TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Information Technology

David Landau

DIGITIZING TEXT HERITAGE

Master of Science Thesis

The subject was approved by the
department council on the 4th of June, 2003.
Supervisors: Professor Karen Egiazarian
Professor Ari Visa

Preface

This master's thesis was prepared in the Institute of Signal Processing at Tampere University of Technology. My research was funded by the Digital Media Institute (DMI), and my trip to Uppsala materialized with the help of a grant from the Finnish-Swedish Cultural Foundation. I am very grateful to these institutions for their support.

This work is part of my search for the elusive etymon of the Finnish word *juhla* 'celebration.' The research is an interdisciplinary quest which includes, in this instance, topics in Germanic philology, digital image processing, multimedia, and software engineering.

I am grateful to my supervisors, Professors Karen Egiazarian and Ari Visa for their support, comments and assistance. I would also like to thank Professor Reino Kurki-Suonio, my tutor for the doctoral dissertation, for his input. During the work I received help and comments from several people whom I wish to thank for their contributions: Radu-Ciprian Bilcu, Jaco Geldenhuys, Jukka Huhtamäki, Mika Katara**,** Lars Munkhammar, Vesa Peltonen, Ari-Pekka Suntionen, Jarmo Toivonen, Tomi Vesanen and Krister Östlund. Lastly, special thanks go to my wife, Tuula, and my children, Tiina and Daniel, for their support.

Tampere, December 2003.

David Landau
dla@cs.tut.fi

Table of Contents

Preface....ii
Table of Contents....iii
Abstract....iv
List of Abbreviations....vi
1. Introduction....1
2. Digitizing Cultural Heritage....4
3. Database and Data Mining: Introducing XML....8
  3.1. Working with XML....10
4. Digitizing Manuscripts and Old Books....14
  4.1. The Scanning Process....15
  4.2. Scanning Resolution....16
  4.3. File Formats....17
  4.4. Compression Considerations....20
  4.5. A Comparative Study....21
5. Digitizing Text Heritage at Uppsala University Library....23
  5.1. Scanning Calculations....26
  5.2. On-line Display....28
6. The Use of Digital Filtering for Enhancing Text Images....31
  6.1. Histogram Equalization....34
  6.2. Mean Filters....36
  6.3. Median Filters....38
  6.4. Highpass and Lowpass Filters....40
  6.5. Derivative Filters....43
  6.6. Thresholding....45
  6.7. Image Fidelity Criteria....47
7. The use of X-rays, Ultra-violet Lightening, etc....49
8. Surface Analysis....52
9. Conclusions....55
Notes....57
References....59

Abstract

TAMPERE UNIVERSITY OF TECHNOLOGY
Department of Information Technology
Institute of Signal Processing

Landau, David: Digitizing Text Heritage
Master of Science Thesis, 60 pages.
Supervisors: Prof. Karen Egiazarian, Prof. Ari Visa
Funding: Digital Media Institute, the Finnish-Swedish Cultural Foundation.
Key words: text (textual) heritage, Gothic language, the Codex Argenteus, XML, multimedia, data mining, digital image processing.

Digital technology offers a new way of preserving text heritage. Instead of copying it by hand, reprinting it, photographing it for microfilms or photocopying it on paper, with the new technology text can be scanned and preserved efficiently as a 'spaceless' and 'timeless' entity, provided that technology is indeed kept up to date at every moment of time.

This thesis details my ideas as to how the digitizing process of text heritage should be approached. The material to be scanned is text concerning the study of the Gothic heritage preserved at the Library of Uppsala University, Sweden. In the first stage, the idea is to digitize books written between 1569-1805. In the second stage, not included in this thesis, the process will be extended to books written till 1927.

The thesis deals with the technical aspects of such an undertaking. The central theme is flexibility; as technology changes rapidly, one must be on constant alert and design the project so that changes can be made swiftly. At the same time, it is important to try to figure out features that may remain useful in the foreseeable future and as such will not need to be altered as technology changes. One such tool is the XML standard which I propose to be used for constructing the database and the multimedia access to the digitized material.

The thesis is divided into three parts. The first one surveys the technical aspects of creating a digital collection of text heritage. The second part deals with the characteristics of the material to be included. The third part is a survey of digital image processing methods relevant to such an undertaking. My conclusion is that such project is indeed feasible and can be accomplished with no great difficulties.

Tiivistelmä

TAMPEREEN TEKNILLINEN YLIOPISTO
Tietotekniikan osasto
Signaalinkäsittelyn laitos

Landau, David: Tekstiperinnön digitointi.
Diplomityö, 60 sivua.
Tarkastajat: Prof. Karen Egiazarian, Prof. Ari Visa.
Rahoitus: Digitaalisen median instituutti (DMI), Suomalais-ruotsalainen kulttuurirahasto.
Avainsanat: tekstiperintö, gootin kieli, Codex Argenteus, XML, multimedia, tiedonlouhinta, digitaalinen kuvankäsittely

Digitaalinen tekniikka tarjoaa uuden mahdollisuuden tekstiperinnön säilyttämiseen. Aikaisemmin ihmiset kopioivat tekstin käsin, tulostivat sitä uudestaan, kuvasivat sitä mikrofilmille tai ottivat valokopiota. Uuden tekniikan avulla teksti voidaan skannata ja säilyttää tehokkaasti 'tilattomana' ja 'ajattomana' kokonaisuutena, mutta samalla tekstin lukemiseen käytettyä tekniikkaa täytyy pitää ajan tasalla.

Tämä diplomityö on suunnitelma pilottiprojektia varten. Sen tavoitteena on esittää yksityiskohtaisesti, millä tavoin tekstiperintöä voitaisiin digitalisoida. Työssä käytettävä tekstimateriaali, jolle suoritetaan skannaus, liittyy gootin kielen tutkimuksiin ja sitä säilytetään Uppsalan yliopiston kirjastossa. Ensimmäisessä vaiheessa on tarkoitus digitalisoida kirjat, jotka on kirjoitettu 1569-1805 välisenä aikana. Toisessa vaiheessa, joka ei kuulu tähän työhön, projekti tulee sisältämään muut samaan aiheeseen kuuluvat kirjat, jotka on kirjoitettu vuoteen 1927 mennessä.

Tämä työ käsittelee tällaisen tehtävän tekemiseen liittyviä teknisiä yksityiskohtia. Keskeinen teema on joustavuus. Kun tekniikka muuttuu nopeasti, on oltava jatkuvasti valppaana ja suunniteltava projekti sillä tavoin, että muutoksia voidaan tehdä nopeasti. Samalla on tärkeää huomioida ja ottaa käyttöön menetelmiä, jotka voisivat olla käyttökelpoisia myös lähitulevaisuudessa. Tällöin niitä ei tarvitsisi muuttaa tekniikan kehittyessä. Tällainen työkalu on, esimerkiksi XML standardi, jota ehdotan käytettäväksi tietokannan rakentamiseen ja digitoidun multimedia-aineiston käsittelemiseen.

Työ on jaettu kolmeen osaan. Ensimmäisessä osassa tarkastellaan teknisiä seikkoja, jotka kuuluvat digitaalisen tekstikokoelman luomiseen. Toisessa osassa esittellään aineistoa, joka kuuluu projektiin. Kolmannessa osassa pohditaan digitaalisen kuvankäsittelyn menetelmien käyttökelpoisuutta tässä projektissa. Johtopäätöksessäni olen tullut siihen tulokseen, että tällainen projekti voidaan toteuttaa onnistuneesti, ilman ylitsepääsemättömiä ongelmia.

List of Abbreviations

AGI	El Archivo General de Indias, Seville Spain
ASCII	American Standard Code for Information Interchange
BMP	Bitmap
CCD	Charge-Coupled Device
CD	Compact Disk
DPI	Dots Per Inch
DVD	Digital Versatile Disc
CIMI	Consortium for the Computer Interchange of Museum Information
DTD	Document Type Definition
ELM	Enables Lucid Models
FFT	Fast Fourier Transforms
GB	GigaByte
GIF	Graphics Interchange Format
HTML	Hypertext Markup Language
ISO	International Standardization Organization
JPEG	Joint Picture Expert Group
LTO	Linear Tape-Open
LZW	Lempel, Ziv, Welch
MB	MegaByte
OCR	Optical Character Recognition
PGM	Portable GreyMap
PNG	Portable Network Graphics
PPI	Pixels Per Inch
PPM	Portable PixMap
RGB	Red Green Blue
SGML	Standard Generalized Markup Language
TIFF	Rag Image File Format
W3C	World Wide Web Consortium
XML	Extensible Markup Language
XSL	Extensible Style Language

1. Introduction

Until the advent of digital technology, the common way to preserve text has been simply to copy it. In antiquity, a scribe would patiently copy a manuscript, and may, at the same time, ‘edit’ it. When the printing machine was invented the printer would use an impression of a raised surface in wood-engravings, sunken lines in copperplate and steel engravings, or flat surface in lithographs, to print the text and images on paper. After the text has been printed the impression has usually been discarded. The invention of photography has enhanced the printing process but has not changed its basic mode; one still has to prepare an impression for the printing process. With photography one can prepare microfilms, however, except for few applications like old newspapers and genealogical records used for family tree construction, the technology is not generally used. In theory, one can always preserve old books by photocopying them on paper.

Digital technology offers a new approach. With the use of a text editor, the writer can type the text using a keyboard, modify it and send it directly for printing. Unlike traditional printing, the digital 'impression' uses very limited space and can easily be preserved. Any time in the future, one can reuse it, providing the appropriate software and hardware are also preserved.

With the universal spread of personal computers, hardly anything of importance is written by hand or with a typing machine. That means that nowadays everything written is eventually being stored on a digital medium and, as such, can easily be retrieved and displayed. For my purpose, I define text heritage as everything written in a non-digital mode starting from antiquity. As such, the amount of existing old text is gigantic however finite, that is, basically it does not grow.

The main theme of this work is how to transfer old text into digital mode for efficient storing and distribution and possibly better reading. The ultimate aim is to incorporate the digitized material into some kind of multimedia format in such a manner that data mining can be easily executed. As a case study I will use the collection of books concerning the study of the Gothic language, starting from 1569 till 1805, which are preserved at the

library of Uppsala University. The aim of this work is to prepare a pilot project for digitizing this collection. Any experience gain from this project can be exploited for enlarging of the scope of material to be scanned.

The work is generally divided into three parts. In the first part (chapters 1-4) there is a survey of various projects conducted around the world, which endeavor to digitize old documents and manuscripts and have them displaced on-line, locally or through the Internet. Next, a case is built for a multimedia format. I argue for the creation of structural documents employing the XML standard for efficient approach to the material and data mining.

This discussion is followed with the details of the digitizing process. In this section I go through the steps to be pursued in order to scan a vast collection of written and printed material. The decisions to be taken are time dependent, that is, what is true and reasonable at this moment, may turn out be dated in a year or two. In this era of fast technological development, it is extremely important to maintain a flexible approach.

The second part (chapter 5) deals with the raw material. I suggest that, as a pilot project, the material concerning the Gothic language at Uppsala University library be digitized. The library preserves, as part of its manuscript collection, the so-called Codex Argenteus, written apparently in the 6th century. For the last four hundred years the institute has collected everything written about this manuscript, as well as other material concerning the study of this old Germanic language.

The third part (chapters 6-8) is concerned with various digital image processing methods and algorithms. In this part I examine whether various filters can improve the results of the scanned material. Several methods, possibly useful in order to decipher damaged text, are also proposed.

Digitizing text has many advantages, but also shortcomings and limitations. Any way one looks at it, resetting an old book and printing it is a way superior to displaying it on a computer screen. However, this alternative is very expensive and as such impractical.

Therefore, digitizing the text is some kind of a compromise between high quality and faithful reproducing and doing nothing. In fact, digitizing enables a better access to the original, prevents damage done to the old documents by repeated exposure and handling, and provides a backup copy.

I conclude the work with some initial observations and suggestions as for how to proceed next.

2. Digitizing Cultural Heritage

Three major problems impede the use of digital technology from being a major tool in preserving cultural heritage. The first one is the enormous amount of financial resources needed for the scanning process, the second one is the quick obsolescence of hardware and software, and the third is the need for standardization of tools and interfaces. As a result, the work done in this area is mostly on a pilot basis, where researchers study the various aspects of certain subjects, build and experiment with modest projects, debate on standards and establish coordinating bodies.

The Consortium for the Computer Interchange of Museum Information (CIMI) is leading an international effort to provide distributed search and retrieval of cultural heritage information. The basic idea is to utilize ANSI/NISO Z39.50-1995 protocol for information retrieval for creating a uniform access to existing and emerging digital collections and the vast repositories of cultural heritage information resources. The aim of the standard is to model search and retrieval of a variety of data and content types, to create a uniform semantics across disparate and heterogeneous systems and databases and to compose unified specifications for the development of appropriate software (Moen 1998: 45).

CIMI membership comprises of museums and research centers. The Consortium has entered into a sponsoring relationship with two software developers to build Z39.50 tools for use in the test-bed. The goal is to demonstrate the capability of the standard to support search and retrieval between multiple server and client implementation of specific types of cultural heritage information resources. If I understand correctly, several test-beds were successfully executed.

In this chapter I will survey several projects aimed at preserving cultural heritage, which are currently conducted.

Memory of the World Program

This is a project run by UNESCO for preserving documentary heritage. The basic premise is that documentary heritage reflects the diversity of languages, peoples and cultures and it

is the mirror of the world and its memory. However, this memory is fragile and every day irreplaceable parts of it disappear forever.

One recent project financed by the program is the Slave Trade Archives Regional Training session, which took place in Dakar, Senegal in February 2002. Among its topics there was an archive project which aimed at enabling participating countries to preserve original documentation relating to the trans-Atlantic slave trade and to improve public access to this material. (http://portal.unesco.org/)

Cultivate

Cultivate is part of Directorate General Information Society of the European Union and is located in Luxembourg. Bernard Smith (2000) defines the aim of the institute as providing "access to scientific and cultural content through the networking of libraries, archives and museums." The ultimate end is directed "towards basic socio-economic needs - job creation, economic growth, quality of life, and in support of the policy and goals of the European Union."

Some of the more specific goals are defining and establishing a framework for heritage information, agreeing on infrastructures, standards and methodologies, supplementing existing national, European and international initiatives, researching and developing the use of state-of-art technologies, and deepening the international co-operation. Among the 'Key Actions' supported by the project there is the topic developing new access modes to cultural and scientific content. This approach gives the opportunity not just to do research but to experiment with new technologies for virtual representations of cultural and scientific objects.

Matenadaran

The Matenadaran is one of the oldest and richest book-depositories in the world. Its collection of about 17000 manuscripts includes almost all the area of ancient and medieval Armenian culture and science. The center has in its collection many works which were lost in their mother languages and are known only from their Armenian translations.

Since 1998 the institute has been included in the register of UNESCO's Memory of the World program (see above). Currently the institute is in the process of updating the Virtual Matenadaran, funded by UNESCO. It will include database of the Armenian manuscript collection with a searching system, concise data, and calligraphic and miniature samples, in Armenian and English (http://www.matenadaran.am/).

Digital Image Archive of Medieval Music (DIAMM)

The purpose of the Digital Image Archive of Medieval Music (DIAMM) is to obtain, archive and, where necessary, enhance digital images of European sources of medieval music (Wathey et al. 2001) The first phase of the project involved the collection of digital images and computer enhancement of 15th-century British fragments. In its second phase the project expanded to embrace all fragmentary and some of the less accessible complete sources of British pre-Reformation music. The project is a collaboration between the University of Oxford and Royal Holloway, University of London.

The DIAMM website portal provides archives with the opportunity to make their collections available to scholars worldwide as a single, accessible, password-protected on-line collection. Anyone who wishes to use the archive must sign a page-long access agreement, designed to safeguard the owners' control over copyright.

