FAKULTÄT FÜR !NFORMATIK

Faculty of Informatics

Digital archiving, processing, and didactic use of historical source documents.

The case of the Stock collection.

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Magister der Sozial- und Wirtschaftswissenschaften

im Rahmen des Studiums

Informatikmanagement

eingereicht von

Matthias Rainer

Matrikelnummer 9804140

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung: Ao. Univ.Prof. Mag. Dr. Karl Fröschl

Wien, 08.04.2012

(Unterschrift Verfasser) (Unterschrift Betreuung)

FAKULTÄT
FÜR !NFORMATIK
Faculty of Informatics

Digital archiving, processing, and didactic use of historical source documents.

The case of the Stock collection.

MASTER'S THESIS

submitted in partial fulfillment of the requirements for the degree of

Mag.rer.soc.oec.

in

Informatics Management

by

Matthias Rainer
Registration Number 9804140

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Ao. Univ.Prof. Mag. Dr. Karl Fröschl

Vienna, 08.04.2012

(Signature of Author) (Signature of Advisor)

Erklärung zur Verfassung der Arbeit

Matthias Rainer
Koppstraße 12 / 5, 1160 Wien

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit - einschließlich Tabellen, Karten und Abbildungen -, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

(Ort, Datum) (Unterschrift Verfasser)

Acknowledgments

Above all I want to thank my wife, Catherine Novak Rainer, and my parents, Elisabeth and Fritz Rainer, for their patience, continuous support, and encouragement during my student years.

I would also like to thank my adviser, Dr. Karl A. Fröschl, for many inspiring discussions and a very positive work atmosphere, and Dr. Gerald Futschek for the opportunity to attend his seminar.

Abstract

A large amount of information still dwells in archives, printed on paper and out of reach of today's search engines. While it might not be necessary (or even possible) to manually process all that data, it could still prove useful to have it indexed and available for search. Such digitized information archives could be used as reference in research and educational projects.

This thesis is the documentation of such a digitization project. The analog media used for this project is the Stock collection: twelve cardboard boxes filled with approximately 1400 articles, information material, and other papers with a focus on automation and information technology in libraries. The state of the art is evaluated based on the U.S. Library of Congress' American Memory Project. The best practices learned in the American Memory Project are freely available online as guidelines for anybody planning to digitize paper media. These guidelines will be evaluated in an attempt to transfer the Stock collection from its original paper form to a hypertext system. The digitized articles of the Stock collection will then be made available online as part of the "Die Informatisierung Österreichs" project, hosted by the Austrian Computer Society. The motivation to do so is to preserve this collection and make it available for further research.

Since only a fraction of the Stock collection was transferred during the evaluation, a manual for further processing will be provided as part of this project, with the intention that the process steps can be applied to other projects as well.

Kurzfassung

Unzählige Information schlummern noch in Archiven, auf Papier gedruckt und außerhalb der Reichweite heutiger Suchmaschinen. Auch wenn es nicht notwendig (oder gar möglich) ist, all diese Informationen händisch zu verarbeiten, könnte es sich doch als nützlich erweisen, diese Daten indiziert und für eine Suche verfügbar zu haben. Derart digitalisierte Archive könnten als Referenzen in Forschungs- und Unterrichtsprojekten dienen.

Diese Magisterarbeit ist die Dokumentation eines solchen Digitalisierungsprojekts. Das analoge Medium, das diesem Projekt zugrunde liegt, ist die Sammlung Stock: zwölf Kartons, gefüllt mit ca. 1400 Artikeln, Informationsblättern und anderen Unterlagen zum Thema Automatisierung und Informationstechnologie in Bibliotheken. Der aktuelle Stand der Technik wird anhand des American Memory Projekts der U.S. Library of Congress ermittelt. Die Methoden, die im Rahmen des American Memory Projekts entwickelt wurden, sind im Internet frei erhältlich und stehen jedem zur Verfügung, der selbst die Digitalisierung von Papierdokumenten plant. Diese Handlungsrichtlinien werden anhand des Versuchs, die Sammlung Stock in eine digitale Form zu überführen, untersucht. Die digitalisierten Artikel werden im Anschluss als Teil des Projekts "Die Informatisierung Österreichs", welches von der Österreichischen Computer Gesellschaft betrieben wird, verfügbar gemacht. Die Motivation für dieses Projekt entspringt dem Wunsch, die Sammlung Stock zu erhalten und für weitere Forschungsarbeiten zur Verfügung zu stellen.

Da im Rahmen dieser Evaluation nur ein Bruchteil der Sammlung Stock digitalisiert wurde, wird außerdem eine Anleitung zur Digitalisierung erstellt, um die Weiterführung der Arbeit zu ermöglichen.

Contents

CHAPTER 1

Introduction

Introduction

This chapter will give an introduction to the origin of this thesis and provide information on the Stock collection, the center piece of the practical project work. Furthermore, the project goals are defined, and a roadmap for the rest of this thesis is laid out.

1.1 On the origin of this thesis

In 2009, the Forum Zeitgeschichte [1] of the Institute of Contemporary History of the University of Vienna initiated the project "Universität Wien im 20. Jahrhundert – Wissenschaftsgeschichte im Kontext". One of the sub-projects [25] was the "Die Informatisierung Österreichs" project [6], originally scheduled from 2009 to 2011 and conducted by Dr. Karl Fröschl, Dr. Siegfried Mattl, Dr. Johann Stockinger, Dr. Werner Kläring, and DI Wilfried Schöfer [7] in cooperation with the Austrian Society of History of Computer Science. The research focus was defined as the history of Austrian computer science and the biographies of its pioneers and protagonists. This endeavor resulted in a Wiki system, hosted by the Austrian Computer Society, and used as an open platform for the collaborative extension of the original set of articles by a community of interested individuals. A considerable part of the research was done by students as part of a lecture on the history of computer science by Dr. Fröschl.

The thesis at hand is the product of such an extension effort. It is the documentation of an attempt to integrate the Stock collection, a moderate collection of printed materials, into a hypertext system, thus preserving the documents and making it accessible to search engines and an audience far beyond the reach of the original printed media.

1.2 The Stock collection

The Stock collection was initiated by Dr. Karl F. Stock, the former head of the library of the Technical University Graz. After his retirement in 1997, Dr. Stock handed the artifacts over to the Austrian Society of History of Computer Science in Vienna, to which they still belong at the time of this writing. At the same time, he also concluded his contributions, adding only a few more items in 1998. Until 2012, the documents were stored in twelve cardboard boxes, sorted by year, and located in the Austrian State Archives. These boxes are shown in figure 1.1. As of late January 2013, the boxes were relocated to the renovated office building of the Austrian Computer Society located at Wollzeile 1 in Vienna.

Figure 1.1: The Stock collection in its original cardboard boxes.

The majority of the collected items was published between 1964 and 1998, with two exceptions dating back as far as 1952, as can be seen in figure 1.2.

The collection contains circa 1400 documents. Approximately 88% of all items are photocopies of original print media, e.g. magazine articles and book extracts. Only about 11% are actual originals, mostly books, information materials, and other forms of product advertisements. The remaining 1% is either a mix (e.g. an original information document with an attached article photocopy) or the state of the document did not allow a definite classification anymore (e.g. black and white documents).

Two noteworthy documents are included, unfortunately as copies only: hand-drawn cartoons about the challenges a librarian faced with early automation systems, signed by Dr. Stock himself and both dated 1982. They are included in the Wiki can be found via a search for “comic”.

The most common document types are:

- articles and essays
- company brochures
- reports
- information pages
- instruction sheets
- speeches and presentations
- product information

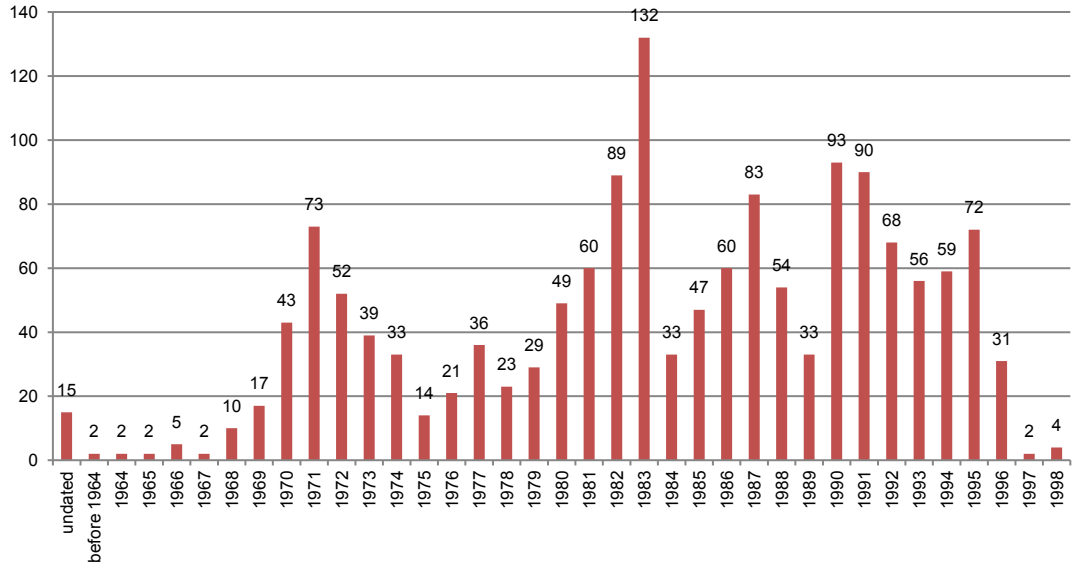


Figure 1.2: Distribution of documents by year.

Articles and essays exceed the other document types in quantity by far, as can be seen in the detailed distribution of all documents by type in figure 1.4.

Although a wide range of topics is represented, a clear trend toward library automation and a general interest in the (then) latest information technology can be observed. The recurring topics include

- automation
- catalog formats
- databases
- data representation
- exchange formats
- usability
- software development
- project reports

These trends were honored in the design of the keywords available for the semantic search in the Wiki system. A list of the keywords available at the time of this writing can be found in section 12.3.

A list of all documents currently available online can be obtained via the Wiki category page for the Stock collection at `http://www.ocg.at/informatisierung/index.php5?`

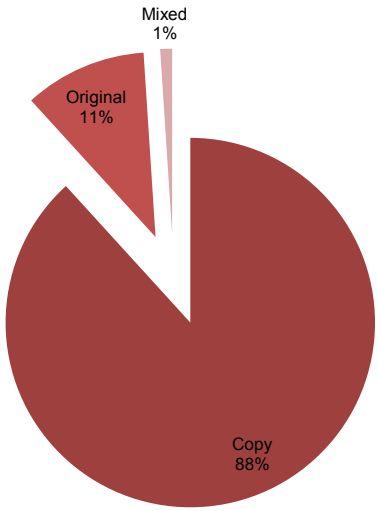


Figure 1.3:

`title=Kategorie:Sammlung_Stock` [online; accessed January 30, 2013]. A complete inventory list of all documents is located at `http://www.ocg.at/informatisierung/images/c/ce/Stock_complete.pdf` [online; accessed January 30, 2013].

1.3 Project scope and goals

The project scope is defined as the evaluation of best practices for digital archiving and representation in a hypertext system. The main goal is the development of a procedure for the

- reprocessing,
- editing,
- integration, and
- representation

of a collection of printed materials into a hypertext system. Existing best practices will be evaluated in a sample project by transferring the Stock collection from its paper form into the "Die Informatisierung Österreichs" Wiki system. The artifacts of the collection should be preserved online. The collection should then be available for further research by an interested audience, using the tools provided by the Wiki system to link information and perform semantic searches. The Wiki should also serve as a platform for the expansion and content-related processing of the digital Stock collection via school projects and lectures by providing the means for collaborative work.

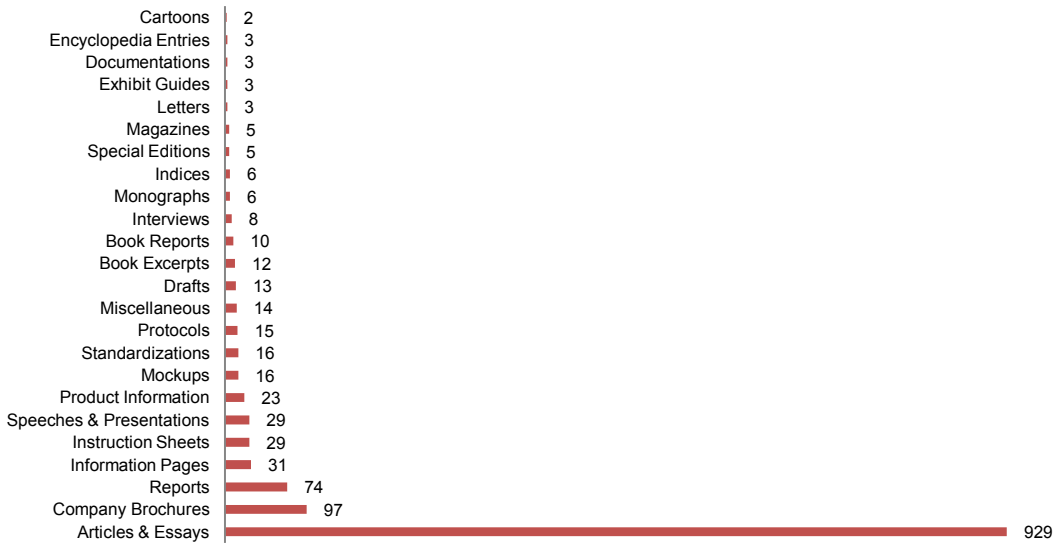


Figure 1.4:

The progress and results of this practical work are documented in this thesis and summarized in a distilled version for the use as a manual for further extensions of the Stock collection Wiki.

The in-depth analysis and content-related processing of the digitized documents is out of scope of this project. The materials are only analyzed as far as it is necessary to get a basic understanding of the content for the selection process to decide which documents form the foundation of the Wiki created by this project. An elimination of redundant materials is also out of scope, although suggestions have been made when the collection was returned to the Austrian Society of the History of Computer Science. The items proposed for recycling dealt with fundamental computer science topics which can be considered general knowledge of the audience. These artifacts were not destroyed but merely returned in an additional container to keep them separated until a decision is made.

Furthermore, legal aspects and copyright issues are out of scope. While the "Die Informatisierung Österreichs" project is still ongoing, access to the Wiki system, which also contains the Stock collection, is protected with a password, and the documents are not publicly available or indexed by public search engines. For the purpose of this project, it is assumed that no copyright laws are violated by the academic use of the materials. Suggestions for possible follow-up projects dealing with copyright issues will be made in section 10.

1.4 Roadmap

Methodology: The methodology applied in this project will be introduced to provide an understanding of the foundation of the practical work.

Hypertext theory: The transition from paper to Wiki system will begin with a short excursion to the theory of hypertext systems and hypermedia learning systems, in order to get a grasp of the greater context of the task at hand. Basic functions of Wiki and hypertext systems will be analyzed, and their use for the Stock collection transfer will be considered.

State of the art: The state of the art will be established by examining recognized best practices for digital archiving. A reference project implemented by the U.S. Library of Congress will be described, and the application of the best practices will be observed. The following steps will be performed with knowledge acquired from this analysis.

Digitization: The criteria for selecting the documents for the Wiki foundation will be defined and used during the manual sifting of the Stock collection. The selected documents will then be transferred to a digital form and stored in the Wiki.

Cataloging: Information will be extracted from the digital documents created in the previous step. This data will be used to create the Wiki articles for each item of the collection.

Presentation: The available Wiki articles will now be connected via meaningful hyperlinks to allow search and navigation in the virtual Stock collection.

Teaching projects: Suggestions will be made for a few projects using of the digital Stock collection in education and research.

CHAPTER 2

Methodology

Methodology

This chapter presents the methods that were applied in the execution of this project.

The theoretical part of the project started with literature research on the theory of hypertext systems. As the Stock collection was to be transformed into such a system, it was necessary to gain an understanding of how hypertext is different from printed media. The power of hypertext navigation poses risks of presenting the content in a confusing and misleading way. Since the digital Stock collection was meant to be used for educational projects, a navigation via meaningful links and well structured paths became a requirement.

To get an understanding for the content that was to be digitized, it was necessary to spend some time browsing through the articles. Even though an organized processing of the document content was out of scope for this project, a basic grasp of the topics covered was required for the initial selection of documents to start the digital collection. Furthermore, this knowledge was required again at a later stage when the digitized documents had to be classified with keywords.

The practical work of digitization and cataloging included several evaluation steps, during which additional information was gathered. The final results (and some of the alternatives considered during the implementation) are documented in the corresponding chapters of this thesis.

The presentation phase built on top of the know-how acquired from the theoretical research phase. The navigation structure created had to be simple to avoid confusion and prevent users from drifting off their path of interest. At the same time, it had to allow the user to reach any other point of interest within the collection with as few steps as possible.

