The Digital Processing of Images in Archives and Libraries Large-scale International Projects

Pedro González

The enormous possibilities which the new image technologies offer in the world of documentation, principally in improving access to information, have scarcely begun to materialize into large-scale document digitalization projects. This is perhaps because these new technologies have not yet matured, or adequate standards do not exist, or because they are unknown to professionals, or are very costly for the normally scarce resources of archives, libraries, museums and research in general.

However, it is clear that the new image and optical media technologies offer positive alternatives to previous reprographic systems:

- The digital image may be reproduced to infinity without any type of loss in quality.
- The digital image makes it easier to obtain copies in any other type of medium (paper, microfilm or computing media).
- above all, the digital image offers extraordinary possibilities re "access" to information and distribution (much greater speed in retrieval, the possibility of copies in paper or magnetic or optical media, electronic editions, the use of communication networks etc.).

There is a growing interest on the part of Archives, Library and Museum specialists in the roads opened up by the new image technologies and information storage systems. However, these professionals find themselves lost in the sea of existing products on offer, without the necessary information or knowing whether what they are being sold is as interesting as it seems or in proportion to its cost.

1. Large Scale Digitalization Projects in Archives

In fact, the new image technologies are being used in a multitude of projects being undertaken in various different centres (archives, libraries, museums), but up till now very little has been done on a massive scale: in most cases these are very limited approaches and are of little interest.

Turning specifically to the area of Archives, it is a fact that some people, encouraged by the increase in the market of multiple systems at relatively low cost, have been initiating or are already working on document digitalization and image storage in optical disk in specific areas. Simple systems made up of a data entry station (microcomputer with scanner), a user workstation (microcomputer with high resolution screen) connected to a data base and optical disk server, have been acquired by some centres for experimental use. However, in general, scarcely the very first steps have been taken, and the range and number of these experiments is very small.

As regards large scale projects, apart from the Computerization of the "Archivo General de Indias" in Spain, we scarcely have any information on any projects in Archives and Libraries which, add digital treatment of large volumes of documentation to pilot projects: some years ago now, the National Archives and Records Administration (N.A.R.A.) in Washington launched a project: "Optical Digital Image Storage System" (ODISS), which

it recently published the results on. We have information of an ambitious projects undertaken in the Municipal Archives of Utrecht (the ARIS project), although according to some recent information, which we have not confirmed, this project has been temporarily abandoned[1]. Nevertheless, in another area — Libraries — the U.S.A.'s Preservation and Access Commission has designed an important project for the preservation of "brittle books."

The ODISS Project (Optical Digital Image Storage System)

The programme started in 1984 and the system was installed in July 1988, in the NARA (National Archives and Records Administration) in Washington. Its aim was to try out the utility of digital image and optical disk technologies for the reproduction, storage and retrieval of archival documents. In March 1991 the report on this programme came out, and in July a draft of the Guide to Digital Image Applications for State and Local Government Agencies was prepared.[2]

The pilot project was undertaken on a series of personal records of a military character: (CMSR), which measures a total of approximately 400 cubic feet, although during the project 220,000 pages were digitalized, some 80 cubic feet.

The resulting system is made up of three subsystems:

1. The Conversion subsystem, which includes the digitalization of original documents with their corresponding indexing and quality control, which creates files of digital images, kept initially in magnetic disk.
   The previously prepared documents pass through a high speed scanner (Photomatrix), able to digitalize in black and white at 200 dpi and 20 pages per minute, using an automatic paper feed and CCITT Group III compression algorithm, controlled by a Unisys-200 microcomputer with a 80286 processor.
   Then quality control is carried out, marking the images which have to be re-digitalized, followed by indexing. This descriptive information is incorporated into the data base, by means of Unify relational model software.
   Re-digitalization is done with a low speed scanner. In fact, two are used, one in black and white and another with 256 levels of grey, until acceptable images are obtained. The low velocity black and white scanner is an IS-400 Ricoh, in flat format, able to digitalize at 200, 300 and 400 dpi For documents which require the use of grey levels, a Xerox Inca-38 is used, capable of working at up to 400 dpi and 256 grey levels.
   A scanner, able to digitalize microfilm, microfiche and window cards of 16 and 35 mm, manufactured by Photomatrix Corporation, is also used.
2. The storage subsystem: transferral of digital images to optical disk for long term preservation.

[1] WANG. *Archive Information System (ARIS)*. Culemborg, 1990.

[2] National Archives and Records Administration (NARA): "Optical Digital Simage Storage System. Project Report". Washington, 1991; National Archives and Records Administration (NARA): "Digital Image Application and Optical Media Systems: Management issues, technical trends, user experience. Guidelines for State and local Gobernment Agencies. A Joint Report by National Archives and Records Administration and National Association of Government Archives and Records Administrators". Washington, 1991.

When the digitalization of a block of documents has been completed, this block is transferred to an optical WORM disk (Sony, 12in), and a backup copy of it is made as soon as both side are full. For the 220,000 pages 5 optical disks were used. The optical disks are handled by a juke-box, also Sony.

3. The retrieval subsystem, responsible for reference to and retrieval of filed images. Made up of two work stations (microcomputers with 286 processor) for the records staff and one more for the public, allowing access to the information contained in the data base through each one of the 13 search fields (name, regiment, company, etc.) and direct, later display of the images on the same screen at 150 dpi resolution. For printing out, a RICOH LP5400 laser printer is used, which reaches 400 dpi. From the Tennessee State Library and Archives in Nashville through a communication line of 1200/2400 bauds one can gain access to information contained in the indexes. The images may be printed out afterwards in the NARA and sent by post.

The conclusions of the report are hopeful. We may sum them up as follows:

- As regards fast digitalization of large quantities of documents, it is concluded that the mechanical transport of documents for automatic feed did not cause damage to the papers. In the case of fragile documents, no difficulties were met with when placing them in polyester sleeves.
- The substitution of digital images is easier than that of electrostatic copies or microfilm. Backup copies can be obtained which are identical to the first, without any need to re-scan.
- Legible images of the documents can be obtained and can even be significantly improved upon. A resolution of 200 dpi in 98
- It can be considered that the stability of the images stored in Sony optical disk will be adequate for 100 years, except in the case of silver salts microfilm and inferior quality paper. Repeated use does not degrade the images.
- The greatest disadvantage is the need to periodically recopy the disks in order to avoid the technology used becoming obsolete.
- Space saving will be significant when digitalization of large quantities of paper is undertaken. With quantities like the ones produced (220,000 pages) the significant saving is lost (80 cubic feet as opposed to less than one cubic foot occupied by the disks) due to the need for equipment: the juke-box occupies, in fact, more than the original paper.
- The time saved in searching for and retrieving information is significant. This can be considered as a "three-fold reduction".
- The conversion of paper documents to any other medium cannot be justified only on the basis of the cost-benefit ratio. There are also intangible benefits to be considered: the improvement in time and quality of access, the improvement in legibility, the reduction of storage space and the fact that originals no longer need to be handled etc.

2. Experience in Libraries

In the libraries field various experiments applying optical technologies to the preservation and distribution of bibliographical material have recently been undertaken. This material has specific characteristics which differentiate it from archival documentation as it is usually printed material.

