Massive, free and reproducible grountruthed document image databases generation with DocCreator

Nicholas Journet, Boris Mansencal, Muriel Visani

Massive, free and reproducible grountruthed document image databases generation with DocCreator

Nicholas Journet[1], Boris Mansencal[1], Muriel Visani[2]
(1) Laboratoire Bordelais de Recherche en Informatique UMR 5800,
Université de Bordeaux, Bordeaux INP, CNRS
Email: {nicholas.journet, boris.mansencal}@labri.fr
(2) Laboratoire Informatique, Image et Interaction - L3i
University of La Rochelle, La Rochelle, France
Email: muriel.visani@univ-lr.fr

***Abstract*—Whether your research is focused on image restoration, layout analysis, text-graphic separation, binarization, OCR, etc. you need a groundtruthed database to train your method or to evaluate it. This article presents DocCreator, a multi-platform and open-source software able to create many synthetic image documents with controlled groundtruth. With DocCreator, you can create complete synthetic images choosing the text, font, background and layout to use, add various realistic degradations (bleed-through, light defect, paper deformation, ink degradation, etc.) on original images, or combine both to increase the size of your database. DocCreator comes as an online (easy to test version) and a desktop solution (fast calculation process, and no need to upload copyrighted data). DocCreator is useful for retraining tasks and to know precisely whether your algorithm is robust. It has already been used favorably and could help other DIAR researchers to produce and share groundtruthed databases.**

I. Introduction

One important challenge for the Document Image Analysis and Recognition (DIAR) community is to increase the proportion of open source software and public datasets. Currently, most DIAR experiments are difficult to reproduce, results are impossible to verify. It should become a habit to share our softwares and also our databases to improve the robustness and reproducibility of DIAR reasearch.

Of course some public DIAR datasets are available. Among the most popular ones, one can cite for printed documents: Washington UW3[1], LRDE[2], RETAS-OCR[3], PaRADIIT[4], etc; for handwritten documents IAM database[5], RIMES[6], etc; for graphical documents: architectural symbol database [7] or musical symbol database CVC-MUSICMA [8] for instance. The International Association for Pattern Recognition, for instance, gathered some interesting datasets[9] mostly used for different conferences competitions over the last two decades. However, very few of them are reliably annotated, copyright-free, up-to-date and easily available for downloading.

In this article we present DocCreator[10], a multi-platform and open-source (LGPL v3) software able to create myriad of synthetic image documents with controlled groundtruth. DocCreator can be especially useful for DIAR researchers or digital curators willing to train or evaluate DIAR tools. DocCreator aims at facilitating datasets exchange and help make DIAR research more reproducible. The first section of this article details a state of the art about groundtruth creation software programs created over the last 20 years. In the second section, we present the architecture of DocCreator and in the last section the features available in DocCreator for data augmentation or synthetic image creation.

II. Data augmentation or groundtruth generation

Many researchers choose to create their own groundtruth. For many basic tasks such as segmentation, it consists in annotating document images by manually cropping different elements in the documents (using rectangles or polygons), and possibly assigning labels to these elements. For text recognition tasks, the groundtruthing work usually consists of two steps: manual transcription of the text and alignment of text lines and image zones. In order to assist them in the tedious task of groundtruth creation, multiple software have been proposed during the last two decades.

[1]http://www.fki.inf.unibe.ch/databases/iam-historical-document-database/washington-database
[2]https://www.lrde.epita.fr/wiki/Olena/DatasetDBD
[3]http://ciir.cs.umass.edu/downloads/ocr-evaluation/
[4]https://sites.google.com/site/paradiitproject/
[5]http://www.fki.inf.unibe.ch/databases/iam-handwriting-database
[6]http://www.a2ialab.com/doku.php?id=rimes_database:start
[7]http://www.cvc.uab.es/ marcal/GREC-SEG/index.html
[8]http://www.cvc.uab.es/cvcmuscima/index_database.html
[9]http://tc11.cvc.uab.es/datasets/
[10]http://doc-creator.labri.fr

