Digital Libraries on Internet: design and exploitation

Rafael C. Carrasco (carrasco@dlsi.ua.es)
Director Adjunto de la Biblioteca Virtual Miguel de Cervantes

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

http://www.cervantesvirtual.com

Contents

1. The Biblioteca Virtual Miguel de Cervantes
2. Text transcription, structural markup and accessibility
3. Tools for internet usage
4. Technological challenges

1. The Biblioteca Virtual Miguel de Cervantes

Background

The Miguel de Cervantes library was created in July 1999 as a result of the cooperation between the Universidad de Alicante and the Banco Santander.

Grupo Santander

Universitat d'Alacant
Universidad de Alicante

Background

The Miguel de Cervantes library was created in July 1999 as a result of the cooperation between the Universidad de Alicante and the Banco Santander.

The library is now supported by

Partners

Universitat d'Alacant
Universidad de Alicante
Fundación Marcelino Botín
Grupo Santander
MINISTERIO DE TRABAJO Y ASUNTOS SOCIALES
MINISTERIO DE CULTURA
MINISTERIO DE EDUCACIÓN Y CIENCIA
GENERALITAT VALENCIANA
C·R·U·E CONFERENCIA DE RECTORES DE LAS UNIVERSIDADES ESPAÑOLAS
AGENCIA ESPAÑOLA DE COOPERACIÓN INTERNACIONAL
MINISTERIO DE ASUNTOS EXTERIORES Y DE COOPERACIÓN
REAL ACADEMIA ESPAÑOLA
Instituto Cervantes
Fundación Germán Sánchez Ruipérez
universia.es
Grupo PRISA
FUNDACIÓN Telefónica
FEDERACIÓN DE GREMIOS DE EDITORES DE ESPAÑA
REPSOL YPF

Objectives

The original objective is to foster Spanish and American cultural content in Internet.

Objectives

The original objective is to foster Spanish and American cultural content in Internet.

1. Documents are selected according to their scientific and cultural value.

Objectives

The original objective is to foster Spanish and American cultural content in Internet.

1. Documents are selected according to their scientific and cultural value.
2. Most texts are transcribed and supervised by linguists.

Objectives

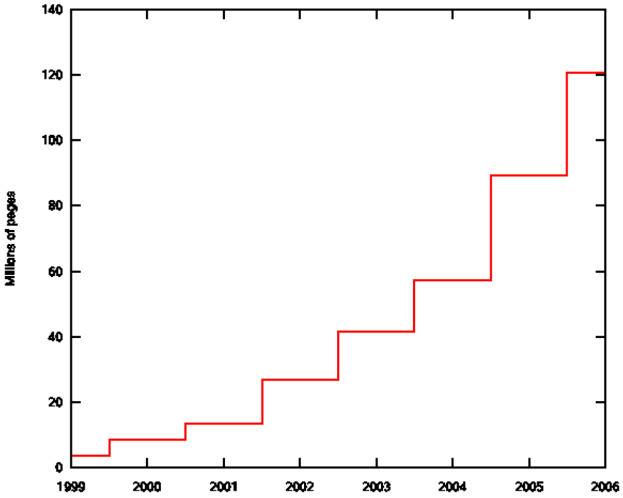
The original objective is to foster Spanish and American cultural content in Internet.

1. Documents are selected according to their scientific and cultural value.
2. Most texts are transcribed and supervised by linguists.
3. Texts include structural *metainformation*.

Impact

Pages server successfully (per year)	121.168.009
Average number of pages (per day)	331.967
Average transfer (per day)	28,6 GB

Impact


Millions of pages
0
20
40
60
80
100
120
140
1999
2000
2001
2002
2003
2004
2005
2006

Top 10 domains

- 🇪🇸 .es (38 %)

Top 10 domains

- 🇪🇸 .es (38 %)
- 🇺🇸 .us (17 %)

Top 10 domains

- 🇪🇸 .es (38 %)
- 🇺🇸 .us (17 %)
- 🇲🇽 .mx (10 %)

Top 10 domains

- 🇪🇸 .es (38 %)
- 🇺🇸 .us (17 %)
- 🇲🇽 .mx (10 %)
- 🇦🇷 .ar (5 %)

Top 10 domains

- .es (38 %)
- .us (17 %)
- .mx (10 %)
- .ar (5 %)
- .cl (4 %)

