ICS 点击此处添加 ICS 号
CCS 点击此处添加 CCS 号

T/CADPA

中 国 音 像 与 数 字 出 版 协 会 团 体 标 准

T/CADPA XXXX—XXXX

古籍（书画类）资源数字出版应用指南 第 1 部分：采集

Application guidelines for digital publishing of ancient books (painting and calligraphy) resources
Part 1: Acquisition

（征求意见稿）

（本草案完成时间：2024-08-28）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 实施

中国音像与数字出版协会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由中国音像与数字出版协会出版融合工作委员会提出。

本文件由中国音像与数字出版协会归口。

本文件起草单位：XXX。

本文件主要起草人：XXX。

古籍（书画类）资源数字出版应用指南
第1部分：采集

1 范围

本文件规定了古籍（书画类）资源数字出版过程中对古籍（书画类）资源的实物数据和数字资源进行采集的基本内容、技术方法和存储要求。

本文件用于出版单位和文博单位对古籍（书画类）数字资源的采集工作。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 3792 信息与文献 资源描述

GB/T 21712 古籍修复技术规范与质量要求

GB/T 31219.2-2014 图书馆馆藏资源数字化加工规范 第2部分：文本资源

3 术语和定义

下列术语和定义适用于本文件。

3.1

古籍（书画类） ancient books(painting and calligraphy)

产生于1911年以前，以及1912年至1949年之间，以线装、经折装等具有中国古代传统装帧形式装帧的书画文献，以及中国传统绘画、中国传统书法艺术等历代创作实物。

3.2

古籍（书画类）资源 ancient books(painting and calligraphy)resources

古籍（书画类）作品及与之相关的所有资料。

3.3

数据采集 data acquisition

对古籍（书画类）资源数据进行收集的方法与过程。

3.4

非接触式采集 non-touchable acquisition

在指定场所中，通过拍摄、扫描等非直接接触方式获取特定古籍（书画类）实体相关信息和测量结果的采集类型。

3.5

接触式采集 touchable acquisition

在指定场所中，通过直接接触或微距观测的方式采集特定古籍（书画类）相关信息和检测结果的采集类型。

3.6

检测数据 measurement data

对古籍（书画类）进行相关检测产生并可供处理与研究的数据。

3.7

常规采集 routine acquisition

为古籍（书画类）建档所做的常规普查采集。

3.8

专项采集 dedicated acquisition

为特定目的而进行的古籍（书画类）整体或局部采集。

4 采集原则

4.1 全面性

采集的内容和过程应全面覆盖古籍（书画类）所有必要数据。

4.2 多维性

依据实际情况，对古籍（书画类）资源进行多角度多方法的采集。

4.3 准确性

采集过程中应保留原始完整数据，且真实反映数据采集所关联的实体。

4.4 可扩展性

当采集数据不能满足后续数据加工需要时，可扩展相应元素集。

5 采集内容

5.1 基本数据

对古籍（书画类）资源的基本数据项进行采集，具体内容详见表1所示。

表1 采集的基本数据项

数据项	描述
名称	指古籍（书画类）的现用名、曾用名、别名。
作者	指在古籍（书画类）的主要创作个人或团体。
时间	指古籍（书画类）创作时间。
地点	指古籍（书画类）的创作地。
材质	指承载古籍（书画类）的材料，可能是实物材料类型或电子存储介质。
类型	中国画按其使用材料和表现方法，分为水墨画、重彩、浅绛、工笔、写意、白描等；按其题材分为人物画、山水画、花鸟画等。中国书法按字体分为篆书、楷书、隶书、行书、草书等。
内容	指古籍（书画类）所承载的主题和内容。
技法	指古籍（书画类）的创作技法，如泼墨、泼彩、写意、工笔等。
题跋	指写在古籍（书画类）前后题记文字。
款识	指古籍（书画类）上的落款信息。
印章	指古籍（书画类）上的作者印章、题跋人印章和收藏人印章等。
流派	中国画流派可按用笔、用墨、用色等实际情况划分。中国书法流派可按媒介、风格、理念等实际情况划分。
形制	指古籍（书画类）的表现形式和样式。
装裱	指古籍（书画类）的装裱情况。
修复	指古籍（书画类）的修复方法、修复内容等相关修复资料。
规格	根据古籍（书画类）具体规格尺寸单独采集。
颜料	指古籍（书画类）所用的矿物颜料和植物颜料。
特色	指古籍（书画类）用笔用墨、结构章法、线条组合、品格内涵等。
存放	指古籍（书画类）的存放要求。
原始功用	指古籍（书画类）被创作时的使用用途与效果。

