Abschlussarbeit
Universitätslehrgang Library and Information Studies
Jahrgang 2016/17

HOFRAT MAG. RAINER STOWASSER, DIPL.TONM.

SEBASTIAN GABLER

SOW: DIGITIZATION AND LONG TERM PRESERVATION OF WEATHER MAPS AT ZAMG

UNIVERSITÄTSLEHRGANG LIBRARY AND INFORMATION STUDIES 2016/2017

PUBLISHED BY UNIVERSITÄTSLEHRGANG LIBRARY AND INFORMATION STUDIES 2016/2017

First printing, September 2017

Contents

This work was created with LaTeXand the Tufte Book Style [1].

[1] [Tuf16]

I want to thank my wife for her patience and my sons for letting Papa work.

Rainer Stowasser

I want to thank my wife for her patience and my sons for letting Papa work.

Sebastian Gabler

Introduction

current state

The daily weather maps at ZAMG start with 1877.Some items are starting with 1861, but records do not exist consitently. The geographical area covers Europe and Austria in its historical borders. The weather parameters are collected in tables. Already on the next day the description is part of the climate coverage. [2]

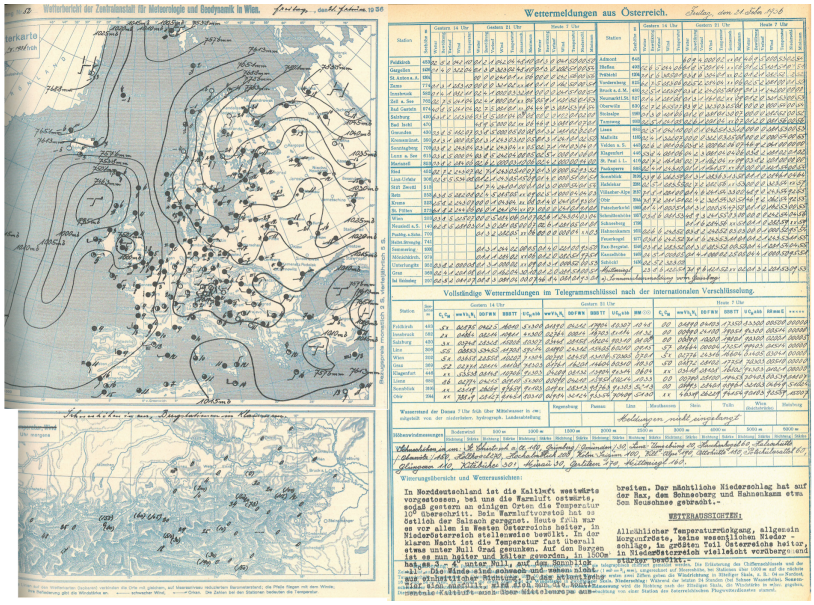


Figure 1: Weather map 21. Februar 1936

[2] Example; the map on the upper left is the special focus of this project. The parameters of the Austria stations are already in a database.

The maps produced untill 1986 where hand drawn with ink-pen on a template. They are unique copies. Every day the person in charge produced the map. Once a year they where bound in book form. Over the time the format changed. However, all objects contain a map and table with the parameters of the stations (obtained over the telegraphic network), isobars, the frontal activity area and[3] comments on the daily weather. Through its long coverage this informations is important for climate research. The fact that during 1942/43 due to war activities some pieces where lost states that its very important to make sure that this inventory is preserved and actions have to be taken to ensure that future generations can use the information.

[3] hand written

The guidelines of the World Meteorological Organisation (WMO) instruct that the data protection should be digital due to the data handling in computer systems.[4] For that analog resources are to

[4] [et.10]

be converted into image files, described with meta data and saved for long time preservation. To ensure operability this digitalisation has to guarantee that in all parts of the data life cycle there are precautions to prevent destruction and obsolescence. This requires action cause these are not easily noticeable for digital objects and just storing is not enough. So preparations have to be taken best starting when the first electronic copy is made.[5]

[5] [Har12]

Required activities:

- weather maps are digitized (scanning process)
- the files are long time preserved
- view and download services are implemented.

The target of this concept are delivering a catalog of requirements, the evaluation of tools and possible file formats (e.g. FITS) necessary (proof of concept) and to estimate the resources needed for an operational program.

objectives

digitation of weather maps
NOT part of the project, just an estimation

As infrastructure, a Ao flat bed scanner and a digital camera with a tripod were available. The pages are to be digitized and converted to FITS.

The Flexible Image Transport System (FITS) is a free format constructed by NASA for astronomical Observations. It is used for research data and by the Vatican Library for the digitalisation of their Objects.

Why FITS ?

→ long time preservation

Its specification allows to describe pixels individually and as a text file is highly compressible.

→ Container

Its possible to include Metainformation directly and to use a file as an independent source.

→ Freeware

All tools necessary are open source.

The daily maps are ordered in individual files and compressed (storage space) .

long time preservation
NOT part of the project.

For the long time preservation infrastructure the OAIS Standard [6] will be used. The concept is part of the project.[7]

[6] OAIS standard, 2012. ISO-14721

[7] Sub Topics:

- Specification of the requirements for the subject of daily weather maps.
- Definition of the Submission Information Package (SIP) as part of the api to the archive.
- Definition of the Metadaten (descriptive, significant properties, provenance, fixity)
- Definition of the data model and possible workflow

This Information shall be developed using the PREMIS 3.0 Standard and documented for further use. The possibility for validation of the representation will be checked with the PRONOM Database (National Archives UK).

The files shall be accessible through a web service.

View und Download Services

NOT part of the project, but specification of the infrastructure.

The catalog and the tools in the DMZ are to be defined in compliance with the IT Services at ZAMG. A link to the parameters in the existing databases has to be tested.

Framework

Theoretical context

In the literature (e.g.[8]) there s a lot of discussion about the need to make sure that resources on paper are preserved and that a conversion to electronic resources is one of the possible ways.

For austria over the years different suggestions where made (e.g.[9] and for an overview for the Austrian National Library see [10]) but to our knowledge in practice [11] the required skill-set and tools are not off-the-shelf resources. Hence, every project needs to find its own path.

[8] [Kuh14]

[9] [BK15]

[10] [Nat15]

[11] [eiP16]

library context

The library at ZAMG is a special library for research purpose and a special unit in the public sector as ZAMG is a governmental institution .

With its foundation in 1841 the "Zentralanstalt für Meteorologie und Erdmagnetismus" started the collection of books and journals with the addition of the private libraries of the directors that contained antiquarian objects so the portfolio spans over 200 years.

The library is located since 1972 at the Hohe Warte in Vienna, Karl-Kreil house and is positioned over 4 floors under ground.

Untill 1994 there was a joined library with the Institute of Meteorology at the University of Vienna. After the seperation most of the more recent publication where integrated in the inventory of the university and the historical holdings stayed at ZAMG. Due to this "the archive" is the main part (about 82.000 objects, more than 60.000 before the second world war).

The catalog was a digitization (re-typing) project of the old index cards and done over years by one Person (Eva Miklos) using a proprietary software specially programmed for ZAMG. As this infrastructure is out of date a conversion to Koha[12] is on the way.

[12] [Koh16]

In recent years the weather parameters documented in the books and papers where used to establish databases for the historical periods with the main focus on Austria and climate research.

To ensure consistency and reproducible research the next step is to document the sources of the data and the objects containing the information.

Content

In meteorology the "weather" was documented over centuries starting with the main parameters temperatur, precipitation and later barometric pressure. But only in very few spots this was done in a systematic way over a long time period . Thus the reconstruction of the historical climate has to use a multiplicity of sources. Starting in the mid of the 19 century with the establishment of the national weather services and the telegraphic network to exchange data it was possible to get a picture over broader areas. For climatic questions this is a very short time span but much better than the climate reference period (1961 - 1990 and 1976 -2005).

milenia if you account for the records in astronomical facilities

e.g. Kremsmünster since 1762

ZAMG being one of the oldest

With the foundation of the WMO standards where applied and data was shared internationally. However, the systematic acquisition was only performed on selected observation stations.

Publications and observation reports are a tremendous resource that could be used. These have to be gathered, recorded, and transformed into information usable for scientific research [13].

[13] [WMO15]

historical Weather data and maps

All national weather services have records of historical weather conditions, but as the "weather of tomorrow" was always more important than the "weather of yesterday" this documentation was not regarded as high priority. With the discussion about climate change, this has changed.

The climate archives focused from the beginning on using data already digital available and then on recovering parameters on paper and typing them into databases. In very few cases the documentation of the original sources was part of this projects..

paper source

But with using this data (e.g. for climate modeling) it became clear that in many cases it was necessary to go back to the originals as the electronic data did'nt make sense (typing errors, file format errors ...). Only after the scans became available, gaps could be easily identified and corrections to the time series could be applied (if the scans where available cause not all of them where archived and shared openly).

Here are some examples of the strategies applied, and a short list of institutions or scientific communities (e.g.[14]) involved. To our knowledge, there are no examples, where long time preservation was undertaken.

[14] [et.16b]

International

Mainly scientific communities started to document parameters into time series. For an overview see [15].

[15] [oEACRU16]

Great Britain

The Met Office provides an archive with daily weather data starting 1860 [16]. It contains pdf-scans and the records. A download is possible but further use of the files is limited due to the low quality of the scans.

[16] [GB10]

Germany

The german weather service (DWD) provides historical weather parameters as grid data monthly with an ftp-service[17] but only few parameter cover a longer time span (e.g. temperature since 1901).

[17] [DWD16]

Swiss

Meteo Swiss has included data of selected stations in an special infrastructure the Swiss National Basic Climatological Network[18] where parameters can be used freely.

[18] [Sch16]

Austria

Historical Data was digitized in many projects at ZAMG. The "Klimabogen" is available as pdf-scan but the originals are only accessible internally. The parameters where processed and are publicly available over a portal (maps since 1971 [19], Monthly Data since 1760).

[19] [ZAM16a]

Due to "data homogenisations" and quality control the measured data is corrected and different "states" of data is available (raw, automatically checked, certified and homogenized) where every step is documented. The question how to "cite" this data correctly are in discussion [20].

[20] [ea16, DCC11, 1114]

The daily weather maps are a next step (but also the seismological archive has to be tackeld).

With the foundation of the CCCA Datacenter[21] in 2016 theres an infrastructure for climate data which is filled with data sets available.

[21] [Aus16]

Records management and long time preservation

With a shelf-life exceeding 150 years, the oldest items of the original resources have survived the re-location of "Centralanstalt" from Wieden to Hohe Warte of 1872, several re-organisations of the parent organisation, not less than six polity changes of Austria between 1867 and 1954, and two World Wars. As the respective resources represent a product of ZAMG's statutory programme, the onus of records management is imposed by the legislation (§22 FOG [22]).

[22] [FOG]

The statutory archive of the ZAMG's records however are the Austrian State Archives (§3 BAG [23]) .The timely and procedural constraints of records management and archiving are defined in the Bundesarchivgutverordnung [24]. Once the resources are no longer required for the statutory operations, they would be submitted to the Austrian State Archives. The minimum records keeping period is seven years, beginning with the operational obsolescence.

[23] [BAG]

[24] [BAV]

Records Management and Archiving are two different fields of action. However, these fields cannot be clearly separated in this case: The original paper resources are being kept in the ZAMG library for their entire life time of up to 150 years as relevant (and active) business records. For the more (fr)agile form of their digital siblings,more intensive access is expected. As even the least record keeping period of seven years is within the horizon of obsolescence of IT-systems (let alone the 150-year operational precedence from history), long-term digital preservation methods have to be applied for the storage from the outset.

The Open Archival Information System (OAIS) reference model [25] is the prevailing methodology in the field of long-term preservation. Mainly coined for digital preservation, OAIS is applicable for analogue and digital resources. For many years, OAIS is established in Austria. Specifically, the PHAIDRA service [26] originating from the Archive- and Library Service of the University of Vienna, highly influential in Austria and the Alpe-Adria Region [27], and the federal digital long-term archive operated by the Bundeskanzleramt, Austria [28] should be mentioned in this context.

[25] [CCS12]

[26] [phaa]

[27] [phac]

[28] [Ös17]

Guidelines specific to the meteorological domain have been examined as well. The methods suggested by the WMO for the safeguarding of digitized climate data are mainly object- and bitstream preservation strategies. I.e., WMO suggests the storage of reformatted climate data on stand-alone media like hard disk or CD/DVD, the use of systematic naming conventions, and the production of

backup copies. [29] These are certainly basic measures that may be included in almost any preservation practise, however they do not suffice as a long-term strategy. According to Harvey: *"Although it is necessary for all digital preservation strategies, bit-stream copying is not a long-term strategy because it does not address the key factors that cause digital deterioration, mainly obsolescence of hardware and software."* [p 142][30]. Moreover, the methods suggested by the WMO may be difficult to sustain for a lot of tens of thousands of documents, as it is for the target collection.

[29] [Gro]

[30] [Har12]

OAIS in turn, pursues an information-centric preservation approach that is not limited to preserving physical or digital objects. The architecture endorses a holistic approach to preservation, combining several strategies mentioned by Harvey, such as encapsulation, long-term formats, normalization, persistent object preservation, and policy development. [p 101][31]

[31] [Har12]

The environment model of an OAIS Archive describes the main functionalities and stake holders

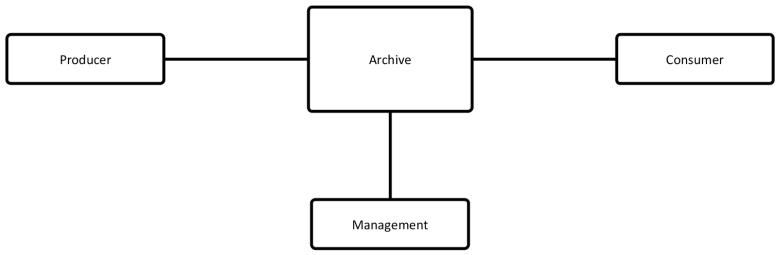


Figure 2: OAIS environment

Producer is the role taken by those persons or resources who provide information to be managed by the archive.

Consumer is a role played by those persons or systems which access the archive for finding and acquiring information of their interest.

An OAIS is specifically providing for a special class of Consumers, the so-called "Designated Community". *"The Designated Community is the set of Consumers who should be able to understand the preserved information."* ([32] 2-1) The ability of persons, institutions, and systems to adopt multiple roles is one of the features that enable the interoperability of an OAIS.

[32] [CCS12]

Specifically, one OAIS may act as Producer or Consumer of another. In the present case, the producer and the OAIS both are under the umbrella of the same parent organization, Zentralanstalt für Meteorologie und Geodynamik. When it comes to an eventual post-business archiving of the records, the ZAMG's OAIS might assume the role of the producer for the Federal Archives.

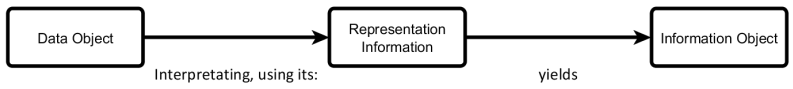
Management is a role that sets the strategic policy for the archive. It is a responsibility in context of global strategy of the parent organization. Management does not entail the day-to-day operations of the archive. For this, the OAIS provides a specific functional entity. Other activities for Management include:

- Source of funding and setting price policies for access
- Setting policies for resource utilization
- Perform regular review processes and risk assessment
- Take a referee role resolving potential conflicts between producers, consumers, and administration.

The OAIS management roles are partially assigned by legislation, i.e. funding and pricing policies are already set by FOG. Other parts may currently best be assigned to the ZAMG library, i.e. review and risk assessment, whereas endorsement for top-level service policies would be required from the ZAMG management.

The concept of Information is essential to the preservation abilities of an OAIS. Schematically, OAIS has a three-tier conceptual approach to information:

Figure 3: OAIS information model



Information is knowledge that can be exchanged between sender and recipient, expressed in some form of data. Representation information may be provided implicitly in the common knowledge base of sender and recipient, or it may be provided explicitly. The extend of representation information required for the Designated Community to understand the information is subject to agreement, and may change over time. However, the minimum set of representation information required to ensure transparent preservation from the information object down to the bitstream needs to be provided. The recursive character of representation information cannot go unmentioned: specifically, high-level representation information consists of data object and representation information. This may be clear for the example of a grammar book. However, it is also valid for every building block of an information technology system utilized for preservation, i.e. hard drive, CPU, file system, and operating system, to name but a few.

Information in an OAIS is organized in information packages. Conceptionally, an information package involves four types of information:

- Content Information
- Preservation Description Information (PDI)
- Packaging Information
- Descriptive Information

Each of these types consist of data and representation informa-

tion. An information package actually contains Content Information and PDI. Packaging information provides the encapsulation and identification information for these data. Descriptive Information provides the means to discover an information package.

