Digital Libraries on Internet: design and exploitation

Rafael C. Carrasco (carrasco@dlsi.ua.es)
Director Adjunto de la 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.
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 a successful experience of cooperation between private and public partners, now supported by
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

About 1 million pages indexed.

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pages server successfully (per year)</td>
<td>142,886,352</td>
</tr>
<tr>
<td>Average number of pages (per day)</td>
<td>220,469</td>
</tr>
<tr>
<td>Average users (per day)</td>
<td>50,000</td>
</tr>
<tr>
<td>Pages(time) per user</td>
<td>4.4 (3 minutes)</td>
</tr>
<tr>
<td>Average transfer (per day)</td>
<td>34 GB</td>
</tr>
</tbody>
</table>
Impact

The Biblioteca Virtual Miguel de Cervantes

Text transcription, structural markup and accessibility

Tools for internet usage

Technological challenges
Top 10 domains

.es (45 %)
Top 10 domains

- .es (45%)
- .mx (12%)
Top 10 domains

- .es (45%)
- .mx (12%)
- .us (9%)
Top 10 domains

- .es (45%)
- .mx (12%)
- .us (9%)
- .pe (5%)
Top 10 domains

- .es (45%)
- .mx (12%)
- .us (9%)
- .pe (5%)
- .ar (4%)
Top 10 domains

- .es (45%)
- .mx (12%)
- .us (9%)
- .pe (5%)
- .ar (4%)
- .fr (3%)
Top 10 domains

- .es (45%)
- .mx (12%)
- .us (9%)
- .pe (5%)
- .ar (4%)
- .fr (3%)
- .ve (3%)
Top 10 domains

- .es (45%)
- .mx (12%)
- .us (9%)
- .pe (5%)
- .ar (4%)
- .fr (3%)
- .ve (3%)
- .cl (3%)
Top 10 domains

- .es (45%)
- .mx (12%)
- .us (9%)
- .pe (5%)
- .ar (4%)
- .fr (3%)
- .ve (3%)
- .cl (3%)
- .co (2%)
Top 10 domains

- .es (45%)
- .mx (12%)
- .us (9%)
- .pe (5%)
- .ar (4%)
- .fr (3%)
- .ve (3%)
- .cl (3%)
- .co (2%)
- .it (1%)
¿Which are the most accessed books?
Top books

¿Which are the most accessed books?

1. El ingenioso hidalgo Don Quijote de la Mancha.
Top 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).
Top 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. Cantar del Mío Cid.
Top 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. Cantar del Mío Cid.
4. La Celestina.
Top 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. Cantar del Mío Cid.
4. La Celestina.
5. La vida de Lazarillo de Tormes.
Top 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. Cantar del Mío Cid.
4. La Celestina.
5. La vida de Lazarillo de Tormes.
6. La vida es sueño (Pedro Calderón de la Barca.)
¿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. Cantar del Mío Cid.
4. La Celestina.
5. La vida de Lazarillo de Tormes.
6. La vida es sueño (Pedro Calderón de la Barca.)
Top 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. Cantar del Mío Cid.
4. La Celestina.
5. La vida de Lazarillo de Tormes.
6. La vida es sueño (Pedro Calderón de la Barca.)
8. La Ética de Aristóteles.
Top 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. Cantar del Mío Cid.
4. La Celestina.
5. La vida de Lazarillo de Tormes.
6. La vida es sueño (Pedro Calderón de la Barca.)
8. La Ética de Aristóteles.
9. Gramática de la Lengua Castellana (RAE)
Top 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. Cantar del Mío Cid.
4. La Celestina.
5. La vida de Lazarillo de Tormes.
6. La vida es sueño (Pedro Calderón de la Barca.)
8. La Ética de Aristóteles.
9. Gramática de la Lengua Castellana (RAE)
10. Marianela (Benito Pérez Galdós).
Top 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. Cantar del Mío Cid.
4. La Celestina.
5. La vida de Lazarillo de Tormes.
6. La vida es sueño (Pedro Calderón de la Barca.)
8. La Ética de Aristóteles.
9. Gramática de la Lengua Castellana (RAE)
10. Marianela (Benito Pérez Galdós).
11. Canto General (Pablo Neruda)
Top 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. Cantar del Mío Cid.
4. La Celestina.
5. La vida de Lazarillo de Tormes.
6. La vida es sueño (Pedro Calderón de la Barca.)
8. La Ética de Aristóteles.
9. Gramática de la Lengua Castellana (RAE)
10. Marianela (Benito Pérez Galdós).
11. 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...
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, poesía española contemporánea Cultura chicana ...
Browsing

About most relevant authors, the library includes: biography, research work, images and sound, external links.
A catalan section called Joan Lluis Vives

La Biblioteca Virtual Joan Lluis Vives és fruit d'un conveni entre la Fundació Biblioteca Virtual Miguel de Cervantes i la Xarxa d'Universitats que formen l'Institut Joan Lluis Vives.