Based on the results of the practical project, some ideas were drafted of how the digital collection can be used.

The details of each phase are documented in the following chapters.

CHAPTER 3

Hypertext theory

Hypertext theory

This chapter presents the navigation concepts on which the tours and paths for the digital Stock collection were developed. The application of these theories in the Stock project is explained as well.

3.1 Hypertext and hypermedia

A hypertext system displays content nodes that are linked via one- or two-directional navigation paths. Early versions of hypertext systems, such as HyperCard, Athena, and Intermedia, required the installation of a software client on the user terminal [23]. Current implementations can be viewed and manipulated in web browsers, which can be considered standard software on any current operating system. Navigation is performed by clicking hyperlinks. Based on Landow, the following types of hyperlinks will be of interest in the scope of this project [11]:

Hard links: Static links that always exist. These links are created by the author of the hypertext document. In the Stock project, this refers to links with a fixed target, such as the PDF file links in each document.

Soft links: Soft links do not exist from the beginning. They are only created on demand when a user clicks on a soft link. This concept translates to links to articles that do not yet exist at the time of the click. In Wiki systems, the default behavior results in the display of an empty article with the option to start editing the new entry.

Based on proposals by Kuhlen, a well-directed navigation path seemed best suited for this project. Considering the goal to serving research and education interests, guided tours and paths appeared to be the appropriate choice [10]. Furthermore, the coherence principle as described by Niegemann et al. [20] suggests reducing the amount of distracting information to a minimum and, thereby, avoiding an extraneous load of irrelevant information on the reader. Hyperlinks should, therefore, only occur where they are necessary to either explain a term in greater details or to provide further information on a certain subject. This theory is supported by the cognitive learning theory as it is described in the context of multimedia learning [20]. In addition, the use of soft links allows users to create their own collections, even for keywords that do not exist yet. Once articles are available that match the criteria, they will appear in the query result.

Building on the concepts of the cognitive learning theory, the hypertext network should be kept simple [20]. A network with too many navigation paths might be confusing for the reader.

A tree structure with one single point of entry seemed to better support a guided navigation concept than a network with every possible navigation path between documents. Unlike books and articles, hypertext does not have one single point of entry per definition, and a well-defined starting point is required, as suggested by Landow [11]. This is achieved by providing a parent article that can also be used as an external reference to reach the Stock collection.

Hyperlinks are used in a way that they depart from a certain area within a page, such as an index or table of contents. The use of hyperlinks from within the text flow of an article was avoided to prevent the creation of confusing network structures. The Stock collection parent page serves as a container for these hyperlink collections.

According to Schulmeister, an objective separation of hypertext and hypermedia cannot be provided, as the criteria to distinguish the two are heavily influenced by previous education and experience of the observer [23]. While some experts argue that a hypermedia system is merely a hypertext system that also displays images, others argue that the ability to display images alone does not suffice for classification of a framework as a hypermedia system. For the purpose of this thesis, the term "hypertext" will be used, even though the Wiki system used in this project is capable of displaying various media formats. Hypermedia in this context is regarded as a tool to unlock the knowledge that can be extracted by combining the information from multiple sources. In the case of the Stock project, the sources are represented by the documents of the collection.

3.2 Wiki systems

A Wiki is a special kind of hypertext system. Besides fulfilling the characteristics of a hypermedia system by supporting various media formats in addition to text, they also provide tools for collaborative work and interaction among users. As with popular websites like Wikipedia, the Stock project is an open presentation platform and also provides the reader with the power to become the editor of the consumed content. According to Niegemann et al. [20], Wikis provide the following capabilities that likewise support the needs of the Stock project:

Collaborative production of information sources: Each reader is allowed to edit the consumed content. If an error is found (e.g. in the text extracted with OCR), the user can fix it right away in the live system.

The product of the collective work is the center of attention: The final digital version of the Stock collection will be the product of a team effort. People will have to work together, and every participant will be able to see the progress on the project by viewing at the articles.

Wikis are popular and well-known: Even though a proprietary syntax has to be learned for the article markup, Wikis are very popular, and there is a chance that project participants already have previous experience with a Wiki system.

A discussion page for self-organization: The Wiki discussion page is available for each entry. Niegemann suggests that this public forum provides the tools required for self-organization in unorganized teams. This might be especially helpful for research teams or

teams of students that are not working in the same time zone. Even though other forms of communications such as E-Mail are available, they are generally separated from the subject matter. The discussion page directly links organizational information with the article it refers to.

No installation required: Wikis can be accessed with a web browser. Users do not need to install software on their local workstations to take advantage of the functionality of the Wiki system.

User is not concerned with technical details: Wiki users do not need to occupy themselves with technical details such as software upgrades. They can focus all their attention on the Wiki content.

Moreover, Schulmeister suggests that a large part of learning is accomplished in the form of dialog [23]. This is another argument in favor of using the discussion page as a communication platform. Furthermore, Tietgens (as referenced in [24]) suggests that in adult education, information should be available when it is needed. The Wiki system provides full text search for all articles out of the box, and the Stock Wiki offers semantic search in addition.

Finally, to maintain a closed platform without dependencies, Niegemann suggests that a Wiki system should not contain links to external resources [20]. As described in section 7.2, it was not possible to follow this recommendation because of the file size upload limit in effect on the "Die Informatisierung Österreichs" Wiki server. Some of the scanned PDF files exceeded the maximum file size of 8 MB and had to be stored on another server. Other than that, all resources and articles are hosted only within the domain of the Stock Wiki.

CHAPTER 4

State of the art

State of the art

This chapter provides an introduction to the technical background of this project.

The National Digital Library Program and the American Memory Project of the U.S. Library of Congress are introduced as reference projects. Recommendations based on best practices and standards for multiple aspects of digital archiving are presented.

Since the Stock collection consists only of printed materials, the focus is the preservation of text and printed documents. Considering the limited resources of the Stock project (when compared to the Library of Congress), resources for personal and home archiving will also be taken into consideration.

The aspects considered in this chapter are based on the "Technical Guidelines for Digitizing Cultural Heritage Materials" [5].

4.1 The American Memory Project

Work on the National Digital Library Program [15] began in 1989 with a survey of research and state library organizations. The survey results showed a strong interest in online libraries, mostly from research institutions and schools. Especially secondary level schools showed a high demand for primary source materials to encourage critical thinking, a change in the educational system based on reforms in previous years.

The Library of Congress then started a pilot project, which ran from 1990 to 1994, to experiment with the digitization of historical sources. The most valuable artifacts of "written and spoken words, sound recordings, still and moving images, prints, maps, and sheet music that document the American experience" [12] were digitized during this period to establish the foundation of the National Digital Library. The first collections were distributed on CD-ROM to selected libraries all over the country, but CD-ROMs soon proved to be too expensive, and by 1994, the collections were made available online.

After receiving 13 million USD in donations from the private sector, the Library of Congress launched the American Memory Project in 1994. The project was funded until the year 2000 by 15 million USD from the U.S. Congress and an additional 45 million USD from private donors [12].

In order to provide valuable means for research and educational use, the National Digital Library had to represent more than the sum of merged library catalogs. The Learning Page [17], published in 1996, remains a helpful tool for browsing collections based on different aspects

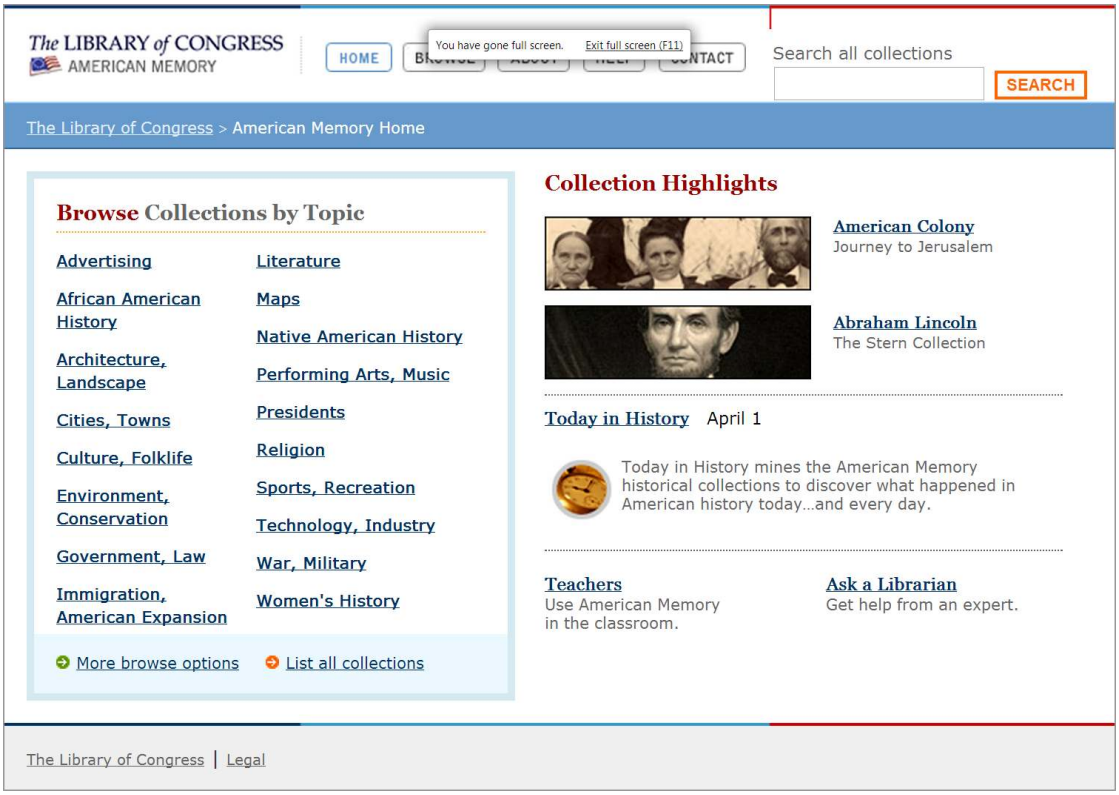


Figure 4.1: American Memory start page

such as the original source or thematic associations. Sample lesson plans, activities and other classroom materials are provided as tools for teachers.

By the year 2000, the American Memory Project [13] contained about 5 million items. It is now an ongoing project which is freely accessible on the Internet.

4.2 Collection Policy

Selection guidelines should be established to define the scope of the project. The scope will vary depending on the purpose of the project, as mentioned in [4]: should the original materials be replaced by digital versions, or should the digital collection be used to enhance access? If historic materials are digitized, conservation treatment might be necessary as well. Digitization must be performed in a way that does not damage the original materials.

4.3 File Format Recommendations

The most important aspect to consider when choosing a file format is the desired quality of the master copy. Future processing techniques might be able to process larger amounts of data

and higher image resolutions. Therefore, for images, photos, and maps, the master copy should be of the highest quality obtainable with current technology. For the purpose of file format selection, historic hand-written documents should be treated as images, if the value lies in the document itself and not so much in the text it contains. Further considerations for printed, text-only documents will be discussed in section 4.4.

While high master image quality is desirable for archiving, the large file size could pose a problem for storage and transfer. Projects with an online presentation like the American Memory Project and the Stock project depend on a compact file size for file transfer over the internet. Some formats do provide a small file size by the means of lossy compression. These files should then be used for representation and viewing only, but the high quality master copy should be kept as primary source.

The following formats are recommended by the Federal Agencies Digitization Guidelines Initiative [5]:

Examples for lossless formats

- TIFF: The Tagged Image File Format is currently the "preferred format for production master files" [5]. It stores uncompressed and lossless compressed images, supports multiple color spaces, and can store metadata within the file. Because of its complex internal structure, it is not useful for streaming images. Images are stored as raster data, and the file size tends to be large. The format is not natively supported by current web browsers.
- JPEG 2000: This format contains additional compression methods compared to the original JPEG format, including lossless compression. It is a rather new and emerging format, and its extended version supports color profiles and layers. Images are stored as encoded data and not as raster data.
- PNG: The Portable Network Graphics format is another lossless format. It is not very common, but as opposed to TIFF, it is natively supported by latest web browser generations. Images are stored in a simple raster format, and alpha channels are supported.

Examples for lossy formats

- JFIF / JPEG: The original Joint Photographic Experts Group format offers a small file size through lossy compression. This results in deteriorating image quality after multiple saving procedures, rendering it useless as a format for master files; it is, therefore, used for representation purposes only. Compared to GIF, JPEG requires more time for decompression.
- GIF: The Graphics Interchange Format offers lossless and lossy compression with a limited color palette and a short decompression time.

Container Formats

- PDF: The Portable Document Format can store raster images in their original format. It serves as a container for multiple objects that should stay combined in one file. The

format supports compression for the whole file or only sections of it and provides a limited number of color spaces.

- PDF/A: A subset of the original PDF, this format includes additional features for the long term archiving of documents that were already created in a digital form (e.g. including used fonts in the file); it, therefore, does not provide an advantage over the original PDF for the purpose of archiving digitized materials.

An additional aspect to consider is the longevity of the selected format.

Proprietary formats are often kept a trade secret and have a stronger dependency on certain software tools to edit and view the files. If the company that owns the format runs out of business or if they decide to cease software support, a (real or virtual) emulation environment is required to process the file format in the future. As long as the hardware architecture does not change too much, the emulation of an older operating system version could be done using virtualization tools such as VirtualBox or VMware. If a completely different hardware platform needs to be emulated, it might be necessary to transform the master copies to another format, e.g. scanning microform documents.

While selecting an open standard format does not guarantee that the format will be around forever, the documentation for open standards is usually available publicly. Should the need arise, a new tool for processing and viewing the format could be implemented based on the format specifications.

Issues to consider for longevity:

- Documentation: the format documentation should be freely available
- Stability and continuity: the format specifications should be mature and not change on a regular basis
- Metadata: metadata should be stored together with the image data to allow future identification
- Complexity: simple file formats should be preferred to allow data extraction if necessary, which might be prevented by complex compression or storage algorithms
- Interoperability: formats that are already supported by multiple software tools should be preferred

4.4 Text Processing

If the value of a document lies in the text content more than in the actual document (as was the case for many items of the Stock collection), the document could be considered for optical character recognition (OCR). Other than with images, lifelike color reproduction is not an issue when the scanned document is processed with OCR tools. In this case, the main criteria are the dots per inch (DPI) of the scanned file. The OCR results vary depending on the quality of the original document. The Library of Congress recommends a standard resolution of 400 DPI for

scanning printed text that will be processed with OCR and 300 DPI for text that is simply stored as an image [18].

4.5 Metadata

Metadata provides the information required for document management and long term preservation. A digital document cannot be considered complete without its associated metadata. The following types of metadata are distinguished in the Technical Guidelines for Digitizing Cultural Heritage Materials [5] and by [4]:

descriptive metadata: The who, what, when, and where of a resource. This metadata type includes bibliographic and physical information such as the media type, dimension, and condition, and it also describes the content. Search algorithms use this information to find documents.

technical / structural metadata: Provides information about the digital document to applications to support correct rendering of the file. The data is used to present complex objects through the representation of relationships between them, e.g. the location of a specific page (as identified on the the original paper document) in a sequence of digital images.

administrative metadata: Information used for the support of internal management of a file.

rights metadata: Data that holds information regarding the rights owner of a resource as well as access rights to the resource.

Metadata extraction, as far as descriptive information concerning the original analog document is concerned, is still a time consuming manual process, but there is a growing interest in automatic solutions, e.g. import from other catalogs if available. It is common practice to create the metadata during the cataloging phase, as described by [8]. Metadata for the American Memory Project was assigned either during the scanning phase or in post processing [18].

There is no single standard format for the storage of metadata, but many organizations use structures similar to the Dublin Core [9] format. The Library of Congress uses MARC [21], as do some other organizations, to define and store their metadata.

Document metadata is often stored separately from document content. This allows organizations to keep only one data set that can be accessed by multiple tools such as online catalogs and citation databases. Content and metadata are usually stored together only for journals. The American Memory Project stores metadata in TIFF or EXIF headers. For this purpose, a standard set of required information was defined; see figure 4.2 for an extract from [18] showing the standard metadata set as used in the American Memory Project. Additionally, a minimum set was defined. This set contains the merged information from the tags from figure 4.2 and tag information from the National Digital Newspaper Project [22].

4.6 Storage

The system of choice for storing archive data are hard drives due to their reliability and storage capacity. Multiple drives can be combined into a RAID (redundant array of independent disks) system where each drive holds the complete data set; if one drive fails, the data is still stored on the other drive. Current file system implementations also allow storing very large files up to 256 TB [19]. Metadata should be stored together with the archive data for future file management and to support search and classification in the archive.

In addition to the hard drives, backup copies should be created on a periodic schedule. At least one backup copy should be stored in another physical location to prevent data loss if one physical data storage location is destroyed, e.g. by a natural disaster. Backups should also be created on tape. As with file formats, the use of an open standard tape format such as LTO (Linear Tape Open) is recommended. Checksums should be created and stored together with the backup copies to check the data integrity over a longer period of time.