One of the pioneers is the work done by the Library of Congress in Washington which, in the early eighties, started a pilot programme with two different facets: the first using optical disk (with digitalization of printed material, in general periodicals) and the second in videodisc (for non-printed material, in general photographs).[3]

The CLASS Project (College Library Access and Storage System)

A very interesting project in the library area due to the new prospects it offers for large scale digitalization, is the joint development by Cornell University and the Xerox Corporation, with the support of the Commission on Preservation and Access.[4] At the same time as researching into the subjects of image digitalization, storage and retrieval, it was attempted to develop a prototype for digitalizing damaged books — the "brittle books"[5] and producing high quality copies later in paper, subject to request from the user. This is precisely one of the basic objectives of the system: to produce paper copies of damaged books for users who request them, at great speed and at a good price.

During the course of the project, up to December 1991, 1,000 "brittle" volumes were programmed to pass through the scanner on one single work station, digitalized in grey levels, and stored in black and white at 600 dpi after compression and interpolation process.

The images resulting from digitalization done at the Olin Library are transmitted to a Xerox Docutech printer located half a mile away, to print out the digitalized pages at 600 dpi, in high quality copies, at 135 pages per minute. With this quality and speed of

---

[3] Only a few references about this pilot project: The Library of Congress: "Optical disk pilot program". Washington, 1986; Price, Josep: " The optical disk pilot program at the Library of Congress" Videodisc and Optical Disk, v. 4 (1984), n.6, p. 424-432; Hahn, Ellen Z.: "The Library of Congress optical disk pilot program. A report on the print project activities" The Library of Congress, 1983; Fleischhauer, Carl: "The Library of Congress optical disk pilot program. A report on the nonprint project activities" The Library of Congress, 1985; Parker, Elisabeth Betz: "The Library of Congress non-print optical disk pilot program". *Information Technology and Libraries*, December, 1985, p. 289-299. More recently: Krayeski, Felix P.: "Transition of an Image System: from paper to microfiche to optical disk". The Library of Congress, June 1990.

[4] This information was personally supplied by Hans Rütimann, Program Officer de la Comission on Preservation and Access, and Stuart Lynn, Vicepresident, Information Technologies, Cornell University. Completed with an article included in the Newsletter, Novenber/December 1991, of this Comission: "Update on digital techniques", by Anne R. Kenney and Linne K. Personius. This article is a summary of a more complete paper which is going to appear in 1992 in a book published by Meckler Corporation *Advances in Preservation and Access*, vol.1 By the kindness of Mr. Stuart Lynn I could also consult the draft of this paper.

[5] "Brittle books. Reports of the Comittee on Preservation and Access". Washington, Council on Library Resources, 1986.

copying there will be no difficulty in sending facsimile copies of any of the "brittle books" to any user who so requests.

In the future the digital images, stored in 12 inch optical disks, will be placed in a juke-box, and will be accessible through the Cornell University Network. Information on the digital files is also in the Cornell catalogue (NOTIS) and in the RLIN (Research Library Information Network) data base, thus providing the first point of access to this digital "library".

Direct access via screen has also been developed with a prototype user station, which shows images of the documents on an 11 x 14 inch screen at 200 dpi.

One of the main ideas analyzed by the Commission is whether optical technology represents a convenient alternative to microfilm. The conclusion was reached that the cost and time needed for digitalization (checking the image on the screen, scanning, storage, and transfer for printing) is similar to that of microfilming, though in the latter case the time for developing the film, the quality control and production of the first generation of films is included.[6]

In addition, possibilities to convert microfilm material to digital image and at the same time to convert digital image to microfilm now exist. It is calculated that it is more profitable to digitalize beforehand in order to obtain the microfilm afterwards for preservation than to do the micro filming first in order to obtain digital images for access by the users.

Lastly, within this area of converting microfilm to digital image, we have been informed that Kodak is carrying out a study on contract from the Genealogical Society of Utah, which is interested in the future possibilities of the new technologies for its millions of rolls of microfilm.

3. The Computerization Project at the "Archivo General de Indias" (AGI)

The special circumstances of Spain holding the V Centennial of the Discovery of America had contributed to the initiation and development of a project to computerize the Historical Archives, using the most advanced technologies for image processing and preservation.

Now, exactly two centuries since its foundation, the Archivo General de Indias has the opportunity to act as a pioneer amongst the world's archives. Three public and private organizations have joined forces (the Spanish Ministry of Culture, IBM Spain and the Ramón Areces Foundation) to provide the A.G.I. with a computerized information system to satisfy the needs of our times, whilst at the same time respecting and following the guidelines laid down by its Enlightenment creators.

Using various different media and modern technologies, the Project focuses on the development of a comprehensive software for processing historical documents in the Archives. To this end, the basic objectives do not only include contributing to wider and deeper knowledge of the documents we have, but also, and with similar enthusiasm, to halting

[6] Waters, Donald J.: "From Microfilm to Digital Imagery... A report of the Yale University Library to the Comission on Preservation and Access". Washington, 1991.

the rapid deterioration caused by the increased handling of original papers in the search room.[7]

3.1. General requirements

As we have said earlier, one of the main characteristics of the AGI project is that it is an "archives" project which is seen from the specific point of view of the archives. And although one of the archives main functions is the dissemination of information, it has other functions which arise from its commitment to "preserving" our heritage. This is what makes the AGI project much more than an "image/ manuscript processing" project, as is understood here, or an "optical disk" project as in the terminology used among archives professionals.

Contributing to the preservation of the original material in the Archive, as well as its greater distribution among and communication to users, are therefore the two principal objectives of the project which could be summed up under the following headings:

1. The design and development of an "integrated" computing system capable of dealing with the major part of a historical archive's functions.
2. The creation of an "Information and Reference System" containing the descriptive information on the Archivo General de Indias. The system must respect the traditional principles of archive handling (especially the principle of provenance). It must also be able to collect all the descriptive information created over the two centuries of the A.G.I.'s history.
3. The creation of an Optical Digital Storage System for the archive documents, in order to substitute consulting original papers with digital reproductions obtained on screen or in hardcopy. Within the term for completion of the project the digitalization of eight million six hundred thousand pages is planned.
4. The integration of these two basic subsystems with a third module — the module known as "User Management" — in charge of controlling researchers' access to docu-

[7] Ministerio de Cultura. Dirección General de Bellas Artes y Archivos: "Computerization Project for the Archivo General de Indias". Madrid, 1990; González, Pedro. "El Proyecto de Informatización del Archivo General de Indias", *Actas de las I Jornadas de Archivística de Euskadi*, 1990. IRARGI, III (1990), p. 261-281; González, Pedro. "Historical Documaentation and Digital conversion of images at the Proyecto de Informatización del Archivo General de Indias, Sevilla". *Microform Review*, v. 18 (1989), n. 4, p. 217-221; Rütimann, Hans and Stuart Lynn. Computerization Project of the Archivo General de Indias, Seville, Spain. A Report to the Commission on Preservation and Access. Washington, 1992. In print: Becerril, José Luis. "Computerization Project for the Archivo General de Indias". *Proceedings of the Quincentennial Conference. Archives and Records for Studying the Hispanic Experience in the United States.* Washington, 1987, ed. by José Luis Becerril, Miguel Latasa and Margarita Vázquez de Parga.; González, Pedro. "Fuentes archivísticas y reproducción de documentos en España" *Proceedings of the Quincentennial Conference. Archives and Records for Studying the Hispanic Experience in the United States.* Washington, 1987; González, Pedro. "Computerization Project for the Archivo General de Indias". *Proceedings of the "International Conference: Archiving and disseminating historical machine-readable data"* Leiden, Nederlands, 1990.

ments (originals and digitalized); monitoring the reading room; monitoring the movement of documents within the Archive etc.

