大数据环境下的档案“数据化”管理研究

金海秋

（海伦市机构编制数据中心 黑龙江 海伦 152300）

摘 要：大数据时代下，各行业与领域加大了对档案管理的重视，并加强了对大数据技术的应用，促进档案管理工作朝着“数据化”的方向发展。特别是机构编制数据中心，在档案“数据化”管理工作中实现了档案管理对象的“数据化”，同时量化档案管理环节，使档案管理的全过程以数据的方式呈现出来，最终达到了“数据管档”的目的。阐述大数据和档案间的密切关系，明确档案数据化管理工作内容，探究未来大数据时代下档案管理工作的发展趋势与数据化管理路径。

关键词：大数据；档案管理；数据化管理；数据存储

【DOI】110.12293/j.issn.1671-2226.2023.15.072

【中图分类号】G251;G271 【文献标识码】A

引言

大数据环境由“数据”“技术”以及“思维”提供支撑，目前已经深入人类社会的各方面，比如导航地图可以实时提供用户位置信息；搜索引擎提供定制化数据；社交软件可分析用户行为，为其提供个性化广告等。大数据理念下，数据为基础，而技术作为支撑，将档案作为基础性数据，实施档案“数据化”管理，有利于提高档案管理效率。

1 档案“数据化”的内涵分析

1.1 存量档案“数据化”，盘活档案库内数据

过去机构内的存量档案主要是结构化数据，通常以纸质档案为主，数据利用和存储耗时且费力。大数据环境下，加强信息化技术的应用，使档案库内的所有数据被盘活，依靠机器读取数据，这是档案“数据化”管理的第一步。将档案与大数据充分结合，以档案数字化为前提，使模拟信号被转为数字信号，同时利用光学字符识别技术进行图像识别分析，使图像内容被机器快速读取，完成数据分析。比如对图书或纸质文件进行“数据化”处理，除了可以满足阅读需求，还能根据用户行为完成深层次数据挖掘与分析。现阶段机构数据中心将纸质档案制作为在线阅读文本。数据中心怎样“用活”档案数据，以及对档案使用过程进行数据记录，挖掘档案信息价值，这是接下来档案管理工作需要关注的重点内容[1]。

1.2 增量档案“数据化”，丰富档案数据库

增量档案“数据化”具体指新档案以“数据化”形式存在，充分满足大数据时代下的数据特征。目前数据中心内的增量档案包含纸质文件、音视频资料、线上网页数据、电子文件等，档案形式与种类繁多，内部存在海量数据信息。增量档案“数据化”的最佳方式是通过接口实现数据在线存档，由此可见，数据在线归档主要包含以下两方面内容：一方面，数据中心对档案管理系统的普及程度；另一方面，系统之间的数据接口。当前已经实现无纸化办公，各办公管理系统与数据平台应用价值提升，纸质文件已经转为电子文件，这对增量档案的“数据化”有着重要意义。

1.3 档案“数据化”管理，为数据管档提供方向

大数据具有预测功能，数据被挖掘后需要进行集中管理，档案管理的关键在于档案自身，同时关注档案管理中形成的数据。虽然数据对档案内容没有影响，但是会影响档案管理与档案利用方式，通过大数据预测潜在的档案管理风险，加强对档案数据收集、整理、统计、利用等环节的风险预测与控制，量化档案管理环节，提高管理质量[2]。

2 实施档案“数据化”管理的意义

2.1 强化档案与数据底层逻辑之间的关联

传统的档案与其他信息特征不同，重视档案内容的“原始记录性”，强调档案信息的记录与保存价值。大数据时代下，自动化设备拥有了数据记录功能，数据可归纳事物逻辑，展示事物未加工的原始素材，将其加工为有利用价值的信息。数据具有原始性，通过对事物的理解，强调数据间的逻辑关系与规则。档案和“离散”的数据不同，主要是指经过管理的规范化数据，数据形式与文档形式的档案同样具有“数字态”，但二者的底层颗粒度不一样。档案“数据化”管理指的是该项工作从粗粒度档案管理过渡到智能化数据管理，实现管理工作的高效化和精准化[3]。

2.2 促进档案来源“数据化”

伴随着人工智能与大数据技术的应用，档案来源呈现出了“数据化”的发展趋势，比如电子证照以“证照模板 + 数据库”为主要模式，将固定的证照模板联合数据管理模式，数据库内无需存放大量图文信息，可高效完成“数据化”信息的检索与统计。“数据化”的便利性让文件管理和档案系统呈现出“数据化”的形式，加强一体化与“数据化”档案管理，使其与档案管理模式相互匹配，促进档案管理的高质量、高水平发

展。

2.3 提高档案服务水平

数据中心对档案管理工作提出了更高的要求，同时档案服务更加多元化，定制化档案服务成为档案“数据化”管理的未来发展趋势。传统档案服务存在无法提供精准服务的问题，利用档案目录进行信息查询，但很难在海量数据内寻找到所需的信息。传统的档案服务只能调出原文，用户自行整合并分析档案内容。档案“数据化”管理则是实施以数据为颗粒度的信息服务，支持档案全文与全数据库的快速检索，可推荐关联信息，生成基于用户需求的定制化内容。