Top 10 domains

- .es (38 %)
- .us (17 %)
- .mx (10 %)
- .ar (5 %)
- .cl (4 %)
- .pe (4 %)

Top 10 domains

- .es (38 %)
- .us (17 %)
- .mx (10 %)
- .ar (5 %)
- .cl (4 %)
- .pe (4 %)
- .co (2 %)

Top 10 domains

- .es (38 %)
- .us (17 %)
- .mx (10 %)
- .ar (5 %)
- .cl (4 %)
- .pe (4 %)
- .co (2 %)
- .ve (2 %)

Top 10 domains

- .es (38 %)
- .us (17 %)
- .mx (10 %)
- .ar (5 %)
- .cl (4 %)
- .pe (4 %)
- .co (2 %)
- .ve (2 %)
- .fr (1 %)

Top 10 domains

- .es (38 %)
- .us (17 %)
- .mx (10 %)
- .ar (5 %)
- .cl (4 %)
- .pe (4 %)
- .co (2 %)
- .ve (2 %)
- .fr (1 %)
- .br (1 %)

Top 10 books

¿Which are the most accessed books?

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).

The Biblioteca Virtual Miguel de Cervantes

Text transcription, structural markup and accessibility

Tools for internet usage

Technological challenges

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).
3. La Celestina.

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).
3. La Celestina.
4. Cantar del Mío Cid.

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).
3. La Celestina.
4. Cantar del Mío Cid.
5. El Periquillo Sarniento (Fernández de Lizardi).

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).
3. La Celestina.
4. Cantar del Mío Cid.
5. El Periquillo Sarniento (Fernández de Lizardi).
6. La vida es sueño (Pedro Calderón de la Barca.)

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).
3. La Celestina.
4. Cantar del Mío Cid.
5. El Periquillo Sarniento (Fernández de Lizardi).
6. La vida es sueño (Pedro Calderón de la Barca.)
7. La vida de Lazarillo de Tormes.

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).
3. La Celestina.
4. Cantar del Mío Cid.
5. El Periquillo Sarniento (Fernández de Lizardi).
6. La vida es sueño (Pedro Calderón de la Barca.)
7. La vida de Lazarillo de Tormes.
8. La Ética de Aristóteles.

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).
3. La Celestina.
4. Cantar del Mío Cid.
5. El Periquillo Sarniento (Fernández de Lizardi).
6. La vida es sueño (Pedro Calderón de la Barca.)
7. La vida de Lazarillo de Tormes.
8. La Ética de Aristóteles.
9. Gramática de la Lengua Castellana (RAE)

Top 10 books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
2. Diccionario del español usual en México (C. de México).
3. La Celestina.
4. Cantar del Mío Cid.
5. El Periquillo Sarniento (Fernández de Lizardi).
6. La vida es sueño (Pedro Calderón de la Barca.)
7. La vida de Lazarillo de Tormes.
8. La Ética de Aristóteles.
9. Gramática de la Lengua Castellana (RAE) .
10. Canto general (Pablo Neruda).

Browsing

The library is organized by themes and authors.

Browsing

The library is organized by themes and authors.

- Some authors:

  Leopoldo Alas, *Miguel de Cervantes*,
  Benito Pérez Galdós,
  Mario Benedetti,
  Calderón de la Barca,
  Alfredo Bryce Echenique,
  Antonio Buero Vallejo. . .

Browsing

The library is organized by themes and authors.

- Some authors:

  Leopoldo Alas, *Miguel de Cervantes*,
  Benito Pérez Galdós,
  Mario Benedetti,
  Calderón de la Barca,
  Alfredo Bryce Echenique,
  Antonio Buero Vallejo. . .

- Themes:

  Fundação Biblioteca Nacional (Brasil), *Biblioteca Nacional de Argentina*, *Biblioteca Nacional de Chile*, *Portal de Venezuela*, *Literatura Gauchesca*, *poesía española contemporánea*. . .

