非结构化数据管理知识与实践

（2023版）

2023年4月

编写组成员

张　群	尹　卓	曹幼林	龙凌云
罗永秀	姚宝敬	闫　述	张　凯
黄永庄	任　歌	陈亚军	彭革非
王　雷	吕艳静	张　程	刘　丹
周兆锋	方　俊	张　治	陆　猛
刘赛赛	徐志东	杨吉云	梁　勇
王长胜			

参编单位

上海鸿翼软件技术股份有限公司

中国电子技术标准化研究院

北京中船信息科技有限公司

华迪计算机集团有限公司

北京数科网维技术有限责任公司

福昕鲲鹏（北京）信息科技有限公司

北京点聚信息技术有限公司

友虹（北京）科技有限公司

永中软件股份有限公司

版权声明

目 录

1. 前言

数据，是当今时代企业生产生存的命脉。企业的持续经营必将产生大量数据，而海量的数据也无时不刻地在影响着企业的经营。无论是在企业的战略层面还是执行层面，数据管理对于企业决策都具有举足轻重的作用。在战略层面，基于数据管理能够有效梳理企业数据资源，支撑企业优化战略决策，提前洞悉业务中存在的潜在问题，把握市场，拓展机遇，抢占竞争先机；而在执行层面，通过数据管理能够帮助企业解决现有业务中的数据责权不清、数据标准不明、管理流程混乱、数据质量低下等常态问题，形成标准化的数据利用流程，提升运营效率，培养企业的核心竞争力。

2018 年，全国信息技术标准化技术委员会大数据标准工作组组织制定的 GB/T 36073-2018《数据管理能力成熟度评估模型》（以下简称 DCMM）国家标准正式发布。在推动 DCMM 国家标准落地应用过程中，当前部分企业已经逐渐形成对数据的管理意识，并陆续开展数据管理相关工作。因此，发布 DCMM 是顺势而为，旨在指导国内企业的数据管理建设与数据文化培养，为企业数字化基础设施的形成与完善提供方向与建议。

根据调查显示，企业数据管理工作目前侧重于结构化数据的管理，已经形成了多种针对企业业务中产生的结构化数据进行管理的专业软件，能够以体系化、动态化、甚至智能化的手段，对企业内的结构化数据进行高成熟度的管理。然而，相比之下，企业针对文档、图片、音视频等非结构化数据的管理方面仍投入不足。这些文件充斥在企业的存储系统与员工日常办公中，大部分企业却依旧处于非结构化数据的局部建设或者初步建设阶段。一方面，大部分企业尚未认识到非结构化数据管理的重要性；另一方面，缺乏成熟的

非结构化数据管理体系和工具的支撑，也缺乏针对非结构化数据实践的专门标准。

为此，上海鸿翼软件技术股份有限公司、中国电子技术标准化研究院联合北京中船信息科技有限公司、华迪计算机集团有限公司，以及北京数科网维技术有限责任公司、福昕鲲鹏（北京）信息科技有限公司、北京点聚信息技术有限公司、友虹（北京）科技有限公司、永中软件股份有限公司等电子文件管理推进联盟会员单位，共同开展对非结构化数据管理相关的技术、应用以及标准化的研究探索工作。

作为 DCMM 在非结构化数据领域的补充与细化，本白皮书立足于非结构化数据管理应用实践，结合 DCMM 国家标准体系框架，提出了**非结构化数据管理能力分级评价模型**，并形成以内容管理成熟度模型 CM³ 为核心的非结构化数据管理解决方案，是鸿翼及电子标准院前期累积的重要研究成果。本白皮书的发布，一方面是为了呼吁各界加强对非结构化数据管理技术、应用及标准化工作的关注，增强社会面的非结构化数据管理意识；另一方面旨在通过分享前期研究成果，支撑各行业及企业开展非结构化数据管理体系建设，实现产业数据管理能力的全面提升。而 2023 年的新版本，则是基于行业近几年的研究重点，聚焦非结构化数据相关的新举措、新实践、新里程碑，对本白皮书进行了更新、勘误、充实。希望本白皮书能够与时俱进，帮助企业精准定位自身非结构化数据管理水平，以正确的手段实现企业数字化转型的目标。

本白皮书由上海鸿翼软件技术股份有限公司和中国电子技术标准化研究院共同组织编写并更新。

2. 非结构化数据管理

2.1. 非结构化数据定义及特征

非结构化数据是指未通过数据模型预先定义的数据，包括关系数据和模型数据。在企业的整体数据架构中，非结构化数据往往是指不适合用数据库二维关系逻辑表来表现的数据，包括所有格式的办公文档、标准通用标记语言下的子集、各类报表、图像和音频视频文件以及工程图文档信息等，约占企业数据存储量的 80%。

存储在计算机系统中的数据被分为结构化数据和非结构化数据。结构化数据与非结构化数据在数据对象、数据格式、时间维度、存储形式、增长速度、信息含量、数据价值等方面存在明显差异，具体如表 1 所示：

表 1 结构化数据与非结构化数据特征差异

	结构化数据	非结构化数据
数据对象	结构化数据以关系型或单一数据属性，如：银行卡号、日期、财务金额、电话号码、地址、产品名称等作为数据对象	非结构化数据以内容或本体，如文件、图像图形、音视频、邮件、报表、网页、各种纸本等作为数据对象
数据格式	强调基于表格的关系型数据值格式类型，如：字符型、整型、日期型、数值型等	由于非结构化数据较多体现在无模式、自描述的文件及内容，其数据格式更为多样，如：png、jpg、mp4、doc、ofd、pdf 等各种类型
时间维度	结构化数据的以单一数据属性为主，需要构建关联，呈现分析结果，应用时效性较短	非结构化数据以文件和内容为主，信息量较大，应用时效性会更长
存储占比	在企业日常运营产生的数据中，结构化数据占存储数据总量的 20%	在企业日常运营产生的数据中，非结构化数据占存储数据总量的 80%
存储形式	结构化数据通常仅存储在软件应用系统和数据仓库中	非结构化数据的存储端多样，可以储存在个人电脑、服务器、应用系统、文件柜或档案室等终端以及数据湖为代表的大数据平台中

增长速度	通常结构化数据占业务数据增长量的 20%	通常非结构化数据占业务数据增长量的 80%
信息含量	结构化数据需要结合上下文语义呈现信息，信息量较小，着重体现在定量数据和关键的业务信息	非结构化数据所包含的信息量较大，可以扩展至情感性、描述性、文档性等更为广泛的信息
数据价值	结构化数据的价值主要体现在假设、明确或已知的数据分析价值	非结构化数据价值拥有更广泛的、探索性、数据挖掘等未知的数据洞察价值

综上所述，非结构化数据与结构化数据是两种差异巨大的数据类型，随着大数据存储和计算能力的增强，非结构化数据由于其丰富的信息量，相较结构化数据拥有更大的数据资产化价值空间。组织应注重非结构化数据在数据管理中的有效管理，着重针对非结构化数据的无序性、分散性开展价值挖掘，对缺乏规则化的非结构化数据，尤其是对分散在个人电脑、服务器、各种应用程序及大数据存储中的非结构化数据开展全面的治理，进一步发挥非结构化数据的资产化价值。

2.2. 非结构化数据管理的发展历程

数据管理的起始可以追溯到 20 世纪 60 年代的数据库技术，当时计算机已经开始在商业环境下获得应用，文件是数据存储的主要介质。文件的存储和访问成为数据管理的核心需求，这也可以看作非结构化数据管理的最初阶段。

20 世纪 90 年代初期，随着无纸化办公技术的发展，传统纸质文档逐步转换为电子化文档，这个时期企业开始构建电子文档库、数字图书馆、数字档案馆，非结构化数据管理体现为对这些数字化文

档的管理。

2000 年以后，随着互联网技术的发展，非结构化数据率先体现在以 WEB 网页为主的内容管理上，随着网站技术的发展，出现了网页内容管理（Web Content Management），这个时期电子商务、电子政务系统也随之快速发展。

2005 年以后，随着企业信息化的不断深入，非结构化数据融入到业务场景中，企业业务流程系统承载了大量文档、图表、报告、音频等形式的非结构化数据。对这类数据的管理需求促进了 ECM 企业内容管理（Enterprise Content Management）的出现，随着 ECM 的出现，非结构化数据开始与业务场景深度融合，发挥出了更大的价值。

2010 年以后，随着云计算，物联网、移动互联网和大数据的不断发展，非结构化数据呈现形式更为多样，如：影像文件、视频文件、工程电子文档、ISO 质量电子文档等，这个阶段 ECM 企业内容管理和非结构化数据应用的发展也越来越趋于规模化。

2015 年以后，随着人工智能技术的成熟与普遍化，非结构化数据开始向着内容服务自动化、文本挖掘、语义分析等方向发展，并形成了非结构化数据管理体系下的内容服务中台化和内容服务智能化。

从上述非结构化数据发展历程可以收获以下几点：一、非结构化数据是随着计算机应用的发展不断丰富起来的，因此任何时代，技术发展都是动力。二、非结构化数据管理的发展历程是非结构化数据逐步从离散文件升级至内容，形成统一的内容服务平台，并进一步构建起融合业务的知识体系，其本质上大大提高了生产运营效率和业务创新能力；三、多层次的非结构化数据平台提供了更为上

层的内容服务，屏蔽了下层的技术实现细节，能够更快速准确地响应业务场景化需求。

因此，随着数字数据管理的成熟，一股专注于非结构化数据管理的浪潮也在悄然崛起，以非结构化数据为研究与发展的重心，掀开了非结构化数据管理的篇章。

国际上，1990 年，Documentum 公司成立，成为了第一家利用标准关系型数据库技术以及面向对象方法提供企业级文档管理解决方案的公司；

2000 年左右，以电子商务和电子政务为代表的门户网站的发展带来了网页内容的指数级增长，促进了网页内容管理的成熟与发展；

2002 年，Documentum 公司正式发布 ECM（企业内容管理）产品；

2006 年，微软发布 Sharepoint Portal Server；

2010 年，OpenText 发布；

2010 年后，云计算、移动互联网、大数据的新技术改变了 ECM 的形式与内容，ECM 的内涵与外延不断更新。

放眼国内，从 2002 年起，航空、核电和工程领域的国际 ECM 一线厂商开始进入我国，在这些行业内，率先掀起了 ECM 的潮流：

2008 年，上海鸿翼软件技术股份有限公司发布国内首款完整 ECM 产品“鸿翼 edoc2 ECM”；

2009 年，拓尔思信息技术股份有限公司针对政府和金融领域推出 WCM 产品；

2010 年，信达雅系统工程股份有限公司在金融领域推出 ECM 影像管理产品；

2016 年之后，以联想企业网盘、石墨文档等为代表的应用层的网盘和功能更全面的 ECM 出现，ECM 系统中的文档协同和服务能力不断提升；

2017 年开始，人工智能（AI）逐渐开始与 ECM 系统进行融合，企业开始利用人工智能手段，赋能非结构化数据管理；

2020 年开始，中国 ECM 行业产品平台化趋势显现，以鸿翼为代表的 ECM 平台开始成熟，基于平台的应用开始在各行业爆发式增长，是为“中国 ECM 元年”。

2.3. OFD—归档用电子文件的标准格式

图文类文档是非结构化数据的常见类型之一，因为贴近决策阅读，这类文档中蕴含着巨大的有用信息。按照是否可以编辑，可以把图文类文档分为流式文档和版式文档。

流式文件支持在任意位置自由编辑，编辑后会按照流式灌排的方式进行版面重新计算与绘制，由于排版计算受操作系统、软件实现版本等影响较大，流式文档可能会出现不同的软件和操作系统平台上内容效果不一致的现象，又称“跑版”。

流式文件一般包含章节、表格、段落、句及图文对象等元素，上述各个层级的对象都有其独特属性。这些内容会按照一定的层次结构进行的描述方式构成流式文件的格式。依托合适的流式文档软件（如 WPS、Office），文档拥有者可以对文件的内容进行编辑、添加、删除等操作，连接文档服务进行辅助校对和创作，并且在此过程中可与其他编辑者协作，是常用的文件类型。

而为了保证文档在各种软硬件环境下的显示、打印等效果高度精准一致，版式文件应运而生。版式文件是版面呈现效果固定的电

子文档，文档内容的分页、换行和图元位置都在文档中直接纪录，在各种设备上阅读、打印或印刷时，可直接读取和使用位置信息，不依赖排版计算确定，因此文档的呈现效果高度稳定。版式文档主要应用于成文后文件的发布、传播和存档，如商务文档、电子公文、电子凭证等。此前，PDF（Portable Document Format）是版式文件的代表实现，经历近了 40 年的发展，在全球范围多个行业内大量应用，已成为了国际标准（ISO 32000-1:2008）。

国内对文档应用有许多独特的应用需求，例如应用国产密码、分段标密或保护、结构语义保留等，由此诞生了许多基于自定义格式或 PDF 的定制应用方案，满足局部需求的同时，也使得国内的版式文档管理更加复杂，为了在应用上兼容各方需求，在技术上统一文档格式，在管理上合理归并冗余，在产业上凝结行业共识，迫切需要出台版式文档格式方面的国家标准。2016 年 10 月 13 日，国家标准《电子文件存储与交换格式—版式文档》（GB/T 33190-2016）正式发布，OFD（Open Fixed-layout Document）由此诞生。

与 Adobe 公司的 PDF 相比，OFD 是我国自主研发的文档格式国家标准，除了可以媲美国际标准的文档静态和动态特性描述能力，在安全性和易用性等方面进行了独特的技术创新。OFD 摈弃了老旧的二进制描述方式，采用 XML 描述文档内容和“ZIP+”方式聚合文档数据，真实地保持文档中原有的文字、图标、公式等版式信息，描述更简洁、信息集成度更高，形成了显著的比较优势。OFD 采用了文档原始内容与附加内容分离保存的策略，有利于相关内容的区分签名及保护，在应用中作为责任区分凭证中发挥独特作用。在文档安全层面，OFD 设计了标准接口内置支持 SM2/SM3 等国产密码，对国外算法也具有很好的适配性，进而全面支持 GB/T

38540、GB/T 35275 等国密算法的签章和签名标准。OFD 设计了元数据、附件、自定义标引等丰富的扩展机制，鼓励在版式文档中携带业务源头的结构化数据，实现了多源异构数据融合，在支持发票、证照和公文深入利用中发挥了独特作用。最后，OFD 未引入动态表单和脚本，更加聚焦于版式文档的优势和职责，切断了病毒与木马的通过文档脚本污染数据、感染系统的路径，消除了最大的文档安全性隐患。

OFD 作为一种后发技术和文档格式，针对 PDF 格式“是‘文档的坟墓’”（意指信息进得去难出来）和对信息安全关注不足等两大痛点做了专门的重新设计和改进，是统一图文类文档格式，消除行业和系统壁垒，提升非结构化数据管理效率的“利器”。OFD 发布以来，在机关办公、政务服务、财税管理等重要领域应用，以公文、证照、发票、回单等不同业务形态，在优化业务应用、凝聚产业力量、防止技术垄断和保障数据安全方面发挥了重大作用。

OFD 作为版式文档领域的新生力量，发展空间巨大，近年来先后发展了党政机关电子公文、电子证照、可入账电子凭证等应用标准，但是网购、保险、金融、企业管理等领域中仍有巨量的电子文件亟待规范化，生产制造、建筑、水利、交通、测绘等领域的专业电子文件则更加具有挑战性。在产业方面，专业技术厂商和开源社区同步发力，除了专用软件外，微信等通用平台对自主格式支持也在加大，应用方获得相关技术支持的门槛降低、服务质量却在不断提升。依托于自主可控的文档格式标准，充分发挥其对于内容管理的友好特性和数据安全特定，在关键基础技术自主、供应链韧性和网络安全得到空前重视的大背景和大环境下，在业务系统应用自主文档格式，通过其实现更懂业务、更高效率和更安全的非机构化数

