Rapport de recherche CEDRIC #99-09

A WWW-based digital library for antiquarian collections

Pierre H. Cubaud, Alexandre Topol
Centre d'Etudes et de Recherche en Informatique (CEDRIC)
Conservatoire National des Arts et Métiers (CNAM)
292 rue St-Martin, 75003 Paris, France.
{cubaud,topol}@cnam.fr

1. Introduction

Digital library technologies have benefited during the last years from the impressive increase of digital data capture, storage and transmission capabilities along with the consequent fall of their cost [7]. The widespread use of the World-wide Web (WWW) also enables digital libraries to meet a very large, international, population. Relying on digital resources for academic researches is now a reality for many scientists, but it is usually restricted to journals, acts or bibliographic databases. For humanistic studies, a step forward is the digitalization of the historical sources themselves. In France, the digitalization project of the Bibliothèque Nationale de France (BNF) opened this trend. The BNF WWW server (http://gallica.bnf.fr) has demonstrated the feasibility of diffusing a large, image-based, digital repository over the internet and time has come now for smaller academic institutions to undertake a digitalization program and a public diffusion of a part of their own antiquarian collections.

Started in Jan. 1998, the "Conservatoire Numérique des Arts et Métiers" is an internal, self-funded project of three CNAM services : the library (which holds a very important collection for French history of science and technology), the research center for technical history (CDHT) and the computer science research laboratory (CEDRIC). The editorial team also includes history of science specialists from other european institutions. We shall only focus in this paper on the information system aspects of the project. Further reports with a wider point of view are expected in the forthcoming months.

We describe in part 2 the digitalization process and discuss the various possibilities for the images delivery. Considerable attention to this issue has already been given (see for instance [13] and [10] for a not-too-outdated survey of still images compression). However, we were not able to find in the literature any survey for the behaviour of compression methods over large corpus of images such as ours. The architecture of the WWW service is described in part 3. The proposed architecture is strongly inspired by the previous experience we have gained in the management of the "Bibliothèque Universelle" (http://cedric.cnam.fr/ABU/), a WWW-based, text-only, repository of public domain French classics. Its corpus is developed on a voluntary basis by internet users, grouped in the ABU non-profit organization. The WWW service itself was developed and is hosted by the CEDRIC laboratory since Sept. 1993. A description of the architecture of the WWW service is given in [4], along with an analysis of its use for the 1996-98 period. We shall first recall the main arguments of [4] that are relevant to the design of the new WWW service. We then describe the current state of development of the Conservatoire Numérique. Finally, we suggest some future directions of studies.

2. The digital collection

2.1. Building the digital collection

For our first numerization batch, the editorial team of the project has selected 42 titles (55 volumes) related to electricity theories in the XVIIIth and early XIXth centuries. Although the books were chosen mainly for their importance to historical studies, we have voluntarily narrowed the choice to medium sized (octavo and small quarto) and well printed volumes. Since most of the volumes are unique copies in the CNAM library, it was decided that the digitalization should be non-destructive, i.e. that the bindings should not be damaged. The books all belong to the CNAM antiquarian collection and only one missing plate within a volume was borrowed from the Institut de France library.

Summaries, indexes and plates tables are the only parts of the books that we converted into textual files. Technical books of the period under consideration usually have very detailed tables and such information converted into textual files certainly ease the consultation of an image-only digital library. This (painful) work was performed by the editorial team. No OCR package has been used since we choose to translate the texts into modern french. Scientists names were also uniformized. The team also established an index of the various names transcriptions found in the books tables.

The numerization process itself has been conducted by an independent contractor. Since no color plates nor texts appear in the selected books, the scanning is bitonal, with a 400 PPI resolution. In order to reduce the processing cost, consecutive left and right pages are scanned together, except for plates and oversized tables.

Each image is labelled with some descriptive information, stored within the corresponding filename and separated with dots. Description fields are :
- the file unique identifier within the volume (4 figures),
- the title reference, using the CNAM library convention : folio size, collection, title number,
- volume number, for titles with multiple volumes,
- the image type (P=Plate, U=Unique page, W=Double page),
- the page or plate number(s). Page and plate numbers are typed "as is", since it may be letters (roman figures or alphabetic notation) instead of arabic numbers.