From the digital image processing point of view, the project's researchers coined a term 'virtual restoration' intended for developing restoration techniques of images in a virtual environment. One of these techniques involves separating and manipulating separately color channels, another is allowing any process applied to the original image to be added as a layer. In one example, MS 144 (Wathey et al. 2001: 237), the authors used those techniques to enhance leaves of a 14th-century music book that were scraped and re-used for writing another text in the 15th century.

El Archivo General de Indias, Sevilla, Spaain (AGI)

This project seems to be an impressive one, well organized and documented and, as such, something to be followed and to learn from. For this reason, I will dwell more on this cultural venture and refer to the experience gathered by its staff later in this work.

The archive houses 43,000 bundles with 86 million pages, the most complete documentation of the Spanish administration in the Americas, from Christopher Columbus until the end of the 19th century. The objective was to offer digital surrogates to reduce the handling of originals, some of them being quite fragile. The project started in 1986 and in 1998 the team produced 11 million digital-image pages. One result is that about one-third of the AGI's on-site consultations were done electronically. In this manner, the digitizing work has greatly reduced exposure of the original documents and, moreover, has reduced the time researchers have had to spend in the archive premise.

One major observation expressed by González (1998), and should be considered by anybody undertaking such a project, is that decisions are bound by time and money, forcing choices that are not always optimal but realistic. One major example is the rapid obsolescence of hardware and software; as work proceeds, swift technology changes render dated and obsolete earlier accomplishments. However, the system created is alive and well, and extensively used by researchers.

3. Database and Data Mining: Introducing XML

While written manuscripts or printed books have attributes of eternity, or at least of many hundreds of years, anything digitized seems to be valid for only a short while. Indeed, the idea of converting manuscripts and books into abstract electrical form is not a self-evident undertaking.

One possible way of overcoming the problem of temporality is creating durable common standards and having everybody follow them. However, this idea is more easily said and agreed upon than actually implemented. Some standards are very good and everybody follows them, some are bad and, as such, universally ignored. Moreover, some standards were very good for a certain period of time, but when technological developments outdated them, they have become obsolete and at times obstacles to advancement.

This chapter deals with the question of creating a system which enables easy access to digitized text. Such a system should include a mechanism for searching data and retrieving it. In addition, the system should be as simple as possible. Since various operating systems and other software programs come and go, the discussion here is restricted to a standard which I consider to be appropriate for this project and strongly believe that it will be used for many years to come – XML (Extensible Markup Language).

XML is a subset of SGML (Standard Generalized Markup Language). Around the late sixties, IBM asked Charles Goldfarb to build a system for storing, finding, managing, and publishing legal documents (Goldfarb & Prescod 1998: 6). Goldfarb found that there existed many systems within IBM that could not communicate with each other as each of them had different representation (file format). So Goldfarb, together with Ed Mosher and Ray Lorie, set up to create a common language. They concluded that the new common denominator should be some form of a markup language.

In the ‘old days’ of typesetting machines and offset printing, a manuscript going to print was "marked up" by hand with written instructions for the printer. These instructions,

written around the text, told the setter how the text should be formatted: “enlarge font, bold it, put in italics, new paragraph,” etc. (see Figure 3.1)

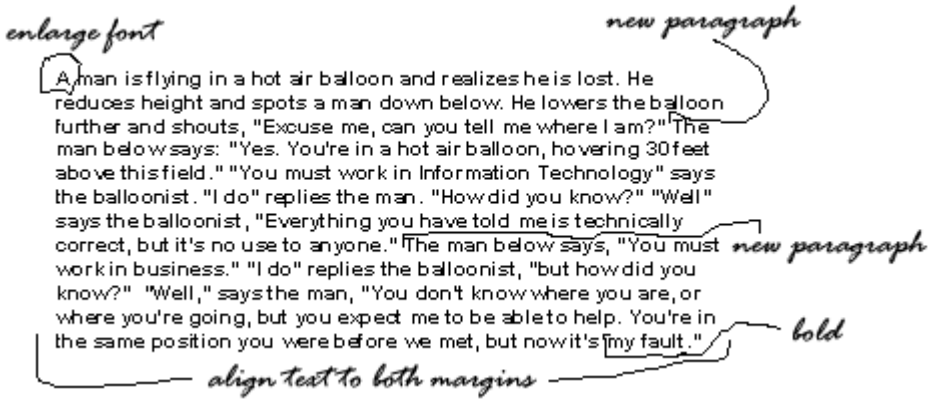


Figure 3.1. Traditionally marked-up text

Formatting a text in a markup language is very much the same; the text is circled with instructions called tags or codes. In 1969, the IBM team developed a language that was not specific to a particular system. They called it Generalized Markup Language which, not coincidentally, has the same initials as the names Goldfarb, Moser and Lorie (Goldfarb & Prescod 1998: 13). Between 1978 and 1986 Goldfarb led a team of users, programmers and academics that transformed the language into an International Standard (ISO 8879), approved in 1986. It received the name SGML – Standard Generalized Markup Language. It has become the de facto standard for the interchange of large and complex documents.

In 1989, a researcher named Tim Berners-Lee proposed that the Cern European Nuclear Research Facility would use hyperlinked text documents for sharing internal information. A colleague named Anders Berglund suggested that they adopt an SGML-ish syntax. Starting from a simple document type in the SGML standard, they developed a hypertext version called HTML – Hypertext Markup Language. Tim called the hypertext system the World Wide Web (Goldfarb & Prescod 1998: 15) and the rest is history.

The biggest strength of HTML is its simplicity. With very little training, one can build basic homepages which can be displayed by all browsers. HTML documents use SGML’s simple angle bracket convention for markup, which means that one can create HTML

documents on any word processor. HTML biggest deficiency is its lack of SGML's capacity of extensibility. In practice, extensibility means that one can define a module or a template which is repeatedly filled by content. For this purpose, a document type specified for a certain need is formally defined and the documents created must be verified for validity against it. If the person who fills the details of a certain database does not follow the specifications set up in the master model, then the browser will not approve it and the syntactically incorrect details must be revised.

In fact, XML has been developed in order to correct the lack of extensibility in HTML. At that time, it was decided that SGML was too complicated for Internet use and a new simpler markup language, XML, should be devised.[1]

3.1. Working with XML

One of the main advantages of the XML standard is that the major players in the on-line arena have agreed to let the World Wide Web Consortium (W3C) to coordinate the efforts for developing and implementing it. Ideally, anybody who composes software following the XML standard, has the responsibility to verify that it is compatible with other software programs that, too, follow the standard. Theoretically, this is indeed a noble idea; however, in practice one should always be on his or her toes. Since big money is involved, one can assume, based on the experience gain by watching the case of Microsoft vs. Netscape, that someone may try to steer the market to its own direction. The main area of possible loopholes is in those applications that are not yet defined, or are not subject for standardization and, as such, are opened to free competition. One way to avoid troubles is to create an XML structure that is as simple as possible, sticking to the rules that are inherently so general that one can quite safely assume that they will be preserved and that nobody will dare to manipulate them.

A major feature of XML is the separation of content and format. In fact, there are two separate languages to accomplish it: XML for Web content and XSL (eXtensible Style Language) for the formatting style. Both languages are large and complicated. The separation of content and style streamlines the creation of structural documents with clear

division of various information categories. This feature enables the use of efficient data mining algorithms.

The first step in any software design is to compose a comprehensive specifications document, which, in our project, has not been prepared yet. For this reason, my intention here is not to create a definite and comprehensive structure but rather present some ideas as to how the database should be constructed.

The first step is content analysis. For the sake of discussion, let us assume that the aim of the project is to have three types of material scanned: old books, handwritten manuscripts and old maps. In the database, those three types should be separately constructed. In addition, the database should include an introduction with general information concerning the project and its aims, description of the structure of the database, and details concerning the participants and their contributions. One popular method to represent the structure of a DTD is called an ELM (Enables Lucid Models) tree diagram. Figure 3.2 displays such a diagram with the main components.

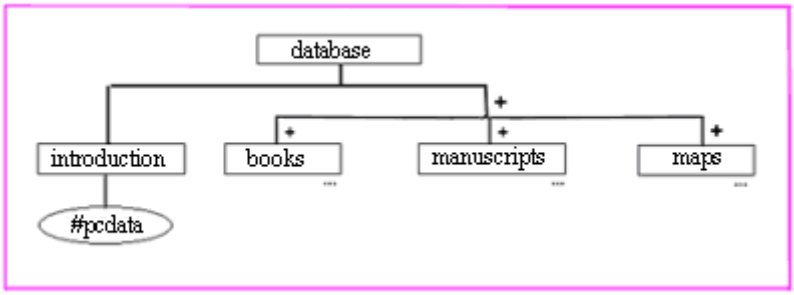


Figure 3.2. The main components

In XML code the structure might look like:

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE DATABASE [
      <!ELEMENT database (introduction, books+, manuscripts+,
maps+)]>
```

This is the beginning of a DTD (Document Type Definition), which is a file that rigorously defines what it means for each type of document to be valid. The basic idea is to guarantee that improper data cannot appear in the database. Another major advantage is that the fields of information are clearly defined, and a searching engine can mine data in specified locations.

Elements are the foundation of XML markup. Element type declaration must start with the string "<!ELEMENT", followed by the name of the element type being declared and content specification. The sign '+' indicates that these elements may be repeated without limit, however there must be at least one element. As a preliminary suggestion, here is a list of elements to be included in the database:

- **Books**: author, title, year of publication, language, number of pages, key words, comments.
- **Manuscripts**: title, subject, language, date, number of pages, key words, comments.
- **Maps**: title, year of publication, language, key words, number of pages, comments.

Each element can be parsed in the following manner:

<!ELEMENT books (author, title, year_publication, language, pages, key_words, comments)>

and each of those elements could still be subdivided, e.g.:

<!ELEMENT author (first_name, last_name)>

The next step is defining the element content:

<!ELEMENT first_name (#pcdata)>

<!ELEMENT last_name (#pcdata)>

According to XML official manual (2000) the keyword **#PCDATA** derives historically from the term "parsed character data." A **character** is defined as "an atomic unit of text as specified by ISO/IEC 10646 [ISO/IEC 10646]."

While XML describes the content in a structured way, another language, XSL, describes the formatting. With XSL one creates a file called a stylesheet, where the human designer

can express creativity and instruct the computer as for how to display the document, either online or in high-quality print. However, as Tanenbaum (2003: 641) comments, “it is not yet clear whether XSL will take over from traditional style sheets.” He also maintains that both XML and XSL can be misused, “You can count on it” (p. 642). Another alternative, which is quite new, is to use XHTML – an application of XML - which enables compiling and creating an independent structure. In any case, one way to avoid the risk of being a victim of a browser war is to stick to general and simple conventions.

4. Digitizing Manuscripts and Old Books

In theory there are two methods for digitizing text; one involves the so-called optical character recognition (OCR) and the other converts pages into digital images. In practice there is only one alternative - the latter one.

According to Mori & al. (1992, 1029) in the early days of pattern recognition research, almost everyone took the subject of OCR, as characters were very handy to deal with and were regarded as a problem which could be solved easily. However, after some initial easy progress, great difficulties surfaced and people diversified their interests over a wide range of topics in the pattern recognition field. In Japan, a series of research and development for OCR had been done nationwide as part of a national project and, as a results, many published papers deal with the Japanese language.

Recognition rate depends heavily on the quality of the print, the preparation of the document, the resolution of the scanner, the font used, the complexity of the layout, and format. So when dealing with manuscripts and old books, using OCR is, in my opinion, impractical. Just for example, figure 5.3 (from 1665, see page 25) includes text in Latin, Hebrew and Greek letters and, in addition, the state of the pages is quite poor. Even with newer printing, where such rates of recognition as 99.9%, when operate under optimal conditions are claimed, they can be very misleading (Mori & al 1992, 1047). The reason why such rates are stated is very simple; if the rate is lower, then it is rather faster for a good and experienced typist to type the text than to scan it and correct the mistakes.

In the beginning of the 1990s, two employees of the Technical Research Center of Finland (VTT) embarked on a research aiming at determining whether OCR is accurate enough for generating automatically full-text indexes for newspaper collections, either from the original newspaper pages, or 35 mm microfilm frames. In a publication from 1994, the researchers Riitta Alkula & Kari Pieskä, describe the research and its results and conclusions.

To get relatively good results, the characters accuracy must be at least 98 percent in order to produce good recognition of words; after all, one spelling mistake cause a misreading of the whole word. As it turned out, the accuracy of recognized words was only 44 percent with scanned microfilmed images. Therefore, the authors concluded that "Microfilm scanning is a promising technology, but currently not adequately established for this project."

Nowadays, as word processors are prevalent, hardly anything of importance is either written by hand or being punched on an old-fashioned, or even electric, typing machines. As a result, my personal observation is that text OCR is practically restricted to mail sorting systems and ballot counting, and possibly several other limited tasks.

The scanning task is a matter of optimization as the process involves a few bottlenecks like screen resolution, printer capabilities, memory storage limitations, etc. As technology changes fast, decisions taken today may turn out to be dated in later stages of the project. Moreover, one should also keep an open mind for other alternatives already available, which may turn out to be useful during the actual work.

4.1. The Scanning Process

The first step in such a project is purchasing and assembling the appropriate instruments. Since we are dealing with old and brittle material, using flatbed scanner is out of question as it damages the books. As an alternative, one can mount a high quality camera on a specially designed apparatus and also construct a special device to hold the book open. At the beginning of the AGI project, they used Rank Xerox 7650 flatbed scanners. As time went by, digitization stations with flatbed scanners have been replaced with Kodak DCS 420 digital cameras. As González (1998) indicates, the new hardware allows easier and more rapid digitization, which boosts productivity and reduces digitization costs. Moreover, the use of a camera instead of a flatbed scanner is safer for the original as there is much less risk of paper deterioration.

Another alternative is to use a camera scanner that can achieve technically almost printable quality. The main difference between camera scanner and a regular digital

camera is that in the former the scanning is done line after line. Basically, there are two principles in arranging the charge-coupled devices (CCDs) with digital cameras: Either on an area or in a line. If the CCDs are arranged in an area, the camera can capture the image in one shot; if the elements are arranged in a line, image capturing happens when this line is moved across the area seen by the lens. This is pretty similar to what happens in a flatbed scanner. The results are apparently better, however, with the advent of digital cameras, it is plausible that the advantage of using the more expensive camera scanner will diminish.

Another component of the hardware is memory storage. After conducting several tests to ensure quality and minimizing storage requirements, the AGI's project management decided to digitize at 100 dpi (dots per inch) and 16 grayscale levels (4 bits per pixels). Obviously this is not a very high quality but according to the report, it well suited the aims of the Seville project – "to digitize for consultation, rather than for replacement of the original." They, no doubt, would have settled for a higher dpi had they not been constrained by the eternal problem of memory storage limitation.

In 1998, while digitizing the facsimile edition, that is an exact copy or likeness, of the Codex Argenteus (printed in 1927) I used a millions of colors mode (24 bits) and 100 dpi. The main reason for this choice was storage considerations; as I used recordable compact discs (CD), I figured that a digitized edition of more than 4 disks would cause too much work to copy and distribute. Nowadays, when the standard recording tool is DVD,[2] I probably would have used 300 dpi.

4.2. Scanning Resolution

Since we are dealing with real-world images, our concern is with bitmap data, which is formed from a set of numerical values specifying the colors of individual pixels or picture elements, as opposed to vector data, which refers to means of representing lines, polygons or curves. A bitmap file contains certain information concerning the file and an exact pixel-by-pixel map of an image. The measure of detail within an image is its resolution and the physical size of a file is determined by its resolution. In practice it means the number of pixel wide multiply by the number of scan line long.

Another factor to be taken into account is the number of dots per inch (dpi), already mentioned above. In computers, dpi is a measure of the sharpness (that is, the density of illuminated points) on a display screen. The amount of dots (pixels) that can be displayed on a monitor is normally limited to only about 72-100 dpi. The amount of dpi for a given picture resolution will differ based on the overall screen size, since the same number of pixels are being spread out over a different space.

Some users prefer the term "pixels per inch (ppi)" as a measure of display image sharpness, reserving dpi for use with the print medium. In printing, dpi is the usual measure of printed image quality on the paper. It actually refers to the dots of ink or toner used by an imagesetter, laser printer, or other printing device used for printing text and graphics. In general, the more dots, the better and sharper the image. As for now, the average personal computer printer provides 300 dpi or 600 dpi. Choosing the higher print quality usually reduces the speed of printing each page. Using dpi rate for either screen display or printing higher than the specifications allows is redundant and adds no extra value to the results.

The size of an image file is determined also by the number of bits used to display one pixel. The number of bits used to define a pixel's color shade is its bit-depth. For example, True color is the specification of the color of a pixel using a 24-bit value (3 bytes), which allows the possibility of up to 16,777,216 possible shades of color. Each byte displays one of the three primary colors: red, green, and blue (RGB), and each of these bytes has a range from 0 to 255 levels. The combination of those colors in various levels enables the creation of any color in the visible spectrum. New color display systems offer a 32-bit color mode. The extra byte, called the alpha channel, is used for control and special effects, e.g. opacity. To make a long story short, better quality entails (much) bigger files.

4.3. File Formats

While collecting material for their book on file formats, Murray and vanRyper (1996: xxvi) found out in that in some cases they simply could not find out who owned the specifications and in two cases they located the caretaker, but the caretaker could not find

the specifications. Still, in other cases they could not get through to the correct person in the organization despite a willingness to provide them with the information. There were also those institutions that wished to restrict the availability of older formats, apparently so they would not be bothered by users calling them up about those obsolete formats. All in all, there are few dozens of file formats, however, in practice only a handful of them are of interest for this project.

The aim of the scanning project is to digitize images of old books and manuscripts for multimedia on-line display and for printing. Each of those targets demands different specifications and solutions: the former entails quick access and the latter a high quality printout. In practice, a multimedia system requires small and compact files with the essential data, for quick and efficient transmission and displaying; in printing environment, time is insignificant, but printing quality is essential. Following the procedure established by the Waller project (see below) I suggest that each image will be scanned twice, in two different modes for two different purposes.

For some years now, Uppsala University Library has been digitizing, with great care and expense, and displaying openly on the Internet a well-known collection of documents assembled by the Swedish surgeon Erik Waller (1875-1955). (http://www.ub.uu.se/arv/waller/eindex.cfm). The material deals with the history of medicine and the history of sciences. According to the latest estimate, there could be around 40 000 items in the collection. Digitizing it is a hard and painstaking task. Digitized images are presented on the Internet in .jpg mode, and, in addition, files in .tif format (see below) are stored separately on CD ROMs.

As argued above, when choosing the file format, eyes should be focused on eternity, which, translated into information technology, means at least a few years and more. The formats I am fond of and use extensively in my filtering software are PPM (Portable PixMap) which supports full-color images, and PGM (Portable GreyMap) for greyscale images. They were primarily written by Jef Poskanzer, and he is the one who owns the copyrights for them. However, he lets both the source and the executable form be distributed freely via the Internet and other channels. Both formats are designed to be as

simple as possible. The header and the data portion are written in ASCII form, which means that one can open the file with any text editor and examine the content. The bad news is that the files of images saved in those formats are almost three times bigger than other heavy formats, and as such those format are impractical for scanning and preserving purposes.

By word of mouth, it seems that TIFF (Tag Image File Format) is considered to be the best file format for the purpose of printing. A TIFF file can be identified as a file with a ".tiff" or ".tif" file name suffix. It was developed in 1986 by an industry committee and is found in most paint, imaging and desktop publishing programs. It has garnered a reputation for power and flexibility, but it is also considered to be complicated and mysterious (Murray & vanRyper 1996: 881). Another advantage of TIFF is that it supports certain types of image compression (see below). The main problem with this format is that some of its substructures are not well defined. An alternative to TIFF is the Microsoft Windows Bitmaps (BMP), which may not be as versatile as TIFF but for all practical purposes, is, as far as I can see, not very much different.

For multimedia purposes, there are three file formats which are being read by Internet browsers: GIF, JPEG and PNG. GIF (Graphics Interchange Format) is a fine and widely used format. The vast majority of GIF files contains 16-colors or 256-color near photographic quality image. However, there is one big problem with this format: The LZW (Lempel, Ziv, Welch) compression algorithm used in the GIF format is owned by Unisys, and companies that make products that exploit the algorithm need to license its use from Unisys. Up to now, Unisys has not required the end users of GIF images to obtain a license, although their licensing statement indicates that it is a requirement. Unisys says that getting a license from them does not necessarily involve a fee. In fact, many GIF downloaders and Web site builders continue to be ignorant of or indifferent to the requirement to get a license from Unisys for the use of their algorithm. Nevertheless, it would not be wise to use the GIF format for any sizable and durable project.

One alternative to GIF is PNG (Portable Network Graphics). It actually was developed by an Internet committee as a patent-free replacement for GIF, after the graphic industry was

shocked to hear the fee requirement. PNG was designed with the goals that it be a simple format, one that is easy to implement and completely portable. PNG is a relatively new format and my observation is that not very many web designers use it, maybe because one cannot compress it as much as JPEG (see below).

A more popular format is JPEG (Joint Photographic Experts Group). It was developed by an International Organization for Standardization (ISO) group of experts for graphic images on the Web. It appears with the file suffix .jpg (ISO standard 10918). A JPEG file is created by choosing from a range of compression qualities. When one creates a JPEG file or converts an image from another format to JPEG, he or she is asked to specify the quality of image he or she wants. Since the highest quality results in the largest file, one can make a trade-off between image quality and file size. However, the standard does not specify the border between the various levels and, as a result, each program has it own specifications. The JPEG algorithm can compress a file to 1/25 of its original size, but some information is lost.

4.4. Compression Considerations

Compression is the reduction in size of data in order to save memory storage space or shorten transmission time. Generally, there are two major kinds of compression: lossless where all the information is preserved, and lossy where part of the information is deleted.

With lossless compression, the size of the file can be reduced with the help of various algorithms by up to 40-50%. Those algorithms remove all extra space characters, insert a single repeat character to indicate a string of repeated characters, and substitute smaller bit strings for frequently occurring characters. One very clever algorithm is LZW. It takes each input sequence of bits of a given length and creates an entry in a table for that particular bit pattern, consisting of the pattern itself and a shorter code. As input is read, any pattern that has been recorded before is substituted with a shorter code, effectively compressing considerably the total amount of input. As mentioned above, the LZW can be used to compress TIFF format and it is part of the GIF format, however it has also acquired added notoriety.

Microsoft's BMP format does not have a special compression algorithm but one can use such a program as WinZip to reduce the size of the file. I maintain that in our project using WinZip is more of a headache than a real saving. This compression procedure is more practical when a file is mailed electronically.

As a matter of fact, there may not be a special need to compress the TIFF images since they are included for printing purposes. As our primary memory storage devices are DVD disks and LTO (Linear Tape-Open) tapes (see below), even a substantial reduction of files' size is not that crucial.

Compression is of significant importance in the domain of on-line communication and the method used for that purpose is, to a great extent, the lossy one. The procedure for the lossy compression is first to transform the image into the frequency domains. Next, frequencies above certain threshold are deleted and the remains are coded, compressed and sent to the destination. There, the signal is decompressed and decoded, and the lossy image is restored and put on display. The result of this process is that the size of the final file is only a fraction of the original one. The saving in transmission time is significant but the price is in the quality of the image. For reading the image on the monitor's screen, the difference might be negligible, but when the file is printed the quality may turn out to be relatively poor.[3]

4.5. A Comparative Study

Following the list of books to be scanned (see below 5.1), four various sizes of the original images were scanned, with varying degrees of resolution and three file formats. The scanning was conducted on a flatbed HP ScanJet 5300C color scanner. Table 4.1 presents the results of this study.