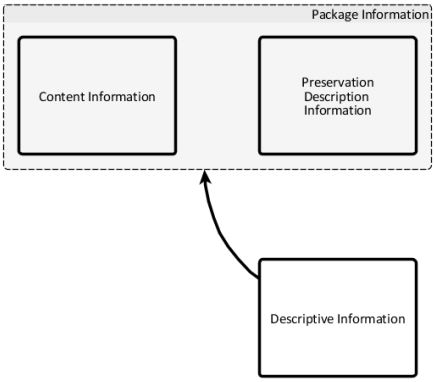


Figure 4: OAIS Information Package model

The definition of PDI cannot be done more concise than by OAIS own words:

"The Preservation Description Information applies to the Content Information and is needed to preserve the Content Information, to ensure it is clearly identified, and to understand the environment in which the Content Information was created. The Preservation Description Information is divided into five types of preserving information called Provenance, Context, Reference, Fixity and Access Rights. Briefly, they are the following: (33 2-6)

33 [CCS12]

- *Provenance describes the source of the Content Information, who has had custody of it since its origination, and its history (including processing history).*
- *Context describes how the Content Information relates to other information outside the Information Package. For example, it would describe why the Content Information was produced, and it may include a description of how it relates to another Content Information object that is available.*
- *Reference provides one or more identifiers, or systems of identifiers, by which the Content Information may be uniquely identified. Examples include an ISBN for a book, or a set of attributes that distinguish one instance of Content Information from another.*
- *Fixity provides a wrapper, or protective shield, that protects the Content Information from undocumented alteration. For example, it may involve a checksum over the Content Information of a digital Information Package.*
- *Access Rights provide the terms of access, including preservation, distribution, and usage of Content Information. For example, it would contain the statements to grant the OAIS permissions for preservation operations, licensing offers (for distribution), specifications for rights enforcement measures, as well as access control specifications.*

OAIS has a clear understanding that the information packages submitted to the archive by producers, those preserved by the archive, and those disseminated to the Consumers may have a different structure. Therefore, different types of information packages are defined. The packages preserved by an OAIS are called Archival Information Packages , which contain the full PDI and representation information. AIPs are produced from Submission Information Packages , i.e. by transformation and completion. Dissemination Information Packages are provided to a Consumer as response to a request. They may contain one or many AIPs, and they may contain different representation information, and/or only a subset of the PDI .

AIP

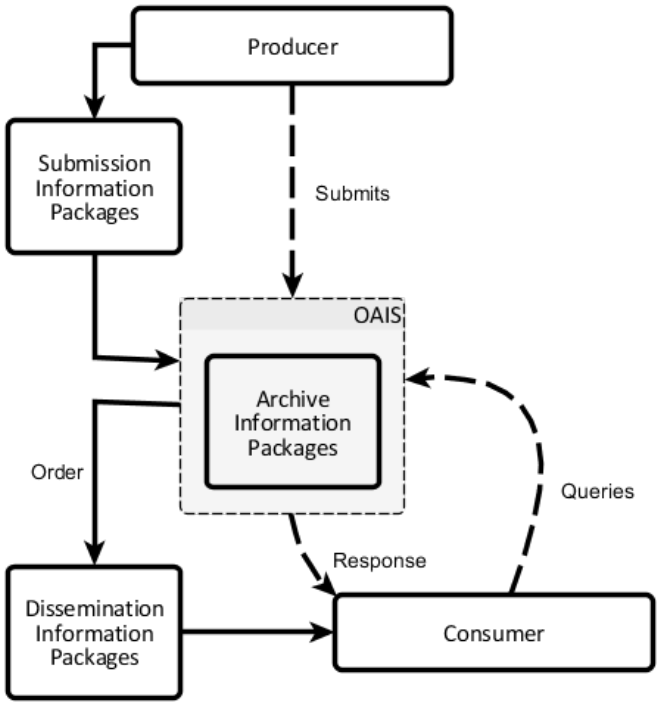
SIP

DIP

Preservation Description Information

The above definitions form the foundation of the internal and external interactions with an OAIS. An illustration of the external high-level interactions is best suited as a summary.

Figure 5: OAIS high-level interactions



As initially hinted, reformatting of paper resources is raising essential questions when it comes to the long-term safeguarding of holdings. For paper resources, a passive preservation strategy (inaction, or benign neglect) may both common and appropriate. *"The major focus for preserving this [analogue] information has been to ensure that they are on media with Long Term stability and that access to this media is carefully controlled."* ([34] p.2-1)

[34] [CCS12]

Digital objects require a pro-active preservation strategy entailing the entire life cycle. As Ross Harvey puts it: *"One thing we understand about information in digital form is that actions must be applied almost from the moment it is created, if it is to survive."* (p 8 [35])

[35] [Har12]

Much of the information required to support long-term preservation strategies is more conveniently or only available at the time when the original information was produced. Specifically, authenticity of a digital record only can be assessed with a seamless chain of trust in place. This may be achieved by producing and maintaining Fixity information at the production of the digital originals. For the preservation strategy of the ZAMG weather maps this means that the long-term preservation begins within the re-formatting process. *"Participation in these efforts will minimize the lifecycle costs and enable effective Long Term Preservation of the information."* ([36] 2-1)

seamless chain of trust

[36] [CCS12]

The re-formatting process outlined in this paper includes the specification of PDI generation, specifically Fixity and Provenance information. Context information is provided by existing bibliographic metadata.

To this date, ZAMG doesn't dispose of an infrastructure for digital preservation. Under the umbrella of the re-formatting of the weather maps hands-on experience may be obtained which may be useful for different business procedures likewise.

The main modules of an OAIS compliant repository are ([37] 3-6, ff):

[37] [CCS12]

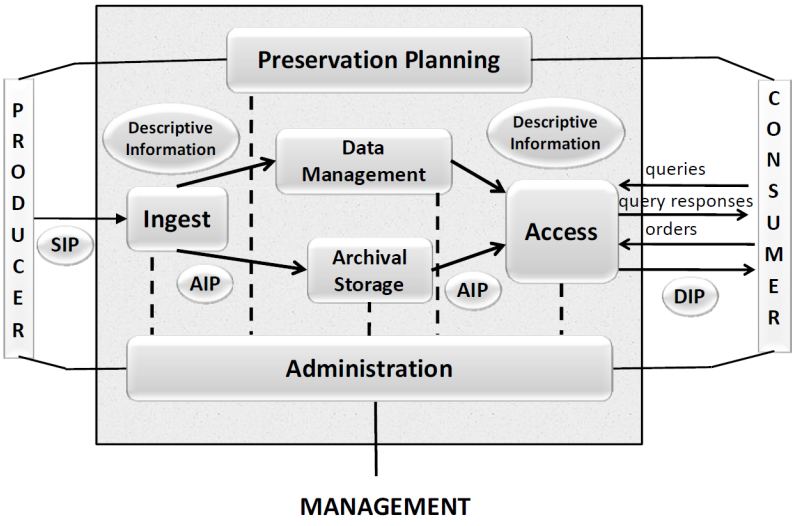


Figure 6: OAIS overview, source: CCSDS

- Ingest Functional Entity (labeled 'Ingest' in the figures in this section)
- Archival Storage Functional Entity (labeled 'Archival Storage' in the figures in this section)
- Data Management Functional Entity (labeled 'Data Management' in the figures in this section)
- Administration Functional Entity (labeled 'Administration' in the figures in this section)
- Preservation Planning Functional Entity (labeled 'Preservation Planning' in the figures in this section)
- Access Functional Entity (labeled 'Access' in the figures in this

section) ([38] 3-6, ff)

[38] [CCS12]

A substantial motivation for operating a subsidiary repository is are the specific user profiles required by the WMO. These are fundamentally different from those currently offered by the Austrian State Archives, and from those offered by a typical university library. However, they may be set up with acceptable effort in a separate system.

ZAMG's own repository?

The market for OAIS- compliant standalone repository systems is a niche. There are many products which claim to be OAIS compliant, starting from the Oracle Database, the IBM Tivoli storage system, and Oracle Storage Works products (See, among other [39]) . Further investigation however clarifies that such claims may be helpful when using these products as building blocks for a repository, but it often doesn't entail support for all OAIS functional modules. We have chosen three different systems which have significant footprint in the market, among those two commercial (proprietary) systems, and an open source platform. All three systems offer support for the OAIS modules Ingest, Access, Preservation Planning, and Administration, which we in short summarize as "complete OAIS support". All systems need additional storage, database, and computing services, so that the implementation of a repository system is substantially complex. The Fedora platform is mainly a middle ware, focussed on repository- and integration services which complement with a separate application framework.

[39] [Ora16]

All systems chosen support Open Access Publishing of the resources (OAI-PMS service) and API-based web services (i.e. REST, SOA). All systems support Dublin Core, METS, MODS, and PREMIS for metadata, and thus allow for signing off metadata requirements on a high-level basis.

Name	License	Complete OAIS support	Access service framework included
Fedora	Apache 2.0	x	Requires application framework
Preservica Works	Proprietary	x	x
Proquest Rosetta	Proprietary	x	x

As we are aiming at an industrialized digitization project, we would be interested in a repository system that allows for workflow management, including a Business Procedure Modelling module that is aware of the parameters required, and the data produced. Some of the platforms investigated include this.

Fedora

Fedora project has started in 1997 as a scholar project of Cornell University. The name is an acronym for Flexible and Extensible Digital Object Repository Architecture. Fedora essentially provides

a repository middleware layer that complements to an application layer (or framework) such as Phaidra, Hydra, or Islandora.

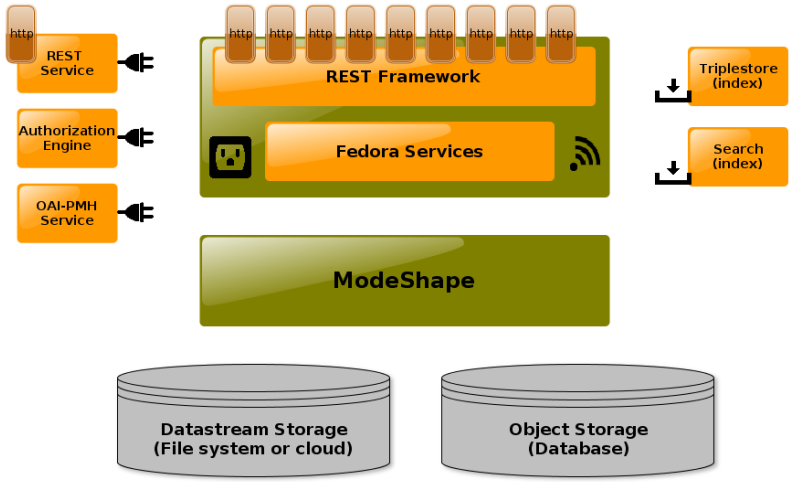


Figure 7: Fedora 4.x Architecture

The current major version 4.x of FEDORA [40] was released in 2015. It brings fundamental changes in respect to the object model, hierarchical file system support, and path-based storage management. Replacing the proprietary Fedora XML(FOXML) [41] object model by ModeShape [42] is probably the most significant change. ModeShape content objects are built on an industry-standard, the Java Content Repository (JCR) API. JCR is supported by Oracle, Adobe, HP, IBM, and SAP, among other. JCR is content-agnostic, and it functions as a federated repository, rather than as a silo. JCR joins functionalities which are required for the handling of structured and unstructured content, i.e. in combination of databases, file systems and repositories:

"The JCR 2.0 API provides a number of information services that are needed by many applications, including: read and write access to information; the ability to structure information in a hierarchical and flexible manner that can adapt and evolve over time; ability to work with structured, semi-structured, and unstructured content; ability to (transparently) handle large strings; notifications of changes in the information; search and query; versioning of information; access control; integrity constraints; participation within distributed transactions; explicit locking of content; and of course persistence. [43]

Introducing ModeShape brings a switch from XML objects to RDF nodes, and thus a switch to graph-based data management from XML objects. This entails support for stream-based storage platforms, and graph-based indexing, facilitating new application scenarios based on W3C Semantic Web standards, and using cloud architecture for operations.

FEDORA seems to be a promising track to follow for the realisation of a repository service. The new features in version 4 are attractive for building a new solution today. A short summary of

[40] [FED]

[41] [dura]

[42] [mod]

[43] [int]

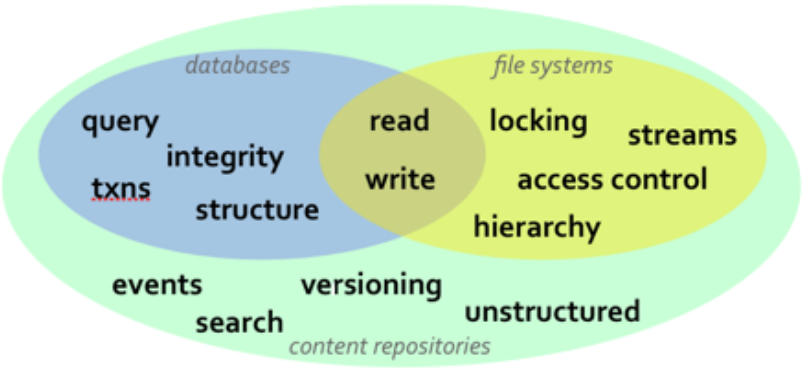


Figure 8: JCR functional overview http://docs.jboss.org/modeshape/2.6.0.Beta2/manuals/reference/html/jcr-features.png (Copyright: RedHat Inc.)

the main aspects of a Fedora 3/4 migration can be found on the Dspace Wiki [44]. Specifically, the knowledge management infrastructure building on graph databases is promising, which falls within the current and future requirements of climate data management.

[44] [durb]

Phaidra

PHAIDRA is an open-source project initiated and maintained by the University of Vienna. The platform already has quite a number of adopters from Austria and surrounding countries [45]. PHAIDRA is a toolkit- and application layer based on Fedora Repository, currently limited to FEDORA 3. Specifically, there is an mod- perl- (Apache http-deamon) / Catalyst-MCV based application [46], an upload client called "Phaidra Importer", and an API. The PHAIDRA project Github page is the only publicly available resource [47] . The information provided there however rather seems to reflect special interests of the core project team instead of informing potential adopters. There are as well no obvious activities from the community adopters on the project page, nor is there any visible interaction with the Fedora core team. PHAIDRA supports specific application scenarios, which are:

[45] [phac]

[46] [phad]

[47] [phab]

- E-book publishing (Phaidra Importer)
- Book viewer
- Audio- and video streaming
- Solr- based search

As formats, PHAIDRA supports a focussed choice of object types, which are: Picture, Video, Audio, and Documents. For each object type, 2-3 file formats are supported as "recommended" or "optional". For raster images, only TIFF and JPEG are supported. Other formats may be used, but they will be only available "as is", without any conversion or publishing services.

Fedora SWOT

Strengths	Weakness
• Complete feature set for basic requirements • Mature project • Academic background • Active and significant community • Significant user base • Based on technology standards • Full OAIS support • Standards-based APIs • Open Source license • Cloud technology support	• Significant integration effort
Threats	**Opportunities**
• Complexity • Knowledge-intensive platform • Drupal change management	• Comprehensive solution for a repository service • Adaptation effort limited to niche functions • Independence from a specific manufacturer • Networking with the relevant developer community • Networking with the relevant user community • Acquisition of ZAMG-specific knowledge

Phaidra SWOT

Strengths	Weakness

• Low overhead architecture
• Clearly defined services
• Good user documentation
• Austrian user base
	• State of technical information
	• Application scenarios not matching
	• Outdated backend support
	  – Support of an outdated FEDORA version
	  – Significant changes in the FEDORA object and storage model
	  – Significant changes in the FEDORA data management model
	  – Significant changes in the FEDORA client/server model
	  – Interaction with the FEDORA community is not visible
	  – No obvious quality management policies
	  – Technical team is an internal UVIE- resource (ZID)
	  – Support for raster images limited to TIFF and JPEG
	  – User base regionally limited and rather small
	  – Cloud technology support unclear

Summarizing, there are substantial risks for a repository realisation with PHAIDRA for this project. The main risks are lack of personal resources, and the outdated technology stack excluding the new key features of FEDORA 4.x, such as byte stream storage and RDF support for metadata, which are attractive. Last but not

Threats	Opportunities
• Migration risk to current Fedora versions • Adaptation effort unclear • Availability of resources for change requests • Potential change requirements on deprecated core service	• Networking with Austrian peers. • Phaidra key persons known to project staff • Personal contact may outweigh documentation deficits • Possible convergence with UB-Maps project

least, it should be mentioned that the current application scenarios for PHAIDRA have not very much in common with those required for the ZAMG, and this project specifically.

Hydra

Hydra is a popular application framework for FEDORA, based on Ruby. The project seems to have arrived at the end of its life cycle. It is currently being replaced by the Samvera platform.

Figure 9: Deprecation note for Hydra, https://wiki.duraspace.org/display/hydra/The+Hydra+Project

With one project fading out, the new one not fully in place, we have decided to skip an analysis for this framework.

Islandora

Islandora is currently the most comprehensive application framework for FEDORA. It is very well integrated in the FEDORA community, with the documentation available from the same place as FEDORA. There is full support of the current FEDORA release and (limited) support for the deprecated F3 versions. It integrates into the popular Drupal general-purpose CMS ecosystem [48], bringing reviewed Drupal modules. A further pillar is the so-called integration layer, bringing interfaces for microservices [49], and JMS interfaces i.e. for the JPEG 2000 engine Kakadu, the Google OCR solution Tesseract, and the Harvard University preservation tool FITS (not to be mixed up with the raster image file format) which can be used for QC and preservation planning.

[48] [dru17]

[49] see wikipedia

Islandora SWOT

Strengths	Weakness
• Support for current and past FEDORA versions • Workflow engine included • Industry-standard user interface • Built on a popular CMS • Established quality management methods • Modular approach • Interfaces for many tools discussed specifically for this project • Wide user base • Academic user base • Excellent documentation • Good interaction with the FEDORA team • GPL license	• No existing contacts with stake holders • No known implementations in Austrian academia
Threats	**Opportunities**
• Change management requirements for CMS components • Complexity	• Basic functions already existing • Workflow engine included • Networking with international academic user base • Networking with international developers

The main risk using the Islandora frame work is the dependency to the Drupal CMS. While Drupal has an excellent track record in security maintenance, the life cycle of a Drupal main version does not extend beyond 8 years. Change management has to provide for plans that allow version migration that may affect all related business procedures. This is specifically relevant if the service is exposed on a public network, i.e. for delivery.

This part has to be evaluated carefully before exposing services. On the bonus side, Islandora currently supports past FEDORA versions, so that chances are that there are no strict dependency on layer-specific migrations also in the future. This also includes the fact that building a repository based entirely on open standards (as it would be the case with F4 and Islandora) leaves no concern about being on the safe side with platform obsolescence and lock-in scenarios.

Islandora offers excellent opportunities for the planned project, as it already includes integration for many relevant tools discussed in other parts, and specifically the inclusion of workflow modelling is highly relevant.