The contractor delivered the 11087 images in the TIFF format using the ITU-T Group 4 (T6) code. The total storage needed is 2.6 GB and this fits into 5 CDROM. Errors inevitably occurred during the numerization process. These errors can be summarized as follows :
- corrupted file (1 image)
- image crop too large (7 images)
- image crop too small, but no text missing (10 images)
- incorrect syntax for image label (2 images)
- bad title reference (2 vol., 302 images)
- bad volume number (6 images)
- bad image type (9 images)
- incorrect syntax of page number (15 images)

Page numbers remain to be checked at the time of this writing, but their proper incrementation has been verified. The low defect ratio (3%) is comparable to the sample defect ratio used for batch rejection during the BNF digitalization program [3]. It could have been reduced substantially by more systematic use of verification scripts by the contractor. One should take great care in deciding how many descriptive fields should be added to the image. Simple probability calculus shows that if errors within fields are assumed to be independent Bernouilli random variables, then the probability of an image label being wrong tends to 1 exponentially with the number of fields.

Since most of the files include two pages, we had to cut them in two different files. A first attempt using the half width of the images proved to be unsatisfactory because the book pages were not always kept centered during the scanning process. We therefore developed a more flexible procedure based on pixels vertical projection, a well-known technique for texte/image segmentation [11,6]. Most of the images contain a shadow in the middle because the books cannot be totally opened. This is clearly visible on the example of fig. 1. Locating this shadow within the text columns is eased by the fact that most of the books we are dealing with have a very simple typographic structure (fig. 2). Once the shadow position is known, one can choose to split it equally into the left and the right page. The shadow itself could also be eliminated. However this erasing cannot be performed crudely by changing all the pixels values within the shadow bounding rectangle, for it only substitutes a grey, irregular, shadow into a all-white regular shape. A simple method would be to reduce grey intensities inside the shadow bounding rectangle to an arbitrary ratio (1/10, for instance) of the average grey intensity of the text columns. Another consequence of the partial opening of books during scanning is that the pages appear slightly rotated. The

rotation angle seems to be small for our image collection. It is never above 3° and the example shown in fig. 1 is among the worst cases. However, the continuous change of rotation angle may be felt annoying when browsing through the book pages. To correct this, one can apply a rotation of positive angle for pages on the left side, and a contrary rotation for the right side page (fig 3).

We must admit that all these issues are somewhat out-dated since today's professional book scanning equipments are able to provide on-the-fly page curve correction, automatic centering, page center erasing and image rotation. However, scanned images using older technologies could be improved using segmentation and a public domain, GNU-like, software tool would certainly be useful to the academic community.

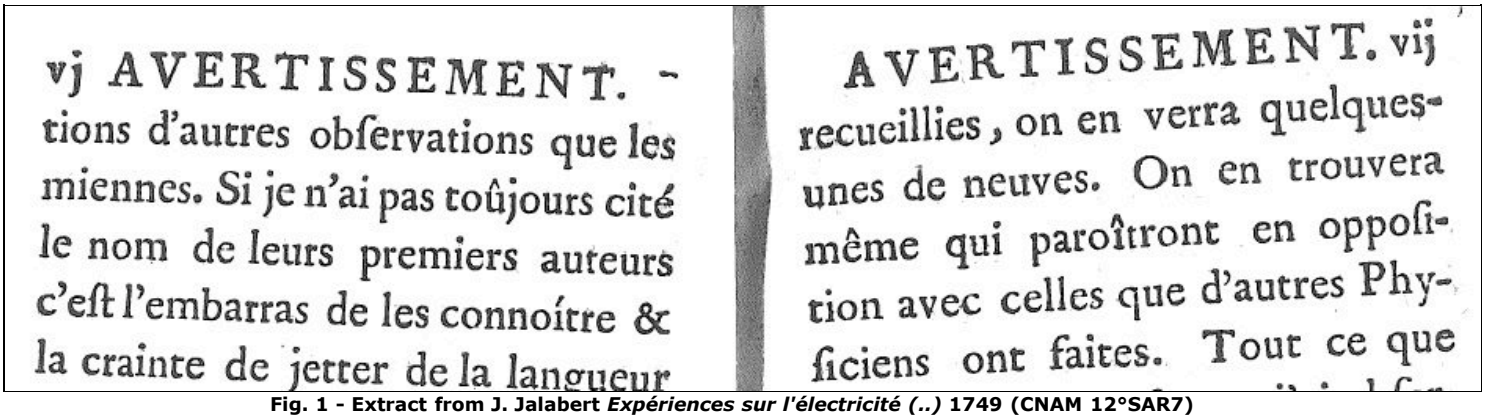

Fig. 1 - Extract from J. Jalabert *Expériences sur l'électricité (..)* 1749 (CNAM 12°SAR7)