据管理，是大势所趋，更是时代的“必答题”。

3. 非结构化数据管理体系

组织构建非结构化数据管理体系，需要基于顶层设计及战略开展非结构化数据治理，落实非结构化数据管理的各项职能活动，最大程度开发非结构化数据的资产价值。非结构化数据管理体系（如图 1 所示）由五大核心方面及十六个重点领域组成。其中核心领域包括：

（1）非结构化数据顶层设计及战略；

（2）非结构化数据管理能力成熟度；

（3）非结构化数据治理，包括组织与职责、制度与流程、评估与审计和数据文化；

（4）非结构化数据价值，包括非结构化数据协作、非结构化数据流转、非结构化数据服务和非结构化数据洞察；

（5）非结构化数据管理，包括非结构化数据集成、非结构化数据标准、非结构化元数据管理、非结构化数据质量、非结构化数据安全和非结构化数据合规。

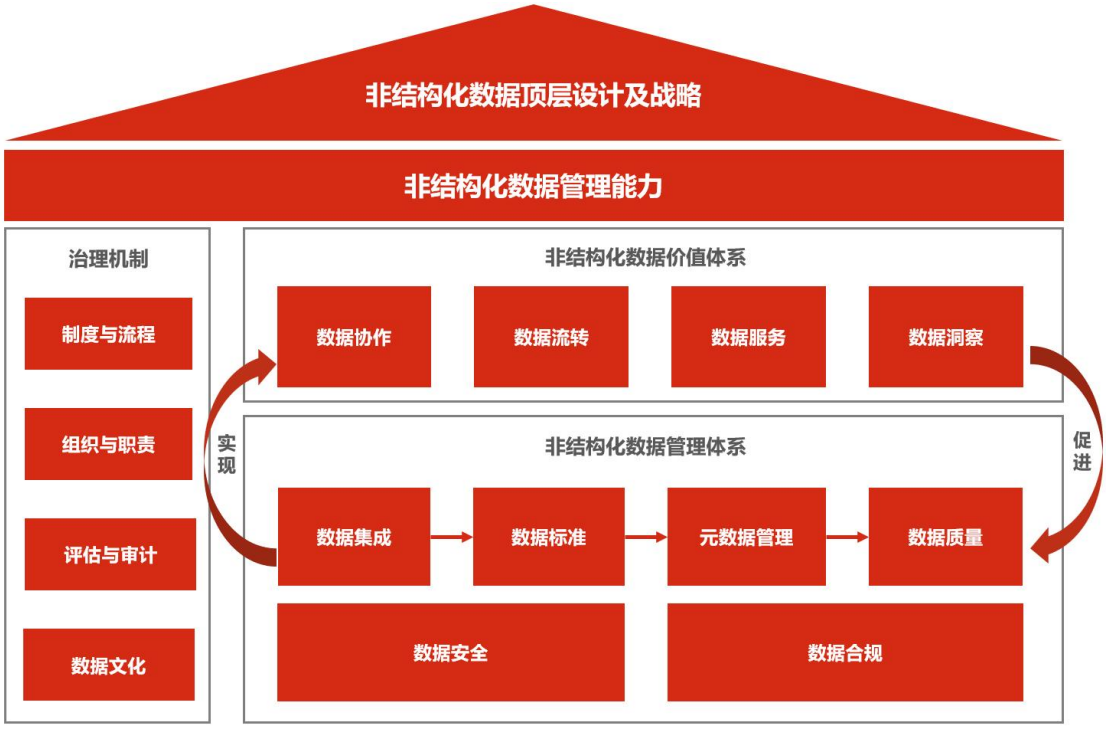


图 1 非结构化数据管理体系框架

其中非结构化数据价值体系与非结构化数据管理体系之间能够起到相互促进的作用，完善的非结构化数据价值体系能够推动企业非结构化数据管理体系的逐层建立与制度完善，而企业的非结构化数据管理体系的规划从某种程度上来说，也是对其非结构化数据价值体系的体现和落实。

3.1. 数据管理能力成熟度模型

能力成熟度模型（Capability Maturity Model）最初源自软件开发管理程序，由美国卡耐基梅隆大学软件工程研究所于 20 世纪 80 年代提出。这一模型将软件开发划分为五个成熟度级别，除了初始级（第一级）以外，每个级别都由关键的过程域组成。关键过程域能够标识组织应该关注的领域，以改进软件开发过程。每个关键过程域分为五个部分，称为共同特征。共同特征指定了关键实践，当这些关键实践被实行，就可以实现关键过程域的目标。

能力成熟度模型认为软件开发并不是一蹴而就的过程，需要组织首先明确工作开展的方向以及工作的优先级顺序。因此，每个成熟度级别都对应着持续改进过程中组织达到的全新阶段。根据能力成熟度模型架构，组织可以标准化、模块化地判断软件当前的成熟度，并将其与行业内其他组织的实践状态进行横向对比。同时，组织也可以使用能力成熟度模型来制订软件开发的改进规划。

因此，借鉴国内外成熟度相关理论思想，数据管理成熟度模型（DCMM）涵盖了数据战略、数据治理、数据架构、数据标准、数据生存周期、数据应用、数据质量、数据安全共 8 个方面（能力域）（如图 2 所示），并根据数据管理过程的有效性、完整性、协调性等因素，划分了各能力域及整体数据管理能力的 5 个成熟度等级，给出了不同成熟度等级的指标要求，旨在为企事业单位评估和持续改进自身数据管理能力提供科学指引。

图 2 DCMM 数据管理能力成熟度模型

DCMM 着眼于数据管理领域，旨在规范和引导组织的数据管理过程，遵循能力成熟度模型的基本理念和结构，再进一步细分关键

过程域，以区分不同成熟度水平。此外，该模型定义了一系列指标，从而指导组织开展数据管理现状评估，组织可以在数据管理实践方面结合自身关注的领域，选取相关数据管理领域开展评估，不同的数据管理域内所包含的关键要素也不尽相同。

3.2. 非结构化数据应用分级要求

3.2.1 DCMM 在非结构化数据领域的细化与补充

基于非结构化数据自身的特征与相关软件产业的发展程度，不难发现与结构化数据相比，非结构化数据的管理与应用的发展整体较为滞后。这不仅因为其每年超数据总量 80%的增长速度，更因为其蕴含着极其丰富的信息和知识，以结构化数据的常规与评估管理方式无法精准地对企业的非结构化数据应用管理能力进行评判。

因此，非结构化数据应用能力分级模型基于 DCMM 数据管理能力成熟度模型的五个成熟度等级构建，参照 DCMM 给出的数据管理能力成熟度模型与等级，并且考虑到非结构化数据特征、组织在非结构化数据管理领域的具体实践，构建起了一套完整的非结构化数据应用能力的评判标准，而对应的非结构化数据管理能力成熟度级别体现为：“初始级”的文件零散化；“连接级”的内容协作；“可度量级”的内容统一管理；“融合级”的内容服务与“智能级”的内容智能等特征（如图 3 所示）。

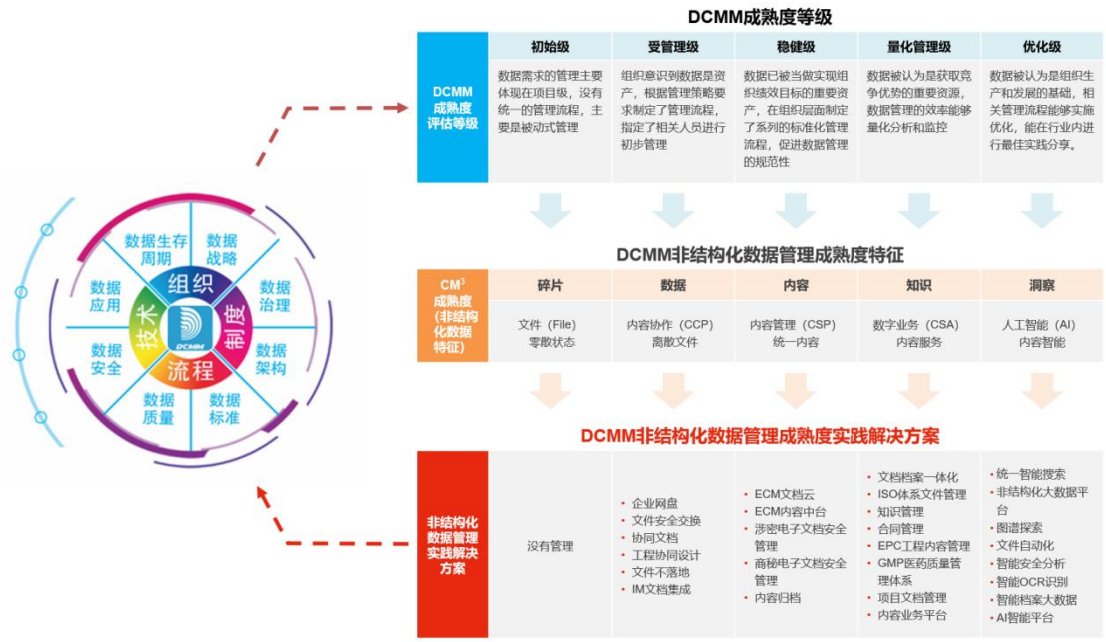


图 3 非结构化数据管理成熟度解决方案

3.2.2 非结构化数据应用分级要求

《非结构化数据应用分级要求》中规定了非结构化数据应用的能力模型与分级要求，主要包括管理制度、管理技术、业务支持、决策支持和安全合规 5 个能力域，适用于企业与组织的非结构化数据应用能力的评估。

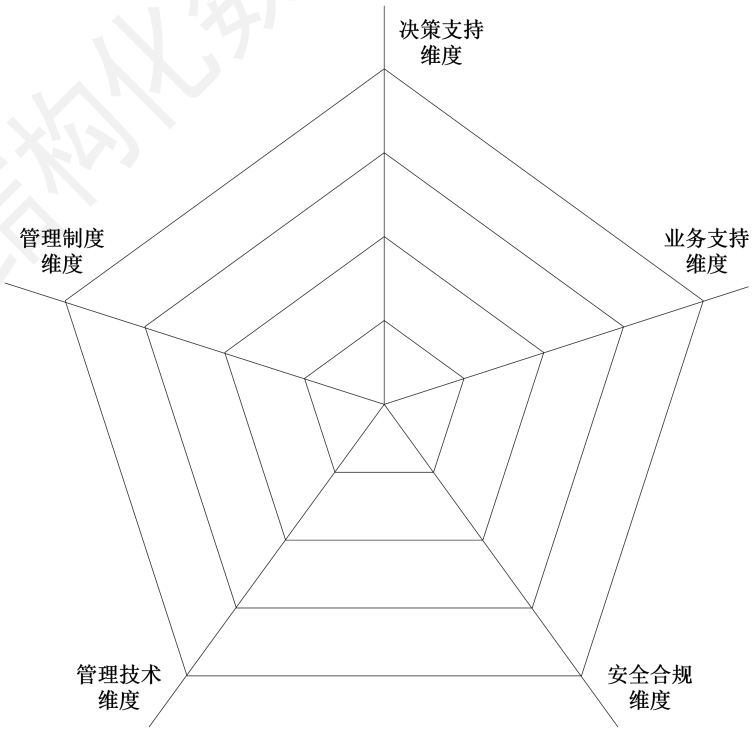


图 4 非结构化数据应用能力模型

根据该模型，能够对企业非结构化数据的管理与应用能力进行全面的分析与评估，通过判断组织的每个维度处于哪个阶段，将组织的非结构化数据应用能力分为五个等级：

a）第一级——**初始级**，组织机构基本上不存在有组织的非结构化数据管理；

b）第二级——**连接级**，非结构化数据管理在组织机构内初步发挥作用；

c）第三级——**可度量级**，非结构化数据管理与组织机构的业务系统深度结合；

d）第四级——**融合级**，利用各类数据来辅助工作成为组织机构文化的一部分，并且这种文化沿着供应链外延到上下游合作伙伴；

e）第五级——**智能级**，组织机构内业务全面实现数字化转型。

每个等级都会分别对企业、组织的五大能力域进行评判，并给出相应的描述与特征，方便企业管理者对自身现阶段非结构化数据应用管理能力有一个清晰的认知，从而调整自身的非结构化数据战略，对于薄弱环节进行加固与重视，帮助企业提升其非结构化数据管理能力，平稳地过渡到非结构化数据管理建设的更高层级。

非结构化数据应用分级包括的五个阶段具体描述如下：

第一级：初始级 非结构化数据管理和相关系统在组织机构内不存在或者未充分发挥作用，应符合如下特征：

a）组织机构内非结构化数据管理的机构、制度和资源配置尚不完善；

b）组织机构内各项业务离散程度较高，主要依赖传统的资料传阅等方式实现业务协同；

c）非结构化数据管理工具和系统装备简单，缺乏成体系的数据管理活动；

d）非结构化数据管理对组织机构内业务生产、经营和决策的支撑能力较弱；

e）主要依赖物理隔离实现访问控制和安全存储。

第二级：连接级 非结构化数据管理和相关系统在本机构的业务经营和决策过程中发挥了作用，应符合如下特征：

a）组织机构内建立了非结构化数据管理制度，并对全员进行了适当培训；

b）各类文件在所属业务、主题和操作环境等维度建立了关联；

c）装备了非结构化数据管理系统，内外数据和文件可在必要时以全内容形式对本机构内提供服务；

d）通过非结构化数据管理系统的权限配置和集中管控来控制安全风险。

第三级：可度量级 非结构化数据管理与业务系统深度结合，在本机构生产经营决策过程中发挥了基础作用，应符合如下特征：

a）数据和文件作为重要资产纳入本机构发展战略，在治理结构中占据重要地位；

b）大多数业务活动通过信息系统开展，数据和电子文件业务伴生、网状联络的特征明显；

c）建立了机构内部的非结构化数据管理系统，部分内外数据在

系统内统一管理，管理的颗粒度精细到文件以下；

d）非结构化数据管理系统中数据可较为全面地反映本机构运营情况，数据和文件在经营决策中发挥重大作用；

e）采用边界检测、内容安全等一系列技术手段强化非结构化数据管理系统安全。

第四级：融合级 组织机构内大部分业务活动实现了数字化，用数据决策、用数据监督成为组织机构文化，应符合如下特征：

a）非结构化数据管理成为组织机构发展的战略性支撑，数据治理成为本机构日常工作的重要部分；

b）本组织机构产出的计算机文件（如设计图纸、产品说明、报告和各类文书）和业务凭证（如合同和财务凭证等）大量实现了结构化，采用国家标准格式并实现了“视读机读双支持”，其文件类型在地区、行业或国家对应注册中心注册，可高速检索匹配文件内容，可直接接收和使用外部生成的同类数据；

c）非结构化数据管理系统广泛、深入接入各业务系统，准确反映本机构运营情况和外部环境，为组织机构领导层提供决策支持；

d）非结构化数据管理和业务系统均由专业人员通过运维系统进行运维保障。运维系统能实时反馈目标系统的运行情况，预警可能出现的异常和故障，发生异常或故障时记录诊断数据，及时按照应急方案恢复系统运行；

e）非结构化数据管理和业务系统达到计算机信息系统安全保护三级以上。

第五级：智能级 组织机构内业务全面数字化转型，数据成为本机构的核心生产要素，应符合如下特征：

a）数据和文件成为组织机构的核心资产，内部定期开展数据资产评估并将其资产增值作为其重要发展目标；

b）通过强化非结构化数据管理促进业务的开展。组织机构内非创造性工作均具备自动处理功能，业务系统对上下游和利益相关方的带动和促进效应明显；

c）在非结构化数据管理中引入自然语言处理、知识图谱和深度学习等新技术，对数据的开发利用产生可度量的效益；

d）在非结构化数据管理基础上全面实现智能管理和智能决策，可基于已发生的活动和内外经验，预测未来发展趋势，各业务在组织机构内外得到显著优化。

每一级的能力要求均是在上一级别基础之上递进增加，通过使用配套的《非结构化数据应用能力分级测试》工具，帮助企业在每个能力域的细项上进行自我评估与自我定位，依托模型逻辑，得到对应的企业非结构化数据应用能力等级，查漏补缺，从而调整自身非结构化数据管理战略，帮助组织更好地对非结构化数据进行系统、科学地管理与应用，达到降本提效的目的。