The use of optical storage devices is not recommended. The materials used to create CD-Rs tend to be too unstable for long-term archiving. If CD-Rs have to be used, special archival quality CD-Rs are recommended. Their materials are more stable and durable. A cyclic redundancy checksum (CRC) should be created for each disc to check the integrity.

DVDs and Blu-ray Discs tend to have similar production shortcomings as CD-Rs. Furthermore, the formats for these larger storage discs are not as standardized as for CD-Rs. While CD-Rs work with well-established formats, some DVD and Blu-ray formats might be discontinued and become obsolete in the near future. As pointed out in Personal Archiving recommendations of the Library of Congress [16], old drives and software versions should be kept for accessing old storage media.

As storage technology and operating systems change, it is necessary to migrate archive data from one system to the next. According to [8], most organizations expect data migration every 3 to 5 years. Depending on the amount of data, the migration process might require 6 to 12 months and might even turn into a continuous process. If archive data is stored on hard drives over a longer period of time, the bits should be refreshed as suggested in [4] by simply copying the content from one physical location to another.

4.7 Quality Assurance

Quality assurance describes the measures taken to ensure the quality and consistency of the digitized goods. These measures should follow a defined process that ought to be designed at the beginning of the project. All results, e.g. of visual evaluations, need to be documented in standardized forms and reports. External contractors should be required to run a set of quality checks before delivery. The logged information has to be stored together with the digital collection for future reference, e.g. as metadata on file or project level as suggested in [5].

It is recommended to perform quality control as a two-step process. For large-scale projects that use contractors for the scanning work, the initial check could already be done by the scanning technician, as suggested in the Technical Guidelines for Digitizing Cultural Heritage Materials [5]. Alternatively, the first step could also be done by an automated software solution,

as done by the Library of Congress [18]. An automated software solution allows the first check to be faster and more detailed. It could include checks such as the correctness of the used file format, the completeness of metadata, and adherence to file naming conventions. In addition to the automated check, the Library of Congress also requires the scanning contractors to perform quality control before delivery, although the extent of these checks is not specified in detail. The second step needs to be a manual check by another person (if the first check was done by a human). All digital deliveries should be checked to make sure that the expected number of files was actually included.

Digital image files also require some form of visual inspection. At least 10 files or 10% of the delivered images (whatever is the larger quantity) ought to be subjected to a visual check. The following aspects, based on the inspection list for digital image files described in [5], should be considered for quality control:

Checks concerning the file

- open the file to check for read errors
- ensure that the digital image is stored in the desired format
- check the file compression and encoding
- make sure that the desired color mode (RGB, gray scale, bi-tonal) and according bit-depth was used and that the color profile was stored

Checks concerning the source document:

- check the resolution and dimensions of the image
- make sure that the proportions are correct, and check for distortions and skew
- check if the orientation is correct
- the image must not be cropped; the original content must be complete

Checks concerning the metadata:

- the data in the header tags must be complete and correct
- the descriptive, technical, and administrative metadata must be complete and correct

Checks concerning the image quality:

- the image color, brightness, and contrast must reflect the original image
- the overall noise and artifacts must be tolerable
- detail must not be obscured

The Technical Guidelines suggest that all images contained in a delivery are checked if more than 1% of the test set fails quality control. If the images contain text that will be extracted with OCR tools, the Library of Congress requires 90% word accuracy.

Tag	Name	Description	Tag Type
256	ImageWidth	The number of pixels per row	Baseline Tag
257	ImageLength	The number of rows of pixels in the image	Baseline Tag
258	BitsPerSample	Number of bits per component	Baseline Tag
259	Compression	Compression scheme used on image data	Baseline Tag
262	PhotometricInterpretation	The color space of the image data	Baseline Tag
273	StripOffsets	For each strip, the byte offset of that strip	Baseline Tag
277	SamplesPerPixel	The number of components per pixel	Baseline Tag
278	RowsPerStrip	The number of rows per strip	Baseline Tag
279	StripByteCount	For each strip, the number of bytes in the strip after compression	Baseline Tag
269	DocumentName	Document Name (path/filename)	Extension Tag
282	Xresolution	Horizontal pixel count per resolution unit (inches, centimeters)	Baseline Tag
283	Yresolution	Vertical pixel count per resolution unit (inches, centimeters)	Baseline Tag
296	ResolutionUnit	Unit of measurement for X and Y Resolution (inches, centimeters	Baseline Tag
306	DateTime	Date and Time image was scanned	Baseline Tag
315	Artist	Person who created image (default LoC)	Baseline Tag

Figure 4.2: Library of Congress: standard metadata for TIFF tags

CHAPTER 5

Digitization

Digitization

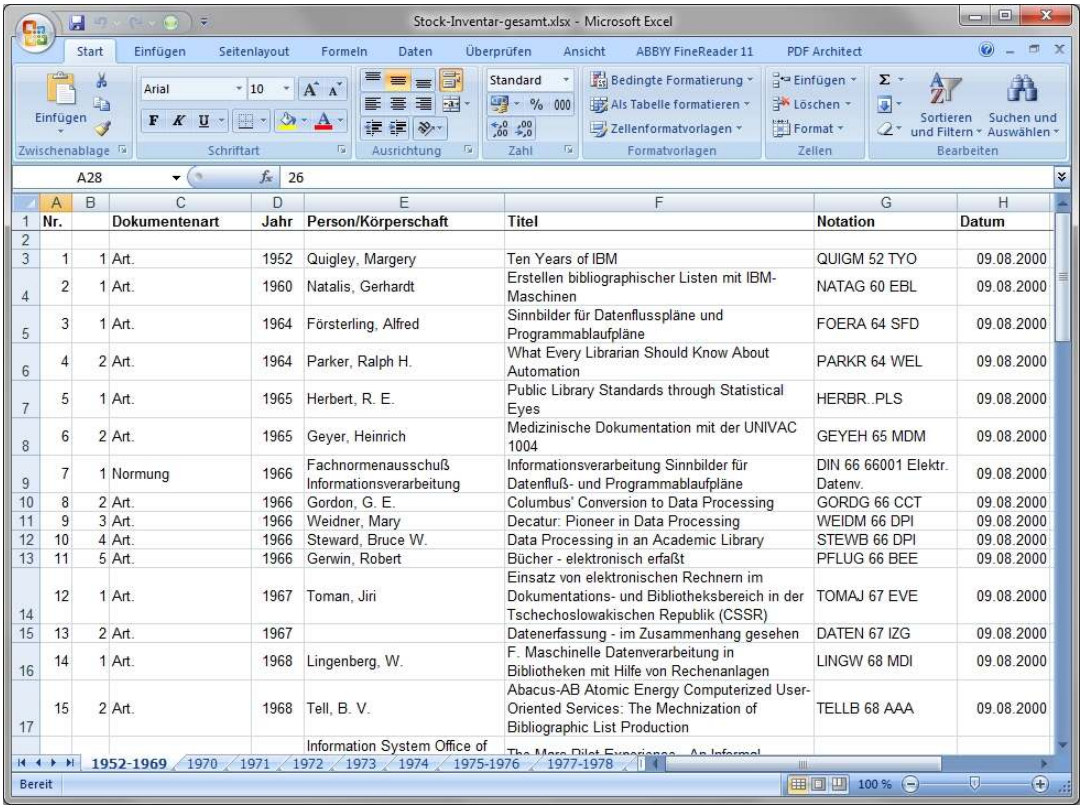
This chapter describes the process of digitization of the Stock collection. The initial state in which the collection was received for this project is described, as well as the criteria of the document selection process. The used hardware and software equipment is documented together with the scanner settings.

5.1 Initial state of the Stock collection

The Stock collection was originally stored in the office building of the Austrian Computer Society, located at Wollzeile 1 in the first district of Vienna, Austria. During a moving period, the collection was transfered to the Austrian State Archives. As part of this transition, the documents were put into twelve cardboard storage boxes provided by the State Archives. At the beginning of the project, the boxes were picked up from the State Archives and transferred to the home of the author of this paper; all further processing would take place there.

Eleven of the twelve boxes were labeled with a large white sticker and an identification number, as can be seen in figure 1.1. One box was mislabeled and did not possess any form of identification. The only available index of the Stock collection at that time were lists created during an internship at the Austrian Society of History of Computer Science. These index lists were stored in Microsoft Excel sheets and contained the most important administrative information, such as the author, document title, and year of publishing. However, there was no reference to the boxes or how to locate a certain document in the collection. As can be seen in figure 5.1, the lists were organized in tabs by publishing year. It turned out that the boxes contained sets of publishing years, but again, there was no clear reference. Only after opening a box and determining the publishing year of a few sample documents was a rough estimation of the content possible.

Even though the Excel lists did now allow a quick localization of a certain document in the collection, the data provided by these lists proved to be very reliable; all documents selected and scanned during the following steps were compared to the data in these lists and, with the exception of a few minor spelling errors, the information was valid. These lists would later serve as a source for metadata in the cataloging stage of the project.



Stock-Inventar-gesamt.xlsx - Microsoft Excel

Nr.		Dokumentenart	Jahr	Person/Körperschaft	Titel	Notation	Datum
1	1	Art.	1952	Quigley, Margery	Ten Years of IBM	QUIGM 52 TYO	09.08.2000
2	1	Art.	1960	Natalis, Gerhardt	Erstellen bibliographischer Listen mit IBM-Maschinen	NATAG 60 EBL	09.08.2000
3	1	Art.	1964	Försterling, Alfred	Sinnbilder für Datenflusspläne und Programmablaufpläne	FOERA 64 SFD	09.08.2000
4	2	Art.	1964	Parker, Ralph H.	What Every Librarian Should Know About Automation	PARKR 64 WEL	09.08.2000
5	1	Art.	1965	Herbert, R. E.	Public Library Standards through Statistical Eyes	HERBR..PLS	09.08.2000
6	2	Art.	1965	Geyer, Heinrich	Medizinische Dokumentation mit der UNIVAC 1004	GEYEH 65 MDM	09.08.2000
7	1	Normung	1966	Fachnormenausschuß Informationsverarbeitung	Informationsverarbeitung Sinnbilder für Datenfluß- und Programmablaufpläne	DIN 66 66001 Elektr. Datenv.	09.08.2000
8	2	Art.	1966	Gordon, G. E.	Columbus' Conversion to Data Processing	GORDG 66 CCT	09.08.2000
9	3	Art.	1966	Weidner, Mary	Decatur: Pioneer in Data Processing	WEIDM 66 DPI	09.08.2000
10	4	Art.	1966	Steward, Bruce W.	Data Processing in an Academic Library	STEWB 66 DPI	09.08.2000
11	5	Art.	1966	Gerwin, Robert	Bücher - elektronisch erfaßt	PFLUG 66 BEE	09.08.2000
12	1	Art.	1967	Toman, Jiri	Einsatz von elektronischen Rechnern im Dokumentations- und Bibliotheksbereich in der Tschechoslowakischen Republik (CSSR)	TOMAJ 67 EVE	09.08.2000
13	2	Art.	1967		Datenerfassung - im Zusammenhang gesehen	DATEN 67 IZG	09.08.2000
14	1	Art.	1968	Lingenberg, W.	F. Maschinelle Datenverarbeitung in Bibliotheken mit Hilfe von Rechenanlagen	LINGW 68 MDI	09.08.2000
15	2	Art.	1968	Tell, B. V.	Abacus-AB Atomic Energy Computerized User-Oriented Services: The Mechnization of Bibliographic List Production	TELLB 68 AAA	09.08.2000
16				Information System Office of			

Figure 5.1: The Stock collection inventory list.

5.2 Criteria for document selection

One of the project goals was to establish an online document foundation that can be extended by follow-up projects. The thematic criteria listed below were defined to select documents for the first digitization wave. Since the Stock collection covers a wide range of topics, some topics were neglected to favor the forming of larger thematic blocks. These blocks will be used for the demonstration of theme-based tours later. The following topics are, therefore, listed in the order of their importance for the selection process; topics on top were preferred since they represent the core of the collection, while the topics near the end of the list can be considered a thematic extension.

- library automation processes and concepts
- reports, requirements, and reviews from automation projects
- library support systems, e.g. user terminals
- library catalog formats, e.g. MARC
- other library-related hardware systems

- documentation for library systems
- user interfaces for library systems
- general data storage and database topics
- general information technology topics and their application in libraries

The following topics were explicitly not considered for digitization:

- articles (mostly from magazines) that give an introduction to fundamental information technology topics, e.g. binary numeral system
- outdated documentation when a current version of the product still exists, e.g. manuals for outdated software versions

No documents were destroyed during this project, but articles that matched the criteria above were sorted out and stored in a separate container. While most artifacts of the Stock collection possess a historic value in terms that they are a snapshot of a moment of the the historical development of information technology, there seems to be little value in spending resources on making documents available online which contain knowledge that is still state of the art, especially when that information is already easily available online. Of the ca. 1400 documents of the Stock collection, less than 10 artifacts fit these criteria for redundancy; only one of them was an original document (an approximately 10 year old manual of the Dreamweaver web design software, which still exists in an updated version at the time of this writing).

As a general rule, original documents were preferred over copies. However, the number of original documents in the Stock collection is relatively small, as can be seen in figure 1.3. Most original documents were advertisements and product information for library systems from hardware supply companies. As part of the preservation of the digital heritage, a special focus was placed on the conservation of old "predictions of the future", e.g. forecasts of how new technologies will perform.

Beyond the thematic scope, the following criteria were defined for document quality and volume:

- the document must be readable and not faded
- the document must not be damaged in a way that prevents processing with OCR tools, e.g. large chunks of paper must not be missing
- it must be reasonable to scan the document by hand, e.g. books and documents with more than 20 pages were left out
- it must be possible to fit the document on the scanner table

5.3 Document selection

Based on the criteria above, the selection process was performed in multiple phases. The selection criteria were extended and refined with each phase as the understanding of the Stock collection grew.

The first phase was a general sifting of all artifacts to get an initial glimpse of the size and volume of the collection, and the general quality of the documents was assessed. Documents that clearly and without doubt did not match the requirements of the project were removed. All artifacts that either matched the criteria or documents in doubt were kept for the refinement phases.

During the second phase, the selected documents where split into original documents and copies. Original documents that matched the selection criteria were then immediately added to the pile for documents that should be scanned. The remaining documents were then checked for copies which matched the criteria, as well as original documents that might not completely match the criteria, but possessed a value great enough to justify adding them to the document foundation. (Examples for documents that were added merely for their historical value are a few cartoons about daily life in a library by Dr. Stock himself.)

The last phase was a final check of all sorted out artifacts to search for documents that might have slipped through in a previous phase. In the end, all documents that were not scanned were placed back in the archive boxes in their original order. Documents that were scanned were moved to another box, also in the correct chronological order, to keep them separated from the non-digitized artifacts. In the end, 140 documents out of ca. 1400 were selected for the foundation for Stock collection Wiki.

5.4 Hardware and software equipment

To evaluate the guidelines and best practices described in chapter 4, the first scanner tests were performed with the already available equipment: a Canon PIXMA MG 5350. This machine is a printer / scanner hybrid device. After a few scans, it was obvious that the scanner cover was too fragile to withstand the constant strain involved in scanning 140 documents with up to 20 pages per document. A more suitable scanner, a Canon CanoScan LiDE 210, was obtained for ca. 80 EUR as a replacement. The CanoScan LiDE 210 also offered a higher maximum resolution of 4800 x 4800 DPI, as opposed to the 2400 x 4800 DPI of the PIXMA MG 5350. Besides that, the CanoScan has some programmable hardware buttons that allow the mapping of certain scanner settings to a single button, a feature that proved to be highly useful as it allowed a more efficient operation of the scanner by limiting the scanning work to the scanner only, instead of dividing the work between the scanning device and a PC.

Due to the previous document selection and sorting process, a large work area was necessary to accommodate the different piles and boxes used to keep the documents in order. As a consequence, a laptop computer was used for the scanning operation. The following hardware specifications proved to be more than sufficient for the task at hand:

- Intel Core i5-2410 @ 2.30 GHz

- 4 GB RAM
- 64 GB SSD drive
- Windows 7 64-bit

5.5 Scanner settings

The software used for scanning was Canon MP Navigator EX 4.0, which was shipped with the CanoScan LiDE 210. The software allowed scanning via the software interface or direct control of the scanner driver settings. The wide range of color and tone related driver settings did not produce a visible improvement of the quality. On the other hand, accessing the driver settings prevented the scanner hardware buttons from working. Having to choose one or the other, the choice fell on the use of the hardware buttons, as this promised to speed up the scanning process by handling the scanner only and being mostly independent of the computer, especially when handling multi-page documents.