	Export	OpenSource	Desktop/Online	Level of Automation	Collaborative	Year
Software for creating groundtruth						
Pink Panther [1]	ASCII	n/a	desktop	no	no	1998
TrueViz [2]	XML	yes	desktop	no	no	2003
PerfectDoc [3]	XML	yes	desktop	?	no	2005
PixLabeler [4]	XML	no	desktop	no	no	2009
GEDI [5]	XML	yes	desktop	+	no	2010
DAE [6]	no	yes	online	++	yes	2011
Aletheia [7]	XML	no	online/desktop	+++	no	2011
Transcriptorium [8]	TEI-XML	no	online	++	yes	2014
DIVADIAWI [9]	XML	n/a	online	+++	n/a	2015
http://recital.univ-nantes.fr/	no	n/a	online	+	yes	2017
Algorithms for synthetic data augmentation						
Baird *et al*[10]	no	n/a	n/a	+	no	1990
Zhao *et al*[11]	no	n/a	n/a	+	no	2005
Delalandre *et al*[12]	no	n/a	n/a	+	no	2010
Yin *et al*[13]	no	n/a	n/a	++	no	2013
Mas *et al*[14]	no	n/a	n/a	++	yes	2016
Seuret *et al*[15]	no	yes	n/a	+	no	2015
Software for creating groundtruth with data augmentation capabilities						
DocCreator	XML	yes	online/desktop	+++	no	2017

TABLE I

TECHNICAL AND FUNCTIONAL CHARACTERISTICS OF EXISTING ANNOTATION OR DATA AUGMENTATION SOFTWARE. SOFTWARES ARE COMPARED ACCORDING 6 FEATURES : EXPORT FORMAT, DISTRIBUTION LICENCE, DESKTOP/ONLINE SOFTWARE, EASE-OF-USE (ONE “”+“” MEANS THAT THE USER NEEDS TO SET A LOT OF PARAMETERS OR TO UNDERSTAND THE ALGORITHMS TO USE THE SOFTWARE , FOUR “”+“” MEANS THAT ITS SIMPLE TO USE THE SOFTWARE ; MOST FEATURES ARE WORKING AUTOMATICALLY), COLLABORATIVE/CROWDSOURCING SOFTWARE , YEAR OF DISTRIBUTION

As detailed in Table I, some are fully manual standalone software, while others provide semi-automatic annotation modules. Some of the most recent solutions are based on an online collaborative platform. These software assist the user in creating the groundtruth associated with real documents, intrinsically limited in number because of acquisition procedures.

Even with a very good groundtruth creation software, this task remains tedious and time-consuming and most of these manually groundtruthed database are not shared by their owners. This inherently limits the amount of groundtruthed data available.

Another solution is available for getting (quickly and with lower human cost) large groundtruthed document image datasets. This solution, investigated since the beginning of the 90's [10], is to generate, from real images, synthetic images with controlled groundtruth. The authors of [11], [16] propose different approaches. Some consist in using a text editor (*e.g.* Word-office, Latex IDE) to automatically create multiple documents with varied contents (in terms of font, background, layout) where others consist in applying degradation models to original images [10], [17], [18].

With DocCreator, we propose to go further than existing synthetic generation methods. DocCreator is a semi-supervised software able, from a real document image, to create several synthetic document images that looks like the original one: same character font, same background and same structure but with any text users wish (see Figure 1). Besides, these synthetic images are created with an XML groundtruth.

DocCreator also proposes many image transformation algorithms that mimic the degradations observed in real images. These degradations might be due to poor printing materials, bad conservation procedures or unforeseeable events. For any document image, DocCreator is able to add bleed-through, ink spots, make some parts of ink characters disappear, distort the paper, add holes, etc. Combining these features allows to generate an infinity of document images, directly with the corresponding groundtruth. It should hopefully help to increase the number of shared databases.

III. DocCreator

A. How synthetic images are created

According to the needs of the DIAR researcher, it is possible to generate synthetic document images (and their groundtruth) by different ways. First possibility: if a researcher has real document images but without any groundtruth, DocCreator can generate synthetic images that look like the real ones, and of course, with the associated groundtruth. To do so, layout structure is extracted form the original image using a hybrid segmentation method [19]. Using an OCR, the characters are extracted and labeled (a GUI allows the user to correct the results). A clean background is obtained using an inpainting algorithm[11]. At last, given an input text, everything is assembled in order to build the final output, that is the created synthetic document image and the associated XML groundtruth. See Figure 1 (left part) for an example of a real image and Figure 1 (right part) for the automatic generation of its synthetic version. Desktop and online version proposes a GUI for making some tests before generating a huge amount of synthetic images (see figure 4)