图 5 非结构化数据应用能力分级测试页面

3.3. 非结构化数据战略与顶层设计

非结构化数据顶层设计即企业针对非结构化数据管理与应用的战略规划，需要基于组织的业务战略和 IT 战略开展建设，应确保与业务目标和 IT 目标相一致，同步制定顶层设计及战略规划的实施策略工作。良好的非结构化数据顶层设计会为组织的安全合规、运营效率、客户满意度等多方面提供支持。顶层设计的驱动力通常来自法律遵从性要求、诉讼响应能力、电子取证请求能力和业务连续性要求。

这里将从非结构化数据管理战略制定的关键步骤，即：现状评估、业务效率、洞察创新、安全合规和数据文化等方面进行介绍（如图 6 所示）。

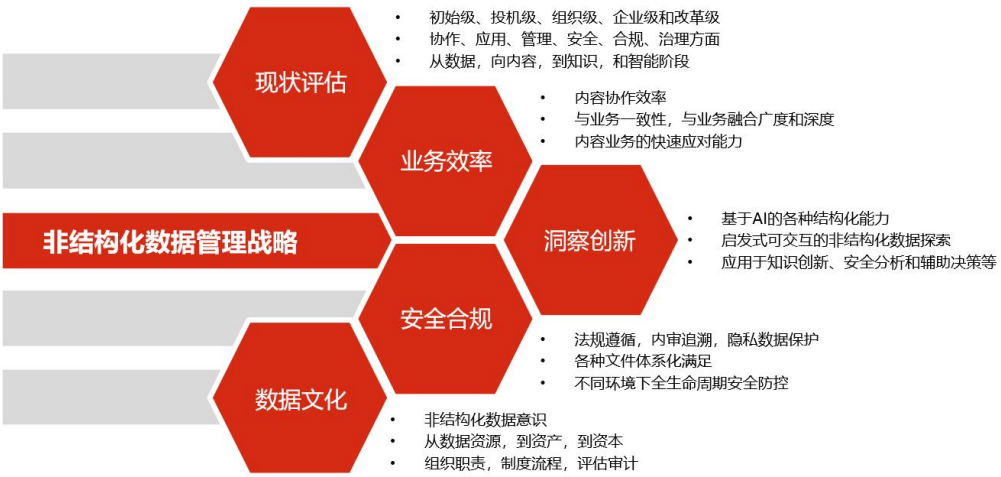


图 6 非结构化数据管理战略图

（1）现状评估

现状评估的重点是对组织内非结构化数据现状开展全面评估工作。通过调研，获得非结构化数据存管现状、技术现状和应用现状的具体情况，并通过能力成熟度模型进行评估，分析出组织非结构化数据管理中存在的问题及所处的级别，识别出根本原因，明确下一步工作重点，为后续非结构化数据管理规划的升级与改革指明方向。

（2）数据文化

数据文化则需要培养组织全体成员由上至下、从纲领到实践的非结构化数据管理意识，理解数据从资源到资产的价值化过程，提升非结构化数据管理能力，落实非结构化数据举措，实现数据资产化的目标。

（3）业务效率

业务运营效率的提升是组织进行非结构化数据管理关注的重点，也是实行非结构化数据管理的首要目标。为了实现业务运营效率的显著提升，组织需要在内容协作效率、内容与业务深度、广度融合与内容快速响应业务变化等方面进行深度建设与更具有针对性

地提升。

（4）安全合规

针对安全合规的需求，则要组织考虑对法律法规、内审追溯、隐私数据保护等规定的遵从性，从构建合规的内容管理体系出发，对数据实施全生命周期的安全管理。

（5）洞察创新

洞察创新作为战略中层级最高的一项，需要组织基于人工智能、知识图谱等前沿技术，进行启发式、交互式的非结构化数据挖掘、数据探索和信息推送，并且关注安全分析和智能决策的场景化应用，以及实现自动化应用与知识创新。

3.4. 非结构化数据治理

非结构化数据治理是开展非结构化数据管理工作的关键，以统筹和协调非结构化数据管理各项工作有序开展为主要内容，其核心要素体现在组织、制度、文化和考核四个方面。

（1）组织

企业需成立非结构化数据的专门管理部门，下设相关的职能小组，如：文件管理组、档案管理组和中台运营组等。相关职责方面，文件管理组主要负责制度文件管理、公文管理和文件管理等；档案管理组主要负责文书档案管理、会计档案管理和科技档案管理等；中台运营组主要负责内容融合管理和内容服务管理等。成立类似的部门与组织有助于企业对非结构化数据的体系化、规范化管理。

（2）制度

组织应制定非结构化数据管理办法、规范和细则等相关制度体

系，如：电子文档管理办法、非结构化数据管理标准等。通过组织对非结构化数据的规范性管理，提高非结构化数据治理水平。

（3）文化

为了更好地推进组织非结构化数据治理，组织应逐步树立起非结构化数据管理的文化，提升全员的数据价值观和数据管理文化素养，培养全员的非结构化数据资产化意识。

（4）考核

组织应通过开展非结构化数据治理的评价与考核，贯彻落实非结构化数据管理战略及目标，跟踪执行过程中的实施情况，及时发现组织非结构化数据管理中的问题，提出优化和改进建议。

3.5. 非结构化数据管理

非结构化数据管理作为组织的重要职能，是落实非结构化数据顶层设计及治理的关键。主要包括非结构化数据标准、元数据、数据质量、数据安全、数据合规和数据集成等六个方面内容。

3.5.1. 非结构化数据标准

非结构化数据标准是组织对非结构化数据提出的规范性要求。组织应以非结构化数据标准体系构建为基础，指导和规范各类非结构化数据管理工作。非结构化标准体系构建包括了对内容模型、内容分类、编码命名、内容格式、内容本体、版本策略、元数据、内容指标和内容接口等规范的标准。

（1）内容模型标准

内容模型标准主要包括子域规划、活动模型规划、文件版本规划、结构化规划、元数据建设、体系合规建设、安全策略等各个方

面。

子域规划从业务过程出发，构建出多层级式的子域结构；活动模型规划从文件发送、共享和外发等流转活动出发，关注文件的动态业务活动；文件版本规划关注文件的主次版本、生效版本、修订版本、版本留存数量等版本规范；结构化规划关注图像类、专业类文件如何基于 OCR 识别、兼容解析等技术实现内容结构化；元数据建设从行业元数据和内容元数据两方面进行规划和建设；体系合规建设是从文件的新增、修订、权限申请以及敏感内容等方面进行规范；安全策略是从授权访问限制、共享外发安全和水印安全等方面进行规范建设。

内容模型作为规范和标准，在非结构化数据管理的过程中尤为重要，内容模型是构建内容库的基础。良好的内容模型需要对各类活动模型、版本模型、权限模型、元数据模型、流程模型和安全模型等各个环节进行全面和规范化的构建。

内容模型与内容库关系紧密，内容库的每个层级都对应着不同业务体系化的数据。不同业务要求不同的数据规范，这些规范的建设都是需要通过内容库的内容模型来实现的。

（2）内容分类标准

内容分类标准是指目录树分类、标签分类、智能分类和分类编号等方面的规则和规范。

目录树分类侧重于体系化内容的分类建设，根据组织维度、业务维度、战略维度等进行内容划分。如果说目录树分类是内容的“垂直分类”，那么标签分类则侧重于内容的“横向分类”。

标签分类是在“垂直分类”的基础上，支持跨业务目录的一种分类方式；同时标签分类是在用户对内容理解的基础上，对内容进

行标签化标注的一种以内容为维度的分类方式。

智能分类是标签分类的延伸，基于人工智能自然语言处理（NLP）技术实现对内容的智能标签分类，辅助人工标签化的过程。

分类编号是通过格式化的编码自动生成分类号进行内容分类，进而延伸出的业务逻辑分类。

内容分类的建设过程主要包括内容分类梳理、分类规范建设、分类执行、分类结果分析评审等关键环节。其中，内容分类梳理需要组织明确内容分类规范；分类规范建设主要遵循漏斗结构原理进行梳理，包括现状调研、现状评估、分析梳理、知识规划、展示设计、用户评价等阶段过程。

内容分类规范性还体现在文档管理的分类分级，需要满足各个层级的用户内容需求，内容分类需要具有清晰的层级结构，从而为后续的多维文档提供良好的支撑。

（3）编码命名标准

编码命名标准是指编码分类、代码表、流水码、手动命名、自动命名等方面的规则和规范。

（4）格式标准

格式标准主要体现在模板库、文件格式、文档尺寸、文件大小、文档期限、文档保管格式等方面。

（5）内容本体标准

内容本体标准主要体现在内容分类分级、内容敏感度、敏感词过滤、内容密级、内容模板和内容审批等方面。

（6）版本标准

版本标准主要体现在主版本（生效版本）、次版本（修订版

本）、历史版本、版本控制、版本配置、版本清理、版本策略管理等方面。

（7）元数据标准

元数据标准主要体现在内容属性、内容扩展属性、内容结构、内容标记、内容类别、元数据分类、元数据格式、元数据检验、元数据追踪等方面。

（8）指标标准

指标标准主要体现在内容指标体系、内容指数、内容维度、内容度量和内容指标项等方面。

（9）接口标准

内容服务接口标准主要体现在接口类型、接口引擎、接口集成、接口配置、接口策略、第三方扩展接口服务、应用程序接口等方面。

3.5.2. 非结构化元数据

非结构化元数据是开展非结构化数据管理的基础，组织应当基于非结构化数据战略构建具体的元数据管理战略。

元数据是描述数据的数据（Data about data），主要是描述数据的上下文信息。非结构化数据的元数据，需要在非结构化数据上下文环境中构建关联，便于对非结构化数据进行发现、使用、管控和洞察。组织中的非结构化元数据管理目标体现在四个方面：

（1）形成统一的信息地图与知识传承平台，有助于解决数据孤岛的问题；

（2）形成整个机构或行业范围的指标库，统一指标和业务内容管理过程；

（3）消除系统与内容平台或电子文件的孤立关系，为规划和设计业务提供数据间的内在联系；

（4）维护业务与数据之间的一致性，如一致的数据使用方式、一致的数据服务输出和一致的企业数据流程规范等。

非结构化元数据管理包括非结构化元数据定义、非结构化元数据策略、非结构化元数据权限、非结构化元数据应用和非结构化元数据分析等。其中，组织需要特别注重非结构化元数据应用、非结构化元数据安全和非结构化元数据治理工作。

3.5.3. 非结构化数据质量

高质量的数据是实现数据价值的前提，非结构化数据质量管理需要从数据质量方针、数据质量策略、数据质量制度、数据质量标准等方面开展整体性的构建，且围绕数据全生命周期开展数据质量持续提升的工作，以确保数据质量满足不同业务的需求。

非结构化数据质量管理需要获得业务、信息和技术的全面支撑，且需要获得相应的资源投入支持。落实非结构化数据质量管理和改进实施工作，主要涉及如下方面：

（1）非结构化数据质量要求，数据中是否包含了足够丰富，容易产生价值的结构化信息，涉及非结构化数据的真实性、完整性、可用性和安全性方面；

（2）非结构化数据的质量控制，反映在模板（规则）、流程、技术和人员等方面；

（3）非结构化数据的质量检查，反映在数据质量审计、智能定密、版本比对、目录清单、文件清单、文件元数据清单、文件权限记录和内容库权限记录等方面；

（4）非结构化数据的质量分析，反映在元数据使用分析、关联统计、文件新增对比图、文件新增趋势图、权限记录报表和最终权限报表等方面；

（5）非结构化数据的质量改进，反映在数据质量改进方案和数据质量改进实施等方面。

而衡量非结构化数据质量则需要从数据的真实性、完整性、安全性、可用性和时效性五个维度入手：

（1）真实性体现在电子文件的来源、元数据、数据内容的真实性检测，元数据与内容管理真实性检测，归档信息包的真实性检测等方面；

（2）完整性体现在应该能够覆盖组织的所有文档，组织可以通过文档清单度量文档数据的完整性，包括对电子文件的数据总量、元数据、内容、归档信息包等完整性检测；并且针对各个阶段化的交付成果，验证和检查非结构化数据的完整性。通过非结构化数据质量管理，可以准确获取内容库中的文件数量以及非结构化的文件是否获得相应审批等信息；

（3）安全性方面包括对归档信息内的病毒检测、载体检测、过程安全检测等；

（4）可用性方面则强调通过工作模板确保非结构化数据的可用性，数据内容包括电子文件元数据、文件内容、文件软硬件环境、归档信息包的可用性检测内容质量，以及通过文档控制流程全面审核非结构化文档数据的内容质量，如文档的编制和操作、相关流程的审批和交付文档的目的与要求等；

（5）时效性方面强调通过对非结构化数据进行全生命周期版本管理，并通过文档版本控制流程，提供非结构化文档数据的生命周

期版本跟踪，如对设计类文档的草稿版本、评审版本、发布版本、停用版本和归档版本等全版本的跟踪。

3.5.4. 非结构化数据安全

非结构化数据安全是数据价值实现的保障，组织需要确保数据的全面安全受控。非结构化数据安全管理遵从信息安全和网络安全体系总体要求，侧重对非结构化数据在行为安全管理、统一存储安全、安全管理方法、事件阶段管理、安全制度标准等方面进行体系化构建。

（1）行为安全管理

基于非结构化数据全生命周期的视角，对非结构化数据全生命周期中的采集、存储、传输、处理、交换、管理、洞察、归档等行为进行安全管理。非结构化数据全生命周期安全管理需要遵从行业级业务数据安全监管标准，基于非结构化数据全生命周期，提供有效的安全管理工具及方法，包括网络隔离、安全预警、权限控制、访问监控、行为管控、内容识别、数据过滤、数据加密、数据脱敏、审计溯源等措施。

非结构化数据全生命周期管理需要构建数据安全体系及策略，包括非结构化数据分级分类管理、访问授权体系、身份认证、行为监控等环节，且需要提供完备的数据安全分析支撑。

（2）统一存储安全

非结构化数据主要分为应用系统文件、体系文件和过程文件三种类型。其中应用系统文件是指各种业务系统中的业务支撑文件和成果文件；体系文件则是已经形成体系的存储于共享服务器中的电子文件；过程文件是指个人电脑的各类文件、电子邮件系统中的附

件等。通过安全防控场景化、手段措施多样化、业务渗透融合化、防控环节串联化的多维技术视角，基于统一存储的非结构化数据安全管理才能更可落地、更可控。

（3）安全运维手段

为了对企事业单位组织内的数据进行全生命周期的操作追溯与风险管控，需要采取符合非结构化数据自身特征的安全管理方式。这就需要考虑到数据防勒索、日志溯源等多个方面，在建立起全面的文档安全体系的同时，持续地对非结构化数据管理系统进行监控与测试，通过数据备份、实施校验与日志记录等技术与功能，实现非结构化数据的长治久安。

