III Implementation of an e-library in the Parliament

Communication from Mrs Adelina SA CARVALHO (Portugal), Marrakech Session (March 2002)

Mr Ian HARRIS, Vice-President, thanked the President and invited her to address the assembly.

Mrs Adeline SA CARVALO, President, spoke as follows:

"It is an honour for me to present to you today the ways in which the Assembly of the Republic is addressing the need to electronically process and disseminate massive amounts of parliamentary information via the *Web {intranet* and *Internet),* in both text and graphic form.

Of special interest are the minutes of parliamentary debates conducted during specific historical periods in the life of the Parliament, from 1821 until present days. These periods can be organized as follows: The Constitutional Monarchy (1821-1910); The First Republic (1910-1925); The "New State" (1935-1974) and The Third Republic (1974-2002).

This ambitious project got under way in May 2000. It is to be accomplished in 2 planned phases:

1. A first phase, already concluded, covering the microfilming, digitalization and OCR (Optical Character Recognition) scanning of parliamentary debates from 1935 till 2002, and the development of a computer application for the storage, search and disseminate the aforementioned Web information.

2. A second phase, which started in July 2001, uses the same processes for inputting the minutes of parliamentary debates for the 1821-1925 period. Processing documents for this earlier period is particularly complex, since not only them are early documents written using XIX century spelling, but some texts are in manuscript form. When this is the case, texts are captured graphically and duly catalogued.

We expect the project to be completed by next year and we strongly believe that it will constitute an important tool for the study of Portuguese parliamentary history, thereby contributing to a better understanding of parliamentary life. We hope that it will bring citizens closer to their Parliament.

For the practical implementation of this project, we decided to work together with acknowledged Centers of Excellence in the field of innovative information technologies. The Portuguese Parliament, within the framework of a Protocol signed with the Portuguese Public Universities Network, has joined in partnership with the University of Aveiro - the ideal partner for the current project's successful completion, as well as for other projects currently under development.

The experience achieved is naturally of mutual benefit: The Parliament profits from the dynamic addition of the university's research capability, its capacity for innovation and its enthusiasm. By the same token, the University enjoys access to the Parliament's wealth of research resources, while establishing a practical link to the working world, specifically in light of the demands of parliamentary activities.

Let me stress that this major national priority - the opening up of the workings of State administration to citizens via the tools of the information revolution - has been retained by Parliament as a one of the core objectives of our information policy. Specifically, the goal is for universal electronic access to parliamentary documents.

The current project will provide both internal and external users with a search gear for consulting these documents, which can be used for accessing digitally-available sources, but also, importantly, the integral texts in graphical format.

An added benefit is that the Assembly of the Republic will be able to provide for the conservation of original documents in optimal conditions in its Historical Archives.

It is now my pleasure to present to you our project partners: Professor Joaquim Arnaldo Martins, Director of the Center for Informatics and Communication Studies at the University of Aveiro, and Associate Professor Sousa Pinto, also from this Center. Afterwards, they will remain on hand to answer any questions you may have.

They will be making their speeches in English, so I would ask that you give them your full attention."

PORTUGUESE PARLIAMENTARY RECORDS DIGITAL LIBRARY

Joaquim Arnaldo Martins - Joaquim Sousa Pinto
Dep. Electrónica e Telecomunicagões /IEETA, Universidade de Aveiro, 3810 Aveiro - Portugal
E-mail: :jam@det. ua.pt, jsp@det. ua.pt