Amb aquest acord es pretén, d'una banda, aprofitar, potenciar i difondre al màxim els fons digitalitzats existents a la xarxa i, d'una altra, proposar i fer efectiva l'edició digital de les obres més representatives de la cultura catalana, valenciana i balear.

El portal rei homenatge a la figura de l'important humanista valencià Joan Lluis Vives, per tot el que representa la seva obra en la cultura europea del Renaixement.
Additional content

1 History.
Additional content

1. History.
2. Signs.
Additional content

1. History.
2. Signs.
3. Children.
Additional content

1 History.
2 Signs.
3 Children.
4 Journals and Ph.D. dissertations.
Additional content

1. History.
2. Signs.
3. Children.

Map
2. Text transcription, structural markup and accessibility
Production workflow

START (selection)

Cataloguing

Catalogue

TEI header

Interface

User

Catalogue

Image

HTML

XSLT Transformation

Supervised XML

Scan

Content created for BVMdC

Facsimile

OCR

Supervision & markup

Automatically processed text
**Catalogue**

Books, web pages and images are catalogued (MARC21).

<table>
<thead>
<tr>
<th><strong>Tit. uniforme:</strong></th>
<th>Poema del Cid</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Título:</strong></td>
<td>Texto modernizado del Cantar de Mio Cid</td>
</tr>
<tr>
<td></td>
<td>Timoteo Riaño Rodríguez y Mª. Carmen Gutiérrez Aja, edición didáctica para el Aula Virtual del Mio Cid</td>
</tr>
<tr>
<td><strong>Publicación:</strong></td>
<td>Alicante : Biblioteca Virtual Miguel de Cervantes, 2007</td>
</tr>
<tr>
<td><strong>Título de serie:</strong></td>
<td>Poema de Mío Cid. Ediciones</td>
</tr>
<tr>
<td><strong>Entradas secundarias:</strong></td>
<td>* Riaño Rodríguez, Timoteo , ed. lit.</td>
</tr>
<tr>
<td></td>
<td>* Gutiérrez Aja, Mª del Carmen , ed. lit.</td>
</tr>
<tr>
<td><strong>Portal:</strong></td>
<td>Cantar de Mío Cid</td>
</tr>
<tr>
<td><strong>CDU:</strong></td>
<td>821.134.2-13. épica.</td>
</tr>
<tr>
<td><strong>Encabezamiento de materia:</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>o Poema del Cid</td>
</tr>
<tr>
<td></td>
<td>o Poesía épica española</td>
</tr>
</tbody>
</table>

Esta obra ha sido consultada en 4215 ocasiones.
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 by applying design templates

([http://www.w3.org/WAI](http://www.w3.org/WAI))
El sombrerero 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.
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.
User tests became an important source of information:

- User does not remember the address. The header included icons to bookmark or make homepage.
- 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.
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 is information about information, that is, about contents:

<table>
<thead>
<tr>
<th>155.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUTes</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
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 española - Siglo 17"/>
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>
Text editing and formatting become differentiated tasks.
Structural metadata: TEI

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

1. Text editing and formatting 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 ] </teiHeader>
  <text>
    <front> [ front matter ... ] </front>
    <body> [ body of text ... ] </body>
    <back> [ back matter ... ] </back>
  </text>
</TEI.2>
Structural metadata: TEI

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

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

```xml
<app>
    <lem wit="1">Per què</lem>
    <rdg wit="1 2">Per què</rdg>
    <rdg wit="3">Per çò</rdg>
</app>
```

The text shown (editor’s choice and 3 sources in the example) is selected and differences highlighted through optional stylesheets.
3. Tools for internet usage
Full text + XML search
Resultados de la búsqueda

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 Sí 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 Sí se apartó
   anoche de vos, es cierto
   que vino por ella.
   MARCO Sí,
   luego que me despedí [ ... ]
Definición según el *Diccionario de la lengua española* (RAE):

**REAL ACADEMIA ESPAÑOLA**

DICIONARIO DE LA LENGUA ESPAÑOLA - Vigésima segunda edición

**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 Miguel de Cervantes:**

- *venir*
Linguistic search engine

Text search will allow soon for combined expressions using surface form, lemmas and part-of-speech tags. Three different indices are created: words, lemmas and categories.