（4）事件阶段管理策略

根据事件发展的事前、事中与事后的三个阶段进行划分，通过事前预防、事中控制、事后审计机制的事件管理策略，进行非结构化数据事件的闭环管理。

（5）安全制度体系

安全制度体系包括对人员安全、场所安全、活动安全、系统安全、数据安全等方面的体系建设。

除以上五方面的安全体系建设之外，非结构化数据安全应用架构技术还包括登录安全、访问安全、传输安全、数据安全交换、内容安全、日志审计、“文件不落地”、终端安全、离线安全、纸质文件安全、存储安全和预警以及安全分析等核心功能。

（1）登录安全

非结构化数据安全应用架构需要保障登录安全，主要保障措施包括确立密码策略、实施网际互连协议过滤、设立验证码、强制设备绑定、设置登录访问协议单点登录和双因子验证等。

（2）访问安全

访问安全需要确保包括权限模板、访问权限、多级还原、密级权限验证、动态安全水印和共享范围等环节的安全可控。

（3）传输安全

传输安全的环节主要包括安全登录、安全隔离、安全绑定、安全限制、套接字协议加密安全传输等。

（4）数据安全交换

数据安全交换是指不同安全域之间的数据安全交换，其交换方式包括流程审批交换、直接触发交换、批量交换和智能交换等。

（5）内容安全

内容安全主要是通过敏感词汇、智能定密、防勒索、安全域、文控流程、历史版本、病毒扫描、隔离区等功能实现。

（6）日志审计

日志审计主要以日志分析引擎、操作日志留痕、审计报告追踪等方式来实现内容安全审计的目标。

（7）文件不落地

文件不落地需要通过虚拟盘、强制采集和文档安全阅读等技术实现，从而保障用户“操作完全本地化，而数据在云端”。用户能够像在本地磁盘中一样，在虚拟盘中操作各种文件，但数据全部存储于企业服务器中，且能够进一步禁止文件在运行时保存于本地磁盘或外设中，通过这种方式提升了对文件安全控制的能力，更易实现专业和严密的安全防护，从而保障了企业数据不泄露。

（8）终端安全

终端安全的保障能够通过终端数据防泄漏的整合、网关数据防泄漏整合、数据安全漏洞排查、数据安全网关和网闸建设、网络数

据安全集成传输、终端防泄漏预警等措施实现。

（9）离线安全

通过透明加密、外发加密、权限管理系统加密整合等措施，实现数据的离线安全。

（10）纸质文件安全

纸质文件的安全在非结构化数据安全应用架构中同样重要，主要体现在对多功能一体机对接、打印留底与审计追溯、光学字符的全文与区域识别等环节。

（11）存储安全

非结构化数据安全应用架构需要依靠多副本存储、切片存储、强制一致性校验、数据加密、自我恢复、数据备份、多数据中心容灾等措施，实现存储安全。

（12）预警及安全分析

非结构化数据安全应用架构需要具备预警与安全分析的能力，需要对敏感操作进行预警，并在运维、安全、业务分析等方向实现相应的安全分析等。

3.5.5. 非结构化数据合规

非结构化数据合规主要是指组织外部环境下的监管和法律约束。组织需要遵从相关法律、法规进行规范管理和建设，同时需要注重合规、隐私等方面的非结构化数据安全保护。

非结构化数据合规管理是确保数据资产保值、增值和价值变现的基础。组织需要构建完备的非结构化数据合规体系，从政策法规、数据资产、利益相关者和基础设施的角度，进行基于数据全生命周期的合规控制，对数据收集、数据处理、数据保管、数据共

享、数据交易、数据披露、数据处置等各个环节进行合规评估和审计。

常见的可参考数据合规要求包括：欧盟《通用数据保护条例》、美国《萨班斯法案》、美国《2018 年加州消费者隐私法案》、中国《中华人民共和国数据安全法》《企业内部控制基本规范》《药品生产质量管理规范》《药物非临床研究质量管理规范》《质量管理体系》、《中华人民共和国会计法》《财政部国家档案局关于规范电子会计凭证报销入账归档的通知》《银行业金融机构数据治理指引》《中华人民共和国档案法》《会计档案管理办法》等。

组织需要确立非结构化数据的合规原则，如：“拥有者自主”原则、“责权利相一致”原则、公开透明原则、确保安全原则、审批受控原则、“例外处理”原则。“拥有者自主”原则是用于保障数据拥有者对数据资产控制的权利。“责权利相一致”原则是用于保障数据责任方和权利方权利平等。公开透明原则、确保安全原则、审批受控原则是用于保障数据资产安全受控。

3.5.6. 非结构化数据集成

非结构化数据集成是数据共享协同和价值挖掘的前提，主要包括数据分布、采集技术、采集策略和数据集成四个方面内容。

（1）数据分布

非结构化数据常见的三种数据分布文件类型是离散文件、体系文件和应用系统文件。其中离散文件的特征体现为个人拥有的大量有价值并且未整理的文档，如各类记录、邮件、参考资料、工作文件等；体系文件主要为体系化文件、合同、纸质文件、网页内容

等，如企业知识、法规规范、各类单据等；应用系统文件特征体现为需要进行归档与索引构建以及长期保持利用的文件，如审批单、财务报销单、图纸、项目资料、技术资料、产品资料等。

（2）采集技术

非结构化数据采集技术主要包括业务系统适配器、集成开发平台和捕获工具。其中，业务系统适配器是指已经形成的与各种应用系统的连接器，基于这些适配器，可以实时或通过计划任务采集各种类型非结构化数据。集成开发平台包括软件开发工具包、业务组件、应用编程接口、可开发组件。捕获工具则包括打印一体机采集器、电子邮件监控、页面抓取工具、爬虫工具、虚拟打印等。

为了实现采集的有效管理，采集平台需要可视化、可配置化和可监控化，也需要对全内容进行采集，其中包括主业务文件、附属文件、关联文件、元数据、日志信息和数据权限等。

（3）采集策略

非结构化数据集成的数据采集策略从非结构化数据源头出发，将非结构化数据管理系统与业务系统深度融合，将采集策略前置到业务中去，以实现采集的时效性、准确性和内容完整性。

（4）数据集成

非结构化数据集成主要分为两方面，一是为各种应用系统提供实时的、平台型的非结构化数据统一存储服务；二是为新业务应用输出各种非结构化数据服务，从而形成数据与业务的双向融合。

3.6. 非结构化数据价值

创造非结构化数据价值本质上是数据资产化的过程，体现在数据的共享交换与服务开放。这里从价值实现技术角度关注非结构化

数据协作、流转、服务和洞察。

3.6.1. 非结构化数据协作

非结构化数据协作是数据价值体现的基础。结构化数据与非结构化数据在协作方面具有一定的相同性，它们都需要频繁地被多个用户进行编辑和协作；但是，也存在差异性，主要体现在结构化数据的颗粒度过细，在具体协作过程中无法表达和解释具象内容，协作者需要借助上下文环境才能够进行协作，应用场景固化且灵活性不够。而在非结构化数据协作中，由于非结构化文件能够呈现对某一事物更为完整的描述和说明，所以协作过程可以从整体视角出发，涵盖大量信息及丰富表现形式，其应用场景更为广泛、协作频次更高。

目前，国内非结构化数据的协作以网盘和微办公环境场景体现，较多企业也已经将企业网盘等软件视为必备的数据协作组件。而数据协作可以对事件和项目等进行及时、有效的协作支持。例如：工程总承包的工程项目公司，可以依靠工程项目过程中产生的非结构化文件对整个项目进行管控，其中包括开工报告、施工过程文档、竣工资料等不同阶段的文件。

组织在进行完备的非结构化数据协作体系构建时，需要基于对内容协作能力的提升需求、跨组织库的内容协作，以及融合业务场景的内容协作，而这些能力可体现在如下方面：

（1）内容协作能力

内容协作能力主要体现在提供协作相关方专属的团队库、对内容的协同编辑、对内容的签入签出、对内容的版本管理、内容变更过程的消息提醒和支持对内容评论批注等。

（2）跨组织库的内容协作

跨组织库的内容协作包括在企业内容库、团队内容库和个人内容库之间的协作。其中，企业内容库能够满足企业组织条线和业务条线的内容管理需求；团队内容库能够满足团队、项目等临时组织的内容管理和协作需求；个人内容库能够满足个人的内容管理和备份需求。内容库的构建需要注重协作模式的构建方式，目前主要有企业内容库的固化式协作，团队内容库的松散式协作，以及基于个人内容库通过共享和外发与其他用户进行的临时性协作。

（3）融合业务场景的内容协作

融合业务场景的内容协作是指实现业务与内容的深度融合协作，将内容嵌入到业务场景中，如项目协同管理、文档档案一体化、合同管理等业务系统，在支撑相关内容场景协作的同时，也与内容库的管理进行深度融合。

3.6.2. 非结构化数据流转

非结构化数据流转是企业数据价值释放的关键，其技术实现需要涵盖数据流转过程中的数据安全和数据流转方式两个方面。

（1）非结构化数据流转安全

非结构化数据流转的安全需要考虑访问权限、透明加密、敏感检测、脱敏、智能检测、流程审批、查杀病毒等环节。流转过程中的安全需要遵从数据分类分级原则，数据流转的安全性则可通过文件安全交换解决方案中的流转安全手段、方式、安全交换等技术实现。

（2）非结构化数据流转方式

非结构化数据流转主要通过共享、链接外发、联邦外发、附件

发送、逻辑安全交换、网闸安全交换等方式实现。非结构化数据流转的技术实现方式可划分为推送式流转和发现式流转两个类型：推送式流转是根据一定规则对文件进行自动化的渠道派送，根据组织对内容的组织规范和要求进行流转；发现式流转最常见的形式是“知识管理”，因其具备多维度扁平化组织能力，发现式流转具有较高的数据组织性，用户获取数据和使用数据会更完整，从而使数据价值得以充分释放。

3.6.3. 非结构化数据服务

非结构化数据服务基于业务导向、以用户为中心开展具体技术实现，并不断通过业务流程重构，将流程与组织结构解耦、去职能化，以服务和用户角色为原点进行流程设计，构建以文档内容和数据服务为核心的非结构化数据应用，实现以内容为核心的业务构建，创造端到端的内容价值实现，全面将内容与业务紧密整合，以保证使信息系统成为业务战略的载体，使业务能力在端到端的贯通中得以全面地呈现。

在进行非结构化数据服务的应用时，组织需要基于业务进行整体设计，切入业务场景中去，以构建基于内容业务的端到端实现；将服务固化到流程引擎和表单引擎之中，通过业务流程和活动，将业务行为模型、角色行为模型、事物行为模型与主数据、业务规则、参考数据、经营指标、管理指标、绩效指标等核心要素平台化，以门户页面的方式对外发布非结构化数据服务，强调持续迭代，长久地进行以内容为主体的业务优化和服务监控管理。

3.6.4. 非结构化数据洞察

非结构化数据洞察可以提供更广泛的业务价值实现。但对非结构化数据的洞察需要大量的准备工作，首先，需要获得大数据集的海量数据支撑；其次，关注技术实现的细节，如：知识图谱、文件关联图谱、主题图谱、自然语言处理和人工智能引擎等方面的应用；然后，通过智能关联、智能推荐、智能搜索、智能识别、智能分类、智能问答、智能定密等多种应用，提供基于文本、单据、物体和人脸等为主体的多种应用场景，获得多种形态下的数据洞察；最后，将洞察能力与业务场景进行结合，发挥洞察创新，从而通过数据洞察实现业务价值转化的目标。

非结构化数据洞察能力主要包括以下几方面：

（1）统一搜索

连通各业务系统、数据源，实现结构化知识、非结构化知识、内部知识、外部知识的集中与统一，通过一站式统一搜索挖掘数据价值。

（2）智能搜索

基于自然语言处理、机器学习技术，结合点击反馈模型等搜索排序算法，利用大规模分布式索引与算法模型的计算与分发，构建强大的知识内容搜索引擎。同时全面整合人工智能能力和自然语言处理技术：结合识别分类转化、聚类回归分析算法、机器学习、用户画像、文本图像、深度学习等技术进一步提升能力。

（3）智能推荐

智能推荐可以基于用户属性、用户行为、业务场景进行分析，通过大数据技术，整合奇异值分解、支持向量机等尖端算法，构建数据挖掘系统，生成用户画像，为用户主动推荐其感兴趣或与当前

工作相关的知识内容，同时也可以激活整个知识库，发挥长尾效应。

（4）知识图谱

通过构建大规模语义网络，发掘实体之间的关联，将数据进行整合，帮助机器理解数据、解释现象、知识推理，从而发掘深层关系、最终实现智能交互。

（5）数据分析与挖掘预测

非结构化数据系统可以基于数据挖掘，可视化地呈现数据总量、数据变化走向与数据分析成果，为用户提供数据预测能力，从而以数据辅助用户的战略与决策。

（6）数据洞察阶段

以非结构化数据系统，结合大数据分析能力，深度融合以上五点能力，从而实现非结构化数据的深度洞察。

4. 非结构化数据管理解决方案

4.1. 非结构化数据管理与 ECM 企业内容管理

非结构化数据管理在企业实践中主要体现为 ECM 企业内容管理，其解决方案是通过企业内容管理系统，统一协调、管理各项非结构化数据应用工作，并保证其具体的落地与实施。

内容是指各类文档中包含的数据，其中以文本、图像、音频、视频等非结构化数据为主。ECM 企业内容管理是一种战略、方法和基础设施，来帮助企业获取、管理、存储、保护、利用和洞察企业组织流程相关的非结构化数据（如图 7 所示）。

图 7 企业内容管理

ECM 企业内容管理是一种专注于非结构化数据领域的软件类型，涵盖了企业网盘、文档管理、知识管理、文件安全交换、工程协同设计、文件安全外发、档案管理、影像文件管理、电子文档安全管理、文档云、ISO 质量文件体系管理、GMP 质量文件体系管理、非结构化数据管理平台、工程内容管理等应用软件，以及基于 AI 智能和 Graph 知识图谱技术的智能推荐、智能搜索、智能定密、智能安全分析等内容智能应用。

ECM 企业内容管理系统可以帮助企业内容管理战略完成从理念到实践的转变，通过内容获取、管理、存储、保护、利用等方式挖掘和释放内容价值，最终促进企业数字化转型，提升企业运营效率，并获得企业商业洞察能力与长远竞争优势。

Gartner 于 2017 年修正了企业内容管理的定义：企业内容管理是一种服务和微服务（如图 8 所示），包括内容协作平台（Content Collaboration Platform）、内容服务平台（Content Service Platform）和内容业务平台（Content Service Application）。具体表现为一个集

成的产品套件或具有通用 API 接口和多储存库的平台型软件，利用不同的内容类型且服务于多分支组织机构和各种应用场景。

在《内容服务平台魔力象限 2021》中，Gartner 还提出：嵌入式智能已经成为内容服务平台的主要趋势——人工智能已经成为内容服务的关键，从通信管理到案例管理，它将越来越多地嵌入到真实的业务解决方案中。

