PHAIDRA

Guidelines on digitisation of two-dimensional documents

Version 2011
Lorisa Andreoli
Marina Cimino

Revision 2022
Lorisa Andreoli
Gianluca Drago

July 2022

Università degli Studi di Padova

SBA Sistema Bibliotecario di Ateneo

Index

Premise

Through various strategies and instruments and in compliance with current legislation on copyright, the University of Padova Library System aims to preserve and make accessible on Internet individual documents as well as important digitised collections related to a broad spectrum of disciplines, in the wake of scientific and experimental tradition that has always characterised our University.

To promote its own ancient, prestigious documentary heritage and to meet customer needs for quick and easy access to digital information content, the University of Padova Library System has established the Phaidra platform: a digital object management system with long-term archiving functions to meet specific aims for the conservation and access to digital collections.

This document, in line with good practices and national and international standards for a quality reproduction of documents, indicates the set of procedures for the digitisation of two-dimensional documents, with particular reference to their archiving in Phaidra. The document is also available in Italian and is periodically updated. It was first published in 2011 and subsequently revised in 2013, 2014, 2017, 2019.

1 Objectives

Defining the objectives of a digitisation project allows you to establish the project management framework from the beginning. In general, digitisation projects must be consistent with one or more of the following general purposes:

- enhance the documentary heritage of the University of Padua and other cooperating institutions,
- promote access to documents relevant for historical, scientific and cultural value or with distinctive features of uniqueness or rarity
- improve services to users, with the possibility of consulting resources collected and sorted into virtual collections, physically distant, inaccessible, little known
- reduce the consultation of original documents in particular conditions (ancient and valuable documents, fragile, in poor condition, in high demand, difficult to handle)
- ensure that the documentary material is available to future generations
- develop collaborative activities with other institutions

2 Selecting documents

Selecting the documents to be digitized, attention must be paid to the laws on copyright and privacy. It is necessary to consider the ownership of the rights, the presence of personal and sensitive data, the final destination of the digitization. To resolve any critical issues it will be necessary to resort to the appropriate solutions (for example, a request for permission to publish), possibly also resorting to legal support.

Documents must be chosen on the basis of the objectives and selection criteria[1] defined by the project. Works already digitised in other collections or already accessible to the public via the network are generally excluded, in order to avoid duplication and contain costs.

In some cases (e.g. heterogeneous, uncataloged or inventoried material), it may be useful to carry out an inventory of documents for identifying the quantity, type, size, state of preservation, and the net asset value of documents and any other information. The *Digitisation project information sheet* (see Attachment 2) may be used as a reference. In addition to digitisation, this information can also be used for storage and cataloguing activities.

3 Preservation

The care and preservation of the originals are activities integrated with digitisation. It is important that an evaluation of the original state of preservation be undertaken before the digitisation and that any treatment on documents be performed after a survey by expert staff; any restoration methodology will be outlined by qualified professionals.

The restoration of documents must be authorised by the competent bodies in the field, depending on the territorial context in which they operate[2].

4 Digitisation

Digitisation is the process of transformation/conversion of an analogue object (text, image, audio, video) into a digital format, interpretable by a computer.

The nature and size of the originals determine the choice of the recording system, the lighting system and methods of treatment (transport, opening of the volumes, handling).

The quality of images defined in the project determines the hardware and the recording software requirements, the acquisition times and image processing, and the memory usage in the storage media to manage and maintain.

4.1 In-house or outsourced digitisation

The choice of digitisation within the institution (in-house) or the use of outside services (outsourcing) has to consider the advantages and disadvantages of the two methods.

---

1 See *Selection for digitizing: a decision-making matrix* https://www.clir.org/pubs/reports/hazen/pub74/matrix/, in Dan Hazen, Jeffrey Horrell, Jan Merrill-Oldham, *Selecting Research Collections for Digitization – Full Report*, 1998 https://www.clir.org/pubs/reports/hazen/pub74/

2 With regard to the Veneto area, please refer to Soprintendenza archivistica e bibliografica del Veneto e del Trentino-Alto Adige and to Soprintendenza Archeologia, Belle Arti e Paesaggio per l'area metropolitana di Venezia e le province di Belluno, Padova e Treviso.

	In-house digitisation	Outsourced digitisation
Advantages	- having direct control of the whole process - learning by doing - improving on-going requirements rather than establishing them in advance - ensuring security, proper handling and accessibility of the material	- the institution pays for the product, usually at an established price per image - containment of costs and limited risks - the supplier can handle large amounts of material - the supplier absorbs the costs of expertise, training and technological obsolescence
Disadvantages	- the institution pays for expenses instead of for products, which include training costs, technological obsolescence and downtime - investment in purchasing and maintaining equipment - need for specialised staff - cost per image not defined	- the institution eliminates one phase of the process; it does not develop in-depth knowledge on digitisation - issues of security, transport and handling of specimens
Recommendations	The in-house service is recommended if: - the collection cannot be moved outside the institution - the digitisation work is very easy - there are already specialised staff and existing equipment	Outsourcing is recommended if: - it is not possible for the originals to be digitised within the institution - the planning involves a large quantity in a short timeframe - there are constraints of space, infrastructure and personnel

Outsourced digitisation can be performed in the premises of the library or at the selected company's location.

The flow of outsourced digitisation activities includes:

- examination of the technical and logistical aspects
- definition of the scanning parameters
- preparation of a market study or a tender
- training of staff and operators involved for quality control
- possible arrangement of the digitalisation set
- preparation of documents
- creation of a prototype
- digitisation
- quality control, with correction of defects and errors
- relocation of documents

The flow of in-house digitisation activities includes:

- examination of the technical and logistical aspects
- definition of the scanning parameters
- market investigation or tender for the purchase of the necessary equipment
- arrangement of the digitisation set
- training of staff and operators involved in quality control
- creation of a prototype
- digitisation
- quality control, with correction of defects and error
- relocation of documents
- product delivery
- final quality control

4.2 Choice of equipment

The data acquisition system (light source, optics, sensor, capture and calibration software) should ensure the image quality required by the project and not damage the original documents.
In particular, the lighting system must be cold-light without emission of UV and IR. For ancient or valuable documents the use of suitable supports is required in order to not damage the document (facing the surface to be scanned upwards and using a tilting platform or V support).

These are some general indications on scanning systems[3]:

Flatbed scanner: for single-sheet documents, or bound documents that can be opened easily: printed materials (e.g. leaflets, posters, printed music, brochures), manuscripts (e.g. letters), maps in good condition, printed music, prints (e.g. engravings, etchings, lithographs), pen and ink drawings without added watercolour or gouache (e.g. cartoons), photographic material (e.g. gelatin prints in black and white and in colour, albumen prints).

Film Scanner, negatives and slides: for films, negatives and slides.