5. The system designed must be transportable to other historical archives, since the Spanish Ministry of Culture sees it as a "pilot project" for the computerization of the rest of the State Historical Archives, each one with their own characteristics and above all with their differences in the volume of the documentation which they hold.
6. The system designed must be easily utilizable for the end "user" of the historical archives; the archivist who continues his work of organizing, describing and distributing document files; and the researcher who comes to the reading rooms or who requests reproduction of documents.

3.2. Present state of the Project

It is with these basic objectives that we have been working for the last 6 years, up to the point of now being able to offer the significant result being commenting upon, and which, in summary, at the present date of January 1992, is as follows:

1. All the development of the integrated computing system has been completed with its different modules, and tests are being carried out to make improvements before the project is concluded.
2. The initial loading of the data base which makes up the Subsystem of Information and Reference has been done, including:
   - all the "finding aids" (inventories, catalogues and indexes) existing up to the present date in the Archivo General de Indias.
   - the whole "description" done within the framework of the project, which includes going into more depth in the description of the documentary series to be digitalized.
3. Seven million pages have been digitalized and stored in optical disk, and work is continuing at a rate of more than 250,000 pages per month in order to achieve the final goal (eight million six hundred thousand pages at the end of '92.)
4. In the Archivo General de Indias a prototype of the system has existed for some time, at the disposal of researchers, with a part of the Information and Reference System and the Optical Digital Image System.
5. From June 1988 a first non-integrated version has been functioning in the A.G.I. with the rest of the User Management modules.
6. The prototype is at present being substituted by the final system, including the complete Information and Reference System and the second integrated version of User Management. It is planned that in summer of this year the compete System will be fully functioning in the Archive.
7. On the Ministerio de Cultura's part a four-year plan has been worked out for the transfer and installation of the system designed in the rest of the State Historical Archives ("Archivo Histórico Nacional" in Madrid, the "Archivo General de Simancas", the "Archivo de la Real Cancillería in Valladollid", the "Archivo de la Corona de Aragón" in Barcelona, and the "Archivo de la Guerra Civil" in Salamanca).
8. The system which has been designed — the property of the three institutions which have participated in its development — is available for installation in other Archives

and cedes its rights free of charge on a non-profit-making basis to the institutions who so wish.

9. The preparation of a CD-ROM is under way, which will contain a selection of documents from the AGI and a small proportion of the possibilities of the system for access to information and image handling.
10. Tests on access through data communication lines are being carried out. An important experience, albeit specific to one area, was begun in February 1992: from the Huntington Library it is possible to consult the whole Information and Reference System, although one can only gain access to a dozen digitalized files, the copies of which are to be found in Pasadena. Other experiments on transmission of digital images will probably be done during the Expo in Seville.

3.3. The Architecture of the System

As was insisted above, with the creation of this system it is hoped to mechanize the day-to-day functions of the historical archives, functions which together may be summed up as:

- intellectual control of archival documentation through the different tools or "finding aids" of the files, thus permitting access to information contained in the documents.
- consulting the documents in the archives, both in their original form in the Reading Room and through reprographic copies.
- monitoring the documents service for consultation and reproduction, monitoring the reading room, movement of files within the archives, etc.
- printing and reproduction of documents and preparation of finding aids for publication.

Taking into account these functions, which the system plans to computerize, as well as the need for flexibility in order to make the future system usable in other archives of greater or lesser scale than the A.G.I., the system was thus designed in a modular form, according to the concept of "distributed processing". (Figure 1)

The whole system is made up of several subsystems, each one of them covering a set of functions which are logically related and which are undertaken within their corresponding operating environments, that is, computers with operating systems and communication facilities.

The different subsystems are as follows:

- Link interface: gives the rest of the subsystems the necessary communication, sending messages between them.
- User interface: facilitates users' dialogue, converting their questions into messages to be sent to the corresponding subsystem and converting the received messages into easily intelligible answers. The aim is to make the system as user friendly as possible.
- Textual data base: gives support to the Information and Reference System, including descriptive information which can offer the user the necessary tools for access to both the original documents and reproduced ones.
- Optical Digital Image System: on the one hand, it covers the tasks of document digitalization and storage in optical disk, and on the other, it offers to the user the

necessary resources for consulting the digitalized images with all possibilities of improving legibility.

- User Management: enables the research, the reading room, the requests for copying and the movement of documents within the archives etc. to be monitored.
- Printing subsystem: offers possibilities for printing information in the form of text (Information and Reference System, User Management) and image (Optical Digital Image System).

The operating environment in which these subsystems function is as follows:

- One LAN Token Ring of 16 Mbits with APPC communication protocol backs up the communications.
- An IBM AS-400 minicomputer, with SQL400 as server for the data base and the User Management system.

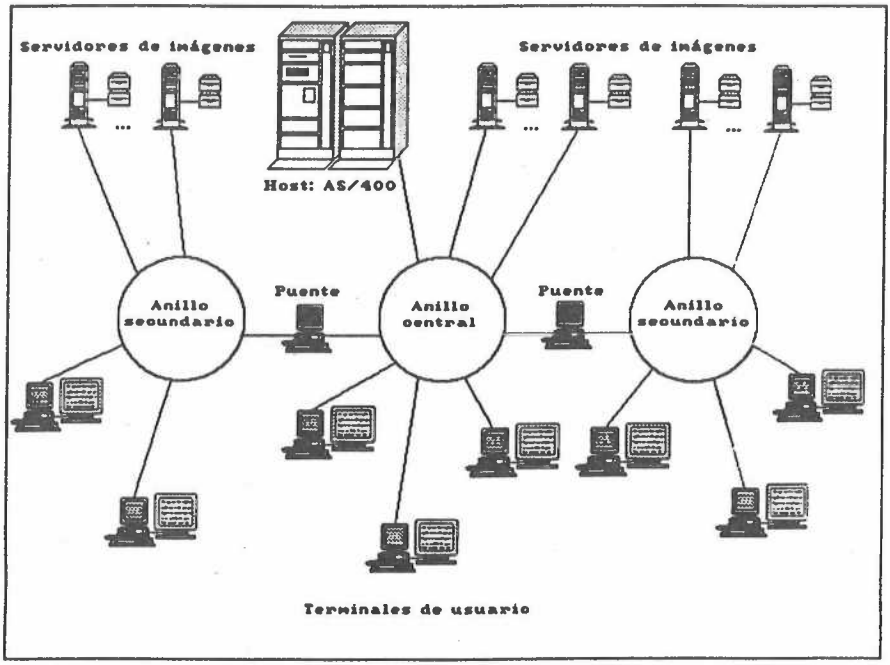

Figure 1: The architecture of the system

- High range PS/2 microcomputers with an OS/2 operating system, as optical disk servers. They control Reflexion Systems RF-5010C optical disk units.
- PS/2 high range microcomputers with an OS/2 operating system, and Dialog Manager and Presentation Manager for the user interface, as work stations. They control two monitors, one conventional one for text (IBM 8513), and another high resolution one for handling images (IBM 8508 for images in grey, or IBM 6091 for colour images).
- Digitalization stations initially with IBM AT microcomputer, DOS operating system, Rank Xerox scanner and Reflexion Systems RF-5010C optical disk unit.
- Maps and plans digitalization stations, consisting of PS/80, with 16 Mb memory, Optic Reflexion Systems RF-5010C, optical disk unit, and Nikon LS3500 slide scanner.