Key words: Digital Library, Portuguese parliament

Abstract: The first phase of the digitalisation of the Portuguese Parliament proceedings will cover the historical period between 1935-2000 and will sum up to around 200.000 pages. In this stage we intend to develop the mechanisms as well as the tools for the entire project of the Parliament Digital Library, which we predict will have up to one million pages. The process involves the microfilming of all the available material, the scanning of the material so as to obtain the image of the page as well as the use of an OCR - Optical Character Recognition - that recovers the original text, allowing to search in the original documents. The project concerns printed material only, manuscripts will be handled separately in a later stage. This project was initially aimed at the disposal of the Parliament Intranet, however this information is already accessible on the Internet. At this stage, the material is organized in small brochures with an average of 40-50 pages each and contains the speeches of members of Parliament. Each one of the pages is treated individually, so when the user is looking for specific information, he is drawn to the page or pages in which the expression is used. The visualization of the pages can be either in text mode or in the original text through digitalized image. Despite the granularity of the system being "page", which means that each page is treated as a complete element, it is possible to print the entire document, in text or image, obtaining therefore a copy of the original document, because each page is wrapped with metadata.

Introduction

The organization and structure of the information is closely related to the last two centuries of the Portugal history. There is a huge amount of information to process, so it was split into two main groups: from 1821 to 1934 and from 1935 to 2000. The first phase of the Parliamentary Digital Library focuses on three historical periods: the "New State", the "Constitution Assembly" and the present political structure. The 2nd phase will go back to the beginning of the Portuguese parliamentary history. Other phases will follow with the inclusion of other relevant information about the Portuguese parliamentary history; an estimated figure of 1 million documents will soon be a reality.

For further details about the Portuguese parliamentary history, visit the web site:

System architecture

The system architecture, presented in figure 1, has four basic modules: the Information System, the Indexing Service, the Web Server and the Interface with other systems.

The Information System is responsible for the information repository. It stores the image files, scanned from the original brochures, the text files, after being post processed with the OCR and added with metadata, and it maintains the relation between a text file and the correspondent image. This module also sends information to the Web Server Module when requested by the user. The Indexing Service when queried by a user should answer with a list of documents stored on the Information System containing the query string. To perform such task the service pre indexes all documents and stores such information in internal tables. To disallow the system to index all words, this module has a stop word list that is controlled by the system administrator. This word list is not the same in all scenarios and must be tuned to promote smaller indexes and consequently faster answers to user queries.

The Web Server module makes the interface with the end users and with the Information System and the Indexing Service.

The Interface with Other Systems allows this system to exchange information with foreign

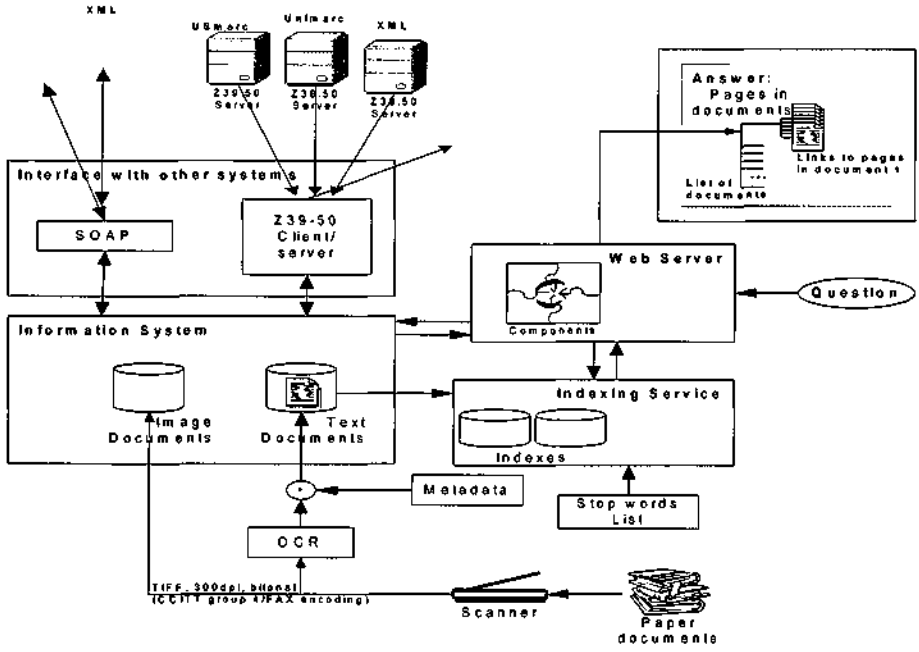

Figure 1 - Overall architecture

systems: the Z39.50 Server module allows other systems to query the Information System and a Z39-50 Client module to retrieve information on foreign systems; client and server modules to allow interaction with other applications using SOAP are also under development.