Additional content

1. History.

Additional content

1. History.
2. Signs.

Additional content

1. History.
2. Signs.
3. Video and sound.

Additional content

1. History.
2. Signs.
3. Video and sound.
4. Journals.

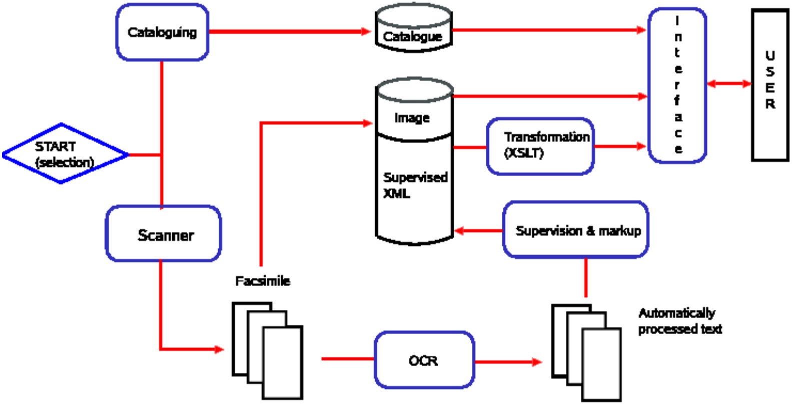
Additional content

1. History.
2. Signs.
3. Video and sound.
4. Journals.

Map

2. Text transcription, structural markup and accessibility

Production workflow



Text transcription and accessibility

Most books are transcribed and supervised: this is a costly process but improves accessibility and usability. Multimodal access is under development and W3C recommendations are followed

(http://www.w3.org/WAI)

Text transcription and accessibility

El sombrero del cura (Clarín)

> *El señor obispo de la diócesis, por razones muy dignas de respeto, prohibió, hace algunos años, que el clero rural anduviera por prados y callejas, costas y montañas, luciendo el levitón de anchos faldones y el sombrero de copa alta, demasiado alta muchas veces. Hoy todos los curas de mi verde Erín, de mi católica y pintoresca Asturias, usan traje talar, sombrero de teja, de alas sueltas y cortas; y, a fuerza de humildad y con prodigios de obediencia, consiguen montar a caballo con sotana o balandrán, sin hacer la triste figura y sortear las espinas de los setos, sin dejar entre las zarzas jirones del paño negro.*

Text transcription and accessibility

SE EQUIVOCÓ LA PALOMA

Se equivocó la paloma. Se equivocaba.
Por ir al Norte, fue al Sur. Creyó que el trigo era agua. Se equivocaba.
Creyó que el mar era el cielo; que la noche la mañana. Se equivocaba.
Que las estrellas eran rocío; que la calor, la nevada. Se equivocaba.
Que tu falda era tu blusa; que tu corazón su casa. Se equivocaba.
(Ella se durmió en la orilla. Tú, en la cumbre de una rama.)
Rafael Alberti.

Accessibility

1. Enhances impact (easier for search engines, faster downloads, multilingualism, clearer design...).

Accessibility

1. Enhances impact (easier for search engines, faster downloads, multilingualism, clearer design...).
2. Simplifies procedures (e.g. updates).

Accessibility

1. Enhances impact (easier for search engines, faster downloads, multilingualism, clearer design...).
2. Simplifies procedures (e.g. updates).
3. Improves efficiency (e.g., server load).

Improving usability

User tests became an important source of information:

Improving usability

User tests became an important source of information:

- User does not remember the address.

Improving usability

User tests became an important source of information:

- User does not remember the address.
  The header included icons to bookmark or make home page.

Improving usability

User tests became an important source of information:

- User does not remember the address.
  The header included icons to bookmark or make home page.
- Users do not know where to start.

Improving usability

User tests became an important source of information:

- User does not remember the address.
  The header included icons to bookmark or make home page.
- Users do not know where to start.
  A guided tour was created.

Improving usability

User tests became an important source of information:

- User does not remember the address.
  The header included icons to bookmark or make home page.
- Users do not know where to start.
  A guided tour was created.
- In some sections, it is unclear what the content is.

Improving usability

User tests became an important source of information:

- User does not remember the address.
  The header included icons to bookmark or make home page.
- Users do not know where to start.
  A guided tour was created.
- In some sections, it is unclear what the content is.
  Descriptive texts were added.

Improving usability

- Some works were external links.

Improving usability

- Some works were external links.
  External content is open on a separate window.

Improving usability

- Some works were external links.
  External content is open on a separate window.
- User quits when too much text is presented.

Improving usability

- Some works were external links.
  External content is open on a separate window.
- User quits when too much text is presented.
  Texts are abbreviated.