- **sf**: looks for surface forms, e.g., *sf:se sf:dignará*.
- **lem**: looks for lemma, e.g, *lem:dignarse* or *lem:línea lem:aéreo*.
- **tags**: seeks categories and morphology, e.g, *tags:prep* or *tags:vblex.fti*.

For instance: dignarse (all forms) + preposition = *lem:dignarse tags:prep*. 
"todos a una" (1 apariciones)

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

1. [3.0]- de hacer.| ESTEBAN- Concertaos todos a una| en lo que habéis de decir
Learning objects

Educative resources are an important added value to the library: a new tool allows to create learning objects.

Then, they can be downloaded as XHTML or SCORM.
Learning objects

Some examples of produced activities:

- Sort.
- Associate.
- Assign.
- Word seek.
Learning objects

The tool also integrates full text search so that passages can be easily reused for activities.
4. Technological challenges
New role for users

New services must be created in order to help users to have a more active role, for instance,

- Content personalization.
- Cooperative tagging and folk ranking.
- Adaptive linguistic resources (including dictionaries, machine translation services, speech synthesis).
- Wikis and content integration services.

The priorities must be set by potential users
Most frequent words
## Most frequent words

<table>
<thead>
<tr>
<th>$n$</th>
<th>palabra</th>
<th>$f(n)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>de</td>
<td>5,952,871</td>
</tr>
<tr>
<td>2</td>
<td>que</td>
<td>4,294,496</td>
</tr>
<tr>
<td>3</td>
<td>y</td>
<td>3,887,331</td>
</tr>
<tr>
<td>4</td>
<td>la</td>
<td>3,473,934</td>
</tr>
<tr>
<td>5</td>
<td>en</td>
<td>2,521,954</td>
</tr>
<tr>
<td>6</td>
<td>el</td>
<td>2,463,429</td>
</tr>
<tr>
<td>7</td>
<td>a</td>
<td>2,348,470</td>
</tr>
<tr>
<td>8</td>
<td>los</td>
<td>1,689,770</td>
</tr>
<tr>
<td>9</td>
<td>se</td>
<td>1,305,932</td>
</tr>
<tr>
<td>10</td>
<td>no</td>
<td>1,261,456</td>
</tr>
</tbody>
</table>

Dictionary size: 995,855 words.
Zipf’s law

Zipf:

\[ f(n) \approx \frac{C}{n} \]
Zipf’s law

Zipf:

\[ f(n) \approx \frac{C}{n} \]

<table>
<thead>
<tr>
<th>( n )</th>
<th>palabra</th>
<th>( f(n) )</th>
<th>( C/n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>no</td>
<td>1.261.456</td>
<td>1.000.000</td>
</tr>
<tr>
<td>100</td>
<td>da</td>
<td>93.619</td>
<td>100.000</td>
</tr>
<tr>
<td>1000</td>
<td>penas</td>
<td>9.837</td>
<td>10.000</td>
</tr>
<tr>
<td>10000</td>
<td>francamente</td>
<td>841</td>
<td>1000</td>
</tr>
</tbody>
</table>
Zipf’s law
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 95% coverage.
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 95% coverage.

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

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

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

1. Syntactic analysis.
Reducing the number of possibilities reduces markup mistakes (sourceforge.net/projects/dtdprune).

- XML docs
- minicervantes.dtd
- cervantes.dtd
DTD simplification

- Reducing the number of possibilities reduces markup mistakes ([sourceforge.net/projects/dtdprune](sourceforge.net/projects/dtdprune)).

- TEI header is imported from catalogographic 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.
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.

4. A more active role of users is foreseen. New services and tools will be added according to users demand.
Digital Libraries on Internet: design and exploitation

Document available at:

http://www.cervantesvirtual.com