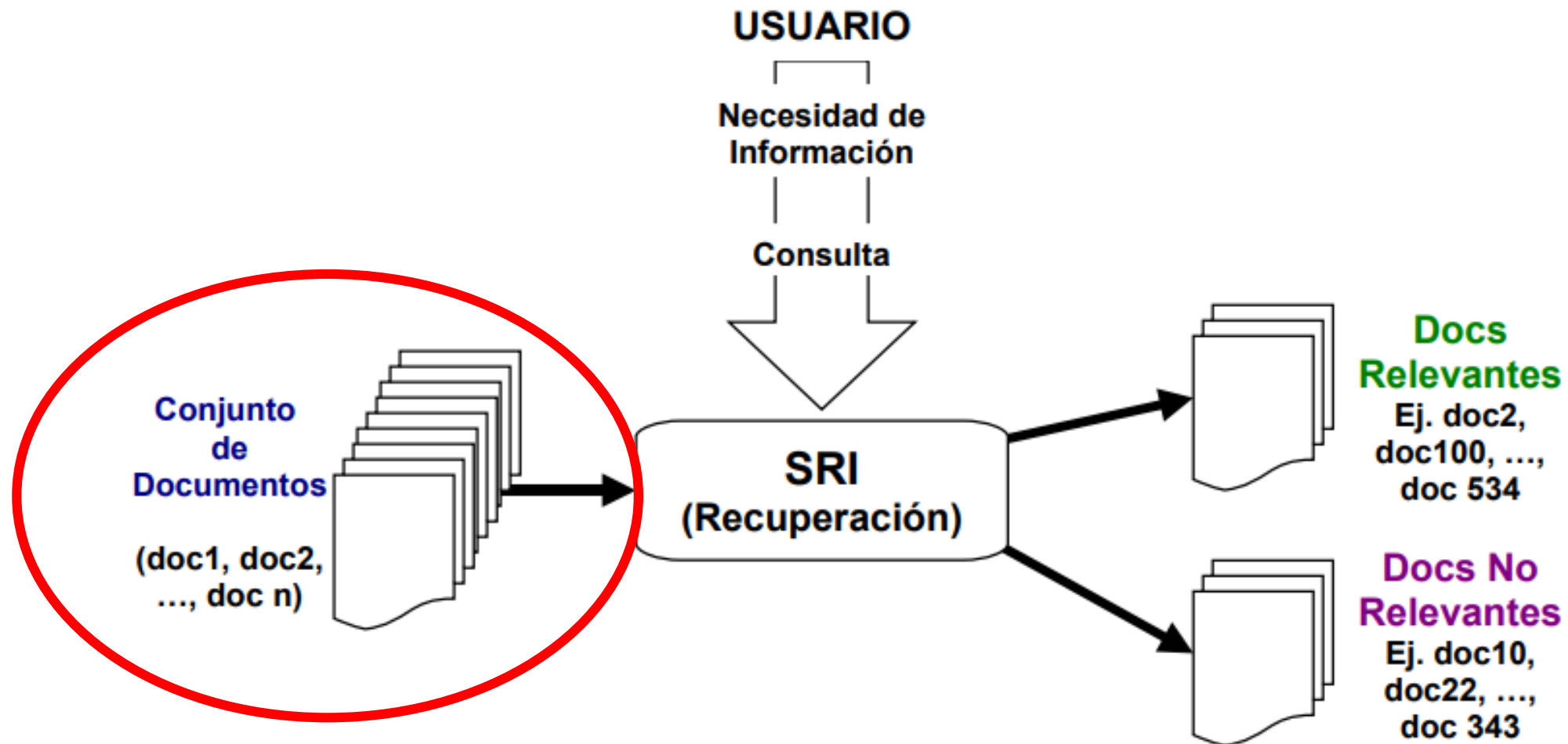
---

- Información hipermedia: Documentos multimedia e hipertextuales
  - Cualquier persona puede publicar una página web sobre cualquier cosa:
    - La calidad de la información no está garantizada
    - Pocas páginas tienen “crítica editorial” o peer reviews
    - A veces se obtiene información maliciosa o equivocada
  - Información dinámica/inestable: Los sitios web aparecen y desaparecen
- Problemas éticos y legales: censura, propiedad intelectual

# Motores de búsqueda



# Sistema de Recuperación de Información



# Recopilación de documentos

---

- *Spider, crawlers, robots* o Agentes de Búsqueda son los nombres que recibe el software que recopila los documentos.
- Funcionamiento
  - ☐ Comienza en una página (A) y recopila todas sus URL
  - ☐ Envía la página (A), comprueba que no está indizada y que no se tiene una versión menos actualizada, indiza la página (A)
  - ☐ Recupera la página (B) que está primera en la lista
  - ☐ Envía la página (B)...

La recuperación de documentos multimedia o no textuales.

---

# Multimedia

---

- En cuanto al concepto “multimedia” (Multi – media = Muchos Medios) etimológicamente su significado es claro, pero, en la práctica, este vocablo no es tan sencillo de determinar y depende del punto de partida en que nos situemos.
- Desde la perspectiva informática, multimedia es la **integración de diferentes tipos de medios en un solo soporte**: pueden componerse de texto, gráficos, sonido digitalizado, vídeo y otros tipos de información, contenidas en soporte informático.

# Reconocimiento de Imágenes



(a)



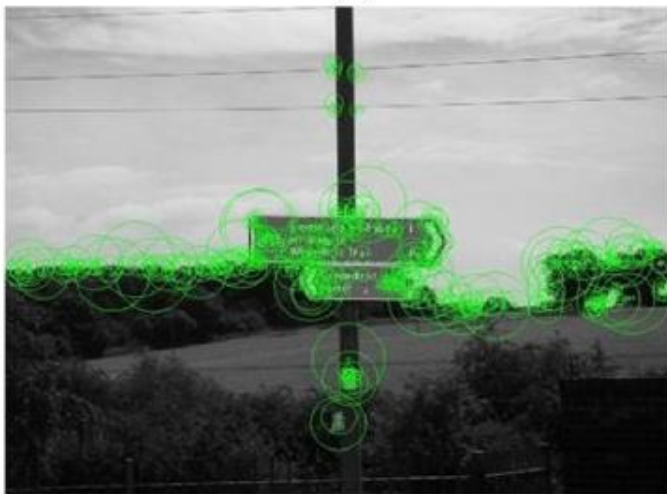
(b)



(e)



(c)



(d)

- Extracción de características
- Clasificación
- Agrupación
  - K-means
  - DBSCAN

# Reconocimiento de Texto en Imágenes

---



(a)



(b)



(c)



(d)

# Otro esquema

---

- El modelo presentado se basa en la combinación de evidencia del contenido del texto y la estructura de hipertexto de la Web.
- El empleo de metadatos, datos de los datos...