Journal of Universal Computer Science, vol. 17, no. 1 (2011), 151-161
submitted: 8/4/10, accepted: 12/12/10, appeared: 1/1/11 © *J.UCS*

Nabuco

Two Decades of Document Processing in Latin America

Rafael Dueire Lins
(Federal University of Pernambuco, Recife, Brazil
rdl@ufpe.br)

Abstract: This paper reports on the Joaquim Nabuco Project, a pioneering work in Latin America on document digitalization, enhancement, compression, indexing, retrieval and network transmission of historical document images.

Keywords: Historical documents, Document engineering, Image processing, Back-to-front interference, Show through, Bleeding
Categories: H.3.3

1 Introduction

Joaquim Nabuco was a Brazilian statesman, writer, and diplomat, one of the key figures in the campaign for freeing black slaves in Brazil (b.1861-d.1910). Nabuco exchanged letters with the most prominent people of his time and the file of over 6,500 documents and about 30,000 pages of active and passive correspondence (including postcards, typed and handwritten letters), kept by the Joaquim Nabuco Foundation [FUNDAJ] (a social science research institute in Recife, under the responsibility of the Ministry of Education of the Brazilian Government), is a bequest of historical documents of paramount importance to understand the formation of the political and social structure of the countries in the Americas and their relationship with other countries.

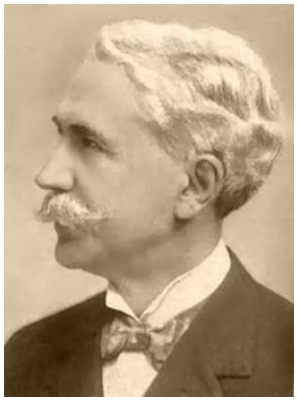
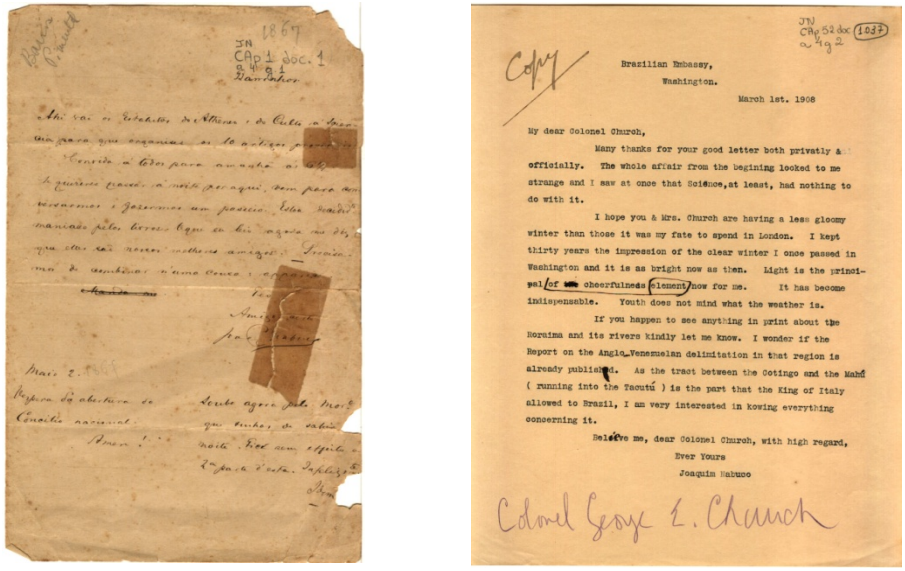


Figure 1: Photo of Joaquim Nabuco in 1870

The Joaquim Nabuco Project, or simply Nabuco Project, was conceived as a way to preserve this important heritage, as the chemical process used in producing paper in the late 19th century used too much beach and the papers are in a fast decomposition process. Graziela Peregrino, the head of the informatics department of the Joaquim Nabuco Foundation, challenged the author in 1991 demanding a better, more flexible and efficient way of keeping the memory of the Nabuco file, which at that moment was microfilmed, a very limited and deficient technique worldwide used to keep the contents of historical documents of the Nabuco file. A number of constraints were imposed by the circumstances, taking into account the Latin American economical reality, and the technological restrictions of those days:

- The physical documents ought not to suffer any damage.
- The project had to use equipments off-the-shelf.
- The platform had to be low-cost (under US$ 5,000 with sales price of the Brazilian market).
- Operators were supposed to undergo minimum training and to be non-specialized.
- The digital file had to be as easily accessible as possible.
- The project ought to digitalize documents in such a way to allow future enhancements of the platform keeping the bequest for future generations.