Because almost all documents consisted of more than one page, the decision was made to use PDF as file format, as it provides easier access for multi-page documents than TIFF. A further reason for PDF was the planned integration in a Wiki system: Media Wiki and also most current browsers provide support for PDF, whereas plug-ins are necessary to display TIFF files. The downside of this decision was that PDF does not provide the metadata flexibility of TIFF; however, since the metadata storage is handled by the Wiki system, where the metadata is also used for the semantic search functionality, this trade-off was considered acceptable.

Test scans with different settings were performed based on the recommendations from chapter 4. The average results for some sample A4 pages taken from the Stock collection are listed in figure 5.2. Despite the recommendation to use at least 400 DPI for document scans if OCR processing is planned, the file sizes for 400 DPI and 600 DPI proved to be too large for the intended use. The existing Wiki system has a maximum upload size of 8 MB, which is the default value for a vanilla Media Wiki system. Although there is no technical limitation to the upload limit other than available hard drive space and connection speed, the average document file size would have been larger than 8 MB, thus potentially limiting future migration to another Wiki system or access for users with slower internet connection. As a trade-off, the scan quality was set to 300 DPI. Only 18 of the 140 scanned document still did not meet the 8 MB file size limit because they contained too many pages. Since only printed documents were contained (as opposed to handwritten documents), the first OCR test results showed a very high rate of character recognition, making 300 DPI an acceptable choice for the purpose of this project. As a side effect of selecting 300 DPI instead of a higher quality, the scan time was considerably reduced to an average of 10 to 15 seconds per page (see figure 5.2).

The final scanner settings are shown in figure 5.3. The document type was set to document. With that setting, the scan software automatically stores recognized text in the PDF file instead of saving the file as image only. However, there is no way to edit or even review the OCR results during the scanning process, therefore additional and reliable text extraction is still required later.

quality	300 DPI	400 DPI	600 DPI
size per page	< 1 MB	ca. 4 - 8 MB	ca. 8 - 12 MB
scan time	10 - 15 s	30 - 40 s	ca. 60 s

Figure 5.2: Comparison of scan performance.

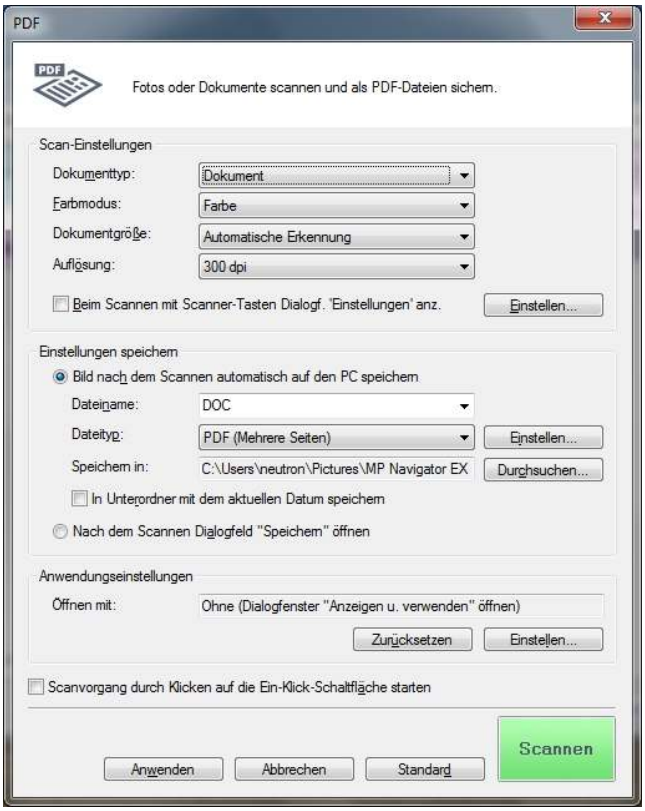


Figure 5.3: Scanner settings for Canon CanoScan LiDE 210 in Canon MP Navigator EX 4.0.

5.6 Scanning process

Dr. Stock used staples to keep all pages of a copied document together. For most items, he also added a cover page (see figure 5.4 for an example including a identification code assigned by Dr. Stock). In order to scan the artifacts, the staples had to be removed, which meant potential damage to the collection items. After discussing the issue with the Austrian Society of History of Computer Science, the current owner of the Stock collection, it was decided that removing the staples for the scanning process would be acceptable; this also meant a great improvement of the quality of the scanned versions (compared to folding the documents to fit the pages on the scanner). After each document was scanned, staples were refastened on the pages, as close as possible to the original holes, in order to restore the original state.

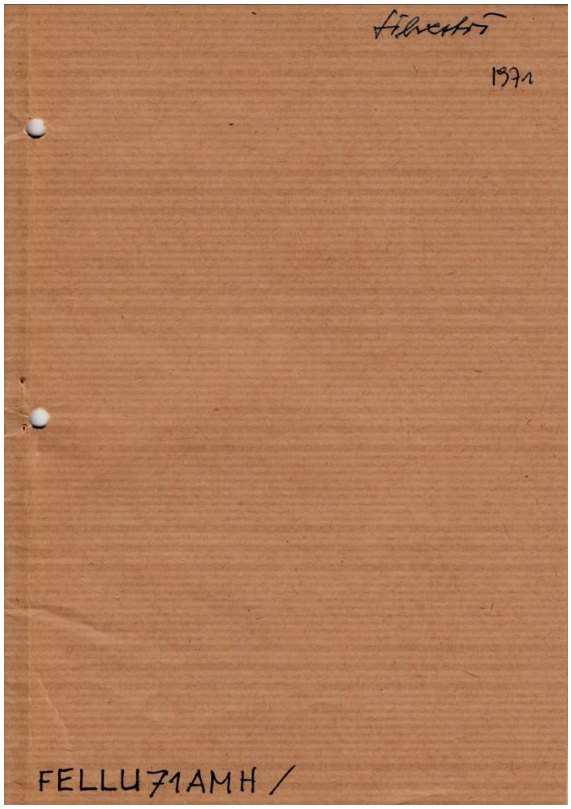


Figure 5.4: A sample cover page.

During the scanning procedure, the digital documents were automatically saved on the target computer in a folder labelled with the current date. File naming was also carried out by the scanning software, which automatically increased a counter attached to the default file name. At the end of the digitization phase, all documents were renamed manually. In preparation of the cataloging phase, all file names were changed to a unique name based on the document name, the publishing year, and the identification code (if available) as assigned by Dr. Stock on the cover page. Not all items were marked with such a code; it was not possible during this project to determine the system behind this code or whether there is one at all, but all codes seemed to be unique.

5.7 Quality control

Although a two-step process was suggested in chapter 4.7 for digitization quality control, it was not possible to perform this due to the limited resources of the project. Instead, the person scanning the documents had to check the quality of the document immediately after the scan. Unsatisfactory scans simply had to be redone.

The quality standards for digitization during this project were relatively low:

- the PDF document had to preserve the original look and feel, including the cover page
- image quality had to be high enough for OCR processing during the cataloging phase
- the document had to be aligned correctly
- the document orientation had to be preserved
- the document color and tone had to represent the original document

As a flatbed scanner was used for scanning, it was rather easy to maintain the correct document alignment and orientation. Color results were also stable and satisfactory, and only very few scans had to be repeated.

5.8 Results of the digitization phase

At the end of this phase, the following milestones were achieved:

- a set of documents was selected for digitization
- the paper documents were converted to a digital form
- the digital files were named after the document they contained

CHAPTER 6

Cataloging

Cataloging

The following chapter describes the handling of the digitized files and how they are prepared for online presentation. This procedure is called cataloging in reference projects, as it is the integration of the digital artifacts into the library catalog. For this step, it is necessary to extract content data for later use in the Wiki article, as well as metadata for administration of the online collection. Accordingly, this chapter includes an evaluation of OCR tools for automated content extraction. For the purpose of thematic separation, this chapter will end with a local catalog of the Stock collection, consisting of the digital documents, the extracted text and images, and the associated metadata for each artifact.

6.1 Evaluation of OCR tools

The documentation used in chapter 4 to establish the state of the art did not contain definite information about the used OCR tools. To get an overview of the existing software solutions, the social software recommendations site Alternativeto.net [3] was consolidated. It turned out that relatively few end-user solutions exist, and many of them are web-based services. To avoid being dependent on the availability of a third party service for an essential step of the process, web service solutions were left out. Of the remaining free and commercial applications, the following were selected for a closer evaluation.

Adobe Reader

The free Adobe Reader for PDF files allows text selection, if text was stored in the PDF when the file was created. As described in section 5.5, the PDF document type was set to document during the scan process, thus making use of this feature possible. Th Adobe Reader version used for this evaluation was 10.1.6. The results proved to be surprisingly accurate and even surpassed the character recognition of the other tested solutions for some special characters (e.g. list bullets).

However, text extraction was only possible by selecting the text in the PDF and manually copying it to a text file. There was no way to automate this procedure. Direct image extraction also was not possible. The only available workaround taking a snapshot of the PDF (a function provided by Adobe Reader) and then manipulating the image in another software tool.

Figure 6.1 shows that the user interface is not designed for OCR processing. There is no integrated editor for manipulation of the recognized text. In addition, as this solution heavily relies on high quality OCR during the initial scan, it cannot be recommended. The lack of an automated solution also prevents this method from being used for a large amount of files.

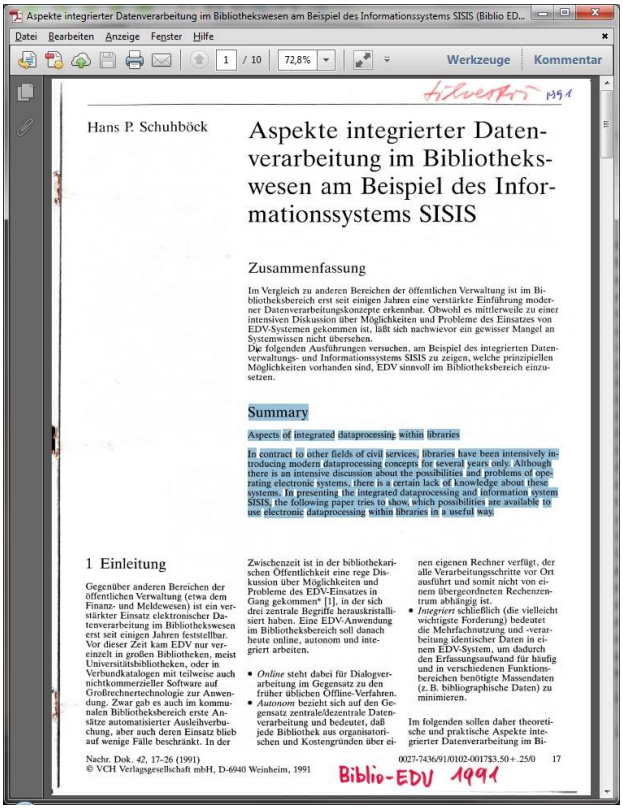


Figure 6.1: OCR via text selection in Adobe Reader.

FreeOCR / Tesseract

FreeOCR builds on top of the open source Tesseract OCR engine. It provides a graphical user interface but no option for automation. (The Tesseract engine is also available separately, but it not considered for a separate evaluation since it is included in FreeOCR.) As can be seen in figure 6.2, the user interface provides an editor for corrections on the extracted text.

The OCR results could be improved significantly by setting the correct document language, e.g. special characters in a German paragraph might not be recognized at all if the document language is set to English. This solution already provides more options and a more suitable user interface than the Adobe Reader. It could be a valuable tool if a tight budget does not permit the purchase of a more sophisticated commercial solution. The tested version, FreeOCR 4.2, did not offer an interface for an automated solution, but the Tesseract engine alone could be used in a script for batch processing. Neither the Tesseract engine nor FreeOCR provide the means for image extraction. Text selection is only possible on a per page base, which will also include header and footer texts that might not be of interest. The OCR performance was slow compared to FineReader.

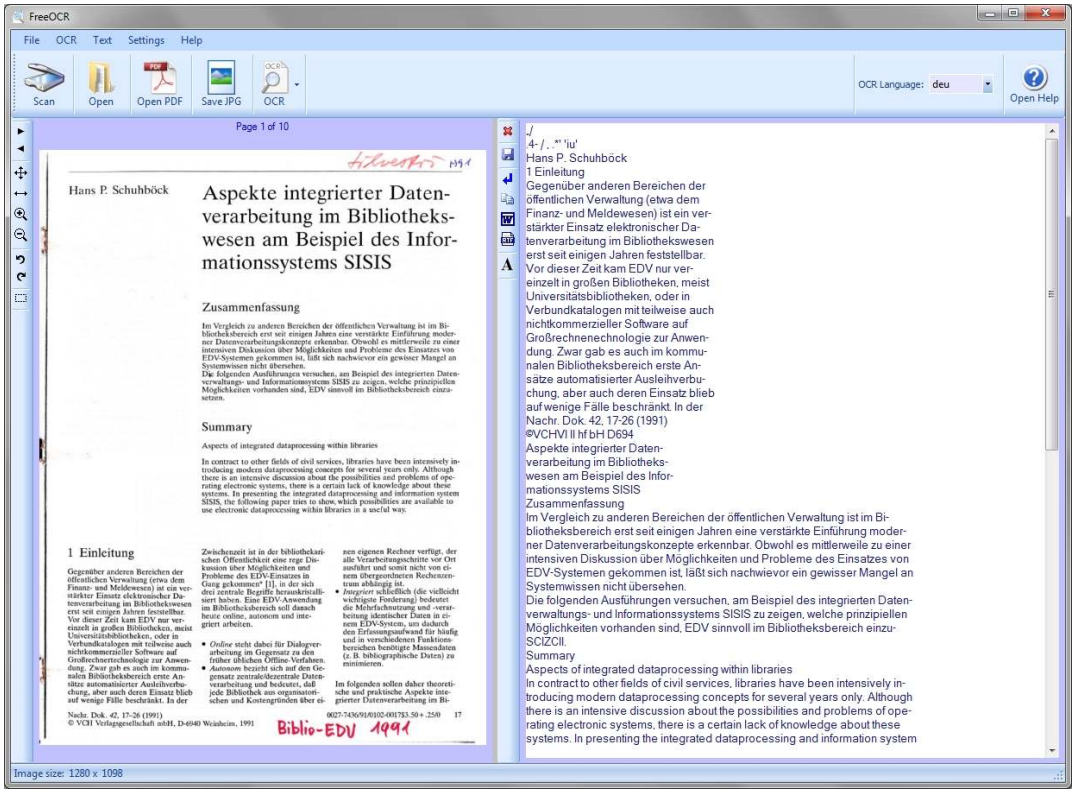


Figure 6.2: OCR with FreeOCR 4.2.

Abbyy FineReader

The third evaluated OCR solution was the commercial software Abbyy FineReader; the tested trial version was FineReader 11. As can be seen in figure 6.3, this tool provides a user interface similar to FreeOCR. In addition, it is possible to select areas in the document view on the left side. The selected area can be set either to text or image, and it is thereby possible to extract images; this is the only evaluated solution that supports image extraction. Certain page areas, such header and footer, can be ignored by reducing the selection area. The OCR results, including images, are displayed on the right side. Paragraphs are split by line separators. If text is edited, the associated selection box is highlighted. OCR processing was much faster than in FreeOCR (both, FineReader and FreeOCR, were tested with the same input file, the 10 page document "Aspekte integrierter Datenverarbeitung im Bibliothekswesen am Beispiel des Informationssystems SISIS (Biblio EDV, 1991)" of the Stock collection). As already experienced with FreeOCR, text recognition results were much better when the correct document language was selected. In addition to manual selection, FineReader offers an automatic language selection, making it a very comfortable tool for processing many files by hand. Furthermore, FineReader offers a command line interface which allows automated text extraction.

