数字保存的持续完整性风险检测*

□臧国全　朱晓庆　李哲　金燕

摘要　针对数字保存风险之一的持续完整性风险设计检测方法，并针对检测方法进行实验研究。(1)调研相关文献，找出研究的切入点；(2)界定持续完整性的含义，析出产生持续完整性风险的因素；(3)设计持续完整性风险型元数据，设置该类风险的检测点；(4)基于一个实际保存系统，利用分层随机抽样法，采集数字对象及其持续完整性风险型元数据内容的实验样本；(5)编制代码，检测数字对象样本集的持续完整性风险点，统计检测结果，分析可能的产生原因，制定可能的降低风险措施；(6)基于实验结果，分析检测方法的局限性，说明检测方法的使用事项。

关键词　数字保存　持续完整性风险　风险检测
分类号　G250
DOI　10.16603/j.issn1002－1027.2018.02.010

1　文献综述

1.1　风险识别方面

数字保存的风险管理研究已有 20 余年。康威(Conway)[1]是该领域的较早研究者，在其《数字世界的保存》中将数字保存活动识别为风险管理过程。之后，相关研究可归为三类：

专用型风险模型。主要有数字对象文件格式的风险[2][3][4][5]、保存介质的风险[6]、特定类型数字资源(如 Web 数字资源)的保存风险[7]等。这类模型适合于相应领域的风险识别，尽管它们具有一定的互补性，但无法替代综合型风险模型。

综合型风险模型。结构上有等级式风险模型[8]、同位列表式风险模型[9][10][11][12]、网状式风险模型[13]。另外，有些综合型风险模型也描绘了风险、数字对象、保存环境之间的关系[14]。《成功的数字保存威胁识别：用于风险评估的 SPOT 模型》[15]一文从数字保存核心职责角度识别保存风险。由于这类模型的应用环境和目的不同，导致它们在风险的种类、数量及模型展现形式等方面存在差异。

数字保存风险的实证研究。如基于罗森塔尔(Rosenthal)模型并进行适当改造，对美国国会图书馆数字保存的风险检查[16]；基于莱特(Wright)模型的对大英图书馆数字保存介质的风险评估[17]。这类研究是对已有模型的实证分析，有助于数字保存项目选择合适的风险评估模型。

上述风险模型的优缺点分析见表 1。基于表 1 的分析，已有的风险模型都存在不同程度的缺憾，还没见到完全满足表 1 列出指标的模型的报道。

表 1　已有风险模型的比较分析

模型作者	概念清晰性【注 1】	风险专指度的合适性与一致性【注 2】	全面性【注 3】	使用方式【注 4】
劳伦斯(Lawrence)，阿姆斯(Arms)，斯特内斯库(Stanescu)，罗格(Rog)等 4 个模型	好：前 2 个模型基于原因，后 2 个模型基于结果	好：四个模型中列出的文件格式风险在专指度方面比较合适且一致性较好	差：仅列出文件格式风险	仅用于文件格式风险的定性评估
莱特(Wright)	一般：基于原因，但技术风险有重叠	好：列出的保存介质风险在专指度方面比较合适且一致性较好	差：仅列出保存介质风险	仅用于保存介质风险的定性评估

* 国家自然科学基金面上项目“数字保存的风险型元数据与风险监控研究”(批准号 71673255)的研究成果之一。
通讯作者：金燕，ORCID：0000-0001-8566-3097，zhwang@zzu.edu.cn。

模型作者	概念清晰性【注 1】	风险专指度的合适性与一致性【注 2】	全面性【注 3】	使用方式【注 4】
麦格文(McGovern)	差:混合型	一般:总体上专指度合适,但与其他风险相比,产权风险明显过细	差:仅针对 Web 数字资源的保存风险	仅用于 Web 数字资源保存风险的定性评估
巴拉泰罗(Barateiro)	好:基于原因	差:产权风险和经济风险无细分,太宏观。技术风险较微观。缺少针对满足用户需求方面的风险。	好:包含所有类型风险	用于全面风险定性评估(包括保存系统和数字对象)
克利夫顿(Clifton)	好:基于原因	差:针对数字对象的风险划分过细(如保存技术的风险),但针对保存系统的风险定义过于宏观(如组织机构、经济方面)	一般:缺少产权风险	用于全面风险定性评估(包括保存系统和数字对象)
罗森塔尔(Rosenthal)	好:基于原因	差:产权风险和经济风险较宏观,技术风险较微观	好:包含所有类型风险	用于全面风险定性评估(包括保存系统和数字对象)
托马兹(Thomaz)	差:混合型	差:列出的风险较宏观,专指度欠佳	差:缺少产权风险、经济风险、用户可理解性风险	用于全面风险定性评估(包括保存系统和数字对象)
视点解析(PARSE.Insight)	差:混合型	好:列出的所有风险在专指度方面都具可比性	好:包含所有类型风险	用于全面风险定性评估(包括保存系统和数字对象)
数字管护中心(Digital Curation Center)	差:混合型	差:针对数字对象的风险划分过细(如保存技术的风险),但针对保存系统的风险定义过于宏观(如产权、组织机构、经济方面)	好:包含所有类型风险	用于全面风险定性评估(包括保存系统和数字对象)
戴波特(Dappert)	好:基于原因	一般:将风险归为 8 种类型,在每种类型下列出具体风险,风险类型之间的专指度具有可比性,但在一些类型下的具体风险之间,专指度可比性较差。	好:包含所有类型风险	用于全面风险定性评估(包括保存系统和数字对象)
萨利(Sally)	一般:基于结果,但风险分类清晰度较低	一般:列出的大多数风险的专指度具可比性,但可用性风险较宏观。	好:包含所有类型风险	用于全面风险定性评估(包括保存系统和数字对象)

【注 1】:指应按照一种方法识别风险,避免歧义和重复。识别方法有两种:一是基于风险发生的原因,如存储介质退化;二是基于风险发生的结果,如二进制数据流序列被破坏。混合式风险列举方法将影响概念的清晰性。

【注 2】:指列举的风险在概念外延上适中。外延太大的风险较难测度,也较难识别产生风险的具体原因;外延太小的风险可能会导致重复检测;概念模糊的风险会导致检测结果的误差。另外,在任何结构的模型(等级式、同位列表式、网状式)中,同位风险的概念外延应大致相当,且外延之和应与上位风险大致吻合。

【注 3】:应列出模型界定范围内的所有主要风险。比如,一个基于原因的针对数字对象的风险识别模型,若无存储介质风险,则保存系统无法识别和管理这种风险。

【注 4】:使用方式有两种:定性评估和定量检测。在所列模型的使用说明中,均表示可用于定性评估,但若用于定量检测,需针对每个风险点设置检测项目。

1.2 相关标准

ISO14721:2003[18]。《OAIS 参考模型》(Open Archive Information System),制定了数字保存系统的框架结构和概念规范。起源于国际空间数据系统咨询委员会(Consultative Committee for Space Data Systems),目的在于维护数字对象的长期有效存取。