The fact that Drupal has a well-organized quality policy for modules is potentially helpful for the code quality as well.

Preservica Works

Preservica Works is the current commercial product, formerly known as Tesella Safety Deposit Box (Tesella SDB, the platform used for the repository of the Austrian States Archives) It is available in two editions, Cloud and Enterprise. The ability to integrate 3rd party services is limited to the Enterprise Edition which offers an optional Software Developer Kit (SDK) at extra cost.

Preservica SWOT

Strengths	Weakness
• Standard Software • Off-the shelf cloud service available • Includes user interface • Cost-efficient entry model • Includes workflow engine • Low complexity (relative)	• Extensibility limited to Enterprise Edition • Proprietary license • Format support unclear
Threats	**Opportunities**
• Long-term cost • Reliability of cloud provider • Vendor lock-in • Adaptation cost • Exit cost	• Relatively easy entry (cloud edition)

It is difficult to assert the actual viability of Preservica. The information publicly available does not allow to assert the support for the required file formats, specifically for the standardised cloud offering. As the decision if a cloud-based offering would be acceptable (basically from the political and regulatory perspective) has not been taken, we decided not to enter in a dialogue with the manufacturer within this project. It may be worth an attempt after clearing this aspect to do so.

A talk with Mag. Jonas Kerschner and Dr. Berthold Konrath, responsible for the implementation of the Digital Archive Austria at the Östereichischen Staatsarchiv showed that it would be possible to enter as a partner and use the available infrastructure as ZAMG is a public institution, but before any technical matters can be accessed the bugetary side would habe to be clear talking about a lot of money here. But not only the licensing cost is relevant but also personal resources needed to set up a project and for the public sector this is the real killer argument.

Proquest Rosetta

Rosetta is the current repository product from ExLibris/Proquest

Rosetta is the current repository product from ExLibris/Proquest. The product is standard software that requires considerable on-

premise data centre resources (storage, network, servers, workstations) for operations. The product includes a storage abstraction layer that allows the integration of read-only and read/write storage instances.

Rosetta SWOT

Strengths	Weakness
• Standard Software • Using established technologies (i.e. Oracle Database) • Access service framework included • Comprehensive storage management options	• Complimentary Proquest products required (i.e. discovery system) • Proprietary license • Primary focus in traditional libraries • Integration cost • Format support unclear
Threats	**Opportunities**
• Long-term cost • Vendor lock-in • Product life cycle unclear	• Using traditional technology

The main attraction for Rosetta is the use of traditional technology. Probably, there will not be unexpected surprises from this end. A Rosetta project still may bring substantial complexity, i.e. there are separate physical servers required for the core service alone, and there is no complete off-the-shelf offering from the manufacturer, as in contrast to Preservica. As well, the product is considerably different to the current Proquest/Exlibris portfolio (ALMA, PRIMO), which is completely cloud-based. This makes it very probable that sooner than later, the manufacturer will declare the product obsolete and replace it by a SaaS offering. The negative track record of transition from Rosetta's predecessor DigiTool (i.e. affecting the ANL) leaves room for questions about the adoption and vendor lock-in risks.

Platforms analysis summary

Implementing and operating a digital repository appears to be a challenge, and it probably will not work without involvement of external resources.

The present analysis is very much in favour of a solution based on open-source software. This is mainly because only these solutions allow for a clear perspective on the feature support required for this programme. All proprietary solutions will need adaptation, only that it is not clear to which extend, and at which cost.

The most attractive solution currently would be building a repository based on F4 and the Islandora framework. However, this approach is also the most complex.

Alternatively, the proprietary Preservica system could be investigated, once a business model for the preservation programme is in place, and if the programme would allow for proprietary and cloud-based solutions.

Metadata structure

Defining the required metadata for the project assumes considerable resources throughout the project definition and operation. The main standards used are DublinCore, Metadata for Images in XML Schema (MIX 2.0)[50], MODS, and PREMIS. As all these schemas are endorsed by the Metadata Encoding and Transport Standard (METS), is possible to use these data sources within a common METS wrapper [51]. Defining a METS profile for this project is a substantial undertaking and thus would have blown the envelope of this project. In the following, we describe the most relevant metadata fields in the preservation project, leaving aside the Administrative Metadata and Structural Metadata aspects for this report.

[50] [oC15]

[51] [oC16a]

Descriptive Metadata

For descriptive metadata, the common standard of all studied publications (WMO, Library of Congress, FADGI, NARA, OAIS) is DublinCore. This is well aligned with the feature set of the technology stack of all repository systems analysed. It is actually not necessary to include full MARC21 support within the project, and it also seems to be reasonable to limit complexity in this field. In the following table, we have combined the minimum NARA set [52] with the respective DublinCore terms:

[52] [NAR04]

NARA lable	DC Terms	Comment, values
Identifier	dcterms:identifier	Date.extension
Title/Caption	dcterms:title	Record date
Creator	dcterms:creator	ZAMG, Deutscher Wetterdienst or Allied Forces authority
Publisher	dcterms:publisher	ZAMG in all cases

It should be mentioned that the decision for DublinCore within a long-term preservation system does not limit the application of a full set of bibliographic data. The two data sets may be matched by using a common identifier in MARC21 and DublinCore.

Technical Metadata

Technical metadata primarily serve the purpose of correct rendering. TIFF embeds technical data as tags. Additionally, the Adobe

XMP system allows storage and transport of these data as Resource Description Framework (RDF)-based XML. For TIFF images, we follow the 2009 FADGI guidelines for the TIFF minimum data set [53]. Additionally, we choose the following parameters as relevant for this project:

[53] [oC09]

- Scanner Make, Model, Software: this is relevant process information. It is desirable that this data is auto-provided from the digitization process.
- Light source: The light source is to be documented for the use with reference targets.
- Color space
- ICC Profile. Using an ICC color profile is an important quality management tool, thus ICC data are to be used throughout the process.

For 'extra samples', we include this parameter to state explicitly that alpha information (transparency) is not conveyed. Scanner data allow tracking of change management information, i.e. it is possible to identify changes of model and software updates as critical parameters. At this time, we have not identified target values for all parameters. These have to be specified in a follow up. Some values will change among batches; i.e., different document sizes should render different image width and image length.

A table in the backmatter names all required parameters. The table is a minorly edited copy from [54].

[54] [oC09]

The storage of these data may be done embedded in TIFF or FITS files. Another alternative is the use of the Metadata for Images in XML schema standard (MIX) which allows to store this information independently of the resource. The version 2.0 of the MIX schema is available at the Library of Congress. [55]

[55] [oC15]

Preservation information

Preservation information is paramount to ensure the authenticity of digital information, and it is important to enable and control a quality-controlled digitization process. Preservation information establishes a context about a preserved object that remains attached to it over time. The key standard for preservation information is PREMIS, PREservation Metadata: Implementation Strategies [56]. PREMIS sets a core set of preservation-related metadata elements. It is a data dictionary that is independent of the implementation. The primary fields are provenance, preservation activity, and technical environment. In PREMIS 3, entity attributes are called Semantic Units that have a distinct meaning, and have data constraints wherever required. Thus, it is possible to validate PREMIS data against the standard, i.e. using XML schema or an Web Ontology Language (OWL) ontology. (Currently, an OWL ontology is available for the PREMIS 2.2 version [57] , PREMIS3 OWL ontologies are work in progress.)

[56] [CC16]

[57] [Con]

Premis Entity model

The PREMIS entity model describes the context of an Object in respect of the actions that have been performed on it during an Event executed by an Agent, and the rules by which this action was performed by the Agent may be formulated in a Rights Statement.

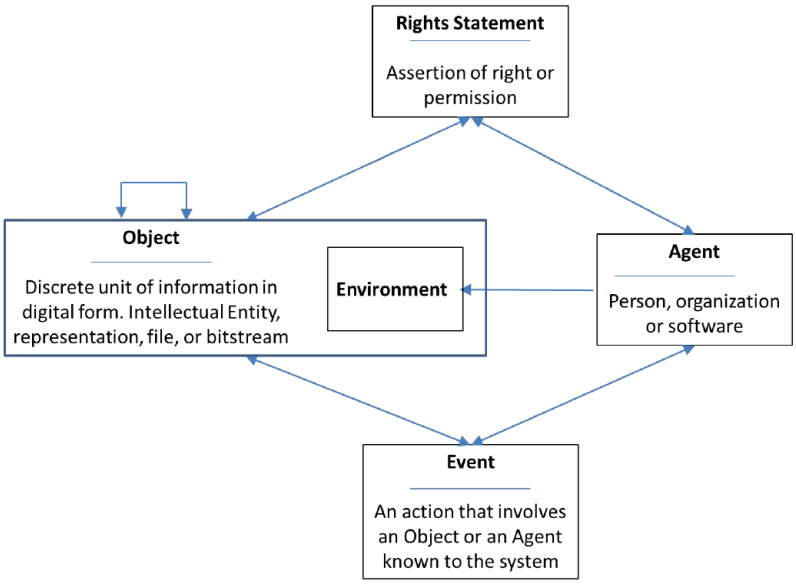


Figure 10: PREMIS data model (From: Understanding PREMIS), https://www.loc.gov/standards/premis/understanding-premis.pdf

Object entity

The PREMIS Object entity entails four different object types, File, Bitstream, Representation, and Intellectual Entity. It includes information, such as:

- Unique identifier
- Fixity information, i.e. a message digest (MD5)
- Object name, and where it is stored
- Object creation information
- Object format and structure
- Object size
- Object rendering information (i.e., which tools are required to access the information)
- The object's relation with other objects

The object type Bitstream typically refers to a subset of the File object type, i.e. we may want to describe separate Bitstream objects for the audio- and video track of an MP4 video. Bitstreams are also helpful for fixity purposes. I.e., the FITS conversion may be helped by message digest that only refer to the actual bitstream raster data, excluding the header information. This information is persistent in a format conversion.

The Intellectual Entity (IE) conceptually refers to the set of information relevant for description and administration. An IE in

PREMIS 3 may be directly described in Preservation Metadata, or outside in descriptive metadata [58]. In previous versions of PREMIS, IEs could only be linked with their UID. An Intellectual Entity type may have many Representations. I.e., in this project the set of TIFF files created from the daily resource represent the daily weather map in the same way as a FITS container created from them.

[58] container or reference

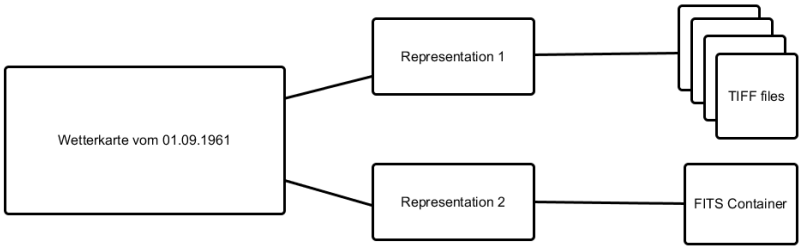


Figure 11: PREMIS reference project model

Events

Events aggregate all actions performed on Objects. An accurate and seamless record for every interaction that changes information is required for maintaining the provenance and authenticity of Objects. Events convey:

- Unique ID
- Event type (creation, ingestion, AIP accession, migration, de-accession)
- Event date & time
- Event description
- Involved Agents
- Involved Objects
- Event outcomes

Events that change an Object should be permanently recoded [59]. For other actions, this is optional.

[59] Documentation of the "state"

Agents

Agents are actors that have a role in Events, Rights Statements, and Environment Objects. They can be individual persons, organisations, or machines. They have the following Semantic Units:

- Unique identifier
- Agent name
- Agents designation (person, organisation, software, hardware)
- Agent version
- Associated events
- Associated objects

The requirement to log an Agent's events and objects makes it near-fetched to provide a central and structured storage facility for

this information. It is obviously insufficient to collect Agents in logs with the Objects if one is to provide reports later on [60]. As well, one Agent may assume heterogeneous roles. I.e., the same person may be the author of a work, and a contributor. This aspect is easier to model in a graph structure than with a conventional relational database.

[60] reproduction of the "state" of an object in a give time

Rights

PREMIS Rights Statements may entail copyright, license, statutory, or institutional policy. Rights Statements are primarily designated for actionable information, that is information that can be acted upon by a machine. Information includes:

- Unique ID
- Allowed actions
- Restricted actions
- Grant or restriction term (time)
- Related Objects
- Related Agents

Provenance step 0 – relation to the physical originals

As PREMIS data mainly focus on the digital preservation aspects, often no description of the physical originals is included. However, only with an adequate description of the originals, the understanding and authenticity of the digital resource will be established. Such a relation is well-established in the map digitization project of University of Vienna.

The MODS physical description construct is suitable for addressing this issue: [61]

[61] From: [UVIE 0:440180]

```
<mods:physicalDescription>
    <mods:extent>1 Kt.</mods:extent>
    <mods:form>Kupferst., mehrfarb.</mods:form>
    <mods:extent>47 x 55 cm</mods:extent>
</mods:physicalDescription>
```

This translates to a raster image of 11,896 x 13,646 pixel (@600 ppi) (503 x 577 mm), which is plausible [62].

[62] [Wie]

File Format

The choice of the file format depends on many factors and theres a hudge literature on this subjects [63].

For our purpose the main distinction would be

- to store the metadata with the picture in one file [64]
- or to have references to a unique id [65]

The list here takes a look on the most commonly used formats.

[63] e.g. see http://www.data-archive.ac.uk/create-manage/format/formats-table

[64] container

[65] seperated infrastructure for metadatamanagement

TIFF

As documented by the TI/A [66] Iniative the TIFF file format is widley used, but *" the specification of TIFF is complex and some of its features are proprietary an therefor not suitable for long-term archival purposes"*. Our test also showed some problem, e.g. the specification of the scan dpi to picture size are not necessarily consistend when using a full scan on A0 where the object is app. A2 size.

The Österreichische Staatsarchiv also showed in its publication ([67], TIFF- Preservation Process in der Praxis) that TIFF is a intermediary format used for the scanning process but for (long term) archiving purposes a clear definition of the concrete TIFF Implementation is necessary and conversions from TIFF to "standard" TIFF would be necessary.

[66] Tagged Image for Archival

[67] [et.16a]

dejavu

The National Oceanographic and Athmospheric Administration Central Library [68] uses the file format for the archive on historic documents [69]. It was originaly intended for the easy handling of scan files but is no longer developed [70]. Special Viewers are needed to access the files.

[68] NOAA

[69] [Lib16]

[70] [djv14]

pdf A

Although the file format is propriatary by Adobe Cooperation there are ISO Standards [71] that define the properties and it is used especially for archiving textual information [72]. There are extensions available also for pictures, but they are implemented as embedded files. A Community [73] is there and there are tools to validate ([74]

[71] ISO 19005-1:2005 for pdf/A, ISO 32000-2:2017 for pdf 2

[72] https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml

[73] https://www.pdfa.org/

[74] [vc16]

still in development as part of the PREFORMA Project [75]), but our test showed that most of the software (even the products of Adobe) do not support all features of the format and a subset has to be specified when trying to use it. As our objects are for a great deal pictures (that should be useable for further processing) it would not be the format of choice for this project but for e.g. the digitalisation of the "Jahrbücher" ([76]A4 book bound, yearly report, text and tables) the format to use (pdf A/1b plus OCR) as proved by our sister institution Geologische Bundesanstalt [77].

[75] [FI16]

[76] [ZAM93]

[77] https://www.geologie.ac.at/services/bibliothek-archiv/

jpeg2000

This format is used by the Österreichischen Nationalbibliothek. Compared to TIFF and standard jpeg its more useable for archival purposes[78]. It is possible to compress the file but it is preserved as a image and no parts of it are accesible individually so a processing would be necessary to use individual object contained and especially the OCR can not be implemented as a container so you have the picture "as a whole" for "display" but not easily usable as information.

[78] [PB08]

FITS

The FITS Format was developed over 40 years ago for astronomical observations.

Flexible Image Transport System

Standard 3.0

For a Reference of FITS see Library of Congress [79] and for the Standard[80] the NASA pages and for a good description [81].

[79] [oC16b]

[80] [NAS08]

[81] [ea10]

FITS for weather maps

For our purpose the different parts of the pages (e.g. map europe, map austria, tables, ...) would be separated into Header/Data Units (HDUs) to be accessed individually, each Header Unit would include the metadatadescription of the parts and in the Data Unit the content of the scan would be present.

For the tables this could be a binary array to preserve the "original", a conversion with OCR or a link to already present parameters in a database. At the time, as we are still talking to users about what they need, its not clear what would be preferred.

Also the color coding of the templates proofed of no significance as the relevant information can be retrieved by black and white the file size can be reduced significantly [82].

[82] [BAV16]

So a preprocessing of the scans will be necessary. Its possible for the different formats of templates to define areas on paper where the information unit should be present and to extract the parts automatic. Then, the parts would be included in ONE FITS file to

have a container for a certain day. With that the SIP would consist of a file structure ordered by date.

As a part of the post processing the comments on corrections have to be handled (some times later a correction of parameters are part of the daily weather map reaching back into the past). This has to be included into the original file of that certain date and be done BEFORE it gets in the archive for long-term preservation.

So the workflow would be similar with the BAV workflow but with more steps and control cycles before a file is stored in the archive.

see below

Application of FITS for archiving purpose

BAV Vatikanische Bibliothek

Since 2010[83] the Vatikan Library is using FITS. In an Kooperation with ESA[84] the archive Format [85] and its application [86] was developed and technical implementation was done[87].

Through a personal contact to Dr. Luciano Ammenti [88] it was possible to get more details from the project. Especially the discussion with the software firm [89] that did the implementation of the IT Infrastructure at BAV was very helpful to get a feeling about the tools used.

A conclusion for our purpose was that doing this on a bigger amount of objects needs some special software not freely available but a implementation of a similar workflow like BAV would be affordable as a PREMIS Database is part of it and a lot of objectives would be similar to the existing solution.