Based on this evaluation, Abbyy FineReader 11 Professional Edition was purchased for

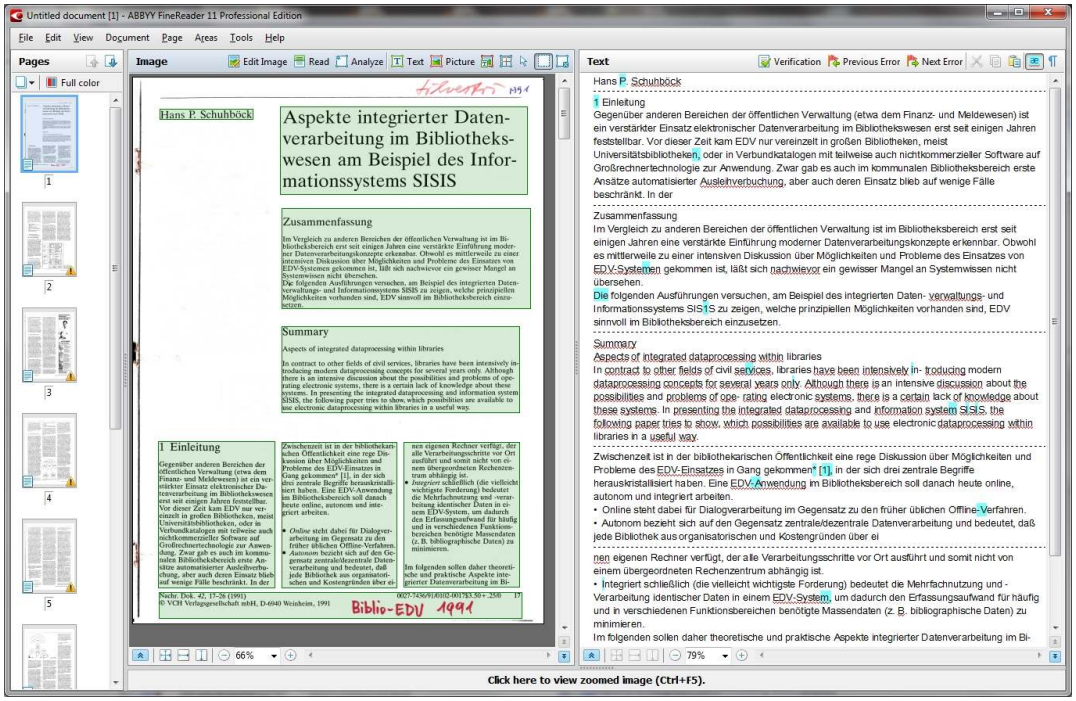


Figure 6.3: OCR with FineReader 11.

about 80 EUR, and this software was then used for all text and image extraction in the Stock project.

6.2 Hardware equipment

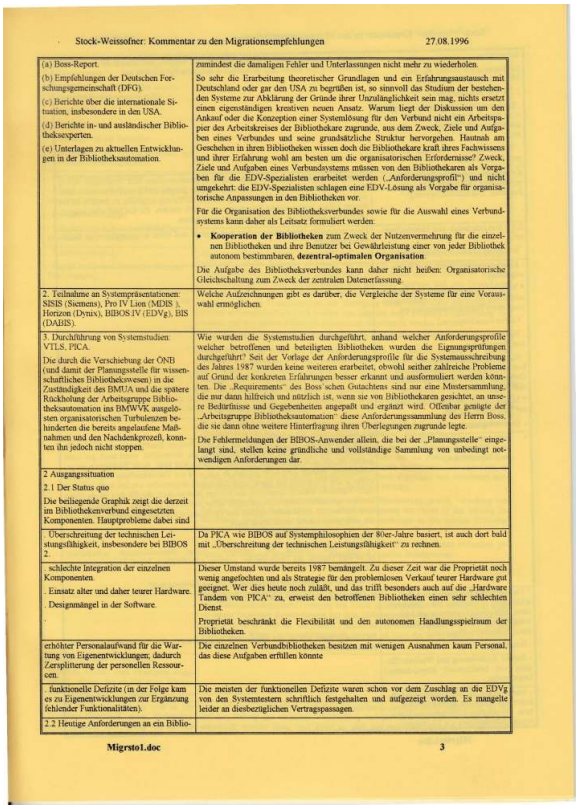
All OCR processing steps were performed on a machine with the following specifications:

- Intel Core i7-3770K @ 3.50 GHz
- 8 GB RAM
- 128 GB SSD drive
- Windows 7 64-bit

This hardware was well-suited for the job, processing 10 A4 pages with text and images in fewer than 15 seconds.

6.3 Content extraction

The content data extraction process requires the PDF files that were created during the digitization phase. They will serve as input for the selected OCR tool: Abbyy FineReader. Since there



Stock-Weissofner: Kommentar zu den Migrationsempfehlungen 27.08.1996

(a) Boss-Report. (b) Empfehlungen der Deutschen Forschungsgemeinschaft (DFG). (c) Berichte über die internationale Situation, insbesondere in den USA. (d) Berichte in- und ausländischer Bibliotheksexperten. (e) Unterlagen zu aktuellen Entwicklungen in der Bibliotheksautomation.	zumindest die damaligen Fehler und Unterlassungen nicht mehr zu wiederholen. So sehr die Erarbeitung theoretischer Grundlagen und ein Erfahrungsaustausch mit Deutschland oder gar den USA zu begrüßen ist, so sinnvoll das Studium der bestehenden Systeme zur Abklärung der Gründe ihrer Unzulänglichkeit sein mag, nichts ersetzt einen eigenständigen kreativen neuen Ansatz. Warum liegt der Diskussion um den Ankauf oder die Konzeption einer Systemlösung für den Verbund nicht ein Arbeitspapier des Arbeitskreises der Bibliothekare zugrunde, aus dem Zweck, Ziele und Aufgaben eines Verbandes und seine grundsätzliche Struktur hervorgehen. Hautnah am Geschehen in ihren Bibliotheken wissen doch die Bibliothekare kraft ihres Fachwissens und ihrer Erfahrung wohl am besten um die organisatorischen Erfordernisse? Zweck, Ziele und Aufgaben eines Verbundsystems müssen von den Bibliothekaren als Vorgaben für die EDV-Spezialisten erarbeitet werden („Anforderungsprofil") und nicht umgekehrt: die EDV-Spezialisten schlagen eine EDV-Lösung als Vorgabe für organisatorische Anpassungen in den Bibliotheken vor. Für die Organisation des Bibliotheksverbundes sowie für die Auswahl eines Verbundsystems kann daher als Leitsatz formuliert werden: • **Kooperation der Bibliotheken** zum Zweck der Nutzenvermehrung für die einzelnen Bibliotheken und ihre Benutzer bei Gewährleistung einer von jeder Bibliothek autonom bestimmbaren, **dezentral-optimalen Organisation**. Die Aufgabe des Bibliotheksverbundes kann daher nicht heißen: Organisatorische Gleichschaltung zum Zweck der zentralen Datenerfassung.
2. Teilnahme an Systempräsentationen: SISIS (Siemens), Pro IV Lion (MDIS), Horizon (Dynix), BIBOS IV (EDVg), BIS (DABIS).	Welche Aufzeichnungen gibt es darüber, die Vergleiche der Systeme für eine Vorauswahl ermöglichen.
3. Durchführung von Systemstudien: VTLS, PICA. Die durch die Verschiebung der ÖNB (und damit der Planungsstelle für wissenschaftliches Bibliothekswesen) in die Zuständigkeit des BMUA und die spätere Rückholung der Arbeitsgruppe Bibliotheksautomation ins BMWVK ausgelösten organisatorischen Turbulenzen behinderten die bereits angelaufene Maßnahmen und den Nachdenkprozeß, konnten ihn jedoch nicht stoppen.	Wie wurden die Systemstudien durchgeführt, anhand welcher Anforderungsprofile welcher betroffenen und beteiligten Bibliotheken wurden die Eignungsprüfungen durchgeführt? Seit der Vorlage der Anforderungsprofile für die Systemausschreibung des Jahres 1987 wurden keine weiteren erarbeitet, obwohl seither zahlreiche Probleme auf Grund der konkreten Erfahrungen besser erkannt und ausformuliert werden konnten. Die „Requirements" des Boss'schen Gutachtens sind nur eine Mustersammlung, die nur dann hilfreich und nützlich ist, wenn sie von Bibliothekaren geschätzt, an unsere Bedürfnisse und Gegebenheiten angepaßt und ergänzt wird. Offenbar genügte der „Arbeitsgruppe Bibliotheksautomation" diese Anforderungssammlung des Herrn Boss, die sie dann ohne weitere Hinterfragung ihren Überlegungen zugrunde legte. Die Fehlermeldungen der BIBOS-Anwender allein, die bei der „Planungsstelle" eingelangt sind, stellen keine gründliche und vollständige Sammlung von unbedingt notwendigen Anforderungen dar.
2 Ausgangssituation 2.1 Der Status quo Die beiliegende Graphik zeigt die derzeit im Bibliothekenverbund eingesetzten Komponenten. Hauptprobleme dabei sind	
. Überschreitung der technischen Leistungsfähigkeit, insbesondere bei BIBOS 2.	Da PICA wie BIBOS auf Systemphilosophien der 80er-Jahre basiert, ist auch dort bald mit „Überschreitung der technischen Leistungsfähigkeit" zu rechnen.
. schlechte Integration der einzelnen Komponenten. . Einsatz alter und daher teurer Hardware. . Designmängel in der Software. .	Dieser Umstand wurde bereits 1987 bemängelt. Zu dieser Zeit war die Proprietät noch wenig angefochten und als Strategie für den problemlosen Verkauf teurer Hardware gut geeignet. Wer dies heute noch zuläßt, und das trifft besonders auch auf die „Hardware Tandem von PICA" zu, erweist den betroffenen Bibliotheken einen sehr schlechten Dienst. Proprietät beschränkt die Flexibilität und den autonomen Handlungsspielraum der Bibliotheken.
erhöhter Personalaufwand für die Wartung von Eigenentwicklungen; dadurch Zersplitterung der personellen Ressourcen.	Die einzelnen Verbundbibliotheken besitzen mit wenigen Ausnahmen kaum Personal, das diese Aufgaben erfüllen könnte
. funktionelle Defizite (in der Folge kam es zu Eigenentwicklungen zur Ergänzung fehlender Funktionalitäten).	Die meisten der funktionellen Defizite waren schon vor dem Zuschlag an die EDVg von den Systemtestern schriftlich festgehalten und aufgezeigt worden. Es mangelte leider an diesbezüglichen Vertragspassagen.
2.2 Heutige Anforderungen an ein Biblio-	

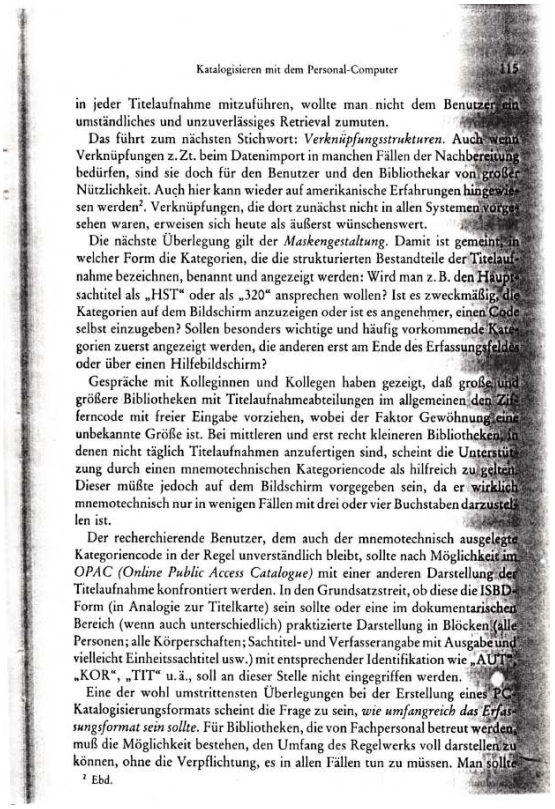
Migrsto1.doc 3

Figure 6.4: Example for layout that has to be handled manually.

was no uniform layout that was used in all documents, every document had to be handled manually. After the file was opened in FineReader, the software automatically initiated the character and image recognition. The result was displayed as shown in figure 6.3. The green selection boxes on the left side highlight recognized areas that are displayed in their extracted text form on the right side. Uncertain characters are highlighted for manual review, and a spell checker marks unknown and misspelled words.

Some layouts favored OCR processing more than others. Figure 6.4 shows an especially complicated layout that could not be processed automatically. Each table cell was extracted as text, but all text segments were displayed in a row, thus completely scrambling the original design. In such cases, the original design had to be copied in the final presentation framework, for example, as an HTML table in a Wiki or hypertext system. Until the document reached its final presentation form, layout information had to be preserved in the temporary storage form. Most documents used a simple text flow layout, allowing the extracted text file to be stored without additional layout information after manual modifications were completed.

Additional problems occurred with some of the older documents. For instance, the ink faded, causing bad OCR results; some copies were damaged from hole punching machines; other copies were of poor quality, as shown in figure 6.5. There was no automated solution possible



Katalogisieren mit dem Personal-Computer 115

in jeder Titelaufnahme mitzuführen, wollte man nicht dem Benutzer ein umständliches und unzuverlässiges Retrieval zumuten.

Das führt zum nächsten Stichwort: *Verknüpfungsstrukturen.* Auch wenn Verknüpfungen z. Zt. beim Datenimport in manchen Fällen der Nachbereitung bedürfen, sind sie doch für den Benutzer und den Bibliothekar von großer Nützlichkeit. Auch hier kann wieder auf amerikanische Erfahrungen hingewiesen werden[2]. Verknüpfungen, die dort zunächst nicht in allen Systemen vorgesehen waren, erweisen sich heute als äußerst wünschenswert.

Die nächste Überlegung gilt der *Maskengestaltung.* Damit ist gemeint, in welcher Form die Kategorien, die die strukturierten Bestandteile der Titelaufnahme bezeichnen, benannt und angezeigt werden: Wird man z. B. den Hauptsachtitel als „HST" oder als „320" ansprechen wollen? Ist es zweckmäßig, die Kategorien auf dem Bildschirm anzuzeigen oder ist es angenehmer, einen Code selbst einzugeben? Sollen besonders wichtige und häufig vorkommende Kategorien zuerst angezeigt werden, die anderen erst am Ende des Erfassungsfeldes oder über einen Hilfebildschirm?

Gespräche mit Kolleginnen und Kollegen haben gezeigt, daß große und größere Bibliotheken mit Titelaufnahmeabteilungen im allgemeinen den Zifferncode mit freier Eingabe vorziehen, wobei der Faktor Gewöhnung eine unbekannte Größe ist. Bei mittleren und erst recht kleineren Bibliotheken, in denen nicht täglich Titelaufnahmen anzufertigen sind, scheint die Unterstützung durch einen mnemotechnischen Kategoriencode als hilfreich zu gelten. Dieser müßte jedoch auf dem Bildschirm vorgegeben sein, da er wirklich mnemotechnisch nur in wenigen Fällen mit drei oder vier Buchstaben darzustellen ist.

Der recherchierende Benutzer, dem auch der mnemotechnisch ausgelegte Kategoriencode in der Regel unverständlich bleibt, sollte nach Möglichkeit im *OPAC (Online Public Access Catalogue)* mit einer anderen Darstellung der Titelaufnahme konfrontiert werden. In den Grundsatzstreit, ob diese die ISBD-Form (in Analogie zur Titelkarte) sein sollte oder eine im dokumentarischen Bereich (wenn auch unterschiedlich) praktizierte Darstellung in Blöcken (alle Personen; alle Körperschaften; Sachtitel- und Verfasserangabe mit Ausgabe und vielleicht Einheitssachtitel usw.) mit entsprechender Identifikation wie „AUT", „KOR", „TIT" u. ä., soll an dieser Stelle nicht eingegriffen werden.

Eine der wohl umstrittensten Überlegungen bei der Erstellung eines PC-Katalogisierungsformats scheint die Frage zu sein, *wie umfangreich das Erfassungsformat sein sollte.* Für Bibliotheken, die von Fachpersonal betreut werden, muß die Möglichkeit bestehen, den Umfang des Regelwerks voll darstellen zu können, ohne die Verpflichtung, es in allen Fällen tun zu müssen. Man sollte

[2] Ebd.

Figure 6.5: Example for poor document quality.

for such documents, and it was often necessary to manually type the missing paragraphs.

Once the manual review in FineReader was finished, the manipulated text was stored in a plain text file, using the same file name as the original PDF document and the appropriate text file extension. Images were stored in JPG format and again, the PDF file name was used with the image extension. Since most documents required more than one associated image file, a counter was added to the file name to maintain the correct order of images.

A few documents, such as the example in figure 6.6, required a layout that was not reproducible with the means of a Wiki system. These documents were, therefore, stored as an image file only, which would then be displayed in the Wiki article instead of the document text. To support search for such documents, keywords were created and stored in each article, as will be explained in the following section.