ISO 16363:2013[19]。《可信任数字保存的审计与认证》,制定了保存系统的质量标准。其中的《可信任数字保存审查表》及其认证程序间接地展示了数字保存的风险。

ISO 16919:2014[20]。《可信任数字保存的审计和认证机构要求》,制定了审计和认证的程序,以及对认证机构的基本要求。

由上可知,相关标准都不是真正的数字保存风险列表。ISO14721 提供一个数字保存功能框架模型,目的是保证数字对象在长期保存过程中规避可能的保存风险,但不是一个风险列表。ISO 16363

的《可信任数字保存审查表》实际上是数字保存的质量评价指标体系，虽然本质上每个指标隐含一种或多种风险，但并不是风险列表。ISO 16919 是对数字保存质量认证机构的要求，也不是风险列表。

1.3 本研究的切入点

从风险发生的角度。数字保存的风险有两个范畴：数字对象风险、保存系统的风险。前者有数字对象的获取、存储、维护和传播等方面的风险，后者有保存系统经济方面、产权管理方面的风险。已有的标准和风险识别模型（除专用型）都包括上述两个范畴。本研究限定在第一个范畴，即数字对象产生的风险，当然第二个范畴的风险也会影响数字对象的风险，但这种影响是间接的，尤其是针对本研究的持续完整性风险。另外，一些保存活动也会导致数字对象产生风险，但这类风险常需要依据保存政策来判断。所以，本研究的风险检测点以风险型元数据形式呈现，包括数字对象方面的、保存事件方面的和保存政策方面的三种。

从风险类型的角度。数字对象的风险有多种，如持续完整性风险、可用性风险、可呈现性风险、真实性风险、可识别性风险、可理解性风险等。已有的标准和风险识别模型都囊括了数字对象的所有类型风险。但作为一篇学术论文，本研究仅限定在持续完整性风险，其他类型的风险后续研究。

从风险识别方法的角度。上述模型中的风险识别方法有三种：基于风险发生的原因、基于风险产生的结果、同时包括这两种方式的混合型识别法。本研究首先基于全面风险管理理论划分风险的范畴，然后针对每个范畴的风险，基于风险发生的原因，识别出风险点。

从风险评估的角度。已有的评估方法都是定性的，本研究的评估方法是定量的。为此，本研究对每个风险点设计检测项目，编制代码进行定量检测，统计并分析检测结果。

2 持续完整性及其风险检测思路

2.1 持续完整性及其风险

持续完整性指构成数字对象的比特流持续存在且没有被破坏，处于可使用、可操作状态，并可从保存介质中完整检索出来以实施浏览等操作。因此，确保数字对象比特流没有发生任何形式的改变，并能从保存介质中被完整阅读，是实现数字对象持续完整性的两个必要条件。

持续完整性风险指保存系统中妨碍实现数字对象持续完整性的各种因素发生的可能性。这些因素包括：(1)数字对象的不适宜存储，如保存条件不足导致无法实现所需的保存级别，致使长期保存过程中数字对象比特流可能被破坏且无法恢复，出现难以被操作使用的情况；(2)存储介质超出有效期，导致介质自然退化，致使存储的数字对象比特流序列可能被破坏，出现无法被完整检索、浏览的情况；(3)存储介质被破坏，或病毒导致，或操作人员失误导致，致使保存的数字对象比特流不再持续完整；(4)用于判断数字对象持续完整性的信息没有被记录，如信息摘要、密钥信息等，导致长期保存过程中无法验证数字对象是否被破坏，致使其持续完整性可能出现风险；(5)保存系统没有按照保存政策的要求实施必要的保存活动，如存储介质刷新、固定性检查、病毒检查等，导致数字对象的持续完整性亦可能出现风险。

总之，数字对象持续完整性风险主要存在于存储介质的管理、保存系统的保存能力、保存事件的实施、数字对象相关信息的记录、数据安全方针的制定等方面。

2.2 检测思路

本文设计的检测思路是：(1)界定持续完整性的含义，由此析出产生持续完整性风险的因素；(2)基于风险产生因素，设计持续完整性风险型元数据，由此实现该类风险检测点的设置，并设置每个风险点的检测项目；(3)基于一个实际保存系统，利用分层随机抽样法，采集数字对象及其持续完整性风险型元数据内容的实验样本；(4)编制代码，检测数字对象样本集的持续完整性风险点的各个检测项目，统计检测结果，分析可能的产生原因，制定可能的降低风险措施。

3 持续完整性风险型元数据

根据全面风险管理理论，企业风险产生于企业整个运营过程，不仅来自生产经营的对象，还来自生产经营的活动以及相关政策。针对数字保存，“生产经营对象”是数字对象，“生产经营活动”是保存事件，“相关政策”是保存政策。因此，可从数字对象、保存事件、保存政策等角度来分析数字保存风险的产生因素，设置风险型元数据。

3.1 数字对象方面的持续完整性风险型元数据

数字对象是保存系统存储和用户访问的独立知识单元。有四种：一是知识实体，是描述一项特定知识所需的内容集合，如一本书、一幅地图、一张照片、一个数据库等；二是表现，是将一个知识实体实例化的一个数字化对象，一般由多个数字化文件及结构化元数据组成，用于知识实体的展现，一个知识实体可以有多个表现；三是文件，是可以被操作系统识别的一组有序的字节；四是比特流，是文件内连续或非连续的数据。针对持续完整性，只需检测文件和比特流，因为其他两类数字对象均由多个文件或比特流组成，若其中一个文件或比特流的持续完整性出现风险，对应的知识实体或表现的持续完整性自动出现风险。

数字对象方面的持续完整性风险型元数据是用于描述与持续完整性相关的数字对象属性，是持续完整性风险的检测点。这类元数据的元素有：

(1)数字对象标识符(Object Identifier)。数字对象被赋予的唯一标识符，以供检索和发现，亦方便参考和引用。该元素内容可由保存系统收录数字对象时创建，也可由生产者创建并与数字对象一起提交给保存系统。赋值方式有保存系统自动生成和人工赋值两种。该风险点的作用为：该元素内容缺失导致无法识别对应数字对象，也就无法进行后续风险点的检测。

(2)数字对象类型(Object Category)。用于描述数字对象的类型(知识实体、表现、文件、比特流)。该风险点的作用为：筛选用于检测的文件和比特流对象；该元素内容缺失导致无法判断数字对象是否适合持续完整性风险的检测。