At that time very few initiatives for document digitalization were set in the whole world and, to the best knowledge of the author, the Nabuco Project was the pioneer of its gender in Latin America, and as will be described later on in this paper, a number of new problems and solutions not previously described in the technical literature were addressed in Nabuco. Figure 02 presents two examples of documents from the Nabuco file.



Brazilian Embassy,
Washington.

March 1st. 1908

My dear Colonel Church,

Many thanks for your good letter both privatly & officially. The whole affair from the begining looked to me strange and I saw at once that Science,at least, had nothing to do with it.

I hope you & Mrs. Church are having a less gloomy winter than those it was my fate to spend in London. I kept thirty years the impression of the clear winter I once passed in Washington and it is as bright now as then. Light is the principal cheerfulness element now for me. It has become indispensable. Youth does not mind what the weather is.

If you happen to see anything in print about the Roraima and its rivers kindly let me know. I wonder if the Report on the Anglo-Venezuelan delimitation in that region is already published. As the tract between the Cotingo and the Mahú (running into the Tacutú) is the part that the King of Italy allowed to Brazil, I am very interested in kowing everything concerning it.

Believe me, dear Colonel Church, with high regard,

Ever Yours

Joaquim Nabuco

Colonel George E. Church

Figure 2: Two documents from Nabuco bequest.
(Left) Handwritten letter. (Right)Typeset letter.

2 Document digitalization

The very first step in the generation of a digital library of historical documents is document digitalization. Although this seems to be a trivial step, the reality is far from being so, as a number of decisions have to be made and their short, medium and long-term implications are of paramount importance to such a project. The very first point to be addressed is what device to use. Today, a number of options are available which range from scanners to cameras (portable and special devices). The technological reality of the dawn of the Nabuco Project in 1991, offered a table scanner as the only solution to meet the restrictions imposed by the project of being developed with off-the-shelf hardware and low cost in the Brazilian market. While today an Officejet HP 3-in-1(scanner, copier and printer) costs around US$ 200.00, those days a desktop HP A4 scanner was priced around US$ 2,000.00 in Brazil.

A number of questions had to be answered:

- Which resolution to use?
- Should digitalization be made in colour, gray scale, or monochromatic?
- In which file format should the images be saved?

The three questions above are intertwined. The higher the resolution and palette, the larger the file. Consequently, the more important the choice of the file format and the use of a loss or lossless compression scheme. Does a higher resolution image mean a better image, or an image with a higher potential use? The technical literature offered no answers, then. There was no other way to follow than to set experiments and answer the questions ourselves.

As "the proof of the pudding is in the eating", the final quality of a document is in visual inspection. A sample of the different documents in the Nabuco bequest (handwritten, typeset, and postcards) was chosen and scanned under 75, 100, 150, 200, 300, with a scanner manufactured by Epson model ES-300C with maximum color resolution of 24 bits/pixel and reading area either A4 or Letter. Then the documents were printed using a HP inkjet printer in white. A team of experts in documentation from the Joaquim Nabuco Foundation inspected the original and printed documents to set the minimum accepted resolution. The conclusion reached by the board of experts was that 75 and 100 dpi images were not satisfactory, but 150 dpi was good enough to visualize all the relevant details of the documents. The difference between the printed documents scanned with 200 and 300 dpi was hardly noticed. The choice made was to digitalize the documents in 200 dpi and one of the important aspect in this decision was compatibility with the resolution of FAX devices, as often the request made by researchers for documents had to be answered via FAX, as there was no Internet available then.

Answered the first of the three digitalization questions, one had to decide if the documents were to be digitalized in colour, gray scale or monochromatic. In principle, monochromatic images were enough to preserve all the document information for researchers of most documents (postcards would lose some of their beauty, but all the information of the written part would be kept). At this time, data CDs were starting to become available, although not popular yet. There was the idea of putting the whole Nabuco bequest in one or two CDs with monochromatic images, together with a simple database search engine, and making this set of CDs available to researchers. In order to make this possible one had to generate good quality monochromatic images.