Improving usability

- Some works were external links.
  External content is open on a separate window.
- User quits when too much text is presented.
  Texts are abbreviated.
- Some pages need long download times

Improving usability

- Some works were external links.
  External content is open on a separate window.
- User quits when too much text is presented.
  Texts are abbreviated.
- Some pages need long download times
  A maximal size is established.

Improving usability

- Some works were external links.
  External content is open on a separate window.
- User quits when too much text is presented.
  Texts are abbreviated.
- Some pages need long download times
  A maximal size is established.
- Too many decorative elements.

Improving usability

- Some works were external links.
  External content is open on a separate window.
- User quits when too much text is presented.
  Texts are abbreviated.
- Some pages need long download times
  A maximal size is established.
- Too many decorative elements.
  The design was simplified.

Metainformation

Metainformation is information about information, that is, about contents:

155.2

NUTes Nutlin, Joseph

La estructura de la personalidad / Joseph Nutlin.
Buenos Aires : Kapelusz, 1973.

237 p. : il.- - (Biblioteca de Psicología Contemporánea ; 27)

1.- PSICOLOGIA 2.- PERSONALIDAD

Descriptive metadata: Dublin Core

Objects include descriptive metainformation according to the Dublin Core standard. (http://dublincore.org)

```
<meta name="DC.title"
      content="El ingenioso hidalgo Don Quixote de la Mancha"/>
<meta name="DC.creator"
      content="Cervantes Saavedra, Miguel de"/>
<meta name="DC.creator"
      content="Cuesta, Juan de la"/>
<meta name="DC.subject"
      content="Novela espa&ntilde;ola - Siglo 17"/>
```

Structural metadata: TEI

- Texts also include structural metainformation: the Text Encoding Initiative provides a vocabulary for literature (http://www.tei-c.org).
- TEI is used also at PERSEUS, Oxford Text Archives, Women Writers Project and many others.

Structural metadata: TEI

```
<title> Doña Perfecta </title>
<author> Benito Pérez Galdós </author>
<text>
   ....
 <note>
   Doña Perfecta, como mujer de su época, ....
 </note>
</text>
```

Structural metadata: TEI

1. Text editing and formating become differentiated tasks.

Structural metadata: TEI

1. Text editing and formating become differentiated tasks.
2. Context sensitive searches are possible: e.g., Sevilla as caption.

Structural metadata: TEI

1. Text editing and formating become differentiated tasks.
2. Context sensitive searches are possible: e.g., Sevilla as caption.
3. Presentation format can be updated easily for all works.

Structural metadata: TEI

```
<TEI.2>
    <teiHeader> [ TEI Header information ]  </teiHead
    <text>
        <front> [ front matter ... ]   </front>
        <body>  [ body of text ... ]   </body>
        <back>  [ back matter ...  ]   </back>
    </text>
</TEI.2>
```

Structural metadata: TEI

The Biblioteca Virtual Miguel de Cervantes

Text transcription, structural markup and accessibility

Tools for internet usage

Technological challenges

- `p`: paragraph (prose);
- `l`: verse line;
- `sp, speaker`: speech and character;
- `hi`: highlighted text;
- `pb`, `lb`: page and line breaks;
- `div`,`div1`,...: document sections.