[11]http://docs.opencv.org/2.4/modules/photo/doc/inpainting.html

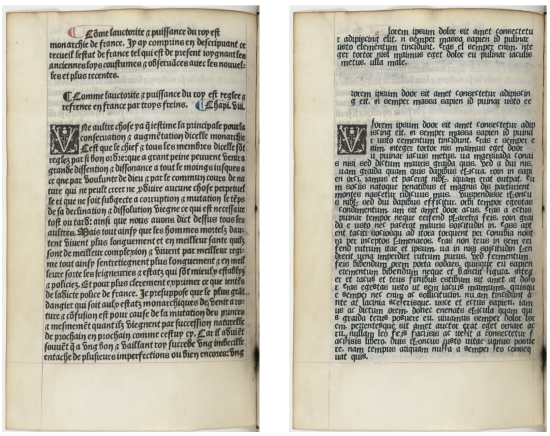


Fig. 1. An original ancient document image (left) and its synthetic version created using the font, background and layout of the original image. The text has been replaced by Lorem Ipsum. During the font extraction process, several instances of each character are extracted. Thus, when the font is used to write a character, DocCreator will randomly choose an image among different versions. This will make the final output look more realistic by reducing the strict uniformity between similar letters. A font extracted by the user can be used later to compose a new document

Second possibility: a researcher has a groundtruthed database but it is too small or not heterogeneous enough. DocCreator provides several degradation algorithms to augment the dataset. By degrading text ink, paper shape or background colors it is possible to create a representative document image database where many defects are present. This complete database is finally useful for precise performance evaluation or to provide multiple cases for retraining processes (in algorithms embedding a learning step).

B. Degradation models

1) Ink degradation: this model (presented in [20]) simulates the apparition of ink spots or the disappearance of ink pixels within the characters. It simulates either a bad print process or an age-related wear. See Figure 3.a for an ink degradation example.

2) 3D deformation: this model reproduces common paper distortions: small or big curvature, rotation, fold, etc. In [21] we detail how several real books have been digitized with a 3D scanner and how our algorithm is able to map any 2D document image on these 3D meshes. See Figure 3.b for results.

3) Adaptive blur: generating blur on image is quite easy. In DocCreator the originality is that the user provides himself a real image with blur. Using an adaptation of [22], DocCreator compute a "blur intensity" value. Then, using a dichotomic algorithm, we compute the size of the kernel of a Gaussian blur that once applied on the input image produces a blur similar to the chosen real blur image.

4) Bleed through: In [23] the authors proposed a method to erase the show-through defect (step by step ink from verso sheet appears on the recto side). We decided to adapt their model. By just giving an input recto image, an input verso image and the amount of wished degradation, DocCreator can reproduce the natural bleed-through defect. See Figure 3.c for an example.

5) Phantom Character: When a document is manually printed using wooden or metal character, after a long time, the characters are usually damaged. The printing process is thus not as clean as it should be: small ink parts are unintentionally printed when the character is pressed up against a sheet of paper. To simulate this defect, we have manually extracted many different phantom characters form real document images. The degradation algorithm is an adaptation of a patch algorithm [24] where a zone from another part of the document image is selected and copied within the patch (place where the phantom character is pseudo-randomly set). See Figure 3.d for an example.

6) Paper holes: Due to poor book storage conditions, holes may appear on documents. These holes might have different shapes, sizes and locations. DocCreator provides the possibility to add different kinds of holes in a real document image. As with the phantom defect, we manually collected many different hole examples from document images. According to the user wishes, black holes from the database are applied on a given document image. See Figure 3.e for an example.

C. Architecture

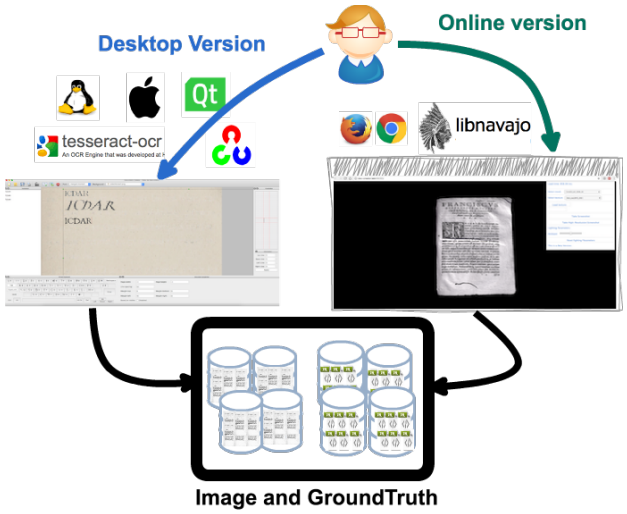


Fig. 2. Architecture of DocCreator.