The direct digitalization setting the scanner to generate monochromatic images has proved unsuitable because as the tested images had dark background due to aging the digitized image obtained was far too dark. Besides that, it was important to have colour images of the documents to keep the content of documents for future possibilities, thus scanning all documents twice would mean handling them two times and double work. For this reason the project team opted in scanning the documents in true-colour (24 bits per pixel) and binarize the image afterwards. As a page of A4 paper has size 210 x 297 mm (8.4 x 11.9 inches approximately) if scanned with 200 dpi resolution it yields almost 4 Mpixels, which in true colour yields 12 Mbytes. Thus, uncompressed in a raster file, such as BMP format, each document image would be that big, being completely unviable to store the images in that format, even for preservation, as a CD would be able to store only about 60 pages of documents (the whole file would be stored in about 500 CDs). The preliminary study made showed that JPEG file format with 1% loss yielded images with excellent quality and whose degradation level was imperceptible to the eyes of the specialists in documentation of the Joaquim Nabuco Foundation whenever compared to the BMP-stored document. Thus, the whole bequest of documents was scanned in 200 dpi, true colour (24 bits per pixel) with a scanner manufactured by Hewlett-Packard model HP Scanjet 4c (optical resolution 600 d.p.i., maximum color resolution 4 bits/pixel, reading area 8.5”x14”) and stored in JPEG with 1% loss, yielding a set of 32 CDs. Two copies of such set were made. One was kept with the Joaquim Nabuco Foundation and another is kept by the Universidade Federal de Pernambuco for research purposes.

Prior to deciding which file format was most suitable to store the digitalized image information, the technical literature was surveyed to search for criteria for such a choice. No reference was found that could help. At the initial phase of the Nabuco Project, back in the early 1990s, a preliminary analysis was made to allow a rapid choice in order to make progress in document digitalization, but this problem remained as a concern to the author, who much later re-addressed the problem with a much wider scope and a better methodology. The results obtained, which was not restricted only to document images but also several other image “clusters” such as people, landscape, documents (colour and monochromatic ones), objects, etc. Several file formats were also analyzed including, BMP, GIF, TIFF, JPEG, PNG, and JPEG2000. Besides that iterative and static file formats were also compared. The results were presented in the paper in reference [Lins and Machado, 04] which amongst other things showed that the decision made for Nabuco was an adequate one as JPEG with 1% loss exhibited a high signal-to-noise ratio compared with the resulting file size. Another pioneer aspect addressed in reference [Lins and Machado, 04] is the analysis of the performance of the iterative file formats (PNG, JPEG, and JPEG2000) which surprisingly enough could yield a file whose final size was up to 10% smaller than the standard (non-iterative) versions. This research was further refined in [Lins and Rodrigues, 06] which analyses the quantity of information carried in each iteration, which is transmitted through computer networks, allowing that a user that is accessing a web-page could decide if the image that is being formed is the one he/she is interested into. This information is important for network browsing and saving computer bandwidth. The experiments reported that JPEG2000 is by far the champion in that category and that the transmission of as little as 10% of the final file

size is enough to carry almost 90% of the image "information", and the remaining 90% of the bytes transmitted only conveyed information "fine" details.

3 Document binarization

To allow researchers to have some access to the important bequest of letters of Joaquim Nabuco, the Joaquim Nabuco Foundation published a set of catalogues in chronological order with the date, sender, addressee and main subject of each letter or postcard. The most important letters were also summarized by the researchers. As already mentioned, the researchers who had access to those catalogues either went to the Joaquim Nabuco Foundation to acquire a copy of the microfiche of some of the documents they were interested into, or requested them to be sent via FAX. The quality of those images was very bad and only a few researchers, due to aforementioned paper degradation problems, had granted the permission to see the original documents. To solve this problem with the technical limitations of those days, the simplest way was to generate a digital library of binary images in one or two CDs. Then, the true color images had to be binarized. Image processing environments (such as Jasc Paint Shop Pro™ [Adobe], ImageJ [ImageJ]) offer a great variety of binarization filters. However, such softwares require specialized operators and that is not feasible to handle large quantities of documents. Besides that, about 10% of the scanned document images presented a feature not previously described in the technical literature, which was called *back-to-front* interference [Lins *et al*., 94]. The back-to-front interference occurs when the back face content of a document becomes visible on its front. Such interference appears on a document, whenever it is written (or printed) on both sides of translucent paper. In the case of historical documents, ageing is a complicating factor as paper darkens overlapping the RGB-distributions of the ink on each side and the paper.