(3)固定性信息(Fixity Information)。描述数字对象在长期保存过程中是否被改变的验证所需信息。固定性检查需要计算数字对象的信息摘要，并与系统收录时产生的信息摘要对比，如果两个摘要相同，则该数字对象在保存过程中没有改变，否则说明发生了改变。因此，固定性检查是一个保存事件，记录该保存活动的实施时间和检查结果。但作为数字对象的一个属性，固定性信息的描述项有：(a)信息摘要算法，如消息摘要算法第五版(Message Digest Algorithm，MD5)、可变长度的哈希算法(Hashing Algorithm with Variable Length，HAVAL)、安全散列算法(Secure Hash Algorithm，SHA－256)等；(b)信息摘要，信息摘要算法运行的结果。固定性信息的赋值可由数字对象提交者产生，但需保存系统验证，否则需由保存系统在收录数字对象时产生。

该风险点的检测项目有：(a)若信息摘要算法内容为空，则无法计算新的信息摘要，无法判断数字对象是否改变；(b)若信息摘要内容为空，则基于原始算法计算出的新信息摘要缺失对比的基准值，也无法判断数字对象是否改变；(c)基于原始算法计算出的新信息摘要与原始信息摘要比较，若不同，数字对象发生改变。上述三种情况均归为在该风险点上产生风险。

(4)签名信息(Signature Information)。常用于信息传输过程中接收者确认信息来源的真实性。在数字保存中，可借用来判断数字对象在长期保存过程中是否改变。基于数字签名的持续完整性验证方法为：(a)数字签名值的生成，保存系统收录数字对象时，采用一种哈希算法生成信息摘要，再使用保存系统私钥对信息摘要进行加密生成签名值；(b)持续完整性验证，采用相同哈希算法生成数字对象的新信息摘要，使用保存系统的公钥对数字签名值解密获取原始信息摘要，对比两个信息摘要，若不同，则数字对象发生改变。

签名信息的描述项有：(a)签名者，若数字对象提交时已有签名值，则签名者为提交者，否则保存系统需生成签名值，签名者为保存系统；(b)签名方法，生成签名值所使用的加密方法和哈希算法，如数字签名－安全散列算法(Digital Signature Algorithm-Secure Hash Algorithm，DSA-SHA1)，前者为加密方法，后者是哈希算法；(c)信息摘要，基于签名方法中哈希算法生成数字对象的摘要；(d)签名值，使用私钥对信息摘要加密生成的值；(e)密钥信息，验证数字签名所需的签名者公钥信息。

该风险点的检测项目有：(a)若签名方法的内容为空，则无法计算新的信息摘要，导致无法判断数字对象在保存过程中是否改变；(b)若密钥信息或签名值的内容为空，无法还原原始信息摘要，导致新信息摘要缺失对比的基准值；(c)基于签名方法计算出的新信息摘要与基于密钥信息和签名值还原的原始信息摘要比较，若不同，则数字对象发生改变。上述三种情况均归为在该风险点上产生风险。

与固定性信息相比，数字签名增加了信息摘要

的加密和对加密的信息摘要进行解密的过程，这种方法虽较复杂，但更准确，消除了同时恶意修改原始信息摘要和数字对象内容使基于固定性判断结果数字对象没有被改变的可能性。

(5)文件大小(Size)。数字对象的字节数量。若保存系统采用一个计量单位(如G,M,K)，该元素只需记录数字对象大小的值，无需记录计量单位。

该风险点的检测项目有：(a)将数字对象文件大小的检测值与该元素的描述值比较，若不相等，数字对象发生变化；(b)若该元素内容为空，数字对象大小的检测值缺失对比的基准值，无法判断数字对象是否改变。上述两种情况均归为在该风险点上产生风险。

另外，如果数字对象的检测值与该元素的描述值相等，也不能确保数字对象没有发生改变，但为简便起见，本文作为无风险处理。因此，该风险点的检测结果具有一定误差，遗漏了虽数字对象大小没有改变但内容已变化的情况。

(6)保存级别[21](Preservation Level)。描述针对一个数字对象实施相应保存功能的保存决策信息，以及实施这些保存功能所需的保存环境信息。

保存系统可以根据数字对象的特征(如数字对象的价值和唯一性、格式的可保存性、法律法规的要求等)提供多个保存级别。保存级别的描述项有：(a)保存级别类型，描述选择的保存级别期望对数字对象实施保存功能的类型，如“基于字节安全的保存”(即“比特保存”)。(b)保存级别值，描述对应类型的保存级别期望实施的保存功能，如“比特保存”级别类型的保存功能可为：“低”(无备份)、“中”(异地一个备份，不定期实施完整性检测)或“高”(异地三个备份，定期实施完整性检测，备份之间高度独立)。(c)保存系统的胜任状态，描述保存系统能否实现保存级别值定义的保存功能，比如“有能力”(指能够实现且已实现)，“需要”(指期望实现，但现在无法实现)。(d)保存级别的赋值原因，当数字对象的保存级别值与常规不同时，需描述其原因，如根据法律规定或合同约定，对一个数字对象的保存级别的赋值要高于同类型的其他对象时，该元素的值是“法律需求”或“合同约定”。(e)保存级别指定日期，随着时间变化，需对数字对象的保存级别进行评估和修改，以适应保存系统的保存需求、策略或能力的变化。

该风险点的检测项目有：(a)检查“保存级别值”与实际实施的保存功能的相符性，如一个数字对象的保存级别值为上例的“中”，但数字对象在“存储位置”元素中的描述仅有一个位置(即无备份)，表明数字对象在遭到破坏情况下期望恢复，但实际上无法实现恢复，则判定该数字对象在该风险点存在风险；(b)检查“保存系统的胜任状态”，若为“需要”，表明保存系统目前无法实现确保数字对象持续完整性所需的保存功能，则判定该数字对象在该风险点存在风险。该元素的检测结果属于间接相关风险，即可能产生风险。

(7)存储位置(Content Location)。存储系统为数字对象分配的存储定位，通常情况下，通过程序分配。存储位置的描述项有：(a)存储位置类型，如物理存储、URI、绝对路径、相对路径；(b)存储位置值，存储系统使用的用于描述数字对象存储位置的具体值，可以是一个完整的绝对路径，也可以是解析系统中与物理路径相对应的信息，还可以是存储系统使用的相对路径信息。根据保存级别，若数字对象存在多个备份，存储位置也应有多个，可采用重复该元素的方式分别描述。