[83] [Rad10]
[84] [Chi12]
[85] [Lib15]
[86] [Amm12]
[87] [All12]
[88] provided by Dr. Paolo Budroni
[89] seretat.com

ZAMG

For ZAMG using FITS would be a synergy as we are the Sentinels National Mirror Austria [90] and with it the LTDP Strategy of ESA [91] is relevant and in it FITS plays a major role.

[90] [ZAM16b]
Long Term Data Preservation
[91] [et.13]

Digitization

Tests

To do the actual work of scanning and converting the pictures into a file was not part of the project. What has been done was to test the tools available and to evaluate the necessary precautions that have to be taken.

see **Software**

Test on file format

We did some test on:

- TIFF versus pdf as scanformat
- TIFF versus jpg for Fotos (RAW →)
- conversion of picture files to FITS
- conversion of FITS files (Raster) into bitmap
- FITS Implementations (color or grayscale)
- compression of FITS files (jpg, gz, ..)
- Picture and Text as HDU

Header and **D**ata **U**nit

The aim was to find a work-flow that could be managed with the infrastructure already there.

By using freeware tool it was possible to get the job done but it is time intensive and a lot of manual work. This was OK for the testing purposes but would be an impossible task for > 100.000 pages .

Scanning process

They way to get the pictures was tested with

Camera A camera (Canon MarkII Digital, EOS-1) with a special objective on a tripod was used to take the pictures. The resolution was aceptable but the bending on the plane proved to be troublesome.

Flat bed scanner The test where done on a Zeutschel A0 flat bed scanner. The weather maps are bound into books so the handling of the oversized objects and the positioning of the fold was a problem. We did some tests on the resolution (starting with 1200 dpi) and settled on 600 dpi high enough to get a acceptable picture and good OCR results. The handling of the fold was done by the scanner

software but some distortions (especially with the maps) where visible.

Professional book scanner We took some weather map books to the digitalisation department of the austrian national library and tested with a Zeutschel OS 14000 A1. With a speed of 15 seconds for a double page, correction of the fold and adjusting the picture on the fly we got perfect results for 600 dpi.

Findings

The planned digitization of the weather maps requires an industrial transfer process. The entire project covers app. 200 bound volumes, rendering an app. 150.000 A3-page equivalent. In uncompressed 24 bit RGB mode @600 ppi, this amounts to app. 25 TB data. [92]

[92] see Inventory table

For digitization processes, traditionally there was a clear distinction between the production of surrogates (for access purposes), and re-formatting for preservation. The 2004 NARA guidelines put this at the very beginning of their document: *"The NARA Technical Guidelines for Digitizing Archival Materials for Electronic Access define approaches for creating digital surrogates for facilitating access and reproduction; they are not considered appropriate for preservation reformatting to create surrogates that will replace original records."* ([93,94] p1) In more recent publications, such delimitations no longer seem to exist. The 2016 FADGI guidelines [95] while sharing large text passages with the 2004 NARA offer guidance independently from re-formatting or surrogate production. This trend is in line with our own experience which seem to point at increasing activities of re-formatting efforts for information that tend to have either electronic access and use in the future, or which are endangered by obsolescence in their current format, such as audiovisual materials.

[93] (

[94] [NAR04]

[95] [FAD16]

In fact, this project has some criteria that lead to the conclusion of doing it rather under "reformatting" auspices than of that of "digital surrogates". These are:

- It is a retro-digitization effort, not an on-demand effort
- Not all final use scenarios can be defined
- Present use scenarios include the creation of machine-readable information
- The scale of the project is substantial
- The present digitization technology is mature
- There are no indications for a possible, second attempt in the future

Quality Management

Quality Management according to ISO 9000:2015 consists of different activities, including quality planning, quality assurance, and quality control ([96] Section 3.3.4). The main objectives of quality planning involve the planning of the digitization infrastructure and –services. It includes the choice of equipment and the design of the process. As well, specific targets are defined in respect to the digitization products.

[96] [ISO15]

Quality assurance is a function of the digitization business procedure, providing evidence that the quality targets are being met.

Quality control is a group of tasks within the digitization process, providing insight if the agreed quality targets are being met. These may include eyeball checks, process modelling, and programmatic approaches, among other.

Strategic Quality Targets

Retaining information of the originals is the primary target of quality planning in the reformatting workflow. Strategic quality targets are:

- The equivalent appearance of the image files versus the originals.
- The ability of a human to retrieve the original information from the raster images
- The ability of a machine to retrieve the original information (i.e., processing the raster images through OCR and vectorization device)
- The files can be used for a standardized reproduction and access process.

There are several technical and procedural parameters that influence the above targets. A file may be regarded equivalent versus the original, if:

- The file is portraying the original resource. This means, it is indeed displaying the expected content, references have not been mixed up with a different resource, and the original resource is unabridged, unaltered in the sense of absence of distortion and artefacts.
- Using an appropriate rendering aid (i.e. a printer or a display), all of the information contained in the original resource can be retrieved by a human.

- Viewing the raster images at 1:1 scaling leads to the verdict that the raster images are equivalent in appearance

While the second target (human readability) is already included in defining the criteria of equivalence, the ability of automated information processing depends on strictly quantified technical parameters that must be assured during the digitization process.

Critical raster image related quality parameters

The relevant system parameters are specified in ISO 19263-1:2017, "Best practices for digital image capture of cultural heritage material". This norm provides practical advice for the reproduction of two-dimensional, reflecting objects.

The image quality characteristics provided there entail: White balance, tone reproduction curve, gain modulation, noise, dynamic range, banding, defect pixels, colour accuracy, sampling rate, limiting resolution, sharpening, MTF 50, illumination non-uniformity, colour mis-registration, distortion, and reproduction scale. [97]

[97] [ISO17]

The ISO standard is complemented by United States FADGI Guidelines and European efforts to the same target, the Metamorfoze guideline of the Netherlands State Library. We are considering these in context.

Aberration and focus problems

Aberration may occur from several sources in the digitization process. Being optical systems, scanners and cameras introduce optical errors in reproduction. The main variation we expect is even placement of the originals during reproduction. However, as georeferencing is immune against any projection errors because it refers to real-world coordinates, the main problem that could be created by distortion and focal problems caused by placement is for OCR. A further issue could be a focus defect of the capturing system.

Spatial resolution

Spatial resolution is defined by pixels per square unit, conventionally by pixels per inch, or ppi. One dimension of a square pixel at 300 ppi is 0.085 mm, or 0.042 mm at 600 ppi. The actual resolution of a scanner system depends on the implementation of the sensor. CCD line sensors are being used for high-quality scanner devices. CCD array sensors are being used by digital cameras. In the first case, capturing is done by progressive exposure. With an array sensor, the original is being captured at once. In either case, the actual resolution depends on the dimensions of the sensor, and the reproduction factor. For the size of our originals, only a few scanners provide the required resolution. These are top models of the most expensive manufacturers. The number of pixels in a sensor

is a best-case limit. Actual resolution is limited by several factors. I.e., specifically Bayer-matrix sensors are typically affected by mosaic effect. De-mosaic artefacts are typically created by the signal processing used to compensate for the dislocation of the spectral sensors for a single pixel. [98] NARA and ISO recommends to look at the Modulation Transfer Function (MTF) or Sampling Frequency Response (SRF) of a system. A capturing system blurs edges of high contrast, depending on the performance characteristics of this indicator. Metamorfoze assumes 85% sampling efficiency from top-level scanners, as of 2012, for a 300 ppi device. [99] For a 600 ppi scan, app. 700 ppi physical resolution would be required at that efficiency.

[98] see Wikipedia Bayer Filter

[99] [Met12]

Density

Density is the property of reflection (or transmission, for films) of incident illumination units. Density is the negative, decadic logarithm of reflectiveness:

$-10log_{10}\frac{reflected}{incident}$

As the originals are non-photographic, reflecting material (ink on paper), the possible density range of <1,5 is well within the capabilities of CCD sensors (d2...d4) [100]

[100] [Met12]

Color Mode

Scanning technology exists for black & white, greyscale and Red/Green/Blue (RGB) colour mode. The additive RGB colour space is the prevalent technology in today's mainstream scanners. Some scanners include a separate lightness channel (8...14 bit). Generally speaking, for the purpose of this project (OCR, vectorization, these are B/W signal based processes) the inclusion of a lightness channel would be desirable, however we haven't found a system so far that allows to produce RGB and lightness signals simultaneously. When assuring a neutral grey scale behaviour in calibration, conversion is almost lossless, and will result in a high-fidelity B/W signal stream.

Signal resolution

A color space is a specific organization of colors. Device dependent and device independent color spaces exist, i.e. eciRGBv2 (device dependent) and CIELAB (universal) color spaces are relevant for this project1. The color space is required for the correct interpretation of the quantized component values, and for the correct transformation of color to lightness. In the domain of device-dependent color spaces, ICC profiles are required for correct interpretation. [101]

[101] [Met12]

Essential Characteristics

For a well-organized definition of the digitization process, the critical characteristics have to be defined in the curatorial and technical aspects.

Curatorial

The requirements are:

- Evidence of provenance of the raster images from the originals
- Maintain the accessibility of the originals on an acceptable level during the digitization process: The users will be informed about the planned absence from the library
- Re-instate the original accessibility of the originals after the digitization
- De-binding of the originals shall be avoided as far as possible
- Only reversible manipulations are allowed on the originals

Technical

Technical requirements include:

- Complete scanning of the originals
- Production of raster images of equivalent appearance of the originals
- Production of raster images technically suitable for the extraction of machine-readable information (Geo-referencing, Vectorisation, OCR)
- Production of preservation information data (PDI)

Regarding these requirements, there may be a conflict of the requirement to produce complete scans, and the requirement to leave the originals intact. Some sources are bound in a way that hides some information in the gutter. These targets may require a trade-off, limited by the curatorial interdict of non-reversible manipulations on the originals. De-binding thus is only allowed to the extent that the original state can be re-instituted.

Metamorfoze suggests one scan per opening. [102] Thus, the physical context of the originals can be easily preserved. No decision has been taken if this advice will be followed in this project, or if for larger originals one scan per page will be allowed. The context trade-off is probably acceptable, metadata allow for the context information anyhow.

[102] [Met12]

Quality assurance methods and objectives

The main axioms for this digitization project are:

- Picture quality assertion vs. process quality assurance
- Systematic quality assurance vs. sample quality assurance

- Automated QA vs. manual QA

In a preservation process, we do not assert if the upstream results are aesthetically pleasant, or if they are suited for a specific downstream process, i.e. for web publishing or printing. Of course, the technical parameters that have been established for information retrieval need to be met. Else, we are primarily interested in accurate reproduction of the sources. If the originals have defects, we do not aim at correcting these in the upstream process, and the target of eventual restoration would be information retrieval. The primary target of quality assurance thus is the reproduction with the highest fidelity possible. For this, the transfer system itself needs to be monitored for transfer parameters, and the individual scans have to be tested for operational errors or malfunctions. The failure rate allowed for mass-digitization in a re-formatting process is very low, maybe 1-5 items in 10.000 documents at most (0,01...0,05%).

The cost of testing can be lowered by minimising the test effort, using statistical methods. Another approach is to use automated test routines. As automated tests scale well, sample testing is usually employed for manual quality control operations, and automated tests can be done systematically. We will discuss how both approaches could be utilized.

Transfer perfomance indicators

We may assume that the picture quality of the originals is adequate for the context they have been created for, as the originals have been produced by a certified specialist. We cannot make an assertion if the meteorological information provided is accurate in each and every case. However, we do assume that the provided information is readable. Hence, subjective assertion of the picture quality of the digital raster images is not a quality objective in this project.

The main quality objective of this project is indeed the transfer process. We may assume that excellent originals, processed by the defined technical parameters, assuring the absence of transfer errors, will provide adequate digital raster images.

The critical high-level transfer errors symptoms include:

- Mismatch of file vs. original
- Focus error
- Alignment errors
- Cropping error
- Data rot

These symptoms are critical, because they either may lead to categorical (i.e. cropping error), or progressive (i.e. focus error) information loss. Thus, they need to be excluded and / or minimised.

Q/A Methodology

Error- testing pattern may be categorized, so that they form a 2x3 matrix:

Testing pattern	Error Pattern	
	Systematic Error	Random Error
Retrievable by sample testing	x	random
Continuous testing	x	x
Severity	high	low

We can conclude that systematic errors will be detected, using sample testing. An error that is occurring with every item will be detected for sure within the very next sample. For any error that occurs at random, sample testing may or may not detect the error. The probability of detection of random errors depends on the occurrence and testing rate [103].

[103] see -> Quality Checking section

Regarding severity, systematic errors are more severe than random errors. For severity, we also categorize reversible an irreversible transfer errors. A document mismatch may be reversible, if "Document A" by error has the file name for "Document B". The same error may be irreversible, if "Document A" is provided two times, once under the name of "Document A" and another time under a different name, while that document is missing.

Error	Systematic	Random	Automated: Detection	Manual	Reversible
Mismatch	o	+	+*	+	+
Bitrot		+	+	-	-
Focus error					-
Alignment error	o	+	?	+	o
Cropping error	-	+	+	+	-

For the image parameters from ISO 1963-1, the following can be classified as primarily prone to systematic errors:

- Reproduction scale
- Noise
- Banding
- Defect pixels
- Sampling rate, and MTF-50/ MTF-10
- Colour mis-registration
- Tone reproduction curve, gain modulation, dynamic range, and colour accuracy.

These parameters may be controlled by periodical alignment of the system, and validation of reference images. This process should take place at the beginning and end of a batch, systematically.

The following parameters may render single-item specific variances because of operation errors:

- Illumination non-uniformity and white balance: the parameter may be influenced i.e. by environmental light sources or operation errors, even if the light conditions in the laboratory are typically controlled
- Resolution (limiting) may be detrimentally affected by focus adjustment errors

- Distortion could be caused by misplacement of the originals

The most critical of these aspects is to avoid resolution defects caused by focus problems. Distortion is the least critical problem, as geo-referencing is to a high degree resilient against this by the means of normalizing the original projection characteristics. Random illumination uniformity and white balance errors are mainly the result of an operation error. The required environmental characteristics of the laboratory however make it rather unlikely that an error is introduced that will lead to a non-reversible information loss.

Quality checking

NARA recommends to do a minimum of the higher of 10 or 10% sample tests, entailing a 100% back tracking if a 1% error rate is detected [104]. While giving some guideline, this is not sufficient as foundation for a quality policy.

[104] [NAR04]

The confidence (or reliability) of finding erroneous outcomes in a digitization batch is comparable with a draw from a pool of blue and red balls of infinite size. After aggregation of all quality requirements, and executing Q/C as an atomic (entire) procedure, this is an adequate comparison, as the final verdict is "error: true/false". If we want to ensure with at a certain confidence that there are not more than a specified proportion of red balls in an infinitely large pool, we need a certain minimum of draws:

$p^x = C$

With p = proportion of blue balls,
x = number of draws,
and C = confidence

$x = \frac{log(C)}{log(p)}$

With p = 0,99 (proportion of blue balls), and C = 0,01 (1% probability allowed of randomly missing one red ball in 100 items), we need to do 459 draws. If we are fine with 90% confidence, we only need 43 draws. If we had to ensure that there are no more than 0,1% errors (1 red in 1000 blue) at 99% confidence, we need 4603 draws. (Note, that this simplification only gives insight for the general behaviour of sample testing. In real-life scenarios, we would define that the draws empty the pool, as we keep drawn balls separate.)

Obviously, specifying sample rate at a certain percentage is not significant for the possible outcome. The actual requirements are the tolerable error rate, and the required confidence.

Digitization in lots

Dividing the digitization project into lots is a paramount requirement. Lots allow:

- Efficiency gain by grouping of critical parameters, such as format and template

- Originals can return to the after a short period at the digitization lab
- Handover and acceptance procedures happen at pre-defined targets
- Delimitation of risks

For the present project, possible batch sizes could be:

- Number of documents per unit (here: volume)
- Documents digitized, per day or week, and person

Assuming the digitization of a document takes 3-5 minutes, there are 96 – 160 documents digitized per 8-hour shift. The typical number of documents per volume for this project is about 180 (half year). By example, 18 draws in a 180-batch would lead to 83 % confidence for the 1% error threshold (which is not very good). If we look at the total entire of the project of 41.000, 10% sample testing would only approximately allow for 99% confidence of an error rate not exceeding 0,1% (Excluding with a probability of 1:100 that the number of defects could be higher than 41 failed transfers).

As much as the NARA backtracking requirement for an error rate found > 1% seems to be a meaningful demand at first sight, but we find that it is problematic in some ways:

- It is not obvious what the 1% threshold means in respect to an underlying quality policy. Ideally, the backtracking threshold is below the acceptable error rate. If it is equal or higher, a batch affected has failed, and qualifies for immediate rejection.
- In the case of outsourcing parts of the digitization process, backtracking is not a useful tool for takeover tests. If at all, back tracking could be a task for the service provider if a batch is deemed to be beyond quality limit.
- The cost of back-tracing for the given sample rate of 10% is factor ten (10x) versus the sample-testing base line. However, to be acceptable as a risk, confidence rates of » 90% would be required to be acceptable. (10x the cost at 10% probability bring a cost increase of 100% over 10% sample testing!) This is hard to achieve for the typical batch sizes in this project and probably also generally.
- At any rate, backtracking must be limited to a single batch. Finding out after the acceptance of a batch that detected errors potentially have affected the total population is inacceptable.

It should be mentioned that additional checks are required to safeguard the acceptance of a batch. I.e., the first and last item typically will be checked, and those items cannot be included in the random sample count. Other obvious efforts include counting and name pattern checking, which apply to all items of a batch.

Summarizing, the suggested sample testing method is not suitable for the present project. The suggested sampling rate does not lead to sufficient sample sizes aligned with the typical lot sizes.