4. The Information and Reference Subsystem

4.1. The Archives Descriptive information

Before going into detail on the Information and Reference system created we must devote a few lines to a minimal comment on the peculiarities of archival description.[8]

A significant part of the archivist's work is devoted to the task of "description", that is to say the creation of the "finding aids" (inventories, catalogues, indexes), which can provide the necessary physical and intellectual monitoring of the holdings and which open up the necessary roads to access the information. Compared with another type of description — bibliographical — archival description has certain differences arising, above all, from the origin of the documents. The "principle of provenance" is the basic archival principle: archival documents reflect the functions and activities of the person or body who produced and accumulated them. For this reason these must always be preserved in relation to their origin or provenance, respecting as much as possible the original order, preserving them along with the documents which were created together, and without mixing them with the documents of another individual and body. Although the description is laid down in the archival practice, as we have indicated, in the individualized "finding aids", in fact, archival description is, due to the peculiarities of the principle of provenance:

- multilevel: it includes descriptions or entries from different units or archival groups considered at their different levels: series, subseries, bundle, file...
- in practice, archival documentation may be seen as a hierarchical, tree structure, in which the different archival units or groups are encompassed within others at a higher level.
- the individualized description of the different documents or items does not taken on its full meaning and is not completely intelligible without being considered in relation

[8] Duchein, Michel: "Le 'respect de fonds' en archivistique: principes théoriques et problèmes pratiques". *La Gazette des Archives*, 1977, # 92, p. 71-96; International Council on Archives: Statement of principles regarding the archival description. By the Ad-hoc Comission on Descriptive Standards. 1992; Keteelar, Erick: "Exploitation of the new archival materials" *Proceedings of the 11th International Congress on Archives.* Paris, 1988, p. 189-199; Lytle, Richard H.: "Intellectual access to archives..." *The American Archivist*, v. 43 (1), 1980, p. 64-76, y v.43(2), 1980, p. 191-208; Lytle, Richard H. and David Bearman: "The power of the principle of provenance" *Archivaria*, 1985-86, # 21, p. 14-27.

to the units of which it forms a part, and the institutions or persons which produced this documentation.

- this means that access to archival information follows different roads from those usual in documentation or library centres. The principle of provenance is not only the basic norm for organizing documents, it is also the main criterion which permits intellectual monitoring of the archives and which controls access to the information.
- given the importance of access by means of provenance, the formulas of direct access to information through indexes have always been of secondary importance in the archives. The indexes are in general "auxiliaries", complementary to the main finding aids, the inventories and the catalogues.

4.2. The Information and Reference system

The Information and Reference System was designed in order to deal with the peculiarities of archival description. This system gives access to the Archives descriptive information, and through it, on the content of its original documents.

The information and reference System contains in one single data base "all" the descriptive information existing on the AGI(the data input is practically completed), with a marked increase in the description of the digitalized documents.

At present the data base contains entries for some 300,000 archival units, units which run from the Archives itself, as a whole, to documents or individual items. Of these 300,000 entries, more than 100,000 have been done during the project, the rest have been taken from the already existing finding aids.

Taking into account the possibilities of computing in the construction of integrated information systems, and making the most of the possibilities of relational model data bases it was attempted to:

- create a single hierarchical, tree system of all the descriptive information which, from the "root" (the Archive) up through the different branches, permits us to reach its "leaves" (units or documentary item).
- to be able to "navigate" through this hierarchy, that is to say, to go up or down the different leaves and branches.
- to be able to include and handle the "real" levels which exist in this structure, or have been described in the Archives, in such a way that the levels are not pre-established: the needs of the Archives are those which determine the number of levels necessary. The system may handle whichever levels are considered opportune at each moment. It is the Archives needs and the possibilities for work which determine their real number.
- to offer, at the same time, direct access to information, either through free text methods, above all through indexing with key words or descriptors.

As has already been said, the Project has paid particular attention to achieving a system which would be as user friendly as possible. For on line access to descriptive information a relatively simple user interface has been developed which makes it possible to:

- access information following the Archive's hierarchical structure.
- to access information through key words or descriptors.

- to access information through a reference number, or element which physically identifies the document in the reprository.
- The system enables the results of the different queries to be printed, it enables the results of a query to be "kept" in the user station for successive days, and enables the archivist to prepare "publications", that is, to create a special format which acts as "filter" when it comes to carrying out searches and sending their content to a list or publication. The result of the work can be printed out directly or transferred to a data processor or a desktop publishing system for more finished product.

For the generation of all the "descriptors" or "key words", which permit "direct" access to the documents, the indexes already existing in the AGI have been also used as far as possible. Logically, in all the new descriptive work done, the production of the indexes has also been included.

The Information and Reference System is supported by one of the "servers" of the whole system, a server which resides physically in the IBM AS-400 computer.

For the design of the data base the relational model has been followed, using the SQL (Structured Query Language) which belongs to the AS-400, although the end user has access only through the designed menus which the User Interface provides him with, prepared with the Presentation Manager and Dialog Manager as was commented earlier.

5. Optical Digital Image Storage System

We have already said that our project aims at the integrated processing of all archival functions and it therefore entails much more than mere digitalization of documents and their storage on optical disk. However, it is clear that digitalization is the most immediately impressive part of the project, due to the possibilities it offers for preserving and providing access to the documents, as well as the state-of-the-art technology used and the spectacular results obtained.[9]

5.1. Digitalization parameters

The main practical objective of document digitalization is to substitute consulting original documents with a view to better preserving them. There are documents in the AGI which go to the reading room more than 100 times throughout the year, which obviously poses the very serious problem of how to preserve them adequately. Faced with the dilemma of good preservation-access to documents the best alternative found is that of reprography:

[9] Rütimann and Lynn, op. cit, p. 11: "The interface to the Image Data Base is outstanding. It is designed primarily to enable researchers to display selected documents and to scroll or otherwise navigate through a bundle of documents stored on the optical previously mounted on the image server using the scanning control information for referencing purposes; and to provide researchers with a set of computational tools for enhancing purposes... The tools have been carefully tailored to the particular characteristics of these documents, taking into account their reflectance and optical contrast, and particular types of artifacts encountered. In this case, such tailoring is appropriate since it occurs at the end user's workstation, not at the image server... The speed and esae of use of these tools are impressive. There is something almost magical in seeing a badly stained section of a 300-year old manuscript cleaned up before one's eyes and become legible again."

the use of copies of the documents for consultation instead of the originals. It is clear that access to information must be maintained and increased, but that also the heritage which has been passed down to us must be protected, even if this means in many cases that the researcher is deprived of the "fetishist" pleasure of physical contact with the original.

If it is intended to install consultation of the document in digital form in the reading room, then a standard of quality must be offered which satisfies the majority of the researchers and which requires recourse to the original only in very limited cases. In the case of old documents, this means excellent quality of image provided thanks to a set of tools which facilitate the work and which provide some "advantages": more speed in access, improvements in the legibility of documents, new possibilities for copying, etc.