In the next chapters we will present the information structure used in the Information System and Indexing Service, the user interface developed for the Web Server.

Hardware

There are several solutions for balancing loads of work required by web servers. Ingham's article [6] describes the various techniques that exist along with the advantages and disadvantages related to each one. The hardware structure we use lies on two web servers with different tasks: one is merely associated with the queries of users and the other is associated with the presentation

of information to the user. To share the data among the servers we have used a shared SCSI bus configuration.

All the hyperlinks that are presented as a result of a search and the navigation hyperlinks point to storage servers; the hyperlinks that allow the user to make new searches point to the query server. This method may allow the task division and if there are major asymmetries between the load distributions, the possibility of the HTTP redirect technique is simultaneously used with a different number of machines in each one of the sub-systems.

Information structure

The data refers to the three historical periods presented earlier. During the period designated as the "New State", another assembly was conceived in 1954, this assembly was responsible for elaborating studies and technical reports for consultative purposes. This chamber was known as the "Consulting Chamber" (Câmara Corporativa). The proceedings that resulted from this chamber are a different sort of parliamentary information, so it has been separated from the latter. In the last three historical periods, there are three kinds of different information, which are placed in four different catalogues:

- National Assembly - refers to the parliament records in the historical period of the "New State" (1935-1974);
- « Consulting Chamber - refers to the records produced by the Consulting Chamber in the historical period of the "New State" (1954-1974);
- Constitution Assembly - refers to the records produced by the Constitution Assembly in the period of June/75 - April/76;
- Parliament - refers to present records produced by the Assembly of the Republic, the recent name of the Portuguese parliament (1976-99).

The main advantages are that the amount of information in each catalogue is less and searches are quicker and more precise. Besides these advantages, each one of the catalogues is closely related to the social and political organization of a country in a certain historical period.

The main disadvantage is that when we need information about a subject related to several historical periods, the same question must be restated in each one of the catalogues and only then may all the answers be combined.

The information in each catalogue is presented in a tree diagram and according to the following:

Legislature (period up to 4 years);

Legislative session (period of a legislative year during a legislature);

Number of documents produced in each legislative session (in each legislative session the numbering of documents restarts);

Page in a document.

As far as the organization of information is concerned, the granularity of the system is "page". This is one of the system key points, which advantages are:

- The overload of the server is less since only the text that corresponds to one page, according to the user's request, will be sent, instead of 40-50 pages that normally form the entire document. The first trials of the system revealed that the user did not read the entire text; the user normally reads one or two pages, rarely are more than three pages of one diary read. These figures are similar to those referred in [1].
- The mental model is still the "diary page" and not an electronic document. There is a mapping process between the electronic document and the document on paper.

The disadvantages of this organization of information are:

- More requirements in management and organization of information in the server, therefore the system becomes more vulnerable to faults. Instead of having one document with 40 or 50 pages, we have 50 electronic documents - one per page.
- It is necessary to add extra data in each page, this will allow the system to recreate the diary each time a user accesses a page.

From the list of advantages and disadvantages, it is clear that the advantages will benefit the user of the system while the server must deal with the disadvantages; this is what we consider the correct option. Every page is stored simultaneously in two formats: text and image. For the text the HTML format was adopted, which allows keeping the original structure of the text, as well as to add a set of extra information after having accessed any page. A TIFF (Tagged Interchange File Format) with a resolution of 300 dpi, bi-tone with the CCITT Group 4 /FAX encoding was adopted for the images of each one of the pages.