Planetary scanner or digital camera on stand: for bound volumes (e.g. books, albums, printed music, atlases), fragile documents, oil paintings, most works of art on paper (e.g. watercolours, drawings), graphic material and artworks made with flaked and friable substances (e.g. crayons, charcoal, soft pencil), watercolours with thick drafting, tempera or with paints, large or fragile maps, manuscripts (e.g. bound diaries, folded documents), parchments, photographic material (e.g. large prints, historical photographic processes, such as daguerreotypes and ambrotypes), three-dimensional material (e.g. textiles, sculptures, objects).

Alongside conventional shooting systems such as camera, flatbed scanner, planetary scanner, the University Library System also employs a robotic scanner for large formats, created as a hardware/software prototype by a spin-off of the University of Padua[4].

[3]For further information on acquisition systems see Ministry of Culture, *Linee guida per la digitalizzazione del patrimonio culturale*, 2022 (draft version 2022-2023), paragraph D.1.A

[4]For information on the equipment employed by the University of Padua Library System, see the web page *Attrezzature per la digitalizzazione* (authentication required).

4.3 Digital acquisition

The following specifications are to be taken as general guidelines, to be tailored in each case to achieve the best compromise between quality and cost.
High quality images, both in terms of resolution and in terms of colour depth, also imply higher costs of acquisition (equipment and qualified personnel) and of management (file size to be kept). On the other hand, the choice of the digital parameters must be sufficient enough to faithfully re-construct the level of detail of the document.
The sampling density, or the number of pixels that represents the unit of length, must therefore be assessed not only based on the size of the document, but also based on the importance of the original document and the available resources.

"It is important to keep in mind that there are multiple factors that influence image quality: among these, in addition to the sampling density, we maintain colour accuracy, dynamic sensor and its noise.
Establishing a certain sampling density is therefore conceptually wrong because, depending on the shooting system that is used, equal to pixels-per-inch, the final quality of the scan can be very different."[5]

The result of digitisation is the creation of files intended for long-term storage, "master" files, and files resulting from further processing, "derived" files, intended for use by users, typically via the Web.

With regard to the file formats suggested below, the provided indications concern exclusively the production of image files generated by the digitisation of textual, graphic or three-dimensional documents. The file formats suggested for other types of materials, such as films and sound recordings, are analysed in the document *Recommended file formats for long-term archiving and web dissemination in Phaidra* [6].

Master file

The **master** file ("preservation master file" or "archival master file") is the file that represents the best-copy output from digitisation, where "best" means that it meets the objectives of a particular project. These objectives may vary depending on the type of document.
The criteria to be used in creating the master file must ensure faithful reproduction of the document in view of its long-term digital preservation or the need for high-quality printing, ensuring that there be no need to repeat the digitisation in the future.

The master files are intended for long-term archiving and the images are archived as they have been captured by the scanning instrument, without undergoing any re-processing.

Recommendations for the acquisition of documents:

[5] F. Lotti, M. Lunghi, G. Trumpy, *Digitalizzazione di beni artistici e documentari. Manuale di procedure per un laboratorio fotografico digitale*, 2009. In particular, see Chapter 4 *"Densità di campionamento"*.
[6] https://phaidra.cab.unipd.it/static/EN-file-formats.pdf

- The document must be taken in its entirety. Around the document, it is necessary to leave a border of a few millimetres in order to make it possible to read the contours of the document.
- For books, an image file is produced for each page: each side, recto and verso, of each page, including flyleaves, even if there is no information, and blank pages; all parts of the binding: endpapers, spine, textblocks, (in order to show headbands, clasps, hinges, borders). For maps, photographs and archive material, the verso is scanned only if there is information present.
- If the original is mounted on a support which contains information (e.g. a photograph mounted on cardboard with the photographer's trademark), digitisation must also include the support.
- Each document must be scanned alongside a chromatic scale, a greyscale and a metric scale, placed outside of the reproduced image and within the overall frame. In the case of volume, it is sufficient to place the scale once on a paper or page (which will be scanned two times, one with the scale and one without).
- In the presence of scratches, wormholes or oxidation of the inks, the papers must be masked with white paper in order to avoid capturing the underlying content.

Digital Camera Capture Workflow

The acquisition of the digital image from a physical document will turn out more or less detailed depending on the type of camera and the quality of the optical systems in use[7].

The digital acquisition can be implemented in two different ways:

1. establishing the sampling density you want to achieve and using it for all scans regardless of the size of the physical document. This will result in maintaining always the same distance between the camera sensor and the object to be shot, therefore obtaining large images in the case of oversize physical documents and small images in the case of small size documents.
2. exploiting the entire sensor matrix by varying the distance between the camera sensor and the object to be shot according to the size of the physical document; the result will be images that, regardless of the size of the physical document, will all have the same size in terms of captured pixels, but different sampling density.

Phaidra Service employs the second digital acquisition method.

A digital camera can produce files of different formats.
The raw format – also called "digital negative"[8] – is normally the best format produced by cameras. Each manufacturer has its own type of proprietary raw format, with its own specifications, which produces files with a typical extension (for example .nef for Nikon, .cr2 or .cr3 for Canon). There is an open raw format, the DNG format, but it is not normally supported by cameras.

[7]For types and characteristics of cameras see Ministry of Culture, *Linee guida per la digitalizzazione del patrimonio culturale*, 2022 (draft version 2022-2023), paragraph D.1.A.7. For workflow with the camera see paragraph D.1.B of the same document.
[8]Unlike JPEG and TIFF formats that store images already processed by the camera, raw files capture raw or minimally processed data directly from the camera sensor. Because raw formats are similar to film negatives in a photographer's workflow, they are often referred to as "digital negatives" (Adobe Digital Negative (DNG) Specification Version 1.4.0.0).

The raw masters produced by the cameras must be maintained as they come out of the capture tool, without further processing. However, as they are in a proprietary format, not suitable for long-term storage, they must be converted to the DNG open raw format, uncompressed. Files in the proprietary raw format should now be deleted. At the time of shooting camera settings should ensure adequate colour quality (e.g., ProPhoto RGB or Adobe RGB colour space and 16- or 8-bit colour depth per channel)[9].

Digital scan workflow

The word scanner refers to very different technologies. Also shooting systems that actually use a camera mounted on a stand are often referred to as scanners, but in this document with "scanner" we refer to the flatbed scanner[10].

The best format produced by scanners is usually the TIFF format.