3 基于大数据环境下的档案“数据化”管理路径

3.1 确认档案数据凭证

数据中心的档案主要是经过整理的标准化与规范化数据，符合哪种条件的数据可以被称为档案，这是档案“数据化”管理的关键所在。单纯的数字序列无法让人理解它的具体含义，面对数据集合，需清理数据逻辑关系，以往的文件与表达会有一定制式，可代表数据背后的逻辑关系。《档案法》中明确了针对电子档案的法定要求，需落实档案数据信息的凭证性，提前定义怎样的逻辑描述下的数据是电子档案，不同类型的档案，其背后逻辑关系不同，需要按照具体业务来定义。确认电子档案数据凭证地位时，应同步数据逻辑关系标准，以找到通用语言描述逻辑结构，要求用到的语言必须简洁，且扩展性较强，可对任何类型档案进行定义。

机构数据中心在档案“数据化”管理环节可以采用 XML 的方式，即统一方法描述与交换独立在应用程序外的结构化数据，XML 的通用性与扩展性较强，可定义所有数据结构类型。利用阅读器实现 XML 展示，特别是在一般文本数据的应用场景中，XML 可支持全文检索，根据用户阅读习惯完成展示，将有保存价值的档案内容存在 XML 中，使档案可以以数据态形式存在，避免后续 OCR 识别期间档案数据提取工作量过大。

3.2 保持媒体档案原貌

当前电子档案信息种类丰富，除了文字档案，还有图像、音视频等媒体档案，这类档案的价值在于其自身的媒体性，比如图像承载了书法作品，识别作品中的文本数据，使该书法方便被查找与利用，但需要保留媒体数据原始样貌，防止作品失去价值。依靠人工智能提高 OCR 识别准确率，不管档案提取的数据怎么精确，要求原始媒体必须保留，建议将媒体数据存储于 XML 标签里，长期保存档案内容时，需要关注原始数据的格式，加强对数据格式的转换，使其被转为通用格式。媒体档案在内容识别方面存在困难，随着大数据技术的发展，档案“数据化”管理模式对内容识别精准度不断提升，比如图像识别技术将视频中某一帧内标记的人物信息精准记录下来。档案管理部门或数据中心可按照档案的原貌接收保存，随后再对数据进一步处理与开发，最大程度上提高档案内容的利用价值[4]。

3.3 加强档案数据治理

针对档案数据的“治理”起源于公共管理方面，具体指规范多元主体参与管理，提高管理效率。从“数据化”条件来看，档案管理需要多元主体参与其中，在协同作用下完成对档案数据的高效治理。在档案全生命周期范围内实施档案数据治理，基于相应的行动规则对档案数据进行管理，治理的主体不仅是档案部门与数据中心，还需多元主体参与其中，共同完成档案的“数据化”治理，实现档案协同治理。

3.4 落实档案数据存储

任何数据的管理都要做好数据保存工作，保障档案数据可以被长期保存。大数据时代下，档案爆炸式增长与类型多样化，使档案保存管理方式更加复杂。长期保存档案，其中会涉及到格式与数据组织形式的保存，以及软硬件环境保存，数据中心应寻找一种通用的保存格式，要求该格式依赖关系不强。比如 XML 格式和 TXT 格式，随着档案形式的多样化，特别是媒体档案的应用，DWG 格式、DXE 格式、WRL 格式等被广泛用于档案“数据化”管理工作中。长期保存档案数据时，需要理清各档案数据的逻辑关系。相比之下，XML 格式更符合多数档案长期保存的格式通用性要求，XML 可以在应用层使用，底层数据采用原始格式即可。对于必须定期迁移的数据，或者无法使用通用格式的数据，不仅要将数据迁移管理，还需定期更新档案格式，保持档案内容的可读性。采用定制转换策略，自动化定期更新档案格式，使用阅读器和数据迁移转换工具，利用人工智能进行数据分析，完成批量化档案格式转换，并预判当前是否有数据需要迁移管理[5]。

档案数据在使用期间需要验证器可用性，有必要加强对数据的使用与整理，提高数据治理效果，谨防数据在保存中的风险问题。传统观念下，档案数据更像是一种“冷”数据，大数据时代下，档案数据经过长期的保存与备份，此时数据更加稳定性。以永久的“活”性保存数据为前提，同时采用多种介质进行数据备份，整合各类介质当中的数据存储特性，为数据存储提供保障。“数据态”档案数据拥有更加广泛的来源，长期保存前，利用大数据或人工智能完成数据清洗，剔除无用数据，避免无效档案数据占用数据库存储空间。采用数据监测与对比机制，及时发现任何数据篡改问题，保障数据和存储时一致，这种监测手段以数据摘要技术为基础，可根据数据内容进行人工智