The first observation is that .jpg is much more efficient than .png. Scanning an A4 image in .png with a resolution of 300 dpi yields a file of the size 5.39 MB and a .jpg with the maximum quality option, a file of 2.23 MB. This difference is not crucial as far as memory storage is concerned but is significant for the on-line transmission. Since the .jpg file is less than half of the .png one, it would be more appropriate to use the JPEG file format.

Table 4.1. A comparative study of resolutions and file formats

dpi	600			300			150		
file format / page size(cm)	.tif	.png	.jpg	.tif	.png	.jpg	.tif	.png	.jpg
15x9	7.543	4.052	1.217	1.869	1.174	0.452	0.487	0.334	0.092
15x22	18.591	10.309	3.147	4.527	2.803	1.247	1.183	0.768	0.239
19x23	24.878	14.401	3.979	6.135	3.816	1.679	1.569	1.039	0.321
21x30	34.209	19.952	11.892	8.532	5.394	2.342	2.183	1.444	0.449

File size is in MB

The second observation is that the resolution of 600 dpi produces too big files for present technology. Advanced contemporary scanners can usually handle an image of that resolution, however, printing a file of the size of 34.2 MB (.tif, A4) may take quite a long time. Moreover, transmitting a file of the size 11.9 MB (.jpg) on-line may, at times, be problematic.

The other resolutions, 150 and 300 dpi may both serve well. I have used quite extensively images of 150 dpi and have not encountered any quality problems. However, with the advent of image processing technology, the use of 300 dpi resolution may be more suitable. In the calculations below, both magnitudes will be examined.

5. Digitizing Text Heritage at Uppsala University Library

On the 4th of September 1827 the library of the Royal Academy in Turku with all its 40,000 books, acquired during almost two centuries, was destroyed in a great fire that devastated the whole town, including the university and the cathedral. Only some 800 books, that were lent out were saved and today form the Aboaica collection. Apparently exemplars of all the books destroyed at that fire, and many others, can be found at the library of Uppsala University, Sweden's oldest research library with 5 million books, old and new. If one does not live in Uppsala and still wants to study those old books, since the library does not let old books and manuscripts leave its premises, he or she must travel to this city and stay there as long as the research lasts, possibly many weeks or more. In fact, in the library itself one has to study the material in a specially designated room closely guarded. Digitizing the vast collection of old books, manuscripts and maps will enable an alternative to coming to Uppsala, and, in addition, will reduce the use of the original material. One more benefit is a creation of back-up copies.

The process of converting society's heritage of old books and manuscripts in general and at Uppsala University Library in particular into a multimedia form may take many dozens of years; the goal of this project is to jump-start the process at that library. Indeed, there must be a first step even in a long march, and I propose to start with the Gothic heritage. Uppsala University Library preserves a world-famous manuscript - the 6th century Codex Argenteus, the "Silver book" - written in silver and gold letters on purple vellum. It contains fragments of the Four Gospels in the fourth-century Gothic version made by Bishop Wulfila. For a few hundred years now, Uppsala has been a center of Gothic studies.[4]

The existence of an old manuscript with Gothic text in it is first mentioned by the Dutchman Johannes Goropius Becanus in his *Origines Antwerpianae* (1569). Figure 5.1 displays the lines, written in Latin, where the author mentioned the existence of a Gothic manuscript in the monastery at Werden. *Origines Antwerpianae* is a history book whose book seven (liber septimvs) deals with the history of the Goths.

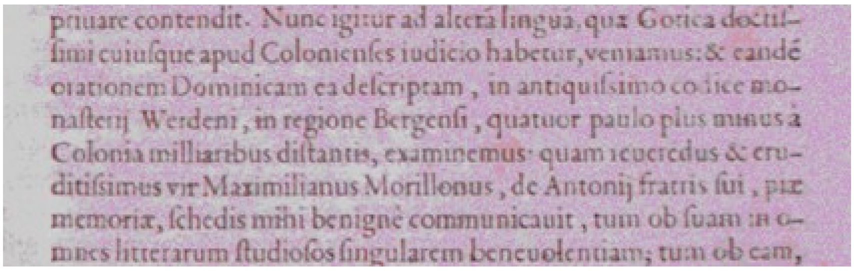
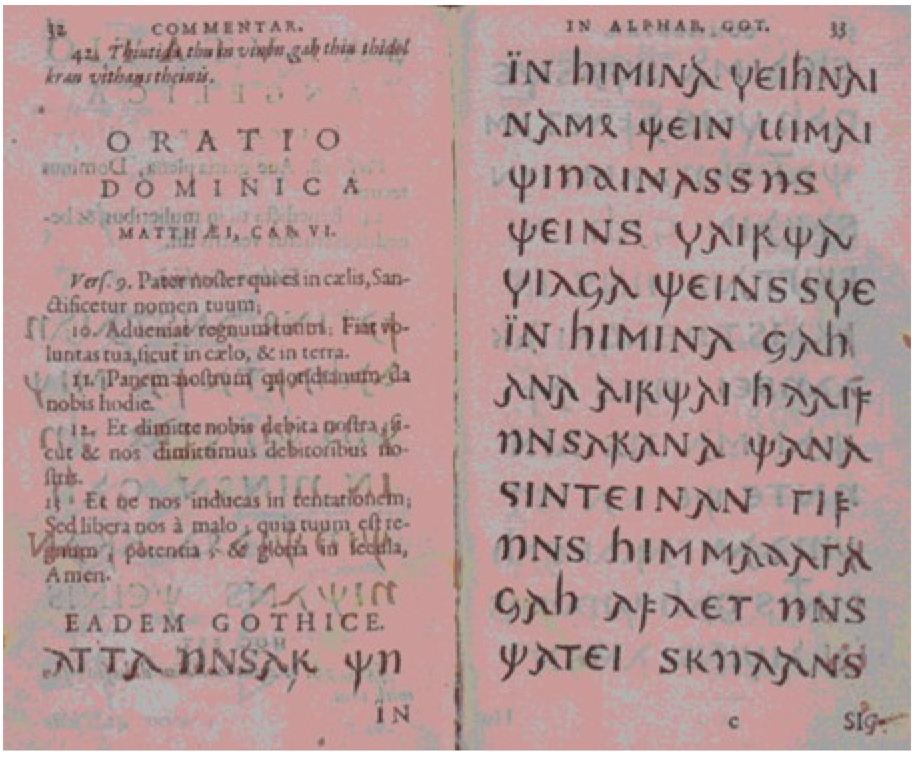

priuare contendit. Nunc igitur ad alterá linguá, quæ Gotica doctissimi cuiusque apud Colonienses iudicio habetur, veniamus: & eandé orationem Dominicam ea descriptam, in antiquissimo codice monasterij Werdeni, in regione Bergensi, quatuor paulo plus minus à Colonia milliaribus distantis, examinemus: quam reuerêdus & eruditissimus vir Maximilianus Morillonus, de Antonij fratris sui, piæ memoriæ, schedis mihi benignè communicauit, tum ob suam in omnes litterarum studiosos singularem beneuolentiam; tum ob eam,

Figure 5.1. Becanus 1569

Here is the text, translated by Lars Munkhammar and Krister Östlund (Uppsala):

> Then, let us turn to the other language, which is considered to be Gothic, according to the scholars in Cologne, and let us in this connection examine the Lord's Prayer that is written in this language. It is in a very old manuscript, belonging to the monastery of Werden in the Bergian region, scarcely four miles from Cologne. The most learned and venerable Maximilian Morillon has sent it to me from the papers left by his brother Antonius of blessed memory ...

The next entry is from 1597. In the second part of the book *De literis & lingua Getrum, siue Gothorum*, the Dutchman Bonaventura Vulcanius mentioned for the first time the term ‘argenteus’ (“silver”). The book deals with the Gothic script and the Gothic language. In Figure 5.2 there is a portion of the Lord’s Prayer as rendered by Vulcanius.


32 COMMENTAR.

ORATIO DOMINICA

MATTHÆI, CAP. VI.

Vers. 9. Pater noster qui es in cælis, Sanctificetur nomen tuum;
10. Adueniat regnum tuum. Fiat voluntas tua, sicut in cælo, & in terra.
11. Panem nostrum quotidianum da nobis hodie.
12. Et dimitte nobis debita nostra, sicut & nos dimittimus debitoribus nostris.
13. Et ne nos inducas in tentationem; Sed libera nos à malo, quia tuum est regnum, potentia & gloria in secula, Amen.

EADEM GOTHICE.

IN ALPHAB. GOT. 33

Figure 5.2. Vulcanius, 1597

In 1602 three pages of Gothic text appear in a monumental book *Insciptiones antiquae totius orbis Romani* composed by Janus Gruter. The author pointed out the difficulty of reading the text because the letters have faded with age:

> As our friend Michael Mercator told us. Arnold, Michael's father, said that there was in the library of the Abbey of Werden a very old codex written in gold and silver on parchment more than a thousand years ago, containing the work of the four evangelists, but, as is to be deplored, torn, ripped apart, and collected together in no order because of the ignorance of the binder ...

(Translated by James Marchand, http://www.florin.ms/aleph2.html#argenteus)

The first major book concerning the Gothic language is Franciscus Junius' monumental work (1665) which included the text of the Codex Argenteus and a dictionary. Glimpsing through the books I have been very much impressed with the examples written in Hebrew (e.g. Figure 5.3), which I found to be highly scholarly. I am indeed looking forwards to study Junius' book thoroughly.

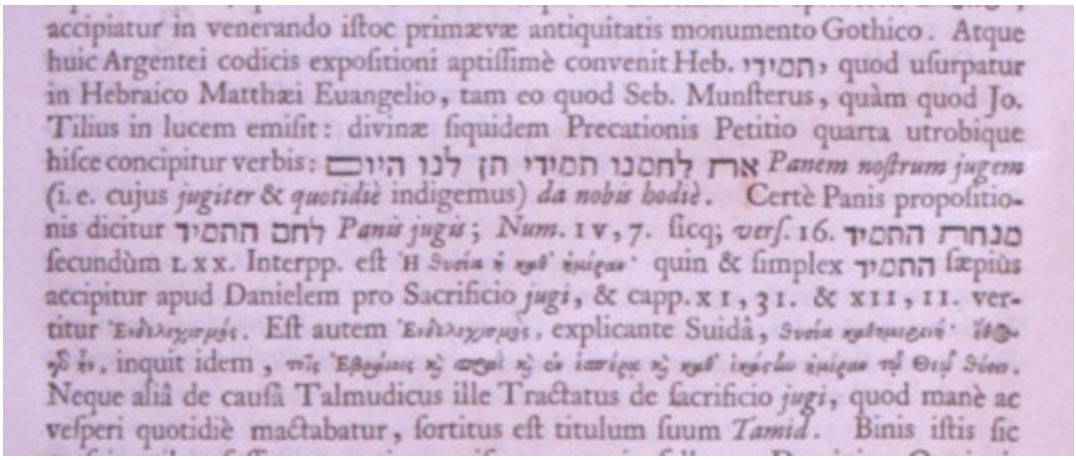
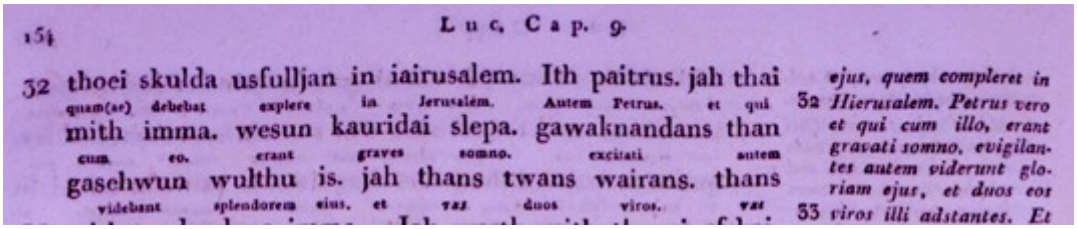

accipiatur in venerando iſtoc primævæ antiquitatis monumento Gothico. Atque huic Argentei codicis expoſitioni aptiſſimè convenit Heb. תמידי, quod uſurpatur in Hebraico Matthæi Euangelio, tam eo quod Seb. Munſterus, quàm quod Jo. Tilius in lucem emiſit: divinæ ſiquidem Precationis Petitio quarta utrobique hiſce concipitur verbis: את לחמנו תמידי תן לנו היום *Panem noſtrum jugem* (i. e. cujus *jugiter* & *quotidiè* indigemus) *da nobis hodiè*. Certè Panis propoſitionis dicitur לחם התמיד *Panis jugis*; *Num.* IV, 7. ſicq; *verſ.* 16. מנחת התמיד ſecundùm LXX. Interpp. eſt Ἡ θυσία ἡ καθ' ἡμέραν· quin & ſimplex התמיד ſæpiùs accipitur apud Danielem pro Sacrificio *jugi*, & capp. XI, 31. & XII, 11. vertitur Ἐνδελεχισμός. Eſt autem Ἐνδελεχισμός, explicante Suidâ, Θυσία καθημερινή· ἤτοι ἡ ἡμέρα, inquit idem, τῆς Ἐβραίων καὶ ἀπὸ τῆς εἰς ἑσπέραν καὶ καθ' ἑκάστην ἡμέραν τῷ Θεῷ θύου. Neque aliâ de cauſâ Talmudicus ille Tractatus de ſacrificio *jugi*, quod manè ac veſperi quotidiè mactabatur, ſortitus eſt titulum ſuum *Tamid*. Binis iſtis ſic

Figure 5.3. Junius, 1665

Some years later, in 1671, a Swedish adaptation of Junius' book was published in Stockholm: *Evangelia ab Ulfila Gothorum*, prepared by Georg Stiernhielm. While Junius presented the Gothic text in Gothic letters, Stiernhielm used Latin ones.

A dictionary, prepared by Eric Benzelius, the Archbishop of Uppsala, was published in 1734. An edition of the text and the dictionary made by Benzelius was published in

Oxford in 1750. A comprehensive Gothic dictionary appeared in London in 1772. Edward Lye, who had prepared this dictionary, died in 1767 and his friend Owen Manning completed and published it. Manning added to the *Dictionarium Saxonico et Gothico – Latinum* two grammars, an Old English one and a Gothic one. Thirty years later, in 1805, Johan Christian Zahn published a new edition of the Gothic texts. In it he incorporated material written by Friedrich Karl Fulda (Figure 5.4).


154 Luc, Cap. 9.

32 thoei skulda usfulljan in iairusalem. Ith paitrus. jah thai
quam(se) debebat explere in Jerusalem. Autem Petrus. et qui
mith imma. wesun kauridai slepa. gawaknandans than
cum eo. erant graves somno. excitati autem
gasehwun wulthu is. jah thans twans wairans. thans
videbant splendorem eius. et vas duos viros. vas

ejus, quem compleret in
32 Hierusalem. Petrus vero
et qui cum illo, erant
gravati somno, evigilan-
tes autem viderunt glo-
riam ejus, et duos eos
33 viros illi adstantes. Et

Figure 5.4. Fulda/Zahn, 1805

Many books concerning the study of the Gothic language were published since the beginning of the 19th century until these very days, however they will be left out of this discussion. The idea is to scan the books published until 1927 in the second part of the project.

5.1. Scanning Calculations

Table 5.1 lists the books to be digitized in the first stage of the project. In general, the pages have acquired a slightly yellowish color and in many cases the pages lost their flatness. In some cases traces of the back page are seen and here and there one notices signs of slight decay. The column of the 'level of condition' indicates the state of the original manuscript on a scale of 1-10.

I propose double scanning: each image be scanned in two modes: .tif for high quality purposes like printing on a laser printer and .jpg, for quick surveying for multimedia presentation, e.g. over the Internet. As both images have the same file name and differ only by the ending, they appear one after the other in the file collection. The database, written in XML, will display the .jpg mode. For printing, the .tif file will be retrieved with an appropriate program.

Table 5.1. The scanning list (150 dpi)

Year	Author	Title	Level of condition	No. of pages	Size of a page (cm)	No. of scans	Memory space (.tiff+ .jpg. MB)
1569	Becanus	Origines	6	133	20x30	133	877
1597	Vulcanius	De Literis	3	110	9x15	55	81
1602	Gruter	Inscriptiones	6	3	23x38	3	19
1665	Junius	Quatuor	5	565	19x23	565	2689
1671	Stiernhielm	Evangelia	4	855	15x22	423	1505
1734	Benzelius	Dictionarium	5	400*	21.5x28	400*	2640
1750	Benzelius	Sacrorum	5	421	22x28	421	2778
1772	Manning/Lye	Dictionarium	4	900*	25x37	900*	5940
1805	Fulda/Zahn	Ulfilas	4	548	24x29	548	3616
Total:						**3884**	**20145**

*These are dictionaries without page numbering. The number given is based on estimation.

Anther consideration that should be taken into account is memory storage space. The calculations in table 5.1 are based on a resolution of 150 dpi (see below calculations for 300 dpi). The basic image size is A4 (21x30cm) as this is the size of pages used by standard printers. If one opening fits into a page of A4 size, then one scanning is executed. In some cases, the size of the page is greater than A4 and in those cases the image size is scaled to fit into an A4 page. The .jpg image quality chosen is the maximum one, however since there is no standard concerning the level of quality and, as a result, each image processing software may produce slightly different images. The scanning sampling for this work was done not with the originals but with surrogate papers on a deskbed scanner.

The table above displays the size of each image in both modes. Multiplying the sum of both files with the number of scans gives a rough idea of how much memory storage is needed. For all the books on the above list, with 150 dpi, one needs a memory storage space of 20145 MB (20.145 GB). The basic storage space I propose is a DVD, which holds 4.7 GB. All in all, one needs 5 such disks for the digital rendering of all those books. For a back-up storage it is possible to mount the digitized material on a tape. One such tape is Ultrium LTO for high-capacity backup, restoration and archival needs that features a native data capacity of up to 200 GB. The list price for one such tape is $80. The digitized material will, then, fill around 10% of such a tape. In fact, with the declining price of hard disks, it might be more prudent to use them as a backup storage facility. The main advantages of hard disks over tapes are easier access.

Better quality of scanning can be achieved by raising the dpi level of the digital image; however the price is slightly more work and much more memory storage space. In table 5.2 the same list as above is calculated with a resolution of 300 dpi. The size of the image files sums up to 85008 MB (85.008 GB) – 19 DVDs, and still fits comfortably into one back-up tape.

Table 5.2. The same list as table 5.1 with a resolution of 300 dpi

Year	**Author**	**Title**	**Size of a page (cm)**	**No. of scans**	**Memory space (.tiff+ .jpg, MB)**
1569	Becanus	Origines	20x30	133	3750
1597	Vulcanius	De Literis	9x15	55	321
1602	Gruter	Inscriptiones	23x38	3	84
1665	Junius	Quatuor	19x23	565	10836
1671	Stiernhielm	Evangelia	15x22	423	6032
1734	Benzelius	Dictionarium	21.5x28	400*	11280
1750	Benzelius	Sacrorum	22x28	421	11872
1772	Manning/Lye	Dictionarium	25x37	900*	25380
1805	Fulda/Zahn	Ulfilas	24x29	548	15453
Total:				**3884**	**85008**

Just for rough estimation of labor time, for starting I suggest that it might take around 15 minutes to accomplish a rendering of one page. This includes scanning, writing file name, adding basic comments, and final checking – after a complete book is scanned. Following this calculation and assuming that a day of labor lasts 7 hours and in a month there are 21 days of work, digitizing 3884 pages may take 6-9 months.

The final checking is indeed a time consuming process. The AGI project conducted this kind of quality control for only a brief period and gave it up. It was left for the users to detect possible errors, such as omissions and repetitions. I do not recommend such an approach, as the users may not notice errors or may notice but would not bother to report them. Moreover, it is always cheaper to correct mistakes in the manufacturing process than to recall products for revision.

5.2. On-line Display

The goal of the project is to broaden access, facilitate retrieval and reduce handling of the originals. As a comparison, the AGI project, which holds about eight kilometers of

shelving, has digitized more than eleven million pages of documents (1998). As a result, about one-third of the on-site consultations were done electronically, greatly reducing exposure of original documents. It was also shown that the average time researchers need to conduct their work in the Archivo nowadays is much less than before the project has started. González (1998) remarks that when they embarked on the project in 1986, the explosive growth of the Internet use was still in the far future, so the electronic version has been limited to the building where the archive is located.

The system employed by the AGI project is based on the idea of a distributed processing and client/server architecture. It consists of three modules: a database with descriptive information that enables a search for data, a digital image storage system, and a user management unit run by the staff of the archive, which includes all the functions relating to the retrieval of documentation. In practice, optical disk servers were installed in the AGI Optical Disk Room, connecting more than a dozen optical disk readers. The optical disks are installed and organized in shelves. An operator at the site handles the requests sent by users by dispatching the requested images through the local area network to the user's station. The time required for sending the images is brief, no more than a few minutes.

The massive emergence of the Internet prompted the AGI administrators into new thinking about the possibilities for long-distance access. However, there have been some issues to deal with. First, there were technical problems: How to send good quality images, how to distribute specific tools developed for image treatment or enhancement, how to communicate between the server and the remote clients, how to prevent crackers from playing havoc with the system, how to protect intellectual property, should the service be free or whether fees are to be charged, etc. According to González, "the decision should be carefully considered, ideally with the participation of all interested parties."

Obviously, the details of the books digitized in our project should be incorporated into the library's general database, known as 'Disa.' As for the images themselves, it will be up to Uppsala University Library to decide what policy to adopt. I would imagine that, at least in the first stage, the library might create a client/server system for studying the material in

the premise itself. Since the digitized books will be mounted on DVD disks, the library will be able to sell them. Each DVD will include an XML file with full details of the content of the disk. In this manner, one will be able to examine the content of the DVD with any Internet browser or a specially developed local browser, using the .jpg files. For printing the material, one could use any photo-editing program that can read .tif files.

6. The Use of Digital Filtering for Enhancing Text Images

A group of IBM's employees have been involved in the digitization of the El Archivo General de Indias (AGI), Sevilla, already mentioned above. According to Gladney et al. (1998) between 30% and 40% of the original documents pose legibility problems, due mainly to their great age and rough handling. Damage includes faded ink, stains, and seepage of ink from the reverse side of documents. The authors indicate that sometimes such damage "make it extremely difficult for a scholar to read the document." To solve the problem, the team investigated various procedures involving information filtering. In addition, they provided the end users with software for improving the legibility of the document.

One such tool is zooming, which enables closer look at certain details. The second available interface enables modifying the color palette in order to reduce the effect of stain ink fading and bleed-through. In addition, assorted nonlinear spatial filtering, which can be performed on any area of a displayed page or a selected portion, are available and enable intensification of faded ink or the removal of distracting background. Figure 6.1 displays a document before and after modification.

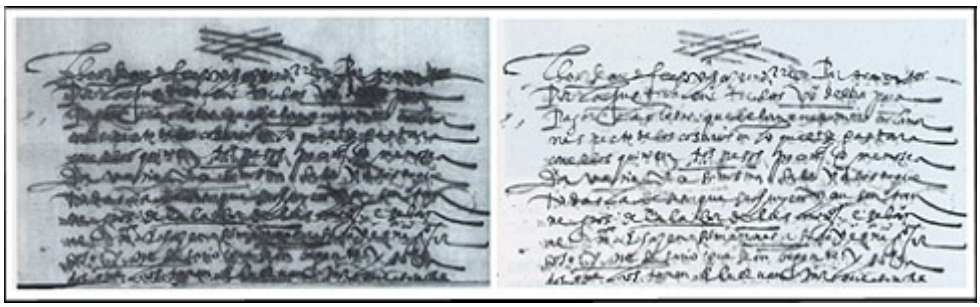


Figure 6.1: An example of ink bleed-through reduction and stain removal (AGI)

Noise in digitized text images may be a result of the scanning process itself, the quality of the paper used for the original writing or printing, or various kinds of contamination and distortions occurred throughout the centuries. Digital image processing discipline offers a wide range of filters: linear, non-linear, derivative, etc., which may help in removing certain types of noise.

The software used is either self-made or various commercial and other programs. Some of the programs use Portable Bitmaps Utilities (PMB) file formats already mentioned above. Those formats are designed to be as simple as possible. Here is an example of a header of a very small (2x1mm.) image in .pgn (‘portable graymap) file format:

```
P2
# CREATOR: XV Version 3.10a  Rev: 12/29/94 (PNG patch 1.2)
4 3
255
57  39  55  91
37  14  17  39
79  57   2  27
```

The header is always written in ASCII and the items are separated by white space and carriage returns. The first line displays the MagicValue, which identifies the file type. The second line is a comment and the third gives the width and height of the image in pixels in decimal values. The fourth line indicates the maximum values of grayscale, starting from 0 (black) up to the highest number, in this case 255 (white). The data portion can be written in either ASCII or binary form.

For converting the image into the PMB format one can use the XV (http://www.trilon.com/xv/), a program developed by John Bradley and for reading the results one tool used is the GIMP - the GNU Image Manipulation Program, which is a freely distributed piece of software suitable for such tasks as photo retouching, image composition and image authoring (http://www.gimp.org/the_gimp.html), as well as Adobe Photoshop.

For some of the experiments Matlab ("MATrix LABoratory") has been used. This is a commercial program widely used in the sphere of digital signal processing. It serves as a tool for doing numerical computations with matrices and vectors. With its large variety of ready functions and toolboxes, it indeed saves a great deal of programming work. However, the conventional user does not know what’s in the code and must trust that the

authors indeed tested properly the tools they offer. In the discipline of software engineering, a rule of thumb is that only around 90% of the lines of a program are actually ever tested (Haikala & Märijärvi 2000: 277).

As a test image I use a portion of figure 5.2 which displays two pages from Vulcanius' book from 1597 (Figure 6.1). The original text was first photographed into a 40 mm/56 (1.57") slide, which was digitized with a flatbed scanner.

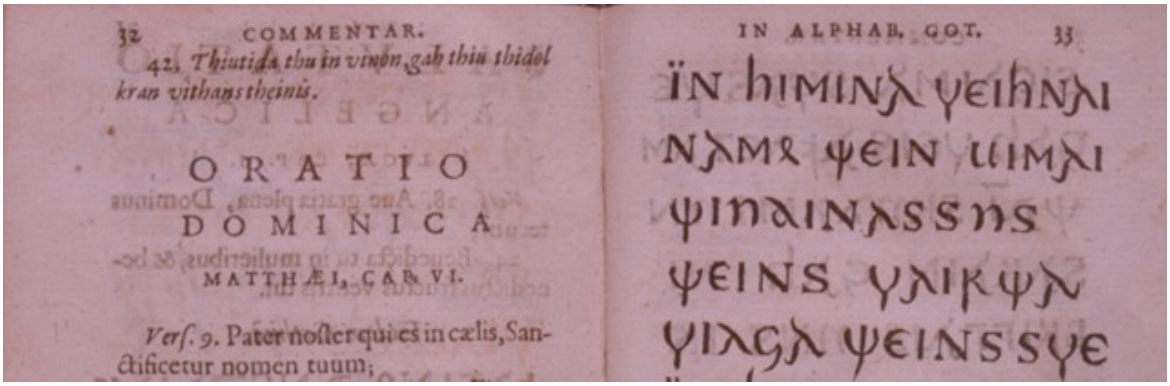


Figure 6.1. The image to be filtered

There are various types of noise in this image and the aim is to examine how different filters succeed in removing them.[5] Since we are dealing with image processing, the ultimate criterion for success or failure is in the eyes of the beholder. The mean square error or other objective criteria are irrelevant, since, for example, an image with a large mean square error might be visually more acceptable and pleasant than another with a smaller mean square error.

Before embarking on the examination of the various filters, here are two techniques widely available. The first one is brightening or darkening the image. In order to achieve these effects, the pixels are scaled with a certain factor. If the multiplier is less than one, the values of the pixel are getting closer to 0 (black). Otherwise the image brightens (Figure 6.2). The results of the scaling should be in the range of 0-255 and in a mode that the file format can read them. The result of this process is not necessarily improvement in reading over the original, as the calculation is done across the matrix indiscriminately.

However, it may result in aesthetic improvement, which, although subjective, is an important criterion in evaluating filtering quality.

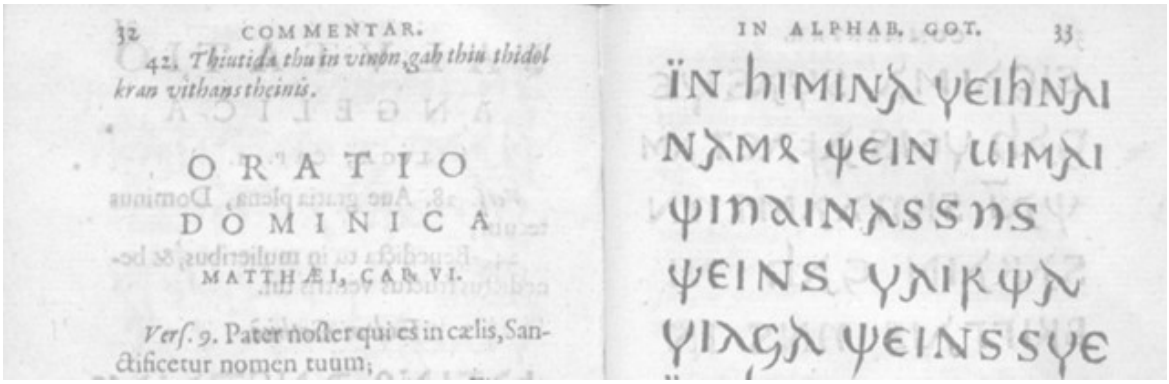


Figure 6.2. The image being brightened

The second common technique is zooming in or out. In this case, each pixel is incremented several times as each zooming is done with certain multiplication, e.g. 4. Figure 6.3 displays the enlarged left-bottom portion of the test image.

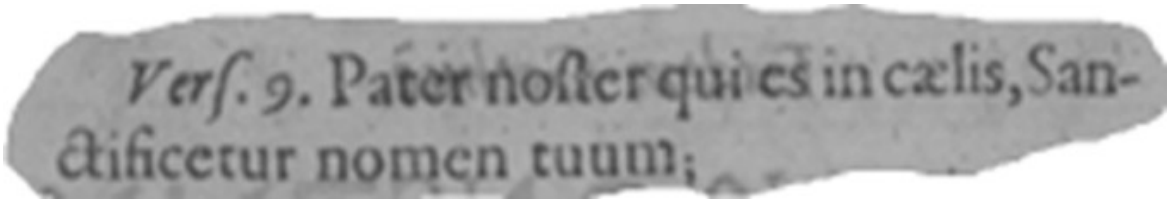


Figure 6.3. A portion of the test image being enlarged

In digital image processing, zooming is a very powerful and useful tool. The discussion below deals with various procedures that are available for image enhancement.

6.1. Histogram Equalization

A histogram is an approximation of the probability density function of a random variable. In a 256-grayscale image, it shows how many times a particular pixel intensity occurred in a certain image. Histogram equalization is mostly used as a method of enhancing low contrast images. Figure 6.4 displays the histogram of the sample image (Figure 6.1). As one can see, the distribution of pixels is concentrated in one block.

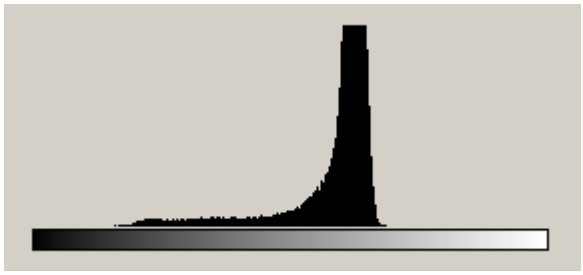


Figure 6.4. The histogram of the sample image

The histogram of the input image is analyzed and transformed to create an output image which has, as nearly as possible, the same number of pixels in at each gray level (Figure 6.5).

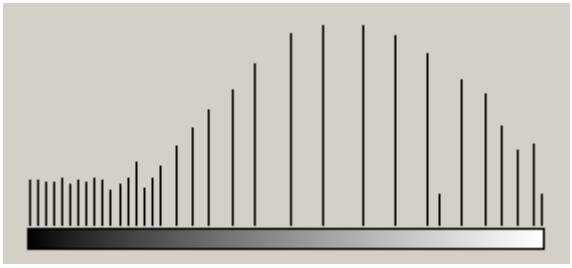


Figure 6.5. The histogram after equalization

As a result, the image displays a higher contrast (Figure 6.6).

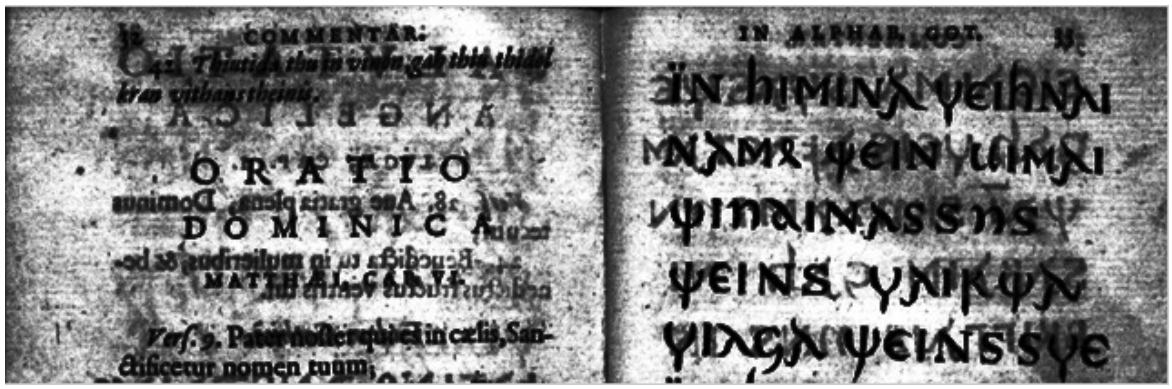


Figure 6.6. The modified image

The problem is, in this case, that both the relevant information and noise has been strengthened and no improvement has been achieved.

Huttunen & Yli-Harja (1999) suggest a fast algorithm for updating the local histogram of multidimensional signals. According to the authors, the algorithm is applicable quite generally, also for updating data structures related to image enhancement transforms.

6.2. Mean Filters

In general, a filter in the spatial domain is an array of numbers which is moved over the image to be processed. The value at the center is replaced by the result of multiplying the contents of each cell by the corresponding pixel and dividing the result by the number of cells. The set of values used in the filter is often called the kernel of the filter. In the case of mean filter, each pixel in the image is replaced by the average of itself and its nearest neighbors. Filters are very often 3 x 3:

1/9	1/9	1/9
1/9	1/9	1/9
1/9	1/9	1/9

but can also be 5 x 5 or more. The mean filter is a linear operation, which means that the value of an output pixel is a linear combination of the values of the pixels in the input pixel's neighborhood.

The filter can work separately on the original signal and the noise, however, one result is that also the underling signal is distorted. Figure 6.7 displays the results of filtering the original image with 5x5 window.

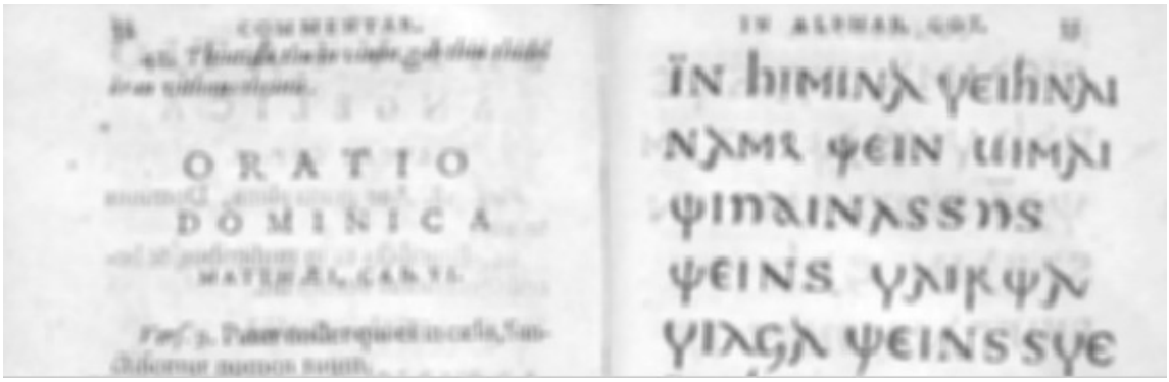


Figure 6.7. The image filtered with 5x5 mean filter

In general, an averaging filter is useful for removing grain noise from a photograph, as local variations caused by grain are reduced. However, as can be observed, the image has been blurred.

A 3 x 3 filter will not return a valid result for the pixels right at the edge of an image (for 5 x 5 this extends to 2 pixels). One way to deal with this problem is to leave the row(s) and column(s) at the edges unaltered.

A version of the mean filter is (r,s) fold-trimmed mean filter. This filter ignores the extreme values, low (r pixels) or high (s) and calculates the mean of the rest, a method used also in some sport competitions. In figure 6.8 the window used is 7x7 and the extreme value in each side was deleted.

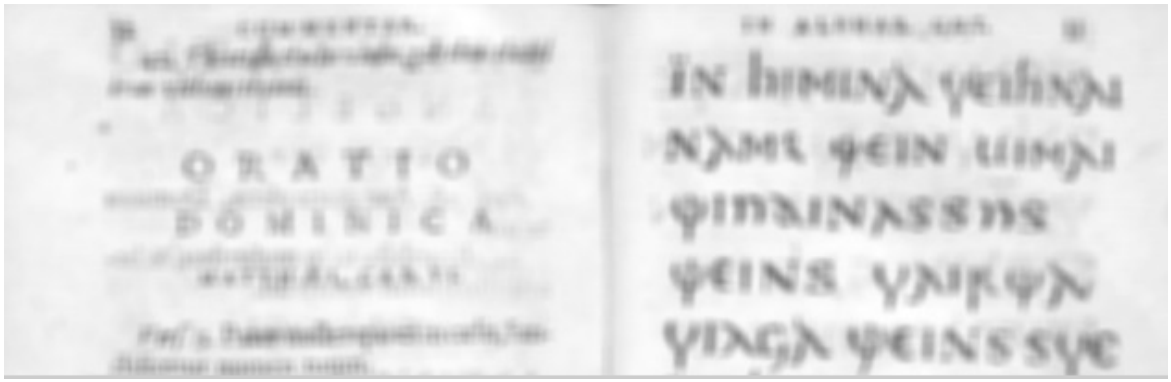


Figure 6.8. The image filtered with 7x7 fold-trimmed mean filter

Another version is the (r,s) fold-winsorized mean filter. Like the (r,s) fold-trimmed mean filter, this filter omits the extreme values, low (r pixels) or high (s) but instead replaces the low values with X(r+1) the values of the r samples and X(n-1) the s largest samples. If r=s, the amount of the trimmed element is often marked as a proportion of half of the window: $\alpha = j/N$ $(N/2 \geq j \geq 0)$ where αN samples are trimmed at each end. In case $\alpha = 0.5$, trimmed- or winsorized-mean filters are in practice median filters (see below). In figure 6.9 the window used is 9x9 and the two extreme values in each side were replaced.

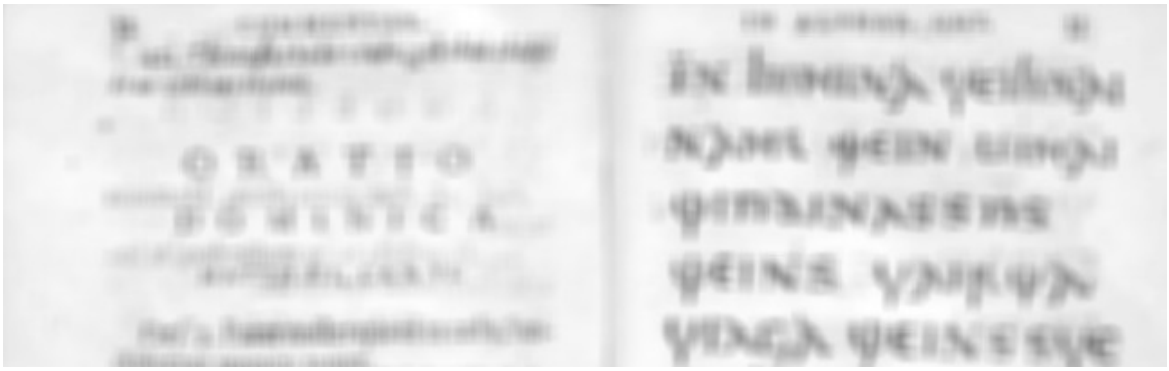


Figure 6.9. The image filtered with 9x9 fold-winsorized mean filter

The mean filter blurs the image. The noise is gone but so are the steep edges, as the mean filter cannot keep steps. All the details are preserved, but they are fuzzy and blurry.

6.3. Median Filters

The Median filter looks at all the pixels in the, e.g., 3 x 3 area around the target pixel and ranks them in order. The center pixel is replaced by the 5th brightest (out of 9). The aim of this operation is to iron out random noise but not to lose edge information or contrast.

The median filter is nonlinear and cannot be applied to the signal and the noise separately. Figure 6.10 displays the original image after being filtered with a median filter of 5x5 window.

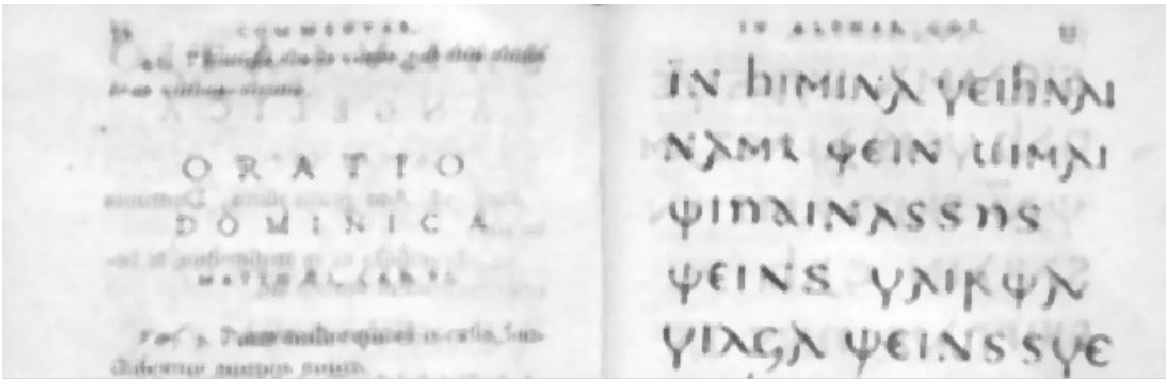


Figure 6.10. Filtering with a median filter of 5x5 window

The filtered image is quite blurred, although a median filter is supposed to preserve sharpness. One major feature in median filtering is the ability to remove impulses. Figure

6.11 illustrates what happens when a window of length 7 is used; all impulses of length less than 4 are completely removed.

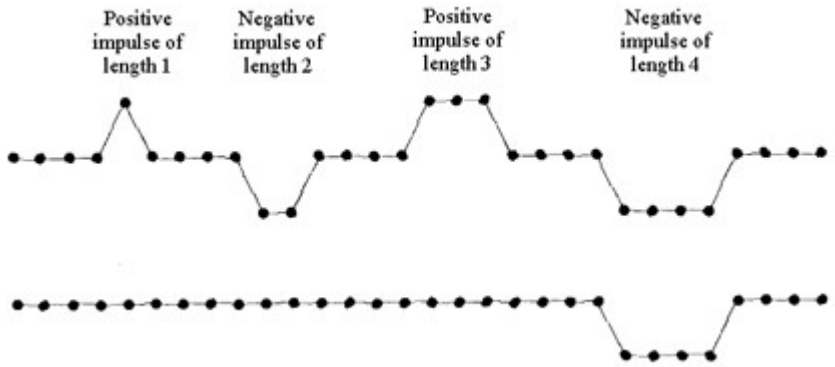


Figure 6.11. Removal of impulses in median filtering (Astola & Kuosmanen 1997: 9)

However, since the image includes letters of different sizes, one danger in using median filters in text images is lose of information, which, of course, is an unacceptable result.

One of the major problems with the median filter is that it is relatively expensive and complex to compute. To find the median it is necessary to sort all the values in the neighborhood into numerical order and this is relatively slow. The basic algorithm can, however, be enhanced somewhat for speed. A common technique is to use the fact that when the neighborhood window is slid across the image, many of the pixels in the window are the same from one step to the next, and the relative ordering of these with each other will obviously not have changed. If the window is small, then such a technique does improve the speed of calculation much, as almost all the values change; however in large windows the saving in calculation time can be significant.

There are many forms of median filters. Some, like the mean ones, ignore the extreme values on both sides and others use various sizes of windows and inner windows. However, like the various mean filters, median filters are not very good at preserving small details.

6.4. Highpass and Lowpass Filters

