

T/CCUA

中国计算机用户协会团体标准

T/CCUA XXXX—XXXX

版本典藏网站资源长期保存 流程管理要求

Long-term preservation collected website resources Process management requirement

（征求意见稿）

（本草案完成时间：2024-06-18）

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

中国计算机用户协会 发布

目 次

前 言 II

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 目的 2

5 基本原则 2

6 管理流程 2

7 获取阶段 3

 7.1 策划 3

 7.2 接收 3

 7.3 采集 3

8 处理阶段 3

 8.1 整理 3

 8.2 质检 4

 8.3 审核 4

9 存储阶段 4

 9.1 保存 4

 9.2 加密/解密 4

 9.3 压缩/解压缩 4

 9.4 校验 5

 9.5 备份/恢复 5

 9.6 迁移 5

 9.7 销毁 5

10 应用阶段 6

 10.1 应用 6

参 考 文 献 7

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。

本文件由中国计算机用户协会提出并归口。

本文件起草单位：中国计算机用户协会创新技术应用分会、中国国家版本馆、江苏嘉图网络科技有限公司

本文件主要起草人：刘成勇、王志庚、张琦、吕健泳、张颖，马驰、马腾飞、王文清、周长岭

本文件于 XXXX 年首次发布。

版本典藏网站资源长期保存 流程管理要求

1 范围

本文件规定了在网站资源长期保存活动中具有保存价值的网站资源的获取、处理、存储与应用的一般流程。

本文件适用于图书馆、企事业单位和其他组织对网站资源进行长期保存管理的过程。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1 长期 long-term

虑及技术变化(包括支持新载体、新数据格式)和用户群体变化对典藏信息影响的足够长的时间段。这个时间段延伸到无限未来。

[ISO 14721:2012 定义1.7.2]

3.2

长期保存 long-term preservation

一种可靠的长期维护版本典藏网站资源真实、完整、有效的行为。

注：改写DA/T 47—2009 3.3

3.3

真实性 authenticity

网站资源的内容、逻辑结构和形成背景与形成时的原始状况相一致的性质。

注：改写GB/T 18894—2016 3.5。

3.4

可靠性 reliability

网站资源的内容完全和正确地表达其所反映的事务、活动或事实的性质。

注：改写GB/T 18894—2016 3.6。

3.5

完整性 integrity

网站资源的内容、结构和背景信息齐全且没有破坏、变异或丢失的性质。

注：改写GB/T 18894—2016 3.7。

3.6

可用性 useability

网站资源内容可以被检索、呈现或理解的性质。

注：改写GB/T 18894—2016 3.8。

3.7

采集 capture

对网站资源及其元数据进行收集、著录和存储的方法与过程。

注：改写GB/T 18894—2016 3.8。

3.8

迁移 migration

在维护真实性、完整性和可用性的前提下,将网站资源数据从一个系统转移到另一个系统或从一种格式转换成另一种格式的过程。

注：改写GB/T 18894—2016 3.16。

3.9 活动 activity

长期保存流程的组成部分，涉及或影响资源对象或其元数据的行为。

4 目的

版本典藏网站资源长期保存活动流程管理要求确保各类资源能够得到科学、有效且长期稳定的保存，防止资源在长期保存活动中出现损坏、丢失或篡改等情况；同时确保资源的可利用性，为公众提供快速、准确地检索和浏览服务。

5 基本原则

版本典藏网站资源长期保存活动流程管理应遵循以下原则：

- a) 完整性保护原则：资源在保存过程中必须保持其原始性、真实性和完整性，防止任何形式的损坏、篡改或丢失，确保资源在任何情况下都能得到完整保护；
- b) 可持续性原则：资源长期保存要考虑长期性和可持续性，要兼顾当前和未来的需求，确保资源能够持续、稳定地保存下去，包括选择适当的存储介质、定期维护更新以及适应技术发展的变化；
- c) 安全性原则：资源长期保存必须确保安全性，防止未经授权的访问、泄露或破坏，包括严格的访问控制、加密措施和安全审计等，确保资源在存储、传输和应用过程中的安全；
- d) 标准化原则：资源长期保存要遵循相关标准和规范，确保长期保存工作的专业性和一致性；
- e) 知识产权保护原则：长期保存资源的应用不能涉及专利、著作权、版权等知识产权。

