

FROM DOCUMENTS TO PRESENTATION: THE MEMORIAL BOOK OF RAVENSBRÜCK

ВІД ДОКУМЕНТІВ ДО ЇХ ПРЕЗЕНТАЦІЇ: КНИГА ПАМ'ЯТІ РАВЕНСБРЮКА

Wolfgang Schade
Widis GmbH
Albert-Einstein-Straße 16, 12489 Berlin, GERMANY
schade@widis.de

Gerd Stanke
GFal e. V.
Rudower Chaussee 30, 12489 Berlin, GERMANY
stanke@gfai.de

Vladimir Kijko, Vyacheslav Matsello
International Research and Training Centre for Information Technologies and Systems of
National Academy of Sciences and Ministry of Science and Education of Ukraine
Prospekt Akademika Glushkova 40, 252022 Kiev, UKRAINE
matsello@image.kiev.ua

Анотація

Наведено повний технологічний цикл для видання та/або презентації старих архівних документів низької якості. Описано методи, що були успішно застосовані для отримання, обробки та подання таких старих архівних матеріалів, які не відповідають сучасним видавничим вимогам. Було розроблено методи попередньої обробки, аналізу та перетворення зорової інформації для архівних документів 20-го сторіччя. Обговорюються результати відокремлення фону, розпізнавання типу документу, застосування програм розпізнавання тестів, а також подання інформації в історичному контексті. Розроблені методи були успішно застосовані підчас видання "Книги пам'яті Равенсбрюка". Базові дослідження були виконані в рамках спільного проекту за підтримки Сенату міста Берлін. Стаття присвячена повному технологічному циклу. Обробка зображень базується на [1, 4].

Abstract

An overview on the whole production cycle for book edition and/or presentation using old bad conditioned archive material is given. Successfully applied methods for data acquisition, processing and presentation from old archival material not underlying standardised production rules are explained. Different material depending methods for image analysis, pre-processing, transformation of pictorial information etc. for archive documents from the 20th century are developed. Results of background removal, form recognition, OCR-application, data dependent processing methods and representation of data in a historical context are discussed and presented. Solutions developed were successfully tested during the edition of the "Memorial Book of Ravensbrück". The basic research was done in a co-operative project granted by the Senate of Berlin. The paper refers the complete production cycle whereby the image processing part bases on [1, 4].

The historical and technical background

The unhappy 20th century history of Europe and in particular of Germany is well known and no more commented here. But, there are different reasons to rework the history of that time by help of documents left in order to make the individual destinies more clear, to satisfy compensation claims or simple to follow the Way of the Cross. The documents existing for processing show quite different representations, typed lists, hand filled forms, commented maps, records, computer files etc.