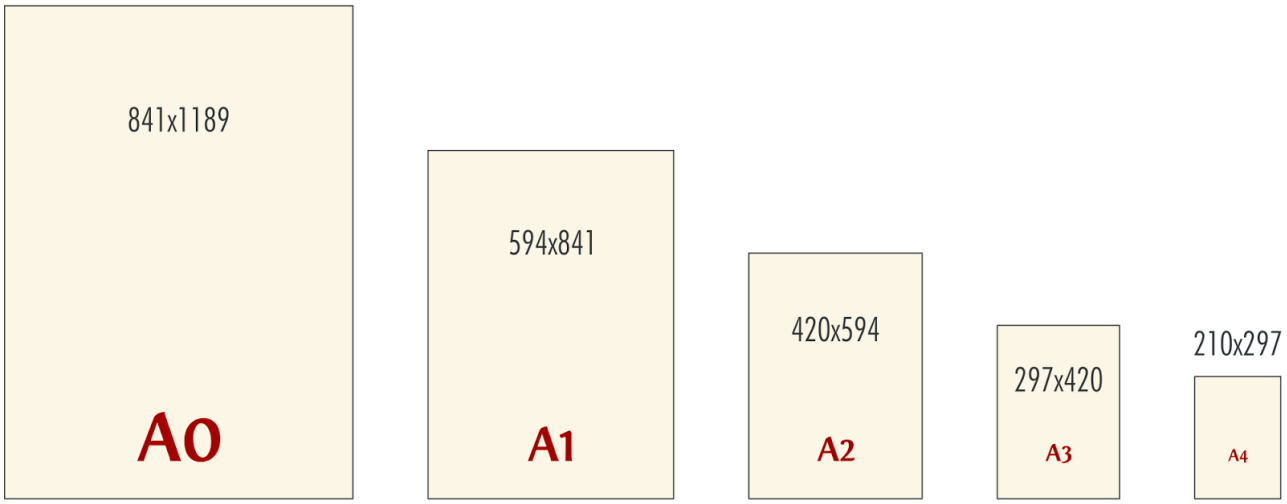
The scanner settings ought to ensure the quality parameters defined in the table below for the different document types.

TIFF master

Type of document	File format	Colour	Optical resolution[11]
Graphic material (Photography, Prints, Drawings, Paintings, Posters, Maps, Geo-graphic Maps...)	TIFF 6.0, uncompressed	Colour profile “Adobe RGB” to 24 bit (8 bits per channel). For documents requiring the highest quality: Colour profile “ProPhoto RGB) to 48 bit (16 per channel)	Format up to A4: 600 dpi. Larger than A4: 400 dpi. For large and small formats, adjust the resolution in order to get the best results
Books, journals and manuscripts, rare or valuable (e.g. illustrated or painted)	TIFF 6.0, uncompressed	Colour profile “Adobe RGB” to 24 bit (8 bits per channel). For documents requiring	Format up to A4: 600 dpi. Larger than A4: 400 dpi. For large and small formats, adjust the

[9]**Colour space** (or colour mode or gamut) is the range of colours that can be displayed on monitors or printers. The widest range of colours is represented by the natural spectrum, containing all the colours the human eye can perceive, but the acquisition devices (cameras, scanners) can have limitations in reproducing the entire colour range, and so do the reproduction devices (monitors, printers); this is why they use limited colour spaces. The most commonly used colour spaces for digital capture, in order of decreasing quality, are ProPhoto RGB, Adobe RGB (1998), sRGB. The latter is a very limited space, but still sufficient for viewing on the monitor (which in most cases, unlike printers, cannot represent a larger colour space).
Colour depth represents the number of shades that can be achieved for each of the three primary colours (red, green, blue). There are 3 levels of depth in order of decreasing quality : 32, 16, 8 bits per channel. The third level is adequate for monitor viewing.
[10]For types and characteristics of scanners see Ministry of Culture, *Linee guida per la digitalizzazione del patrimonio culturale*, 2022 (draft version 2022-2023), paragraph D.1.A.1 and following.
[11]The **optical resolution** is measured in dpi, (dots per inch), a definition that refers to printing methods. Ppi, or "pixels per inch” describes the resolution of images displayed on a computer screen, although the two definitions are often used interchangeably. The overall quality is determined by the pixel size of the final image, so you need to adjust the dpi measurement to achieve the desired pixel dimensions. For example, scanning a 10-inch side photo on a 300-dpi flatbed scanner will produce an image of 3,000 pixels on the side, whilst a 5-inch photo should be scanned at 600 dpi to get the same size as 3,000 pixels. Also consider that, whilst in a flatbed scanner the dpi are preset by the operator, this is not possible when shooting with a digital camera, because the final size of the image and its resolution depend on the physical size of the camera sensor (in pixels / megapixels). If you want to provide print quality or display quality to certain monitors, you should assign a defined resolution in dpi to a camera image.

or with **poor readability** (faded characters, low contrast, margin notes in pencil, stained)		the highest quality: Colour profile “ProPhoto RGB” to 48 bit (16 per channel)	resolution in order to get the best results
Books, journals, manuscripts, typed and mimeographed, not rare nor valuable, easily readable	TIFF 6.0, uncompressed	Colour Profile “Adobe RGB” to 24-bit (8 bits per channel) or to 16-bit greyscale	Format up to A4: 400 dpi. Larger than A4: 300 dpi. For large and small sizes, adjust the resolution in order to get the best results
Negatives, Black and White Slides	TIFF 6.0, uncompressed	16-bit greyscale	From 35 mm to 10x12 cm: 800-2800 with a resolution based on 4000 pixels on the longest side. From 10x12 to 20x25 cm: 800-1200 with a resolution based on 6000 pixels on the longest side. > 20x25 cm: 800 with a resolution based on 8000 pixels on the longest side.
Negatives, Colour Slides	TIFF 6.0, uncompressed	Colour profile “Adobe RGB” to 24 bit (8 bits per channel). For document requiring the highest quality: Colour profile “ProPhoto RGB” to 48 bit (16 per channel)	From 35 mm to 10x12 cm: 800-2800 with a resolution based on 4000 pixels along the long side. From 10x12 to 20x25 cm: 800-1200 with a resolution based on 6000 pixels along the long side. > 20x25 cm: 800 with a resolution based on 8000 pixels on the longest side.



Paper size (mm). A0–A4 in Series A, International Standard ISO 216

Derivative files

The derived files are produced from the master file and optimised for different uses, for example for in-browser viewing, to be converted into text via OCR or for viewing on a dedicated PC. They are normally resized and compressed, even with loss of information (e.g. images in JPEG format), in order to be exploited in the most convenient way without excessive loss of quality.

The choice of what type of derived files to create depends on the needs of the digitisation project, taking into account the availability of tools and skills to process the files, the different uses envisaged, as well as the quality of the images to store in Phaidra.

The following information is meant to be guidance on the characteristics of derived files for different uses.

TIFF. In the case of camera capture in raw format, from the master file a derived file in uncompressed TIFF 6.0 format is obtained, with Intel byte order (PC); these images must be equivalent to the original and therefore not be altered in any way, except for the colour balance, an action that can be done successfully only in raw files.