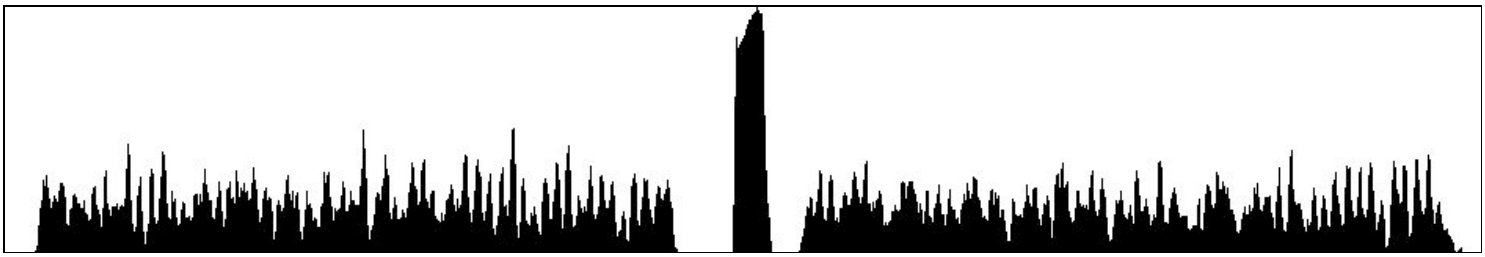

Fig. 2 - Image segmentation using vertical pixel projection

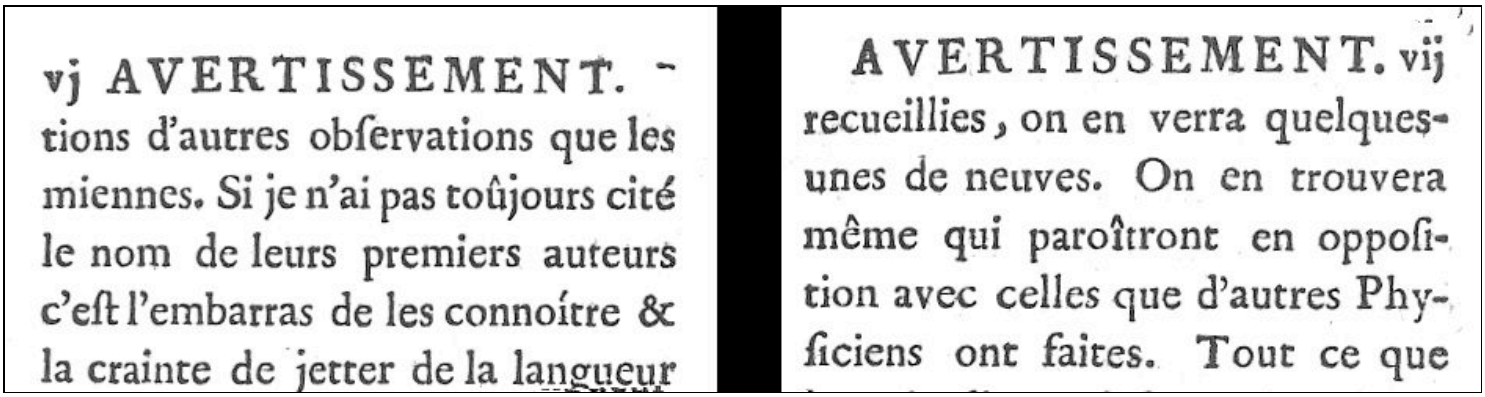

Fig. 3 - Text rotated (left : +1.6°, right : -3.0°) and binding margin removed

2.2. Image format for WWW diffusion

Since today's usual screen resolution is 72 PPI, our 400 PPI images appear 5 time bigger than the original book dimension when displayed on screen. The average image size is 5116 x 4212 pixels (see table 1 below) and this is far superior than what graphic hardware currently handles (1200 x 1000 pixels) so the image cannot be shown at once. Viewing softwares therefore reduce the image resolution and solve the subsequent aliasing problem by increasing its color set size. Deciding whether this resolution decrease should be conducted once and for all on the WWW site or by the user agent, at user will, is a major design issue. The first solution is certainly the most rational, since otherwise the corresponding computations are reproduced over and over by each user agent. It is also a straightforward solution to the difficult problem of uncontrolled rediffusion of images under property rights. Another interesting point is that, for any value less than 2^16 grey levels we can expect the resulting file to be smaller than the original, if the same compression method is used for both. This will result in smaller storage needs and transmission times. On the other hand, one cannot guarantee that the chosen resolution will fit the viewing conditions of all users. Another advantage of using bitonal pictures is the absence of Gamma conversion for the user's viewing equipment. A good solution might be a compromise between the two approaches. For our collection of bitonal 400 PPI images of book pages, we found that quite satisfactory results could be obtained by reducing the resolution to 100 PPI, with 8 grey levels. Reducing the resolution below 100 PPI is in general impossible for the proper reproduction of small typographical signs and engravings that typically appear in XVIIIth century scientific books.