5.2 实物数据

5.2.1 度量数据

对古籍（书画类）资源的度量数据进行采集，具体内容详见表2所示。

表2　采集的度量数据项

数据项	描述
光通量	单位时间内光通过某一表面的总功率的物理量。
照度	光线强度在某一表面上的分布和密度。
光亮度	材料表面反射或发射光线强度。
RGB	以RGB格式采集的颜色信息。
CMYK	以CMYK格式采集的颜色信息。
Munsell	以Munsell格式采集的颜色信息。
白度	物体表面反射光的亮度程度。
匀度	色彩、亮度或纹理的一致性或均匀性。
相对反射率	物体表面相对于标准白色参照物的反射能力。
纸张含水率	纸张中所含水分的百分比。
质地情况	表面的材料、纹理、光泽、手感、细腻度等方面的特性。
厚度测量	物体沿着某一方向的尺寸或测量值。
杂质情况	化学杂质、非纤维杂质、着色剂杂质和其他杂质等。
磨损程度	材料表面的磨损程度或损耗程度。
织物形态	织物材料中经线和纬线的宽度和密度。
纤维分析	材料纤维的长度、宽度、密度、配比、帚化、磨损和断裂情况。
材质表面pH值	材质表面的酸性或碱性程度。
表面层纹	材料表面的纹理、花纹或纹理图案。
渗透痕迹	对象上可见的外来物质渗入或渗出的痕迹。
着色分析	着色均匀性、渐变、分布、明暗、深浅特征。

5.2.2　普通光环境成像数据

古籍（书画类）普通光环境成像数据的采集应按照表3的内容进行。

表3　采集的普通光环境成像数据项

数据项	描述
数字图像	二位数字组形式表示，由模拟图像数字化得到，以像素为基本元素，可以用数字计算机或数字电路存储和处理的图像。
表面三维显微图像	显微镜或表面扫描技术获取的古籍（书画类）表面的立体图像。
霉变部位显微图像	用显微镜观察和记录霉变部位的图像。
污迹部位显微图像	使用显微镜观察和记录污迹部位的图像。
水渍部位显微图像	使用显微镜观察和记录水渍部位的图像。
老化部位显微图像	使用显微镜观察和记录古籍（书画类）上的老化部位的图像。
偏光显微图片	通过偏光显微镜观察和记录材料光学特性的图像。
颗粒分析	颗粒的外观、形状、尺寸、结构特征以及经染色处理后的颜色等。
消光特性	材料对光的吸收、散射、透过和反射能力。
超景深显微图片	使用特殊的图像处理算法来获取具有扩展焦深的显微图像。
边缘附着特征	对象边缘可见的灰尘、污渍、颜料溢出、风化等痕迹。
有色颗粒物	对象表面残留的可见的灰尘、污渍、颜料等小颗粒物。

5.3　关联数据

对古籍（书画类）关联数据的采集应符合表4的规定。

表4　采集的关联数据项

类型	内容
创作	描述古籍（书画类）所涉及创作相关的文字、图像、视频等资源，内容应包括：
	（1）创作工具；
	（2）创作材料；
	（3）创作技艺；
	（4）创作过程；