JPEG. They are files intended for dissemination on the web. The chromatic, grey and metric scales in the master must be removed. In addition, the images must be balanced for colour, brightness, contrast and saturation in order to correct any chromatic aberration, due to the acquisition conditions, based on the samples returned by the colour scales and grey colours. The balancing must be performed starting from the master files, before exporting to JPEG, with a tool (namely with adequate colour checking and photo editing software); it must aim at the identical reproduction of the chromatic characteristics of the original, not at an arbitrary aesthetic improvement. Images will also need to be straightened and cropped for optimal viewing.

As for resizing you will have to make a compromise choice trying to maintain a good image quality, but without producing very heavy files. In Phaidra, even very large images (for example with a side

of 10,000 px) can be easily appreciated thanks to the IIPImage[12] viewer that allows the user to navigate and zoom without slowing down. However, it must be considered that the file size affects both the uploading time in Phaidra and the downloading time by the user.
For documents that require a very high detail display (for example, maps or ancient manuscripts containing minute glosses) it is recommended to follow the specifications indicated in the high-quality JPEG table, whilst in other cases the images can be scaled to smaller sizes than those of the masters, for example by following the indications in the jpeg table of medium quality

The image resolution does not influence the on-screen display, however, if you want good quality printing, it is recommended to use a 300 dpi resolution.

The JPEG compression quality must be assessed on a case-by-case basis, but it should not be less than 75%.

High-quality JPEG

Type of document	File format	Size	Colour	Optical resolution	Use
All documents in the Master File Table	JPEG compressed at medium-heigh quality	The same of master	Colour profile Adobe RGB (1998) and depth of 24 bits (8 bits per channel)	300 dpi	Maps or other documents that require a detailed view.

Medium quality JPEG

Type of document	File format	Size	Colour	Optical resolution	Use
All documents in the Master File Table	JPEG compressed at medium-heigh quality	Approximately 3000 pixels on the longest side	Colour profile Adobe RGB (1998) and depth of 24 bits (8 bits per channel)	300 dpi	Documents that do not require a detailed view

Texts to be subjected to OCR (Optical Character Recognition)

If you want to make text-searchable files available, the digitised images must be subjected to OCR In this case, you can create a searchable PDF, as well as various other formats depending on your needs (TXT, EPUB...)

[12] https://iipimage.sourceforge.io/

Document filed in Phaidra as a book

If you want to upload a “Book” in Phaidra as searchable text:

- the OCR must be performed at the same image size as those that will be uploaded to Phaidra
- for each image (therefore for each page of the book) you must create an XML file, which has the same name as the image file
- the XML file must be formatted as in the example in *Attachment 1. Specifications for XML files of texts to be subjected to OCR*
- a searchable PDF must be created

A document filed in Phaidra as searchable PDF

A possible workflow for creating a searchable PDF is described in the document "Creazione di PDF ricercabili con software libero. Caso studio: i Bollettini-Notiziari della Facoltà di Scienze Statistiche"[13].

Document archived in Phaidra as an image or as a simple PDF and processed by Internet Archive

For objects with a licence allowing their reuse, a workflow has been set up to export them semi-automatically from Phaidra to the Internet Archive with a double result: greater international visibility and the provision in Phaidra of text files created by the Internet Archive (searchable PDFs, TXT ...). In order to use this option, you must make a request to the Library System Helpline[14].

4.4 File Names

In general, the name of each file will be a character string composed of several parts, having therein the information necessary to uniquely identify the project document to which the image refers. File names will be completed with the appropriate extension (tif, jpg, pdf, xml).

In mass storage, image files will be organised in multiple folders, in order to preserve the overall ordering of materials.

The nomenclature of the folders and files is a string of fields (library code[15], shelf mark...) separated by a hyphen (-). Where the shelf mark contains a hyphen (-), spaces or special characters, they are replaced by a dot (.).

[13] After authentication https://bibliotecadigitale.cab.unipd.it/collezioni_navigazione/Members/bibliotecari/materiali_settore_bd/gl-biblioteca-digitale/gruppo-phaidra/factory/creazionepdfricercabili.pdf

[14] https://bibliotecadigitale.cab.unipd.it/aiuto

[15] Originally, the University Library System adopted the SBN registry codes to identify the individual libraries (after authentication, see: https://bibliotecadigitale.cab.unipd.it/collezioni_navigazione/Members/bibliotecari/materiali_settore_bd/gl-biblioteca-digitale/gruppo-phaidra/factory/digitalizzazione/organizzazionepolifondi_ott2018.pdf). Starting from 2020, following the reorganization of the libraries and the new library management software, the University of Padua Library System has adopted a new code system (after authentication, see: https://bibliotecadigitale.cab.unipd.it/collezioni_navigazione/Members/bibliotecari/materiali_settore_bd/gl-biblioteca-digitale/gruppo-phaidra/factory/digitalizzazione/codicifondisbaalephalma.ods). If it is a local hub project, use the local hub code.

To facilitate quality control, it is recommended not to include more than 200 pictures in folders for TIF files, or more than 100 images if they are large format documents. In these cases, subdivide the folder into more, consecutively-numbered folders.

For graphic material and archive material that are scanned on both sides, follow the progressive numbering of “-r” files for the recto, and “-v” files for the verso.

For books, front and back covers are named so that they occur in the same order they have in the physical document. The spine or other parts of the original document (textblocks, binding details ...) must be included at the end.

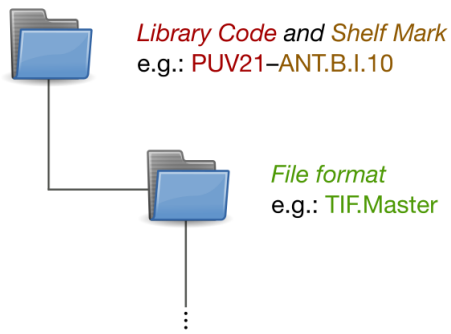
The image that includes the colour scale, the greyscale and the metric scale, must be named so that it is the last file in the folder and a “-c” is added to the progressive numbering of the file.

Books

The main folder, named “Library Code – Shelf Mark”, will contain the following subfolders:

- DNG.Master, if the native format produced by the capture tool is a raw format
- TIF.Alta.Qualita, if the native format produced by the capture tool is a raw format; it will contain copies that are faithful to the original
- TIF.Master, if the native format produced by the capture tool is TIFF format
- JPG
- if OCR is required, PDF and XML subfolders, as well as a folder for each type of text file that may be present (TXT, EPUB...)