For the diffusion of the images via the WWW, or any other network-based application, one must also take into account two other factors :
- The highly variable end-to-end throughput of the internet "at large", which typically ranges from 1 to 100 KB/s. This favors highly compressing formats, but one should also take into account the time to uncompress the file and actually display the image on screen. Progressive compression allows low resolution versions of the same image into one compressed stream.

- The practical availability of decompression and visualization softwares on a wide range of computer platforms. Relying on patented, commercial, software is clearly a difficulty for WWW services such as our, which is intended for a large class of anonymous users.

In order to choose a proper file format for the Conservatoire Numérique, we have considered TIFF-group 4, GIF, PNG and the recent DjVU format from ATT laboratories [1]. JPEG (JFIF) has been discarded since this compression method has been primarily designed for continuous-tone images. The other TIFF compression schemes were also excluded since they are equivalent to those used by GIF and PNG (i.e., the LZW dictionary method). It should be noted that TIFF-group 4 is not handled by the most popular WWW user agents (and in fact, by any browser we are aware of). Commercial plug-in are however available, and group 4 files can also be embedded into Adobe PDF documents, for which a reader software is available for free from Adobe. For DjVU, a plug-in for Windows, MacOS and UNIX is freely available. A UNIX version of the DjVU compressor is also freely available for non-commercial use (non-commercial use includes the diffusion by academic institutions of historical material).

The table 1 summarizes the results for the conversion of our 11087 images in the formats under consideration. Fig 4 and 5 show the histograms for file sizes and compression ratio respectively. All the conversions were performed using the netpbm package (vers. 1mar94.p1-30). The compression factor is defined as the ratio of the original raw data over the compressed result file. The DjVU files are 200 PPI images with 5 greyscales, GIF and PNG are 100 PPI with 16 greyscales (8 grey levels aren't supported by PNG). The large dispersion of TIFF compression factor may be explained by the underlying compression method (RLE with Huffman coding). It is clearly a disadvantage since it will result in an increase of the mean response time for the WWW user (a classical argument in Queuing theory). For our image batch, the GIF and PNG formats have almost equivalent compression ratio, and the claim of PNG superiority for that matter [8] should be minimized. It was felt difficult to choose between GIF and PNG. Both of them are widely supported and GIF is now in the public domain. PNG format handles Gamma information so that grey values appear identical on any type of screen. On the other hand, GIF files with 8 grey levels are always smaller than PNG with 16 grey levels. For that reason, we have decided to retain the GIF format. The total storage size is 1.24 GB for 20864 images (double pages are cut), so it represents half of the original TIFF-group 4 storage size. Mean file size is 62 KB. Average storage for one volume is 24 MB (comparable to the 29.4 MB required for the downloading of Microsoft Internet Explorer 5.0 or many other softwares)

According to [1] and the ATT online FAQ (http://djvu.research.att.com/wid/faq.html), the DjVu image compression technology is specifically designed for scanned document pages and integrates three interesting characteristics :
- a very good compression ratio due to the JBIG2 algorithm for bitonal documents. Indeed we notice in the Table 1 that the average compression factor for DjVu images is by far the best with similar subjective quality,
- a segmentation algorithm that splits the image into a background and a foreground plane and therefore can be used to reduce the scanning noise,
- the DjVU plug-in doesn't need to hold the entire decompressed image in memory.
However, one must admit that DjVu is today a young technology and the tools for using it are still in a prototypal state. For instance, djvuencode ends inexplicably with a segmentation fault for some ordinary files such as pp 210-1 of Sigaut de La Fond "traité de l'électricité" and the ppmtodjvu converter did not work with 4 grey levels (we used DjVU 1.1.5 for Linux/intel (2.0.x) libc.so.5). DjVU was primarily designed for color images and maybe its use for grey pictures has been considered secondary by the DjVU developers. The viewer plug-in is also not easy to install and such an operation can not be asked to average WWW users. Nevertheless, we have decided that we shall provide a DjVU version concurrently to GIF on the Web server, in order to encourage this initiative and analyze the users reactions.

One can see on fig. 5 that some images have a very modest compression factor. Fig. 6 shows a portion of a typical example (the phenomena occurs only with heavily engraved plates). The compression factor for this picture is 3 time smaller than the average for TIFF, GIF and PNG and 6 time smaller for DjVU.