TEI: most usual tags

The Biblioteca Virtual Miguel de Cervantes

Text transcription, structural markup and accessibility

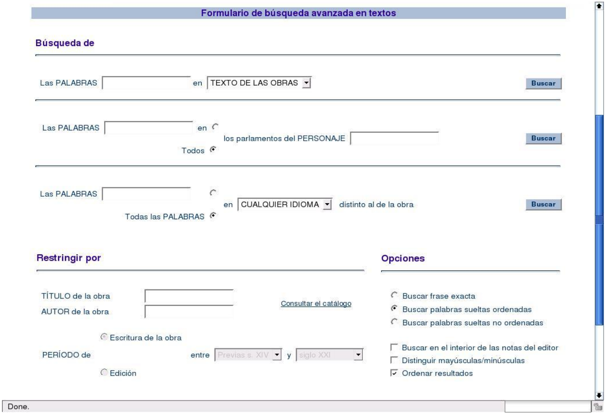
Tools for internet usage

Technological challenges

1	p	450059	26.9 %
2	l	402754	24.1 %
3	hi	249863	14.9 %
4	sp	84453	5.06 %
5	speaker	84277	5.05 %
6	pb	67971	4.07 %
7	cell	41664	2.49 %
8	note	34880	2.09 %
9	head	32020	1.91 %
10	foreign	29178	1.74 %
11	stage	23103	1.38 %
12	name	21922	1.31 %
13	lg	18731	1.12 %
14	row	13379	0.802 %
15	div1	11452	0.686 %
16	item	11252	0.674 %
17	q	10323	0.618 %
18	div0	6568	0.393 %
19	eDate	4981	0.298 %
20	lb	4901	0.293 %
21	figure	4807	0.28 %
22	change	3955	0.23 %
23	resp	3955	0.23 %
24	respStmt	3955	0.23 %
25	div2	3659	0.22 %
26	title	3080	0.18 %
27	bibl	2499	0.14 %
28	language	2474	0.14 %
29	role	2319	0.14 %
30	castItem	2319	0.14 %
31	milestone	1540	0.09 %
32	div3	1144	0.06 %
33	titlePart	1051	0.06 %
34	front	1051	0.06 %
35	docTitle	1051	0.06 %
36	titlePage	1051	0.06 %
37	text	1051	0.06 %
38	body	1046	0.06 %
39	table	1045	0.06 %
40	author	1033	0.06 %

3. Tools for internet usage

Full text + XML search



Formulario de búsqueda avanzada en textos

Búsqueda de

Las PALABRAS [ ] en [TEXTO DE LAS OBRAS ▾] **Buscar**

Las PALABRAS [ ] en ○ los parlamentos del PERSONAJE [ ] Todos ● **Buscar**

Las PALABRAS [ ] ○ en [CUALQUIER IDIOMA ▾] distinto al de la obra Todas las PALABRAS ● **Buscar**

Restringir por

TÍTULO de la obra [ ]
AUTOR de la obra [ ]
Consultar el catálogo

PERÍODO de ○ Escritura de la obra ○ Edición entre [Previas s. XIV ▾] y [siglo XXI ▾]

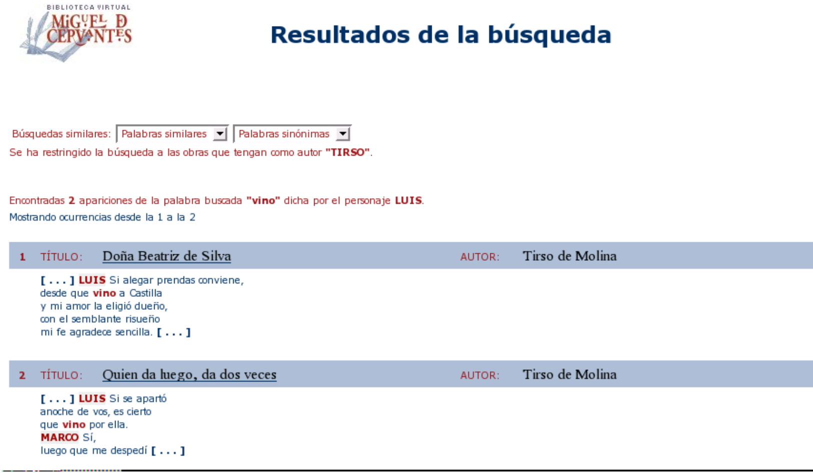
Opciones

- ○ Buscar frase exacta
- ● Buscar palabras sueltas ordenadas
- ○ Buscar palabras sueltas no ordenadas

- □ Buscar en el interior de las notas del editor
- □ Distinguir mayúsculas/minúsculas
- ☑ Ordenar resultados

Done.

Full text + XML search



Resultados de la búsqueda

Búsquedas similares: Palabras similares ▾ Palabras sinónimas ▾

Se ha restringido la búsqueda a las obras que tengan como autor **"TIRSO"**.

Encontradas **2** apariciones de la palabra buscada **"vino"** dicha por el personaje **LUIS**.