类型	内容
	（5）创作相关的活动；
	（6）创作相关的实物。
学科技艺	描述古籍（书画类）所涉及学科及技艺相关的文字、图像、视频等资源，内容应包括：
	（1）相关学科发展情况；
	（2）作者相关学科背景；
	（3）相关技艺发展情况。
文献资料	反映古籍（书画类）相关文献资料的文字、图像等资源，内容应包括：
	（1）古籍按古籍的分类方式标注，包括书名、作者(编纂者)、版本、章节、刊印机构、朝代(或再版时间)、内容简介、收藏情况；
	（2）普通图书包括书名、作者、出版社、出版时间、版别、内容简介、保存情况；
	（3）期刊报纸包括篇名、作者、刊名、文章栏目(或版别)、摘要、发表时间；
	（4）手稿、内部资料等非正式出版物包括名称、撰写人、刊(誊)印机构、年代、收藏或保存；
	（5）采集的文献资料图像应包括：封面、版权页、目录页、重要内容页等。
保护情况	描述古籍（书画类）的保护情况的文字、图像、视频等资源，内容应包括：
	（1）古籍（书画类）保护的长期、中期、短期规划和执行情况；
	（2）政策法规、规章制度；
	（3）负责保护的专门机构和民间机构及其参与形式；
	（4）采集与机构相关的实物照，对于项目保护单位应拍摄项目入选名录证书、所获荣誉的图片资料；
	（5）保护措施情况，包括：经费投入情况；相关组织（或机构和个人）的参与情况；保护单位和传承人权利、义务的落实情况；相关机构参与或组织保护传承活动的情况；保护措施所取得的成效；保护措施的更新情况和更新周期。
组织机构	描述收藏古籍（书画类）的组织机构的相关文字、图像、视频等资源，内容应包括：
	（1）名称、性质、职能、规模、现状等；
	（2）历史沿革；
	（3）主要成员，应包括主要负责人、骨干人员等；
	（4）作用与贡献；
	（5）组织机构相关的活动；
	（6）组织机构相关的实物。
相关评价	描述相关人员对古籍（书画类）的评价相关文字、图像、视频等资源，内容应包括：
	（1）古代相关领域人员对作者和作品的评价；
	（2）现代相关领域人员对作者和作品的评价；
	（3）古籍（书画类）的再创作资料。
	（4）作者的创作资料；
	（5）对书画艺术品做出评价的相关典籍文献；
	（6）现今相关领域人员的资料；
作者社会网络	对作者的亲友、同事、师承派别、社会团体、赞助商、社会支持者等相关社会网络进行描述的文字、图像、视频等资源，内容包括：
	（1）姓名、性别和年龄；
	（2）与作者的关系、称谓；
	（3）与作者共同参与的活动；
	（4）与作者之间的往来通信、谈话、报道、合照等；
作者时空行迹	描述作者创作时空行迹相关的文字、图像、视频等资源，内容应包括：
	（1）创作时间；
	（2）创作地域及相关历史沿革；
	（3）创作的时空变迁过程；
	（4）创作活动对创作地域的社会影响
	（5）作者时空行迹中相关活动；
	（6）作者时空行迹中相关实物。
作品时空行迹	描述作品流转的时空行迹相关的文字、图像、视频等资源，内容包括：
	（1）流转时间；
	（2）流转地域；
	（3）流转事件，例如何人为该作品的流转从事何种活动；

类型	内容
	（4）作品对流转地域的社会影响；
	（5）作品流转相关的实物。

6 采集技术

6.1 文字识别技术

采用人工识别、机器识别等方式对古籍（书画类）实物数据与关联数据中所涉及的手写或印刷文字（如现代汉字、古代汉字、书法字等）进行识别。

6.2 增强识别技术

增强识别是一种通过使用计算机算法和传感器来提高对现实世界中的物体、场景或现象的识别能力的技术，可对图像、音频和视频等信息进行处理，从而提高识别准确率和效率的方法。可运用增强识别技术辅助古籍（书画类）采集：

1) 通过图像增强技术，改善古籍（书画类）图像质量，例如调整亮度、对比度、饱和度等，使古籍（书画类）作品的细节更清晰。
2) 使用字符识别技术识别古籍（书画类）的文字部分，将手写或印刷的文字转换为可编辑文本。
3) 提取古籍（书画类）作品的特征，如线条、颜色和纹理等，辅助作品智能分类。