6.4 Metadata extraction

The documents of the Stock collection did not follow a uniform naming convention. Some documents were marked with an identification code assigned by Dr. Stock (as shown in figure 5.4), while others did not provide a publishing year. An automatic metadata extraction, such as one

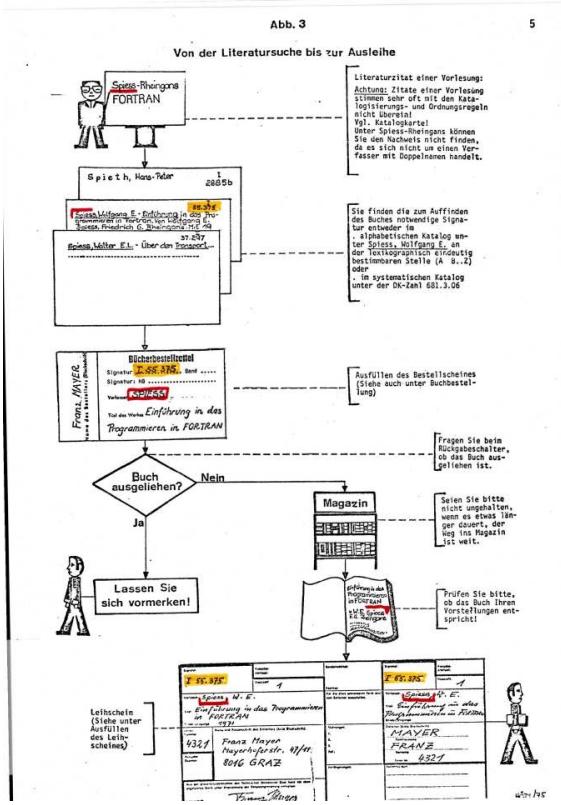


Figure 6.6: Example for a document that cannot be processed with OCR.

based on the provided Excel lists, was not possible, and metadata was extracted manually. As the complete text of each document would be available to full text search via the Wiki functionality, only a rather small set of metadata appeared to be necessary:

- document author
- publishing year
- keywords

The initial approach for the keywords was to create a tag cloud using a web service located at tagcrowd.com. An example for such a tag cloud can be seen in figure 6.7. However, even though the appearance of the Wiki articles was more attractive with the tag cloud, a change in the Wiki system to handle raw HTML was necessary. This resulted in raw HTML code being displayed in all search results. Furthermore, this created a dependency for future extensions of the digital collection on the web service, which then might not be available anymore, preventing the Wiki from maintaining a consistent look and feel. It was thus decided to manually assign keywords to each article.

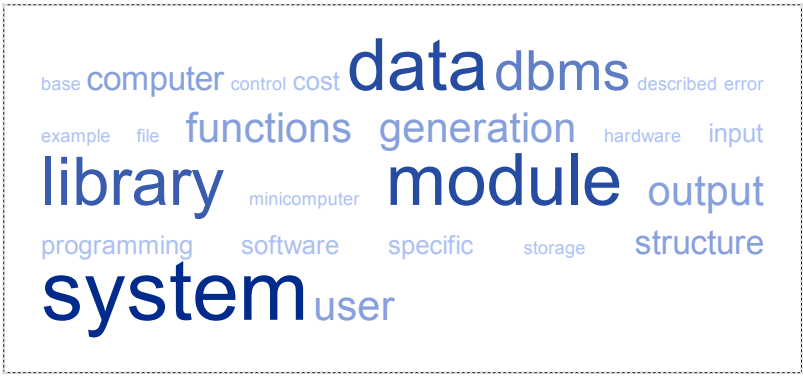


Figure 6.7: Example for a tag cloud from tagcrowd.com.

Once all required metadata was compiled, the information was stored in a plain text file. Then, the file was saved under the same file name as the extracted text (which was based on the PDF file name) with an additional suffix to mark it as the metadata storage file.

6.5 Results of the cataloging phase

At the end of this phase, the following artifacts were available:

- digital versions of the original files stored as PDF files and named after the contained document
- the text extracted from the PDF files, stored in a plain text file named after the source document
- the extracted image files, stored as JPG files named after the source document with an increasing counter
- metadata for each document, stored as a plain text file named after the source document with a name extension to mark the file as metadata

CHAPTER 7

Presentation

Presentation

This chapter outlines how the results from the previous steps were used to build the hypertext system. A short introduction of the requirements is given to provide an overview of the following steps. The implementation of each requirement is discussed in greater detail, together with encountered problems and their solutions.

7.1 Requirements

The following requirements were set for the digital version of the Stock collection at the beginning of the project:

- the collection should be presented in a hypertext system (section 7.2)
- semantic search based of article attributes should be available for focused research (section 7.3)
- thematic tours for random exploration of the collection should be provided (section 7.4)

Each requirement is explained, and implementation details are discussed in the corresponding sections below.

7.2 Presentation in a hypertext system

This sections describes the different aspects that have to be considered to make best use of the extended capabilities of a hypertext system.

Hypertext system

One of the project goals was the integration of the Stock collection into the existing "Die Informatisierung Österreichs" [6] framework; the presentation layer was thereby defined to be a Wiki hypertext editing system. As Wikis are described as suitable platforms for the collaborative production of information sources [20], this appeared to be an ideal choice considering the additional project goal of supporting collaborative work on and extension of the digital collection. In addition, the discussion page functionality available in Wiki systems serves as a communication platform for teams, which is an essential advantage for the self-organization of study and research groups according to [20].

The target audience at this stage of the project is intentionally kept small to reduce the moderation effort. So far, only administrators and participating students have access to the Wiki, but it is not yet available to the general public. Progress in the parent project is documented on the discussion page, which also serves as a form of work progress log.

Article template

The template for articles of the Stock collection was designed to display the most important information for each article at the top of the page. The Wiki article name is the name of the original printed source document. If the publishing year was known, it was added to the article together with the identification code that was assigned by Dr. Stock (if one was available; an example can be seen in figure 5.4). Because of the heterogeneous nature of the source document, there is no standard template that fits the content of all entries. Therefore, it appeared necessary to display all metadata before the actual document text. All metadata information are stored as semantic attributes to be available for semantic search. The following metadata is contained in each article in this order:

1. collection: An attribute to specify the collection to which the article belongs. So far, only articles of the Stock collection are online.
2. Original document scan: A link to the PDF version of the original document.
3. Author: The original creator of the content.
4. Year: The publishing year of the document.
5. Keywords: Keywords based on the content of the article.

These data sets are followed by an index of the full article. This index is automatically created by the Wiki system based on the article markup. The full text of the document, which was extracted via OCR from the scanned PDF file, is displayed as the last segment of the page. A Wiki template was created so that this structure can be reused during future extensions of the Stock collection.

Article integration

Based on the artifacts created during the cataloging phase, the Wiki articles were created online in the live system using the Stock collection Wiki template. The PDF files were uploaded simultaneously and linked to the correct article. The link to the PDF file was checked manually after each article was saved to avoid dead or wrong links.

Although time-consuming, this was a rather straightforward process. The single noteworthy problem encountered was the file upload limit of the PHP installation running the MediaWiki system. Approximately 20 of the 140 PDF files were larger than the 8 MB upload limit. Since the web server administrator had concerns about raising the file upload limit globally, these files were copied to an FTP server and the links for these articles point to their location on that server, instead of using the Wiki functionality for uploaded files.

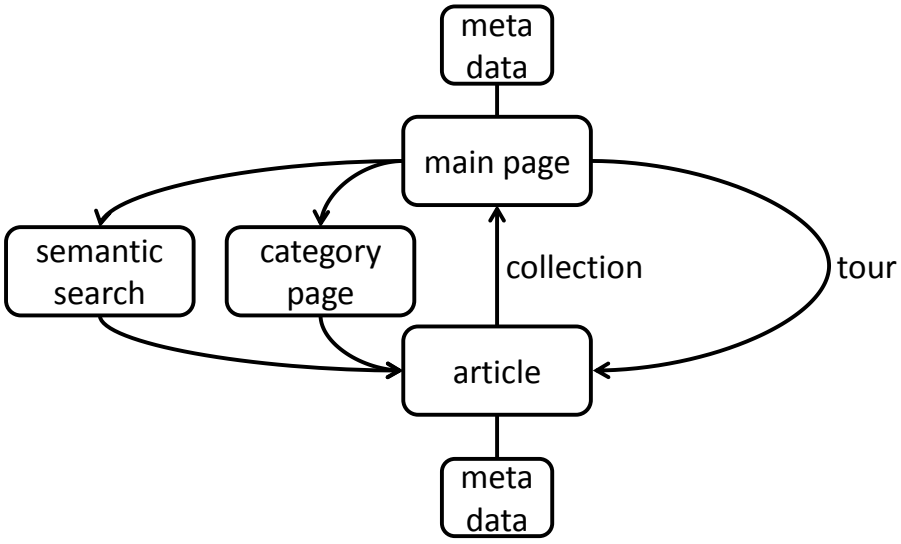


Figure 7.1: The Wiki navigation diagram.

Navigation

The Wiki was intended to be a closed system. Hyperlinks should only reference files and articles within the Wiki. This goal was achieved with the exception of a few PDF files that did not meet the file upload size restrictions and, therefore, had to be hosted on an FTP server.

Documents and files within the Wiki are connected via meaningful links. The following navigation paths, also displayed in figure 7.1, are allowed:

Main page to article via semantic search: This path describes a search as it might be used in a research project. The user combines multiple attributes on the semantic search site and receives a result set containing links to all articles that match the search criteria.

Main page to article via category page: The category page is a shortcut to display all articles that were assigned to a certain category. In this case, all articles of the digital Stock collection are displayed.

Main page to article via theme based tour: Theme based tours are embedded semantic search queries, therefore this path is equivalent to the semantic search path from a technical point of view. From a user's point of view, tours provide a guided path through the collection without the need to know the technical details of the semantic search functionality.

Article to main page via the collection attribute: Each article contains a semantic attribute that identifies the collection to which the article belongs. The attribute also serves as a link back to the collection page, which is the entrance point for the digital collection.

Assuming that Wiki users are familiar with basic browser functions, an explicit return link to the main page was only implemented on article level, thus creating a one directional path

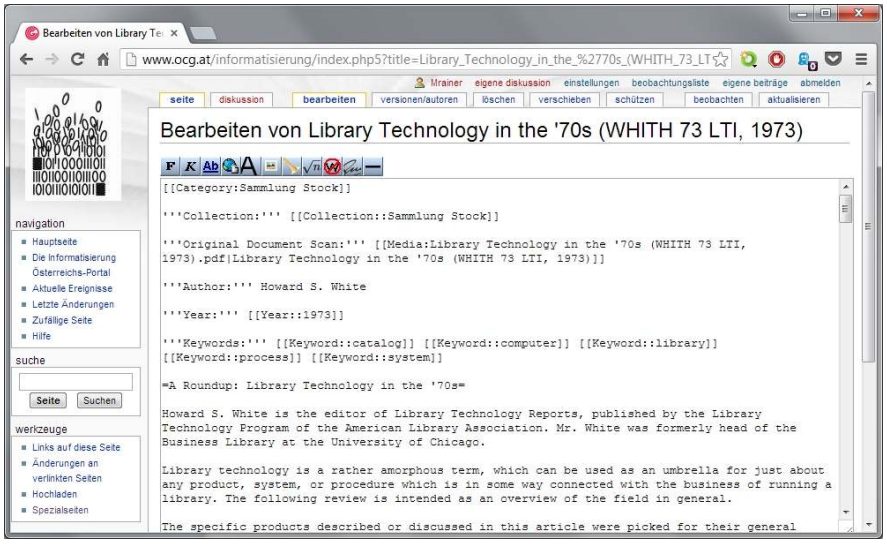


Figure 7.2:

and reducing maintenance efforts, as it is also described by Landow in [11]. An implicit return function at any stage is available via the browser back button.

7.3 Semantic search

To make the Stock collection available as a digital archive for research, the basic search functionality of the Wiki system had to be extended. While the used MediaWiki distribution provides full-text search out of the box, it is not possible to classify articles with keywords or attributes and then perform a search only for certain attribute values. To support this use case, the semantic search functionality was added via the Semantic MediaWiki plug-in [2]. To use this plug-in, the Wiki articles had to be extended with semantic attributes. The syntax is similar to the regular Wiki markup. A regular Wiki link (e.g. a link to the article Stock Collection: [Stock Collection]) can be turned into an attribute by simply adding a second pair of brackets (e.g. a link to the article Stock Collection that is also used as attribute: [[Stock Collection]]). To use semantic attributes for search, a value has to be assigned to the attribute (e.g. assigning the value Stock to the attribute Collection: [[Collection::Stock]]). Figure 7.2 shows an example for semantic attributes in a Wiki article markup.

These attribute categories are extendable by simply adding a new attribute. Once the article is saved, the new attribute can already be used in the semantic search function. The list of all keywords available at the time of this writing is available in the appendix, section 12.3.

The semantic search plug-in provides two search functions:

- search by attribute: search all articles for attributes that have the given value, as shown in figure 7.3

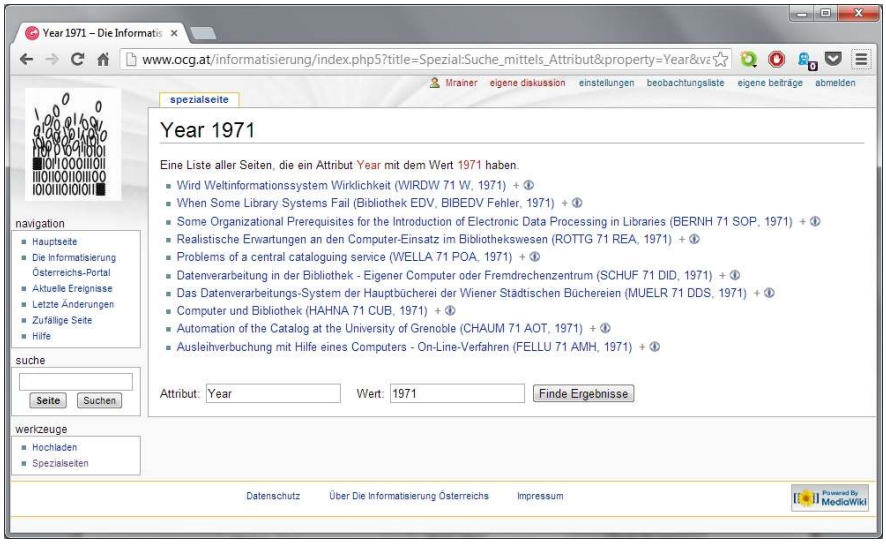


Figure 7.3: Example for a simple semantic search by attribute.

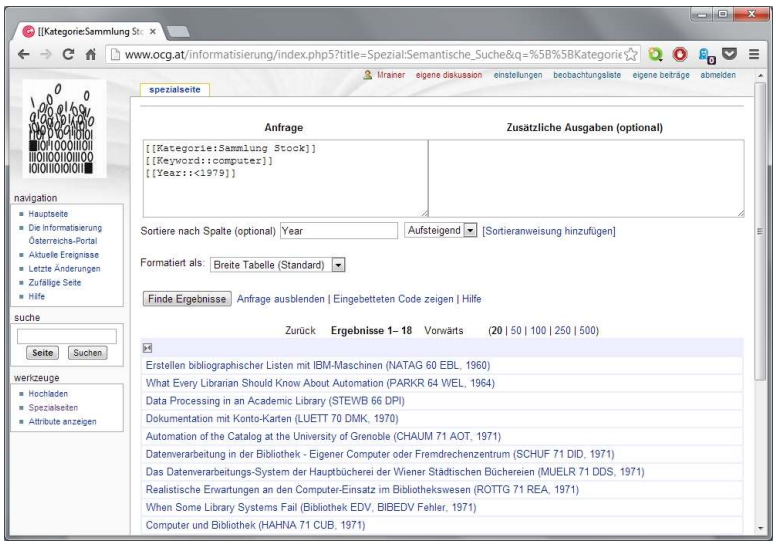


Figure 7.4: Example for a complex semantic search query.

- general search: construct complex queries combining multiple attributes, as shown in figure 7.4

Both search function can be reached via the special pages link in the Wiki system. The search by attribute allows a simple and quick search, but it is not very powerful. If multiple attributes have to be combined, a more complex query has to be defined via the semantic search page. These queries can be saved and embedded in Wiki articles. Multiple formatting options

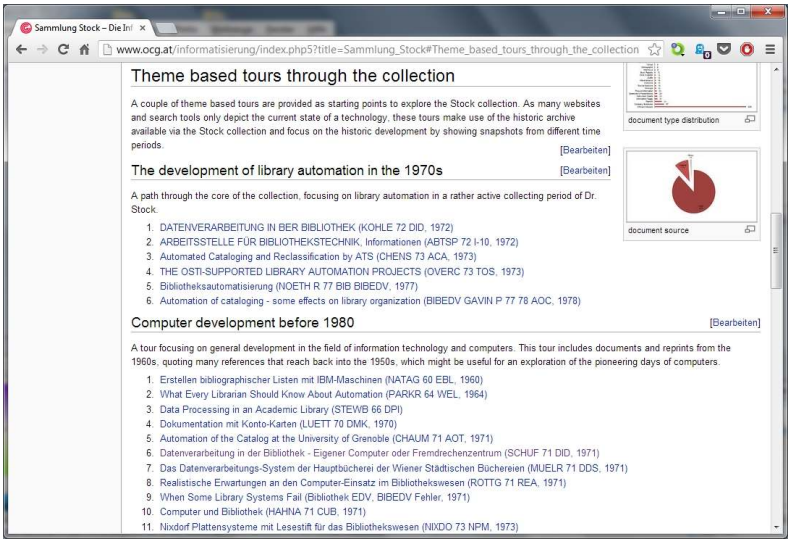


Figure 7.5: Theme based tours of the Stock collection.

for the result table are available. If the query is embedded in an article, each time the article is loaded, the result set always contains the latest information based on the available Wiki articles. These queries, hence, provide a powerful tool for growing collections with constant changes. Such queries are also used for the thematic tours (section 7.4) in the Stock collection.

