

*Mercedes:*  
**detección de términos en textos del CT\_IULA**

Jorge Vivaldi Palatresi  
Raúl Araya Tauler

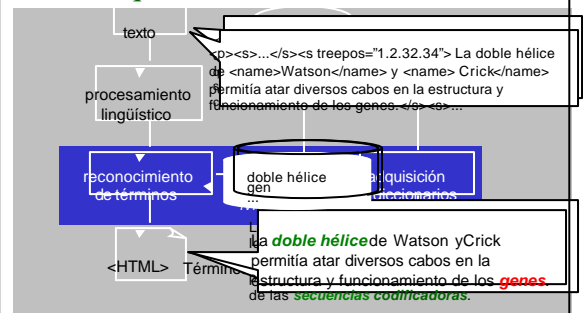
## Extracción vs Detección de términos

- Extracción: buscar términos en un texto (**yate**).
- Detección: encontrar términos (previamente validados como tales) en un texto (**mercedes**).

## Mercedes

- Programa para la detección de términos ya validados.
- Módulos para la:
  - detección de términos
  - gestión de diccionarios
- Se puede aplicar a varios dominios
- Se puede utilizar en textos del CT en castellano, catalán y inglés

## Mercedes : esquema de funcionamiento



## Limitaciones

- No están todos:
  - Sólo marca como término los que están incluidos en los diccionarios. Ej.: *DNA* vs *ADN*,
  - No se trata la coordinación. Ej. *dret constitucional i estatutari*, ...
- No todos son términos
  - Polisemia
  - Ej. (genoma): expresión, tipo, base,... (medicina): virus, ... (derecho): anular, sala, zona, ...

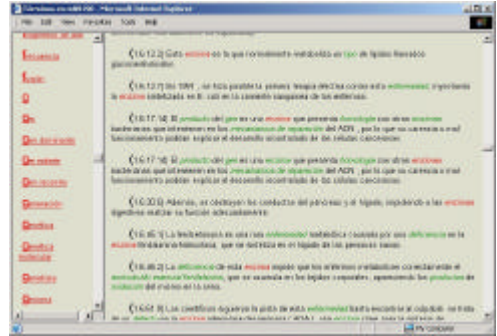
## Limitaciones: no están todos

- La organización de l **genoma mitocondrial** humano , es radicalmente diferente de l **genoma nuclear** , pero tiene grandes **similitudes** con la mayoría de los **genomas** de las **bacterias** ( **celulas** procariotas ): es más simple , está constituido por unos dieciséis mil seiscientos **pares de bases** , conteniendo 37 **genes** y con una disposición circular .
- Se cree que la **célula** eucarótica actual, conteniendo ambos **genomas** nuclear y mitocondrial, procede de la **simbiosis** entre dos **células** diferentes, una nucleada eucariota) y otra sin **núcleo** diferenciado( procariota).

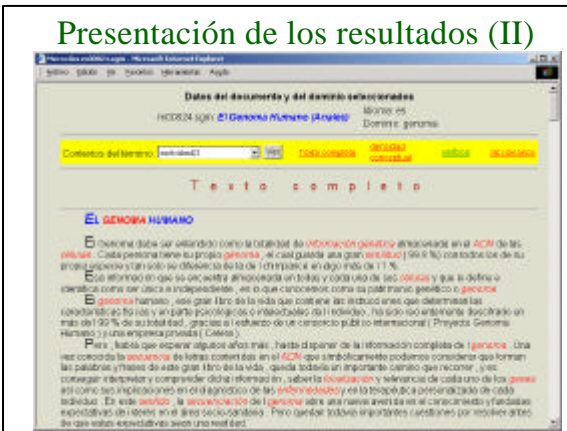
## Limitaciones: no todos son términos

- Franklin había deducido, mediante cálculos precisos, que las **bases nitrogenadas** que entraban a formar parte de la **composición del ADN** debían estar hacia adentro de una **estructura helicoidal**, con el espinazo de azúcar-fosfato en su exterior.
- Estas **enzimas** constituyen la **base** de la **ingeniería genética** que sería impensable sin ellas.
- El hecho de que cada **base nitrogenada** se aparee específicamente con otra **base** de la **cadena** opuesta explica satisfactoriamente la **replicación del ADN**.