6.3 语音识别技术

一种机器自动将人的语音内容转录为文字的技术。必要时，可使用该技术采集古籍（书画类）相关的视频、音频资料中的语音内容，将其转录为文字资料进行采集。

6.4 数字视频转换技术

将某个视频信号从一种数字格式转换为另一种数字格式的技术，可能涉及视频格式（如AVI、MP4、MOV）转换、视频编码（如H.264、H.265、MPEG-2等）转换、分辨率转换和帧率转换等。使用数字视频转换技术，可以将不同格式的数字视频按照需求相互转换，提升古籍（书画类）数据采集和存储的工作效率与质量。

7 采集流程

7.1 流程说明

古籍（书画类）资源的采集需由持有方与采集方共同参与。具体采集流程见图1所示。

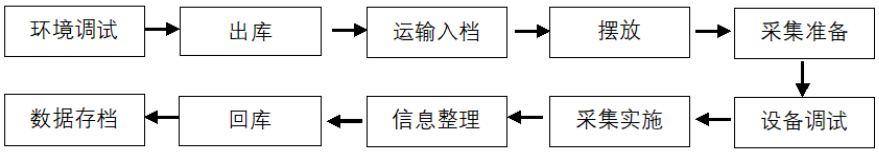

图1 采集流程图

7.2 环境调试

采集方对古籍（书画类）的采集空间和采集环境进行测试、调整。

7.3 出库

古籍（书画类）资源持有方由专人完成出库登记并将古籍（书画类）从库房运出到指定区。

7.4 运输入档

古籍（书画类）资源持有方将需要采集的古籍（书画类）从库房运输至采集区，协助采集方展开采集和入档登记。

7.5 摆放

古籍（书画类）资源持有方将运输完成的古籍（书画类）陈列在出入库存储区，与采集方交接并摆放妥当。

7.6 采集准备

采集方在采集古籍（书画类）前搭建好拍摄环境，使拍摄设备、工作人员准备就绪。

7.7 设备调试

采集方在开始采集前，反复测试与调整设备参数，使设备严格满足古籍（书画类）的采集要求。

7.8 采集实施

采集方对古籍（书画类）进行采集、检查采集效果。

7.9 信息整理

采集方拍摄完成后，对采集得到的各类数据进行整理并标注备份。

7.10 回库

古籍（书画类）资源持有方对完成采集的古籍（书画类）进行登记和入库信息存档，并将其搬运回仓库。

7.11 数据存档

采集方将已采集到的数据存档入库。

8 采集数据存储

8.1 数字文本

数字文本的存储宜使用DOC、DOCX、PDF、XLS 或 XLSX 等主流格式，也可根据需求使用XML、HTML等格式。同时，使用关系数据库或图数据库以结构化地存储数字文本，对古籍（书画类）资源进行高效数据查询和管理。也可使用知识图谱技术存储古籍（书画类）数字资源，进行复杂查询和分析操作。数字文本编排的格式项、页码、字体字号和要求则按实际需求决定。

8.2 数字图像

数字图像存储的类型、精度、规格、编码格式等要求详见表5。

表5 数字图像存储要求

存储内容		存储要求
图像类型	灰度图像、RGB图像、CMYK图像、多色调图像、索引彩色图像等	应对图像进行色彩校正，确保扫描影像的色彩忠实于原影像载体。对于获取的古籍（书画类）数据要根据原件进行修正（纠正歪斜的图像），去除图像中由于原稿的问题所留下的污点、霉斑、刮痕等不属于原件信息的缺陷。对于反映古籍（书画类）信息部分的图像基本不做处理，以避免造成信息误差。
图像精度	图像比例尺	需按原件尺寸的 100%扫描、加工。分为多部分扫描的影像数据应进行拼合，每一部分的色调、对比度、明暗度要保持基本一致。
	图像分辨率	珍贵品需采用光学分辨率 600-1200 dpi（含）以上，普通品需采用光学分辨率 300-600 dpi（含）以上。
图像规格	检测规格	检测应减少误差，提高精度，图像精度应当能够与检测精度相匹配。
	微区数据定位	检测数据于数字图像上的精确定位，对数据位置的描述需采用三维坐标，图像坐标的定位需考虑图像精度和检测精度，并且图像精度应当能够与检测精度相匹配。
编码格式	GIF、JPEG、JPEG2000、PDF、PSB、PSD、RAW、TIFF等	根据需求选择不同的图像编码格式。