DocCreator is an open-source and multi-platform software. DocCreator is available as a stand alone software or as an online demonstrator. The global architecture is detailed in Figure 2.

The user can choose between a desktop or online version. The desktop version comes with a Qt GUI, and has dependencies on OpenCV and Tesseract libraries. It can

be compiled/used on Linux, Apple OSX and Microsoft Windows. The desktop version allows to work locally on (potentially copyrighted) data. The online version is accessible through a web browser (tested on Google Chrome and Firefox). It does not need any installation, but requires to upload images to a server. The main part of the C++ source code is shared between the two versions. The core C++ is executed via libnavajo (C++ web server) for the online version. Some resources for creating a synthetic image (fonts, background, 3D models, etc.) are embedded with the desktop version and downloaded when required for the online version. With both versions, the user can complete his own database.

Source code, video presentation, installation instructions, online demonstrators are available on our website http://doc-creator.labri.fr/.

IV. Conclusion

With DocCreator any DIAR researcher can create complete groundtruthed images *and increase the size of its document image database. After 5 years of several collaborations and test campaigns, DocCreator (or databases created with DocCreator) have been tested by different researchers and used to publish [25], [26], [27], [28] proving its utility for performance evaluation or retraining tasks. DocCreator is on an open source ongoing project. We already plan to integrate new degradation models and another standard output format (TEI).*

References

[1] *B. Yanikoglu and L. Vincent, "Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation,"* Pattern Recognition, *vol. 31, no. 9, pp. 1191–1204, 1998.*

[2] *C. Ha Lee and T. Kanungo, "The architecture of trueviz: A groundtruth/metadata editing and visualizing toolkit,"* Pattern recognition, *vol. 36, no. 3, pp. 811–825, 2003.*

[3] *S. Yacoub, V. Saxena, and S. Sami, "Perfectdoc: A ground truthing environment for complex documents," in* Eighth International Conference on Document Analysis and Recognition. *IEEE, 2005, pp. 452–456.*

[4] *E. Saund, J. Lin, and P. Sarkar, "Pixlabeler: User interface for pixel-level labeling of elements in document images," in* International Conference on Document Analysis and Recognition. *IEEE, 2009, pp. 646–650.*

[5] *D. Doermann, E. Zotkina, and H. Li, "GEDI – a groundtruthing environment for document images," in* 9th IAPR International Workshop on Document Analysis Systems (DAS 2010), *2010.*

[6] *B. Lamiroy and D. Lopresti, "An open architecture for end-to-end document analysis benchmarking," in* Document Analysis and Recognition (ICDAR), 2011 International Conference on. *IEEE, 2011, pp. 42–47.*

[7] *C. Clausner, S. Pletschacher, and A. Antonacopoulos, "Efficient ocr training data generation with aletheia," in* Proceedings of the DAS 2014. *Tours, France: International Association for Pattern Recognition (IAPR), 2014.*

[8] *B. Gatos, G. Louloudis, T. Causer, K. Grint, V. Romero, J. A. Sánchez, A. H. Toselli, and E. Vidal, "Ground-truth production in the transcriptorium project," in* Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on Document Analysis Systems. *IEEE, 2014, pp. 237–241.*

[9] *H. Wei, K. Chen, M. Seuret, M. Würsch, M. Liwicki, and R. Ingold, "DIVADIAWI – a web-based interface for semi-automatic labeling of historical document images,"* Digital Humanities, *2015.*

[10] *H. S. Baird, "Document Image Defect Models," in* IAPR workshop on Syntatic and Structural Pattern Recognition. *Murray Hill, NJ, Jun. 1990, pp. 13–15.*

[11] *Z. Jiuzhou, " Creation of Synthetic Chart Image Database with Ground Truth," National University of Singapore, Tech. Rep., May 2005.*

[12] *M. Delalandre, E. Valveny, T. Pridmore, and D. Karatzas, "Generation of Synthetic Documents for Performance Evaluation of Symbol Recognition & Spotting Systems,"* Int. J. Doc. Anal. Recognit., *vol. 13, no. 3, pp. 187–207, Sep. 2010.*