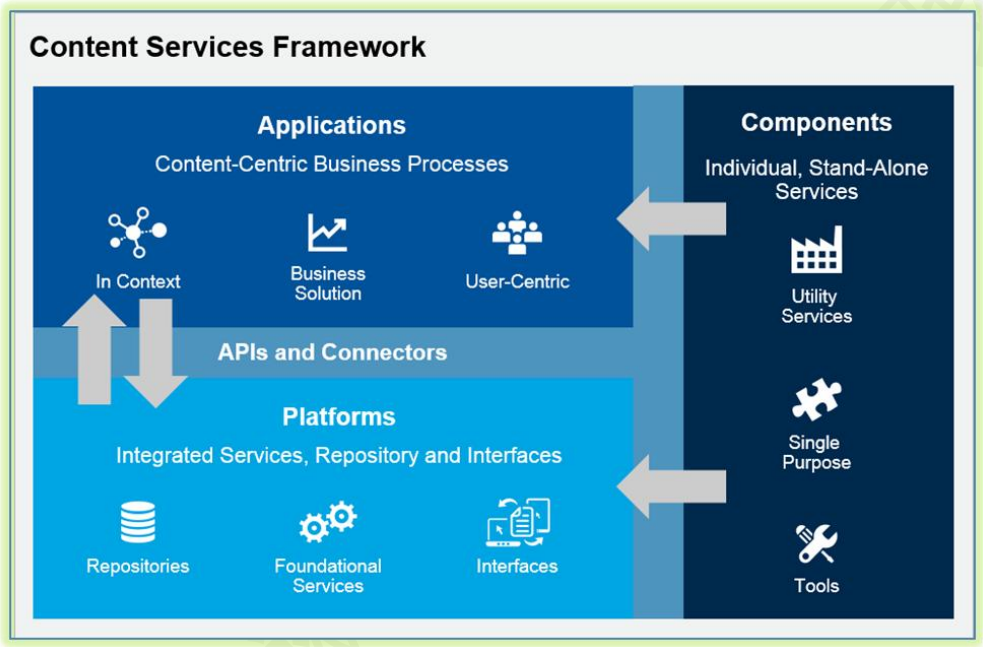


图 8 Gartner 内容服务框架

进一步分析 Gartner 的内容服务框架，其主要包括内容管理平台 CMP（Content Management Platform）、内容服务应用 CSA（Content Service Application）和内容组件（Content Component）。其中内容管理平台是底层内容统一存储和统一管理的基础平台，提供各种 API 接口和 Connector 连接器等集成支撑；内容服务应用强调以内容为中心的业务应用；内容组件则是一种类似转档、预览、编辑等细颗粒的内容服务组件，其能力可输送于内容服务平台 CSP（Content Service Platform）和内容服务应用。

企业内容管理的本质是为企业业务和数字化转型提供内容服务

支撑，并提供内容服务的快速响应能力。基于内容服务平台的内容服务应用 CSA 分为体系化 CSA 和场景化 CSA。其中体系化 CSA 覆盖了垂直业务领域的内容服务，场景化 CSA 则着眼于第三方业务系统的集成和整合。完整的内容服务框架（如图 9 所示）的底座是内容服务平台，中层是基于低代码开发技术的内容业务平台，上层构建起内容协作、内容安全、内容管理、内容治理、内容合规、内容业务、内容智能等各种内容应用场景。

图 9 内容服务框架

4.2. ECM 内容管理成熟度模型 CM³

内容管理成熟度模型 CM³（如图 10 所示）是基于多个行业和领域的非结构化数据实践应用以及不同阶段的内容管理特征总结提出的，其中包括内容协作阶段（Content Collaboration Platform）、内容服务阶段（Content Service Platform）、内容业务阶段（Content Service Application）和人工智能（AI）四个阶段。

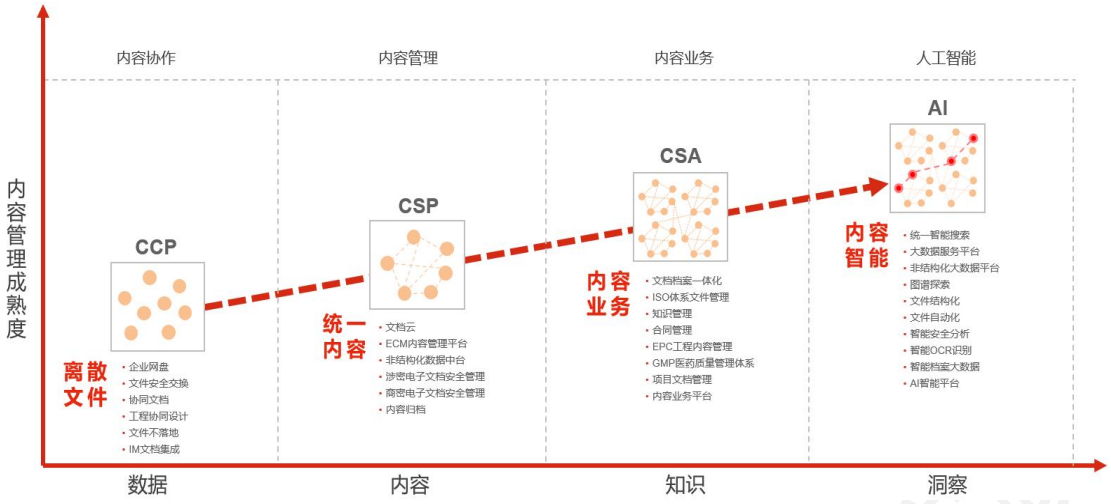


图 10 内容管理成熟度模型

第一阶段是内容协作阶段，此阶段文件呈离散型存储在用户的电脑上，仅能完成文件之间的协作，不能与业务融合。在组织里，有 80%以上的离散文件以此种形态存储，所以此阶段可以概括为离散文件的协作管理阶段。

第二阶段是内容服务阶段，此阶段中的企业数据以业务系统文件、体系文件等形式存在，并通过内容元数据进行网状式广泛关联，使得数据来源于业务，又输出服务于业务。这是企业数据从文件到内容的一个过渡阶段，数据经汇聚、整理、处理后，以全内容服务的形式开放，构建统一的企业非结构化数据管理平台。

第三阶段是内容业务阶段，在这个阶段中，行业数据经过不同维度地整理、提炼，并围绕业务的垂直领域性、体系性进行立体式地关联与聚合，最终形成行业性的知识体系，以支持企业迅速应对市场变化和进行业务创新。

第四阶段是人工智能阶段，这是一个通过深度学习、自然语言处理、知识图谱等技术对海量数据进行深度处理的阶段，使内容管理全生命周期各环节具备智能能力，从而辅助企业的降本增效与精准决策。

总体来说，这四个阶段是企业数据从内容到知识再到智能化的全面提升的过程；同时数据安全性、数据关联性、业务融合度和数字化能力，也随着阶段发展逐阶提升（如表 2 所示）。

表 2：内容管理成熟度四阶段特征

特征类型	第一阶段	第二阶段	第三阶段	第四阶段
数据形态	过程性、离散的用户电脑中的文件	业务系统文件，体系文件	经过整理、提炼后的行业领域性数据、从不同维度视角归纳后的数据	海量数据，标注数据
数据颗粒度	文件级	内容级	知识级	语义级
数据间关系	离散的，不关联	通过内容元数据进行网状式广泛关联	存在围绕业务的垂直领域性、体系性关联，是一种立体式关联与聚合	语义层关联，主题图谱和实体知识图谱
业务融合	满足文件协作，无业务融合	数据来源于业务，又输出服务于业务	深度融合业务，甚至形成业务应用	立足于业务场景上内容智能
数据安全	协作过程中的数据流动，需要在存储、访问、流转等阶段的安全保护能力，属于企业安全防护初级阶段	全形态数据防护覆盖，内容级颗粒度安全措施更精准，基于统一存储的全生命周期全方位安全防护体	业务场景化的渗透性安全防控，安全服务紧随业务，为业务中的数据安全保驾护航	多层面图谱结合用户画像和操作日志，深度追溯数据安全问题，洞察预测安全风险
数字化能力	文件协作层面的数字化	内容中台的数字化，数据可以经汇聚、整理、处理后以全内容服务形式开放	业务数字化，基于低代码平台形成垂直领域的业务体系化应用，支撑企业迅速应对市场变化并进行业务创新探索	智能数字化，利用自然语言处理、人工智能、图谱与大数据技术为企业降本增效，辅助企业决策

4.3. ECM 内容管理平台架构

组织在开展数据管理能力规划和提升时，需要从总体视角考虑大数据环境下 ECM 内容管理平台的总体架构设计。组织需要构建基

于非结构化数据战略、方法和工具的 ECM 内容管理平台，使其提供基于全生命周期的数据采集、存储、保护、管理、使用、交换和归档等能力，并使其与组织业务流程活动中相关内容和文档进行深度融合与应用。

ECM 企业内容管理平台具有数据采集、存储、治理、服务、应用、洞察和安全等全方面的能力支撑，帮助企业对非结构化数据开展全生命周期的管理，其能力具体如下：

（1）数据采集

平台建立起了一套有效的内容数据分类机制和方法，根据内容数据的重要程度，可通过不同采集方式实现资料采集；在管理体系上建立“事前提醒催办，事后汇总分析”的机制，形成对内容数据的全面管控。数据采集形式主要包括用户主动上传、端点强制采集、API 集成采集、外网爬虫采集、邮件内容采集、打印一体机采集等多种采集手段，能够适应企业组织内多样的应用场景。

（2）数据存储

ECM 企业内容管理平台能够承担组织应用系统投产后所有新增非结构化数据的集中存储，基于统一的分布式对象存储方式，具备海量数据存储、高性能读写、加密存储、多副本存储、便捷的水平扩展、冷热数据分离、全类型存储接口支持等特征。

（3）数据治理

数据治理包括对数据标准、元数据管理、数据安全、数据流转、数据质量、内容库模型、权限体系模型、分类模型、数据健康度等的综合治理，能够提供完整的数据治理情况总览和分析。

（4）数据服务

通过功能组件和中间件提供非结构化数据服务，负责将平台底

层的公共能力输出到各业务应用。数据服务范围涵盖全业务服务内容，通过数据服务内容、数据服务技术、数据服务模式，基于统一内容服务总线架构，以实现数据资产可视化、可管理和数据资产的价值变现。

（5）数据应用

数据应用是指通过 ECM 内容管理平台提供各种非结构化数据的协同编辑、共享外发、统一搜索等基础应用，以及基于 ECM 内容管理平台上层的体系化业务应用，例如项目文档管理、合同管理、知识管理等。

（6）数据洞察

数据洞察的核心是基于人工智能和图谱技术实现的非结构化数据知识图谱。通过利用实体图谱、语义主题图谱和文件图谱，构建起非结构化数据完整的知识图谱，将内容深层的逻辑关系进行梳理和呈现，从而实现对非结构化数据的全面洞察。

（7）数据安全

提供访问安全、数据摆渡、离线安全、内容安全等服务。其中访问安全包括权限模板、访问权限、多级还原、动态水印、共享范围、密级权限验证等；数据摆渡包括直接触发数据摆渡、流程审批数据摆渡、批量计划数据摆渡、智能内容数据摆渡等；离线安全包括透明加密、外发加密、DLP 边界防控等；内容安全包括敏感词、病毒扫描、智能定密、安全域、文控流程、历史版本和防勒索模块等安全能力模块，从多方面、多层级、多维度保障数据的安全可追溯。

4.4. ECM 内容管理核心技术

4.4.1. ECM 底层架构技术

现代 ECM 需要支持多种应用场景下的大规模集团化架构，包括集群架构、分区域架构、联邦架构、混合云架构、混合云架构和多中心架构。同时需要支持 PB 级的分布式对象存储，并实现数据的冷温热分层、自我恢复等；支持 10 亿级海量小文件的极速寻址。目前领先的 ECM 系统底层架构需要基于微服务和容器化的云原生（Cloud Native）技术实现，配合 APM 监控运维平台，起到对系统的监控与观测的作用。

4.4.2. ECM 服务技术

ECM 内容服务包括对不同格式文件的转档与预览服务、上传下载与在线编辑等文件操作类服务、文档权限类服务和内容搜索类服务等。实现 ECM 内容服务的主要技术包括：CSB 内容服务总线技术，内容服务可视化技术，服务监控与调度管理，Metadata 元数据建模与服务技术等。通过可视化数据采集技术，ECM 能够将内容数据汇聚一起，进而实现与各种应用系统融合，融合后的内容数据经过治理后，将以服务组件、WebAPI 等方式输出标准的内容服务。

同时，为适应企业数据管理能力的提升，需要企业内容业务平台能够具备表单建模、BPM 流程引擎和 WCM 门户展现引擎，实现“一次拖拽、多端适配”，让业务人员具备应用开发能力。同时提供丰富的业务组件、接口集成平台、支持标准 WCP 控制模式的工作流，可大幅降低开发和维护成本。基于端到端技术的企业数字化能力与内容形成闭环，实现了组织业务快速响应与持续创新。

4.4.3. ECM 文档处理技术

ECM 作为企业非结构化数据“收管存用”的统一平台，能够为使用者提供文档处理与使用的一站式服务。一方面，ECM 能够以标准组件的方式输出服务，实现快速集成，为企业用户提供包括文件的合成、拆分、格式转化与版式副本的创建服务。另一方面，它也能够深度融合 AI 人工智能，赋能用户自主挖掘非结构化数据价值。

4.4.3.1 文件合成、格式转化与版式副本

基于 ECM 标准的文档组件，用户可以将多个文档中的内容进行合并，避免了将文档内容复制粘贴的繁琐工作。为了实现各类文档的在线预览，ECM 能够对多数企业内常见文档类型进行格式的转化，从而保障用户能够流畅地对文件进行浏览与使用。

同时，版式文档作为企业内常见的文档形式，能够防止文档内容遭到篡改，在各类软件与操作系统平台上都呈现出相同的内容效果。ECM 能够将多数流式文档进行转换，补充著录或标签信息，形成 OFD 或 PDF 格式的文档副本，结合 ECM 的权限策略，更好地控制文件的安全流通。

4.4.3.2 OCR 文档内容识别

ECM 中存储着大量的图类型文件（扫描件、图片文档等），这些文件往往蕴含着丰富的信息。为了充分挖掘与利用这些信息，可以通过光学字符识别技术（OCR）将文件中所包含的全文文字进行整体提取，为智能检索、文档抽取等应用场景创造了前提条件。

4.4.3.3 文档标签

基于 NLP 自然语言处理技术，结合知识网格的规划，ECM 能够针对含有文本信息的文件进行自动提取，形成符合业务逻辑的内容标签。通过使用 TF-IDF 等基于统计加权技术的成熟模型，对上传至系统的文件快速提取语义标签，当用户选择检索某一标签名时，就能够找到被打上相同标签的文件集合。

4.4.3.4 内容检测

为了防止敏感信息的泄露，ECM 能够提供 DLP 审批策略，辅助文档的审批流程，结合自定义的机械规则以及基于无监督的 AI 模型算法，共同组成复杂的 DLP 规则策略，从而精确地甄别文件，有效违规风险，辅助文档定密，实现内容敏感监测和合规监测。

4.4.3.5 文档抽取

在 ECM 中，能够利用语义特征提取的 AI 技术把图片、扫描件（包括合同、报告、证件等）中的关键信息要素提取出来，形成结构化的数据，便于检索、统计分析等需要。通过使用训练成熟的语义特征提取算法模型，精准识别文档上下文段落中的语义特征，提炼关键信息要素，再以结构化的方式存储，形成“元数据”，为非结构化数据的分析和洞察提供基础。

4.4.3.6 文档智能审核

通过将文档内容提取技术与自定义的文档审核规则策略相结合，ECM 能够实现对文档中关键信息要素的审核，并且能够针对不合规的内容进行提醒并给出修改建议，智能审核卡证、票据、合同等多种文档内字段、数值的合理性与合规性。