	Width	Height	TIFF	DjVu	GIF	PNG
mean	5116	4212	251	37	155	147
std.dev.	1149	410	126	14	49	47
coef.var.	0.22	0.09	0.5	0.4	0.3	0.3
mean compression factor :			10.5	88.9	4.2	4.5

Table 1 - Images dimension, files size (KB) and compression ratio

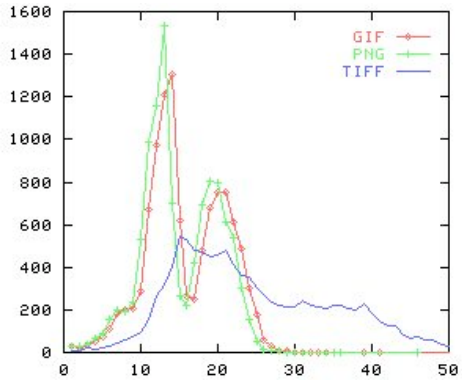

Fig. 4 - Histogram of file sizes (unit = 10KB) for GIF, PNG and TIFF-g4 formats

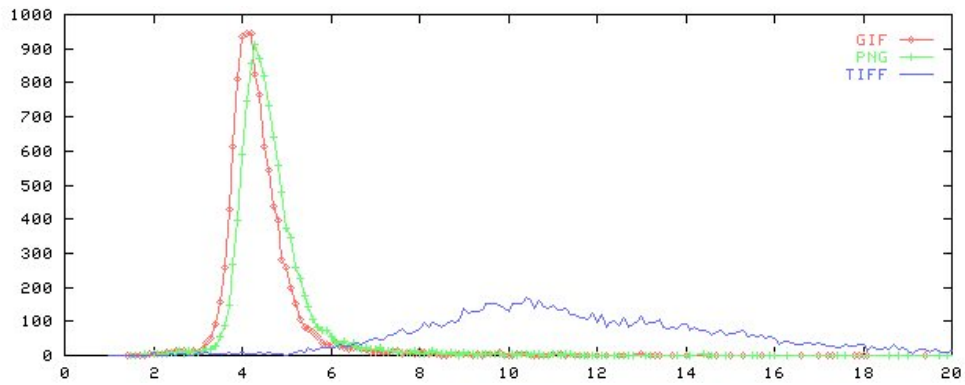

Fig. 5 - Histogram of compression factor for GIF, PNG and TIFF-g4 formats

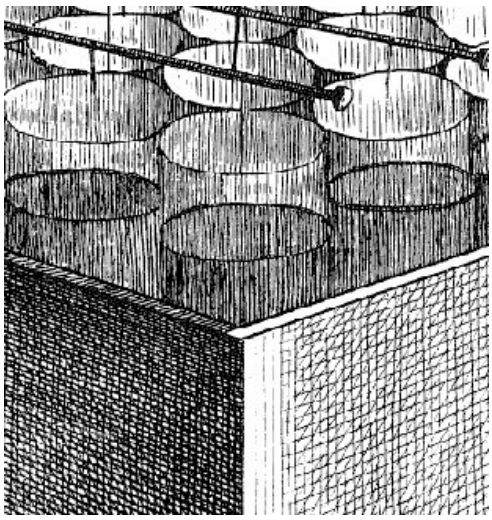

Fig. 6 - Hardly compressable engraving - from Priestley *Histoire de l'électricité*, 1771 (CNAM 12°SAR6.3)

3. The WWW service

3.1. Usage analysis of the ABU service

Any new information system is (or should be) preceded by the analysis of the users needs. This difficult task can be simplified when one can rely on passed experience. During the design phase of the Conservatoire Numérique, we made the assumption that our hypothetical users would behave similarly to those connecting today to ABU. This should help to forecast the server workload and the kind of interaction system we could provide in order to satisfy the largest class of users. WWW users behavior is generally analyzed with the help of the HTTP server log files. Log files generally include the date, type of request, the requested URL and the user machine domain name. Of course, the connection-less nature of the HTTP protocol and the absence of user authentication limit the strength of such kind of analysis. One should also take into account the unstable nature of the internet population, which grows exponentially since years. In [4], we found that a linear growth model matched quite well the observed evolution of the workload.