[13] *F. Yin, Q.-F. Wang, and C.-L. Liu, "Transcript mapping for handwritten chinese documents by integrating character recognition model and geometric context,"* Pattern Recognition, *vol. 46, no. 10, pp. 2807–2818, Oct. 2013.*

[14] *J. Mas, A. Fornés, and J. Lladós, "An interactive transcription system of census records using word-spotting based information transfer," in* 12th IAPR International Workshop on Document Analysis Systems (DAS 2016), *2016.*

[15] *M. Seuret, K. Chen, N. Eichenbergery, M. Liwicki, and R. Ingold, "Gradient-domain degradations for improving historical documents images layout analysis," in* Document Analysis and Recognition (ICDAR), 2015 13th International Conference on. *IEEE, 2015, pp. 1006–1010.*

[16] *E. Ishidera and D. Nishiwaki, "A study on top-down word image generation for handwritten word recognition," in* Document Analysis and Recognition (ICDAR), 2003 7th International Conference on. *Washington, DC, USA: IEEE Computer Society, 2003, pp. 1173–.*

[17] *J. Zhai, L. Wenyin, D. Dori, and Q. Li, "A Line Drawings Degradation Model for Performance Characterization," in* Document Analysis and Recognition (ICDAR), 2003 7th International Conference on, *Edinburgh, Scotland, August 2003, pp. 1020–1024.*

[18] *J. Liang, D. DeMenthon, and D. S. Doermann, "Geometric Rectification of Camera-Captured Document Images,"* IEEE Trans. Pattern Anal. Mach. Intell., *vol. 30, no. 4, pp. 591–605, 2008.*

[19] *J.-Y. Ramel, S. Leriche, M. Demonet, and S. Busson, "User-driven page layout analysis of historical printed books,"* International Journal of Document Analysis and Recognition (IJDAR), *vol. 9, no. 2-4, pp. 243–261, 2007.*

[20] *V. Kieu, M. Visani, N. Journet, J. P. Domenger, and R. Mullot, "A Character Degradation Model for Grayscale Ancient Document Images," in* Pattern Recognition (ICPR), 2012 21st International Conference on, *Tsukuba Science City, Japan, Nov. 2012, pp. 685–688.*

[21] *V. Kieu, N. Journet, M. Visani, R. Mullot, and J. Domenger, "Semi-synthetic Document Image Generation Using Texture Mapping on Scanned 3D Document Shapes," in* Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, *Washington DC, USA, 2013, pp. 489 – 493.*

[22] *L. Lelégard, M. Bredif, B. Vallet, and D. Boldo, "Motion blur detection in aerial images shot with channel-dependent exposure time," in* ISPRS-Technical-Commission III Symposium on Photogrammetric Computer Vision and Image Analysis (PCV), *2010, pp. 180–185.*

[23] *R. F. Moghaddam and M. Cheriet, "Low Quality Document Image Modeling and Enhancement," in* International journal on document analysis and recognition, *vol. 11, no. 4. Berlin, Heidelberg: Springer, March 2009, pp. 183–201.*

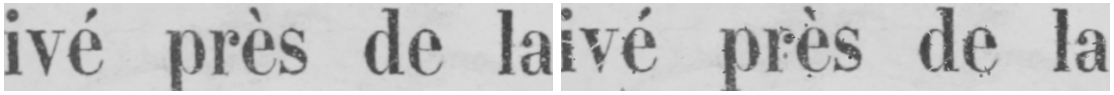
[24] *G. Shakhnarovich, "Learning task-specific similarity," Ph.D. dissertation, Massachusetts Institute of Technology, 2005.*

[25] *M. Mehri, P. Gomez-Krämer, P. Héroux, and R. Mullot, "Old Document Image Segmentation Using the Autocorrelation Function and Multiresolution Analysis," in* Proc. of the 20th DRR, *San Francisco, CA, USA, 2013.*

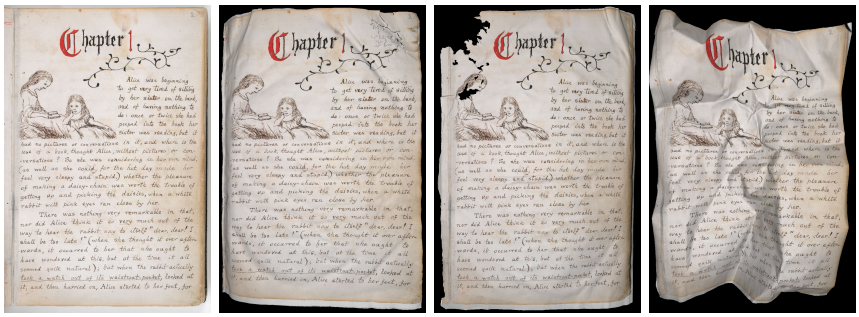
[26] *J. Calvo-Zaragoza, L. Micó, and J. Oncina, "Music staff removal with supervised pixel classification,"* International Journal on Document Analysis and Recognition (IJDAR), *pp. 1–9, 2016.*