The terms highpass and lowpass are usually reserved to the frequency domain. In this sense, the image is transformed into the frequency domain, multiplied by a certain frequency response and then inverse-transformed back into the space domain. However, by using kernels, high- and lowpass filtering tasks can indeed be performed in the spatial domain. This mode of filtering is very useful in environments with restricted computing power, as in mobile devices. In practice, what is actually done in the space domain is convolution, which is parallel to multiplication in the frequency domain. However, in high-power processors filtering in the frequency domain is computationally faster to perform two 2D Fourier transforms and a filter multiply than to perform a convolution in the image (spatial) domain. This is particularly so as the filter size increases.

For implementing a highpass filter in the spatial domain the shape of the impulse response should have positive coefficients near its center and negative coefficients in the outer periphery (see Figure 6.12).

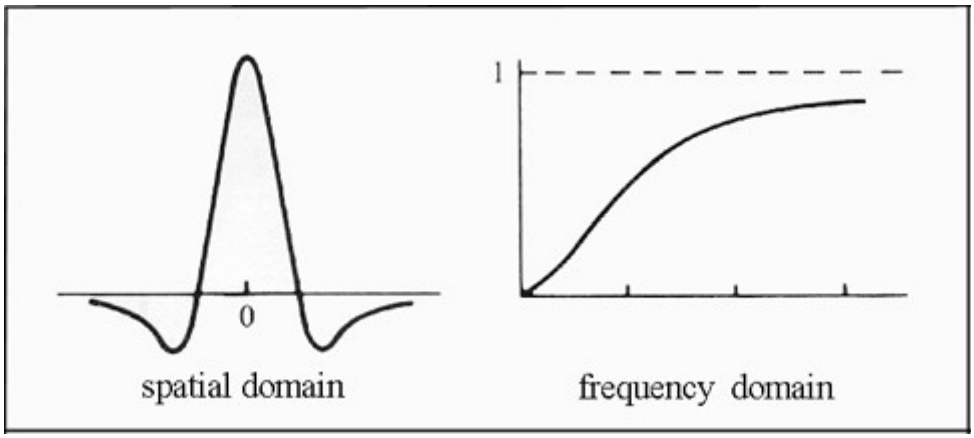


Figure 6.12. Cross sections of basic shapes of highpass filters

For a 3x3 mask, choosing a positive value in the center location with negative coefficients in the rest of the masks meets this condition:

-1/9	-1/9	-1/9
-1/9	8/9	-1/9
-1/9	-1/9	-1/9

Figure 6.13 displays the result of filtering the original image with such a 3x3 highpass window.

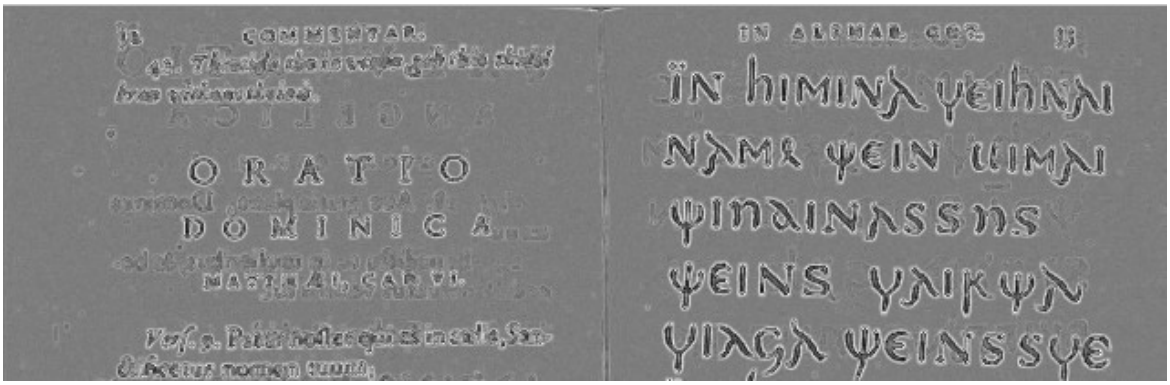


Figure 6.13. The original image after highpass filtering in the space domain

Alternatively, the highpass filtering can be done in the frequency domain. With the help of the Fast Fourier Transform, the image is converted into the frequency domain and a chunk of amplitudes of waves with low frequency is deleted (Figure 6.14).

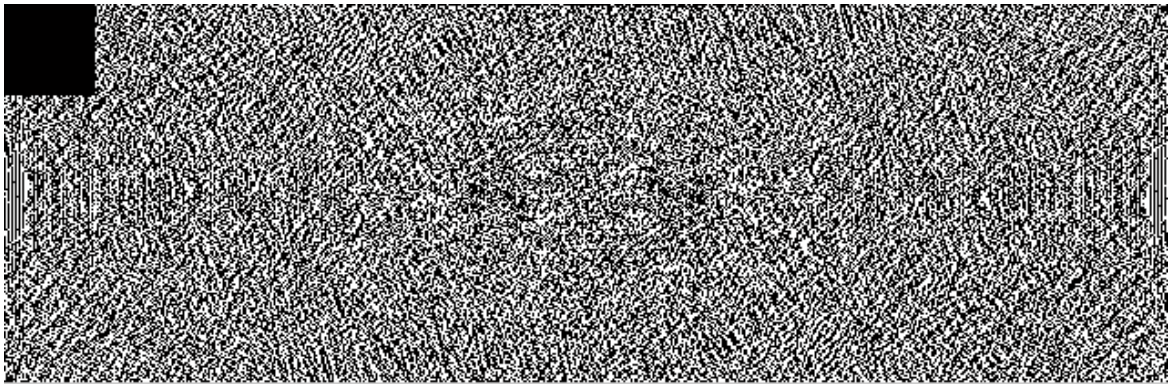


Figure 6.14. The frequency domain

Applying the inverse transform the filtered image is restore (Figure 6.15).

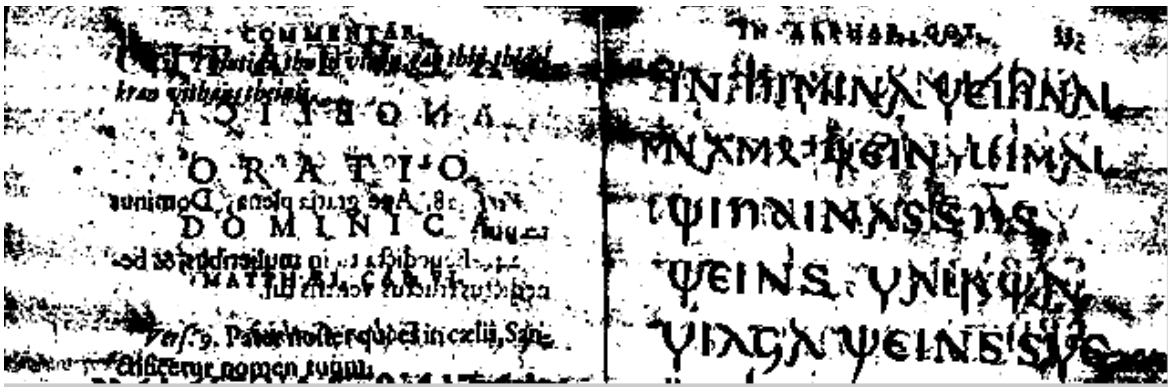


Figure 6.15. The original image after highpass filtering in the frequency domain

As a matter of fact, there is still a third alternative: Highpass = Original – Lowpass. Since a mean filter is actually a lowpass filter, subtracting the mean-filtered image from the original produced an image with the high frequencies (figure 6.16). I find this method of performing highpass filtering quite interesting and worthy of further studies. The filtered image has a very narrow range of intensity values that can be quite easily separated, using multilevel thresholding method (see below). Other kinds of lowpass filtering masks should be examined.

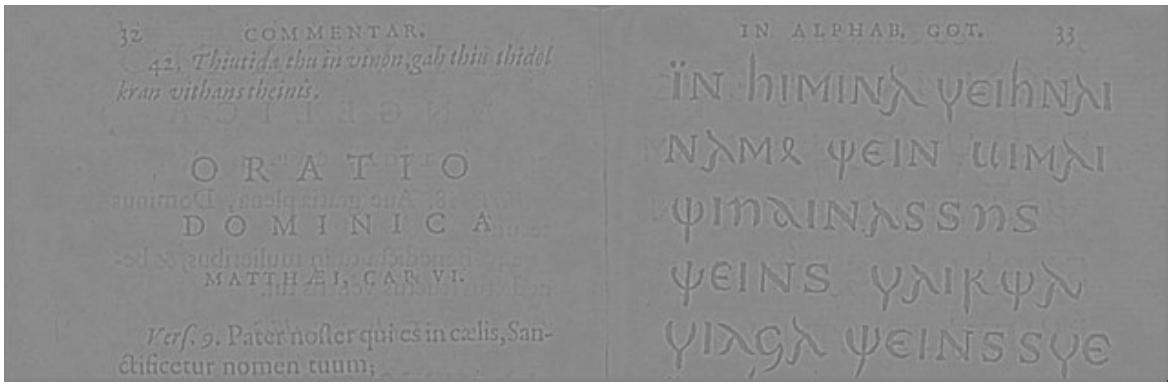


Figure 6.16. High frequencies produced by image reduction

Theoretically figure 6.13, 6.15 and 6.16 should have produced, more or less, the same results and indeed they do, except for the background and slight variations due to quantization errors and effects around the border of an image when convolving it in spatial domain.

As intended, a highpass filter, executed either in the spatial or frequency domains, sharpens the edges in the image and enhances details that have been blurred. However, it has the same effect on both the original image and the noise.

6.5. Derivative Filters

The aim of the derivative filters is to examine whether an edge passes through or near a given pixel. This is done by examining the rate of change of intensity near the pixel; sharp changes are good evidence of an edge.

Averaging the pixels over a region is analogous to integration, as it tends to blur detail in an image; differentiation can be expected to have the opposite effect, that is, sharpen the image (Rosenfeld & Kak, 1982: 280). Differentiating an image, that is a two-dimensional function *F(x,y)*, produces a vector gradient:

$$\nabla \mathbf{f} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix}$$

However, since the gradient is defined to continuous function, another term has been coined - digital gradient:

$$\mathrm{GRAD}(f) = [\mathrm{DX}(f), \mathrm{DY}(f)]$$

where GRAD(*f*) is an ordered pair of images, and the notation [,] denotes the integer spatial coordinates. The function produces the degree of overall change of gray level at the pixel at which it is being evaluated (Dougherty & Giardina, 1987: 79). In practice, instead of calculating the differential, as in continuous functions, one calculates the difference between neighboring pixels. For a convolution mask of 3x3

A_0	A_1	A_2
A_7	*f(j,k)*	A_3
A_6	A_5	A_4

the partial derivative of the two components will be computed following these formulae (Rosenfeld & Kak, 1982: 287):

$$f_x = (A_{6+} CA_5 + A_4) - (A_0 + CA_1 + A_2)$$
$$f_y = (A_{2+} CA_3 + A_4) - (A_0 + CA_7 + A_6)$$

There exist several masks that derivate the image in x and y directions. One of them is called the Sobel operators where the variable C in the above formulae is equal to 2. The image is computed by both horizontal and vertical masks sequentially and the outputs are combined.

$f_x =$

-1	-2	-1
0	0	0
1	2	1

$f_y =$

-1	0	1
-2	0	2
-1	0	1

All mask coefficients sum to 0, indicating a response of 0 in the constant areas, as expected of a derivative operator. Figure 6.17 displays the results of using Matlab's 'sobel' command.

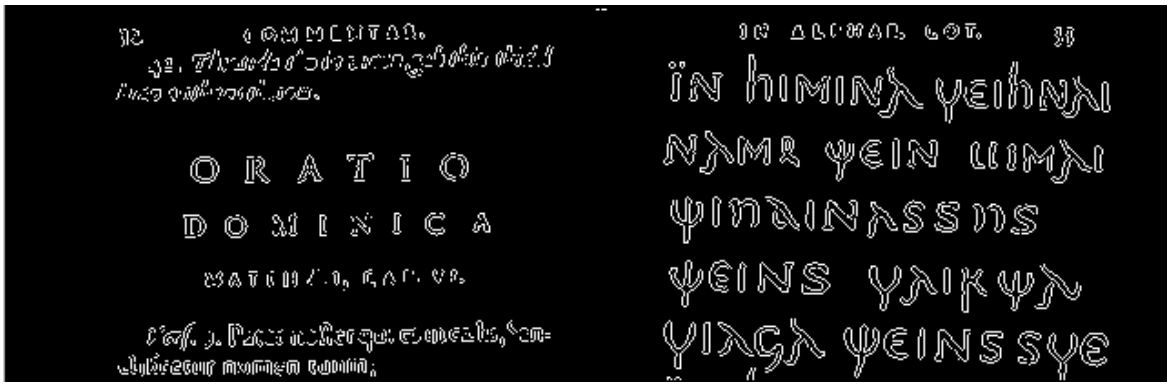


Figure 6.17. The original image after filtering with the Sobel operators

Another derivative filter is the Laplacian. It is obtained by applying one more time the derivative mask to each one of the derivative images f_x and f_y to obtain f_{xx} and f_{yy}. The result is a non-directional second derivative with a convolution mask:

-1	-1	-1
-1	8	-1
-1	-1	-1

It gives a zero result on any uniform or smoothly varying image region but a large positive or negative response at edges, lines and points. Figure 6.18 displays the results of GIMP's Laplace command.

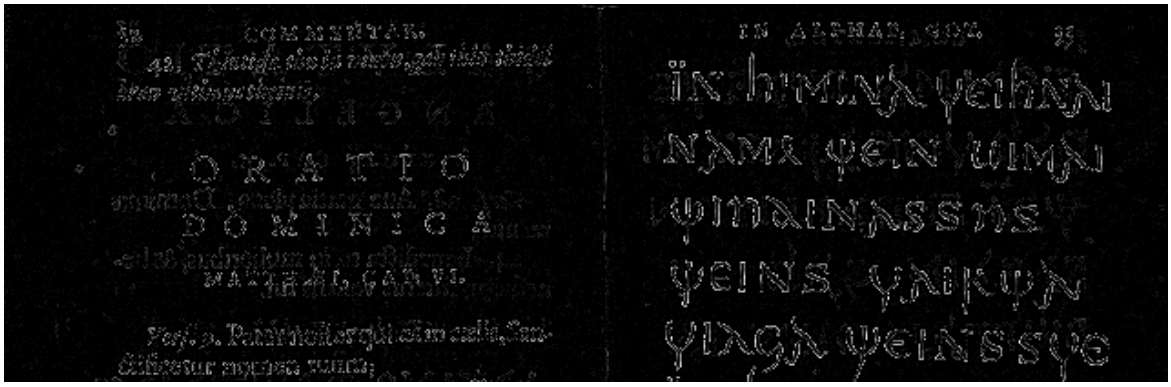


Figure 6.18. The original image after filtering with the Laplacian filter

The main advantage of the various derivative filters is that they enhance prominent edges considerably. They may serve well in situations were the detection of the contour of large objects is important, like in security checking of luggage in airports. However, as far as detecting delicate items like letters and punctuation signs, those filters are too rough.

6.6. Thresholding

Thresholding is an important tool in image segmentation. It is used to distinguish an object from its background in order to enable, for example, machine reading. The basic idea is to create either black or white pixels. All grayscale values under a certain threshold are reduced to zero (black) and above this threshold to white (255). The result is a binary image with, for example, all black pixels are 1 and white 0.

As argued up to now, filtering is a neighborhood operation, in which the value of any given pixel in the output image is determined by applying a certain algorithm to the values of the pixels in its neighborhood. In this respect, a pixel's neighborhood is some set of pixels defined by their locations relative to that pixel. In a certain sense, thresholding,

although not completely conforms to this definition, can be considered as a filter and be used for text images. As an example, creating a binary image where the intensity value of 130 is the threshold, keeps all the information in the image but also some of the noise. Unfortunately, in some cases the text and the noise are indistinguishable (Figure 6.19).

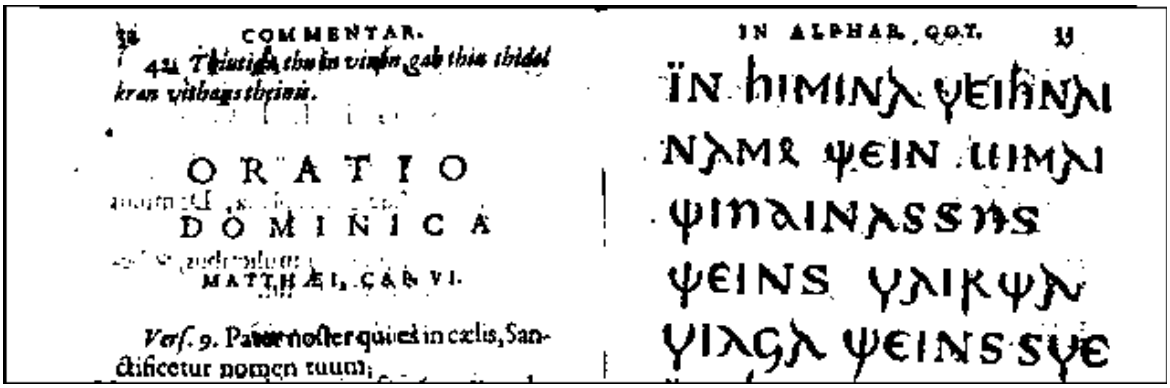


Figure 6.19. Thresholding

Keeping the 256 grayscale level of intensity and doing a multilevel thresholding (Rosenfeld & Kak: 66), e.g. using two threshold points and leaving the pixels in between unaltered, produces better results. Figure 6.20 is a result of leaving intact the pixels between 120 and 130. Another advantage of fine thresholding is that it can be done also with a color image. In this case each of the primary colors, red, green and blue can be handled separately (Landau 2001). To remove the rest of the noise, one must use digital brush and clean the image carefully several pixels at a time.

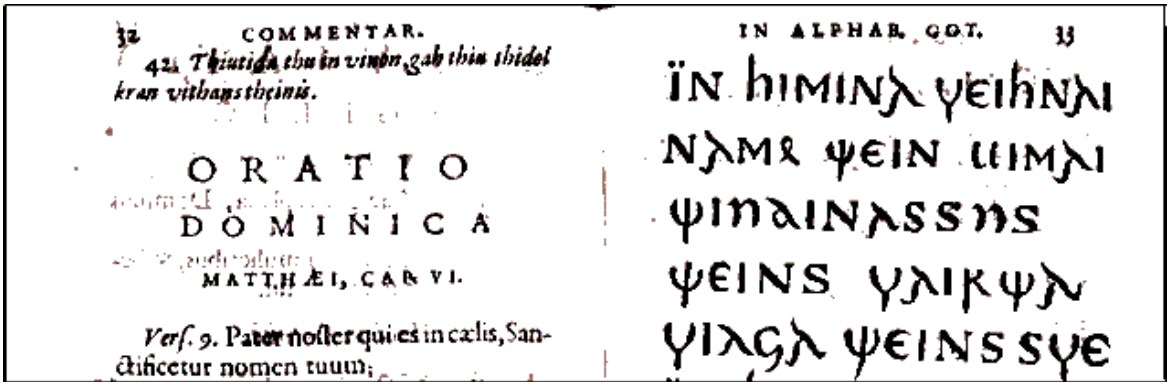


Figure 6.20. Multilevel thresholding

In my opinion, fine thresholding, although a 'simple' technique, is a very powerful tool and can succeed where other, much more mathematical and sophisticated filters do not produce any improvement in legibility.

6.7. Image Fidelity Criteria

Jain (1989: 57) defines rating scales for evaluation of image quality. One such scale is the global goodness scale (table 6.1) which rates image quality on a subjective scale, ranging from excellent to unsatisfactory. The goodness scale is based on comparisons within a set of images.

Table 6.1. Overall goodness scale (Jain 1989: 58)

Excellent	(5)
Good	(4)
Fair	(3)
Poor	(2)
Unsatisfactory	(1)

Table 6.2 details the application of this scale to the results of the various filters used in this chapter. The details to be examined are those which are of significant to the eventual user, that is the philologist. The overall esthetical presentation of the image is of little importance.

Table 6.2. Results of the various filters

Details / Filters	Preserving big letter	Preserving small letters	Punctuation marks	Removal of noise
histogram equalization	3	3	3	2
mean filter (5x5)	2	1	1	3
medium filter (5x5)	2	1	1	3
highpass filter	3	1	1	2
Sobel	4	2	1	4
Laplacian	2	1	1	2
thresholding	3	3	2	3
multilevel thresholding	4	3	4	3

It seems that the only method that produce somewhat satisfactory results is the multilevel thresholding. However, one must be very careful when utilizing this tool as it is quite easy to remove essential information, as small letters or punctuation marks.

Digital image processing has transformed the field of image enhancement from the domains of optics and chemistry into the domain of mathematics. In this domain, image processing is actually matrix calculation and the possibilities for manipulations are numerous. The bad news is that an image with noise does not necessarily fit the plethora of digital filters, it seems that de-noising text images is not an easy task to accomplish and cannot be automated.