6 管理流程

版本典藏网站资源长期保存活动的管理流程包括四个阶段：获取、处理、存储和应用。

每个阶段由有序的不同活动组成，包括策划、接收、采集、整理、质检、审核、保存、备份/恢复、加密/解密、压缩/解压缩、校验、迁移、销毁和应用。

版本典藏网站资源长期保存管理流程如图1所示。

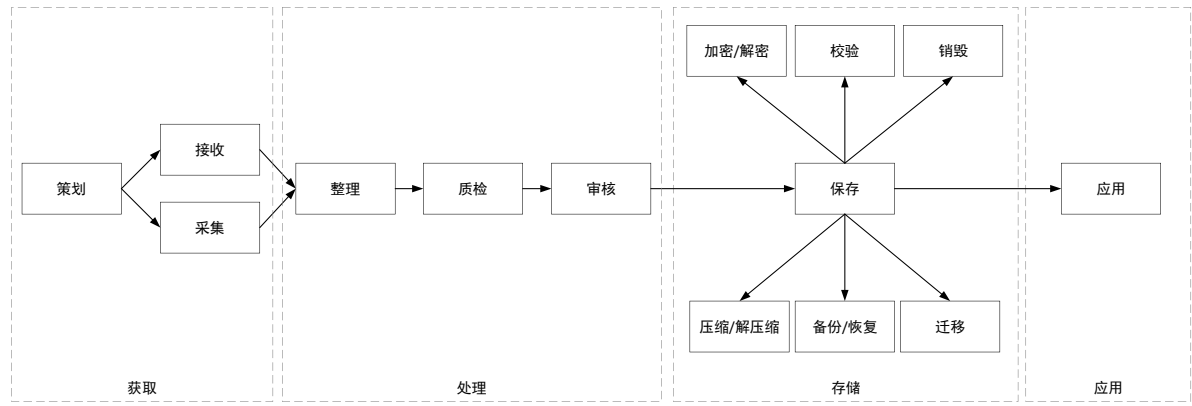


图1 版本典藏网站资源长期保存管理流程

7 获取阶段

7.1 策划

策划活动是长期保存活动的发起端，应做到以下内容：

- a) 策划活动应明确长期保存活动的目标网站，对目标网站资源进行全面的评估，包括网站的类型、规模、价值以及运行状况等，并根据评估结果制定针对性的保存策略，实现目标网站的有效保存和利用；
- b) 应针对不同类型的目标网站，制定相应的保存策略、技术选型和软件工具准备，同时，制定具体的执行措施，确保策略的有效实施；
- c) 应对长期保存活动中可能出现的风险进行预测和评估，制定相应的应对措施，如数据备份、应急预案等，以确保长期保存活动的安全和稳定；
- d) 在长期保存活动结束后，对策划和执行的效果进行全面的评估，总结经验教训，为后续的活动提供借鉴和参考。

7.2 接收

接收活动是从提交者处获取网站资源的活动，应做到以下内容：

- a) 接收活动应明确网站资源提交的流程和要求，如提交格式、质量标准和文件命名规则等；
- b) 接收网站资源时应确认资源接收的完整性和准确性，与提交者进行核对，生成接收报告或记录表，记录资源接收的详细信息，包括接收时间、数量、状态等；
- c) 应对收集到的资源进行初步整理，如去除重复项、处理损坏或缺失的部分，应根据资源的性质、类型或主题进行分类，记录资源的物理状态，如保存状况、完整性等，应对资源的质量进行初步评估，判断其是否满足长期保存的标准；
- d) 接收到网站资源后应将接收的资源正式交接给长期保存团队或部门，并通知相关人员。

7.3 采集

采集活动是主动获取在线网站资源的活动，应做到以下内容：

- a) 采集活动进行前应清晰界定采集的目标网站，明确需要收集的网站类型、领域或主题，应确定采集的时间范围和物理范围，以便有针对性地开展采集工作；
- b) 应根据网站资源的特点和保存需求，制定合适的采集策略，设定采集的质量标准，确保所采集资源的真实性和完整性，应选择适当的采集方法和工具，确保采集网站资源的全面性和有效性；
- c) 采集过程中应详细记录采集活动的过程，包括采集时间、地点、参与人员、方法等，应建立采集资源的元数据文档，记录资源的描述信息、来源、质量评估等；
- d) 采集过程中应严格遵守版权法律和知识产权规定，确保所采集资源的合法性，如有需要，与资源所有者或版权方进行沟通，获取必要的授权或许可；
- e) 应为采集到的资源提供合适的存储环境，采取必要的保护措施，应在采集完成后对采集到的资源进行质量检查，确保其完整性、准确性和可读性。

8 处理阶段

8.1 整理

整理活动是对接收或采集到的网站资源进行预处理的的活动，应做到以下内容：

- a) 应对接收或采集到的网站资源进行清洗和整理，包括清除重复数据、纠正错误数据、填充缺失值等，根据资源类型、主题或领域进行分类，采用统一的编号规则对资源进行编号；
- b) 根据长期保存活动元数据要求对网站资源著录元数据，包括不同粒度的知识对象、表现和文件等资源对象的元数据著录，元数据著录可采用自动提取、人工著录和两者结合的方式；
- c) 应对接收到的网站资源数据进行病毒查杀处理，查杀病毒应使用有效的软件工具以确保网站资源的安全；
- d) 整理过程应进行详细记录，包括整理的时间、人员、方法、资源状态和详情等。