[27] *I. d. S. Montagner, R. Hirata Jr, and N. S. T. Hirata, "A Machine Learning based method for Staff Removal," in* Int. Conf. Pat. Recog. (ICPR), *Stockholm, Sweden, 2014, pp. 3162–3167.*

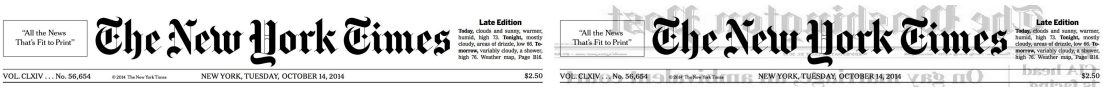
[28] *H. Wei, M. Baechler, F. Slimane, and R. Ingold, "Evaluation of SVM, MLP, and GMM Classifiers for Layout Analysis of Historical Documents," in* Document Analysis and Recognition (ICDAR), 2013 12th International Conference on, *Washington DC, USA, 2013, pp. 1220 – 1224.*



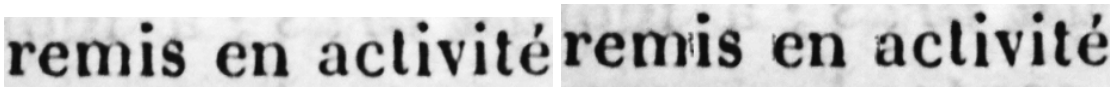
(a) Ink degradation on an old document



(b) 3D deformation: an original image and three 3D deformations



(c) Bleed-Through defect



(d) Phantom character apparition

(e) Hole added on an color document

Fig. 3. Example of degradation model available in DocCreator

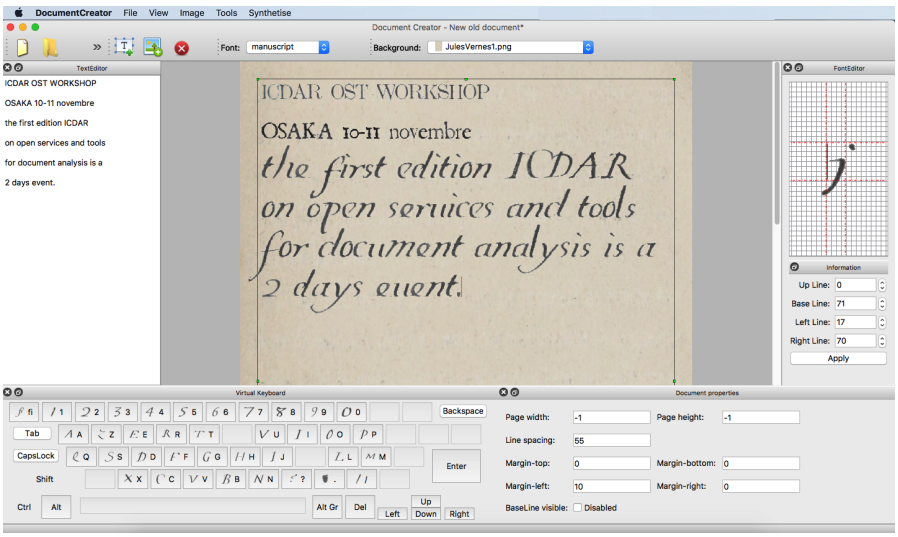


Fig. 4. GUI of DocCreator (desktop version). The central windows allows to input text to compose interactively synthetic document images using aivalable fonts and backgrounds. On the left, the original text is present. On the right, one can edit the properties of a character of the font. On the bottom, the mapping of the font onto the physical keyboard is presented via a virtual keyboard. One can also edit global properties of the document. Defect models are available in the top menu. Batch generation is available through "the factory" icon on the toolbar. An online version is also available at http://doc-creator.labri.fr/. The different features available on online/desktop latests versions are listed on the web site.