Document metadata

It is necessary to add extra information to the document, so that we may handle it as a whole, after we have had access to one of the pages. In order to do so, a set of metadata fields is added to each page. The Dublin Core Metadata Initiative [2] set of qualifiers was used when possible. The qualifiers defined in the Initiative do not constitute a closed set, designed to meet all of the descriptive need of the implementators. The complete set of metadata fields adopted in this implementation, using the DCMI notation, is presented in Table 1:

Numeric field with the document legislature

Element:	**Legislature**
Name:	**Document Legislature**
Identifier.	**Legislature**
Definition:	
Comment:	**This value changes accordingly inside a catalog. In the catalog reporting to the present historical period, this value ranges from 1 to 7.**
Element:	Session
Name:	**Document Session**
Identifier:	**Session**
Definition:	**Numeric field with the document session.**
Comment:	**In each legislative session the diary numbering restarts. As described, a legislature usually has up to four sessions.**
Element:	Number
Name:	**Document Number**
Identifier:	**Number**
Definition:	**Numeric field with the brochure number.**
Comment:	**This element, associated with the previous two elements, completely describes a restarts document.**
Element:	Page
Name:	**Document Page**
Identifier:	**Page**
Definition:	**String field with the document page.**
Comment:	**The data in this metafield is twofold. When converted to integer, this value corresponds to the page number. By adding/subtracting the page number from 1 the next/previous page number is obtained.**
Element:	Limits
Name:	**Document Limits**
Identifier:	**Limits**
Definition:	**String field with the document limits.**
Comment:	**This metafield combined with the previous metafield, allows the navigation through the document.**
Element:	Created

Name: Created
Identifier: Created
Definition: Date field with the date of the parliamentary session.
Datatype: W3C-DTF
Comment: This metadata field is as defined in Dublin Core Qualifiers |3|.

Element: Issued
Name: Issued
Identifier. Issued
Definition: Date field with the issued date of the brochure by the National Press.
Datatype: W3C-DTF
Comment: This date is usually the next weekday after the parliamentary session.

Element: DocSerie
Name: Document Serie
Identifier: DocSerie
Definition: Numeric field with the number of the document series.
Comment: The parliament has several documents series: The Ist series (parliamentary sessions); The 2nd series (parliamentary commissions).

Element: Keywords
Name: Keywords
Identifier: Keywords
Definition: String field containing the keywords for the brochure.
Datatype: Eurovoc Thesaurus |5|.
Comment: This field has all the qualifiers associated with the entire brochure. This allows thematic searches.

Element: PagCategory
Name: Page Category
Identifier: PagCategory
Definition: String filed used to classify the data in different pages of a brochure.
Comment: Up to now the pages are classified as "summary" and "diary" but in the future new values/ behaviors could be added.

Table I - List of metadata fields added to each page.

User interface

Due to the organization chosen for the information (granularity of information is "page") documents are in brochures and it is possible to view them in text and/or image mode, special attention was given to the interface with the user. This will allow the user to easily go to another page of the brochure or switch the visualization mode of the document, therefore switch between text and image mode, without losing the search context initially proposed.

Besides the handling of information, the utmost attention was also given to the search system, which enables the user to have access to the information.

Searching for information

The indexing engine used, indexes documents in several formats and allows Boolean searches in the body of the document and/or in the attributes of the indexed documents. The format used for storage and indexation of documents is HTML. Although the documents are not directly exposed, this format has the advantage of allowing the addition of new fields of information specification (metadata fields), without changing the structure of the information. It is possible for the user to search according to:

- Free text in the documents, with no restriction;
- Free text in the summary pages of the documents. This search looks for the expression we intend only in the pages, in which the attribute PagCategory = summary. In this case the expression we intend to search, is sought for in the attribute Keyword. Eurovoc Thesaurus

defines the terminology dictionary associated to this field. There is a special interface for this type of search; the hierarchy structure of thesaurus is graphically presented and the user only has to choose the expression he intends;

- Brochure. In this search, the user besides indicating the number of the brochure he intends to view, he should also indicate the legislative session because this numbering restarts in each legislative session.