8.2 质检

质检活动是对整理后的网站资源进行质量检验的活动，应做到以下内容：

- a) 应检查网站资源的完整性，使用技术手段或人工抽检验证数据是否完整，检查是否存在缺失的内容和元数据的缺失；
- b) 应检查网站资源的一致性检查，可使用抽检的方式检查网站资源与在线资源的一致性，定位不一致的位置和内容；
- c) 应检查网站资源的合规性，可使用抽检的方式检查网站资源的内容是否存在与国家、行业相关法律法规相悖的情况；
- d) 质检活动应在整理活动后及时进行，以确保网站资源的时效性；
- e) 质检活动应记录数据质检的过程、方法和结果，并生成网站资源质检报告，总结质检的结果和存在的问题，为后续的数据管理和利用提供参考。

8.3 审核

审核活动是对质检后的网站资源进行内容审核的活动，应做到以下内容：

- a) 应根据相关法律法规、行业标准和规定确定审核的标准，确保标准涵盖各类违规内容，如涉政、涉黄、暴力、侵权等；
- b) 应明确内容审核范围，包括文本、图片、视频、音频等各种形式的内容，根据不同内容类型的特点，制定相应的审核策略和方法；
- c) 宜在审核过程中利用智能审核技术，如文本识别、图片识别、语音识别等技术，提高审核效率和准确率，同时应结合人工审核，确保审核结果的准确性和可靠性；
- d) 应对敏感内容和侵权内容进行严格审查，并采取必要的处理措施，如删除、修改或限制访问；
- e) 应全面准确记录内容审核的结果和过程，及时向相关人员反馈审核结果，以便于后续追溯和审计。

9 存储阶段

9.1 保存

保存活动是对审核后的网站资源进行存储操作的活动，应做到以下内容：

- a) 应根据数据类型、大小、访问频率、存储介质的可靠性、容量、性能和成本，选择适当的存储介质，如硬盘、光盘、磁带或云存储等，确保满足长期保存的需求；
- b) 应同时保存网站资源数据及其元数据，并记录数据保存的过程、方法和结果，以便于后续追溯和审计；
- c) 应对存储环境进行定期监测，包括温度、湿度、电磁干扰等因素，定期对存储设备进行维护和检查，确保存储环境稳定可靠。

9.2 加密/解密

加密/解密活动是对正常存储状态的网站资源进行数据加密/解密操作的活动，应做到以下内容：

- a) 应根据存储和业务需要选择需要加密/解密处理的网站资源；
- b) 应根据数据的敏感性和存储需求，选择合适的经过广泛验证和认可的加密算法，如 AES、RSA 等；
- c) 应使用强随机数生成器生成密钥，确保密钥的随机性和不可预测性；
- d) 应妥善管理加密密钥，包括密钥的存储、分发、更新与销毁；
- e) 加密/解密操作应在安全的环境中进行，防止潜在的攻击或数据泄露，可以考虑使用专用的加密工作站或虚拟机，确保加密操作与其他系统和服务隔离；
- f) 加密/解密后的数据应按照保存活动的要求妥善存储，确保其可访问性和持久性；
- g) 应记录所有加密/解密操作、密钥管理活动和访问尝试，以便进行审计和追踪。

9.3 压缩/解压缩

压缩/解压缩活动是对正常存储状态的网站资源进行数据压缩/解压缩操作的活动，应做到以下内容：

- a) 应根据存储和业务需要选择需要压缩/解压缩处理的网站资源；
- b) 应根据数据的类型、特点和需求，选择适合的压缩算法，不同的算法有不同的压缩比率和性能特点，需要根据实际情况进行权衡，优先考虑使用无损压缩算法，以确保数据在压缩和解压缩过程中不会丢失任何信息；
- c) 在压缩/解压缩过程中，应确保数据的完整性不被破坏，可使用校验和或哈希函数来验证压缩前后的数据是否一致；
- d) 宜建立压缩数据的索引和元数据管理机制，以便快速定位和解压缩所需的数据；
- e) 压缩/解压缩后的数据应按照保存活动的要求妥善存储，确保其可访问性和持久性。