7.4 Thematic tours

A set of sample thematic tours was created for random exploration of the Stock collection based on topics of interest. These tours are composed using semantic search queries combining multiple semantic attributes. If the digital collection grows and additional articles match the tour query, the number of articles included in the tour grows as well. The tours build on the concept of the sample lesson plans provided by the U.S. Library of Congress for their American Memory collection [17]. Students have the opportunity to explore the collection's contents on a guided path and develop understanding of subject matter they might need for class assignments, as suggested by the theory of constructivism [24]. Figure 7.5 shows the starting page of the currently available tours.

The following tours are available as of this writing:

The development of library automation in the 1970s: A path through the core of the collection, focusing on library automation in a rather active collecting period of Dr. Stock.

Computer development before 1980: A tour focusing on general development in the field of information technology and computers. This tour includes documents and reprints from the 1960s, quoting many references that reach back into the 1950s. These sources might prove useful for exploring the pioneer days of computers.

Early state of the art: Providing a range of reports from projects before 1990, this tour paints a picture of the state of the art in the 1970s and 1980s.

The development of library catalogs in the late 20th century: Stretching from 1980 to the end of the collection, this tour puts the focus on library catalog developments.

The current set of tours was created manually with knowledge of the content of the Stock collection. The investigation of methods for an automatic creation of tours is left for a follow-up project (see section 10.5). Such a project would have to determine whether tours can be created automatically at all, without knowledge of the article content, and whether such tours would still be meaningful to humans.

7.5 Results of the presentation phase

At the end of this phase, the following artifacts were available:

- the digital versions of the documents were available online
- an article for each document, containing the content data and associated metadata
- a start page with thematic tours through the collection
- semantic search via the keywords attributes

CHAPTER 8

Teaching projects

Teaching projects

This chapter presents ideas of how the Stock collection can be used for educational purposes.

8.1 Collaborative extension

As only ca. 10 % of the Stock collection were transferred to a digital form in the scope of this project, it is necessary to continue the work and extend the digital archive. Based on the example of the host system, the "Die Informatisierung Österreichs" Wiki [6], the extension workload could by split by collaborative work. As it was done for the predecessor project, the Stock collection could be extended as part of high school and university projects.

Digitization phase

A critical aspect of distributing the workload is how to deal with the original paper documents during the digitization phase. Handing out the original documents to students poses the risk of losing the document. Unlike high school classes, university courses have no legal attendance requirements; other than the student's motivation to finish the course to get a positive result paper, there is no legal way to force a student to return to a course. If an artifact from the collection is handed to a student at the beginning of the term and the student decides to drop the course, the collection owner depends on the good will of the student to return the item. If the participants are required to return to the class (e.g. in schools), the risk of loss could be reduced. The risk of (accidental) damage to the document however remains with every form of distribution. In order to deal with the risk of loss, a high school class seems to be more suitable for the digitization phase: attendance is mandatory and the student's parents could be contacted by the school as a last resort to retrieve the original documents.

The digitization phase could be implemented as part of a basic information technology class. Students could learn how to handle computer peripherals such as the scanner. It would be a valuable lesson especially for children who do not have easy access to technology at home. Based on the principles of constructivism [24], students could try different scan settings and evaluate the results for themselves as feedback on how their adjustments worked out. Setting up and using computer peripherals by themselves at an early age could give the participants confidence to set the foundation for their computer literacy and prevent anxiety towards hardware and technology that goes beyond using social networks on a cell phone.

Cataloging phase

Once the documents are available in a digital form, the opportunities for using them are not limited by potential damage to or loss of the original sources anymore. The digital documents could also be used in a basic information technology class on a high school level. While text processing and spreadsheet applications can be considered fairly common, OCR tools go beyond the basic spectrum of standard software for many computer users. This provides an opportunity to try and learn how to use a new tool set in a controlled environment, with guidance if needed.

To keep up with current technology, a re-evaluation of new OCR tools could also be considered for a school project or a homework assignment. The students could be given the task of transferring the paper documents to a digital form and to come up with a solution of their own. The different results could be discussed in class together with their advantages and restrictions.

8.2 Research projects

So far only technical aspects of the Stock project were considered. A content review of the Stock collection documents was only done to the point necessary to allow a classification of the document, but no organized processing of the content has taken place yet. This could be done in an interdisciplinary research project in history and computer science classes as a bridge between digital humanities and engineering. Students could be assigned a batch of documents based on the current keyword classification. To avoid multiple students processing the same document, these assignment lists should be checked for double entries beforehand. The students could then begin with a content analysis of their batch. The results could be documented within the Wiki articles. Research interests might include the following aspects:

- development and use of one key technology using all resources available on the technology of interest, e.g. the MARC catalog format
- general development of the library automation field during a given time period, e.g. automation milestones in the 1970s
- studies on human-computer interaction and user interface design based on information material, e.g. for library terminals

Built-in Wiki tools like the discussion page could be used to document the research progress and further ideas. This would be especially helpful if a team works on different aspects of the same set of documents; different approaches and views could be discussed online, independent of the geographic location and time of the team members.

As this kind of research could benefit from contextual knowledge, an interdisciplinary project seems to be most suitable. While technological problems such as semantic queries can be solved by the participating engineering students, the humanities students with a focus on technological history can provide the context for the information extracted from the Stock collection. The final product of such a research project might also be useful in education , e.g. by designing a lesson plan similar to the examples of the American Memory Project [17]. The sample lessons could then be used in computer science or history classes at high school or university level.

8.3 Software development projects

The presentation phase offers a few technical challenges and seems to be more suitable for a university course or project on software development than a high school setting. Some of the suggested follow-up projects 10 concerning the presentation require programming skills and an understanding of Wiki systems, databases and web servers. As the presentation affects the whole Wiki system, simultaneous changes by many students at the same time should be prevented; these projects are more suitable for a team or a single student assignment as part of a seminar, bachelor or master project.

Development projects could include the following aspects:

- customization of the existing Wiki system to make is easier to use in a high school environment, e.g. with feedback plug-ins to rate the quality of articles
- presentation of the articles with a customized look and feel, e.g. by designing a color scheme and icon set
- design and implementation of a system-wide information layout, e.g. combing semantic attributes in information boxes
- automating OCR tasks, e.g. implementing a system that automatically marks uncertain characters in the Wiki system with a Wiki template for further review

CHAPTER 9

Conclusion

Conclusion

The total runtime of the practical work on this project was approximately 300 hours, divided over a period of three months. The first month, with five work days per week, was used for digitization; the Stock collection was analyzed, documents were selected and sorted out and then scanned. For the second and third month, work time was reduced to approximately two days a week. This time was used mostly for the cataloging phase, with some time reserved toward the end for the presentation phase.

For the Stock collection of ca. 1400 documents, the implementation efforts can be summarized as follows:

- ca. 50% project time for digitization
- ca. 40% project time for cataloging
- ca. 10% project time for presentation

If previous experience with the used presentation framework exists, the amount of time used for the presentation phase can most likely be reduced to 5% or even less.

The total amount of financial expenses for this project was 160 EUR. 80 EUR were spent on a scanner and an additional 80 EUR were spent on a commercial OCR software solution. However, if only printed documents of high quality are digitized, the OCR function of the scanner might be sufficient.

Experience reports and recommendations of the Library of Congress and some other organizations were used as reference for the Stock project. A few modifications of these reference processes were necessary to reflect the needs of this project. Because of the limited resources, quality control had to be done by the same person implementing all phases, which is a procedure that is generally not recommended. Furthermore, it was not feasible to follow some format recommendations. The TIFF format, which is widely used for digital preservation, did not provide the ease of use desired for the Stock project. Because a Wiki system serves as the presentation framework, PDF seemed to be the more suitable format choice in this case.

In conclusion, the Stock project proved that the digitization of a collection can be achieved with limited resources. The guidelines and best practices established by the U.S. Library of Congress and other organizations can also function on a smaller scale. Consequently, it is possible to implement all required steps in a one man project.

CHAPTER 10

Follow-up projects

Follow-up projects

During the implementation of the Stock project, a few issues were discovered that did not fit in the scope of the project. This chapter provides ideas for follow-up projects based on these findings.

10.1 Dealing with copyright issues

A future project should evaluate whether the materials of the Stock collection are still protected by copyrights, since this was not within the scope of this project. The non-commercial, educational, and academic nature of the Wiki should be considered when approaching copyright holders. As far as books are concerned, it should by possible to determine if a company still claims rights. One solution for books might be to use the ISBN code of the book as reference; in addition the index could be displayed. The magazine articles might pose a bigger challenge, especially if the magazine series or the publishing company do not exist anymore. While the Wiki is still password-protected, search engines will not index the articles, and access will be limited to a small, selected group. These questions should be answered before the Wiki is made public. Again, the American Memory Project could serve as a reference: as stated in the frequently asked questions section [14], the "web site does not provide definitive legal advice on particular questions of copyright"; users must make their "own, independent assessment of the legal rights that may exist in the materials".

10.2 Thematic collections

If multiple collections become available online, thematic collections containing documents from different source collections could be set up using semantic attributes. Considering the wide variety of topics already included in the Stock collection alone, a large number of digital items will be necessary to accumulate collections focused on a single topic, e.g. catalog formats only.

10.3 Display of semantic attributes in info boxes

Similar to articles on Wikipedia, the semantic attributes could be displayed in an info box rather than as a plain list. This would be merely a cosmetic change; the semantic search functionality would not be changed. This project would require a new Wiki template to be linked in every

article. It might be necessary to export all articles to XML to allow batch processing, e.g. reworking all articles with the help of regular expressions, and then import the modified articles.

10.4 Automatic OCR solution with quality check

Many OCR tools provide a programming interface that allows batch processing. A tool-set could be compiled to automatically run the OCR software on a PDF file and then determine the quality based on the number of uncertainties in the OCR tool. If the number is too high, the document is kept for a quality check by a human supervisor. If the text extraction did not raise any issues, the document is automatically processed and integrated into the Wiki.

10.5 Automatic creation of tours

The first set of theme-based tours of the Stock collection was created manually. A follow-up project could investigate options to automate this process. The project should investigate, whether it is at all possible to create tours, that are meaningful to humans, with an automated solution. One approach might be an automatic analysis of available keywords via entity recognition. The frequency of the keywords could serve as an indicator of how interesting a topic might be in the context of the Stock collection. Another approach might use social network tools which let the users vote on potential tour topics.

CHAPTER 11

Summary

Summary

The Stock collection is an accumulation of printed documents. It was compiled by Dr. Karl F. Stock, the former head of the library of the Technical University Graz, and the collection contains documents published between 1964 and 1998. Most artifacts are photo copies and only a few original items (such as information material and books) are included. Approximately 10 % of the Stock collection was transferred from paper to a digital collection as part of a proof of concept project to evaluate if large scale industry standards can be applied to small scale and home projects with a low budget.

The state of the art was analyzed based on technical standards and recommendations from the American Memory Project of the U.S. Library of Congress. Additional resources from the Federal Agencies Digitization Guidelines Initiative and the Digital Preservation Coalition were evaluated to gather know-how for the technical implementation part of the project. Literature on hypertext systems and multimedia learning systems provided the theoretical and didactic foundation.

The findings of the project implementation can be summarized in three phases:

1. Digitization phase: This phase starts with the selection of documents based on certain criteria such as content and quality. The scanning equipment is tested to evaluate the best settings for the used hardware. The selected paper documents are scanned, and their digital version is stored for the next phase.

2. Cataloging phase: The digital documents are integrated into an existing Wiki system during this phase. The files containing the scanned documents are processed with OCR tools to extract text and images. All files are then integrated in templates and uploaded to the Wiki server.

3. Presentation phase: The basic presentation design has to be done only once per project. Common use cases and user interest should be considered to provide easy access to the information in the Wiki articles. Semantic attributes have to be defined.

The final product of this project was the foundation of the digital version of the Stock collection. At the current implementation state, approximately 10 % of the Stock collection is online in a Wiki system. These approximately 140 articles are available for full-text search. All scanned documents are linked in the corresponding article to preserve the original look and feel of the paper documents. Thematic tours are available for exploration of the collection based on topics of interest. A more focused research is supported by semantic search via defined keywords.

The lessons learned during the implementation of this project are summarized in a few basic guidelines to support further work on the project.

Examples for the didactic use of the digital Stock collection are provided, with a focus on how to extend the collection and use it for research. A couple suggestions for follow-up projects are also included.

CHAPTER 12

Appendix

Appendix

12.1 Guidelines

The following guidelines should help to identify the tasks that have to be performed for a digitization project and to estimate the effort and required resources. These guidelines are based on the practical experience and lessons learned from the Stock project.

Identify the core of the collection: If the collection is too large to be digitized completely in the scope of the project using the available resources, it is necessary to select a minimum set of artifacts for the first wave of digitization. The digital versions of these items will form the foundation on which future extensions can build. It helps to look at all documents at least once, to get a feeling for the collection and its contents. Take notes of recurring keywords, and start a check list to keep track of how often certain terms reappear. The terms with the highest frequency mark the central topics of the collection.

Select documents based on the core topics: Using the information from the first step, select a set of documents for the first digitization wave to form the foundation of the digital collection.

Evaluate the digitization hardware and software: Select scanning hardware and software based on document types. Text used for OCR processing requires higher quality than text stored as plain image. If the digital collection completely replaces the original collection, the highest possible quality standards should be applied. Run a few test scans to determine the best performing settings.

Prepare the concept for presentation: Create a draft of the concept for the presentation of the digital collection using some representative artifacts. The following questions should be answered:

- What content data has to be displayed?
- What metadata data has to be displayed?
- Is additional metadata required that cannot be extracted directly from the documents (e.g. keywords)?
- How should the artifacts be connected?
- Will the collection require maintenance? How can this be supported?

- Is a suitable editor available for the digital artifacts?
- What will be the entry point of the digital collection?
- How can users navigate the digital collection?
- How can users find certain artifacts?

Select a presentation framework: Based on the requirements found in the previous step, select a solution that fits the needs of the digital collection. Set up a test environment to try out new settings and formats before the system goes live.

Validate the concept: Run the whole process using some sample documents: scan them, extract the data, and put them online. Check the concept for flaws, and modify it if necessary.

Scan the documents: Scan all documents that were selected for the digital collection. Save them with their final file name immediately (rename them if necessary).

Extract data from digital documents: Using the digital versions, extract all required content data and metadata. Store both in files with the same file names as the digital documents. If images are extracted, add an increment counter to maintain the correct order.

Upload the digital documents: Upload all files associated with one source document simultaneously to the presentation system. Complete each document before starting the next to avoid confusion and wrong file links.

Implement the start page: Create a start page for the collection. This page should contain background information, such as history, thematic focus, and ownership. Tours, like those based on the core topics of the collection, should be provided for interested users who want to explore the collection without a special focus. To support research interests, a semantic search or full text search function is recommended.

12.2 Wiki documents

As of this writing, the documents below are available online at `http://www.ocg.at/informatisierung/index.php5?title=Sammlung_Stock`. This list uses the exact name of the article, just as it can be found in the Wiki system.