4.4.4. ECM 安全技术

ECM 安全技术主要包括以下几个方面：

数据存储安全，基于数据块和多副本技术的数据融灾，保障数据存储的安全和可靠；

数据使用安全，通过细颗粒度地访问权限控制、密级权限验证和安全域边界权限等技术保证多层数据防护，基于图权限计算模型对深层亿级海量文件进行毫秒级权限计算；

数据流通安全，基于内核过滤驱动保证文件保存在终端不落地，基于智能 DLP 敏感检测保证敏感数据无法摆渡和外发；

数据审计安全，基于大数据和知识图谱技术，满足各种场景化的数据安全审计和分析需要。

4.4.5. ECM 存储技术

面对企业组织内日益增长的数据量，ECM 具备全对象的存储能力，能够提供包括动态扩展、高性能 IOPS、集群多副本、自我恢复、存储加密等服务；支持多类型的混合存储、单实例存储，同时兼容市面上主流的存储系统，基于规则引擎实现动态存储，给予用户无感地海量文件导入与使用体验；通过多级存储与自动迁移，实现在线存储与归档存储一体化，为 ECM 持续提供优质数据服务打下坚实的基础。

4.4.6. ECM 传输技术

在大规模数据安全存储的基础之上，ECM 作为企业的内容统一管理平台，需要为全员提供数据的访问、上传、下载、编辑等服务，这就对 ECM 平台的传输效率、速率、功能、安全等多个方面提

出了要求。

ECM 能够实现跨国、跨区域的传输加速，并且能够通过就近上传下载缓存，实现文件利用速率的提升。针对大体积文件，采用分块传输，从而能够在本地数据库中记录每一块成功传输的数据块，在发生网络异常或者用户手动暂停传输之后，于下一次启动传输的时候从本地调取未传输的数据块位置，从而实现断点续传的需求。此外，还能够根据权限，做到针对用户、部门的上传下载限速，确保文件利用的优先级。

4.4.7. ECM 与人工智能

在 ECM 系统中，需要将人工智能的关键技术的机器学习、深度学习、NLP 自然语言处理与大数据技术进行深度融合，通过对模型语料、算法、训练、评估、发布和持续迭代的全生命周期管理，实现了对文本和图像的智能分类、智能标签、智能 OCR 识别、智能抽取和生成等。

通过结构化 D2R 技术、半结构化 Wrapper 技术和非结构化 NLP 文本抽取技术构建起 Graph 知识图谱。非结构化数据知识图谱同时融合本体知识图谱、基于语义抽象的主题图谱和文件关联的文档图谱这三大图谱；并结合用户画像与行为日志，实现启发式可交互的非结构化数据探索能力。该能力可应用于智能搜索、智能推荐、智能定密、智能安全分析、知识创新和辅助决策等领域。

4.4.8. ECM 生态融合技术

ECM 平台具有强开放性，能够通过广泛的适配器、多种集成模式、全内容整合等手段，构建起多行业、全生态的融合技术。基于

此，ECM 能够深度集成和融合财务类、ERP 类、OA 类、PDM 类、IM 类、存储备份类、加密安全类等各种企业应用系统，实现了企业和组织非结构化数据的统一存储、统一管理和统一服务，横向打通了组织内的信息孤岛，构建起统一的非结构化数据中台，实现数据互通、业务融合，为企业和组织的业务创新与精准决策提供了完整且有效的非结构化数据支撑。

4.5. 新一代 ECM 平台的发展方向

作为非结构化数据管理的一种通行的软件类型，在国际上，自 2006 年 SharePoint 问世以来，ECM 已经走过了十几个年头，从最初的开源内容管理平台，到取代 DMS 文档管理，成为企业内承托内容的一体化平台。科技飞速发展，内容管理的技术与形式也日新月异，2018 年 Gartner 也提出，将 ECM 拆分为内容服务平台 CSP、内容协作平台 CCP 和数字体验平台（DXP），以三个各有侧重的平台，去支撑企业内容的管理、治理与处理。

而遍观国内，随着平台定位的精细化与企业需求的多样化，ECM 也面对着升级与迭代的强烈需求，针对国内企业的数据管理需求与实际应用场景，ECM 未来的发展方向大概能够分为以下几个方面：

4.5.1 成为企业基础设施

数字化时代，数据始终占据着企业发展的核心地位。企业业务进行中的过程文件、内部日常运营所产生的办公文件、各类场景下的体系化文件与 OA、ERP 等业务系统中的文件，共同组成了企业内流通的非结构化数据的总和。因此 ECM 需要具备统一收集、存储这些非结构化数据的能力，打通企业内部的数据孤岛，发挥底座能

力，帮助企业构筑一站式的企业非机构化数据平台。

数据爆炸的时代，企业每天都在面临着海量新数据的收集与利用的难题。ECM 应当以强大的底层能力，实现企业 PB 级别以上的数据存储，灵活扩容；同时能够轻松处理高并发请求与跨国、跨区域的文件传输需求，支撑大规模的内容利用。

4.5.2 挖掘内容生产力

非结构化数据占据企业内数据总量的 80%，但管理难、数量大、信息量庞杂，相比之下，结构化数据易分析、易利用，拥有更成熟的管理方式。为了提升非结构化数据管理效率，ECM 需要以元数据及各类智能手段，抽取非结构化内容中的实体，打通结构化数据与非结构化数据的转换壁垒，实现了对非结构化数据的统一高效管理。

同时，企业也需要 ECM 具备解析内容文档的大颗粒结构、将非结构化数据庞大的信息量细颗粒化的能力，能够将文件检索精准到关键词与关键实体，从而大幅度压缩内容筛选与检索的时间成本。在帮助企业在提升内容利用效率的同时，解放内容价值，降低内容获取与复用的门槛，促进企业内容管理再升级。

除此之外，通过 ECM 的各类组件与功能，对内容进行整理、提炼与重新组织，企业能够提高自身数字资产管理能力，实现知识武装员工、内容指导业务的战略级目标。将企业内各类知识体系化，然后智能推送到不同部门，形成个性化的知识汲取环境，从而创造学习型组织，构建良好的企业文化。

4.5.3 激发内容应用合力

面对企业日益复杂的内容应用需求，ECM 应当立足于内容管理与文档管理，助力文档协同作业，统一数字资产管理，融合实际内

容业务，智能发掘数据价值，构筑出涵盖企业网盘、文档云、档案管理、知识管理、智能搜索等类型的内容管理解决方案，更有针对医药、制造业等特殊场景的内容管理应用，全面助力企业非结构化数据管理体系建设。

未来，企业内的 ECM 平台不仅要以内容采集、存储、传输、处理、使用、保护、交换、归档为核心，结合企业常见内容场景，逐步构建一站式平台，还需要连接公司营运与业务的各个环节，打通数据隔阂，在保障数据安全合规的同时，以大数据、人工智能等先进手段，实现企业内容的全生命周期管理与价值实现。

企业在数字化转型与挖掘数据价值的过程中，往往会遇到文件离散、系统孤立的数据存储问题，并且伴随着内容质量低下、文件检索困难、数据支撑不足、内容难以合规等多重挑战。以 ECM 作为企业非结构化数据的管理平台，承托非结构化数据的各个使用环节，从而全方位提升企业内容应用能力。

4.5.4 构建内容数字空间

随着数字化发展的趋势愈发清晰，企业需要根据自身特点制定一套全局的数字化目标。未来的 ECM 应当能够引导和帮助企业将业务逐步数字化，打造企业专属内容数字空间，整体性建造统一的内容价值链，避免多个系统造成的数据孤岛。助力企业不仅从业务层面，更从公司日常营运、员工培训提升等多个方面，实现线上线下协同发展。

借助 ECM 平台，汇集企业各类数据，提炼、转化为有价值、可传递的知识内容，统一进行管理与利用。基于企业的业务内容类型，通过内容元数据进行网状式广泛关联，让数据来源于业务，又输出服务于业务。企业数据经过汇聚、整理和处理之后，以全内容

服务形式进行开放，构建起统一的企业非结构化企业管理平台，驱动业务高效发展。

位于竞争快车道的企业，无时不刻需要通过创新发展巩固自身市场地位。而创新从来都不是无根之水，通过 ECM 平台，深度融合智能与知识管理手段，打造企业“内容新基建”，植根于企业过往的案例、经验，充分汲取企业内容中的养分，构筑企业创新内核，赋能企业智能化、规模化、数字化的创新之路。

4.5.5 深化 AI 赋能赋智

基于内容管理底座，ECM 有能力构建起一整套涵盖非结构化数据与各类业务结构化数据的内容智能体系。通过采集相关数据、创建模型，经过深度学习与自主优化，文本智能可以帮助企业实现文字审核、抽取、分类等功能；基于光学字符识别与机器视觉算法，图像智能能够实现各类 OCR 功能与目标的检测与分类；以文档、主题、实体为节点，构建全景的知识图谱，从而反哺文档智能搜索与智能推荐，最大程度释放内容价值。

同时，企业内各类实体之间有着纷繁复杂的关系，以静态的传统目录方式很难完整地对其进行实时展示。ECM 能够通过知识获取、知识融合、知识存储、语义理解、知识检索和可视化展现等多个模块，将经过梳理、总结的知识传递给用户，构建企业知识图谱，激发用户求知欲，实现智能探索。

5. 非结构化数据管理应用实践

5.1. 非结构化数据管理应用类型

随着信息技术全面升级，很多组织不仅关注内容管理层面，更

关注利用内容，助力业务流程，进一步提升生产效率，以及挖掘数据价值，从而获得商业洞察能力。受行业背景及企业业务流程等因素的影响，企业在选择解决方案时需结合行业特色和企业业务过程，以非结构化数据管理作为支撑，梳理内容创建、管理、储存、保护与应用等过程，并制定相关的管理机制和管理体系。典型的企业内容管理系统包括内容协作和交互、内容全生命周期管理、统一的内容管理平台、内容的知识化平台、内容归档和合规管理，以及电子文档安全管理等。

（1）内容协作和交互

企业内容管理系统可以通过完善的共享、外发、扩展编辑、检索等环节实现远程办公，文档协作和交互，并通过完善的权限管理机制，保证文档应用的安全。

（2）内容全生命周期管理

通过企业内容管理系统，能够针对电子文件的全生命周期进行管理，并且根据企业业务过程管理企业内容，建立全程可追踪的管理体系。

（3）统一的内容管理

统一的内容数据管理平台可以通过标准化应用程序接口，与各大系统对接，建立跨系统内容协作，并通过分布式部署等途径，打破物理空间对内容管理的限制，以整合各类数据的资源，提供数据能力。

（4）内容知识化

企业内容管理系统可以对知识进行沉淀与传承，构建知识管理和应用体系，使显性知识规范化，并沉积存储在员工头脑中的隐性知识，助推隐形知识显性化进程。

（5）内容归档和合规管理平台

企业内容管理系统可以自动化和智能化地收集、管理、保存和利用企业内有价值的数据信息，然后基于人工或自动的判断，流程化地对内容进行归档，留存企业宝贵数据。同时保障企业内数据遵循法律法规，满足企业合规性管理，帮助企业从容应对国内外各类质量体系要求。

（6）电子文档安全管理平台

针对非结构化数据进行体系化管理，可以用于涉密或商秘电子文档的集中存储及安全管控，为企业提供全程安全的文档管理业务系统与可追溯可控制的数据应用环境。

5.2. 非结构化数据管理应用实践

5.2.1. 某大型药业集团内容协作案例

应用领域：公司管理

应用场景：公司文档管理

案例提供者：某大型药业集团

（1）案例中存在的问题

某大型药业集团在企业数据管理中面临几个重要的问题：一是各个厂区间的质量车间关于规章制度的互相协作修改问题；二是企业规范文档的在线创建问题；三是分公司之间多人同时编辑文档的问题。

（2）解决方案介绍

打造质量体系文件管理及在线协同编辑体系，实现基于办公组件的文件多人编辑模式，并且在编辑模式下，显示不同用户修订过的记录，标注用户名、时间、内容，给协同编辑工作带来极大的方

便。在有效提高工作效率的同时，压缩了大量会议的成本和其它沟通成本。

此外，对企业实验室进行数据安全备份，使数据完整性、安全性、一致性得到保障，让员工工作过程中对文件的操作有据可循，避免责任推诿，搭建高效易用的在线协作平台，切实提高企业研发效率。

5.2.2 某地铁公司内容全生命周期管理案例

应用领域：公司数据管理

应用场景：数据资料管理

案例提供者：某地铁公司

（1）案例中存在的问题

为了保障地铁安全运营，实现对安全保护区项目合同的全过程管控，提高内部工作效率和管理水平，并更好地对地铁轨道监护办公室累计二十多年的宝贵数据资料进行管理，如技术数据、图纸、文件等。某地铁公司急需建立一个完整的综合管理平台。

（2）解决方案介绍

通过构建文档档案一体化管理平台，实现对文件采集、图片处理、索引分类、传输、海量存储、查询分发、归档和销毁等电子档案的全生命周期管理。该平台能够为监护公司档案室提供服务功能，包括入库、借阅等，切实提高地铁监管部门的档案管理能力。同时，通过信息共享和业务协同业务，实现管理工作的信息化、流程化、无纸化，使内部管理可视化、知识化。

5.2.3 某电力集团统一的内容数据管理平台案例

应用领域：公司内容数据管理

应用场景：数据信息检索

案例提供者：某电力集团

（1）案例中存在的问题

某电力集团目前存在的问题为：不同的业务系统中的数据分散，随着时间的积累，非结构化数据的数据量急剧增长，员工查找和利用数据信息极为不便。同时，针对海量的信息数据，如何快速精确查找所需的信息内容成为首要问题。在当前的各业务系统中均仅提供简单的、基于指定字段的检索，但其指定的检索字段、检索范围、检索方式对使用人员的操作门槛高，无法满足使用人员通过简单的关键字进行跨分类、跨字段的全文检索需求。

（2）解决方案介绍

构建非结构化数据中台以统一数据存储，打破各业务系统的数据孤岛，实现基于统一内容库的统一搜索和分析利用。通过如电力生产管理系统、企业信息继承应用系统、电子商务平台、图纸管理系统等系统将实现组织架构同步、第三方业务系统单点登录，仅维护一套账号体系，大大降低工作量。

系统基于独立性、可靠性、实用性、多信息源、可扩展性、可维护性的原则，强调与其他业务系统的协同和数据共享。实现本平台与电力生产管理系统、企业信息集成应用系统、电子商务平台、图纸管理系统、人事管理等系统的对接，将各业务系统中数据信息进行集成。继承业务系统中的业务权限体系，实现对进入检索系统中的信息内容的权限管理。提供统一的搜索利用平台，快速便捷的综合检索，满足公司员工对公司跨平台海量信息的检索、以及通过简单的关键字进行跨分类、跨字段的全文检索需求，提升检索效率以及准确性。

5.2.4 某金融联合组织电子文档安全管理案例

应用领域：数据管理

应用场景：数据中心

案例提供者：某金融联合组织

（1）案例中存在的问题

某金融联合组织目前处于银行卡产业的核心和枢纽地位，其面临问题为：一是文档分散，无法统一存储；二是缺乏安全有效的外部数据交换手段；三是各业务系统间的信息孤岛；四是文件查找使用不便，利用率低。

（2）解决方案介绍

建立统一非结构化数据中心，实现密级文件管控及防泄露管理，基于文件内容识别，甄别上传文件是否属于国家机密或一级商密文件，同时对上传文件进行删除、转移等安全控制。使各密级文件遵循业务管理规范，无缝整合流程平台，实现流程审批并由数据中心对外发布。