Mostrando ocurrencias desde la 1 a la 2

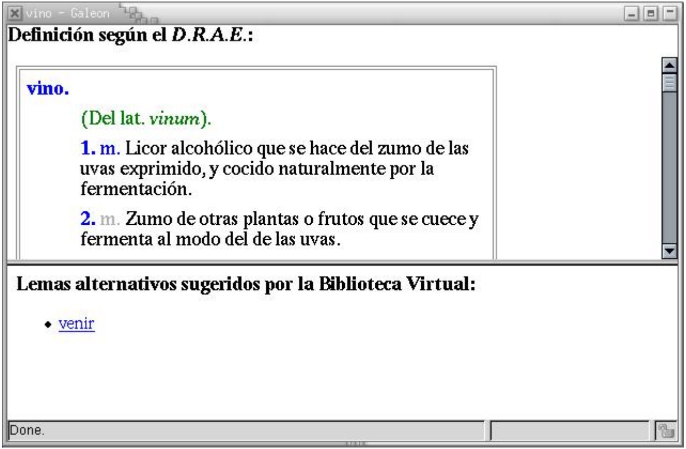
1 TÍTULO: Doña Beatriz de Silva AUTOR: Tirso de Molina

[. . .] **LUIS** Si alegar prendas conviene,
desde que **vino** a Castilla
y mi amor la eligió dueño,
con el semblante risueño
mi fe agradece sencilla. **[. . .]**

2 TÍTULO: Quien da luego, da dos veces AUTOR: Tirso de Molina

[. . .] **LUIS** Si se apartó
anoche de vos, es cierto
que **vino** por ella.
MARCO Sí,
luego que me despedí **[. . .]**

Dictionaries



Definición según el *D.R.A.E.*:

vino.

(Del lat. *vinum*).

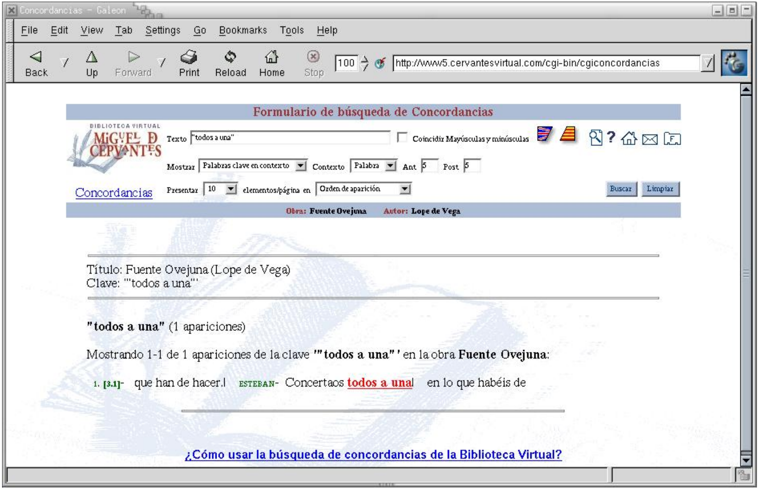
1. m. Licor alcohólico que se hace del zumo de las uvas exprimido, y cocido naturalmente por la fermentación.

2. m. Zumo de otras plantas o frutos que se cuece y fermenta al modo del de las uvas.

Lemas alternativos sugeridos por la Biblioteca Virtual:

- venir

Words in context


Concordancias - Galeon
File Edit View Tab Settings Go Bookmarks Tools Help
Back
Up
Forward
Print
Reload
Home
Stop
100
http://www5.cervantesvirtual.com/cgi-bin/cgiconcordancias
BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES
Concordancias
Buscar
Limpiar
¿Cómo usar la búsqueda de concordancias de la Biblioteca Virtual?

Formulario de búsqueda de Concordancias

Texto "todos a una"
□ Coincidir Mayúsculas y minúsculas

Mostrar Palabras clave en contexto Contexto Palabra Ant 5 Post 5

Presentar 10 elementos/página en Orden de aparición

Obra: **Fuente Ovejuna** **Autor:** **Lope de Vega**

Título: Fuente Ovejuna (Lope de Vega)
Clave: '"todos a una"'

"todos a una" (1 apariciones)

Mostrando 1-1 de 1 apariciones de la clave **'"todos a una"'** en la obra **Fuente Ovejuna**:

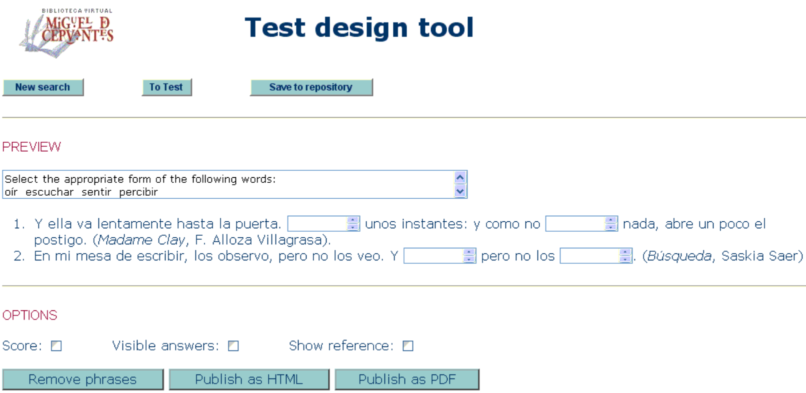
1. [3.1]- que han de hacer.| ESTEBAN- Concertaos **todos a una|** en lo que habéis de

Learning objects

Educative resources are an important added value to the library: we plan to add tools to create, store and use learning objects. Some examples of prototypes:

- Sort.
- Associate.
- Assign.

Learning objects



Test design tool

PREVIEW

Select the appropriate form of the following words:
oír escuchar sentir percibir

1. Y ella va lentamente hasta la puerta. ____ unos instantes: y como no ____ nada, abre un poco el postigo. (*Madame Clay*, F. Alloza Villagrasa).
2. En mi mesa de escribir, los observo, pero no los veo. Y ____ pero no los ____. (*Búsqueda*, Saskia Saer)

OPTIONS

Score: □ Visible answers: □ Show reference: □

Learning objects

Escribe cada una de las siguientes palabras en el hueco correspondiente: “contumaz”, “pertinaz”, “obstinado” “impertérrito” y “perseverante”.

1. “Pero porque no digas que soy . . . y que jamas hago lo que me aconsejas, por esta vez quiero tomar tu consejo y apartarme. (Don Quijote, Cervantes)
2. “D. Luis era . . . , era terco,” (Pepita Jiménez, Juan Valera)
3. Un silencio firme, . . . , fue la única respuesta. (Por la gloria, E. Pardo Bazán).
4. Concha trepaba . . . , unas veces a pie y otras a caballo, según los accidentes del terreno. (De Cartago a Sagunto, B. Pérez Galdós).
5. El aire del mar y el . . . ejercicio me prestaron alguna fuerza (Cuentos nuevos, E. Pardo Bazán).

4. Technological challenges

Most frequent words

The Biblioteca Virtual Miguel de Cervantes

Text transcription, structural markup and accessibility

Tools for internet usage

Technological challenges

Most frequent words

n	palabra	$f(n)$
1	de	5.952.871
2	que	4.294.496
3	y	3.887.331
4	la	3.473.934
5	en	2.521.954
6	el	2.463.429
7	a	2.348.470
8	los	1.689.770
9	se	1.305.932
10	no	1.261.456

Dictionary size: 995.855 words.

Zipf's law

Zipf:

$$f(n) \simeq \frac{C}{n}$$

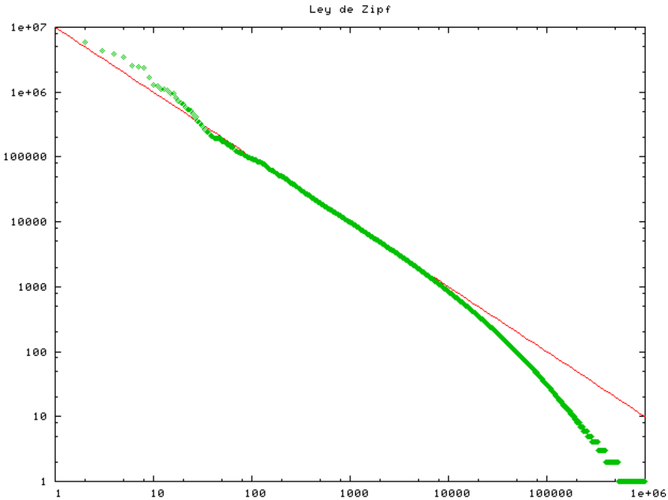
Zipf's law

Zipf:

$$f(n) \simeq \frac{C}{n}$$

n	palabra	$f(n)$	C/n
10	no	1.261.456	1.000.000
100	da	93.619	100.000
1000	penas	9.837	10.000
10000	francamente	841	1000

Zipf's law


Ley de Zipf
1e+07
1e+06
100000
10000
1000
100
10
1
1
10
100
1000
10000
100000
1e+06

Zipf's law and coverage

Consequences:

- About $\frac{1}{2}$ of words are hapax legomena.