During our study of the 1996-98 period, we found that about 80% of the ABU users machines had an identifiable primary DNS code. The FR domain represents 35% of the sites and almost half of the requests. The other half is scattered unevenly between 98 primary domains. 12% of the requests were due to jumps from another WWW service. ABU is referenced in more than a thousand WWW pages, but most of the traffic comes from search services (Yahoo) and specialized catalogs of online text archives (Athena, Clicnet, Gallica). The ABU home page represents 27% of the requests for static HTML documents. Access to the texts and authors catalogs represents also 27%. We were rather surprised by the very low amount (4%) of requests for the information pages that describe the textual corpus and the ABU project. It is however a general and recognized behavior of on-line users not to visit a server as a whole. Emails sent to the webmaster often suggest that the online information was unknown to the sender. There is also a significant disparity between the various commands : text browsing is by far the most used (69%). Full text downloading is 8%. Keywords search within the whole corpus, or within an author corpus amount to 11%. An average user site performs 11 HTTP requests, 4 of which are full-text downloads. These values are remarkably stable over the period under study, but there is a great dispersion between users and we believe that a Pareto-like concentration law applies here. Making sure that users will find their way within an on-line service is still an experimental art (interesting guidelines may be found in [12]). For instance, the ratio of downloadings versus page browsing requests significantly increased (from 10% to 17%) since the webmaster changed in Feb. 1999 the download command name from "copie" (copy) to "texte complet" (full-text)...

For the month of March 1999, 355753 HTTP requests have been processed on the ABU server. 131285 were requests for browsing through a text of the corpus and there has been 23204 requests for full-text download. The corresponding average throughput was 22 Kb/s (a very small portion of the CNAM 2 Mb/s link) One can compute the effect of similar workload on the Conservatoire Numérique if we assume now that each page is a 62 KB GIF file instead of some HTML text and no download is allowed. The average throughput is almost unchanged (24 Kb/s). If we consider (in the worst case) each download to be now a 24 MB archive , the average throughput becomes 1.2 Mb/s. These figures are very reasonable for today's network technology. The platform used is a rather out-dated 40 Mhz SUN Sparc System 600. This is not however a critical issue since on the average, the ABU server is executing an HTTP request every 8s. Most HTTP requests activate Perl cgi scripts that execute in less than 2s. For these reasons, we believe that a standard personal computer can easily handle the forecasted workload of our digital library. The prototypal platform is a 300 Mhz Pentium II with 4 GB SCSI running Linux and the Apache HTTP server.

The "agent_log" file supplied by the HTTP server can give us hints concerning the expected users equipment. We use again the statistics gathered for the month of march 1999. One surprising fact is the great variety of user agents : 2533 different types of user agents have contacted the ABU service during the month under consideration. A user agent is supposed different from an other if, of course, the manufacturers are different but also if the version numbers, the operating systems or the localization differ. It is clear that each new version comes with some new features (and bugs) and therefore has a different behaviour. About 85% of the requests were issued by "Mozillas" user agents such as Netscape Navigator and Microsoft Internet Explorer. Non-Mozilla agents are mostly robots, or non-graphical (Lynx). Table 2 summarizes the number of HTTP requests executed by the various types of Mozilla user agents. We can infer from this table that, at the time of this writing, about one third of the ABU users don't have the ability to access to HTML frames (introduced in version 3) or Java 1.1 applets, which appeared in version 4. The great variety of operating systems (table 3) also suggests a potential source of difficulties when one relies on non-HTML features that require plug-in or helpers extensions. We then believe that the first version of our WWW service should only use the transitional HTML 4.0 DTD.

Major Version	Total	IE	Netscape	Others
5	3	0	3	0
4	228506	131575	96612	319
3	47144	4997	40209	1938
2	43328	36316	6473	539
1	490	344	141	5
0	12	0	12	0

unknown	1	0	1	0

Table 2 - Requests classified by user agent identifier

Platform	Requests
win95	161021
win98	62494
winNT	35958
win32	2267
win16	6031
Mac OS	21765
winCE	16
webTV	735
memoWeb	3054
X11	14734
OS/2	327
Others & undefined	8281

Table 3 - Requests classified by user agent platform

3.2. Server organization and user interface

The navigation within the Conservatoire Numérique can be viewed as a mixture between the navigation model of ABU and that of Gallica. Navigation within Gallica is based on HTML frames, whereas ABU is best used with multiple windows (most user agent provide a shortcut for opening new windows for selected links at user will). We have tried to rely as much as possible on the user agents functions for navigation history, bookmarking and annotation (a useful feature of the late Mosaic).