5.2.5 文档云应用解决方案

文档云是基于企业内容管理平台的应用解决方案，通过网盘应用将个人电脑上的文件全部上传至文档云；借助企业内容管理系统的各种采集和融合能力，将各种业务系统文件和体系文件上传至文档云，从而实现企业所有文档的统一存储、统一管理和统一应用。帮助企业管理采集、存储、保护、治理、使用、交换和归档等过程中的文档以及企业组织流程相关的内容与文档，逐步构建企业非结构化数据内容服务平台。（如图 11 所示）

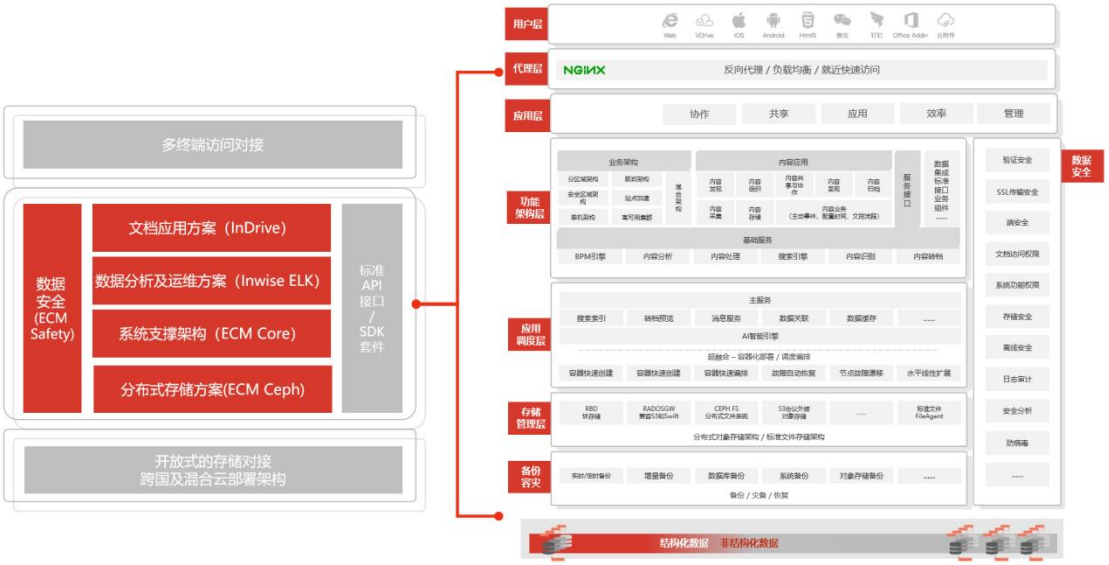


图 11 文档云应用解决方案

5.2.6 信创国产化解决方案

对 IT 基础设施的国产化替代正是当下经济数字化转型、提升产业链发展的关键，而 ECM 能够作为事业单位、党政机关、高等院校、科研院所等组织提供内容管理支撑，成为这些组织的“新基建”，推动相关行业的数字化进程。通过完成国产化硬件的全面适配，ECM 内容管理平台能够解决组织内的数据安全、协作效率、业务流转和管理合规的问题，在保障终端安全的情况下，提升效率、降低成本，以统一平台解决办公文档协作共享、跨平台数据迁移、会务组织、移动办公、文档内部流转、跨网安全交换、业务附件归档等特色场景的需求。（如图 12 所示）

图 12 信创国产化解决方案

5.2.7 某汽车金融企业非结构化数据中台应用案例

应用领域：金融

应用场景：数据中心

案例提供者：某汽车金融企业

（1）案例中存在的问题

某汽车金融企业早年已部署内容管理平台，软件硬件设施老旧，维护成本高、难度大，系统架构复杂，基于原有内容管理平台实施新项目、新系统的成本高，系统稳定性差，严重影响到了用户的日常使用。

（2）解决方案介绍

非结构化数据中台依托于 ECM 内容管理平台的厚实底座，能够为企业提供非结构数据统一存储、统一管理的内容管理平台和内容服务平台。作为传统内容管理平台的升级迭代，通过构建企业非结构数据中台，更新系统架构，为该汽车金融企业提供了方便于前台应用统一输出的各种非结构数据服务，通过门户、建模和表单三大引擎快速构建各种以非结构化内容为主的业务应用，实现企业业务

的内容驱动和精细化运营，从而降低运维成本，以标准化的系统设计，为未来企业的系统更新与迭代提供了更优质的解决方案。（如图 13 所示）

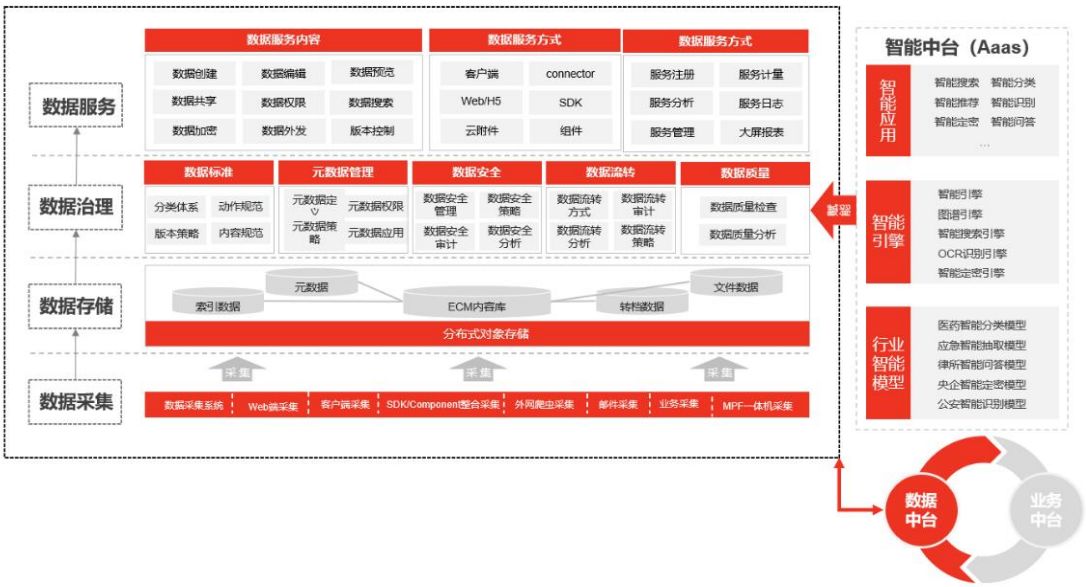


图 13 非结构化数据中台应用解决方案

5.2.8 涉密/商秘电子文档安全管理应用解决方案

商密电子文档安全管理遵循国资委《中央企业商业秘密信息系系统安全技术指引》，完成商密密级文件在形成、流转、存储、脱密以及销毁等阶段的全域全生命周期安全保护。涉密电子文档安全管理遵循国家保密局《涉及国家秘密的电子文档安全保密产品技术要求》标准，以密级识别技术为基础，综合应用电子审批、访问控制等技术手段，对涉密文件的内容和使用权限进行安全控制，防止涉密信息在内部发布与交互过程中被肆意传播和违规使用，最大限度保护涉密文档的使用安全。

该解决方案可以为政府单位、军工企业、科研院所等涉密电子文档的集中存储及安全管控提供应用解决方向。系统遵循分级保护要求，运用身份鉴别、数据加密、细粒度访问控制以及安全审计等多种技术进行全程全域的安全保护，做到事前防范、事中控制和事

后追溯相结合，以确保涉密电子文档全生命周期的安全（如图 14 所示）。

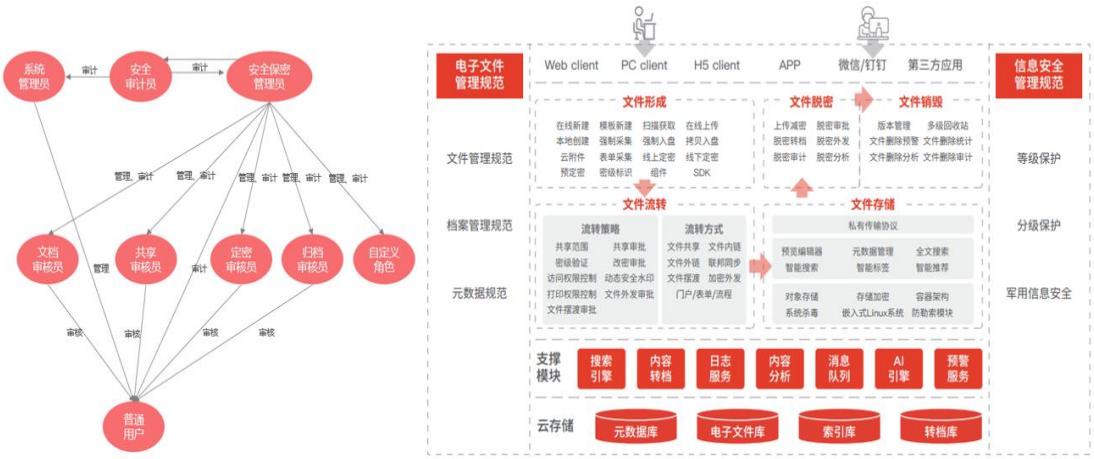


图 14 涉密/商密电子文档安全管理应用解决方案

5.2.9 某技术、产品和解决方案公司内容知识化平台案例

应用领域：公司数据管理

应用场景：数据管理

案例提供者：某技术、产品和解决方案公司

（1）案例中存在的问题

某领域全球最大的技术、产品和解决方案公司，在企业发展过程中，面临诸多问题：一是数据离散存储，公司同时使用多套业务系统，导致重要数据分散在各自系统及员工电脑中，无法做到集中管理；二是历史数据与知识的检索困难，无法对历史数据、成果文档进行全面检索、利用；三是涉密文件管理困难，文档安全无法有效管控，文档越权利用无法做到审批后自动赋权、到期后自动回收权限；四是知识缺少互通，部门间知识和经验缺少统一分享平台，导致协同办公困难。

（2）解决方案介绍

构建企业非结构化数据的统一利用平台，通过标准的数据接口，打通企业内部各个业务系统之间的信息孤岛，将文档及知识在

统一的平台进行存储，遵循统一的数据标准进行管理、应用和发现。在数据中台为业务系统减小负担的同时，实现了企业数据集中存储、安全受控与合规利用。基于微服务架构松耦合的特性，平台底层的所有文档服务也可以反过来支撑业务系统，如元数据、统一预览、统一搜索查询、批量归档和调阅利用等等。

5.2.10 某全球大型制造企业内容归档和合规管理案例

应用领域：公司数据管理

应用场景：公司部门间沟通管理

案例提供者：某全球大型制造企业

（1）案例中存在的问题

作为全球恒温器制造领导厂商，企业目前面临着研发、生产和质量三个部门之间分散、孤立的问题。研发、生产和质量是构成企业管理的三大要素，分别由各个部门负责把控，但这三者之间又是相互关联、密不可分的。当前情况下设计图纸、零部件表单、管理流程等垂直业务内容仍分散、孤立在各部门，亟需实行有效的措施打破各部门间的壁垒，将三个部门有效地串连起来，并进行有效的监督和管理。

（2）解决方案介绍

利用附件归档及统一搜索系统，统一存储，打通信息孤岛，对接各类业务系统数据，同时实现 BMP 业务流程自动化，提升业务内容查找效率及准确性。垂直搜索是目前相对通用搜索引擎的信息量大、查询不准确、深度不够等提出来的新的搜索引擎服务模式，通过实行垂直搜索这一项解决措施，可以使企业部门针对某一特定领域、某一特定人群或某一特定需求提供更高价值的信息和相关服务。

5.2.11 KM 知识管理应用解决方案

KM 知识管理应用解决方案可为用户提供专业的知识管理咨询规划，结合成熟的落地方法论，基于业界领先的 ECM 平台提供知识统一存储平台，并以此构建兼顾稳定性、先进性、实用性、可靠性和可拓展性，面向业务及管理需求的知识管理平台。同时，AI 智能技术也为新技术下的知识管理应用场景提供支撑（如图 15 所示）。

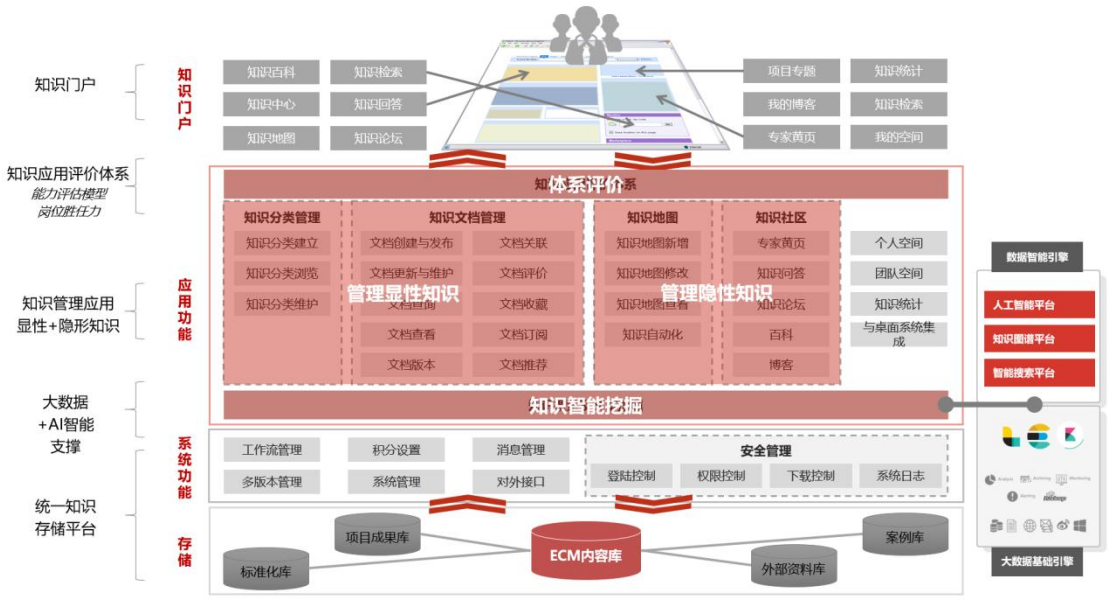


图 15 KM 知识管理应用解决方案

5.2.12 文档档案一体化应用解决方案

文档档案一体化应用模式遵循“文件生命周期理论”，为企业级档案信息化建设提供应用解决方案，提供了基于电子文档全生命周期管理功能，以满足企业非结构化数据的统一采集、存储、保护、管理、使用和交换需求，同时提供了基于档案业务全生命周期的管理功能。

按照企业档案管理规范，将档案收集工作前置，约束各个职能部门在日常文档管理过程中进行预归档整理，定期向档案系统归档，实现文档业务充分融合，同时将档案记忆知识化，为企业再生产活动提供强有力的知识支撑（如图 16 所示）。