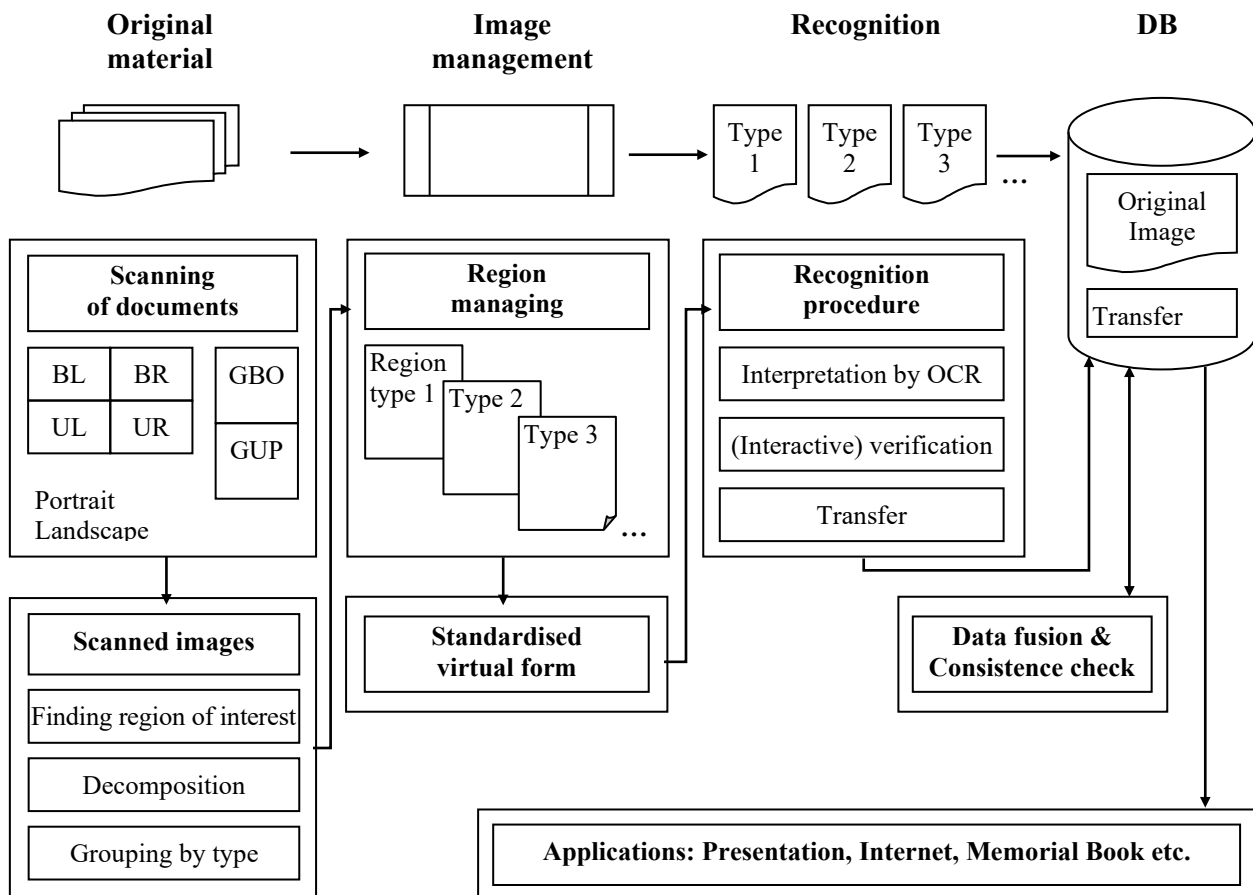
Most of these documents are bad conditioned and need preprocessing for further automatic processing by commercial programs.

There were two points stimulating the investigations. First a project for editing of the “Memorial Book of Ravensbrück” by a group of specialists lead by a German historian and second a cooperative project of five Berlin located institutions (WIDIS, GFaI, Graphikon, IPK, FHTW) for the processing of old archival material, granted by the Senate of Berlin. The results planed were the completion of the memorial book as well as a portfolio of common applicable processing procedures for archival documents from the last century.

Beside the help for the memorial place „Ravensbrück“, a former concentration camp for women in Germany, there came out results as presentation of materials via internet and information points, the possibility to answer letters of prisoners, asking for a proof of their forced labour, and the publication of „memorial book“ of victims. Starting point was the increasing interest of archives and libraries in recording and digital processing of documents for a better access, scientific analysis and presentation. Especially this scientific topic was supported by the Senate of Berlin. The formal obstacles to do the work in the environment of bureaucracy regulation provoked by the federal structure of Germany and the lack of money will be not mentioned in detail. Of interest is, we had to follow the safety regulations of archives and to manage the international co-operation. It was sometimes not so easy.

The workflow to implement

This way a technology for automatically data gathering and managing from typed or hand-written forms exists. The following schema shows the general system architecture.



Workflow:

Scanning – Form Management – Virtual Form – Recognition – Consistence Check – Data Base – Applications

Information Saving and Processing in Data Bases

On request of the involved historians the registered data had to be saved close to the source. That's why for each document class (more than 100!) a separate sub data base was necessary. Including the manual registered information the data base includes up to now about 200 000 data records.

The original saved data were uncomplete and inconsistent, some data are repeatedly registered (prisoners' names e.g. arise in transport lists as well as in new admission lists). To unite all data registered in different data bases, tool for data comparison has been developed.

Just imagine the different style and spelling of names. There are not only several ways to fill in the names into the referring field but also differences in the spelling of the names:

Jozef van der Broek / van der Broek, Jozef / v.d. Broek, Jozef / Broek, van der, Jozef

Therefore possibilities were developed to discover persons through the phonetic pronunciation in order to avoid double registration of persons. Special "Soundex-Algorithm" (see table showing the principle of the „Soundex-Algorithm“) was implemented as well as a comparison procedure for faulty or incomplete name registration.