The storage systems for digital images of documents existing in the market now have been designed for modern documents in standard paper, typed or printed in general, of good reading quality etc. For this reason they use parameters that are not acceptable when dealing with old papers, sometimes with preservation problems, with possibly stains or loss of intensity in the inks, written on both sides of the paper and sometimes with the ink bleeding through the paper. Sometimes the paper is fragile and its size is rarely standard. Sometimes it comes in loose sheets and others in bound books or stitched bundles. This means that the documents can never be put into the scanner with an automatic paper feed but require scrupulously careful handling by the operator. It also means that the time needed for the job is multiplied by unavoidable human intervention in the work, which sometimes leads to a high cost for data input.

Most of the systems for digital processing of documents take the image using only black and white, avoiding grey levels, to make for more economical storage. However, when the problem of ink transparency is acute, it is further exacerbated by black and white digitalization. The techniques used by conventional scanners (contrast changes, histograms, moving threshold, etc.) cannot help us with this problem in most cases.

However, we must repeat, the project is conceived with the idea that reference to the original documents will be highly restricted after they have been digitalized. This means that we must think of the end user of the digitalized images, the archivist, or the researcher who will have to spend long hours in front of the screen. Thus, the copy should be high quality and clearly legible if the system it to be acceptable.

The conclusion of all this and after carrying out pertinent acceptability tests, was that we would digitalize using 16 grey levels and 100 dots per inch. This gives us a trade-off between higher quality and storage space. Far more information needs to be stored and the compression algorithms are not as efficient as when they are used at only two levels (black and white).

5.2. The "archival" work of preparing the documents

We have insisted on this point: the project is specifically directed towards answering the needs of the archives. This means that the documents to be digitalized have to be adequately prepared in accordance with the characteristics of archival documentation and archival methodology. On the other hand, if the documentation is not correctly prepared beforehand, computerization (especially document digitalization) is most unlikely to be successful.

That is why, given that we had no precedents on which to fall back, we have had to develop our own working methodology to take us step by step from the original document to the image of the document, from bundles of papers to an optical disk, with the minimum risk of error possible.

Taking into account the archival nature of the project, the first criterion for selecting documents to be digitalized is only to digitalize complete series of documents, not odd documents selected for their interest or for the subjects they deal with.

Moreover, our aim was to obtain the highest possible level of performance in providing reading room services. Statistical studies were carried out to determine which series had been consulted most in the last few years. The result of this purely statistical analysis indicated to us that by correctly selecting the documents to be digitalized, with only 10 per cent of the AGI the needs of almost 40 per cent of consultations could be covered.

Other criteria such as the documents' arrangement and description and their state of conservation were also taken into account, as well as a geographical consideration due to the project's specific circumstance: it was intended that the documents should affect as wide as possible a geographical area.

Using all these criteria, we drew up a selection process covering over 4,000 bundles form different series within the AGI, which is what we are currently working with. Later, the statistical studies on the performance of the series selected were repeated, using data obtained from the User Management programme that was already operating. The results are very encouraging. If we manage to meet our goal of 10to deal with over 30holdings from the search room. Apart from these document series, the entire section of maps and Plans , consisting of 7,000 items, will be digitalized in colour.

However, during the development of the project we have decided to initiate an important change in this strategy. Since there are document series in other Spanish archives dealing with the colonisation of America, we shall scan them also: papers on the "Secretaría de Guerra" (War Secretariat) Section from the Archivo General de Simancas (AGS) and papers from different sections of the Archivo Histórico Nacional (AHN): Ultramar, Diversos, Inquisición, Ordenes Militares...). The results are clear: the AGI will have a complete copy of the Indies records that it does not hold within its own walls (some two thousand bundles by the end of 1992). The AGS and the AHN will also benefit from the project once the entire system network planned for this and the other historical archives has been installed.

The documents selected must be adequately prepared for despatch to the scanning room i.e. through the purely archival tasks of ordering the documents bundle by bundle, preparing a new description of them, or revising the already existing one, placing them properly in document cases, writing the reference number, etc. Without adequate preparation all the later work will be practically useless.

A special form is also made for each bundle, called the "digitalization guide", with the basic information for later scanning. In time the information on this form will be recorded on a floppy disk which will be delivered to the scanner operator along with the corresponding bundle which he is to digitalize.

5.3. Digitalization

Once the documents have been prepared they are sent to the scanning room, where 15 work stations are available: a microcomputer with DOS operating system, a XEROX 7650 flat scanner able to digitalize up to A3 size at 400 dots per inch, with 256 levels of grey (although, as mentioned above, only 100 dpi are used and only the 16 most important levels of grey are maintained), and a Reflexion Systems Mod. RF5010 optical disk unit, with 940 MB Panasonic or Plassmon 5 1/2 inch optical disk.

The digitalization is done bundle to bundle. Each one of the operators (30 at present, in two shifts) receive a bundle, correctly prepared (in order, filed, with reference number) and accompanied by the diskette which contains the digitalization guide.

For practical reasons one optical disk is used for each bundle (with a calculated 2,000 pages in each bundle and 350 Kb per page: the average occupied space is around 700 Mb per disk, although there are bundles which occupy more than one disk).

Reading the guide floppy, the screen will constantly direct the operator with indications regarding the call number of pages, and the page on which he is currently working, etc. The programme had various facilities : digitalization by scanning in two-page sweeps, size permitting; warning that there are blank pages in the original document, so that the page order can be maintained etc. We have tried to foresee all possible requirements and make the system user friendly. For operators without archival skills will not be expected to make decisions and their work must be as simple as possible in order to eliminate the possibility of mistakes. In actual fact we consider that the resulting quality of the regular digitalization process at 16 grey levels and 100 dpi — once the scanner is adjusted and the most suitable colour cover placed on it — is sufficient, and in practice does not require that the operator decides on the quality, only that he checks that each image is complete and in its correct position. For this reason the screen used is an IBM 8512 i.e. a small format screen which is not very high resolution.

It currently takes one minute per page to carry out all the operations: placing the document carefully (for we should not forget how old they are), giving exact instructions (e.g. page format) switching the scanner on, looking at it on the screen, compressing it and recording it on optical disk.

Despite the compression algorithms (we use Differential Pulse Code Modulation with statistical compression), enormous storage capacity is required: approximately 350 Kb for an average screen image.

5.4. Quality Control

Two types of quality control are done on digitalization during the process: one is purely automatic, through a programme which compares the "digitalization guide diskette", and another manual check done by an Archives professional who checks using statistical criteria the optical disks which are the result of the digitalization process. It is important to indicate that in this process in general the quality of the image is not examined: in accordance with our initial decisions, it is taken for granted that the quality used in digitalization is generally adequate.

If errors are detected during the automatic checking stage, a manual check is carried out in a more detailed manner in order to detect several different errors: pages which have not been digitalized by mistake, pages out of sequence, pages repeated, incomplete pages etc.

The errors detected may be corrected: many of them through the quality control programme (changing the page order, inverted pages etc.) and others through re-digitalization (pages which have not been digitalized, incomplete pages...).

In fact, once digitalization has been done the process may be concluded and the disk placed at users' disposal. Quality control and precise modifications can be done afterwards, but in general it is intended not to make the disks available for consultation until they have been checked.

5.5. The Documents Service

As I have insisted throughout this paper, the final aim of digitalizing documents is to substitute direct reference to the original document with high-resolution on-screen displays or reproduction on paper. This is of interest for several reasons

- better conservation of our historical records,
- greater speed of access to information and thus
- wider availability of such information.

Placing this enormous quantity of information in the reading room for consultation has meant that serious problems have had to be faced up to. One of the main ones was how to actually provide the image "service" in the reading room, given that there are several thousands of optical disks.