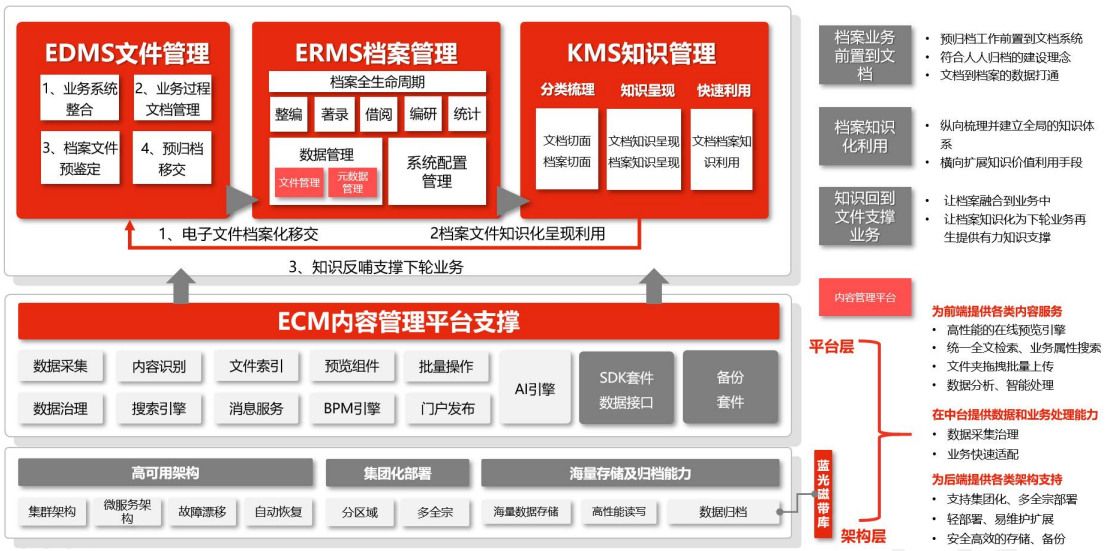


图 16 文档档案一体化应用解决方案

5.2.13 某中级法院文件安全交换应用案例

应用领域：党政机关

应用场景：内外网数据交换

案例提供者：某中级法院

（1）案例中存在的问题

为响应“十三五规划”中“法院信息化 3.0”决策的深化与完善，某中级法院面临着两大核心挑战：需要构建符合等保条件的文档办公环境与内外网安全文件交换系统，法院内部禁止内外网交叉使用移动介质，且需要具备文件备份还原系统与防勒索病毒的手段，全面保障法院内部各类办公资料、文书、音视频、法宣资料的安全。

（2）解决方案介绍

针对某中级法院的需求，采用了容器安全隔离的单套文件安全交换平台。该平台是融合网络隔离技术和网盘技术于一体的专业应用， 文件安全交换具有全方位安全管控的特点，通过授权、审批、审计、查杀毒、敏感性检测和文档标签追踪等方式保障了数据交换全过程的安全性。

基于中院与基层院之间的垂直关系，构建起了中院垂直部署，

全市统一的文件共享交换架构，用户在不同网间进行数据摆渡时，系统会进行严格地交换安全控制，并留有完整日志和摆渡文件以备后续审计追溯，以保障用户在不同网间、不同环境下的数据传输、存储、交换、共享与分享的安全性（如图 17 所示）。

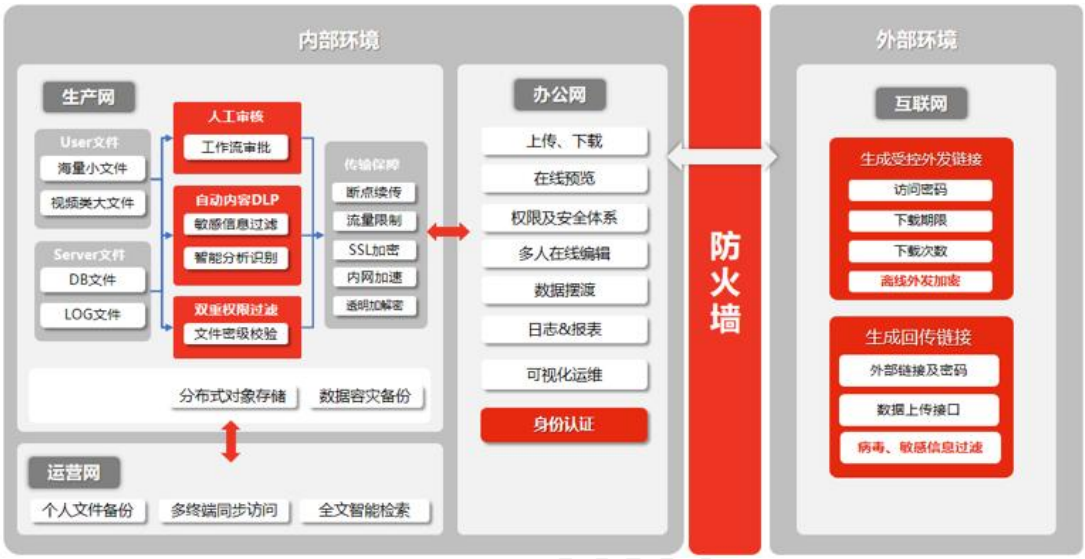


图 17 文件安全交换应用解决方案

同时，配备了细颗粒度的访问权限设置与防勒索+病毒查杀的模块，通过进程识别与日志分析，阻断感染文件上传至服务器，保护文件本地至云端的安全。

5.2.14 某药业集团 GMP 医药质量应用案例

应用领域：医药制造

应用场景：医药行业文件管理

案例提供者：某药业集团

（1）案例中存在的问题

该药业集团是一家跨地区、产学研相结合、科工贸一体化的大型医药企业集团。集团经营过程中的所有文件、数据都需要完整有效地进行保存，随着集团业务量的不断扩大，企业内的文件管理正面临着越来越严峻的挑战。信息量大、文件查阅难；文件审核流程复杂、效率低下，无法实时监控；文件传递缺乏安全管控，数据

完整性难以达到监管要求，缺乏完整的培训体系。

（2）解决方案介绍

基于该药业集团现状，提出了医药质量 GMP 应用解决方案，以高效管理和使用企业运营过程中产生的业务文件、质量文件以及档案等文件，满足 GMP 和 GSP 等医药管理规范，贴近国际和国内 GMP 标准的计算机系统验证服务和安全管理技术和策略，利用医药领域构建集团性文档和档案管理技术，构筑质量文件管理体系、记录管理体系、培训管理体系和集团级业务流程质量管控流程体系（如图 18 所示）。

图 18 医药质量 GMP 应用解决方案

首先，建立统一的内部文件管理中心，以光学字符识别识别技术实现企业内的便捷搜索，降低企业存储成本。其次，以信息化的流程监控，在企业节约时间、人力成本的同时，实现集团的高效协同运作，提升员工能力。

同时，提供文件全生命周期管理解决方案，通过建立全面的文件管理体系，提高企业的协同效应，落实对 GMP 文件的创建、审核、培训、发布签收、执行修改、归档、废止等全流程的有效监控。

该解决方案覆盖了 GMP 文件全生命周期的管理方式，保证了数据的完整准确，更好地稳定产品质量、应对 GMP 的频繁检查，促进企业规范管理。该解决方案也可以通过完整的培训体系管理，实现集团文件及时的上传下达，以及有效监测，实现企业知识和信息资产有效管理和利用。并且建立质量管理要素关系模型，为企业质量管理提供可靠的基础平台，通过安全的质量应用平台及合规管理，实现全面高效的质量信息化管控。

5.2.15 ISO 质量体系文件管理应用解决方案

应用领域：大型制造企业

应用场景：质量体系文件管理

案例提供者：某芯片设计企业

（1）案例中存在的问题

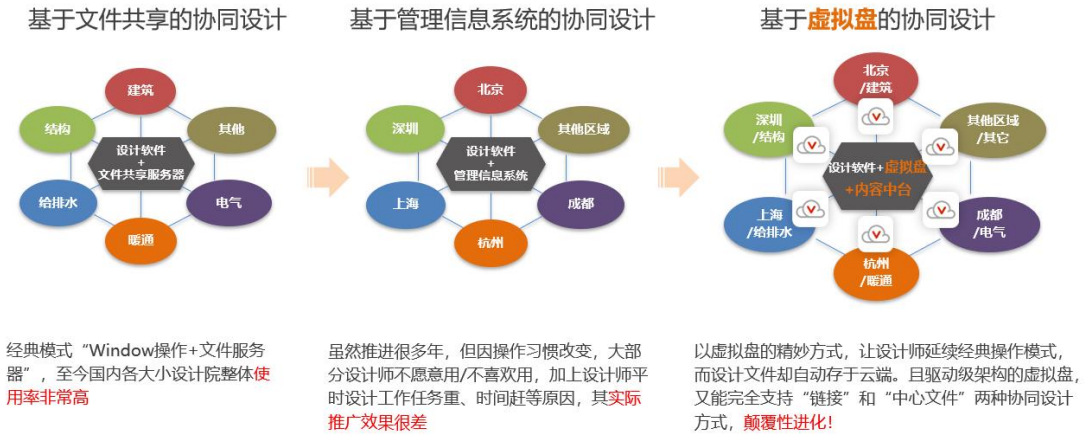
作为高精尖产业中的领军企业，多年以来，某芯片设计企业内累积了大量产品相关的质量手册、程序文件、指导书等文件，但这些文件都分散地存储在企业的各系统与个人电脑之中，文件版本迭代快，缺乏明确的文件、版本、流程管控手段，亟需一套系统应用，遵循国际通行的 ISO 质量体系要求，对所有质量体系文件进行统一管理。

（2）解决方案介绍

建立基于内容管理与流程的 ISO 文档全生命周期管理平台，为该芯片设计企业提供从文件预审、创建与修订、分发、签收与培训、生效，到复审与回收和废止的 ISO 文档全生命周期管理功能。通过对 ISO 质量体系文件的流程化、全生命周期管理，帮助企业明确岗位职责与权限，协调各部门之间的协作关系，实现质量管理体系要求的规范落地建设。

5.2.16 工程协同设计应用解决方案

基于虚拟盘客户端技术，保持设计师现有操作模式，将设计文件自动存于云端。基于驱动级架构的虚拟盘，支持“外部参照”和“中心文件”两种协同设计方式。工程协同设计应用实现基于网盘的日常过程设计协同，校审、提资、收发文等阶段成果流程协同，以及基于 BIM 模型的全专业可视化协同三种协同应用模式（如图 19、20 所示）。



	基于文件共享的协同设计	基于管理信息系统的协同设计	基于虚拟盘的协同设计
底层技术	Windows Server的文件共享服务	基于插件/应用系统的文件共享	驱动级架构虚拟盘实现文件共享，基于内容中台进行统一存储和管理
通过Windows资源管理器访问文件	支持	不支持	支持
文件存放位置	设计文件在服务器上，本地没有缓存，网络要求高	设计文件在服务器上，本地有缓存	设计文件在本地打开，实时同步到云端服务器
增量传输与存储	不支持	不支持	支持
版本管理	不支持	支持	支持
权限管理	受限于Windows系统，权限粗	灵活	灵活，细颗粒度权限
“外部参照”协同设计	支持	支持	支持
“中心文件”协同设计	支持	不支持	支持
局域网协同	支持	支持	支持
广域网协同	不支持	支持	支持
与其他系统整合接口	不支持	支持	支持
其他终端访问	只支持Windows	支持多终端	支持多终端
海量对象存储	不支持	不支持	支持，内容中台支撑

图 19 工程协同设计方式比较

图 20 工程协同设计应用解决方案

5.2.17 EPC 工程内容管理应用解决方案

EPC 工程内容管理是基于 ECM 企业内容管理和内容业务平台构建起的设计协同管理系统，实现了业务流程协同、设计过程及内容协同、设计模型全专业协同。流程协同可实现流程可回查，轻松排查工程、管理问题；提供可定位的关键节点和角色信息；项目进度量化显示，项目细节可管理；即时消息提醒，严格把控任务执行进度。设计过程及内容协同可实现业务流程与文档管理相关联，提升协作效率；文档集中管理，分散使用；系统根据流程运转情况自动分配并同步文档的保密等级；三层病毒防护结构，杜绝感染病毒。设计模型全专业协同可实现全专业设计文档兼容，快速生成模型全貌；全专业在线协同校审，校审成果文档可直接归档并发布（如图 21 所示）。

图 21 EPC 工程内容管理应用解决方案

5.2.18 海量数据快速发现解决方案

基于 ECM 对企业内文档数据的统一管理，能够为企业提供丰富的文件查找定位方式。通过基础搜索、智能搜索和高级搜索三者的有机组合，能够满足企业常见基础搜索需求，并且可以根据企业内具体的场景，提供个性化的检索手段。

基础搜索能够借助分词引擎，提供常规的精准搜索、模糊搜索，针对用户输入内容进行匹配，还能够对文件夹名、文件名、标签进行联想，预测用户意图，提供联想推荐；记录用户搜索历史，提供历史搜索词条；记录企业内热搜词汇，展示组织内的搜索倾向与企业热门信息；另外，还提供切面筛选与搜索排序，帮助用户梳理搜索结果。

通过与 AI 智能深度结合，ECM 能够提供以文搜文、以图搜图与知识图谱的智能检索能力。利用 NLP 语义理解技术，能够快速检索出与当前文档内容语义相同或相似的文件，扩展关键词检索的能力；借助对图像特征的识别能力，能够快速检索与上传图片相似的图片素材，打破只能进行文字搜索的局限；而通过知识图谱技术，

能够理解用户的搜索意图，从而为用户匹配最符合其需求的文件，助力用户完成对知识的挖掘。

高级搜索则可以通过更多检索条件的组合，帮助用户更精确的定位到想要的文件，从而提升用户文件检索的效率。可选择的条件包括：内容、文件类型、创建时间、所在文件夹、创建人等，支持自定义字段与复数条件的组合规则，给用户提供个性化的搜索体验。

5.2.19 文档智能应用解决方案

通过与 AI 智能技术深度融合，ECM 平台能够具备一定的智能能力。其中与文档、内容管理强相关的包括：以图搜图、以文搜文、人像识别、智能搜索、DLP、OCR 整体识别、OCR 区域识别、文档/图像智能标签。除了搜索相关的几项能力之外，人像识别能力能够通过分析人员照片，快速筛选定位与该人员相关的其他图片；DLP（Data leakage prevention）数据泄密防护技术则基于机器学习模型及人工录入规则引擎对捕获到的数据进行风险分析，从而提供文件定密、文件分类、敏感内容检测等安全功能的能力；

而借助 OCR 光学识别技术，能够实现整体识别、区域识别和智能标签的功能。将图像类（含扫描图像形成的 PDF）文件的文字进行提取，生成索引，辅助全文检索，同时可以依据用户配置将识别出的文本结果存放至指定位置；辅以 NLP 技术，能够对图像类文件进行关键内容提取，自动化赋值元数据，可以被应用在合同的关键要素抽取等场景之中；结合知识网络，针对含有文本信息的文件自动提取出更加符合业务逻辑的内容标签；借助图像识别，对上传至系统的图像自动生成符合其特征的标签。

6. 结束语

本白皮书围绕非结构化数据管理展开了体系化、整体性的系统阐述，涵盖了从非结构化数据管理体系、解决方案与应用实践的方方面面。非结构化数据作为占据每年数据增量 80%的一种重要数据类型，值得相关机构与组织进行持续地跟踪调查与深入地研究分析，而对非结构化数据实行科学的管理，需要基于其以下的三个特征：

内容完整：非结构化数据具备很强的描述性，一份文档、一段录音、一部视频往往可以对事件或者人物进行完整的描述，这是结构化数据所不具备的能力；

体量巨大：企业级的非结构化数据大多都是 PB、EB 量级，文件数量超过“亿”、“十亿”量级，大量的非结构化数据占用了企业的存储空间，扩容的需求同时也造成了投资成本的不断攀升；

形式多样：非结构化数据的呈现是多样化、多种类的，也是人们日常最容易接触到的，无论是图文类文件、音视频等多媒体文件，还是图纸报告等专有格式文件，都是非结构化数据的一种，蕴含着日常人类行为活动的规律与轨迹；

因此，专业的非结构化数据管理是每个企业的数字化基础设施，利用好非结构化数据更是每个企业数字化转型的必经之路。非结构化数据在人的活动中产生，又服务于人、贴近于人，是始终围绕着人这一主体进行流转的数据。只有充分挖掘、利用非结构化数据的价值，才能够更好地适应网络和信息化时代，更好地驱动企业革新，释放企业生产力，获得响应时代、弄潮时代的本领。