After the home page, the user can choose between a catalog of primary and secondary authors, a catalog of texts (classified by primary author names), or a global keyword search within all the textual corpus (books summaries and indexes). For each title, one can access either directly to the digitized pages, to the textual summary or its image version. The navigation within the book pages can be performed in many ways. One can use the textual summary, browse sequentially through the images, jump from plates to plates, if any, or specify a page (or plate) number. Jumping between the various sections of the book without using the summary has not been implemented yet. Fig. 7 reproduces a screen dump of a work session with the Conservatoire Numérique.

As with ABU, three types of HTML documents are used : the general information pages are hand-written, but the catalog and authors pages are automatically generated by the site administrator when the corpus is revised. All other pages are dynamically generated by users requests. Authors and volumes description, volumes summaries are mark-up textual files . Images and images descriptive information are simply organized within directories into the file system. The linux file system can handle 4 TB of data and roughly one thousand of elements within each directory.

The visualization of the book pages is of course the most critical aspect of the WWW service. One cannot be sure that all the 100 PPI GIF images will fit the user screen and we would like to be able to navigate easily between more than one page window on the screen. The only solution is to provide zoom functions to the user. The HTML DTD provides an interesting feature inside the IMG element that tiles the image size to a fixed ratio of the containing window. Resizing the window therefore generates an image resize and this is performed very efficiently on most platforms. However, it seems that this mechanism doesn't coexist with tables within a HTML document. Another solution is to specify the IMG width and height to specific values, computed within the script that generates the HTML page. One can rely on the user agent documents cache so that only the HTML text is requested by the user agent when the user performs a zooming action. The cache can be used to provide some kind of progressive downloading when the HTML page contains two images : one being currently read by the user while the next one is being downloaded. This is particularly useful for users with slow dial-up access. A large document cache on the user side also significantly improves the navigation within the passed pages.

For the DjVU images, the interface is modified since it is based on a plug-in that already offers zooming and scrolling functions. We haven't find any elegant way to hide the choice of image format to the end user. It should be noted that HTML 4.0 offers a mechanism for specifying alternate object formats, depending upon the user agent rendering ability. The example below is adapted from the HTML 4.0 specification [9]. It is not fully usable here, since the entire HTML document has to be different.

```
<!-- First, try the DjVU format -->

<OBJECT data="myimage.djvu" type="image/djvu">

 <!-- Else, try the GIF format -->

 <OBJECT data="myimage.gif" type="image/gif">

  <!-- Else render the text -->

  A short textual description of image content.

 </OBJECT>

</OBJECT>
```