The usual systems in the market solve the problem using the juke-box, however no juke-box is known to be capable of handling four or five thousand optical disks. On the contrary, they are designed usually for not much more than a hundred disks.

Another possible solution was the use of a robot: in the market now robots are available which can handle several miles of magnetic tape cartridges, but they require adaptation to handle optical disks. So, this is an alternative which is still under study for the future.

We also analyzed the possibility of keeping a copy of the disks in the reading room, so that the researcher himself could take the disk directly from the shelf, or ask the member of staff in charge for the one required and return it after use (the decision to make a disk for each bundle would facilitate this operation): this is what is usually done with microfilmed documentation in certain places. But this meant a lot more risk due to the movement of the disks and the need for a disk drive for each user station.

The solution adopted was slightly different: at this moment of the system's operation human intervention comes into play. From the user station, where a search has been undertaken and the images of a document have been requested, a message is sent to a monitor situated in the optical disk room: an operator receives the message and puts the corresponding optical disk into one of the disk drives to send the images through the network. The images of the document are stored in the hard disk at the user station and when the delivery is completed, the optical disk may be removed, leaving the disk drive available for another request. With one single person attending to the different requests

handled through the monitor, and a not excessively high number of disk drives, the whole process can be carried out in a relatively convenient way.

5.6. Document display

At his work station (consisting of a high range PS/2, with two monitors: one for access to the Information and Reference System, and User Management, and the other: a high resolution IBM 6019 for presenting the images) the researcher receives the set of pages which make up the document or file he has requested. These images are stored in his hard disk and with them he may work locally.

In the time required for access to the information the human operator's time must be evaluated particularly. This does not have to be particularly high, but logically may have a negative effect on the process if his work is carried out in too "flexible" a way. We must also consider the access to optical disk time and the time for image decompression, as well as the size and, above all, the number of pages, very varied, which make up each document or file. The time it takes to go through the Local Area Network (16 Mbits per second) is practically negligible. It is known that access to the optical disk is fairly slow (about 5 seconds are needed for each page and a few more to decompress it once it is in the user station).

The pages are sent in order, beginning with the first, however it has been attempted to give the system a certain amount of intelligence in sending the images, above all if the documents are extensive. If the researcher goes through, looking at the first, second and third pages etc in order, the system continues to send the document in order until it is completed. On the other hand, if the researcher "jumps" from the first, for example, to the twentieth page before this page reaches the user station, the system will jump to number 20 and continue sending 21, 22 etc. until the end, or until another jump.

On the screen used for handling the system all the possibilities shown in Figure 2 will appear. Below are listed only some of these:

- The system uses the traditional numbering system by foliation, indicating recto and verso pages, informing the user of the total number of sheets and the total number of images which may be displayed.
- If in the process of preparing the documents this has been anticipated, each document or file may be broken down into various "blocks" which, if they are marked in the process of digitalization as such, allow a kind of "skimming" of the file. Really what is being done is the marking of "jumping points" within the document which enable it to be consulted quickly. When one takes a document or file in paper form, one does not usually read from the first line of the first page to the last line of the last page, on the contrary, one "skims" leafing through and collecting general information before deciding where to begin, enabling one to jump rapidly from one part of the document to another. This is what is intended, in somewhat similar way. The Information and Reference System only has one entry for the whole document: therefore, all its pages will be obtained, and the document can be seen from the first page to the last, but if the beginning of each one of these blocks has been marked beforehand (for example, the beginning of each one of the documents which make up a personal file) it is possible

to "skim" down the first page of each block in order to gain a quick grasp of the whole and later concentrate on the pages of greatest interest.

- A page which is of interest may be marked in order to go back more quickly to it later, without having to remember what page it is.
- The image may be enlarged or reduced in order to obtain a better display. An image may be rotated or inverted and a scroll made with it.
- A page on the screen may be printed out and, in particular, different tools may be used to improve the legibility of the documents. As has already been stated, the aim is not to store the best possible image during the digitalization process: a generally good quality image is obtained, but, above all, different operations to improve this image may be carried out afterwards at the user station.

This is really one of the most spectacular aspects of the system, as can be seen in the reproductions which are included with this report. Throughout the process several research projects have been undertaken which have enabled different algorithms to be developed to improve the quality of the image.[10] We have finished two different lines of study, oriented towards improving the legibility of the documents.

The first is based on prior digitalization with grey levels, which make it possible to provide the users with a set of facilities so that they themselves can select the range of tonalities which best suit each image they are trying to read. Several functions are already included in the system, which separates levels, blackens the letter and separates it from the background. The process is done in real time at the user station, giving spectacular results.

The second line is somewhat more complex and aims at giving individualised treatment to each type of ink degradation, using specific algorithms to compensate for the degradation as much as possible. Many of the methods used are based on local algorithms which process each zone of the document separately as a function of contrast and other problems it might offer. The user may select any of these methods (See Figure 2 quoted above) and apply them to a whole page or a part of any image previously marked. Examples of the results of such methods seen in figures 3 – 5.

5.7. Documents in colour

For colour documents a different working strategy has been followed. In this case we have taken it that, in a historical Archivee such as the AGI, colour treatment is only

[10] Some papers published by members of the team about these lines of research: Bescos, Julián. "Image processing algorithms for readability enchancement of old manuscripts" *Electronic Imaging'89*, v. I. Pasadena, 1989, p. 392-397; Bescos, Julián (et al.). "Reflectance and optical contrast of Old Manuscripts: wavelength dependence" Julián Bescos, Francisco Jaque, Luis Montoto. *SPIE*, (1988), V. 1028, p. 258-262; Bescos, Julián (et al.). "Filtering and compression of old manuscripts by adaptive processing techniques". Julián Bescos, Juan Pedro Secilla, Juan Navarro. *Society for Information Display International Symposium*. Las Vegas, 1990, p. 384-387; Bescos, Julián (et al.). "Mejora de legibilidad de documentos antiguos mediante tratamiento digital de imágenes" Julián Bescos, Juan Navarro, Carlos Ramon. *IV Simposium Nacional de Reconocimiento de Formas y Análisis de Imágenes* Granada, 1990, p. 51-58.

required by a small quantity of documents which, usually for reasons of conservation, are part of a specific Maps and Plans Section.

In the case of the Archivo General de Indias there is a Maps and Plans section made up of some 7,000 items. Not all the maps and plans require colour, but a significant number do.

These documents, apart from colour, have other characteristics which differentiate them from conventional documents: as regards digitalization, fundamentally, their variety of sizes — in general very large — and sometimes also the size of the letters or the lines drawn in them.

Most of the sizes come between A3 and A1 (in other words between 420 x 297 and 841 x 594 mm), with a greater number in the A2 category (594 x 420 mm).

These data of a general nature (large size and need for colour, fundamentally because these factors may be very significant in maps and plans) pose the immediate problem of storage, added to the problem of scanning equipment — which cannot be a flat scanner in this case —, and the screen's display capacity in order to be able to successfully present maps of greater size.