The file name will follow the following schema:
“Library Code – Shelf mark – Progressive Number.extension”



Example of folders structure and file name:
PUV21–ANT.B.I.10\TIF.Master\PUV21–ANT.B.I.10–0001.tif

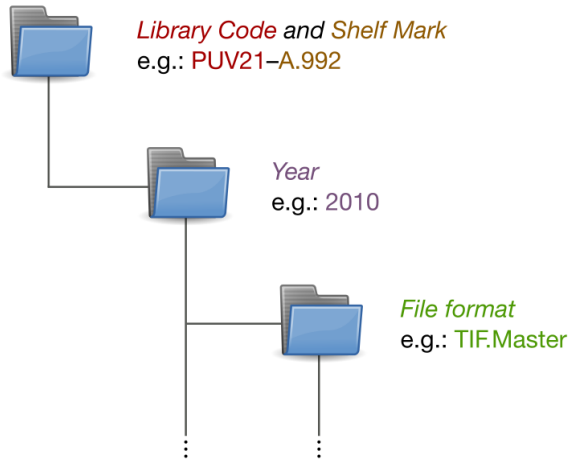
Journals

The main folder, named “Library Code - Shelf mark”, will contain a subfolder for each year of the journal.

Within individual years, there will be different folders for different types of files:

- DNG.Master, if the native format produced by the capture tool is a raw format
- TIF.Alta.Qualita, if the native format produced by the capture tool is a raw format; it will contain copies that are faithful to the original
- TIF.Master, if the native format produced by the capture tool is TIFF format
- JPG
- if OCR is required, PDF and XML subfolders, as well as a folder for each type of text file that may be present (TXT, EPUB...)

The files will be named as follows:
“Library Code – Shelf mark – Year – Month – Issue – Progressive Number.extension”



Example of folders structure and file name:
PUV21–A.992\2010\TIF.Master\PUV21–A.992–2010–12–24–0001.jpg

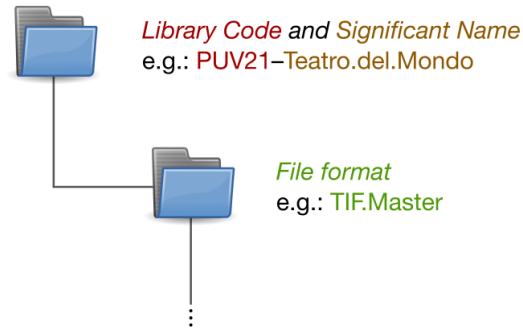
Photos, posters, maps (not bound in an atlas), parchments and other materials in loose sheets

The main folder will be called “Library Code – Significant Name”. The significant name will be created case by case at the time of digitisation. This folder will contain the following subfolders:

- DNG.Master, if the native format produced by the capture tool is a raw format
- TIF.Alta.Qualita, if the native format produced by the capture tool is a raw format; it will contain copies that are faithful to the original
- TIF.Master, if the native format produced by the capture tool is TIFF format
- JPG

- if OCR is required, PDF and XML subfolders, as well as a folder for each type of text file that may be present (TXT, EPUB...)

The file name will follow the following schema:
“Library Code – Significant Name – Progressive Number.extension”



Example of folders structure and file name:
PUV21–Teatro.del.Mondo\TIF.Master\PUV21–Teatro.del.Mondo–0001.tif

If necessary, distinguish recto from verso (e.g.: photography with information on the back):
PUV21-IB.Y.1\TIF.Master-3\PUV21-IB.Y.1-0001-r.tif
PUV21-IB.Y.1\TIF.Master-3\PUV21-IB.Y.1-0001-v.tif

Archive material

The main folder, named “Library Code – Collection Code – Series or Subseries Number – File or Subfile Number”, will contain the following subfolders:

- DNG.Master, if the native format produced by the capture tool is a raw format
- TIF.Alta.Qualita, if the native format produced by the capture tool is a raw format; it will contain copies that are faithful to the original
- TIF.Master, if the native format produced by the capture tool is TIFF format
- JPG
- if OCR is required, PDF and XML subfolders, as well as a folder for each type of text file that may be present (TXT, EPUB...)

The file name will follow the following schema:
“Library Code – Collection Code – Series or Subseries Number – File or Subfile Number – Progressive Number.extension”

4.5 Quality control

Quality control is aimed at ensuring good screen readability of the entire information content present in the original, this should be documented and maintained during the entire digitisation pro-

cess. Besides the on-screen control, it can be useful to do print tests to verify the quality of the image on paper.

Quality control planning includes:

- proper preparation of the environment (hardware configuration, visualisation software, viewing conditions, etc.)
- a priori definition of “acceptable” and “unacceptable” characteristics (e.g. whether optical deformations and/or aberrations may be acceptable or not)
- verification mode (any product or a sample, all files or only the master, visual screen quality and printing quality, instrumental verification, etc.).

The instrumental and visual inspection of an image usually involves:

- completeness of digitisation and correctness in the order of files
- correctness of framing and exposure
- control of the chromatic tolerance
- depth and colour profile
- digital size and format
- the presence of any elements which compromise the fidelity of the reproduction (light reflections, etc.);
- files and folders naming conventions

4.6 Data storage and conservation

It is essential to maintain digital assets created over time in order to avoid repeating the costly work of scanning, so procedures must be put in place to ensure that digital objects remain usable and accessible regardless of future changes in technology.
The usability and accessibility of digital objects over time is guaranteed by file format[16], by media storage and by the digital repository. It is essential to use open standards.

The Library System has defined a project archiving procedure which requires the project manager to deliver the digitization product to CAB (The University Library Centre) for secure storage at the ASIT data center (the Computer and Telematic Services Area of the University of Padova)[17].

For ease of access, the digital collections are also saved on optical or magnetic storage media (Blu-ray Disc, removable discs), in two copies, one to be kept in the library and the other at CAB. When copying files from one physical medium to another, integrity must be verified with the help of integrity verification software[18].

[16]On file formats suitable for long-term storage see the document *“Recommended file formats for long-term archiving and for web dissemination in Phaidra”* http://phaidra.cab.unipd.it/static/EN-file-formats.pdf
[17]After authentication, see the document *Archiviazione dei progetti di digitalizzazione SBA*.
[18]For example, in a Windows environment, you can use the open source software WinMerge.

5 Archiving in Phaidra

Archiving in Phaidra consists of uploading digitised files and entering the necessary data for the identification and description of the digital item.

For how to archive and compile metadata, see “Guida all’archiviazione” (Archiving Guide)[19].

It is possible that the object being archived is catalogued in other systems, such as an online catalogue or other platforms, so it is recommended to contact the Library System Helpline[20] to determine the procedure for the possible migration of data.

6 Further details

Selection of resources divided by topic.