## Presentación de los resultados (I)



## Presentación de los resultados (II)



## Tratamiento de términos incrustados

- Una misma secuencia de palabras incluye mas de un término.
- Ej. “ADN” y “ADN recombinante”
- Mercedes reconoce las dos cadenas como dos términos diferentes
- Cuando se trata de un contexto, únicamente señala el mas largo



## Otros sistemas similares

- Harkema et al. (2004). *A large-Scale Resource for Storing and Recognizing Technical Terminology*. LREC2004
- Término: “*dictionary-based*” *term recognition*
- Características:
  - Capaz de almacenar gran número de términos
  - A cada término, asocia diferentes informaciones :
    - Morfosintáctica (categoría y clase morfológica)
    - Semántica (forma lógica y conceptos de la ontología)
    - Origen
  - Rapidez y eficiencia en el reconocimiento
  - Precisión: ~60%

## Mercedes: especificaciones

- Texto de entrada: textos del CT -IULA (ES, CA, EN)
- Sólo detecta términos nominales
- Dominio: Genoma
- Dicionarios de referencia en Genoma: 13
- | Idioma  | Nro. entradas |
|---------|---------------|
| Catalán | 1051          |
| Español | 5950          |
| Inglés  | 5847          |
- Visualización de los resultados en páginas HTML
- Otros resultados:
  - Verbos indicadores de relación en contexto
  - Distribución de términos en un texto
  - Cálculo de la densidad terminológica (tentativo)

## Mercedes: requisitos de ejecución

- Acceso al servidor Mordred (k:\Utils) o bien tener la aplicación en forma de ejecutable
- Acceso a un intérprete Perl
- Acceso a Internet
- Entorno MS-Windows
- Obtener los ficheros del documento a procesar
- Indexación
- Ejecución de Mercedes
  
- Manual de utilización en  
J:\USUARIS \Iulaterm\Projectes\Mercedes\Mercedes.doc

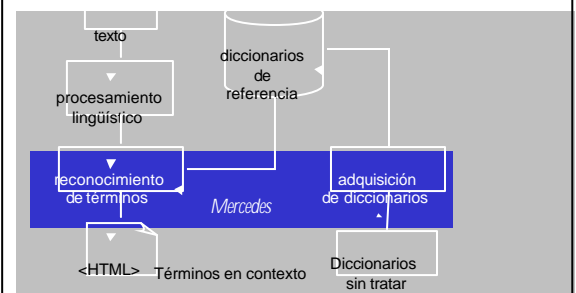
## ¿Qué es el módulo de diccionarios?

- Es la parte de Mercedes que se encarga de la gestión y mantenimiento de los diccionarios.
- Es importante tener en cuenta que un “diccionario” en Mercedes puede estar constituido por varios fragmentos de otros diccionarios.

## ¿Cómo funciona?

- Servidor de bases de datos MySQL.
- Gestión de la BD: scripts en Perl.
- Archivos en formato de texto.
- Archivos en formato XML.

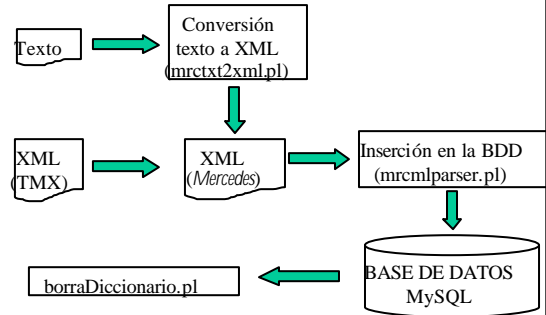
## Modalidades de inserción de diccionarios



## Programas de inserción de diccionarios

- mrcml2txt.pl: transformación de texto a XML.
- mrcmlparser.pl: comprobación e incorporación de datos a la bdd desde XML.
- borraDiccionarios.pl: eliminación de diccionarios de la base de datos.
- Transformación mediante XSLT.