该风险点的检测项目有：(a)若该元素内容为空，即使数字对象的唯一标识符存在，也无法获取具体的数字对象，故也无法对数字对象实施相应检测；(b)比较存储位置的描述个数与保存级别中要求的数字对象备份数量是否相符，若不同，则可判定该数字对象在该风险点存在风险。由于持续完整性与存储的具体位置无关，因此在对数字对象进行持续完整性风险检测时，只需判断其是否有存储位置以及存储位置的个数，无需检查其具体的位置。上述两种情况均归为在该风险点上产生风险。

(8)存储介质[22](Storage Medium)。描述数字对象所存储的物理介质(如磁带、硬盘、CD-ROM、DVD等)。若数字对象有多个备份，存在多个存储介质，可采用重复该元素的方式分别描述。

该风险点的检测项目有：(a)基于保存政策中保存介质的使用寿命，判断数字对象的保存介质是否过期，有多个存储介质时应分别判断，若过期，保存的数字对象可能因为介质自然退化而遭到损坏。(b)判断该元素的描述值是否为空，若为空，无法识别数字对象的存储介质，导致无法知晓存储介质的状况，难以判断保存的数字对象是否遭到破坏。(c)

基于该元素的描述值，寻找保存政策中设置的相应存储介质的刷新周期，判断保存事件“介质刷新”的执行是否符合保存政策的要求，若不相符，保存的数字对象可能因为介质损伤没有得到及时发现和修补而遭到破坏。上述三种情况均归为在该风险点上产生风险。

3.2 保存事件方面的持续完整性风险型元数据

用于描述对数字对象实施保存操作的信息有两种类型，一是执行结果产生新数字对象的事件，如数字迁移；二是执行结果不产生新数字对象的事件，如固定性检查。由于持续完整性风险仅需对数字对象（包括相关保存环境）的检查，所以这类风险检测仅限在第二类事件。这类元数据的元素有：

（1）固定性检查（Fixity Check）。根据保存政策对数字对象进行固定性检查。如果没有执行该事件或虽执行但不符合保存政策要求，该风险点产生风险。

（2）信息摘要计算（Message Digest Calculation）。保存系统通过计算获得数字对象的原始信息摘要（若数字对象提交者提供原始信息摘要，保存系统需计算予以验证）。如果没有执行该事件，原始信息摘要缺失，无法执行固定性检测事件，也无法进行固定性信息、数字签名信息风险点的检测，该风险点产生风险。

（3）保存介质刷新（Storage Medium Refresh）。根据保存政策对数字对象保存的介质进行刷新。如果没有执行该事件或虽执行但不符合保存政策要求，该风险点产生风险。

（4）病毒检测（Virus Check）。根据保存政策进行病毒检测。如果没有执行该事件或虽执行但不符合保存政策要求，该风险点产生风险。

3.3 保存政策方面的持续完整性风险型元数据

保存政策主要是数字保存操作的指标设置。分为两类：一是保存系统对保存事件实施规则的描述信息，比如保存介质刷新的频率、固定性检测周期、病毒检测周期等。二是保存系统对数字对象质量判断的指标描述信息，比如数字对象的容错率、数据丢失的允许率、内容失真的允许率、数字迁移的准确率等。与持续完整性相关的保存政策仅限在第一种类型，另外，判断数字对象的存储介质是否过期，需要参考存储介质的使用寿命。因此，这类元数据的元素有：

（1）存储介质的使用寿命（Media Life）。用于存储介质风险点的检测。

（2）保存介质刷新频率（Media Refresh Rate）。用于保存事件“保存介质刷新”风险点的检测。

（3）固定性检测周期（Fixty Check Period）。用于保存事件“固定性检测”风险点的检测。

（4）病毒检测周期（Virus Check Period）。用于保存事件“病毒检测”风险点的检测。

4 检测实验

4.1 数据采集

数字对象样本来源于中国知网（CNKI），样本采集量1万件。

数字对象的样本采集。（1）层次单元划分。基于CNKI数字对象的时间区间、文献类型、学科类型三个属性，将其划分为504个层次单元，即：7（时间区间数）×9（文献类型数）×8（学科类型数）＝504。其中，时间区间：1990年之前、1991－1995年、1996－2000年、2001－2005年、2005－2010年、2011－2015年、2016年之后；文献类型：期刊、硕博论文、会议论文、年鉴、统计数据、专利、标准文献、古籍、工具书；学科类型采用CNKI大类划分：基础学科、工程技术、农业科技、医疗卫生科技、哲学与人文科学、社会科学、信息科学、经济与管理科学。（2）各层次单元样本量计算。计算各层次单元的数字对象数量与《中国知网》数字对象总数量的比例，乘以1万（设定的样本总量），获得各层次单元的抽样数量，这样，各层次单元数字对象以接近的概率被抽样。（3）样本集形成。基于无重复抽样的简单随机抽取法，从各层次单元中抽取样本，通过套录形成有代表性的数字对象样本集。如依据上述方法计算出层次单元为“1995－2000年时间区间工程技术的专利文献”的样本抽取量为100篇，通过对《中国知网》检索，该层次单元共有512833件，在1到512833之间随机生成100个不重复的数，套录该100个数对应的检索结果数字对象，获得该层次单元的样本。

元数据元素的赋值。纯粹用于科研目的，CNKI帮助提供上述样本对象的保存型元数据、管理型元数据和描述型元数据的内容。针对本文制定的持续完整性风险型元数据的每个元素，若出现在上述任一种元数据中，则该元素内容直接套录，否则，该元素内容置空。上述过程由代码实现，但需人工干预，

比如对名称与 CNKI 不同但含义相同的元素的赋值需人工甄别和转换。

4.2 代码编制

代码功能:针对不同维度的风险检测(见 4.3.1,4.3.2,4.3.3,4.3.4),检查、统计并以可视化形式展现相应层次单元中数字对象在每个持续完整性风险型元数据元素的风险点上产生风险的概率。代码设计过程中涉及下述问题:

(1)数字对象的选择。在样本集中,删除标识符内容为空的数字对象,因为这类数字对象无法识别。删除类型不为“文件”和“比特流”的数字对象。

(2)元数据元素的选择。数字对象类型、数字对象标识符、所有保存政策的元数据元素等三类元素无需检测。因为第一类元素用于筛选文件对象,第二类元素用于数字对象识别,第三类元素为保存事件元数据检测提供参考基准值。

(3)元数据元素赋值内容的编码。这是实现自动检测的基础。另外,需对表达不同但含义一样的赋值内容归并,赋予一个编码,以提高检测的准确性。

4.3 风险检测

4.3.1 单维度风险检测

检测并统计整个数字对象样本集在持续完整性风险型元数据每一个元素的风险点上产生风险的概率。结果见图 1。风险概率较高的风险点依次为:签名信息、固定性信息、固定性检测事件、信息摘要计算事件、保存级别。