6.1 Planning

ATHENAWP3 (edited by), *Digitisation Standard Landscape for European museums, archives, libraries,* 2009
https://phaidra.cab.unipd.it/o:6785

Cohen, Daniel J. – Rosenzweig, R., *Digital history : a guide to gathering, preserving, and presenting the past on the web,* [2005?]
https://chnm.gmu.edu/digitalhistory/

International Federation of Library Associations and Institutions (IFLA), *Linee guida per pianificare la digitalizzazione di collezioni di libri rari e manoscritti*, 2015
https://repository.ifla.org/handle/123456789/458

Istituto centrale per il catalogo unico delle bibliotyeche italiane e per le informazioni bibliografiche (ICCU), *Linee guida e standard*
https://www.internetculturale.it/it/1131/linee-guida-e-standard

Istituto centrale per il catalogo unico delle bibliotyeche italiane e per le informazioni bibliografiche (ICCU), *Normative catalografiche, linee guida e standard*
https://www.iccu.sbn.it/it/normative-standard/

Istituto Centrale per la Digitalizzazione del Patrimonio Culturale – Digital Library, *Il Piano nazionale di digitalizzazione. Versione in consultazione*, 2022
https://digitallibrary.cultura.gov.it/il-piano/

[19] https://phaidra.cab.unipd.it/static/guida-all-archiviazione.pdf
[20] https://bibliotecadigitale.cab.unipd.it/aiuto

Ministerial network for valorising activities in digitization (MINERVA), *Linee guida tecniche per i programmi di creazione di contenuti culturali digitali*, 2006
https://www.minervaeurope.org/publications/technicalguidelines_it.htm

Ministerial network for valorising activities in digitization (MINERVA), *Manuale di buone pratiche per la digitalizzazione del patrimonio culturale*, 2004
https://www.minervaeurope.org/publications/buonepratiche.htm

National Information Standards Organization (NISO), *A Framework of Guidance for Building Good Digital Collections*, 2007
https://www.niso.org/publications/framework-guidance-building-good-digital-collections

Northeast Document Conservation Center (NDCC), *Handbook for digital projects*, 2000
https://www.nedcc.org/assets/media/documents/dman.pdf

The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials, 2003
https://chnm.gmu.edu/digitalhistory/links/pdf/chapter1/1.17.pdf

6.2 Preservation

International Federation of Library Associations and Institutions (IFLA) Core Programme, Preservation and Conservation, *Principi dell'IFLA per la cura e il trattamento dei materiali di biblioteca*, 2004
https://repository.ifla.org/handle/123456789/1269

Northeast Document Conservation Center (NEDCC), *NEDCC Preservation Leaflets*
https://www.nedcc.org/free-resources/preservation-leaflets/overview

The Library of Congress, *Preservation, Collections Care*
https://www.loc.gov/preservation/care/

6.3 Digitisation

Association for Library Collections & Technical Services (ALCTS), *Minimum Digitization Capture Recommendations*, 2013
https://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations

Besser, Howard (revised by S. Hubbard, D. Lenert), *Introduction to Imaging*
https://www.getty.edu/research/publications/electronic_publications/introimages/intro.html

Federal Agencies Digitization Guidelines Initiative (FADGI) - Still Image Working Group, *Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files*, 2016
https://www.digitizationguidelines.gov/guidelines/FADGI Federal_Agencies Digital Guidelines Initiative-2016 Final_rev1.pdf

Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), *Linee guida e standard*
https://www.internetculturale.it/it/1131/linee-guida-e-standard

Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), *Linee guida per la digitalizzazione e metadati*
https://www.iccu.sbn.it/it/normative-standard/linee-guida-per-la-digitalizzazione-e-metadati/

Istituto Centrale per la Digitalizzazione del Patrimonio Culturale – Digital Library, *Il Piano nazionale di digitalizzazione. Versione in consultazione*, 2022
https://digitallibrary.cultura.gov.it/il-piano/

JISC, *JISC Digital Media* [archiviato 2016]
https://www.webarchive.org.uk/wayback/archive/20160101151305/http://www.jiscdigitalmedia.ac.uk/

Lotti, Franco – Trumpy, Giorgio, *La digitalizzazione documentaria*, CNR IFAC (Istituto di Fisica Applicata "Nello Carrara") e Fondazione Rinascimento Digitale, 2007
https://phaidra.cab.unipd.it/o:469492

Lotti, Franco – Lunghi, Maurizio – Trumpy, Giorgio, *Digitalizzazione di beni artistici e documentari. Manuale di procedure per un laboratorio fotografico digitale*, CNR IFAC (Istituto di Fisica Applicata "Nello Carrara") e Fondazione Rinascimento Digitale, 2009
https://phaidra.cab.unipd.it/o:469494

Osservatorio Tecnologico per i Beni e le Attività Culturali (OTEBAC), *Schema di capitolato per attività di digitalizzazione*, 2008
http://www.otebac.it/index.php?it/127/capitolato-tecnico-digitalizzazione

Research Libraries Group (RLG)*, RLG Model Request for Proposal (RFP) for Digital Imaging Services*, 1997
https://www.oclc.org/content/dam/research/activities/digimgtools/rlgmodelrfp.pdf

University of North Texas Libraries (UNT), *Digital projects unit*
https://library.unt.edu/digital-projects-unit/

U.S. National Archives and Records Administration (NARA), *Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images*, 2004
https://www.archives.gov/preservation/technical/guidelines.html

6.4 Metadata

Baca, M. (edited by), *Introduction to metadata*, 2008
https://www.getty.edu/research/publications/electronic_publications/intrometadata/index.html

Dublin Core Metadata Initiative, *Dublin Core User Guide*
https://www.dublincore.org/resources/userguide/

IEEE, *Standard for Learning Object Metadata*, 2009
https://ieeexplore.ieee.org/document/1032843

Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), *Linee guida e standard*
https://www.internetculturale.it/it/1131/linee-guida-e-standard

Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), *Linee guida per la digitalizzazione e metadati*
https://www.iccu.sbn.it/it/normative-standard/linee-guida-per-la-digitalizzazione-e-metadati/

National Information Standards Organization (NISO), *Understanding Metadata: What is Metadata, and What is it For?*, 2017
https://www.niso.org/publications/understanding-metadata-2017

The Library of Congress, *Standards*
https://www.loc.gov/librarians/standards

6.5 Guide di Phaidra

Andreoli, Lorisa [et al.], *University of Vienna Metadata UWmetadata*, 2019
https://phaidra.cab.unipd.it/static/phaidra-uwmetadata.pdf

Bettella, Cristiana, *PHAIDRA_DC Metadata Element Set*, 2018
https://phaidra.cab.unipd.it/static/phaidra_dc-metadata-element-set.pdf

Cappellato, Linda [et al.], *Guida all'archiviazione*, 2019
https://phaidra.cab.unipd.it/static/guida-all-archiviazione.pdf