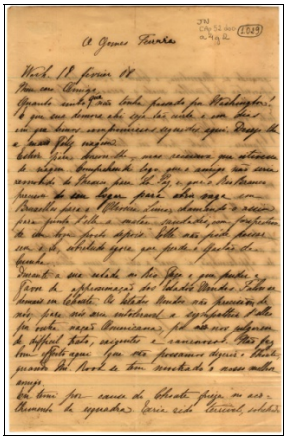


Figure 3: Historical Document with back-to-front interference.

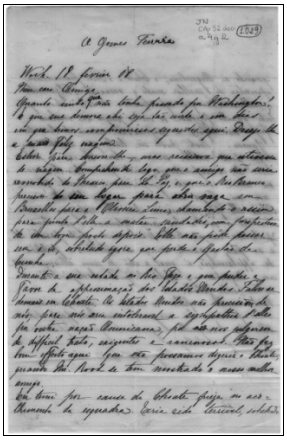


Figure 4: Gray-scale version of Figure 3.

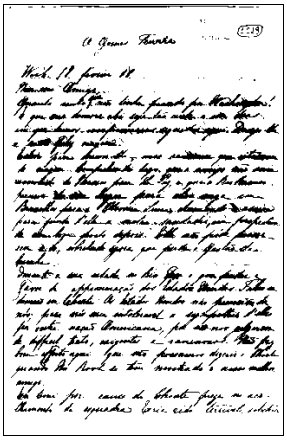


Figure 5: Binarized document of Figure 3.

Figure 3 provides an example of a document from Nabuco bequest with back-to-front interference and Figure 4 is the gray-scale version of the same document. The binarized version of this document generated by the direct application of the binarization algorithm by using Jasc Paint Shop Pro™ version 8 (Palette component: Grey values, Reduction component: nearest color, Palette weight: non-weighted) is completely unreadable, as one may observe in the part of Figure 5 in Figure 6.

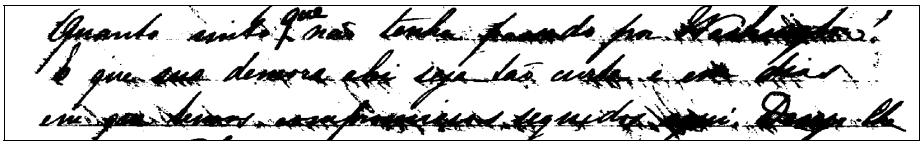


Figure 6: Zoom into part of document of Figure 5.

3.1 Binarizing Documents with Back-to-Front Noise

As already mentioned the author pioneered addressing the problem of back-to-front interference. Much later, other authors addressed the same phenomenon and called it *bleeding* [Kasturi *et al*., 02] and *show-through* [Sharma, 01]. In a document such as the one presented in Figure 3, one expects to find three color clusters corresponding to the ink in the foreground, the paper background and the trespassed ink (the back-to-front interference). Unfortunately, no image representation provided such clustering to allow the easy filtering out of the back-to-front interference. In the paper the noise was first described [Lins *et al*., 94], there is a suggestion of using mirror-filtering to remove it. In that case, both sides of the documents would be scanned and analyzed simultaneously, the dark pixels of higher intensity correspond to the foreground (ink or print) and are mapped into black, while the ones of lower intensity are back-to-front interference and are mapped together with background pixels into white. The inconvenience of this technique, as explained in [Lins *et al*., 94] is that aligning the two images is far from being a trivial task. If the document was folded as a letter, the perfect alignment is almost impossible. The mirror filtering technique was followed by [Sharma, 01] to remove the back-to-front noise that he called show-through, but he provides no solution to the image alignment problem, thus that is not an effective way of filtering it out.

Thresholding algorithms [Sankur and Sezgin, 04] have a wide range of applications in image enhancement and binarization. The first attempt to use thresholding to remove the back-to-front interference is reported in reference [Mello and Lins, 00] whose cut-off point is calculated based on the entropy [Abramson, 63] of the grey-scale version of the document. Such a strategy is still one of the most successful techniques for filtering out back-to-front interference are and [Mello and Lins, 02]. Although recent advances were made in finding efficient algorithms that yield good quality images [Silva *et al*., 08], a final solution to the filtering of back-to-front interference is still sought off. Several papers in the literature present different ways of solving the back-to-front interference problem. Some authors use waterflow models [Oha *et al*, 05] [Hyun-Hwa *et al*., 05], other researchers have used wavelet filtering [Tan *et al*, 02], but the technique of most widespread used is thresholding [Kavallieratou and Antonopoulou, 05], [Leedham *et al*., 02] and [Wong, 01]. All