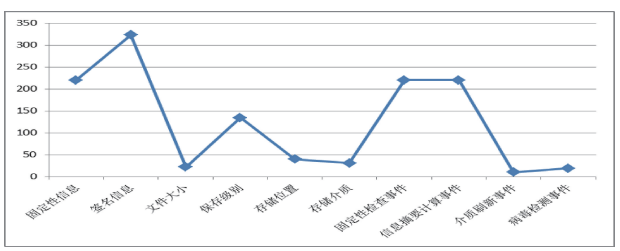

图 1 单维度风险监测结果

4.3.2 二维度风险检测

基于数字对象的一个属性,加上持续完整性风险型元数据的元素,建立一个二维空间坐标系,检测并统计每个坐标点上的数字对象集合在该坐标点上持续完整性风险型元数据元素的风险点上产生风险的概率。有下述三种类型:

(1){时间区间,风险型元数据元素}二维风险检测。检测并统计由持续完整性风险型元数据元素、数字对象的时间区间属性所建立的二维空间坐标系中,每个坐标点上的时间区间所覆盖的数字对象集合,在该坐标点的持续完整性风险型元数据元素风险点上产生风险的概率。结果见图 2 的左图。主要风险点有:(a)签名信息,主要分布在 3 个层次单元:2000 年之前的 3 个时间区间文献。(b)固定性信息,主要分布在 2 个层次单元:1995 年之前的 2 个时间区间文献。(c)固定性检查事件,主要分布在 2 个层次单元:1995 年之前的 2 个时间区间文献。(d)信息摘要计算事件,主要分布在 2 个层次单元:1995 年之前的 2 个时间区间文献。(e)保存级别,主要分布在 7 个层次单元:所有 7 个时间区间文献。

(2){文献类型,风险型元数据元素}二维风险检测。检测并统计由持续完整性风险型元数据元素、数字对象的文献类型属性所建立的二维空间坐标系中,每个坐标点上的文献类型所覆盖的数字对象集合,在该坐标点的持续完整性风险型元数据元素风险点上产生风险的概率。结果见图 2 的中间图。主要风险点有:(a)签名信息,主要分布在 2 个层次单元:期刊文献、会议文献。(b)固定性信息,主要分布在 1 个层次单元:期刊文献。(c)固定性检查事件,主要分布在 1 个层次单元:期刊文献。(d)信息摘要计算事件,主要分布在 1 个层次单元:期刊文献。(e)保存级别,主要分布在 1 个层次单元:专利文献。

(3){学科类型,风险型元数据元素}二维风险检测。检测并统计由持续完整性风险型元数据元素、数字对象的学科类型属性所建立的二维空间坐标系中,每个坐标点上的学科类型所覆盖的数字对象集合,在该坐标点的持续完整性风险型元数据元素风险点上产生风险的概率。结果见图 2 的右图。主要风险点有:(a)签名信息,分布在 8 个层次单元:所有的 8 个学科文献。(b)固定性信息,分布在 8 个层次单元:所有的 8 个学科文献。(c)固定性检查事件,分布在 8 个层次单元:所有的 8 个学科文献。(d)信息摘要计算事件,分布在 8 个层次单元:所有的 8 个学科文献。(e)保存级别,主要分布在 5 个层次单元:5 个学科文献(基础学科、工程技术学科、农业科技、医疗卫生科技、信息科学)。