However, one should not give up. Reading digitized images of text on a screen, like reading microfilms, is not a pleasant undertaking. Moreover, the printout of a digitized text image is usually of low quality. The noise in a digitized image may have its genesis in the book being scanned, or is a result of the digitizing process, or both. One possible alternative to improve the outcome is to go back to the good old methods and to try to combine digital technology with various optical filters, special lightening, and various forms of radiation (see below).

7. The use of X-rays, Ultra-violet Lightening, etc.

The idea of producing a facsimile edition of a valuable document is not a new idea. One example is the history of the Codex Argenteus. Already in 1706, Lars Roberg, an Uppsala physician, drew and made a woodcut of one page of the manuscript. The woodcut is still extant (see Munkhammar 1998: 178).[6] In the 19th century, with the advent of photography, there were plans to produce a facsimile edition by photographic means. Those plans finally materialized in 1927, when Uppsala University celebrated the 450th anniversary of its foundation (Kleberg 1984). In this chapter I survey the process of preparing this facsimile edition and try to see whether some of the techniques used can be applied also to modern digital technology.

One of the people involved in the facsimile project was Theodor Svedberg (1884-1971), a chemist who in 1926 won the Nobel Prize in chemistry. The fact that a chemist was involved in the project was quite natural, as photography involves various chemical processes and compounds. Another aspect, which required involvement of a chemist, was to examine the influence of various modes of radiations and light absorption on the ancient manuscript. As the text was written apparently in the 6th century on a parchment, then exposing it to ultraviolet and infrared radiation may reveal information that cannot be seen or photographed with white light, however, the material might be somehow damaged too.

According to the report attached to the facsimile edition (1927), Svedberg and Ivar Nordlund started to investigate the possibility of making a legible reproduction of the manuscript by photographic means already in 1917. For the study they used five leaves that had been earlier torn off the Codex. The leaves exhibit all the typical damages of the manuscripts, such as discoloration, falling off the silver and gold used for writing the text, and penetration of ink from one side of the leaf to the other.

Two of the methods they used proved to be decidedly superior to the rest. The first one, which came to be referred to as the ultra-violet method, involved photographing with reflected ultra-violet light of the wavelength 366 μμ (picometer, mikromicron - one trillionth of a meter). In the second method, the fluorescence one, the fluorescence is

excited with the same wavelength of 366 μμ. As these two methods complement one another in certain respects, it was decided to give a reproduction of each page of the Codex by both these methods.

As it turned out, the ultra-violet and the fluorescence methods did not always produce satisfactory results and a complementary section was added. It includes photographs of some of the pages produced in three other different methods. The first method was photography with yellow filter, the second involved using secondary X-rays and the third used oblique illumination. According to the authors, photography with secondary X-rays, invented by Svedberg and Nordlund themselves, turned out to be the most fruitful for the reconstruction of the original text. Working on restoring the text from the facsimile edition (1927), I can testify that indeed this method produced the best results (Landau 2001). I can only regret that, unlike the ultra-violet and fluorescence methods, not all the pages were photographed in this method.

Modern photography is also making use of various electromagnetic radiations.[7] The basic idea is to cause emission of photons with relatively short wavelength. One result is fluorescence, which is the emission of electromagnetic radiation in the visible region that occurs during excitation caused by ultraviolet energy and light. This kind of radiation can be produced by, for example, a lamp with two electrodes in an atmosphere of very low-pressure mercury vapor. Fluorescent lamps find widespread use because they generate very little heat. Ultraviolet and infrared radiation can sometime reveal information that cannot be seen or photographed with visible light.

While reading the report written by those who produced the facsimile edition in 1927, one noticed that technology has apparently changed since that time. For example, the time of exposure “exceeded one hour only in exceptional cases.” The average time for St. Matthew was 9 minutes, St. Luke 8 minutes, St. John 27 minutes and St. Mark 22 minutes. The negatives were printed on daylight paper and from those prints the phototypes were then prepared. The photographs with secondary X-ray radiation were taken at the Röntgen-ray department of the University Hospital in Uppsala. The stabilivolt apparatus of the hospital was used in conjunction with an A.E.G tube for depth therapy. The authors

maintained that “where destroyed letters in the manuscripts have at some time been restored with ordinary ink, the Röntgen-photographs can often verify the original text.”

Using X- ray radiation as a general tool for digitizing text heritage may be overkill and probably not suitable but for extreme cases. In this category one should also examine other technologies, such as magnetic resonance, ultrasound, etc. However, proper illumination is an indispensable part of doing the work properly and may also serve as a filtering device to remove certain noise.

From my experience with the images of the facsimile edition, I would like to point out another advantage of using radiation – a sort of useful intensity slicing technique. The basic idea is to slice the range of intensity levels of grayscale image into certain number of regions, giving each region a different color. The technique can be used, as an example, for emphasizing a certain feature in a medical image, or possibly, for transforming a black and white old movie into a modern one with colors. Another application is converting continuous elevation data in maps into discrete intervals (0-100, 101-200 meters, etc.)

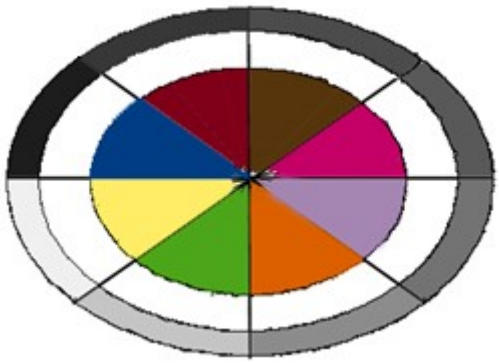


Figure 7.1. Intensity slicing

While working with the photos of the facsimile edition, I have noticed that the various radiation techniques have substantially reduced the number of colors in the original manuscript into a manageable range of colors and that enabled me to do some fine thresholding and clear the rest of the noise with a digital brush.

8. Surface Analysis

One of the methods used in preparing the facsimile edition of the Codex Argenteus (1927) was photography with oblique illumination, where only one source of light was used. The source was so placed that the light struck the paper at an angle of 45°. The camera was also placed obliquely so as to receive the light reflected directly from the manuscript. In this way, the indention left in the parchment by the writing which has fallen off, were revealed quite distinctly against the smooth reflective surface of the parchment. By choosing a suitable kind of illumination, it was possible to accentuate even more the contrast on the negative.

The pulp and paper industry uses certain methods for automatic quality control, which involve the examination of the surface of the paper. It is possible that the same methods, with certain adaptation, might be useful for deciphering ancient manuscripts.

One usage of this technology is detecting pulp fibers. In general, for automated machine reading the image must be segmented. However, for our purpose, the person who is familiar with the ancient language by which the text was written in, is the decipherer using his or her eyes and knowledge.

The following brief survey[8] illustrates the technology and its possible application for ancient text deciphering. The most common textural measurements in paper science are those of formation and surface roughness. For many purposes, it is necessary to do some pre-processing on the image before making textural measurements. This often takes the form of filtering: Highpass filtering for reducing large scale variations and lowpass filtering for eliminating small scale variations like dirt or dust.

One technique used in textural measurement is Contact Beta Radiography. It is done by making a sandwich of an extended beta source, the paper sample and a piece of X-ray film which are all in good contact. A wedge consisting of a range of known basis weights of mylar film is also included. Sufficient time is allowed to elapse for the film to be exposed and then the film is developed. The study of the results is based on statistical methods.

Typically variance and coefficient of variation are reported. The degree of unevenness compared to what would be expected for a random sheet can be calculated. It is also possible to calculate values for formation from transmitted light images but care must be taken because light transmission depends on density as well as mass. Fillers and coatings can lead to misleading results.

Another possible technique involves X-Rays. Low energy X-rays have been shown to be usable for paper formation measurements and do not suffer from scattering to the same extent as beta rays. This might well provide a good method but, apparently, has rarely been used in practice.

Surface roughness can be analyzed using image analysis techniques if a matrix of surface heights has been obtained by laser triangulation or similar methods. A study of such methods is described in details by I'Anson (1998). Figure 8.1 displays laser surface profiler scan of a surface of a sample of white-lined packaging board and its 3D representation.

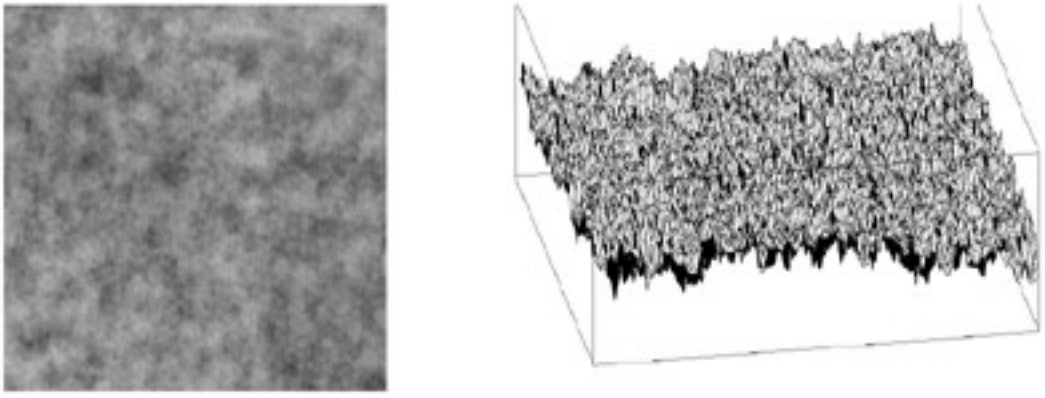


Figure 8.1. Laser surface profiler scan and its 3D representation (l'Anson 1998)

Surface roughness can be split up using FFT (Fast Fourier Transforms) band pass filtering to give a spectrum of roughness at different scales. A number of components of any paper or board machine have periodic variations which can lead to periodic non-uniformities in the structure of the product. It has been shown that such marks can be detected, identified and quantified using image analysis and two-dimensional FFTs. This procedure involves