9.4 校验

校验活动是对存储状态的网站资源进行数据解压缩操作的活动，定期校验数据有助于预防潜在的数据丢失或损坏，检查和验证数据的完整性。校验活动应做到以下内容：

- a) 应通过特定算法计算数据的特征值来检测数据在传输过程中是否发生更改；
- b) 应检查存储数据的介质或系统是否稳定可靠，没有故障或损坏的迹象，检查数据的存储环境是否满足其长期保存的要求，如温度、湿度等；
- c) 校验活动结束后应生成校验报告，向相关人员提供数据质量和完整性的反馈，包括发现的问题、采取的措施以及处理结果；
- d) 应根据数据的重要性、业务需求合理安排定期校验数据。

9.5 备份/恢复

备份/恢复活动是对存储状态的网站资源进行数据备份和恢复的活动，应做到以下内容：

- a) 应根据数据的重要性、访问频率和更新速度，制定合适的备份/恢复策略，包括备份范围、备份频率、备份类型、工具软件、备份数量、介质和是否压缩等；
- b) 应定期检查或更换备份介质，确保备份介质的稳定性和可用性；
- c) 在备份完成后，应验证备份数据的完整性和准确性，确保备份数据没有损坏或丢失；
- d) 应定期进行备份恢复测试，验证备份数据的可用性和恢复流程的有效性；
- e) 恢复数据前准备适当的恢复环境，包括硬件、操作系统、网络等，以确保恢复过程能够顺利进行；
- f) 在恢复完成后，应对恢复的数据进行验证和测试，确保其完整性和准确性；
- g) 应记录恢复过程、操作、结果和遇到的问题，以便于后续审计和追溯。

9.6 迁移

考虑到技术升级与兼容性、提升存储性能与可扩展性，应根据需要对存储的数据进行迁移，迁移活动应做到以下内容：

- a) 迁移前，应对源系统和目标系统进行详细的评估和分析，理解数据的规模、结构、复杂性和质量，分析迁移的潜在风险并确定可能的数据冲突和格式不兼容问题；
- b) 应根据评估结果，制定合适的迁移策略及详细的迁移计划，包括迁移的时间表、步骤、顺序、风险评估以及迁移后的验证和测试计划；
- c) 在进行迁移之前，务必备份所有要迁移的数据，确保备份的完整性和可恢复性；
- d) 迁移时，如果源系统和目标系统的数据格式不兼容，应进行数据格式的转换，确保数据能够正确映射到目标系统；
- e) 迁移完成后，对迁移的数据进行验证，确保其与源数据一致。

9.7 销毁

对于不再需要或已经过期的网站资源数据，定期销毁将有助于组织优化资源配置，提高管理和运营效率。销毁活动应做到以下内容：

- a) 应对数据进行分类和评估，明确销毁的具体范围，通常基于数据的敏感性、过期时间、法律法规要求或业务需要销毁真正不再需要的数据；
- b) 应根据数据类型、存储介质和安全性要求，选择适当的销毁方法，常见的销毁方法包括物理销毁（如磁盘破碎、磁性消除等）和逻辑销毁（如数据擦除、数据加密等）；

- c) 对于物理销毁的存储介质，应确保介质被彻底破坏，无法再次使用，对于逻辑销毁的存储介质，虽然数据已被擦除或加密，但仍需妥善管理这些介质，防止数据被意外恢复；
- d) 应对数据销毁过程进行详细的记录和审计，包括销毁的时间、地点、方法、操作人员等信息。

10 应用阶段

10.1 应用

为充分利用长期保存的数据资源、提升数据价值，应在允许的范围内提供网站资源应用服务。提供服务时应做到以下内容：

- a) 应考虑数据安全性，必须确保存储的数据免受未经授权的访问、泄露或破坏，应采用先进的加密技术、访问控制机制以及定期的安全审计来保障数据安全；
- b) 需要充分考虑用户的需求和场景，不同的用户可能对数据有不同的需求和应用场景；
- c) 应按照网站资源所涉及的知识产权相关规定或授权方式来提供网站资源服务；
- d) 应满足合规性要求，避免所提供的网站资源内容存在与国家、行业相关法律法规相悖的情况；
- e) 可提供数据访问和检索服务，便于用户能够根据需要快速获取所需数据；
- f) 可提供不同粒度的网站资源下载服务；
- g) 可提供网站浏览回放服务。

参 考 文 献

- [1] GB/T 18894-2016 电子文件归档与电子档案管理规范
 - [2] GB/T 26162-2021 信息与文献 文件（档案）管理 概念与原则
 - [3] GB/T 35559-2017 技术产权交易服务流程规范
 - [4] CY/T 158-2017 数字出版业务流程与管理规范
 - [5] Digital Preservation Activities and Workflows（数字保存活动和工作流） EMBL Archive, 2021
 - [6] Guidance for Digital Preservation Workflows（数字保存 workflow 指南） Kevin Bolton, Jan Whalen, Larysa Bolton and Rachel Bolton 2023-12-20
-