Adam	Adam	A4m
Adamska	Adam3ka	A4m3k
Bednarkiewicz	Bednarkiewicz	Be4nrkiewi3

These problems indicate, that the solution depends on quite powerful image processing procedures, capable to handle different types of forms, forms with substantially background noise, hand filled information etc., these problems will be considered in detail.

The material accessible for the development

The materials came from different sources showing different technical quality and were partly in a very bad preserved. In particular there were: certificates of the registry office Fürstenberg, file cards of the special registry office Arolsen (International Red Cross), transport lists of prisoners from Besançon, index file cards about prisoners (German Central Archive Berlin-Hoppegarten), files of data bases from Ukraine and Belorussia and others.

KL: Konzentrationslager Stutthof

Häftlings-Personal-Karte

Fam.-Name: Lubowska Überstellt am: _____ an KL: _____
 Vorname: Hanna am: _____ an KL: _____
 Geb. am: 1879 in: _____
 Stand: Kinder am: _____ an KL: _____
 Wohnort: _____ am: _____ an KL: _____
 Strasse: _____ am: _____ an KL: _____
 Religion: mos. Staatsang.: Judin am: _____ an KL: _____
 Wohnort d. Angehörigen: _____ am: _____ an KL: _____

Eingewiesen am: 21. 11. 43 am: _____ an KL: _____
 durch: Sipo Bialystok am: _____ an KL: _____
 in KL: Stutthof am: _____ an KL: _____
 Grund: _____ Entlassung: _____ durch KL: _____
 Vorstrafen: keine am: _____ durch KL: _____
 mit Verfügung v.: _____

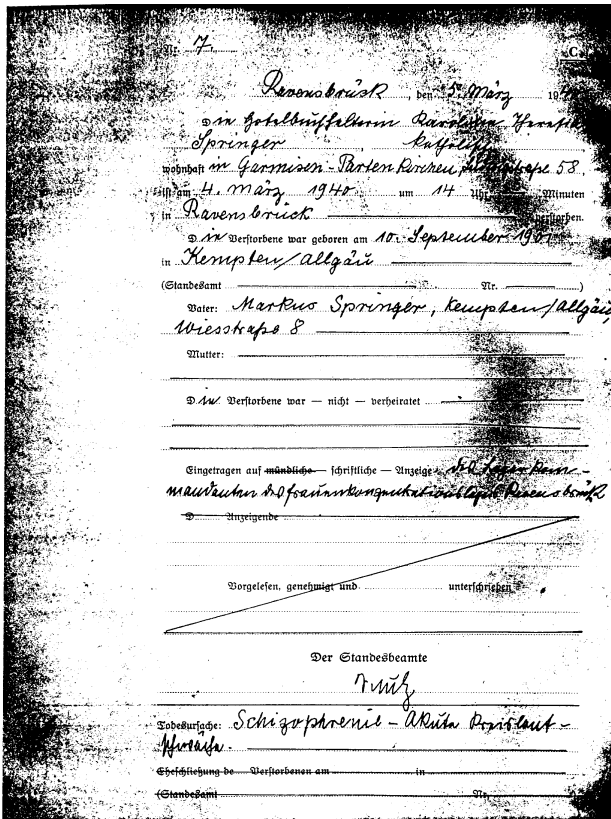
Strafen im Lager:
 Grund: Am 10. Auf dem Transport verstorben Art: _____ Bemerkung: _____
am: 21. November 1943
 eingearbeitet am: 24. - - -

KL 54 43-500000

Among this material were typewriter written lists, typewritten forms, hand filled forms, pre-printed forms, maps, computer files. The following images give an idea about the different types of documents and their poor quality.

Among this material were typewriter written lists, typewritten forms, hand filled forms, pre-printed forms, maps, computer files. The following images give an idea about the different types of documents and their poor quality.

File card from Stutthof:
 typewritten, handwritten remarks, stamps
Mixed font recognition



Certificate of the registry office Ravensbrück:
No chance for OCR



File cards of Arolsen, different colors, typed:
OCR hopeful

All these documents show a special history based on the order from the German War Ministry, to record the prisoners and their profession in all concentration camps for controlling the forced labour. Hollerith-machines were used since the second half of 1944, 120.000 index cards in Hoppegarten and 40.000 cards in Warsaw, mentioned in the book „IBM and the holocaust“, are seen as a „proof“ for the co-operation between IBM and the fascist Germany.