能分析，随着数据的不断更新与完善，数据除了可以保障与进入存储池时一致，还应定期完成在管理库内的摆渡。

3.5 强化档案数据服务

档案“数据化”管理环境下，以数据为颗粒度的档案服务能够解决现阶段档案服务问题。以“收管存”为基础，加强对档案数据的“用”。对此，以下建议可供参考：(1)以数据分析为前提，按照需求提供数据管理服务，根据用户实际情况，分析数据在该场景下能否对用户开放，以及判断用户是否有权限获得数据，保障档案数据在应用过程中的安全性。档案管理工作中，不能因整个档案开放性而影响用户对档案数据的获取。(2)定制化生成信息，并将其准确推送给用户，按照用户提出的需求，智能化整理档案信息，掌握档案内容或多媒体数据，重视对档案的语义理解，将其转为当前语言来回应，将大数据、人工智能技术与 3D、AR 等技术相融合，最大程度上丰富数据的展现形式，促进档案数据互动。(3)跨区域数据共享服务。不同地区会建立专门的数字档案管理或机构数据中心，尤其是在大数据时代下，数据汇集会产生新的数据，数据档案能否按照标准自动生成，有必要基于统一化交换格式，发挥数据之间的共享作用。以数据颗粒度为基础实现数据协同共建，创建共识机制，方便理清档案数据主体权利与义务，保障档案数据共享的安全性，提高各方主体对档案共享的内在动力。(4)促进档案数据服务的便携化，提高数据服务效率，转变传统档案审核方式，让更多有价值的数据被开放使用，并经过文件与档案的在线共享模式，实现档案文件快速归档，便于用户远程查看档案，提高档案服务效能。

3.6 承担档案管理使命

大数据环境下，档案“数据化”管理是一项复杂且系统化工作，需要获得技术、数据以及思维的支持，明确档案“数据化”管理的使命。以此，以下建议可供参考：

(1)以守护数据为基础性工作。迎合大数据环境，档案管理的客体就是将海量数据存储于存储载体内，和以往纸质文件不同，纸质文件主要保障纸张不损坏，那么档案的内容就会安然无恙。大数据模式下，数据的存储方式不可见，如何保持数据的持续可获取是一项难题，档案管理人员应保障电子档案数据的可读性与可用性，做好数据守护管理，实现档案“数据化”管理内容的延伸。

(2)以捕获数据作为档案“数据化”管理的延伸。过去档案管理主要以档案收集、整理。存储为关键，以纸质文件为数据捕获对象。大数据背景下，档案“数据化”管控以“数据管档”为主，通过对各环节数据的精准捕获，利用大数据技术优化档案管理模式。除了对档案数据本身进行管理，还需要延伸到捕获档案的过程，对过程数据加以管理。大数据背景下过程管理数据一般有两方面来源，分别为“传感器感知到的数据”与“应用系统中的数据”，系统内既有数据主要为系统日志。部分数据虽然存在被捕获的条件，但是无法被展开，出现该情况与档案部门管理能力不足有关，没有形成归档管理需求。未来，数据管档将会成为数据中心实施档案“数据化”管理的重要工作部署。

(3)以数据挖掘为重点。接下来档案“数据化”管理主体将会承担起信息专家的角色，工作任务不再局限于简单的档案装订与维护管理，为用户提供的服务也不再是简单的信息查阅，而是要在档案数据群中按照用户的需求，利用大数据技术快速挖掘并检索出有价值的档案数据。一直以来，档案数据挖掘难度较大，受专业能力与数据挖掘工具的限制，大数据背景下档案部门所提供的数据服务将会更加先进，经过用户行为识别与分析，为其提供更精准且快捷化的大数据服务。档案部门以智能化管理为基础，此时挖掘数据已经成为了一项常规性工作，基于知识与智力支持实现数据的高效共享。

总结：总而言之，大数据时代背景下，档案的“数据化”管理类似于创建数据网，随后经过大数据挖掘与分析，加强对档案从收集到使用的全过程管理。强化档案与数据底层逻辑关系，促进档案来源数据，进一步确认档案数据凭证，加强档案数据治理与数据存储环节的管理，全方位提升档案数据服务水平，保障档案“数据化”管理质量。

参考文献：

[1]苏永芬,吕晨曦.大数据背景下档案数据化管理助力企事业单位发展的策略探讨[J].四川档案,2022,No.227(03):34-35.

[2]黄若非.智能城市档案管理大数据化[J].中华建设,2022,No.287(06):32-33.

[3]徐钦梅,戴敏.档案数据化管理的实现路径研究[J].浙江档案,2021,No.488(12):32-35.

[4]于英香,滕玉洁.大数据背景下档案管理数据化转型探析[J].中国档案,2021,No.567(01):81-83.

[5]邱川燕.大数据时代档案数据化管理与建设[J].福建电脑,2020,36(10):66-68.