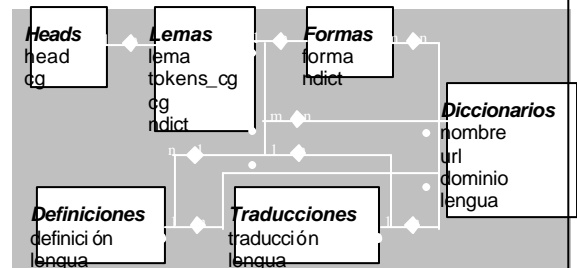
## Esquema de funcionamiento



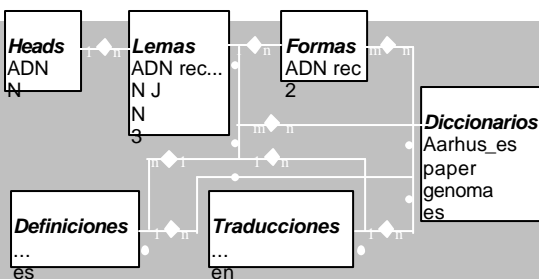
## La base de datos

- Diccionarios:
  - Nombre
  - Origen
  - Dominio
  - Lengua
- Entradas:
  - Forma
  - Lema (pendiente morfológico).
  - Cat. gramatical.
    - Opcionales:
      - Definición
      - Traducción

## Esquema de la base de datos



## Esquema de la base de datos: ejemplo



## Formatos de los diccionarios

- Texto plano:
  - Archivo de texto delimitado por caracteres.
  - Se puede crear manualmente.
  - Se puede obtener desde MS Access o MS Excel o similares.
- Formato XML.
  - Archivo de texto marcado con etiquetas.
  - Sigue una DTD propia.
  - Se puede generar con un programa a partir de los archivos de texto u otras fuentes.

## Formato de texto:

- Cabecera:
  - Nombre.
  - Origen.
  - URL base.
  - Dominio.
  - Lengua.
  - Lista de campos codificados.

## Formato de texto:

- Las entradas:
  - Término.
  - Lema.
  - Categoría gramatical.

## Ejemplo de formato de texto:

```
#nombre: Diccionario de pruebas FOLDOC.  
#origen: http://wombat.doc.ic.ac.uk/foldoc/index.html  
#urlbase: http://wombat.doc.ic.ac.uk/foldoc/  
#dominio: informática  
#lengua: en  
#"TERM"; "LEMA"; "CG"  
"bit"; "bit"; "N"  
"mouse"; " mouse"; "N"  
"PHP"; "PHP"; "N"
```

## Formato XML.

- Cabecera:
  - Nombre.
  - Origen.
  - URL base.
  - Dominio.
  - Lengua.

## Formato XML

- Entradas
  - Término.
  - Lema.
  - Categoría gramatical.

## Ejemplo de formato XML:

```
<mrc-dictionary domain="dret"  
  lang="ca"  
  name="Subdiccionari de dret extret del diec"  
  url="paper">  
<entry>  
  <term term-cg="N">abanderament</term>  
  <term-lemma tokens-cg="N">abanderament  
    </term-lemma>  
  <definitions>  
    <def lang="ca">Acció d'abanderar.</def>  
  </definitions>  
</entry>
```

## De los archivos a la base de datos:

- Texto:
  - mrcxt2xml.pl: produce un archivo en formato XML conforme a la DTD de Mercedes.
- XML:
  - mrcmlparser.pl: comprueba el archivo en formato XML e inserta los datos en la base de datos.

## Eliminar un diccionario:

- borraDiccionario.pl: nos permite borrar diccionarios. Lo puede hacer:
  - Uno a uno.
  - Por pares dominio-lengua.

## Mercedes: posibles ampliaciones

- Discriminación entre diccionarios de un mismo dominio
- Añadir información semántica
- Cálculo de la densidad a lo largo del documento
- Cálculo real de precisión y cobertura
- ...
  
- Otros usos de los diccionarios:
  - Detección de términos, ...