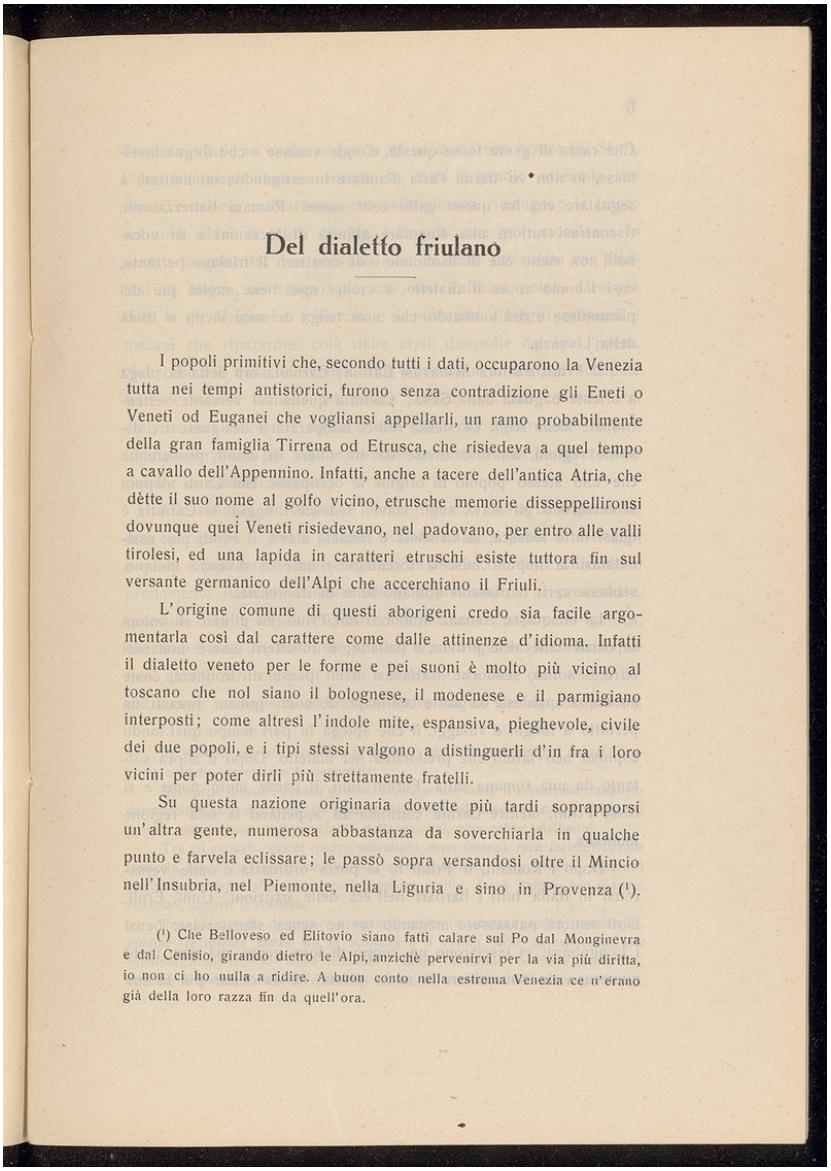
Drago, Gianluca (a cura di), *Formati dei file raccomandati per l'archiviazione a lungo termine e per la disseminazione web in Phaidra*, 2019
http://phaidra.cab.unipd.it/static/EN-file-formats.pdf

7 Contacts

For more information on digitisation and Phaidra write to the University of Padua Library Helpline.
Ask the Library Helpline.

Attachment 1. Specifications for XML files of texts to be subjected to OCR

The XML file must have the same name as the image file to which it refers (e.g. the image *page1.jpg* must correspond to an XML file named *page1.xml*).



Del dialetto friulano

I popoli primitivi che, secondo tutti i dati, occuparono la Venezia tutta nei tempi antistorici, furono senza contradizione gli Eneti o Veneti od Euganei che vogliansi appellarli, un ramo probabilmente della gran famiglia Tirrena od Etrusca, che risiedeva a quel tempo a cavallo dell'Appennino. Infatti, anche a tacere dell'antica Atria, che dètte il suo nome al golfo vicino, etrusche memorie disseppellironsi dovunque quei Veneti risiedevano, nel padovano, per entro alle valli tirolesi, ed una lapida in caratteri etruschi esiste tuttora fin sul versante germanico dell'Alpi che accerchiano il Friuli.

L'origine comune di questi aborigeni credo sia facile argomentarla così dal carattere come dalle attinenze d'idioma. Infatti il dialetto veneto per le forme e pei suoni è molto più vicino al toscano che nol siano il bolognese, il modenese e il parmigiano interposti; come altresì l'indole mite, espansiva, pieghevole, civile dei due popoli, e i tipi stessi valgono a distinguerli d'in fra i loro vicini per poter dirli più strettamente fratelli.

Su questa nazione originaria dovette più tardi soprapporsi un'altra gente, numerosa abbastanza da soverchiarla in qualche punto e farvela eclissare; le passò sopra versandosi oltre il Mincio nell'Insubria, nel Piemonte, nella Liguria e sino in Provenza (1).

(1) Che Belloveso ed Elitovio siano fatti calare sul Po dal Monginevra e dal Cenisio, girando dietro le Alpi, anzichè pervenirvi per la via più diritta, io non ci ho nulla a ridire. A buon conto nella estrema Venezia ce n'erano già della loro razza fin da quell'ora.

From an image like the one above (https://phaidra.cab.unipd.it/o:83943) an XML file formatted like this should be obtained:

```
<?xml version='1.0' encoding="UTF-8"?>
<ns0:ocrtext xmlns:ns0="http://phaidra.univie.ac.at/XML/book/ocrtext/V1.0">
 <ns0:page pid="o:4" abspagenum="9">
```

<ns0:ocrword word="Del" x1="542" x2="620" y1="471" y2="515">
<ns0:ocrchar char="D" x1="542" x2="581" y1="471" y2="514"/>
<ns0:ocrchar char="e" x1="585" x2="606" y1="492" y2="515"/>
<ns0:ocrchar char="l" x1="609" x2="620" y1="475" y2="515"/>
</ns0:ocrword>
<ns0:ocrword word="dialetto" x1="652" x2="821" y1="475" y2="517">
<ns0:ocrchar char="d" x1="652" x2="676" y1="475" y2="515"/>
<ns0:ocrchar char="i" x1="681" x2="690" y1="478" y2="514"/>
<ns0:ocrchar char="a" x1="694" x2="718" y1="492" y2="515"/>
<ns0:ocrchar char="l" x1="722" x2="731" y1="475" y2="514"/>
<ns0:ocrchar char="e" x1="735" x2="757" y1="492" y2="516"/>
<ns0:ocrchar char="t" x1="761" x2="774" y1="478" y2="516"/>
<ns0:ocrchar char="t" x1="778" x2="792" y1="477" y2="516"/>
<ns0:ocrchar char="o" x1="794" x2="821" y1="493" y2="517"/>
</ns0:ocrword>