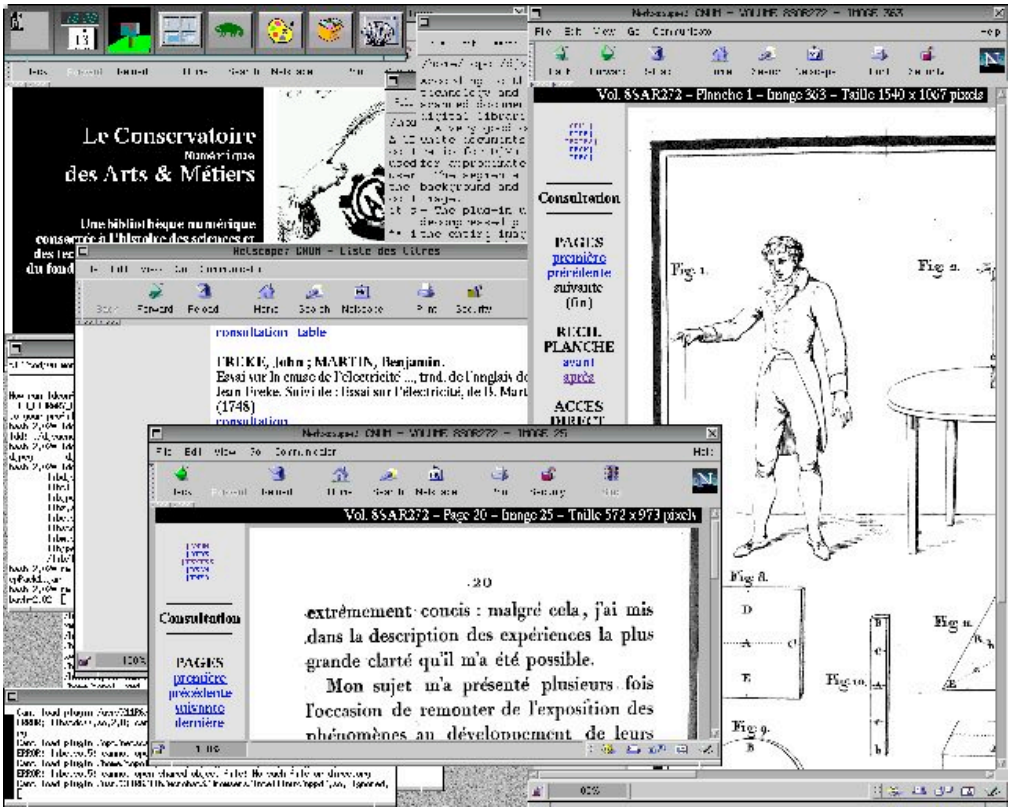

Fig. 7 - The WWW service in use

4. Conclusion

Many desirable features are currently lacking in the Conservatoire Numérique, in order to facilitate the navigation between the documents. Plates for instance could be precisely described, so that it would be possible to find the figures they contains more easily. Maybe a generalized document annotation and cross-referencing mechanism would be necessary for the progressive "textualization" of the repository. A participation of the users in that process would certainly be very interesting. Also missing are downloading and authoring tools, either for extracts of pages (with cut-and-paste functions, for instance), for sections of books or maybe entire volumes. It is regrettable that the annotations functions of Mosaic have not been fully developed within today's commercial user agents.

A step beyond would be the study of interfaces that take advantage of the high throughput of new network infrastructures and low-cost 3D graphic hardware on users platforms. Virtual reality interfaces can greatly improve the user's navigation within a digital collection [5]. The textual summaries could be used to generate "on-the-fly" 3D organizations of the books based on user requests. Books themselves may be translated into 3D metaphors (like the Xerox PARC WebBook experiment [2]). Collaborative work between library patrons (or staff) could be incorporated into this framework.

We believe however that these functionalities should be progressively added, following the users ability and needs. A clear distinction should be made between a minimal, largely usable, service and more advanced, prototypal features. The growth of the corpus should also be considered. A thousand books (i.e. 24 GB) is certainly within reach for the current platform (although maybe not within the CNAM digitalization budget). A network of a few hundred of such modest systems would constitute a powerful, fully distributed, inter-institutions digital library and a cost effective alternative to the centralized architectures favored in the passed years. Disseminating the knowledge of such technology and editorial responsibilities into the local level would certainly benefit to the academic community as a whole.

References

[1] L. Bottou, P. Haffner, P.G. Howard, P. Simard, Y. Bengio, Y. Le Cun. "High Quality Document Image Compression with DjVu". Journal of Electronic Imaging, vol 7, no 3, pp. 410-425, 1988.
http://www.research.att.com/~leonb/DJVU/adl98/ (DjVu format)
http://www.research.att.com/~leonb/PS/haykin.ps.gz (gziped postscript format)

[2] S. K. Card, G. G. Robertson, and W. York, "The WebBook and the Web Forager: An Information Workspace for the World-Wide Web," in Proceedings of CHI '96, ACM Conference on Human Factors in Software, 1996.

[3] G. Cathaly. Interview (June 1998) and "Numérisation - procédures de travail v.2" EpBF internal report, sept. 1993.

[4] P. Cubaud, D. Girard. "ABU : une bibliothèque numérique et son public", Document numérique, vol. 2, no. 3/4, pp. 13-30, 1998.

[5] P. Cubaud, C. Thiria, A. Topol. "Experimenting a 3D interface for the access to a digital library", Proc. of the third ACM conf. on Digital Libraries, Pittsburgh, June 1998, pp. 281-282.

[6] D.G. Elliman and I.T. Lancaster. "A review of segmentation and contextual analysis techniques for text recognition", Pattern Recognition, vol. 23, no. 3/4, pp. 337-346, 1990.

[7] M. Lesk. Practical digital libraries - Books, bytes and bucks. Morgan Kaufmann. 1996.

[8] PNG (Portable Network Graphics) Specification. Version 1.0. W3C Recommendation 01-October-1996 http://www.w3.org/TR/png.html

[9] D. Raggett, A. Le Hors, I. Jacobs. "HTML 4.0 Specification, W3C Recommendation", April 1998. http://www.w3.org/TR/REC-html40/

[10] D. Salomon. Data compression, the complete reference. Springer Verlag. 1997.

[11] S.N. Srihari. "Document image understanding", Proceedings of ACM-IEEE C/S Fall Joint Computer Conference, Dallas, November 1986, pp. 87-96.
http://www.cedar.buffalo.edu/Publications/TechReps/DocImage86/paper86.html

[12] R. Tennant. "The Art and Science of Web Server Management", Untangling the Web UCSB conf., April 1996. http://www.library.ucsb.edu/untangle/

[13] I.W. Witten et al. Managing gygabytes: compressing and indexing documents and images. van Nostrand Reinhold, 1994.