All these searches are associated with a catalogue and subdivided in legislatures. A search through all legislatures in the entire catalogue presents a considerable degree of possibilities/ freedom and shows all the pages, which contain the occurrence. It is essential to bear in mind that the catalogues are organized in a tree diagram by legislatures, legislative sessions and by the number of brochures.

If the user intends to make a search related to two or more legislatures, the effectiveness of such a search is compromised by the internal organization of the catalogues and the search in all legislatures is too vast. However, we must bear in mind that crossing and relating information is decisive and imperial when it comes to history. A search that allows crossing information in a catalogue between two dates was developed. This search operates without the difficulty imposed by the internal chronological structure of the catalogue.

Any given expression we may intend to search is sought for as a whole, as a phrase. In order to search for the word mode the user will have to select the checkbox presented on the right of the search expression. The system will look for the exact combined expression "constituição revolucionária", whereas in word search mode the search expression would be equivalent to "constituigão or revolucionária".

Information presentation

After a search is implemented the user may immediately begin the exploration of information through the activation of hyperlinks accessible in the response window. For example, if the user chooses a hyperlink from page 116 from diary number 7, he will view the contents in text mode. In the upper part of the window there is a bar that locates the document in time (legislature, legislative session, record number, date of session and publication date) and also reveals the location inside the brochure (present page and limits of the brochure). This bar allows the navigation inside a document, through the hyperlinks to the first page, the last page and the previous or following page. There is a checkbox on the right hand side of the control bar that allow the user to switch between the visualization of the page in text or image mode as well as print either one or both of these options.

As it is obvious, the pages are neither stored with a predefined control bar nor with the research text introduced by the user. This processing has to be done in real time. On the other hand, there is no real interest in sending information from the metadata fields to the client. In order to process a text page and allow the presentation of the control bar, an ActiveX component was developed. The full pathname, the expression we intend to search and the search mode (phrase or word) for the text highlighting procedure are the input parameters of the component. As output parameters there is the highlighted text, which varies according to the search expression and the search mode, a list of names of the metadata fields and another list of the values of those same fields.

Owing to its architecture, this component adapts to any possible present or future scenario and presents the advantages of the omission of information of the metadata fields for the client and allows that the physical location of the documents being processed is out of range of the web server as well as the net spiders.

For the storage of images, because the original documents only contain text, we have adopted the format TIFF (Tagged Interchange File Format) with a resolution of 300 dpi, bitone with the CCITT Group 4 /FAX encoding. This resolution is considered a compromise between the size presented by files and the success rate of document OCR, after some exhaustive tests. The main advantage of this format is lost less. The original documents have the size of 2472 x 3489 pixels and occupy and average size of 90KB after coding.

The main disadvantage of this format is that the browsers do not automatically interpret it. In order to overcome this limitation, there are two possible approaches: to convert the format to one supported by the browser or to develop a plug-in capable of visualise TIFF.

The first approach is followed in JSTOR [1]. They produce screen-resolution page images in GIF format, which are readable down to about 60 dpi. We made some experiences about this last approach and we have decided to simply convert the format and send the client the image in its original size. Two reasons led to the abandon of this option: the first is associated with the increase of the amount of information transmittable to the client and the second and most important is related to the way the browser handles images. When the browser is required to reduce the size of an image the rendering algorithm is applied to the image and it is resized, which results in a poor image. The suitable algorithm for such a situation is one that resamples the image and develops intermediate grey levels; this increases the visual comfort of the user. Due to the inadequate way the browsers treat the redimension of images, we have decided to build a Plug-In/ActiveX Viewer that supports several formats of the image (BMP, TIFF, JPEG and GIF) and that is able to resize or resample the image, according to what is requested.

The Plug-In/ActiveX Viewer transfers the image via HTTP to the client's disc and then presents it in its native format. The component was designed to support different graphic resolutions (640x480, 800x600 e 1024x768) and several types of image zooms of the image. When compared to the procedure followed in JSTOR, the amount of information transferred is bigger but the effort in server management is smaller. On the other hand, the resample of an image is a task that involves image processing; with the methodology used, the effort is passed on to the client.