- Über Fehlentwicklungen beim Aufbau von EDV-Systemen für Bibliotheken (ABELR 73 UFB, 1973)
- Überblick über den Stand der Einsatzvorbereitung von elektronischen Datenverarbeitungsanlagen im Bibliothekswesen der sozialistischen Länder (STEIG 70 UUD, 1970)
- ALEPH - AUTOMATED LIBRARY EXPANDABLE PROGRAM
- ARBEITSSTELLE FÜR BIBLIOTHEKSTECHNIK, Informationen (ABTSP 72 I-10, 1972)
- AUFGABENBEREICH UND DERZEITIGE ARBEITEN (ABTSP 75 AUD, 1975)

- A LIBRARY CATALOGUING SYSTEM USING MICROCOMPUTERS. ANALYSIS OF AN EXPERIENCE. (PC Katalogisierung, 1983)
- Anforderungen an ein EDV-Programm für öffentliche Bibliotheken
- Anforderungen an ein EDV-Programm für öffentliche Bibliotheken (1990)
- Aspekte integrierter Datenverarbeitung im Bibliothekswesen am Beispiel des Informationssystems SISIS (Biblio EDV, 1991)
- Aufbau der Österreichischen Zeitschriftendatenbank (ÖZDB) (Ztschr Datbk, 1986)
- Ausleihverbuchung mit Hilfe eines Computers - On-Line-Verfahren (FELLU 71 AMH, 1971)
- Austrian Academic Libraries Survey (Bibliotheken Öst, 1983)
- Automated Cataloging and Reclassification by ATS (CHENS 73 ACA, 1973)
- Automation of cataloging - some effects on library organization (BIBEDV GAVIN P 77 78 AOC, 1978)
- Automation of the Catalog at the University of Grenoble (CHAUM 71 AOT, 1971)
- Automatisierung der Literaturverwaltung in einem Consulting-Betrieb (1996)
- Bücher - elektronisch erfaßt (PFLUG 66 BEE, 1966)
- Bücher für die Wissenschaft (1994)
- BIBLIO-DATA - die nationalbibliographische Datenbank der Deutschen Bibliothek
- BIBLIOTHEKS-VERBUND-SYSTEM BIBOS (1985)
- BIBLIOTHEKSAUTOMATION UND OPTISCHE BILDPLATTE
- BVS - für die Einzelbibliothek und den Verbund (SIEMENS 81 JTI, 1981)
- Bericht über den Stand der Bibliotheksautomatisierung in Österreich (KAMPH 78 BUD, 1978)
- Bibliographie der österreichischen Bibliographien (1984)
- Bibliographische Expertensysteme und Online-Kataloge
- Bibliothek Datenverarbeitung (Bibliotheken EDV, 1990)
- Bibliothekarisches EDV-Wissen (BIBEDV AUSB, SCHUJ 79 BEW, 1979)
- Bibliothekarisches Grundwissen (Bibliothek EDV, BIBEDV Hacker, HACKER R 83, 1983)

- Bibliotheksautomatisierung (NOETH R 77 BIB BIBEDV, 1977)
- Bibliotheksautomatisierung im Verbund (SIEMENS 78, 1978)
- Bibliotheksautomatisierung in den USA (USA EDV Bibl. 1987, 1987)
- Bibliotheksautomatisierung in den Vereinigten Staaten (BIBEDV USA, 1993)
- Bibliotheksautomatisierung und Bibliotheksnetzwerke in den USA (EDV-BNETZ NOET R 80 BUB, 1980)
- Bibliotheksnetze und elektronische Medien (LAN 94, 1994)
- Bibliotheksverbundsystem als Mittel zur Errichtung eines nationalen Verbundes (Bibliotheksverband 1988, 1988)
- Checkliste für die Retrospektive Konversion (RECON, 1994)
- Comic Ansuchen Magnetplatte (1982)
- Comic OCLC Database (1982)
- Computer und Bibliothek (HAHNA 71 CUB, 1971)
- Conference of European National Librarians (CENL) (1994)
- DATENVERARBEITUNG IN BER BIBLIOTHEK (KOHLE 72 DID, 1972)
- DIE ÜBERFÜHRUNG DER DEUTSCHEN NATIONALBIBLIOGRAPHIE AUF ELEKTRONISCHE DATENVERARBEITUNG UND DIE PERSPEKTIVEN DES EDV-EINSATZES FÜR DIE BIBLIOGRAPHISCHE LITERATURINFORMATION (ROSTG 70 DUD, 1970)
- Das Datenverarbeitungs-System der Hauptbücherei der Wiener Städtischen Büchereien (MUELR 71 DDS, 1971)
- Das Library Management System (JEDWB 72 DLM, 1972)
- Das Library of Tomorrow Projekt der Library of Congress (Maschinenlesbare Daten, 1990)
- Data Processing in an Academic Library (STEWB 66 DPI)
- Datenverarbeitung in der Bibliothek - Eigener Computer oder Fremdrechenzentrum (SCHUF 71 DID, 1971)
- Datenverarbeitungskonzepte für die italienische Nationalbibliothek in Rom (SKACF 72 DFD, 1972)
- Der Zettelkasten als Informations-Retrievalkonzept (SCHNP 70 DZA, 1970)
- Der elektronische Archivar (Archivar ScanView 1991, 1991)

- Die Arbeit des Hochschulbibliothekszentrums (RAU P 78 DAD, 1978)
- Die Rolle der Bibliotheken im Informationsnetz (Bibliothek Netz, PFLUG G 83 DRD, 1983)
- Die Rolle des Personal Computers in Bibliotheksverbundsystemen (1988)
- Die Vermittlung von Kenntnissen in der Bibliometrie (Bibliometrie, 1983)
- Die Zukunft der Enzyklopädie (KLOFAT R 81 DZD, 1981)
- Dokumentation mit Konto-Karten (LUETT 70 DMK, 1970)
- EDV-Systeme in deutschen Bibliotheken (EKZ, 1995)
- Eigenes Bibliotheksverbundsystem für Österreich überflüssig? (1982)
- Ein Gespenst geht um (Bibliotheken Kosten, 1983)
- Einsatz eines Arbeitsplatzcomputers an Spezialbibliotheken mittlerer Größenordnung (PC Kleinbibliothek, 1988)
- Einsatzmöglichkeiten von schlüsselfertigen EDV-Systemen im Rahmen der bibliothekarisch-dokumentarischen Ausbildung (DUGALL B 81 EVS, 1981)
- Elektronische Bibliotheksdienste - lokal und weltweit (OPAC Gopher, 1993)
- Empfehlungen des Wissenschaftsrates zur retrospektiven Katalogisierung an wissenschaftlichen Bibliotheken (Katalogisierung RECON, 1988)
- Erste Ergebnisse der Anwendung der Elektronischen Datenverarbeitung in der Nationalbibliothek der Sozialistischen Republik Serbien (STAMD 76 EED, 1976)
- Erstellen bibliographischer Listen mit IBM-Maschinen (NATAG 60 EBL, 1960)
- Film als Buch - Hyperdokumente zur Filmanalyse (Hypertext, 1990)
- Format für den Austausch von bibliographischen Daten (ISO 2709 DIN 1506, 1978)
- Fourth Generation Systems for Libraries (GROSA 77 FGS, 1977)
- Grundkonzept für den Einsatz der elektronischen Datenverarbeitung im Österreichischen wissenschaftlichen Bibliothekswesen (1972)
- Hypertext- & Hypermedia-Systeme - Ein Überblick (Hypermedia 1992, 1922)
- Hypertext und Information Retrieval (Hypertext, 1990)
- Integrated, Dedicated Minicomputer-Based Applications In South African Libraries (PC Bibliothek, 1983)

- Ist die retrospektive Katalogkonversion ein Zauberwort (Retrokonversion 1991, 1991)
- Katalogisieren mit dem Personal-Computer (PC Katalogisierung, 1989)
- Kommunikation in Bibliotheken (BIB-Net, 1991)
- Kooperative Bibliotheksautomatisierung - das niederländische Pica-System (PICA Bibnetz, 1986)
- LIBRARY AUTOMATION (Bibliotheksautomation MARC, 1980)
- LIBRARY LEADERSHIP AND NETWORKING IN NEW YORK STATE (GEDDES A 76 LLA, 1976)
- Library Automation Status and Trends (1993)
- Library Local Network With Shared Resource Manager (1984)
- Library Technology - The Black Box Syndrome (STEVENS N 83 LTT, 1983)
- Library Technology in the '70s (WHITH 73 LTI, 1973)
- Linking Library Automation Systems in the Internet - Functional Requirements, Planning, and Policy lssues (LAN, 1996)
- Möglichkeiten und Grenzen elektronischer Datenverarbeitung in Bibliotheken und lnformationssystemen (Bibliothek EDV, NIEDERMEYR W 83 MUG, 1983)
- Maintenance of Automated Library Systems (Bibliothek EDV Wartung, 1983)
- Maschinelle lnformation und Dokumentation am Beispiel der Österreichischen Historischen Bibliographie (Bibliographie EDV, BIBL EDV, HÖDL G 81 MIU, 1981)
- Maschinelles Austauschformat für Bibliotheken Version 1 (MAB1) (MA B1, 1978)
- Microcomputer-Based Bibliographic Information Storage and Retrieval System (Bibl EDV, KARARIA K 81 MBB, 1981)
- Microcomputer Hardware and Software in Libraries in the Nordic Countries (PC Software Bibliotheken, 1986)
- Microelectronics And The Communications Revolution (PC Kommunikation, 1983)
- Mikrocomputer der Universitätsbibliothek Braunschweig druckt jetzt auch Katalogkarten (Katalogkarten EDV, 1983)
- Mikrocomputer in Bibliotheken - Erfahrungen der Herzog August Bibliothek Wolfenbüttel (Biblgrf Kleincomp, 1984)
- Minimalkatalogisierung und Katalogkonversion (Minimalkat. 1991, 1991)

- Modular und intelligent - das Terminal 6730 von Siemens (SIEMENS 78 TERM 6730, 1978)
- Neue Medien und Technologien in Wissenschaftlichen Bibliotheken (Trends Neue Medien, 1985)
- New media, new messages - innovation through adoption of hypertext and hypermedia technologies (Hypermedia, 1990)
- Nixdorf Plattensysteme mit Lesestift für das Bibliothekswesen (NIXDO 73 NPM, 1973)
- OCR SCANNING (OCR 80, 1980)
- ON-LINE-AUSLEIHVERBUCHUNG IN EINER MAGAZINBIBLIOTHEK (FELLU 70 OLA, 1970)
- ORGANISATIONSVORSCHLAG für eine EDV - Unterstützung beim ZENTRALKATALOG
- On-Line Bibliographic Services Selected British Experiences (KIDDJ 77 OLB, 1977)
- Online-Bestellungen von Bibliotheksdokumenten (Biblio EDV 1991, 1991)
- Organizing Babylon (UNICODE, 1996)
- Persönliche Bibliothek (PC Biblio 1992, 1992)
- Probleme des Online-Benutzerkatalogs (OPAC, 1989)
- Problems of a central cataloguing service (WELLA 71 POA, 1971)
- Prozesse der Bibliothek der Montanistischen Hochschule Leoben
- REMARC (REMARC, 1983)
- RETROSPEKTIVE KATALOGISIERUNG (RECON, 1993)
- Realistische Erwartungen an den Computer-Einsatz im Bibliothekswesen (ROTTG 71 REA, 1971)
- Retrospective Conversion (RECON, 1987)
- Round Table - Regeln für maschinenlesbare Dokumente (1994)
- SABRE - a novel software tool for bibliographic post-processing (Bibliographie EDV 1989, 1989)
- SOME CHARACTERISTICS OF FUTURE INFORMATION SYSTEMS (Infosys Trends, 1983)
- Scanning für Altdatenumsetzung in maschinenlesbare Katalogisate (1990)

- Shifting Gears - Information Technology and the Academic Library (De Gennaro, Zukunft Bibliotheken, 1984)
- Some Organizational Prerequisites for the Introduction of Electronic Data Processing in Libraries (BERNH 71 SOP, 1971)
- Special Libraries and the Development of a Canadian Library System and Network Policy (MACLH 74 SLA, 1974)
- Strategisches Informationsmanagement in Bibliotheken (Infomanagement, 1991)
- Systembeschreibung FRIDEN 4300 Magnetband-Datenerfassungssystem (FRIDE 70 MDF, 1970)
- THE OSTI-SUPPORTED LIBRARY AUTOMATION PROJECTS (OVERC 73 TOS, 1973)
- Techniken und Methoden wissensbasierter Systeme - Expertensysteme in Bibliothek, Information und Dokumentation (1995)
- Test BIBOS (1987)
- Test McDonnell Douglas - URICA (1987)
- The British Library's Catalog (1995)
- The British Library automated information service (BLAISE BIBEDV HOLMP77TBL, 1977)
- The Impact of Automation on Professional Catalogers (EDV Katalogis, 1990)
- The Library of Congress: Our Master Or Servant in the Network Age? (GOODC 74 TLO, 1974)
- The Role of Mini-Computers in Libraries (PC Bibliotheken, 1983)
- Toward a Global Library Network (1994)
- Verbundsysteme und Scanningverfahren (1995)
- Von der Literatursuche bis zur Ausleihe (1975)
- Von der Lochkarte zum Online-Informationssystem
- Vor Beginn der Datenverarbeitung in (wissenschaftlichen) Bibliotheken (STOLTZENBURG 83 VBD, 1983)
- Wahl eines Bibliothekssystems für die Schweizerische Landesbibliothek (1993)
- What Every Librarian Should Know About Automation (PARKR 64 WEL, 1964)
- What is the Internet? (Document Categories 1992, 1992)

- When Some Library Systems Fail (Bibliothek EDV, BIBEDV Fehler, 1971)
- Wird Weltinformationssystem Wirklichkeit (WIRDW 71 W, 1971)
- Zur Nachfolgefrage eines Verbundsystems der österreichischen Bibliotheken aus der Sicht funktionierender und oder geplanter lokaler Systeme (1996)

12.3 Wiki keywords

The following keywords are available for the semantic search at `http://www.ocg.at/informatisierung/index.php5?title=Spezial:Semantische_Suche`:

- ALEPH
- application
- automation
- bibliography
- BIBOS
- catalog
- computer
- conversion
- data
- development
- documentation
- information technology
- format
- library
- LOC
- management
- MARC
- OCLC
- OCR
- OPAC

- organization
- PICA
- problem
- process
- project
- REMARC
- report
- requirement
- software
- statistics
- system
- tool
- user

Bibliography

[1] Forum Zeitgeschichte. http://www.univie.ac.at/universitaet/forum-zeitgeschichte/. [online; accessed January 29, 2013].

[2] Semantic MediaWiki. http://semantic-mediawiki.org/. [online; accessed April 30, 2013].

[3] Alternativeto.net. Alternativeto.net. http://alternativeto.net/. [online; accessed April 09, 2013].

[4] Caroline R. Arms. Keeping Memory Alive: Practices for Preserving Digital Content at the National Digital Library Program of the Library of Congress. *RLG DigiNews*, 4(3), 2000.

[5] Federal Agencies Digitization Guidelines Initiative (FADGI). Technical Guidelines for Digitizing Cultural Heritage Materials. Technical report, 2010. http://www.digitizationguidelines.gov/guidelines/digitize-technical.html. [online; accessed January 29, 2013].

[6] Karl Fröschl, Siegfried Mattl, Johann Stockinger, Werner Kläring, and Wilfried Schöfer. Die Informatisierung Österreichs. http://www.ocg.at/informatisierung/index.php5. [online; accessed January 29, 2013].

[7] Karl Fröschl, Siegfried Mattl, Johann Stockinger, Werner Kläring, and Wilfried Schöfer. Die Informatisierung Österreichs - Projekt Team. http://www.ocg.at/informatisierung/index.php5?title=Projekt_Team. [online; accessed January 29, 2013].

[8] Gail M. Hodge. Best Practices for Digital Archiving. *D-Lib Magazine*, 6(1), 2000.

[9] Dublin Core Metadata Initiative. Dublin Core Metadata Initiative. http://dublincore.org/. [online; accessed March 31, 2013].

[10] Rainer Kuhlen. *Hypertext. Ein nicht-lineares Medium zwischen Buch und Wissensbank.* Springer, 1991.

[11] George P. Landow. *Hypertext 3.0: Critical Theory and New Media in an Era of Globalization.* Johns Hopkins University Press, 2006.

[12] Library of Congress. About the American Memory Project. http://memory.loc.gov/ammem/about/index.html. [online; accessed February 17, 2013].

[13] Library of Congress. American Memory Project. http://memory.loc.gov/ammem/index.html. [online; accessed February 17, 2013].

[14] Library of Congress. American Memory Project FAQ. http://memory.loc.gov/ammem/help/faq.html. [online; accessed February 24, 2013].

[15] Library of Congress. National Digital Library Program. http://memory.loc.gov/ammem/dli2/html/lcndlp.html. [online; accessed February 07, 2013].

[16] Library of Congress. Personal Archiving. http://www.digitalpreservation.gov/personalarchiving/. [online; accessed January 31, 2013].

[17] Library of Congress. Teacher Resources. http://lcweb2.loc.gov/ammem/ndlpedu/. [online; accessed February 17, 2013].

[18] Library of Congress. Technical Standards for Digital Conversion of Text and Graphic Materials. Technical report, 2006. http://memory.loc.gov/ammem/about/techStandards.pdf. [online; accessed January 31, 2013].

[19] Microsoft. File System Algorithms. http://msdn.microsoft.com/en-us/library/ff469400%28v=prot.20%29.aspx. [online; accessed March 03, 2013].

[20] Helmut M. Niegemann. *Kompendium multimediales Lernen*. X.media.press. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[21] Library of Congress. MARC Standards. http://www.loc.gov/marc/. [online; accessed March 31, 2013].

[22] Library of Congress. National Digital Newspaper Program. http://www.loc.gov/marc/. [online; accessed March 31, 2013].

[23] Rolf Schulmeister. *Grundlagen hypermedialer Lernsysteme. Theorie - Didaktik - Design.* Oldenbourg, 4. auflage edition, 2008.

[24] Ewald Terhart. *Didaktik.* Reclam, 2009.

[25] Forum Zeitgeschichte. Universität im 20. Jahrhundert - Projektliste. http://www.univie.ac.at/fileadmin/user_upload/forum-geschichte/Aktuelles/Projektliste.pdf. [online; accessed January 30, 2013].