Abstellung von K. L. Mittel

Arbeiter D. G. 18

1.	Leonard	13 491	5. 4.97
2.	Erre	13 492	25. 2.20
3.	Leonard	13 493	4. 9.9e
4.	Theodor	13 494	18. 4.05
5.	Leonard	13 495	2. 9.2e
6.	Victor	13 496	14. 8.20
7.	See	13 497	10. 4.18
8.	Jean	13 498	27. 2.24
9.	Elmer	13 499	3. 5.08
10.	Edmond	13 500	5.11.20
11.	Vanilij	13 501	15. 3.07
12.	Vanilij	13 502	28. 2.19
13.	Roman	13 503	30. 7.86
14.	Alexander	13 504	22. 5.91

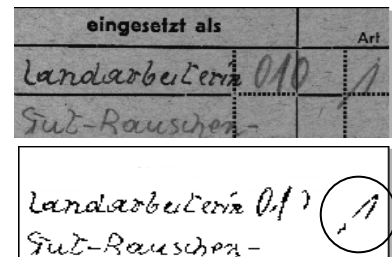
Besançon, typed Xeroxcopy of 3rd carbon copy
 Background (very) dirty:
A chance for OCR?

The technical (image processing) problems

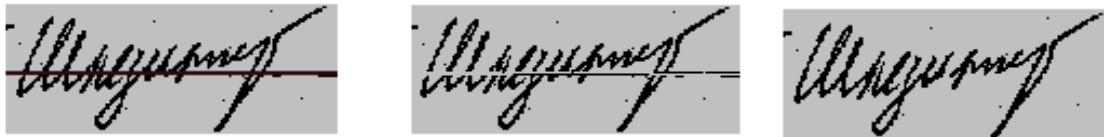
The reasons why image processing has to play an important role for automated analysis of old documents base in different states of imperfection. Many of existing basically powerful OCR systems for typed and/or for hand written documents rely on the presumption that these documents are produced meeting a certain minimal quality level. But, this is usually not true for documents from archives. Therefore pre-processing procedures must be developed in order to transfer the original document into a form and quality manageable for commercial OCR-systems.

Content extraction for hand written and/or typed forms

As mentioned above, there exist powerful OCR-systems on the market. Most of them integrate modules for form-processing. That is necessary because the entire recognition modules handle only letter or word like information without noise, disturbing information as lines or dashed lines etc. The example in the following figure is taken from [4] and gives a good imagination what background removal and line removal can perform. But, there remain imperfections (see the disruption in figure "1").



A similar problem appears if the pre-printed form contains lines with colors similar to the color of writing tools used. Therefore additional solutions were developed and adopted to tackle even such cases, it means to remove color lines, but to preserve the lines of the filled in text. The figure below shows the result for a color form where the filled in information have the same color as the form lines. A quite good reconstruction of the original filled content is reached, no line disruption remains after pre-processing (right). These and further results were reached by a substantial usage of knowledge provided by scientists of the International Research and Training Centre of Information, Technologies and Systems Kiev in the framework of [2].



A second useful procedure was developed by a more or less intelligent processing of forms using an online adjusting of the processing area for recognition. This is relevant because in pre-printed forms occur several significant words as "name", "place of birth", "date" etc. which describe the content expected in this row. But sometimes synonyms are used, that's why areas with identical semantic information will be found at different places in the row or even in the whole document. A special procedure recognises the semantic meaning of the pre-printed text, defines the location and afterwards transfers the found area to the standard recognition procedure. Index card files of Arolsen, one of the greatest archives, confronted us massively with this problem. A simi-

Familienname:	van der AA, geb. Meertens
Vorname:	Bramine Felcoline
Beruf:	Relig.:
wohnhaf in	
Sterbefag	19.2.45.
Uhr	Min.
Sonderstandesamt Arolsen Sterbebuch Jahr 1964 Abt. BB Nr. 795	
Sterbeort:	Bergen-Belsen Ravensbrück
geboren	17.9.94. in Batavia

lar question arose according the detection of erased information (comp. figure), which is handled by a skipping module for erased information.

One of the hardest problems came up, when masses of documents were processed automatically by powerful and flexible OCR-systems capable to handle filled forms. Why, the system must be acquainted with the expected form structures. These structures were more or less constant in all the thousands of forms, but the detailed geometric positions on the sheets differ. See fig. below, the "confusion" of different grey levels represents in principle identical but in reality different forms.