Allowing higher error rates within a batch than for the total, or allowing lower confidence thresholds within a batch are not an option, as it might adversely affect the total quality level, and brings the risk of backtracking events. Lot sizes for outsourcing as well must be kept below a size where acceptance testing can be completed within a week after delivery. Both, increased sample rate and backtracking could impose an inacceptable cost burden or cost risk.

We also think that it is not necessary to treat all error patterns in the same way. 1% error rate for misaligned scanners may be acceptable, as this does not necessarily entail information loss. On the other hand, 410 lost documents by mismatch, bit rot, or any other fatal error pattern would be on the high side. This however leads to substantial challenges in the Q/A implementation: An allowed error rate between 0,01% and 0,1% (99,9% ... 99,99% error-free) would result in 5...41 lost documents. For sample testing, this would incur a sample rate close to the entire total.

For defining a final quality policy, the requirements are:

- Setting acceptable error and confidence rates per error pattern
- Establishing methods for in-process and handover Q/A procedures

The practical options are:

- Using technical means to assist manual quality assurance
- Mix and match. Outsourcing certain parts of the Q/A process to a service provider
- Implement automated Q/A tests, specifically for quality aspects that require low error rates

Automated Q/A approaches

Automated Q/A procedures are established for preservation workflows for an ongoing period. In the field of digitization of audiovisual materials, machine-assisted Q/A is paramount due to the high amount of manual work required otherwise. See [105], [106], and [107] for some sample publications in this field. Key criteria of automated Q/A include:

[105] [Gab07]
[106] [Gab16]
[107] [LH99]

- Pro-active quality testing strategy instead of back tracking risk
- Assembly-line work-flow approach: Transcription and testing are separate tasks
- Q/A is metadata-driven. Metadata are used for quality control, and metadata is systematically controlled.

As we have not found and existing, comprehensive automated quality control initiatives for raster images, we would like to come up with concepts that could be evaluated in the follow-up of this project:

File and metadata testing

For the TIFF-files in the acquisition process, JHOVE2 presents a comprehensive toolset to analyse the essence data for well-formedness and conformity [108]. All parameters mentioned in the section Technical Metadata can be validated. Metadata files can be tested against their respective XML schemata. Values are tested against expected values. Critical parameters are, among other:

[108] [LoC]

- IFF 6.0 well-formed
- Expected Image dimensions (width, length)
- Bits per components (8, 8, 8 RGB)
- Compression (uncompressed)
- Color space (sRGB)
- ICC profile
- Scanner data
- File names
- Checksums

Some target values are constants for the entire project, i.e. file format, pixel format, color space, and ICC profile. Other parameters change with the batch, i.e. expected image dimensions. Any deviation from the desired target value, or target range mean an error signal.

File names and checksums have to be unique for the entire project. That is, if doublets occur, this is a possible error signal.

Alignment, focus, and cropping error testing could be automated by using information from reference data which are introduced in the digitization process.

- Test targets: Each scan is accompanied by a reference target which is added to the scan plane. The raster area of the target may be automatically extracted in the process, and the size and histogram are matched by a known reference (i.e. a reference scan of the test target alone). This test will reveal alignment problems on the item level
- Reference OCR test: Each scan is accompanied by a reference text in critical size (< 8pt). The text is printed in B&W on a paper strip and placed in a specified zone aside the resource. In Q/A, the text area will be extracted, processed by OCR, and compared to the reference value. The expectation is that a critical reference text will be unreadable for the OCR processor, if the scanner is out of focus. Optionally, the text may contain item-specific information (i.e. resource date) that allows the detection of matching errors.

The placing of two reference items in specified areas (i.e. top-left and bottom right of the resource) allows the automation of cropping error tests: If the reference items represent a closed rectangle, entailing the entire resource, the presence of both reference items in the prescribed places excludes unwanted cropping of the resource.

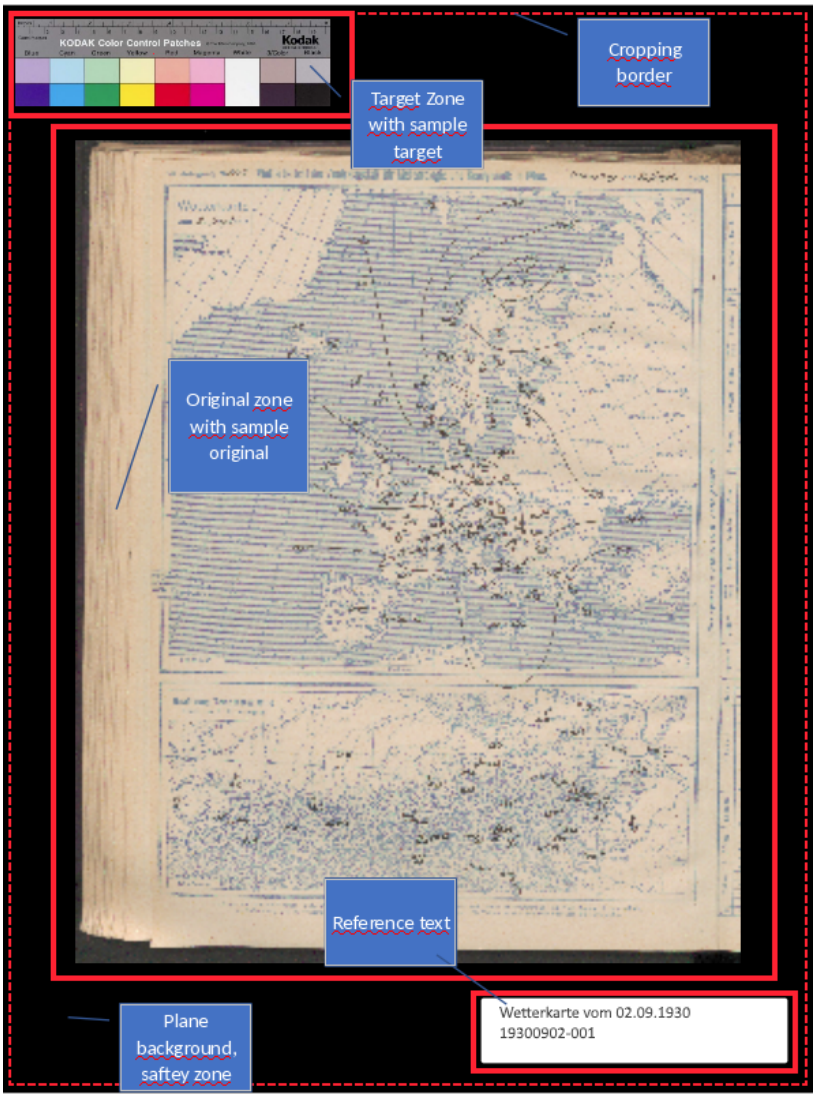


Figure 12: Sample zone composition

The implementation of the Q/A tests may be done with standard tools. I.e., the open-source image processing library OpenCV may be used to extract pre-defined image areas, process, and evaluate them [109].

The design of the reference text may be done incrementally, so that negative results remain below the accepted error margin for a properly aligned scanning process.

[109] [Opeb]

Digital image quality assurance methods

Relevant digital image quality assessment methods for mass-digitization transfer systems for raster images have been established by the National Library of the Netherlands. The Metamorfoze guide lines establish a systematic quality assertion system that observes the dependency of the image quality parameters. I.e., it stresses that correct value interpretation requires the color space to be defined, and that color accuracy, opto-electric conversion func-

tion (OECF), and exposure only can be assessed if the entire gray scale has neutral reproduction.

Metamorfoze specifies target and deviation values for these parameters, and suggests daily and per-image control routines to assess image quality [110]. A very useful information included in Metamorfoze are extensive reference tables for all relevant quality parameters in (device independent) L*a*b*-values and 8-bit quantisation levels, plus a variety of conversion routines, i.e. from RGB to lightness. Moreover, Metamorfoze suggests the use of a test target that unifies greyscale, color- and MTF assessment features and reference files in a single product, the so-called UTT target. The use of conventional Kodak Q13, Color Checker SG, and QA-62 reference targets is supported by Metamorfoze as well, as the features of these targets are unified in UTT. UTT apparently cannot be used as object-level reference targets.

[110] [Met12]

UTT establishes a software-supported QA process in the form of daily references, which may be widely automated, and is available in commercial products from Zeutschel [111], IQ- Analyzer from Image Engineering, Germany [112], and a cloud-based solution with the brand name Delta-E [113]. All three software solutions allow the quality assurance of a transfer system, and they allow to a certain degree the automated monitoring of the calibration process.

[111] [Zeu]
[112] [iqm]
[113] [del]

A different approach is pursued by the DICE software, developed by the Library of Congress. DICE is any acronym for Digital Image Conformance Evaluation Program. The software allows the assertion of the following parameters against the FADGI Star system (FADGI 4-star is equivalent to Metamorfoze) (see [114] p 11, ff):

[114] [FAD16]

- Sampling Frequency
- Tone Response
- White Balance Error
- Illuminance Non-Uniformity
- Color Accuracy (ΔE2000)
- Color Channel Mis-Registration
- MTF/SFR (Modulation Transfer Function / Spatial Frequency Response)
- Reproduction Scale Accuracy (Future Implementation)
- Sharpening
- Noise
- Skew (Future Implementation)
- Field Artifacts (Future Implementation)
- Geometric Distortion (Future Implementation)

A blog post on the Library of Congress web site allows for the conclusion that DICE can be used in the actual transfers. This is a major benefit versus the above-mentioned solutions, which only allow for ensuring a periodical alignment process.

The implementation is using object-level reference targets to determine the quality.

DICE seems to be the most promising approach. It is available

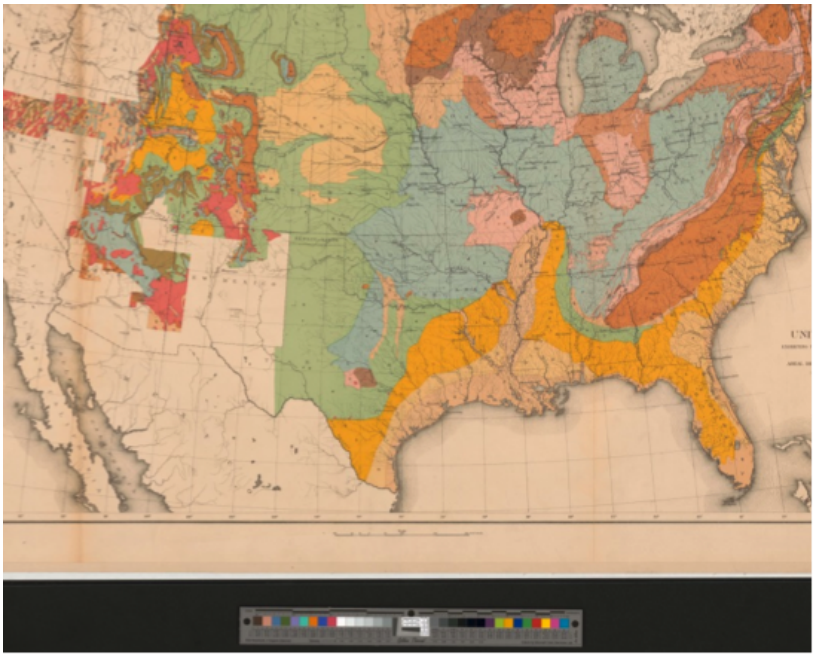


Figure 13: DICE enabled scan, https://blogs.loc.gov/loc/files/2015/08/Untitled.png

as open source software. Further evaluation seems to be appropriate, though. It may be noted that FADGI does not impose the systematic rigor on the sequence of quality parameters, and it is also not clear how DICE will practically assess certain parameters with object-level targets. I.e., it is not obvious how illuminance non-conformity and other global parameters can be asserted using a small target strip on the bottom. Some parameters, such as Field Artifacts are not yet implemented in DICE, so we need to find separate solutions for securing the quality targets in this sector.

The main difference of monitoring a calibration process and monitoring the actual transfers is that the former mainly aims at quality assurance without controlling the actual transfer, and the latter assures the process quality by controlling the actual operation. The DICE approach also has the benefit avoiding operational overhead – which is introduced by an extensive calibration monitoring process.

Ensuring object integrity

A critical quality check is to ensure that the captured image is complete. This could fail because of the following reasons:

- The original is not placed correctly
- The capture is incomplete
- The resulting image was inadequately cropped

In a visual check, an operator acquainted with the originals may easily assess pass / fail of this condition. However, we believe that this check also can be automated. The parameters that allow so are:

- Horizontal and vertical dimensions: With the physical dimensions of the originals and the required resolution known, the dimensions of the raster images are within a certain range of expectation. Dimensions cropped below the threshold raise an error.
- Using reference items: A color target and a serialized (i.e. containing specific values for each item) reference text are placed top-left and bottom-right to the originals, in a pre-defined position. These can be identified using object detection software. (See: Fig:12)
- Zone delimitation: Reference items and originals delimit clearly-defined zones of the raster image, determined by their approximate co-ordinates. Each of the zones has specific characteristics, expressed by their histograms:
  - The color target zone renders a standardized histogram which accurately reproduces with the calibrated scanner.
  - The reference text renders a histogram in very narrow boundaries, moreover using OCR it can be validated for the actual information
  - The original zone renders a histogram that is sufficiently specific to determine if it contains an original, or not. Specifically, it may be compared to the histogram of the scan plane background material, which is known. This is suited to spotting completely missing originals.
  - The zone set may secure, with a certain confidence, against the absence of field artefacts. I.e., the reference items create a security zone around the originals that make the intrusion of hands (See: Fig:12) , or other disturbances less likely. As well, if a field artefact covers any of the zones, it may be detected by changing the characteristic picture parameters

The resulting performance indicators can be evaluated automatically.

Respective test implementation requires the following actions which should be part of a work package:

- Identification of a suited software library for object detection and image statistics
- Reference measurements and validation
- Validation of critical thresholds
- Evaluation of batch-specific parameters (i.e. dimensions, batch-specific reference histograms)
- Implementation of check task in the workflow engine
- Monitoring of the results

The strength of the fully-automated test method is higher reliability, and cost benefits in respect to testing effort. The weakness is required implementation cost.

Note: the aforementioned library OpenCV allows zone extraction, based on co-ordinates or even on reference histograms. It also

allows for programmatic evaluation, i.e. comparison of two histograms by quantitative and statistical methods, including the Bhattacharya Distance of two histograms, see: [115]. This allows for fuzzy matching methods, including the definition of error thresholds.

[115] [opea]

The role of checksums

Checksums, or fixity-information in OAIS / PREMIS jargon, are paramount for assuring the authenticity of a digital archival. Created immediately upon the creation of a file, they need to be linked persistently, but separately from the essence data. That is to facilitate the concurrent access, and to avoid parsing of the essence for access to the fixity.

Fixity is being used for several outcomes in Quality Assurance. It allows the assertion that no data rot has occurred, it presents a unique pointer to the item (as no two items have the same information), and it even may be used to automate any migration that is per se lossless, i.e. the re-packaging from TIFF to FITS.

The MD5 of the raster data in a TIFF file cannot be predicted from the MD5 that includes the TIFF metadata and headers. However, if we can validate the global MD5 on access, and parse the TIFF successfully, extracting the raster data bitstream, we can create a checksum entailing only the raster data. If the unaltered raster data is inserted into the FITS container, we can validate that, and by the consistency of the pairwise checksums conclude that the conversion was authentic – without requiring manual interference.

Source matching

The number of expected files and their file names can be provided as input values, as they are rule-based. The comparison with the found values at the Q/C stage allows for automated checking. Batch number, batch range, file name match, and item count per batch can be automated, and will give some hint for further Q/A. I.e., it is sufficient to make sure that remaining matching errors are reversible. If a batch of 100 scan jobs contains 100 discrete file sets (by hash, set range, and file name), and the start- and end files have been checked for consistency, the probability of an irreversible matching error (i.e. by repeating items) is very low.

Moreover, an automated evaluation of the serialized reference text will provide further information about the source matching. (See: Fig:12) The most obvious approach is to include input metadata, i.e. the original date in the reference text strip. Processed with OCR, the values can be automatically compared with the input data, and validated. The reference text has a unique constraint, so it cannot occur repeatedly. All reference values must be found for set completeness. This makes mixing up of sources very unlikely.

Critical parameters for the efficiency of this approach are the quality of the input data, i.e. that initially asserted file numbers and file names are correct, and the ability to insert correction values

along the process. I.e., if the digitization operator finds discrepancies with the source data (additional pages to be scanned, expected pages missing), this must not put the process to a halt. A feed-back loop should allow the correction of the data in a qualified process.

Assertion of quality parameters on practical scans

The information of the originals consists of tables, text, and discrete line drawings. The information conveyed with colour is limited to dyed patches, or ink pen. Colour or lightness gradients do not exist as information component. Thus, equivalent appearance with the original is obviously provided with a RGB scan of 8 bit component depth.

For local resolution, the quality margin is set by machine requirements. As the human eye is in average limited to discriminate structures of 0,15 mm … 0,3 mm , a 150 ppi scan would probably suffice to regard a raster image equivalent to the original. However, for the purpose of OCR and vector processing, requirements quite different. The recommended resolution for 8-point or smaller fonts is 400-600 ppi [116]. NARA recommends to choose a resolution that covers the finest line with at least 2 px [117]. Below standing example illustrates the typical relations of information units relevant for vectorization (600 ppi scan, 800% zoom, pixel grid enabled, 8x3 pixels selected on finest line) The segment shows the Atlantic Ocean coast line of the Iberian Peninsula near Bilbao:

[116] [ocr]

[117] [NAR04]

Figure 14: ZAMG weather map test scan, excerpt enlarged

- Water surface hatches have a width of app. 3 px. (see black test patch)
- Ink-pen produced lines have a width of app. 10 px.
- Printed station marks and grid lines have a width of > 4 px.
- The dynamic range is well within the capabilities of the scanner and the format. Lights and shadows have plenty of head- and footroom.