Document printing

The pages that form the brochure are handled individually, the printing of the entire document would require the user to handle each one of the pages and then press the print command. This procedure is acceptable in some cases, especially when the user needs extracts of precise information. However, when the user needs a complete copy of a document, this task can be very difficult because the average number of pages per document is 40-50 and some documents have up to 120 pages. Once more a Plug-In/ActiveX was developed, it allows the user to print the document in text format or the images of the original format, at once. In any case what the component does is to transfer all the files to the client and prints them one by one just as the user would have to do. Navigation elements are taken from the bar situated on the top of the page when printing text.

For printing images, the possibility of an application of a visible watermark over the image, as is referred to in [6], is now being tested. Images are transferred to the client side; all watermarking procedures are located on the client side. We may always question the safety of the process if it is done on the client side; the watermark and the files are transferred to temporary files in the user's disc and are immediately erased after printing in a continuous process. The possibility of catching images on the client side is completely out of the question.

Conclusion

This communication has described the implementation of an information system in the Portuguese Parliament which will allow the access to parliamentary diaries since 1821. Electronic records of this sort do not exist; therefore the development of a page-by-page image and text file is being implemented.

As far as the organization of information is concerned, it is separated into historical periods, designated as catalogues. In each catalogue, the information structure is organized by page, there are, nevertheless, functionalities in the system that allow the user to navigate inside a brochure. This form of partition of information allows quicker responses for the user because less information is transferred each time and the mental structure is kept the same. The documents present a similar structure to those of the original paper documents, this factor is extremely important in order to avoid rejections of the system.

The low granularity of the system based on the page metaphor, and the associated metadata model, proved to be the key point of this flexible and scalable system.

The system is not closed and there are functionalities that allow exporting or importing data to other systems as well as presenting information from remote systems.

References

[l]Thomas, Spencer W. et all, 1999, 'Technology Choices for JSTOR Online Archive', In: *Computer*, Volume 32, Number 2, February, Pages 60-65.

[2]dc,2000, *Dublin Core Metadata Element Set, Version 1.1: Reference Description,* http://purl.org/DC/documents/rec-dces-19990702.htm, last visited: July/2000.

[3]dc,2000, *Dublin Core Qualifiers,* http://www.purl.org/DC/documents/rec/dcmes-2000071 l.htm, last visited: July/2000.

[4] W3C, 2000, *Date and Time Formats,* W3C Note, http://www.w3c.org/TR/NOTE-datetime, last visited: July/2000.

[5]EC 1995, *Thesauro Eurovoc,* Luxemburg, EC, ISBN 92-77-86361-7.

[6]Mintzer, Fred, 1999, 'Developing Digital Libraries of Cultural Content for Internet Access', In: *IEEE Communications,* Volume 37, Number 1, January, Pages 72-78.

[7]Ingham, David B., et all, 2000 January/ February, 'Constructing Dependable Web Services', *IEEE Internet Computing,* Pages 25-33.

[8JXML, 1998, February, *Extensible Markup Language (XML) 1.0,* W3C, http://www.w3.org/TR/REC-xml, last visited: July/2000.

* * *

Mrs Adelina SÁ CARVALHO, President, thanked the speakers for their contribution. She noted that the project had involved working with very old texts on extremely thin paper. These were wearing out very quickly. The library had always been the poor relation of Parliament, and many papers had been allowed to become quite spoilt. This was regrettable as the whole part of the history of Parliament was possibly going to be lost. She welcomed the new project because it made research in this area possible and would make up for the eclipse of much parliamentary history during the period of the dictatorship in Portugal.

Mr Ian HARRIS, Vice-President, thanked Mrs SA CARVALHO.