Zipf's law and coverage

Consequences:

- About $\frac{1}{2}$ of words are hapax legomena.
- About $\frac{2}{3}$ appear less than 3 times.

Zipf's law and coverage

Consequences:

- About $\frac{1}{2}$ of words are hapax legomena.
- About $\frac{2}{3}$ appear less than 3 times.
- A dictionary containing 10 of words covers 85 of text; however, 50 is needed for a 95ċoverage

Zipf's law and coverage

Consequences:

- About $\frac{1}{2}$ of words are hapax legomena.
- About $\frac{2}{3}$ appear less than 3 times.
- A dictionary containing 10 of words covers 85 of text; however, 50 is needed for a 95ċoverage

A dictionary with millions of words needs automatic supervision. State of the art: about 10 % of words may need manual correction.

Markup correction

Shallow markup: simple, automatizable ⇔ Deep markup: expensive, handcrafted.

TEI y TEI-Lite

Even TEI-Lite DTD is far too complex, as it includes markup for:

1. Syntactic analysis.

TEI y TEI-Lite

Even TEI-Lite DTD is far too complex, as it includes markup for:

1. Syntactic analysis.
2. Multilingual parallel texts.

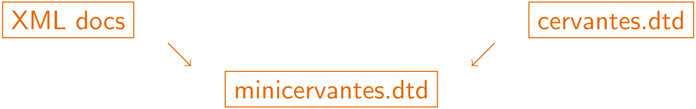
TEI y TEI-Lite

Even TEI-Lite DTD is far too complex, as it includes markup for:

1. Syntactic analysis.
2. Multilingual parallel texts.
3. Critical editions.

DTD simplification

- Reducing the number of possibilities reduces markup mistakes (http://www.dlsi.ua.es/~carrasco/RCCsoft.html).



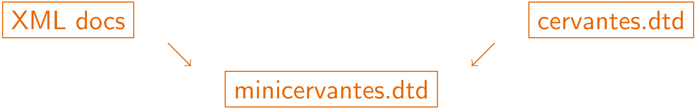
XML docs

cervantes.dtd

minicervantes.dtd

DTD simplification

- Reducing the number of possibilities reduces markup mistakes (http://www.dlsi.ua.es/˜carrasco/RCCsoft.html).



XML docs → minicervantes.dtd ← cervantes.dtd

- TEI header is imported from catalographic data.

An example

Element speaker in teilitex.dtd:

```
<!ELEMENT speaker
      (#PCDATA | ident | code | kw | abbr | address
      | date | name | num | rs | time | add | corr
      | del | orig | reg | sic | unclear | emph
      | foreign | gloss | hi | mentioned | soCalled
      | term | title | ptr | ref | xptr | xref | s
      | seg | gi | formula | index | interp | interpGrp
      | lb | milestone | pb | gap | anchor)* >
```

After simplification:

```
<!ELEMENT speaker
      ((#PCDATA | name | note | foreign | hi | lb)*) >
```

Conclusions

1. Text transcription and structural markup help to build, exploit and preserve digital libraries.

Conclusions

1. Text transcription and structural markup help to build, exploit and preserve digital libraries.
2. Large scale production needs tool for:

Conclusions

1. Text transcription and structural markup help to build, exploit and preserve digital libraries.
2. Large scale production needs tool for:
   - reliable automatic transcription (and markup);

Conclusions

1. Text transcription and structural markup help to build, exploit and preserve digital libraries.
2. Large scale production needs tool for:
   - reliable automatic transcription (and markup);
   - search information using also metadata.

Conclusions

1. Text transcription and structural markup help to build, exploit and preserve digital libraries.
2. Large scale production needs tool for:
   - reliable automatic transcription (and markup);
   - search information using also metadata.
3. Digital libraries call for cooperation between scholars, developers and practitioners.

Digital Libraries on Internet: design and exploitation

Document available at:
http://www.cervantesvirtual.com/documentos/talk.pdf

BIBLIOTECA VIRTUAL
MIGUEL DE CERVANTES

http://www.cervantesvirtual.com