algorithms report limitations in different kinds of images (too dark paper background, too faded printing, interference restricted to part of the document, etc.). A complex filtering scheme is proposed by Nishida and Suzuki [Nishida and Suzuki, 03] where first the foreground components are separated from the background and interference through locally adaptive binarization for each color component and edge magnitude thresholding [Cumani, 91]. Background colors are estimated locally through color thresholding to generate a restored image, and then corrected adaptively through multi-scale analysis along the comparison of edge distributions between the original and the restored image. Due to the nature of the documents in Nabuco's bequest edge detection seems to be of little help in eliminating show-through noise, thus the Nishida-Suzuki method seems to be unsuitable for this kind of document, although this is still to be borne out by experiments.

In general, analyzing the quality of images produced by filtering algorithms is far from being a trivial task. Visual inspection of the filtered images provides a weak quantitative assessment of the performance of the algorithms for binarizing documents with back-to-front interference. Thus, a quantitative method to assess the quality of algorithms for binarizing such documents was introduced in reference [Lins *et al*., 08], which generalizes and provides better comparison grounds than the one presented in [Lins and Silva, 07]. Reference [Lins *et al*., 08] simulates the back-to-front interference by adding an image with another for which the intensity of pixels is decreased by a value called the "*fade*". That paper analyzes the performance of eight algorithms ([Johannsen and Bille, 82], [Kapur *et al*., 85], [Mello and Lins, 00a], [Otsu, 79], [Pun, 81], [Silva *et al*., 06], [Yen *et al*., 95], [Wu *et al*., 98]). Their performances vary according to the strength of back-to-front interference. The algorithms by Yen-Chang-Chang [Yen *et al*., 95] and Kapur-Sahoo-Wong [Kapur *et al*., 85] only produce reasonable filtering for images with medium-to-weak back-to-front interference (*fade*≥110). In the most frequent noise region (*fade*≈80) these algorithms are unable to filter out significant amount of the back-to-front interference.

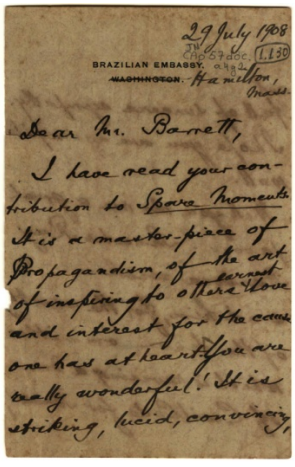


Figure 7: Document from Nabuco bequest.

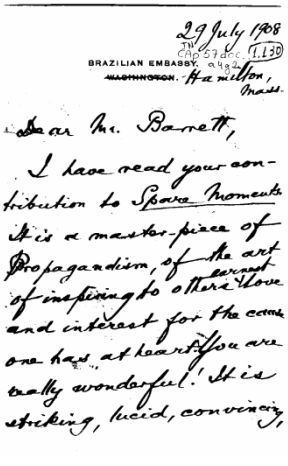


Figure 8: Algorithm by Silva-Lins-Rocha.

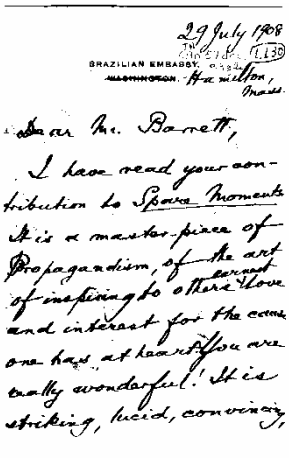


Figure 9: Algorithm by Silva-Lins-Martins-Wachenchauzer

The algorithms of Mello-Lins [Mello and Lins, 00a] and Otsu [Otsu, 79] are able to filter images with *fade* greater than 90 and 100, respectively, enhancing their performances as the noise weakens. For images with strong back-to-front noise ($30 \leq fade \leq 60$), the algorithm proposed by Wu-Songde-Hanqing [Wu *et al.*, 98] has good chances of performing well, however it tends to be greedy and remove part of the foreground information. The steadiest good performance in filtering out the bleeding noise is provided by Silva-Lins-Rocha algorithm. The eight algorithms studied in [Lins *et al*., 08] performed well in the cases of images with very low back-to-front noise ($fade > 120$). Otsu algorithm worked better than the others, however.