[omissis]

<ns0:ocrword word="fin" x1="528" x2="557" y1="2038" y2="2058">
<ns0:ocrchar char="f" x1="528" x2="536" y1="2038" y2="2057"/>
<ns0:ocrchar char="i" x1="536" x2="543" y1="2038" y2="2058"/>
<ns0:ocrchar char="n" x1="545" x2="557" y1="2044" y2="2057"/>
</ns0:ocrword>
<ns0:ocrword word="da" x1="575" x2="604" y1="2038" y2="2059">
<ns0:ocrchar char="d" x1="575" x2="589" y1="2038" y2="2058"/>
<ns0:ocrchar char="a" x1="592" x2="604" y1="2045" y2="2059"/>
</ns0:ocrword>
<ns0:ocrword word="quell’ora." x1="621" x2="746" y1="2038" y2="2064">
<ns0:ocrchar char="q" x1="621" x2="635" y1="2045" y2="2064"/>
<ns0:ocrchar char="u" x1="639" x2="651" y1="2045" y2="2058"/>
<ns0:ocrchar char="e" x1="654" x2="666" y1="2043" y2="2058"/>
<ns0:ocrchar char="l" x1="669" x2="674" y1="2038" y2="2057"/>
<ns0:ocrchar char="l" x1="678" x2="683" y1="2038" y2="2057"/>
<ns0:ocrchar char="’" x1="686" x2="691" y1="2038" y2="2046"/>
<ns0:ocrchar char="o" x1="697" x2="711" y1="2044" y2="2058"/>
<ns0:ocrchar char="r" x1="714" x2="723" y1="2045" y2="2058"/>
<ns0:ocrchar char="a" x1="725" x2="738" y1="2045" y2="2060"/>
<ns0:ocrchar char="." x1="741" x2="746" y1="2056" y2="2060"/>
</ns0:ocrword>
</ns0:page>
</ns0:ocrtext>

Attachment 2. Digitisation project information sheet[21]

This sheet indicates the set of prerequisites that should be highlighted in digitisation projects aimed at archiving documents and collections in Phaidra.

Project Information	
Project title	
Short description of the collection	
Structure (Department, Centre, Library, etc.)	
Scientific manager	The scientific manager (an expert or scientific committee) is the person who assumes responsibility for the selection of the materials and defines the quality of the metadata. In the selection phase, he/she is supported by the project manager, particularly for the examination of the materials and for legal aspects.
Name	
Phone	
E-mail	
Project Manager	The project manager cooperates with the scientific manager, supports the scientific manager in the analysis of the legal issues, coordinates the activities related to the digitisation and guarantees the quality of the metadata.
Name	
Phone	
E-mail	
Technical coordinator	As coordinator for technical-operational activities, he/she collaborates with the Phaidra Team, which in turn provides technical assistance.
Name	
Phone	
E-mail	
Short description of the collection and members involved	

[21]The Digitisation project information sheet can be downloaded, after authentication, from https://bibliotecadigitale.cab.unipd.it/collezioni_navigazione/Members/bibliotecari/materiali_settore_bd/gl-biblioteca-digitale/gruppo-phaidra/factory/digitalizzazione/elenco-dei-documenti-relativi-alla-digitalizzazione

Duration of the project	
Information on the original documents (books, collections of journals, atlases, maps, photographs, etc.)	
Dating	from_____to_____
Type	Estimated Quantity
▢ printed text	
▢ handwritten text	
▢ printed and handwritten music	
▢ map	
▢ poster	
▢ postcard	
▢ drawing	
▢ painting	
▢ print (engraving, etching, etc.)	
▢ parchment	
▢ negative b/w	
▢ negative col.	
▢ photograph b/w	
▢ photograph col.	
▢ slide b/w	
▢ slide col.	
▢ other (specify) _____	
Presentation of documents	▢ loose sheets ▢ rolled sheets ▢ bound ▢ album ▢ mounting on cardboard or other material ▢ mounting in a frame ▢ envelopes ▢ folders ▢ boxes ▢ other (specify) _____
Document Dimensions	< di A4 _____ A4

	_____ A3 _____ A2 _____ A1 _____ A0 _____ > di A0 _____ other (specify) _____
Total number of documents	
Information on digital objects (files produced)	
Estimated number of digital objects	
Intended use of digital objects	▢ web, open access * ▢ web, restricted access * ▢ access through local network ▢ CD-ROM or DVD ▢ print ▢ other (specify) _____ * Please note that digital objects located in Phaidra may be open access (any user can look at the preview, metadata and files) or restricted access (view metadata and a preview).
Preliminary checks	
Document sources	▢ acquisition ▢ donation ▢ I do not know ▢ other (specify) _____ _____

	_____
Has a selection of documents been made?	▢ yes ▢ partially ▢ no If so, what are the selection criteria? ▢ historical and cultural value ▢ uniqueness and rarity ▢ high demand ▢ material without legal constraints or with digitisation permits obtained ▢ access restricted due to the state of conservation, the value, or the location ▢ added value through online access, the creation of virtual collections, increased research interest for little known or unknown material, etc. ▢ other (specify) _____ _____
Was a review made?	▢ yes ▢ partially ▢ no
Is there a digitised version?	▢ yes ▢ no If not, which organisations, websites, catalogues, etc. have been checked? _____ _____ _____ _____
Are there legal restrictions (copyright, privacy protection, the donor's rights, etc.)?	▢ yes ▢ partially ▢ I do not know Any additional information: _____ _____
Are the documents	▢ yes, all ▢ yes, partially ▢ no ▢ I do not

described/catalogued?	know If yes, how? ▢ printed list ▢ digital list ▢ printed catalogue ▢ electronic catalogue ▢ printed archive inventory ▢ electronic archive inventory ▢ other (specify) _____ _____
In the case of printed text, is it intended to implement OCR (Optical Character Recognition)?	▢ yes ▢ no ▢ I do not know
In the case of handwritten text, is it intended to transcribe the documents?	▢ yes, all ▢ yes, partially ▢ no ▢ I do not know
Estimated project costs	If digitisation is in-house, indicate: • cost of equipment specifying the type of instrumentation _____ _____ • operator costs _____ If digitisation is outsourced, indicate: • unit cost _____ • total cost _____
Notes	

Sheet compiled by	
Date compiled	

The undersigned are aware of having to operate in compliance with current legislation on copyright.
The documents of this project, subject to the legislation in force on copyright, fall under one of the following conditions (tick one or more options):

▢ the rights holder is the University of Padua
▢ the rights holders have granted the University of Padua the right to use them
▢ are in the public domain

Signature of Scientific Manager _____

Signature of Project Manager _____