Mr Mwelwa Ng'ono CHIBESAKUNDA (Zambia) asked what the total cost of restoring the information was and asked whether it came from the annual budget of parliament or from the university. He asked whether parliamentary staff was being trained in conservation work.

Mrs Adelina SÁ CARVALHO, replied that she was not sure where the money was coming from. She had a bad memory for figures but thought that the cost was about 0.5 Escudos per page. Parliament paid the university to do the work. She hoped that in the next financial year, she would see further work being carried out. It was a very easy system to use for anyone with a reasonable education in computing. She referred to the number of searches that had been carried out in the last months. On average, seventy-five searches had been made per day via the Internet. She also referred to the Intranet system where there was an unknown number of hits.

Mrs Marie-Andrée LAJOIE (Canada) thanked Mrs SA CARVALHO and noted that the Parliament in Canada was attempting something similar. In Canada previously there had been various different systems for producing documents and these had now been replaced by one common system. She asked what method was being used to produce the parliament's record and asked whether the system could produce text and image. She also asked whether a copy of the presentation was available.

Mrs Piedad Garcia ESCUDERO (Spain) noted that a similar process was being followed in Spain. At first in the Senate, the system had originally been too rigid and the new system had been created. She referred to the importance of conservation and said that whenever old documents were asked for the text was automatically digitilised. The Cortes had also taken on digitilisation of reports of debates. She thought it was a good idea to include images on the system.

Mr Bas NIEUWENHUIZEN (Netherlands) asked two questions. The first was how the project was managed and the second was how many staff were involved in updating information.

Mr Everhard VOSS (Germany) made three points. He noted that Mrs SA CARVALHO hoped the project would bring citizens closer together but asked what provision was being made for older, less computer-literate people. The second question was whether the system took into account the fast changes in the IT world and he asked how long the scheme would last. The third question was to what extent it would be possible in African Portuguese speaking countries to take advantage of this system.

Mrs Stavroula VASSILOUNI (Greece) said that the Greek Parliament had also taken similar work. A private company was undertaking it under the direction of the Greek Parliament. She asked what plans there were to extend this to other work and whether all work would be put on open access on the Web. She also asked how much would be translated into foreign languages.

Mrs Adelina SÁ CARVALHO, President, replying to the questions, said that in reply to the Canadian member the Portuguese used the same system as the Spanish Parliament which was very expensive. It took a year to find out how many people actually bought the document. There was no attempt to replace this by public access on the Internet. Professor PINTO commented that

the photographs were high resolution on the system but low resolution on the Internet. As far as the treatment of text was concerned it was possible to copy or select text and put it on a notepad.

Mrs Adelina SÁ CARVALHO, President, in replying to her Dutch colleague said that there was an agreement between Parliament and the universities by which universities could be asked to report on specific subjects. It was even possible to select the particular people who would do it. This was done by the Speaker on the suggestion of the Secretary General. This decision was very important and it was equally important that the university respected the time limit set. There were about twenty staff devoted to the project. Access to archives was becoming more and more electronically based and so there was less need for access to books and less photocopying.

Mr VOSS had put an interesting question and obviously there was a generational difference in the use of electronic equipment but she noted that many people who were considerably older than she was had become very expert on computers. It was still open to people who did not like computers to use the books. It was just that there was more chance for people to get what they wanted if there different forms of access. She thought that Portuguese universities were the best organisations for supporting this system because a system which involved hiring outside providers would mean that they would quickly become like civil bureaucrats. She thought that if it was required to have advanced level work, it was necessary to go either to universities or to business. She noted that Portuguese Africa was catered for by helping them preserve their historic documents

Turning to Mrs VASSILOUNI's question, she said that other plans were being developed although there were budgetary problems. As far as translation was concerned, it would cost a great deal of money to translate all the documents. There was, however, translation into English, French and Spanish of the website only. Anything further would be too expensive. CD Roms were also being produced.

Mr Ian HARRIS, Vice-President, thanked Mrs SA CARVALHO and Professors MARTINS and SOUSA PINTO.