A more recent algorithm by Silva-Lins-Martins-Wachenchauzer, presented in reference [Silva *et al*., 08] enhances the results obtained by Silva-Lins-Rocha both in quality of binarized image as well as in performance. Figure 7 presents an example of another document from the Nabuco file with back-to-front interference. The results of its binarization by both algorithms are presented in figures 8 and 9.

The binarizarion schemes developed by the author and his research colla

4 Information retrieval

The historical file of documents of Joaquim Nabuco is of great importance to whoever studies the freedom of black slavery and the formation and the formation of the nation in the American continent. One way of allowing the access to some information on those documents was by printing catalogues with the summary of the documents and their related information. The recent book [Bethell and De Carvalho, 2009] is an undeniable testimony of the importance of Nabuco´s correspondence. In that book the authors focus in the "little-studied aspect of the struggle to abolish slavery in Brazil in the 1880s is the relationship between Joaquim Nabuco, the leading Brazilian abolitionist, and the British and Foreign Anti-Slavery Society in London. The correspondence between Nabuco and Charles Harris Allen, secretary of the Anti-Slavery Society, and other British abolitionists throughout the decade and beyond reveals a partnership consciously sought by Nabuco in order to internationalize the struggle. These letters provide a unique insight into the evolution of Nabuco's thinking on both slavery and abolition. At the same time, they offer a running commentary on the slow and (at least until 1887-88) uncertain progress of the abolitionist cause in Brazil." Those letters are kept by the Library of Rhodes House, Oxford, England.

At the time of the start of the Nabuco Project there were the catalogues [Andrade *et al.*, 1980] organized by the historians of the Joaquim Nabuco Foundation with the letters in their files. In [Lins *et al.*, 94] there is a report of a database which was initially developed by the Joaquim Nabuco Foundation in MicroIsis© [MicroIsis, 1988] a simple database system for libraries whose copyright belongs to the UNESCO, which can be used for free. The MicroIsis Nabuco database had all the information (kind of document, sender, addresee, date, keywords, language, summary, and transcription of the most important letters) in [Andrade *et al.*, 1980]. Although MicroIsis© could run image handling tools, the results obtained were not satisfactory. Experiments were made with vgif, a tool to show gif files. Unfortunately,

it was not flexible enough to present images that do not fit one screen. Another tool called Cshow allowed scrolling images of gif files, but due to its large space consumption it was not able to run under MicroIsis©. An extension to MicroIsis© was developed at the early days of the Nabuco Project allowing the document images to be part of the database.

The advent of the Internet yielded a real revolution in all areas of the human activity and the access to information was made much simpler. Storage space saving and faster network transmission time are the name of the game now. The technological restrictions met at the beginning of the Nabuco project are now overcome and the original motivation of making the historical documents available is reality today.

5 Conclusions

The aim of preserving the memory of the great thinker and statesman called Joaquim Nabuco for future generations was reached. His documents are now available at the site of the Joaquim Nabuco Foundation as a result of joint effort of two educational institutions and the vision and perseverance of Graziela Peregrino. The Nabuco Project was, to the best of the author knowledge, the pioneer work in Latin America in the development of a digital library of historical documents. The steps needed to accomplish its targets generated research motivation and several original and pioneer contributions to the area of document engineering of historical files. Although each file has his own particularities the techniques developed for the Nabuco bequest were used in the generation of other digital libraries and the set of tools generated form the HistDoc platform [Silva *et al*., 10a], which is publically available.

Acknowledgements

Research reported herein was partly sponsored by CNPq – Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Brazilian Government.

References

[Abramson, 63] N. Abramson, "Information Theory and Coding", McGraw-Hill Book Co, 1963.

[Adobe] Adobe Systems Inc. http://www.adobe.com, accessed on 20/01/2011.

[Andrade *et al.*, 1980] A. I. de S. L. Andrade, C. L. de S. L. Rêgo, T. C. de S. Dantas, Catálogo da Correspondência de Joaquim Nabuco 1903-1906, volume I 1865-1884, volume II 1885-1889, volume III 1890-1910, Editora Massangana, ISBN 857019126X, 1980. (Available at: www.fundaj.gov.br/geral/2010anojn/catalogo_nabuco_v2.pdf)

[Bethell and De Carvalho, 2009] L. Bethell, J. M. De Carvalho. Joaquim Nabuco, British Abolitionists, and the End of Slavery in Brazil: Correspondence 1880-1905, Institute for the Studies of the Americas, 2009. ISBN-13: 978-1900039956.