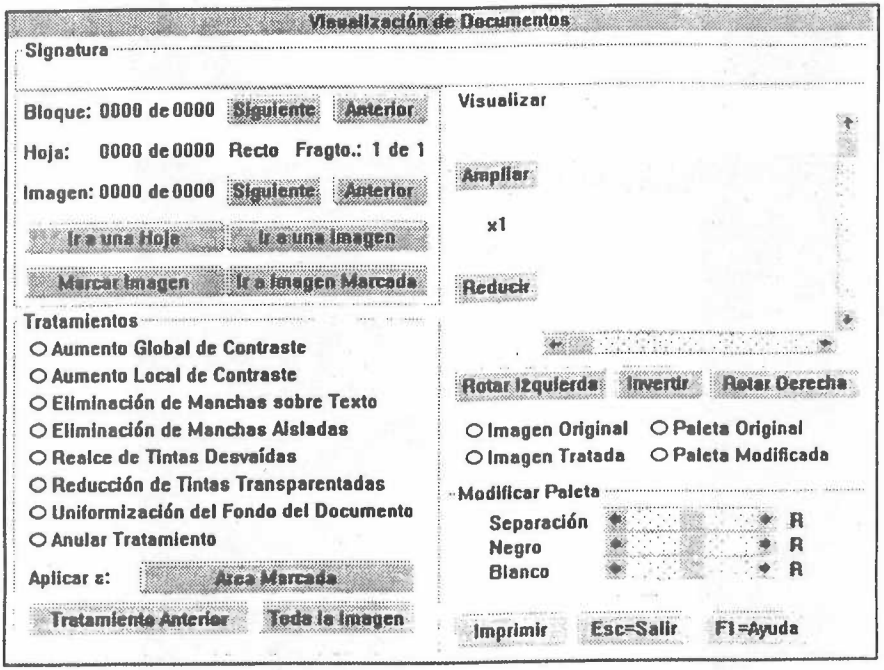

Figure 2: The user screen for managing the images

Figure 3: A document before and after processing to remove the bleeding of ink through the paper

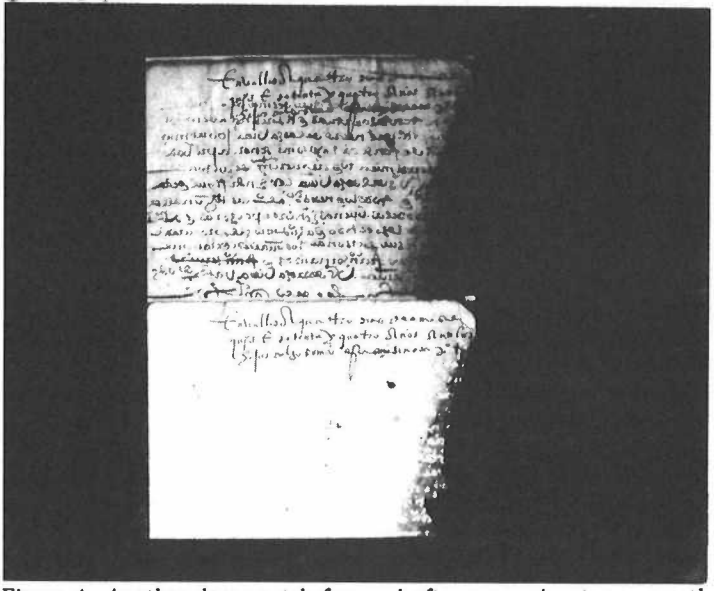

Figure 4: Another document before and after processing to remove the bleeding of ink through the paper

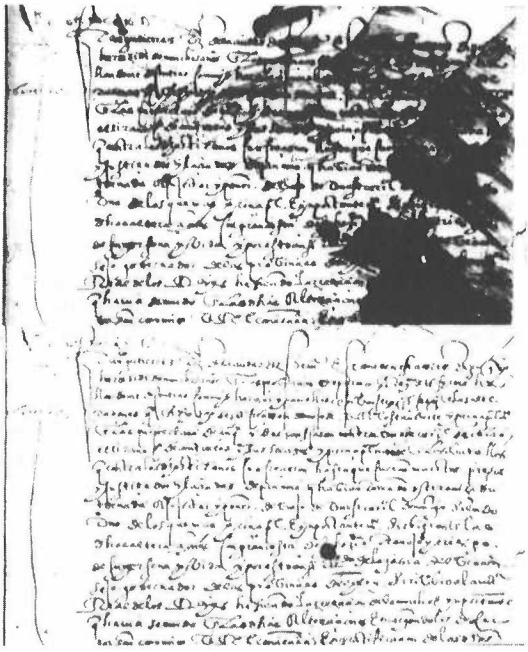

Figure 5: A document before and after processing to remove blotches

In order to solve these problems the following solutions were opted for:

1. To search for a balance between quality of image, storage needs and the time required for later display.
2. Digitalization of maps and plans through prior microfilming in colour. Advantages:
   - a copy in photographic medium with excellent resolution and perdurability is obtained, which may be used for other purposes where high quality reproduction is required.
   - The quality obtained (up to 200 pairs of lines in 35 mm photograms) enables the high quality reproduction needs in printing to be satisfied.
   - At the same time microfiche colour edition of the different series of maps and plans in the AGI made be done, which could be continued by editing series of maps and plans from other Archives.
   - The illumination of the original is done in a very short time, avoiding exposure to the intense, prolonged light required by digital cameras.
   - Practically all maps and plans can be microfilmed in their entirety, without recurring to fragmentation.
   - Moreover, the possibility remains of repeating digitalization in the future as digital technology advances without recurring to original maps.

3. Later digitalization is done by a microfilm or slide scanner. In the specific case of the AGI project, the photograms are then mounted in slide frames and then digitalized with the Nikon LS3500 scanner.
4. The use of the Nikon scanner permits a maximum resolution of 4096 x 6144 dots for 35 mm. photograms. Although this scanner enables rolls of film to be digitalized without mounting, it is used its digitalization facility in slide format, with a useful image size of 22 x 34 mm.
5. The resolution which is stored through digitalization is around 100 dpi, equal to the rate for black and white documents. The number of colours is 256.
6. As the process of displaying the resulting images in colour is relatively slow in any case, another digital copy of the same image is stored in black and white for much faster access by the user who consults the image in colour only if required.

The results will therefore be:

- a copy in black and white, for more rapid and less demanding handling
- a high quality, digital copy considered sufficient for direct consultation in the reading room, when consultation in colour is required, although access time is penalized.
- a quality copy in colour microfiche.

5.8. Document reproduction

Only a brief reference to reproduction of documents on paper. This will be done in two ways:

- hard copy, made directly from the image displayed on screen, with the legibility enhancers introduced by the processing algorithms used at that moment, and
- deferred printing, i.e. printing done directly by the system, following a request sent through the User Management System, providing reproductions of entire documents or selected pages from them. In this case, the images will be reproduced in the form in which they were digitalized, without any sort of enhancement in legibility.

6. New Perspectives

It is a well-known fact that one of the main problems in any computing system is the obsolescence of equipment and technology due to the very rapid speed of new advances. For that reason the improvement of particular aspects is already being anticipated for which new possibilities have appeared during the last few years but which could not be incorporated at the time.

One of these aspects is that of obtaining an adequate backup copy system. For some time we have been awaiting the moment at which "optical tape" would become a reality. Manufactured by a Canadian company, CREO, its use is now possible, with a capacity calculated at 1,000 Gb (Gigabytes) on tape i.e. one Tb (Terabyte). Some steps forward are being made and it is possible that this system may be available in the not too distant future.