The image shows a 'Häftlingskarte' (Prisoner Card) with various sections. The top section contains the title and some administrative fields. Below that, there are several rows of handwritten text and numbers, including what appears to be a name, a date, and several identification numbers. The bottom section of the form is a grid-like structure with multiple columns and rows, likely used for tracking or recording specific data points for each prisoner.

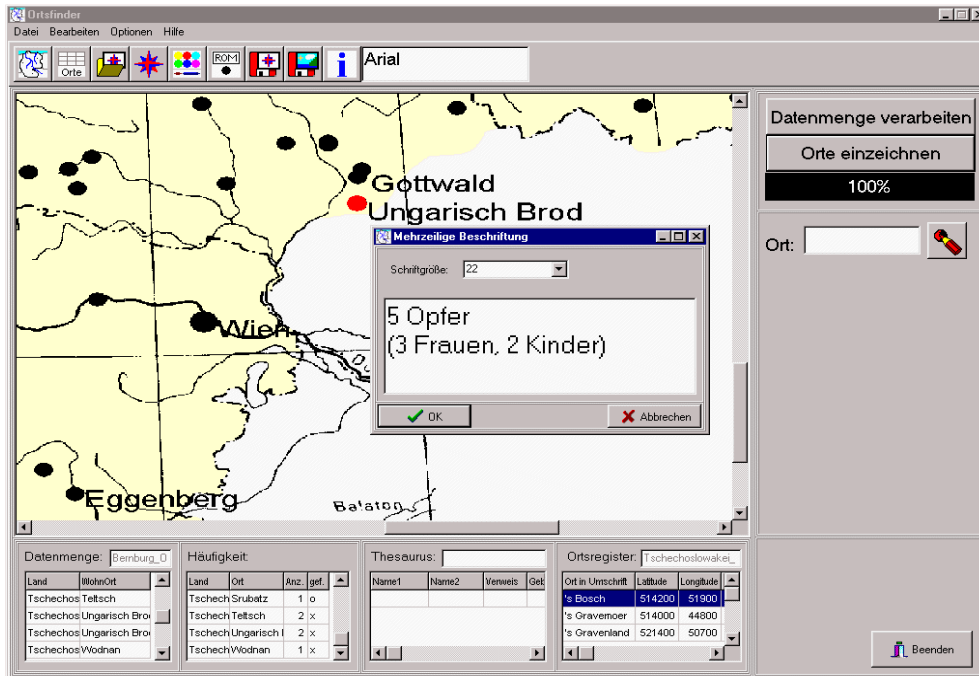
The solution developed was a reliable form recognition pre-processor. This processor allows to classify the types of forms as well as to cut out single forms of combinations of several forms scanned at once. It was necessary to use such combinations in order to scan several forms in one pass and by this to speed up the scanning process. The problem behind this will be more clear in the figure. If the OCR gets the shadowed (too large) area for the recognition of the code number there will be a wrong result 1094352 instead of 094352. The reason lies in the left stroke caused by distorted and shifted imprint position when a false form type is used. Such a bad coding can cause big troubles in the database. The form classification module avoids the reasons for such misinterpretations by adaptive imprint classification and alignment.