Our tests for the retrieval of machine-readable text and vectors

have shown that vectorisation yields considerable quality improvements for resolutions of 600 ppi. [118]. The excerpt supports these findings by the following observations:

[118] See Appendix

- Generally, a line coverage of more than two pixels is required, as lines may be slanted. At an angle of 45°, a line of only 2 pixels width may be difficult to trace. (see: Fig:15)

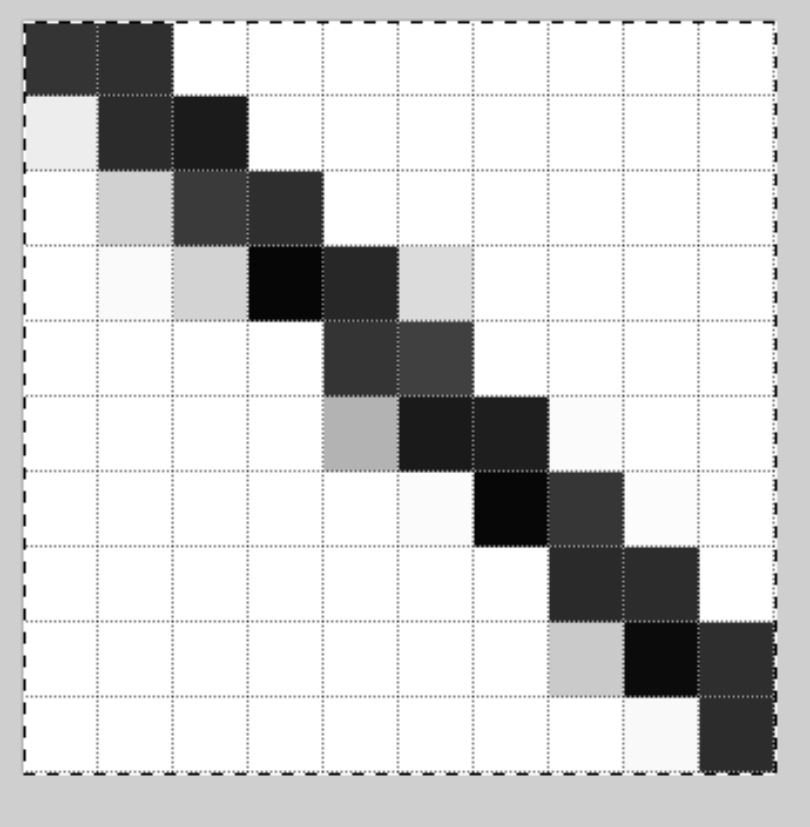


Figure 15: 2-px synthetic, slanted line at 45°

- The edges in Fig:14 show slight sample frequency response artefacts even with the moderate contrast of the original. In our test scan, there is an occasional hue of overly light pixels around some edges that is 1 pixel wide
- As a typical issue of the originals, ink coverage is occasionally patchy for all manual entries. Typically, tracing tools can accommodate any resulting gaps easily, however MTF distortion or de-mosaic issues potentially may exacerbate the problem.

The sample scan is free of de-mosaic issues and focal issues. The pen-drawn details render lines that are app. 10 pixels. wide, and expose the maximum contrast. These results in some vector tracing headroom (also valid for OCR operation), which would not be provided with 300 ppi resolution. Focal issues as well would easily create problematic conditions for vectorisation, and higher resolutions offer some additional headroom for feature extraction.

Model digitization workflow

All workflows modelled hereunder are based on the approach of task distribution and Business Process Modelling (BPM). That means that tasks are processed by specialized operators. Parallelization can be reached by working to and from central repositories that allow non-blocking, concurrent access. Dependencies exist from the order of the tasks modelled, and some exist by the constitution of the originals. I.e., only one person can work on the items bound in one volume.

At this stage, we can build two business procedures illustrating relevant high-level functions. The two sections, Inventory Data and Digitization may run asynchronously. It also needs to be stated that their logistical link depends on several indicators that have not been specified so far. Prominently so on the question of in-house and outsourcing of digitization itself. However, they are valid in both scenarios. A dependency in any case exists for the individual item under treatment: Before an item can be digitized, inventory data must be created.

Outsourcing i.e. would require a specific lot-based takeover procedure. We have not modelled this so far, as no outsourcing has been specified. As well, several other procedures have not been modelled so far, i.e. the AIP creation, and access. These procedures are on the one hand clear in their requirements from the OAIS standard perspective; on the other hand, the individual performance criteria are not yet known.

Creation of inventory data

The creation of inventory data is paramount for the operation of a distributed, template-based, and quality-controlled digitization process. It requires inspection and autopsy of the originals to determine the effective set of items to be digitized. The process itself needs to be quality-controlled [119]. The process works against a central repository, an inventory dataset in a database. The sequence of the tasks is relevant; however, the task groups may be distributed and parallelized, i.e. two people may work on unit level, and five persons may work on item-level. This is provided by the existence of a unit-based and an item-based loop.

[119] not explicitly modelled here

The workflow creates data about each unit, and each item. The result is a clear understanding about the identifiers and names on unit (volume) and item level, including pagination, physical measures, and template type. Any observations need to be annotated (i.e. special entries, damage). This provides a hierarchical set information from bottom-up.

Item-based digitization

The digitization workflow is modelled on item-basis. An item (resource) are the maps and tables for one date. It may consist of sev-

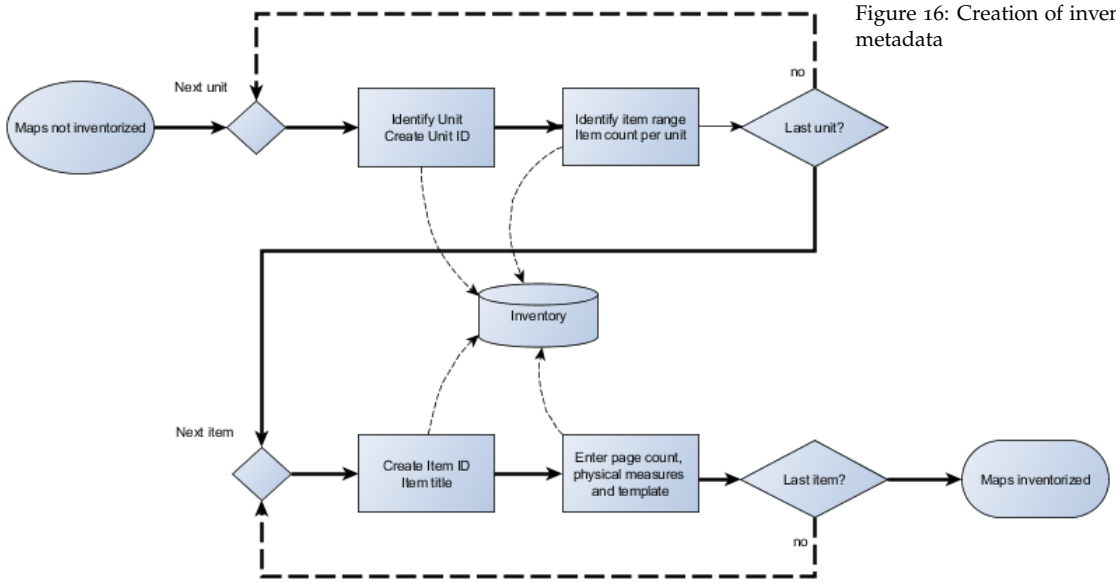


Figure 16: Creation of inventory metadata

eral pages. The workflow has a feedback-loop after the manual scan task. If more than one unit is processed, the entire workflow can be parallelized. In contrast to the inventory, this workflow leverages automatic tasks for file check (JHOVE), Zone-and image testing (Test-lib), and automatic quality control based on data and metadata (Auto QC). These tasks may be parallelized per se, and that is also true for the manual QC step. SIP creation does not need a systematic QC task, as it is fully formalized.

The entire workflow runs off a Workflow database that contains inventory data, plus the operational data which may vary depending on the template, physical dimensions, and conservatory considerations, and the status information. The Workflow DB also collects all the data and metadata created in the workflow (not pictured, analogous to the inventory flow each task creates metadata). These data contribute to the PREMIS set, which is the central piece of the preservation description information. Thus, from the outset of the life cycle of the digital object, consistent authenticity and provenance information is provided.

The workflow is initialized using the operational parameters and the inventory data for each item / page. The completed workflow contributes data and metadata to the OAIS repository, transactionally. During the workflow operation, the OIAS repository is not updated.

Automatic QC is a pure evaluation task, driven by the initialisation data, reference data from the operational pattern, and all data collected during the workflow. These are for example:

- All inventory data
- Essence data
- Fixity
- JHOVE parameters

- Test-lib results
- OCR results
- Reference data created per lot, i.e. histograms
- Rules for a complete and successful digitization

The further processing depends on result codes. In case the check is negative, a manual QC step will be invoked. The same happens for successful tasks, evaluated by a random picker, following the specified test rate, for sample testing.

Manual tasks (scan, QC) will be at least instructive. That means, specific suggestions are made to the operator what steps he has to follow, and pick lists will be offered for all pre-defined values, such as file names, and result codes. If possible, API-based integration may be attempted, i.e. pre-filling file names, and file navigation, driven by workflow data.

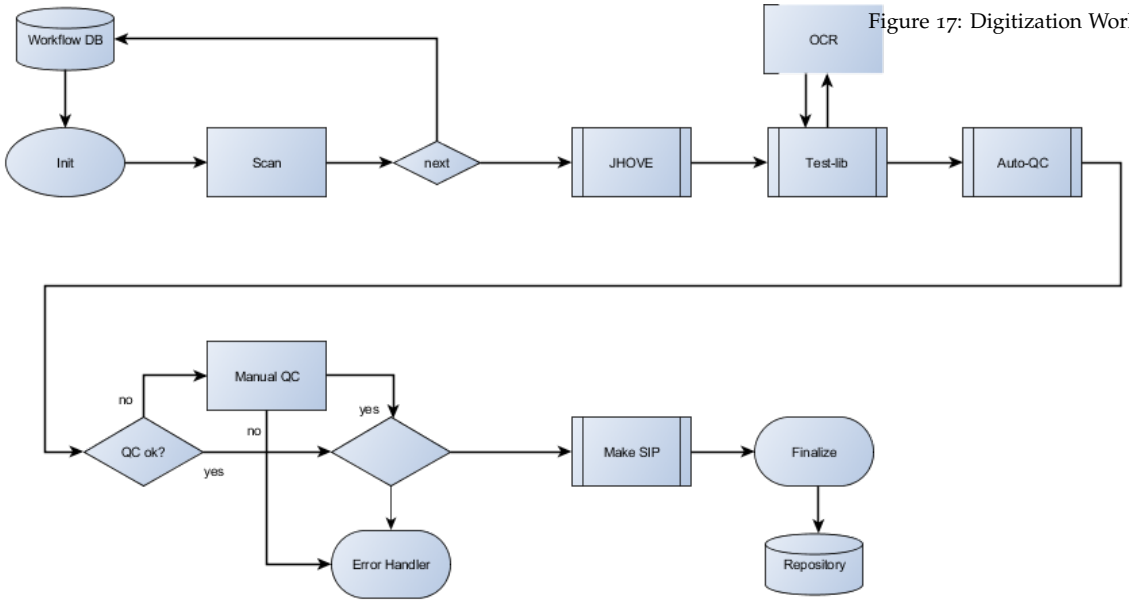


Figure 17: Digitization Workflow

Use of the output files

As our monitoring of the work done in the area of digitation of weather map has shown a broad range of strategies may be applied.

For a simple "solution" the scans could be made available[120] as a picture (e.g. djavu, pdf, tiff, jpeg) for visual evidence.

A more "advanced" way of dealing would be to digitize the paramters in the source and make the data available[121].

The "perfect" solution would be to provide the information and the source with the documentation of all metadata involved in the production[122].

No standard software exists for the full process from source to data and providing an infrastructure to use the ressources that are there. So we are on a new path here and this project was the first step on the journey.

[120] Access

[121] Information

[122] Content

Climate Data (Rescue)

The WMO is active in the climate domain and has through GCOS and other Initiatives a special interrest in historical Data [123].

The Departments "Datenprüfung offline" and "Klimaforschung" at ZAMG have already started on bringing the historical parameters into databases. The existing gap to the originals can be closed with this project. This will create significant synergies.

The CCCA Datazentrum [124] provides already some data for the use of the scientific community and is in discussion what should be provided in what way spanning the whole spectrum from analytical maps to raw data. Also here synergies are to be expected.

Wold Meteorological Organisation
Global Climate Observing System

[123] [WMO15]

[124] [Aus16]

Vectorisation of the weather maps

It is possible [125] to convert the weather maps and acquire the isobar lines automatically. With this time series comparison of the weather situation would be possible.

The Isobarlines are drawn on a template. The grid of the template is a geographic projection. Through algorithms the lines can be captured and extracted as vectors.

[125] see chapter Automatic vectorization

georeferencing maps
the proof of concept is part of the Project.

Vectorisation of Isobarlines
NOT part of the project

Conclusion

In the year of the Universitätslehrgang Library- and Information Studies (Grundlehrgang), it was possible to leverage exitisting expertise of one team member (Sebastian Gabler) in the field of long-term preservation, and transfer it into the re-formatting of paper resources. Mutually, it was possible to leverage the domain- and organisational expertise of the other team member (Rainer Stowasser). This has produced bi-directional knowledge transfer, which was mutually satisfactory.

This statement of work (SOW) establishes in-depth foundations to start a digitization effort for the collection of the historical weather maps of the ZAMG. The findings are of substantial value for the organisation, fulfilling another section in the wide scale of its regulatory duties.

The main fields of this SOW were:

- What is the context of the collection?
- What is the future use scenario for the information stroed in the collection?
- What are the governing standards?
- What are suitable file- and data formats?
- What are the available tools?
- What are governing quality standards?
- What are procedural requirements?
- What are suitable automation approaches?

After designing this concept, a program has to be set up to tackle the steps necessary and to set the strategic goals and define what should (and can) be done.

The main questions that would have to be answered:

- Should these ressources be made available?
- What is the required investment cost?
- What is the required operational cost?
- What are the required skills?
- Outsource or inhouse efforts?

As we have documented here there are different approaches, "here they are do as you please " [126] to "we have use cases and provide a solution" [127].

Lets see what the future brings along.

[126] picture on a website

[127] a system with user support

Automatic vectorization - Sebastian Flöry B.Sc.

In order to use the information contained in the scanned weathermaps it is necessary to georeference and vectorize the relevant features. The automatic vectorization can be summarized in three steps:

1. Detection of the relevant map within the whole scanned image
2. Geo-referencing of the map
3. Extraction of the features and theire attributes

The goal of the automatic vectorization is to retrieve the features as georeferenced datasets which further can be used in various other products.

Various software packages have evolved through the need of software being capable of processing und working with geographical data efficiently. This software packages are known as GIS software - geographical information system software. Besides commercial GIS (ArcGIS), many open-source GIS are developed (GrassGIS, QGIS, SagaGIS). QGIS is one of the most used open source GIS and is especially known for the possibility to develop own plugins and easily include them. Therefore QGIS is the perfect environment for the task of automatic vectorization.

One big advantage by including the process into QGIS is the benefit of using all the already realised and implemented functions. Especially in the end of the process, where a human user needs to correct errors and add missing information this will be a major advantadge.

Detection of the map

In figure 18 an example scan is shown. It can be seen that the relevant map, including the relevant features, is only a small part of the whole image. Therefore in the first step it will be necessary to detect and clip the image to the relevant map. This step will reduce the overall processing time and further eliminate possible error sources. On the right side of the scanned image we can see that the raw measurement values are listed in tabular form.

Figure 19 shows the clipped map. In the following we will be only looking at this part of the scan.

The background information in blue is clearly distinguishable from the drawn information in black. The background information

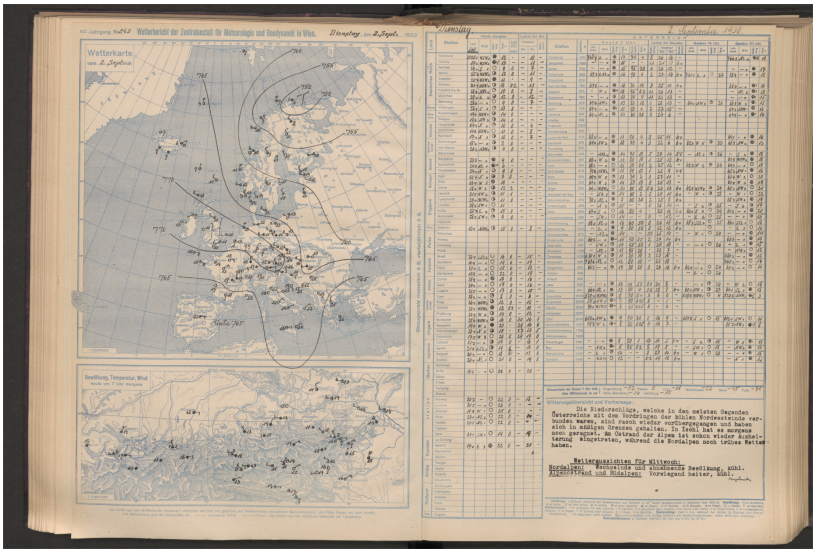


Figure 18: Selected example scan. Relevant map only on the left

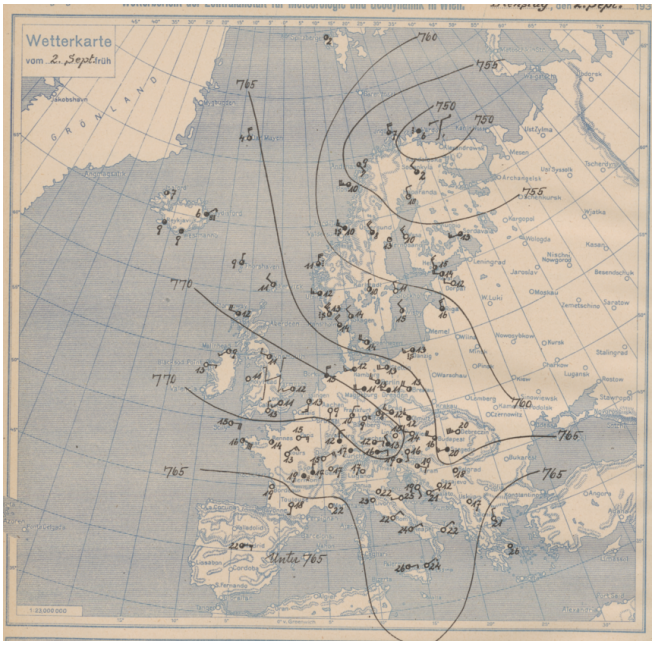


Figure 19: clipped map

mainly shows geographical data: coast lines, latitude, longitude and cities (as blue empty circles). Our features of interest are the weather stations (symbolized as circles with weather vane) and the isolines.