illuminating the paper or board sample so that the mark can be clearly seen and obtaining an image of an appropriate size using a CCD camera linked up to a frame-grabber in a computer.

An identical analysis method can be used to analyze height data taken directly from the surface of the paper or board sample using a laser triangulation sensor. The incident laser beam is vertical with the sensing optics operating at an angle. In order to study paper surfaces, it is usually necessary to examine areas of 50 mm x 50 mm to 100 mm x 100 mm and it is always necessary to have at least 256 lines of 256 data points each to get a useful FFT spectrum. With a proper instrument, an area of 50 mm x 50 mm at 256 x 256 points takes approximately 7 minutes and an area of 100 mm x 100 mm at 512 x 512 points takes approximately 30 minutes.

Reading I'Anson's article is almost like reading the report written by Svedberg and Nordlund more than 75 years ago; the sophistication and precision is impressive. Obviously, the technology applied for examining paper production cannot be directly utilized for examining priceless and irreplaceable ancient manuscripts, but, taking as an example a work of the old masters, one would imagine that contemporary technology might be also applicable for our study. One such case is the study of the palimpsests, that is parchments or other writing-material written upon twice, the original writing having been erased or rubbed out to make place for the second. Figure 8.2 displays a portion of a manuscript where in antiquity someone had tried to erase the original text in Gothic, but fortunately did not completely succeed, and wrote over it a text in Latin.

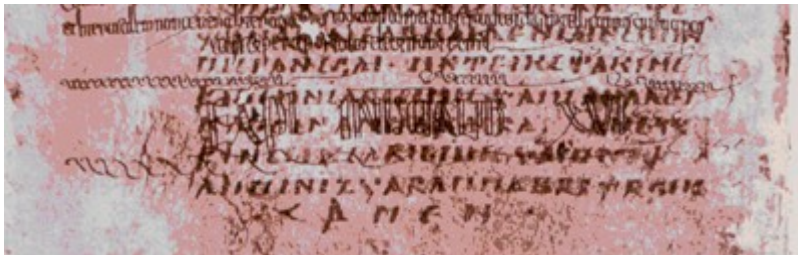


Figure 8.2. A palimpsest

9. Conclusions

Digital technology offers a new approach for preserving and displaying text heritage. The tools and methods are already available and, in principle, work can start right away.
However, as digital technology changes so rapidly, a project such as transferring a huge collection of old books and manuscripts into digital form and displaying it, or part of it, on-line can be planned only in broad outlines. As Pedro González of the AGI project in Seville writes, " …institutions learn by doing and learn from each other." In fact, the immediate task here is to be convinced and to convince others that such an undertaking is indeed reasonable, plausible and feasible, and that work can start right away.

Indeed, the main theme of this work is flexibility. There are hardly any precedents for this kind of undertaking, and as technology changes so fast, the main hurdle is how to keep an open mind and adaptable objectives, in order to avoid needless extra work and distractions. The intention here is to sketch some broad ideas as for what sort of scanning equipment, memory storage facilities and file formats be used.

The structure of the database I propose has two main components: the multimedia part, which enables smooth access to the material and the printing mechanism, which provides a relative high quality paper version of the digitized material. With due respect to computer technology, there is still room for a researcher reading a book or a printed paper laying on a desk in front of him or lying on bed while holding the printed material in his or her hands, instead of sitting in front of a monitor's screen.

The digitizing process is not without problems. There can be noise caused by a wide range of sources, such as traces of decay in the paper, seepage of ink from the other side of the page, etc. I examine several sorts of filters and other techniques and conclude that once the digitization has been completed, there is not much one can do to improve the results. Therefore, the emphasis should be on proper digitizing and illumination techniques. In special cases, like in deciphering old manuscripts, digital technology may offer some very useful tools.

In my opinion, the best approach will be just to raise enough money to complete the pilot project, purchase the appropriate hardware and software and to get to work. Problems, no doubt, will surface and they will be solved, one after the other, as the work proceeds.

Notes

1. I would like to stress that, in my opinion, HTML is still a more convenient tool for casual web pages; it is much easier to construct and maintain HTML files than XML ones. New text editors, like MS Word 97, are able to create automatically hypertext pages using XML, but if there is a mistake, unlike automatically created HTML files, it is very hard to find it. However, when creating structural documents, XML may serve as a better choice, although a great deal of work has to be invested in order to create an appropriate system.

2. Originally an acronym for "Digital Video Disk," although the subsequent introduction of DVD-ROM drives and extensive software programs has caused this to also mean "Digital Versatile Disk." As of now, there is no one actual meaning for the acronym DVD.

3. The lossy method has been developed purposely for the use in the Internet where short transmission time is essential. Another usage is in digital video compression where many photos are transmitted each second. Additional ingenious algorithms are designed to detect motion so that only certain parts of the consecutive photos are sent. With due respect to all those researchers and other professionals who labor hardly on the problem, digital video in narrow-band transmission does not work satisfactory.

4. This survey is based on Ebbinghaus (1983), Kleberg (1984), Munkhammar (1998), as well as my own examination of the books during my stay at the Library of Uppsala University in May 2003.

5. The discussion concerning the various filters and methods for image enhancement is based upon Gonzalez & Woods (1993), Astola & Kuosmanen (1997), as well as my own observations.

6. A digitized image is displayed at:
http://www.students.tut.fi/~dla/Cod_Arg/C_A_roberg.html.

7. The discussion here is based on Langford, 1989 and Stroebel & al., 1986.

8. The survey is based on the material displayed on the Internet for a course on the application image analysis methods for the paper industry, given by Dr. Stephen J. I'Anson (http://pygarg.ps.umist.ac.uk/ianson/image_analysis/Paper_egs.html), Department of Textiles and Paper, the University of Manchester Institute of Science and Technology.

References

Alkula, Riitta & Pieskä, Kari. 1994. Optical Character Recognition in Microfilmed Newspaper Library Collection: A Feasibility Study. VTT Research Notes 1592. Espoo: Technical Research centre of Finland.

Astola, Jaakko & Kuosmanen, Pauli. 1997. Fundamentals of Nonlinear Digital filtering. Boca Raton: CRC Press.

Codex Argenteus Upsaliensis Jussu Sentus Universitatis Phototypice Editus. 1927. Upsaliæ & Malmogiæ.

D. N. Jesu Christi SS. Evangelia Ab Ulfila. 1671. [Ed. Georg Stiernhielm.] Stockholmiæ.

Dougherty, Edward R. & Giardina, Charles R. 1987. Matrix Structured Image Processing. Englewood Cliffs: Prentice-Hall.

Ebbinghaus, Ernst A. 1983. Gothic Lexicography, part 1. General Linguistics, Vol. 23, No. 3, 202-215.

Extensible Markup Language (XML) 1.0 Second Edition, W3C Recommendation 6 October 2000, http://www.w3.org/TR/REC-xml.

Gladney, Henry M. et al. 1998. Digital Access to Aniquities. Communications. Volume 41, Number 4, pp. 49-57.

Goldfarb, Charles F. & Prescod, Paul. 1998. The XML Handbook. Upper Saddle River, NJ: Prentice Hall.

González, Pedro. 1998. Computerization of the Archivo General de Indias: Strategies and Results. http://www.clir.org/pubs/reports/gonzalez/contents.html.

Gonzalez, Rafael C. & Woods, Richard E. 1992. Digital Image Processing. New York: Addison-Wesley Publishing Company.

Goropius Becanus, Johannes. 1569. Origines Antwerpianae sive Cimmeriorvm Beccesel-ma. Antverpiae.

Gruter, Janus. 1602. Insciptiones antiquae totius orbis Romani. Antwerp.

Haikala, Ilkka & Märijärvi, Jukka. 2000. Ohjelmisto tuotanto. Helsinki: Satku-Kauppakaari.

Huttunen, Heikki & Yli-Harja, Olli. 1999. Fast algorithm for updating the local histogram of multidimensional signals International Symposium on Nonlinear Theory and its Applications (NOLTA), Hilton Waikoloa Village, Hawaii, pp.65-68.

Jain, Anil K. 1989. Fundamentals of Digital Image processing. Englewood Cliffs, NJ: Prentice Hall.

Kleberg, Tönnes. 1984. Codex Argenteus, the Silver Bible at Uppsala. Uppsala: Uppsala University Library.

Landau, David. 2001. The study of old Texts with the aid of Digital technology: the Gothic Manuscripts. Tampere University of Technology: Institute of Software systems. Report 26.

Langford, Michael. 1989. Advanced Photography. London: Focal Press.

l'Anson, Stephen. 1998. Analysis of Paper and Dryer Fabric Surfaces using 3D Laser Technology: Wire, Roll and Fabric Marks, Roughness of Paper and Surface Volume of Fabrics. Proceedings of PTS Symposium: Image Analysis for Quality Assurance and Enhanced Productivity (Code: IA-SY 810 MUC), 12-13 October. (http://pygarg.ps.umist.ac.uk/ianson/Papers/pts98_3html.pdf).

Moen, William E. 1998. Accessing Distributed Cultural Hertage Information. Communications of the ACM. April, Vol. 41, No. 4, 45-48.

Mori Shunji & Suen Ching Y. & Yamamoto Kazuhuko. 1992. Historical review of OCR Research and development. Proceedings of the IEEE, Volume 80, NO. 7 July.

Munkhammar, Lars. 1998. Silverbibeln: Theoderiks bok. Stockholm: Carlsson Bokförlag.

Murray, James D. & vanRyper, William. 1996. Encyclopedia of Graphic File Formats. Bonn: O'reilly & Associates.

Quatuor D. N. Jesu Christi Euangeliorum … Franciscus Junius F.F. 1665. Dordrechti.

Rosenfeld, Azriel & Kak, Avinash C. 1982. Digital Picture Processing, Volume 2. New York: Academic Press.

Sacorum Evangeliorum Versio Gothica … Cum Interpretatione … Erici Benzelii. 1750. Oxonii.

Smith, Bernard (2000). Digital Heritage and Cultural Content in the New Information Society Technologies Programme. http://www.cultivate-int.org/issue1/ist.

Stroebel, Leslie & Compton, John & Current, Ira & Zakia, Richard. 1986. Photographic Material and Processes. London: Focal Press.

Tanenbaum, Andrew S. 2003. Computer Networks. Upper Saddle River, NJ: Pentice Hall.

Ulfilas Gotische Bibelübersetzung die älteste Germanische Urkunde nach Ihre'ns Text … herausgegeben von Iohann Christian Zahn. 1805. Leipzig.

Vulcanius, Bonaventura. 1597. De literis & lingua Getarum, siue Gothorum. Lvgdvni Batavorvm.