4.3.3 三维度风险检测

基于数字对象的两个属性,加上持续完整性风险型元数据的元素,建立一个三维空间坐标系,检测并统计每个坐标点上的数字对象集合在该坐标点上

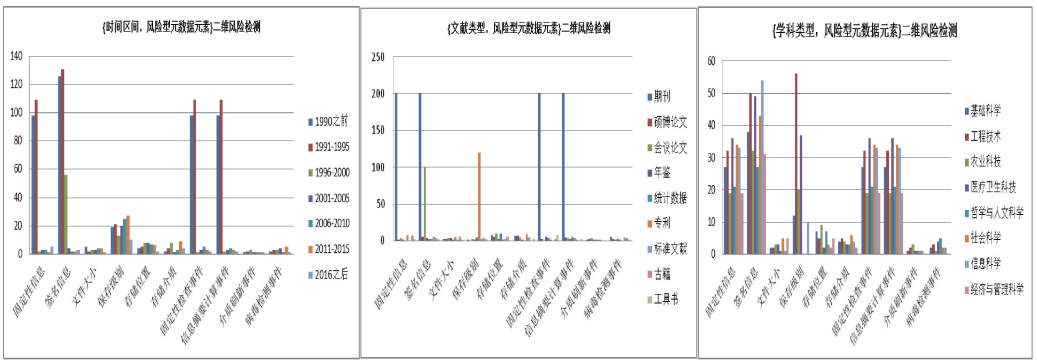

图 2　二维度风险监测结果

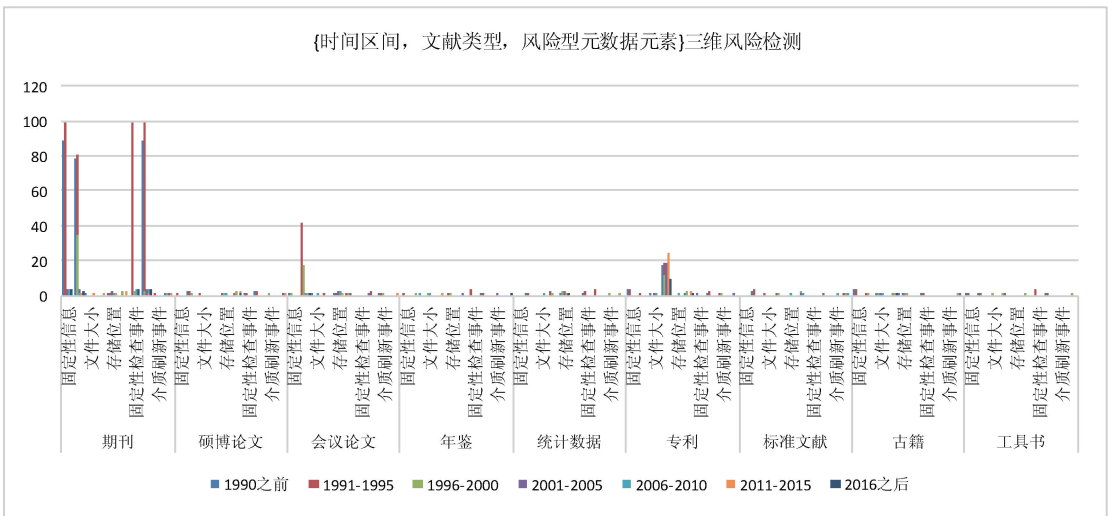

图 3　{时间区间，文献类型，风险型元数据元素}三维风险检测结果

持续完整性风险型元数据元素风险点上产生风险的概率。有下述三种类型：

(1){时间区间，文献类型，风险型元数据元素}三维风险检测。检测并统计由持续完整性风险型元数据元素以及数字对象的时间区间、文献类型两个属性所建立的三维空间坐标系中，每个坐标点上的[时间区间，文献类型]所覆盖的数字对象集合，在该坐标点的持续完整性风险型元数据元素风险点上产生风险的概率。检测结果见图 3。主要风险点有：(a)签名信息，主要分布在 6 个层次单元：2000 年之前的 3 个时间区间的期刊文献、会议论文。(b)固定性信息，主要分布在 2 个层次单元：1995 年之前的 2 个时间区间的期刊文献。(c)固定性检查事件，主要分布在 2 个层次单元：1995 年之前的 2 个时间区间的期刊文献。(d)信息摘要计算事件，主要分布在 2 个层次单元：1995 年之前的 2 个时间区间的期刊文献。(e)保存级别，主要分布在 7 个层级单元：所有 7 个时间区间的专利文献。

(2){时间区间，学科类型，风险型元数据元素}三维风险检测。检测并统计由持续完整性风险型元数据元素以及数字对象的时间区间、学科类型两个属性所建立的三维空间坐标系中，每个坐标点上的[时间区间，学科类型]所覆盖的数字对象集合，在该坐标点的持续完整性风险型元数据元素风险点上产生风险的概率。检测结果见图 4。主要风险点有：(a)签名信息，主要分布在 24 个层次单元：2000 年之前的 3 个时间区间的所有 8 个学科文献。(b)固定性信息，主要分布在 16 个层次单元：1995 年之前的 2 个时间区间的所有 8 个学科文献。(c)固定性检查事件，主要分布在 16 个层次单元：1995 年之前的 2 个时间区间的所有 8 个学科文献。(d)信息摘要计算事件，1995 年之前的 2 个时间区间的所有 8 个学科文献。(e)保存级别，主要分布在 35 个层级单元：所有 7 个时间区间的 5 个学科文献(基础学科、工程技术学科、农业科技、医疗卫生科技、信息科学)。

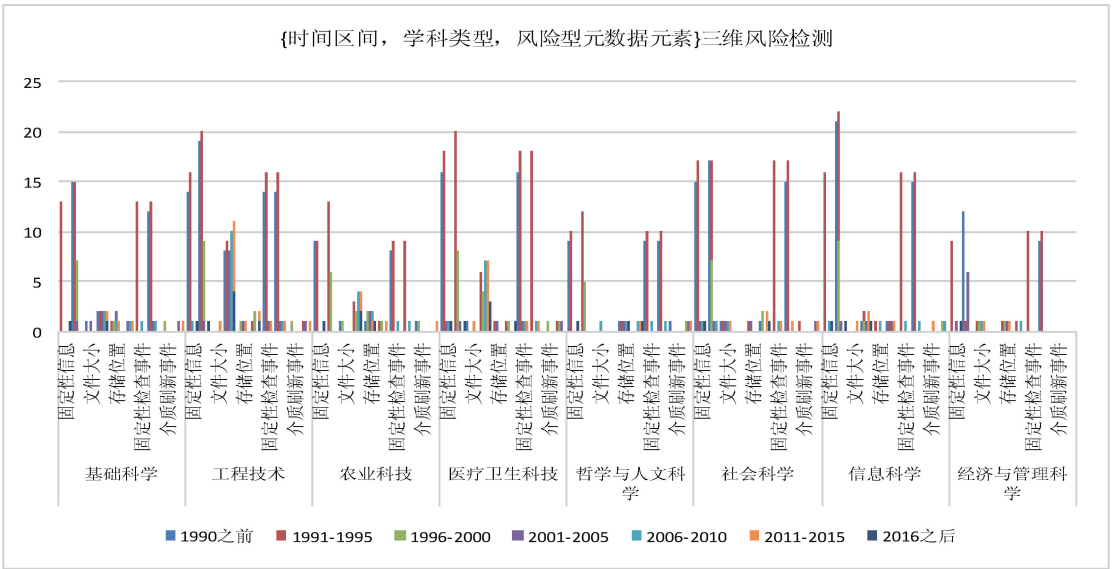

图 4 {时间区间,学科类型,风险型元数据元素}三维风险检测结果

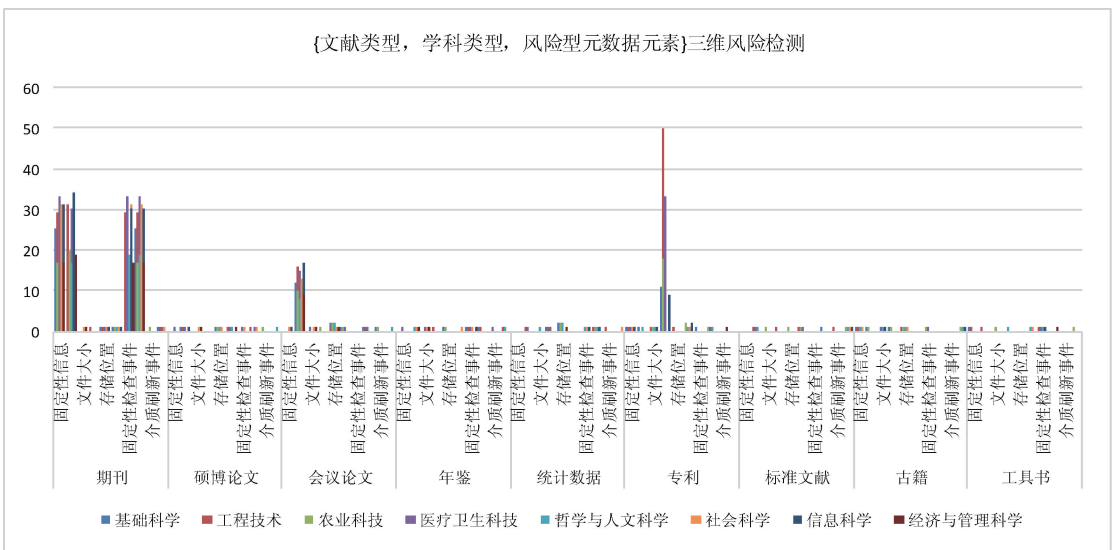

图 5 {文献类型,学科类型,风险型元数据元素}三维风险检测结果

(3){文献类型,学科类型,风险型元数据元素}三维风险检测。检测并统计由持续完整性风险型元数据元素以及数字对象的文献类型、学科类型两个属性所建立的三维空间坐标系中,每个坐标点上的[文献类型,学科类型]所覆盖的数字对象集合,在该坐标点的持续完整性风险型元数据元素风险点上产生风险的概率。结果见图 5。主要风险点有:(a)签名信息,主要分布在 16 个层次单元:所有 8 个学科的期刊文献、所有 8 个学科的会议论文。(b)固定性信息,主要分布在 8 个层次单元:所有 8 个学科的期刊文献。(c)固定性检查事件,主要分布在 8 个层次单元:所有 8 个学科的期刊文献。(d)信息摘要计算事件,主要分布在 8 个层次单元:所有 8 个学科的期刊文献。(e)保存级别,主要分布在 5 个层级单元:5 个学科(基础学科、工程技术、农业科技、医疗卫生科技、信息科学)的专利文献。