Geo-referecing the map

As mentioned before, the background information in blue contains mainly geographical information. We will use both the crossings of the latitudes and longitudes and the displayed cities for geo-referencing the scanned map.

Figure 20 shows the automatically detected cities in red. We can see that especially in central europe most cities are covered by the

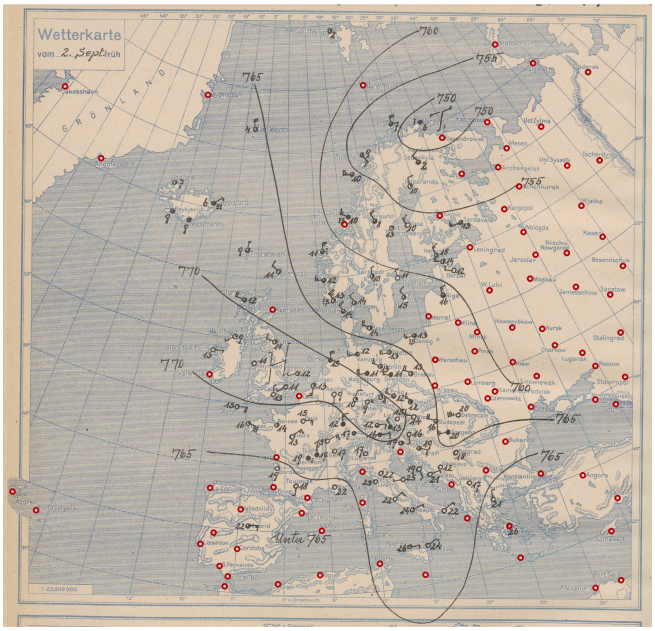


Figure 20: Automatically detected city symbols (red)

symbols of the weather stations. Therefore to guarantee a good georeferencing also in these part of the map the inclusion of the longitude and latitude is vital.

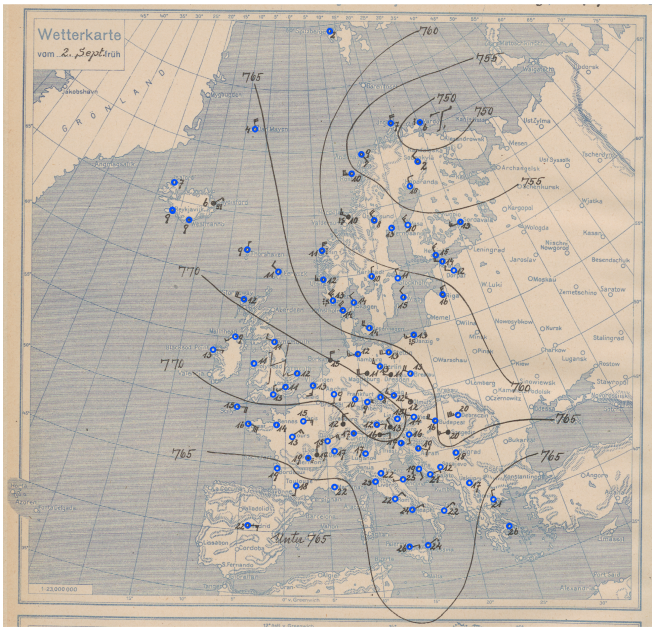


Figure 21: Automatically detected positions of weather stations (blue)

Extraction of the relevant features

The position of the weather stations is marked with a circle. The filling of the circle itself indicates the degree of cloudyness. Attached to the circle a weather fane indicates the direction of the wind and its strength. In a first step, the position of the weather stations itself is detected.

Figure 21 shows the result of the automatic detection highlighted

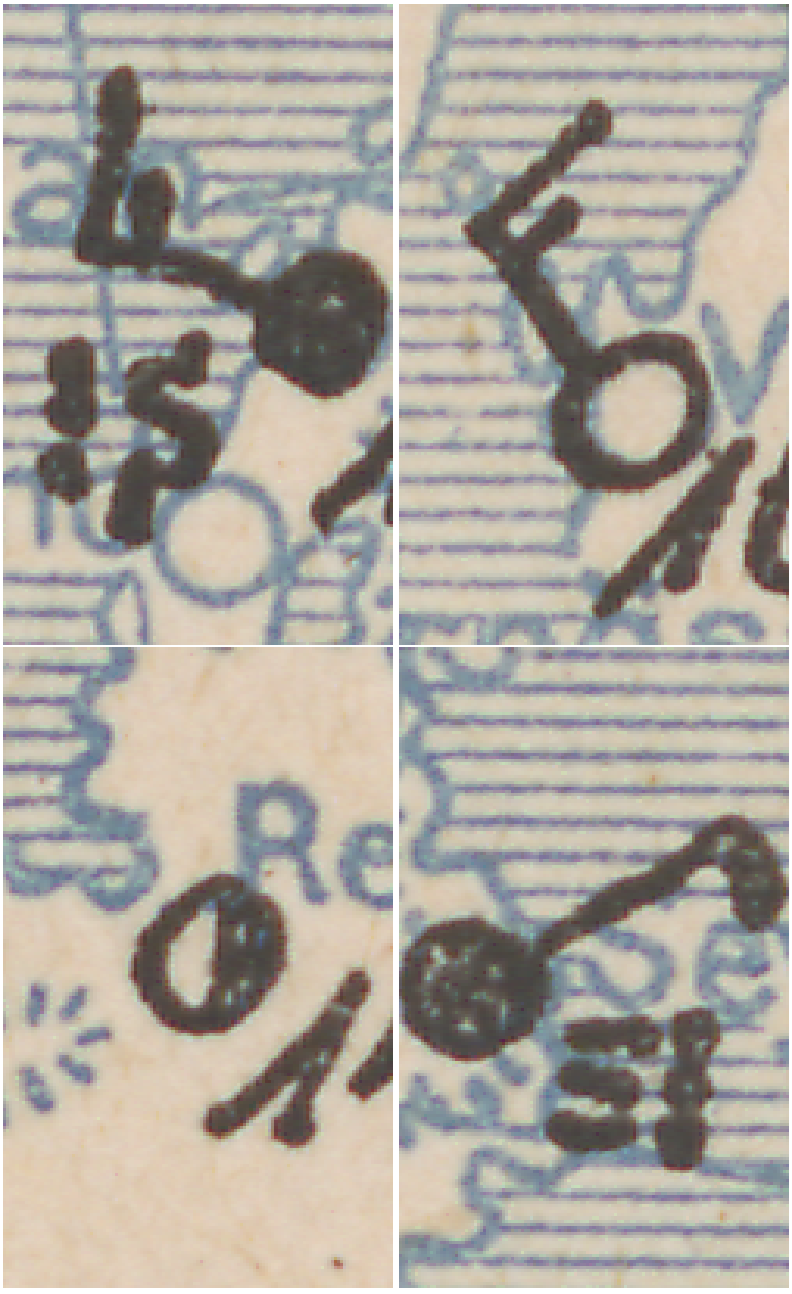


Figure 22: Selected weather station symbols showing different weather conditions

in blue. As soon as the position is known the surrounding at every station is processed in detail to capture the additional attributes as well.

Additionally to the weather stations also the isolines are vectorized. 23 shows the result in green.

Postprocessing and quality control

As mentioned in the beginning by embedding the automatic vectorization within a GIS-software it is possible to use various additional functions and tools. Especially for the post-processing and quality control this is of major importance. Therefore the result of the georeferencing can be directly controlled using comparative geographi-

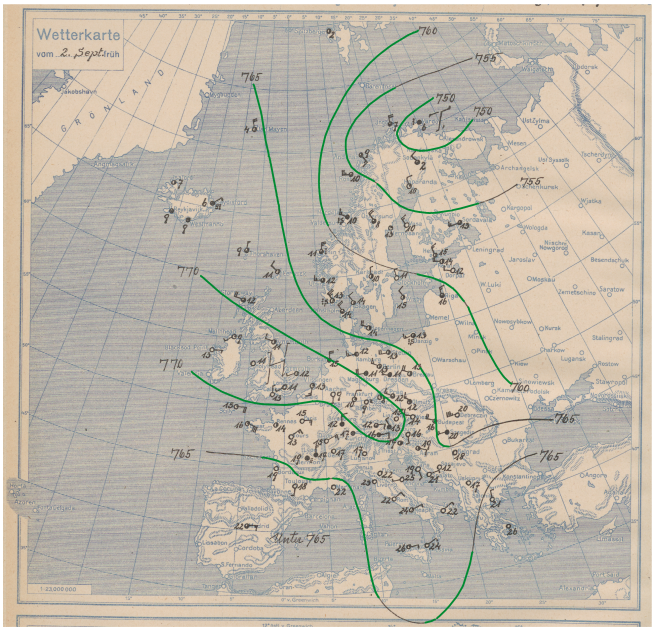


Figure 23: Automatically detected isolines (green)

cal data. Additionally missing or wrongly classified geometries and attributes can be easily adjusted and corrected.

Tables

Formats of the objects

All objects contain tables with the parameters of selected stations of austria and international (different ones over time).

We have produced test scans for every template but do not include all of the in this work due to the sice of the pictures.

Especially the number of pages per day where used to access the amount of available information but the different "objects" contained make a specification necessary as only the map of europe and in the tables the parameters of some stations are consistent over the whole time span.

The format of the books varies and the paper dimensions do not exactly comply to the ISO 216 A - Standards and in some cases where cut in the binding process, so every size mentioned is an approcimation.

start	end	format	volume	description
1877	1883	A3, landscape	24	map europe template grey
1884	1908 I	A3, landscape	2 per year	map europe template blue, date on top, tables
1908 II	1914 I	A3, portrait, 2 pages	2 per year	map europe template blue, date upper left
1914 II	1916 II	A3, portrait, 2 pages	2 per year	map europe template blue, date upper left, tables, map air pressure, temperature
1917	1920	A2, landscape	2 per year	map europe template blue, date upper left, map air pressure, temperature
1921	1923 II	A3, portrait, 2 pages	2 per year	template bluedate upper left, maps morning 7h, midday 2h, evening 7h UTC
1924	1928 II	A3, landscape, 2 pages	2 per year	map europe template blue, Date upper left, map Austria , Atlantik, Isotherme, yesterday evening, Isobare
1929	1934 II	A3, portrait, 2 pages	2 per year	left page map europe template blue, map Austria, right side tables
1935	1940 I	A2, landscape, 1 page	2 per year	map europe template green, below map Austria, right tables
1940 II	gaps	A3, portrait, 6 pages	2 per year	template europe grey, Radiosonde report
1945	1948	A3, portrait, pages varying	2 per year	reports of the occupying powers, tables, sometimes maps
1949	1953 II	A3, portrait, 2 pages	2 per year	right page map europe template blue, map Austria, left page tables
1954	1962 I	A3, portrait, 2 pages	2 per year	left page template europe blue , map Austria , right page Tables , map air pressure 500mb

1962 II	1968 II	A3, portrait, 2 pages	2 per year	left page map europe template grey, map Austria, right page Tables , map air pressure 500mb
1969	1970 II	A3, portrait, 2 pages	2 per year	left page tables, map Austria, right side map europe template grey, map air pressure 500mb
1971	1977 II	A3, portrait, 2 pages	2 per year	left page tables, map Austria, map air pressure difference yesterday 1h today 1h, today 4h today 7h, map air pressure 500 mb, right page map europe template green, comments
1978	1985 II	> A3, portrait, 2 pages	2 per year	left page tables, map austria, right side map europe template green
1986	...	> A3, portrait		Computer print out

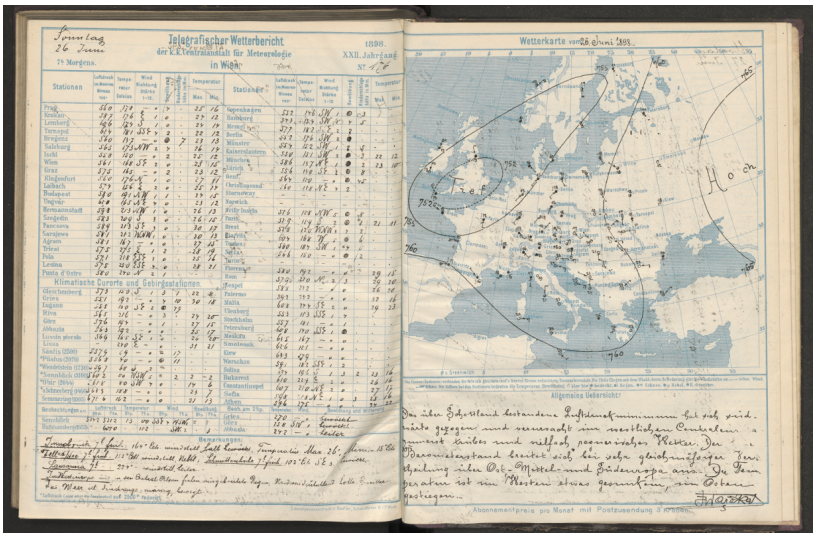


Figure 24: Weather map, 1898 26 July, paper damaged, Scanner Zeutschel OS 14000 A1, Original TIFF 600dpi

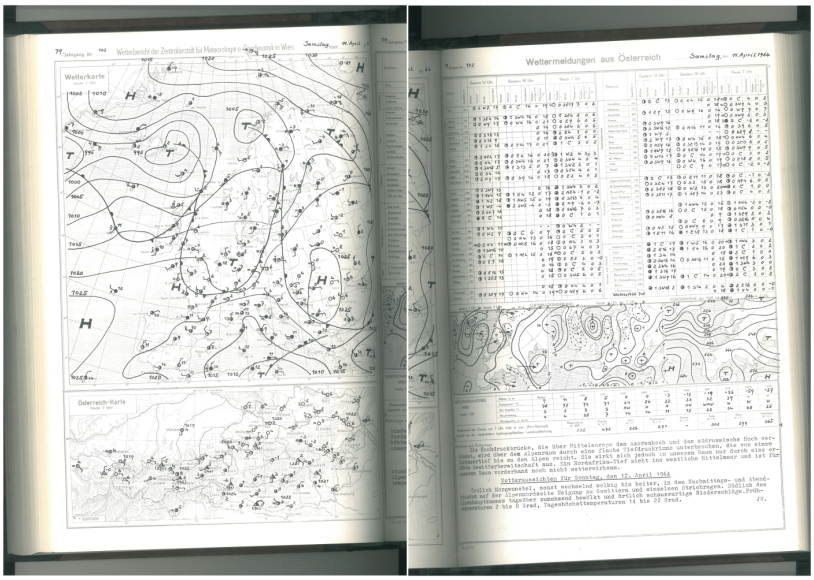


Figure 25: Weather map, 1964 11 April, Scanner Konika Minolta bizhub 223 A3, Original TIFF 600dpi

Metadata Table

TIFF Metadata.