[Cao, 01] R. Cao, C. L. Tan and P. Shen, A wavelet approach to double-sided document image pair processing, Proc. Int. Conf. Image Proc. Oct. 2001.

[Cumani, 91] A. Cumani, "Edge detection in multispectral images", G. Models and Image Processing, 53(1):40-51, 1991.

[FUNDAJ] FUNDAJ – Fundação Joaquim Nabuco: http://www.fundaj.gov.br, accessed on 20/01/2011.

[Hyun-Hwa *et al*., 05] O. Hyun-Hwa, Kil-Taek Lim, and Sung-Il Chien. An improved binarization algorithm based on a waterflow model for document image with inhomogeneous backgrounds. Pattern Recognition 38 (2005) 2612 – 2625, 2005.

[IMAGEJ] IMAGEJ: http://rsbweb.nih.gov/ij/, accessed on 20/01/2011.

[Johannsen and Bille, 82] G. Johannsen and J. Bille, "A threshold selection method using information measures", ICPR'82, pp. 140–143 (1982).

[Kapur *et al*., 85] J. N. Kapur, P. K. Sahoo and A. K. C. Wong, "A New Method for Gray-Level Picture Thresholding using the Entropy of the Histogram", Computer Vision, Graphics and Image Processing, 29(3), 1985.

[Kavallieratou and Antonopoulou, 05] E. Kavallieratou and H. Antonopoulou, "Cleaning and Enhancing Historical Document Images", Intelligent Vision Systems, Springer-Verlag, LNCS 3708, pp. 681-688, 2005.

[Kasturi *et al*, 02] R. Kasturi, L. O´Gorman and V. Govindaraju, "Document image analysis: A primer", Sadhana, (27):3-22, 2002.

[Leedham *et al*, 02] G. Leedham, *et al*., "Separating text and background in degraded document images—a comparison of global thresholding techniques for multi-stage thresholding", 8th International Workshop on Frontiers in Handwritten Recognition, pp. 244–249, 2002.

[Lins and Machado, 04] R. D. Lins and D.S. A. Machado, A comparative study of file formats for image storage and transmission. Journal of Electronic Imaging (Springfield). , v.13, p.175 - 183, 2004.

[Lins and Rodrigues, 06] R. D. Lins and C. M. de S. Rodrigues. Assessing File Formats for Static Image Transmission In: International Telecommunications Symposium, 2006, IEEE Press, 2006. p.592 – 596.

[Lins and Silva, 06] R. D. Lins and J. M. M. da Silva, "Assessing Algorithms to Remove Back-to-Front Interference in Documents", ITS-2006, Fortaleza, Brazil, IEEE Press 2006.

[Lins and Silva, 07] R. D. Lins and J. M. M. da Silva, "A Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents", In: ACM Symposium on Applied Computing, 2007, Seoul, Korea, 2007. p. 610-616.

[Lins and Silva, 07a] R. D. Lins and J. M. M. da Silva, "Generating Color Documents from Segmented and Synthetic Elements", In: International Conference on Image Analysis and Recognition, 2007, Montreal, Canada. Springer Verlag, 2007. v. LNCS. p. 1217-1228.

[Lins *et al*., 94] R. D. Lins, L.G. Rosa, L.R. França Neto, M.S. Guimarães Neto, "An Environment for Processing Images of Historical Documents", Microprocessing & Microprogramming, pp. 111-121, North-Holland, 1994.

[Lins *et al*., 06] R. D. Lins, B. T. Ávila, and A. A. Formiga, "BigBatch: An Environment for Processing Monochromatic Documents", ICIAR2006, LNCS 4142, pp.886-896, Springer Verlag 2006.

[Lins *et al*., 08] R. D. Lins, J. M. M. da Silva, F. M. J. Martins. Detailing a Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents. Journal of Universal Computer Science. v.14, p.266 - 283, 2008.

[Mello and Lins, 00a] C. A. B. Mello and R. D. Lins, "Image segmentation of historical documents", Visual 2000, Mexico City, Mexico.