4.3.4 四维度风险检测

基于数字对象的三个属性,加上持续完整性风险型元数据的元素,建立一个四维空间坐标系,检测并统计每个坐标点上的数字对象集合在该坐标点上持续完整性风险型元数据元素风险点上产生风险的概率。有下述一种类型:

{时间区间,学科类型,文献类型,风险型元数据元素}。检测并统计由持续完整性风险型元数据元

素、以及数字对象的时间区间、学科类型、文献类型三个属性所建立的四维空间坐标系中，每个坐标点上的[时间区间，学科类型，文献类型]所覆盖的数字对象集合，在该坐标点的持续完整性风险元数据元素风险点上产生风险的概率。检测结果的可视化图太大，省略。主要风险点有：(a)签名信息，主要分布在48个层次单元：2000年之前的3个时间区间的所有8个学科的期刊文献、会议论文。(b)固定性信息，主要分布在16个层次单元：1995年之前的2个时间区间的所有8个学科的期刊文献。(c)固定性检查事件，主要分布在16个层次单元：1995年之前的2个时间区间的所有8个学科的期刊文献。(d)信息摘要计算事件，主要分布在16个层次单元：1995年之前的2个时间区间的所有8个学科的期刊文献。(e)保存级别，主要分布在35个层级单元：所有7个时间区间的5个学科(基础学科、工程技术、农业科技、医疗卫生科技、信息科学)的专利文献。

4.4 检测结果分析

风险检测的目的在于为保存系统的维护提供依据。由上可知，检测维度越高，产生风险的数字对象集合越具体，风险识别的针对性越强，越利于保存系统采取针对性的措施降低或规避风险。针对本实验，持续完整性风险主要集中在下述5个风险点的相应数字对象集合上：

(1)固定性信息、固定性检查事件、信息摘要计算事件。这三个风险点产生风险的概率几乎相同，且都集中在1995年之前各个学科的期刊文献上。由此可以推测，该层次单元中一些数字对象收录到保存系统时，可能没有执行信息摘要计算事件，导致这些数字对象的信息摘要内容缺失，固定性元数据中信息摘要元素内容为空，固定性检查事件因缺失对比的原始信息摘要基准值而无法执行。可能原因是1995年之前的期刊文献的数字化版本大多是通过数字转换获得的，当时可能没有完全执行对收录数字对象计算信息摘要的保存政策。保存系统可对这类数字对象重新执行信息摘要计算事件，并将计算结果赋值到对应数字对象的固定性信息元数据的信息摘要元素中。

(2)签名信息。该风险点产生风险的数字对象主要集中在两个区域：1995年之前各个学科的期刊文献、2000之前的所有学科的会议论文。针对第一个区域的数字对象，由于与本节(1)中文献集合相同，且发生风险的概率值也与本节(1)中的三个风险点比较一致，所以可以推测，该区域的数字对象至少也缺失信息摘要的描述值，可能的原因和保存系统可以采取的措施也同本节(1)。针对第二个区域的数字对象，可能原因是缺失密钥信息或签名值的记录，致使无法计算新的信息摘要，也可能是新信息摘要与原始信息摘要比较结果不同；针对前者，保存系统可以进一步核实数字对象的各项元数据元素的描述值，补充缺失内容；针对后者，保存系统可进一步分析导致数字对象发生改变的因素。

(3)保存级别。该风险点产生风险的数字对象集中在所有时间区间的5个学科(基础学科、工程技术、农业科技、医疗卫生科技、信息科学)的专利文献中。首先，在所有的8个学科中，其他3个学科(哲学与人文科学、社会科学、经济与管理科学)很少产生专利文献，所以专利文献集中在上述5个学科；其次，产生风险的可能原因是专利文献数字对象设置的期望保存级别较高(可能是这类数字资源提交者—国家知识产权局的要求，也可能保存系统认为这类数字资源的价值较高)，而该风险型元数据的元素“保存级别的胜任状态”的赋值为“需要”(意味着保存系统在实现该类数字资源的期望保存级别所需的支撑条件尚不足)。保存系统可以采取的措施是针对这类数字资源，完善保存环境，提升保存条件，满足这类数字资源的保存需求。

5 检测方法的局限性与改进思路

基于检测结果与样本数字对象的对比分析，检测方法还存在下述一些不足，并针对每项不足提出对应的改进思路。

(1)风险识别单元的问题。检测方法中，风险的识别单元是元数据(即风险点)。针对一件数字对象，一个元数据中任一检测项目出现风险，该检测点就产生风险，且有多个检测项目出现风险时，也归并为该检测点出现风险一次。比如“固定性信息”风险点设置了3个检测项目，本实验中，有213件数字对象的“固定性信息”风险点产生了风险，但是具体到每件数字对象，是原始信息摘要算法缺失？原始信息摘要丢失？还是数字对象在长期保存过程中发生了改变？是发生了上述一种情况、二种情况？还是三种情况同时发生了？无从知晓。导致保存系统难

以采取准确措施降低或规避风险。因为不同原因导致的风险，应采用的规避或降低方法不同。如前2个原因导致的风险的规避措施是补齐原始对象的信息摘要或算法即可；第3种原因导致的风险，只有通过本地或异地备份恢复数字对象来解决。

检测方法的改进思路。将检测方法中以元数据为风险识别单元，改变为以检测项目为风险识别单元。这样，可视化展现时，不仅显示每个元数据产生的持续完整性风险的数字对象总数量，还需显示针对一个元数据的每个检测项目上产生风险的数字对象数量，很显然，可视化展示图也会随着增大很多。

(2)统计对象的问题。检测算法中，在一个风险点上产生风险的所有数字对象将形成一个集合，统计该集合中数字对象的个数，形成一个数字，展示在可视化图中相应风险点上。但到底是哪些数字对象？无从知晓，因为缺失数字对象的清单。导致的结果是，无法针对具体数字对象采取风险规避或降低措施。如针对上述例子，本实验抽取的1万件数字对象中，有213件产生“固定性信息”风险，但没有列出这231件数字对象的具体唯一标识符，无法识别出具体的数字对象，也就无法实施风险规避或风险降低的措施。