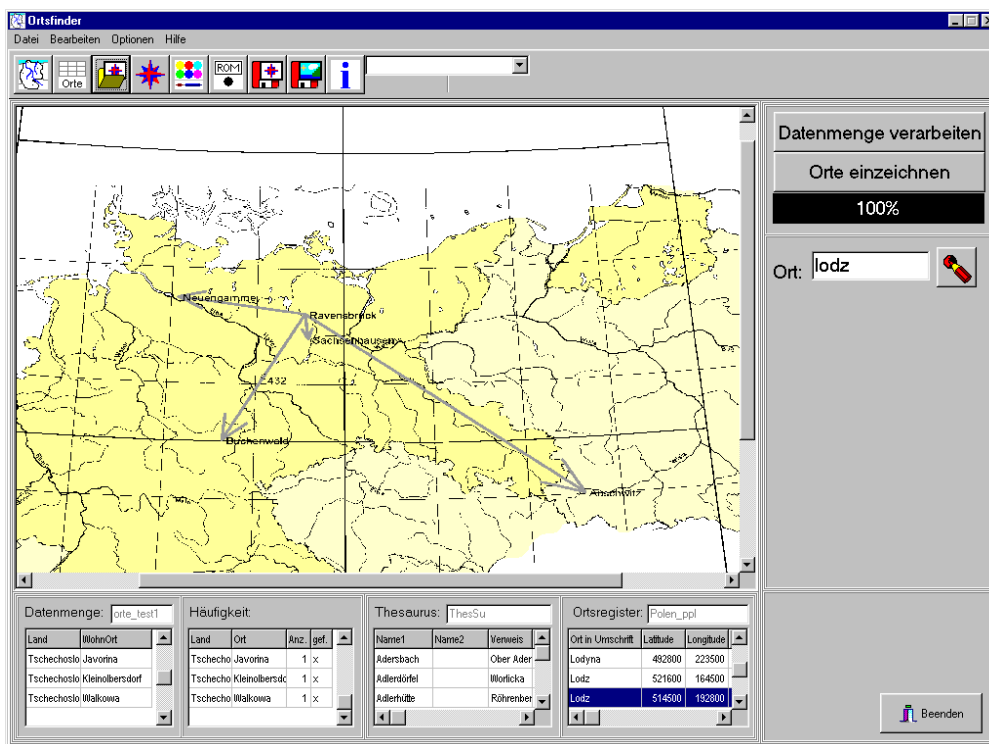
The whole workflow integrating the special pre-processing procedures for the scanned images consist of the following steps: find the areas of interest for recognition, cut out the areas from the images, build up an image as a „virtual file card” with exact known positions of printed areas, filled areas and marks, application of background and line removal algorithms. The resulting "standard image" is given to a standard OCR.

Presentation of the Contents

In result a memorial book for “Ravensbrück” was presented including names of 10,000 victims. Since only half of the available documents is digitised and processed this memorial book proves to be previously and has to be completed. In order to present the data at info points, a surface for the visualisation of the prisoners’ places of birth and residence in historic maps was developed. For this the different place names (e. g. the German names for today’s Polish, Czech or Russian towns) were exploited. The number of prisoners is shown through variable place marks.



The places can be commented with additional information



Visualisation of prisoner transports

The Results and further development

By means of algorithms developed the multiple of the 65000 prisoners records integrated in the database could be compared and combined. The edited resulting memorial book contains more than 10.000 names of victims, for this over 17.000 data sentences were analysed.

There is a lot of further work using the recognised information at the level of databases. For instances tackling of different kinds of name writing, the processing of Cyrillic text from Ukraine sources etc. These results are described already more in detail in [3]. There can be found also the convincing approach how to present results in an geographical and historical context

Actually, in Ravensbrück most of existing source oriented data bases are combined to as so called Meta data base.

Even if a part of work is done there are many open questions addressed to ongoing research projects as (for example) MEMORIAL in the EC's Fifth Framework Programme IST. In cooperation with partners in Germany, Poland, UK, and Israel we are dealing with special problems of layout analysis as recognizing stamps and areas of hand written remarks or signatures. By using special filters during the scan procedure and using the colour information of images the recognition of blurred type written characters shall be improved. Moreover, basic methods for an exchange of information, available in existing data bases on different memorial places will be developed.

References

- [1] Report of the WTZ-agreement Germany/Ukraine UKR-007-97 "Bildinhaltsgesteuerter Zugriff zu Bilddatenbanken", Ukrainian part, Berlin, GFai, 2000
- [2] V. Kiiko, V. Matsello and G. Stanke, "Bildinhaltsgesteuerter Zugriff zu Bilddatenbanken", BMBF-Förderprojekt UKR-007-97, GFai-Jahresbericht 1999 pp. 42.
- [3] W. Schade, "Vom Inhalt zur Präsentation - die Häftlingsdatenbank der Mahn- und Gedenkstätte Ravensbrück", Conf. EVA 2001 Berlin, 14-16 November 2001
- [4] W. Schade, G. Stanke, From Documents Contents to Presentation - the data base of prisoners published as the "Memorial Book of Ravensbrück", Conf. EVA 2002 Florence, 22 March 2002
- [5] Ch. Feist and G. Stanke, "Entwicklung von Verfahren zur automatischen Erfassung, Sicherung und Präsentation von Archivmaterialien und von entsprechenden Datenstrukturen", Berlin, GFai-Jahresberichte 1999 pp. 30, 2000 pp. 26.