8.3 数字音频

数字音频存储宜使用CD、MP3、WAV、FLAC、AAC、OGG等主流音频格式。

8.4 数字视频

数字视频存储宜使用MP4、MOV、MPEG、FLV、WebM、MKV、AVI等主流格式。

9 数据加工

9.1 一般数据加工流程

9.1.1 数据清洗

数据清洗是一个反复的过程，需要不断地发现清洗过程中的各类问题并解决问题。数据清洗的过程中需要做好被清洗数据的备份工作，防止数据的丢失。

9.1.1.1 清洗内容

对采集的古籍（书画类）数据进行的删除、添加、分解或重组等操作，保留有效古籍（书画类）数据。通过比较详细的数据分析来检测数据源中的错误或不一致。对于古籍（书画类）数据（数据样本）的分析一般采用手工检查、借助分析程序检查或两者相结合的方法。

9.1.1.2 定义转换规则

根据数据分析得到的结果定义数据清洗的转换规则。根据数据源的个数及数据源中数据的质量，为数据清洗和转换选定一种算法，提高数据自动转换的效率。

9.1.1.3 验证评估

数据清洗前对预先定义的数据清洗转换规则的正确性和清洗的效率进行验证和评估。一般是在数据源中选择数据样本进行清洗验证，当测试结果不满足数据清洗要求时需要对原有的数据清洗转换规则进行调整和改进。

9.1.1.4 加工清洗

在数据源上执行预先设计好并且已经得到验证的数据清洗转换规则。在源数据上对数据进行清洗前，需要对源数据进行备份，以防源数据的丢失或损坏。

9.1.1.5 干净数据回流

用被清洗的干净数据替换数据源中原有的数据。提高原有数据库中数据的质量，避免再次抽取数据时进行重复的清洗工作。

9.1.2 数据转换

数据从一种系统转移到另一种系统时，需要将古籍（书画类）数据转换到系统能够识别的格式。

9.1.3 数据标注

采用人工标注或机器标注的方式对未处理的古籍（书画类）初级数据，进行图像标注、文本标注、语音标注、视频标注和时空信息标注，为机器学习提供训练数据，支持机器学习的分类、识别或预测。

9.1.4 数据挖掘

从采集到的大量的、不完全的、有噪声的、模糊的、随机的古籍（书画类）数据集中识别有效的、新颖的、潜在有用的以及最终可理解的数据模式。

9.1.5 数据分析

用适当的方法对收集来的古籍（书画类）一手资料和第二手资料进行分析，最大化地发掘数据资料的价值。

9.2 数字图像加工流程

9.2.1 形变矫正

对数字图像进行空间矫正和几何矫正。对数字图像进行空间矫正，由于古籍（书画类）资源形态各异，在采集中可能发生图像空间形变，需要多次拍摄选取最优图像数据。对数字图像进行几何矫正，由于拍摄姿态和扫描非线性可能引起图像几何失真，首先需根据得到的图像数据建立几何数学模型，其次利用已知条件确定模型参数，最后根据模型对图像进行几何矫正。

9.2.2 色彩矫正

采用光学概念的三基色RGB与三补色CMY进行互补纠色，以校正照片和图像的偏色。古籍（书画类）数据色彩校正应照顾到摄影照片色调还原的全局,既要符合人眼在采集现场看到的感受又要遵循摄影成像和成色的科学规律。

9.2.3 图像增强

有目的地配合出版需要，针对领域专家研究需求，将原来较模糊的局部图像或某些特征点进行图像增强，使图像局部得到锐化。处理后的图像与原始图像存在一定的差异，可保留关键可利用的图像数据资源。

9.2.4 图像拼接

将连续拍摄的图片自动进行拼接，合成完整照片。完整的照片大图一般无法按照原精度输出，只做备份用，实际使用需按照1200/600/300/150dpi进行psb、psd、tiff、jpg或png格式输出，同时记录照片存储标准。

参 考 文 献

[1] GB/T 21712-2008 古籍修复技术规范与质量要求
[2] GB/T 31219.2-2014 图书馆馆藏资源数字化加工规范 第2部分：文本资源
[3] GB/T 3792-2021 信息与文献 资源描述