Metadata Set	Tag Nr	Name	Description	Sample Values	Note
TIFF tag, baseline	256	ImageWidth	The number of pixels per row	3616	Typical scanner size
	257	ImageLength	The number of rows of pixels in the image	4418	
	258	BitsPerSample	Number of bits per component	8	Grayscale
				8 8 8	24-bit color
	259	Compression	Compression scheme used on image data	1 = Uncompressed 4 = CCITT Group 4	
	262	Photometric Interpretation	The color space of the image data.	0 = WhiteIsZero.	Specified Color Space for project: eciRGBv21
				1 = BlackIsZero.	
				2 = RGB.	
	277	SamplesPerPixel	The number of components per pixel	1	Grayscale
				3	24-bit RGB color
	282	XResolution	Horizontal pixel count per resolution unit	2400000/ 10000	240 ppi: Rational data type
	283	YResolution	Vertical pixel count per resolution unit	629145600/ 2097152	(inches, centimeters)
	296	ResolutionUnit	Unit of measurement for X and Y Resolution	1	None
				2	Inches
				3	Centimeters
	306	DateTime	Date and Time image was scanned	2008:07:23 17:45:21	24 hour clock UTC
	315	Artist	Image Producer	ZAMG	
	338	ExtraSamples	Description of extra components	0 = Unspecified data	Associated alpha is generally interpreted as true transparency information.
				1 = Associated alpha data (with pre-multiplied color)	

Metadata Set	Tag Nr	Name	Description	Sample Values	Note
TIFF tag, extended	269	DocumentName	Document Name	path/filename	Image referenced: Constructed to uniquely identify the file Document referenced: Document title
	270	ImageDescription	A text string that describes the subject of the image	Path/filename Agency uid	A baseline tag that must be respected by all applications. It may serve the same purpose as the 269 field.
	42016	ImageUniqueID	A unique file identifier	Uuid: 01AF8BA45...	ASCII 128-bit UUID
TIFF tag, baseline	271	Make	The scanner manufacturer	e.g. Zeutschel	Simple ASCII text string
	272	Model	The scanner model name or number	e.g. 14000	Simple ASCII text string
	305	Software	Name and version of the software package(s) used to create the image	e.g. Stokes Software Inc. IWS - Version 02.04.01.01	Simple ASCII text string
Exif	34665	Exif IFD	Pointer to collection of all Exif Metadata. Exif uses field names rather than tags to indicate the field content.		
	37384	LightSource	The kind of light source.	0 = Unknown; 1 = Daylight; 21 = D65	Known light sources are appropriate when using targets containing reference color components.
	40961	ColorSpace	eciRGBv2 specification	1; 256; 65535	eciRGBv2
ICC	34675	ICC Profile	Color profile data		

References

Persons

• Hofrat Mag. Rainer Stowasser	Head of Library ZAMG
• Dipl.-Tonm. Sebastian Gabler	Projektleiter
• Mag. Christa Müller	Betreuerin
• Dr. Paolo Budroni	Phaidra Universität Wien
• Dipl.-Geol. Chris Schubert	Leiter CCCA Datenzentrum
• Erwin Petz	INSPIRE ZAMG
• Dr. Thomas Hofmann	Leiter Bibliothek GBA
• Mag. Martin Forster	Digitalisierung ÖNB
• Dr. Luciano Ammenti	Biblioteca Apostolica Vaticana
• B.Sc. Sebastian Flöry	TU Wien

Abreviations

CCCA	Climate Change Center Austria
DARE	Data Rescue projects and initiatives WMO
DMZ	Demilitarisierte Zone
ESA	European Space Agency
ESO	European Space Observation
FITS	Flexible Image Transport System
GAB	Geologische Bundesanstalt
GCOS	Global Climate Observing System
HDU	Header and Data Unit
LTDP	Long Term Data Preservation
OCR	Optical Character Recognition
SOW	Statement of Work
SIP	Submission Information Package
UTC	Unified Time Coordinates
WMO	Wold Meteorological Organisation
ZAMG	Zentralanstalt für Meteorologie und Geodynamik (eingetragenes Warenzeichen)

Software

- astropy [128] [128] [AT16]
  - a python interface to astronomy packages, FITS File handling (astropy.io.fits)¶
  - Windows, Linux, Mac OSX
- FITS Liberator [129] [129] [ESA16]
  - FITS image processing software
  - Windows
- fv [130] [130] [HEA15]
  - Interactive FITS File Editor
  - Windows, Linux, Mac OSX
- ImageMagick [131] [131] [ISL16]
  - create, edit, compose, or convert bitmap images
  - Windows, Linux, Mac OSX
- Koha [132] [132] [Koh16]
  - fully featured, scalable library management system
  - Linux
- netpbm [133] [133] [Hen16]
  - utilities for manipulation of graphic images
  - Linux, Mac OSX
- scan tailor [134] [134] [AS12]
  - interactive post-processing tool for scanned pages
  - Windows
- tesseract [135] [135] [Smi16]
  - OCR engine, since 2006 developed by Google
  - Linux, Mac OSX, (3rd Party für Windows)
- verapdf [136] [136] [vc16]
  - Free, open-source, Implementation Checker validation software for all parts and conformance levels of the PDF/A specification for archival PDF documents.
  - Linux, Mac OSX, Windows

List of Figures

Bibliography

[1114] Force 11. Data citation principles. https://www.force11.org/group/joint-declaration-data-citation-principles-final, 2014. Accessed: 16. June 2017.

[All12] Stefano Allegrezza. Flexible Image Transport System: a new standard file format for long-term preservation projects? In *EWASS Special Session 12, Long-term preservation from the stars? File format assessment and technical issues in preservation projects for cultural resources* [137]. Konferenz.

[Amm12] Luciano Ammenti. BAV and the FITS (Flexible Image Transport System) format. In *EWASS Special Session 12, Long-term preservation from the stars? File format assessment and technical issues in preservation projects for cultural resources* [138]. Konferenz.

[AS12] Nate Craun Aidan Sawyer. scantailor, version 0.9.11.1. http://scantailor.org/, 2012. Accessed: 16. June 2017.

[AT16] Open Source Community Astropy Team. astropy, version 1.3. http://www.astropy.org/, 2016. Accessed: 16. June 2017.

[Aus16] Climate Change Center Austria. Data center. https://data.ccca.ac.at/, 2016. Accessed: 16. June 2017.

[BAG] Bundesarchivgesetz. *RIS - Gesamte Rechtsvorschrift für Bundesarchivgesetz - Bundesrecht konsolidiert, Fassung vom 29.04.2017*.

[BAV] Bundesarchivgutverordnung. *RIS - Gesamte Rechtsvorschrift für Bundesarchivgutverordnung - Bundesrecht konsolidiert, Fassung vom 29.04.2017*.

[BAV16] BAV. Fits color space. https://www.vatlib.it/home.php?pag=spazioColore, 2016. Accessed: 16. June 2017.

[BK15] et.al Bettina Kann. Bibliothekarstag 2015, slot 6.3: Digitale langzeitarchivierung. https:

[137] *European Week of Astronomy and Space Science Special Session*, Rome, Pontificia Universita Lateranense, 2012. Biblioteca Apostolica Vaticana. Konferenz

[138] *European Week of Astronomy and Space Science Special Session*, Rome, Pontificia Universita Lateranense, 2012. Biblioteca Apostolica Vaticana. Konferenz

//bibliothekartag2015.univie.ac.at/fileadmin/user_upload/k_bibliothekartag2015/pdf/BT15-6_3.pdf, 2015. Accessed: 16. June 2017.

[CC16] Library of Congress and PREMIS Editorial Committee. Premis data dictionary for preservation metadata, version 3.0. https://www.loc.gov/standards/premis/v3/index.html, Nov 2016. Accessed: 10. August 2017.

[CCS12] Reference model for an open archival information system. https://public.ccsds.org/Pubs/650x0m2.pdf, Jun 2012. Accessed: 10. August 2017.

[Chi12] Lucio Chiappetti. The fits format. In *EWASS Special Session 12, Long-term preservation from the stars? File format assessment and technical issues in preservation projects for cultural resources* [139]. Konferenz.

[Con] The Library of Congress. Lc linked data service: Authorities and vocabularies (library of congress). http://id.loc.gov/ontologies/premis.html. Accessed: 10. August 2017.

[DCC11] DCC. cite-datasets. http://www.dcc.ac.uk/resources/how-guides/cite-datasets, 2011. Accessed: 16. June 2017.

[del] Deltae - online image quality assessment. http://delt.ae/index/home. Accessed: 10. August 2017.

[djv14] djvu. djvu. http://djvu.org, 2014. Accessed: 16. June 2017.

[dru17] Drupal - open source cms. https://www.drupal.org/, Aug 2017. Accessed: 10. August 2017.

[dura] Fedora digital object model - fedora 3.6 documentation. https://wiki.duraspace.org/display/FEDORA36/FedoraDigitalObjectModel. Accessed: 10. August 2017.

[durb] Training - migrating from fedora 3 to fedora 4 - fedora repository. https://wiki.duraspace.org/display/FF/Training-MigratingfromFedora3toFedora4. Accessed: 10. August 2017.

[DWD16] DWD. Cdc ftp-server rasterdaten. http://www.dwd.de/DE/leistungen/cdcftprasterdaten/cdcftpraster.html?nn=16102, 2016. Accessed: 16. June 2017.

[139] *European Week of Astronomy and Space Science Special Session*, Rome, Pontificia Universita Lateranense, 2012. Biblioteca Apostolica Vaticana. Konferenz

[ea10] W.D. Pence (et al). Definition of the flexible image transport system (fits) version 3.0. *Astronmy & Astrophysics*, 524, 2010.

[ea16] Mark D. Wilkinson et al. Comment: The fair guiding principles for scientific data management and stewardship. *scientific data*, 2016.

[eiP16] e-infrastructure Project. e-infrastructure. https://www.e-infrastructures.at, 2016. Accessed: 16. June 2017.

[ESA16] ESA/ESO/NASA. FITS Liberator, version 3. http://www.spacetelescope.org/projects/fits_liberator/, 2016. Accessed: 16. June 2017.

[et.10] L.S. Tan et.al. *Guidelines on Climate Data Rescue*. WMO/TD No. 1210, 2010.

[et.13] Ruben F. Perez et.al. Towards Long Term Data Preservation of EO data in Europe and Canada with ESA's Multi-Mission PDGS. *ResearchGate*, 275646159, 2013.

[et.16a] Fröhlich et.al. *Digitale Archivierung, Innovation - Strategie- Netzwerke*. Mitteilungen des Österreichischen Staatsarchiv. Generaldirektion Österreichisches Staatsarchiv, 2016.

[et.16b] I. Auer et.al. HISTALP. http://www.zamg.ac.at/histalp/, 2016. Accessed: 16. June 2017.

[EWA12] *European Week of Astronomy and Space Science Special Session*, Rome, Pontificia Universita Lateranense, 2012. Biblioteca Apostolica Vaticana. Konferenz.

[FAD16] FADGI. Technical guidelines for digitizing cultural heritage materials creation of raster image files. http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf, Sep 2016. Accessed: 10. August 2017.

[FED] FEDORA. Fedora architecture diagram. https://wiki.duraspace.org/download/attachments/67242832/f4-arch.png?version=4&modificationDate=1482446347849. Accessed: 10. August 2017.

[FI16] PREFORMA Consortium EU FP7-ICT. Preforma - preservation formats for culture information/e-archives. http://verapdf.org, 2016. Accessed: 16. June 2017.

[FOG] Forschungsorganisationsgesetz. *RIS - Gesamte Rechtsvorschrift für Forschungsorganisationsgesetz - Bundesrecht konsolidiert, Fassung vom 29.04.2017*.

[Gab07] Sebastian Gabler. Managment und qualitätssicherung von a/v archiven. *24th International Audio Convention - Abstract Service*, Jan 2007.

[Gab16] Sebastian Gabler. Quality management for preservation of analogue and digital video tape. *IASA journal*, (46):50–57, 2016.

[GB10] Met Office GB. Uk observations daily weather report / daily weather summary 1860-. https://digital.nmla.metoffice.gov.uk/archive/sdb%3Acollection|86058de1-8d55-4bc5-8305-5698d0bd7e13/, 2010. Accessed: 16. June 2017.

[Gro] PMB Group. Guidelines on best practices for climate data rescue. https://library.wmo.int/opac/index.php?lvl=notice_display&id=19782.

[Har12] Douglas Ross Harvey. *Preserving digital materials*. De Gruyter Saur, Berlin, 2 edition, 2012.

[HEA15] NASA HEASARC. fv, version 5.4. http://heasarc.gsfc.nasa.gov/docs/software/ftools/fv/, 2015. Accessed: 16. June 2017.

[Hen16] Bryan Henderson. netpbm, version 10.47.63. https://sourceforge.net/projects/netpbm/, 2016. Accessed: 16. June 2017.

[int] Chapter 1. introduction to modeshape. http://docs.jboss.org/modeshape/2.6.0.Beta2/manuals/reference/html/introduction.html. Accessed: 10. August 2017.

[iqm] Iq- analyzer. https://www.image-engineering.de/products/software/376-iq-analyzer. Accessed: 10. August 2017.

[ISL16] Open Source Community ImageMagick Studio LLC. ImageMagick, version 7.0.4-1. https://www.imagemagick.org/script/index.php, 2016. Accessed: 16. June 2017.

[ISO12] OAIS standard, 2012. ISO-14721.

[ISO15] Qualitätsmanagementsysteme – Grundlagen und Begriffe, 2015. ISO-9000.

[ISO17] Best practices for digital image capture of cultural heritage material, 2017. ISO-19263-1.

[Koh16] Open Source Community Koha. Koha library software. https://koha-community.org, 2016. Accessed: 16. June 2017.

[Kuh14] Rainer Kuhlen. *Grundlagen der praktischen Information und Dokumentation: Handbuch zur Einführung in die Informationswissenschaft und Praxis*. De Gruyter Saur, 6 revised edition, 2014.

[LH99] Frank Lott and Siegfried Herla. Phönix aus dem schallarchiv - das broadcast wave file (bwf) 20. tonmeistertagung karlsruhe 1998. bericht vom 20.-23. november 1998. *Verband Deutscher Tonmeister*, 20(20):26 S., 1999.

[Lib15] Vatican Library. B.A.V. FITS Keyword Dictionary: Version 1.0 - 2015. https://www.vatlib.it/home.php?pag=KeywordsFITS, 2015. Accessed: 16. June 2017.

[Lib16] NOAA Central Library. U.s. daily weather maps. https://www.lib.noaa.gov/collections/imgdocmaps/daily_weather_maps.html, 2016. Accessed: 16. June 2017.

[LoC] LoC. Jhove2 project. http://www.digitalpreservation.gov/partners/jhove2.html. Accessed: 10. August 2017.

[Met12] Bureau Metamorfoze. Metamorfoze preservation imaging guidelines. www.metamorfoze.nl/sites/metamorfoze.nl/files/publicatie_documenten/Metamorfoze_Preservation_Imaging_Guidelines_1.0.pdf, Jan 2012. Accessed: 10. August 2017.

[mod] Modeshape - jboss community. http://modeshape.jboss.org/. Accessed: 10. August 2017.

[NAR04] NARA. Technical guidelines for digitizing archival materials for electronic access: Creation of production master files – raster images. https://www.archives.gov/files/preservation/technical/guidelines.pdf, 2004. Accessed: 10. August 2017.

[NAS08] NASA/GSFC. Fits-standard 3.0. https://fits.gsfc.nasa.gov/fits_standard.html, 2008. Accessed: 16. June 2017.

[Nat15] OEsterreichische Nationalbibliothek. Digitale bibliothek. https://webarchiv.onb.ac.at/web/20150701141434/https://www.onb.ac.at/about/digitale_bibliothek.htm, 2015. Accessed: 16. June 2017.

[oC09] Library of Congress. Sustainability of digital formats. http://www.digitizationguidelines.gov/guidelines/TIFF_Metadata_Final.pdf, 2009. Accessed: 10. August 2017.

[oC15] Library of Congress. Metadata for images in xml standard. http://www.loc.gov/standards/mix/, Nov 2015. Accessed: 10. August 2017.

[oC16a] Library of Congress. External schemas for use with mets. http://www.loc.gov/standards/mets/mets-extenders.html, Feb 2016. Accessed: 10. August 2017.

[oC16b] Library of Congress. Fits-description. http://www.digitalpreservation.gov/formats/fdd/fdd000317.shtml, 2016. Accessed: 16. June 2017.

[ocr] ocr - optimal image resolution [technology portal] quick links. https://abbyy.technology/en:kb:images_resolution_size_ocr. Accessed: 10. August 2017.

[oEACRU16] University of East Anglia Climatic Research Unit. Data. http://www.cru.uea.ac.uk/data, 2016. Accessed: 16. June 2017.

[opea] Histogram comparison. http://docs.opencv.org/2.4/doc/tutorials/imgproc/histograms/histogram_comparison/histogram_comparison.html. Accessed: 10. August 2017.

[Opeb] OpenCV. Opencv library. http://opencv.org/. Accessed: 10. August 2017.

[Ora16] Oracle. Oracle storage works. http://www.oracle.com/technetwork/articles/systems-hardware-architecture/oos-digital-media-wp-2587906.pdf, 2016. Accessed: 16. June 2017.

[PBo8] Franco Liberati Paolo Buonora. A format for digital preservation of images, a study on jpeg 2000 file robustness. *D-Lib Magazine*, 2008.

[phaa] P h a i d r a. http://phaidraservice.univie.ac.at/das-system-phaidra/. Accessed: 10. August 2017.

[phab] phaidra. https://github.com/phaidra. Accessed: 10. August 2017.

[phac] Phaidra partners. http://www.phaidra.org/community/phaidra-partners/. Accessed: 10. August 2017.

[phad] Technisches über phaidra. http://datamanagement.univie.ac.at/ueber-phaidra/technik/. Accessed: 10. August 2017.

[Rad10] Vatican Radio. Ankündigung FITS. https://www.vatlib.it/home.php?pag=newsletter_art_00086&ling=eng&BC=11, 2010. Accessed: 16. June 2017.

[Sch16] Meteo Schweiz. Schweizer Klimamessnetz (Swiss NBCN). http://www.meteoschweiz.admin.ch/home/mess-und-prognosesysteme/bodenstationen/schweizer-klimamessnetz.html, 2016. Accessed: 16. June 2017.

[Smi16] Ray Smith. tesseract, version 3.04.01. https://github.com/tesseract-ocr, 2016. Accessed: 16. June 2017.

[Tuf16] Edward Tufte. The work of Edward Tufte and Graphics Press. https://www.edwardtufte.com/tufte/, 2016. Accessed: 16. June 2017.

[vc16] veraPDF consortium. verapdf 0.6. http://verapdf.org, 2016. Accessed: 16. June 2017.

[Wie] Universität Wien. Exif viewer für das objekt 0:440180. https://phaidra.univie.ac.at/exif_viewer/o:440180. Accessed: 10. August 2017.

[WMO15] WMO. Data Rescue projects and initiatives (DARE). http://www.wmo.int/pages/prog/wcp/wcdmp/CDM_2.php, 2015. Accessed: 16. June 2017.

[ZAM93] ZAMG. Jahrbücher der Centralanstalt für Meteorologie und Erdmagnetismus/ Zentralanstalt für Meteorologie und Geodynamik. In ZAMG, editor, *Jahrbuch ff*. ZAMG, Hohe Warte 38, 1190 Wien, 1871 - 1993.

[ZAM16a] ZAMG. Informationsportal Klimawandel, Klimakarten. http://www.zamg.ac.at/cms/de/klima/informationsportal-klimawandel/klimakarten, 2016. Accessed: 16. June 2017.

[ZAM16b] ZAMG. Sentinels national mirror austria. https://www.sentinel.zamg.ac.at, 2016. Accessed: 16. June 2017.

[Zeu] Zeutschel. Os qm-tool. https://www.zeutschel.de/en/produkte/qm-os/index.html. Accessed: 10. August 2017.

[Ös17] Bundeskanzleramt Österreich. Digitales Archiv Österreich. https://www.bka.gv.at/digitales-archiv, Apr 2017. Accessed: 10. August 2017.