DFG-Project Venod –
Processing non OCR-able Documents

by Arved Hübler and Lothar Meyer-Lerbs

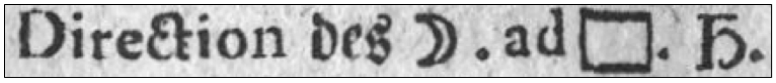

Figure 1: *Typical text, which is not processable with today's* OCR.

Even though today's OCR technologies are very efficient and convenient, there are many types of documents that are not suitable for OCR processing or for which the necessary effort for manual corrections would be disproportionately high. Until now there are no alternative processing mechanisms to integrate documents with mixed languages, with non-standardised character sets, uncommon symbols and special characters, or extraordinary fonts into the digital information cycle (Fig. 1). These documents are usually managed as scanned pixel files and reproduced in digital book systems as facsimile copies. This does not only involve problems with automatic indexing for search functions etc., it is also impossible to change line and page breaks. Especially the last fact becomes more and more important with regard to different formats of electronic reading systems (e-books).

A simple solution for this problem would be document-specific character encoding and preservation – an idea that was developed by the Institute for Print and Media Technology at Chemnitz University of Technology and enhanced in cooperation with the Center for Computing and Communication Technologies at the University of Bremen. The basic idea is to manage the characters of a text as graphic primitives instead of immediately converting them into abstract character values after the separation, as it is usually done by common OCR systems.

First of all, a document-specific font is generated from the recognised characters of the text. For this, the used characters are clustered by combining identical character forms (taking into account scaling and representation functions) in one reference character, vectorizing it and coding it in SVG format.

Hence a document-specific font is created with a pool of SVG characters, from which the entire text can be composed. Depending on the complexity of the text, this might result in alphabets with several thousand characters, especially when texts combine multiple writing systems.

Even though these raw fonts do not contain character values, but are only a graphical representation of the text in a linearly independent elementary character system, this approach has two big advantages:

1. Compared to a TIFF file, a text that is coded with an individual alphabet with SVG font is much more compact. The memory size of the document file is much smaller without data loss.
2. The text allows layout rearrangements. Especially line, column and page breaks as well as the insertion of tags is possible.

In the next step, labels can be assigned to the individual characters of the font. This is usually done manually with a suitable user interface. The values that are assigned to the characters of the individual SVG fonts either comply with existing alphabets or must be newly defined, if necessary. After this step, the texts are not only equivalent to OCR texts in that they are searchable, but also have a document-specific font apart from the transcription into a commonly used alphabet, which allows for direct and flexible representation of the original without reference to the TIFF file. Additionally, the manual assignment of character values to the identified symbols is much easier, more secure and faster than the acquisition of OCR procedures and the correction of substitutes in OCR texts.

A prototype for universal character encoding has been developed in Chemnitz and Bremen. At the moment, tests are carried out with increasingly complex text types.

*

We especially thank our former colleague Stefan Pletschacher, now at the School of Computing, Science & Engineering at the University of Salford/UK, for important contributions to this work.

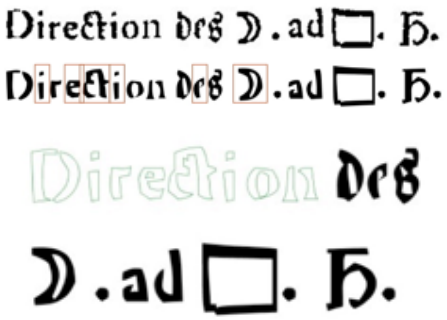

Figure 2: *Upper line: Original pixel image after scanning. Center line: The same text with a document-specific SVG font. Some letters are framed for clarification. Lower line: Changed layout of the text (font size, line break, the first letters as outline)*

Prof. Dr. Arved C. Huebler
Institute for Print & Media Technology (pmTUC)
at Chemnitz University of Technology
Reichenhainer Str. 70
09126 Chemnitz
Germany

Telephone: +49 (0)3 71 5 31-2 36 10
Email: arved.huebler@mb.tu-chemnitz.de

*

Lothar Meyer-Lerbs
Center for Computing and Communication Technologies (TZI)
University of Bremen
Am Fallturm 1
28359 Bremen
Germany

Telephone: +49 (0)4 21 2 18-87 97
Email: lml@tzi.de