New possibilities also exist in the operating practice of the disks' service. Given that at the time there were no juke-boxes capable of handling such a high number of disks, the system's architecture was prepared with the introduction of human intervention into the optical disks service, as has already been commented. At present a robot system

(Comparex) is being adapted, prepared initially to handle several miles of cartridges of magnetic tape and which will be used for the optical disk service.

Another aspect on which work is being done for the future is the new image compression tools. When the project was started up no standard algorithm for image compression with grey levels existed. At the present moment the JPEG (Joint Photographic Expert Group) exists, which is being studied by a development team and which will probably enable the storage necessary for image to be reduced from the present average of 350 Kb to an average of below 120 Kb.

This same lack of standards appears in other aspects of the project. In actual fact, one of the main problems which we face for the future (taking into account that this project began more than six years ago) is the standards problem which is becoming increasingly serious, especially due to the importance of information exchange in general, and OSI systems (Open Systems Interconection) in particular. The exchange of information was not a priority objective for the project (it has not been a priority in general in the Archives world up to now), which was conceived as a system, firstly, for the internal functioning of the Archives Service and, secondly, for its integration withing the future Spanish State Archives Network, providing possibilities for text consultation through the Ministry of Culture's Network (Puntos de Información Cultural, P.I.C.). Really this integration has not yet been undertaken, however, although the problem has been anticipated for the future and it will very probably be tackled soon, when considering possible changes to open systems.

This very reasons for opening the project up towards information exchange and the greater distribution of the software, for its use in other Archival Services, make it advisable for us to produce a version for the future which uses only microcomputers. Taking into account the fact that a fair number of possibilities exist for transferring the system to other Archives, although we are thinking more here about the Information and Reference modules and the User Management System than about image treatment, some steps forward will probably be made in the future to produce a new version which will enable PS/2 microcomputers to be used with the OS/2 operating system as server for the Information and User Management Systems.

For the future other specific developments are being undertaken at this moment. To sum up the project, as a synthesis of what it is possible to do with the new technologies when applied to archive documentation and as an example of what the Archivo General de Indias is: at present a CD-ROM is being developed, which will have the title of "Tesoros del Archivo General de Indias" (Treasures of the AGI) and which will be a small sample of what has been done, at the same time as being a summary of the history of Spanish colonization in America through its principal documents.

7. Conclusions

7.1. Differences and similarities between the different projects.

The three systems which we have presented over the course of these pages offer in general positive prospects for what can be expected of the new image technologies. But the objectives and approaches to the problem are very different, as I believe has been made clear in this work. Let us try to specify these differences:

Clear differences exist In the general thinking or approach to the problem : the ODISS project and the CLASS project are two typical document digitalization, image processing and optical disk storage projects. In contrast to them the AGI project is much more ambitious: it attempts to achieve the development of an integrated computing system for overall treatment of historical archives.

The American projects include the digitalization of documents and the basic tools for access to them through a data base. Here also the AGI project goes further: it attempts to develop a unified information system with all the descriptive information in the Archives, not only what is used to retrieve the digital images, although logically this part of the data base is the most important.

There is nothing in the American projects related to "user management", monitoring the reading room, monitoring the movement of archives and establishing a management system for research, consultation and reprography.

It is clear that the AGI project is much more ambitious in this sense and that it has many more facets, since it is focused towards the treatment of all the tasks involved in a historical Archives.

The different projects analyzed have similar objectives: to preserve and distribute documents, although the documentary holdings they deal with are of a different nature.

Although the historical worth of the documents which are the object of digitalization is different in the different projects, in no case has the possibility of eliminating the original document to be substituted by the digital image been considered. On the contrary, the objective is to collaborate towards their better preservation and avoid handling the originals.

Logically, the documents in the Spanish project present more handling problems and therefore more digitalization problems, since they are from a much greater range of periods (XV-XIX c.) and from documentary series with very different characteristics (the "Informaciones y Licencias de Pasajeros" series has nothing to do with the series of "Registros"). The American projects, in contrast, deal with much more recent documents (XIX -XX c.) and a single series of documents (in the case of ODISS) or books (in the case of CLASS).

This difference is fundamental in the general approach to the problem of digital image. In the Spanish case this has made it necessary to develop extremely powerful tools in order to improve display of the images (problems of transparency of the inks, stains....). This problem, combined with the greater historical value of the Spanish project's documents, has led to a different approach: the improvement of the image is not done at the moment of digitalization but rather at the moment of viewing by the user. The image held in the disk is more "real", more photographic, and the improvements which enable display problems to be eliminated are done afterwards only where necessary.

This all means higher costs for the digitalizing operation (the documents must be very carefully handled due to their great antiquity; it is not possible to use fast scanners; the digitalization process is slower; more storage is needed, etc.) but it offers greater possibilities for later display.

The objective of the CLASS project, to be able to obtain facsimiles at high speed, is not of such high priority in the other two projects, so that the printing systems have less importance in these cases.

7.2. Prospects and problems pending

At the start we talked about the tremendous possibilities of the digital image. However, at the present moment there are still grey areas for which, without doubt, solutions will be found in the coming years:

- the problem of the lack of generally accepted standards is a serious one. There still does not exist an image format which is universally accepted (although progress is being made on this), nor the interchangeability of optical disks... As for compression algorithms, for some time standards have existed for two-level images, but only recently the JPEG has been accepted for grey levels. For this reason it is very difficult, if not impossible at this moment, to interchange information between different systems.
- the problem of the obsolescence of computing equipment (physical and logical) is even more serious in the case of image and optical storage technologies — so new and dynamic — which make it vital to foresee the necessary conversion of systems in the not very long term future.

Nevertheless, the future advantages of the new technologies for digital image processing are clear, as we indicated at the start: advantages for access and distribution of information, advantages for best display, advantages for obtaining copies in other media.

The technologies of "electronic publications" — the CD-ROM mainly — may also be used in an advantageous way to distribute digital images of documents, although at the moment there are no large scale projects in this area and their present capacity is fairly limited for image distribution. However, the preparation of "collections" of CD-ROM disks, with document images could be a valid instrument for the large scale distribution of important document series at a relatively low price.

The advance of technology is also working in our favour: disk capacity, for example, increases continuously, and the cost of storage is reduced significantly (it is considered that this is reduced by half every two years). New tools to improve displaying appear, the possibilities of information interchange grow, the tools of printing improve and ever more powerful processors enable processing time to be reduced. There will be new advantages in the future as the tools of Optical Character Recognition (O.C.R) advance, which enable original images to be transcribed for conversion to text. It is true that these tools have not advanced in the automatic transcription of manuscript text, but the way is open. What will happen when, through the automated analysis of digital images using these O.C.R. tools, all the "Registros" of the Archivo General de Indias can be transcribed — now already digitalized — so that the text of each one of the Indies Council's government orders, dictated daily over the course of three centuries, may be analyzed and retrieved using all the most up-to-date methods for working with the text?

Manfred Thaller (Ed.)

Images and Manuscripts in Historical Computing

Max-Planck-Institut für Geschichte
In Kommission bei
SCRIPTA MERCATURAE VERLAG

St. Katharinen, 1992

Halbgraue Reihe
zur Historischen Fachinformatik

Herausgegeben von
Manfred Thaller
Max-Planck-Institut für Geschichte

Serie A: Historische Quellenkunden

Band 14

Erscheint gleichzeitig als:

MEDIUM AEVUM QUOTIDIANUM

HERAUSGEGEBEN VON GERHARD JARITZ

26

Table of Contents