[Mello and Lins, 00b] C. A. B. Mello and R. D. Lins, "Generating Paper Texture Using Statistical Moments". In IEEE Silver Jubilee International Conference on Accoustic, Speech and Signal Processing – ICASSP 2000, pp. 100-105, IEEE Press, 2000.

[Mello and Lins, 02] C. A. B. Mello and R. D. Lins, "Generation of images of historical documents by composition". ACM Document Engineering 2002, McLean, VA, USA.

[MicroIsis, 1988] MicroIsis Manual, Micro CDS/ISIS, version 3.0, Copyright UNESCO, 1988.

[Nishida and Suzuki, 03] H. Nishida and T. Suzuki, "A Multiscale Approach to Restoring Scanned Color Document Images with Show-trough Effects", Proc. of. ICDAR 2003, 2003.

[Oha *et al.*, 05] Hyun-Hwa Oha, Kil-Taek Limb, Sung-Il Chienc, "An improved binarization algorithm based on a water flow model for document image with inhomogeneous backgrounds". Pattern Recognition 38 (2005) 2612 – 2625, 2005.

[Otsu, 79] N. Otsu, "A threshold selection method from gray level histograms", IEEE Transactions on Systems Man and Cybernetics, 9, 62–66 (1979).

[Pun, 81] T. Pun, "Entropic Thresholding, A New Approach", Computer Graphics and Image Processing, 16(3), 1981.

[Sankur and Sezgin, 04] B. Sankur and M. Sezgin, "A survey over image thresholding techniques and quantitative performance evaluation", Journal Electronic Imaging, 13(1), 146-165 (2004).

[Sharma, 01] G. Sharma, "Show-trough cancellation in scans of duplex printed documents", IEEE Trans. Image Processing, v10(5):736-754, 2001.

[Silva and Lins, 07a] J. M. M. da Silva, R. D. Lins, "Color Document Synthesis as a Compression Strategy", ICDAR 2007. IEEE Press, 2007. v. I. p. 466-470.

[Silva and Lins, 07b] J. M. M. da Silva, R. D. Lins, "A Fast Algorithm to Binarize and Filter Documents with Back-to-Front Interference", In: ACM Symposium on Applied Computing, 2007, Seoul, Korea. ACM Press, 2007. p. 639-640.

[Silva *et al.*, 06] J. M. M. da Silva, R. D. Lins and V. C. da Rocha Jr. "Binarizing and Filtering Historical Documents with Back-to-Front Interference", In: ACM Symposium on Applied Computing, 2006, Dijon. ACM Press, pp. 853-858, 2006.

[Silva *et al.*, 08] J. M. M. da Silva; R. D. Lins; F. M. J. Martins; R. Wachenchauzer. "A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference". Journal of Universal Computer Science, v. 14, p. 299-313, 2008.

[Silva *et al.*, 10] G. F. P. e Silva, R. D. Lins, S. Banergee, A. Kuchibhotla, M. Thielo. "A Neural Classifier to Filter-out Back-to-Front Interference in Paper Documents". In: 20th International Conference on Pattern Recognition, pp. 2415-2419, IEEE Press, 2010.

[Silva *et al.*, 10a] G. F. P. e Silva, R. D. Lins, J. M. M. da Silva. HistDoc - A Toolbox for Processing Images of Historical Documents. ICIAR 2010, Springer Verlag, 2010. LNCS 6112. pp.409-419.

[Su, 07] F. Su and A. Mohammad-Djafari, "Bayesian Separation of Document Images with Hidden Markov Model", 2nd International Conference on Computer Vision Theory and Applications, Barcelona, Spain, 2007.

[Tan *et al.*, 02] C. L. Tan, R. Cao, P. Shen, Restoration of archival documents using a wavelet technique, IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(10), pp. 1399-1404, 2002.

[Wang and Tan, 01] Q. Wang, C. L. Tan, Matching of double-sided document images to remove interference, IEEE CVPR2001, Dec 2001.

[Wu *et al.*, 98] L. U. Wu, M. A. Songde, and L. U. Hanqing, "An effective entropic thresholding for ultrasonic imaging", ICPR'98: International Conference on Pattern Recognition, pp. 1522–1524 (1998).

[Yen *et al.*, 95] J. C. Yen, F. J. Chang, and S. Chang. "A new criterion for automatic multilevel thresholding". IEEE Trans. Image Process. IP-4, 370–378 (1995).