检测方法的改进思路：在可视化展示图中，加入超级链接，将每个风险点链接到具体产生该类风险的数字对象清单上，并设置打印功能，需要时可打印输出。

(3)元数据的相关性问题。检测方法中设计的不同元数据与持续完整性之间的相关性是不一样的。比如数字对象方面，“固定性信息”和“签名信息”与持续完整性直接相关，相关性最大；“文件大小”“保存级别”“存储位置”和“存储介质”与持续完整性都是间接相关，相关性较小；保存事件方面，“固定性检查”和“信息摘要计算”也是直接相关，但“介质刷新”和“病毒检测”则是间接相关。直接相关的元数据的检测结果更准确，间接相关的元数据的检测结果都存在误差，有的误差很大。将检测结果与样本进行对比，“固定性信息”和“签名信息”的2个风险点的检测准确度都大于90%，“固定性检查”和“信息摘要计算”的2个保存事件风险点的检测准确度也都大于90%，但其他间接相关的风险点检测结果的准确度都较低，在10%—45%之间。这样，对间接相关的元数据产生风险的数字对象的识别所需工作量很大。

检测方法的改进思路：采用分级检测，首先使用直接相关的元数据进行检测，将产生风险的数字对象析出，剩余的数字对象再使用间接相关的元数据进行检测。由于后者的检测样本集已减小，所以识别所需的工作量也随着降低。

6 检测方法的使用

本文设计的检测方法针对CNKI进行了实验，结果表明，除了存在上节列出的局限性外，其他方面均具有较好的适用性。CNKI保存的主要是文本型数字对象，针对其他类型保存系统(如多媒体数字对象的保存系统)的实验没有进行。因此，使用该检测方法(尤其是非文本型数字对象的保存系统)时，保存系统需注意下述事项。

(1)元数据的完善。本检测方法的核心是持续完整性风险型元数据的设计，检测结果的准确度和全面性依赖于所设计的元数据方案的科学性。因此使用该方法时，保存系统应该针对其保存的数字对象、保存目标、保存政策、目标用户群体等实际，分析、改造和完善本文设计的元数据方案。

(2)检测项目的完善。本检测方法中，每个元数据均设置一定数量的检测项目，对元数据的检测是通过对其设置的检测项目进行检测而实现的。因此，检测结果的准确度完全依赖于设置的检测项目。使用该方法时，保存系统应在上述完善元数据的基础上，结合实际，改造和完善每个元数据的检测项目。

(3)维度的划分。本检测方法的实验样品来自CNKI，因此实验中的维度划分完全基于CNKI的实际。但应用到其他保存系统时，需根据其收录数字对象的实际，重新划分维度。

(4)动态风险的监控。本检测方法仅局限在静态风险的检测，没有涉及动态风险的监控。可以在静态风险检测的基础上，从时间维度设置一个检测频率(如每天检测一次)，基于该频率进行持续的离散的静态风险检测，结果就形成了动态风险的检测。当然，这种动态非完全连续，而是离散式的。实际上也无需完全连续检测，因为保存系统中数字对象的状态变化不可能完全连续，基于一个合理的检测频率进行离散式检测即可。在此基础上，可设置一个时间区间(如一个星期、一个月、一个季度或一年

等),统计该时间区间内动态风险的检测结果,实现包括集中趋势、离散趋势、分布形状和时间趋势等在内的各种统计,并以可视化形式呈现,最终实现动态风险的监控。

风险检测应是数字保存的一项常规工作,也是规避和降低风险的基础。本文仅对持续完整性风险设计了一种检测方法。实际上,除了该类风险外,数字保存还存在其他类型的风险(比如可用性风险、真实性风险等),识别这些类型的风险并设计其元数据方案,并在此基础上设计相应类型风险的检测方法,乃至进一步整合为数字保存的全风险型元数据并进行全风险的检测方法设计,是本课题的后续研究内容。

参考文献

1 Paul C..Preservation in the digital world[EB/OL]. [2016－10－12].https://www.clir.org/pubs/reports/reports/conway2/index.html.

2 Lawrence G. W..Risk Management of digital information: a file format investigation[EB/OL]. [2016－12－23]. http://www.clir.org/pubs/reports/pub00/contents.html.

3 Arms C.. Digital formats: factors for sustainability, functionality, and quality[EB/OL].[2016－11－09]. http://www.imaging.org/IST/store/physpub.cfm? seriesid＝28&pubid＝692.

4 Stanescu A.. Assessing the durability of formats in a digital preservation environment: the inform methodology [J]. OCLC Systems & Services, 2005, 21(1): 61－81.

5 Rog J.. Evaluating file formats for long-term preservation[EB/OL]. [2017－01－23].http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf.

6 Wright R.. The significance of storage in the "cost of risk" of digital preservation[J]. International Journal of Digital Curation, 2009, 4(3): 20－32.

7 McGovern N. Y., et al. Virtual remote control[EB/OL]. [2017－01－20].http://www.dlib.org/dlib/april04/mcgovern/04mcgovern.html.

8 Barateiro J.. Addressing digital preservation: proposals for new perspectives[EB/OL]. [2017－01－09]. http://cs.harding.edu/indp/papers/barateiro7.pdf.

9 Clifton G.. Risk and the preservation management of digital collections[EB/OL].[2016－12－20]. http://archive.ifla.org/VI/4/news/ipnn36.pdf.

10 Rosenthal D.S.H., et al. Requirements for digital preservation systems[EB/OL]. [2017－01－15].http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html.

11 Thomaz K. P.. Critical factors for digital records preservation [J].Journal of Information, Information Technology, and Organizations, 2006 (1):21－39.

12 PARSE. Insight consortium.. Science data infrastructure roadmap[EB/OL].[2017－01－02].http://www.parse-insight.eu/downloads/PARSE-Insight_D2－2_Roadmap.pdf.

13 DCC.. Digital repository audit method based on risk assessment [EB/OL]. [2016－12－02]. http://www.repositoryaudit.eu/.

14 Dappert A.. Report on the conceptual aspects of preservation [EB/OL]. [2016－12－21]. http://www.planets-project.eu/docs/reports/Planets_PP2_D3_ReportOnPolicyAndStrategyModelsM36_Ext.pdf.

15 Sally V., Brian L., Priscilla C.. Identifying threats to successful digital preservation: the SPOT model for risk assessment[EB/OL]. [2016－12－19]. http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html.

16 Littman J.. Actualized preservation threats[EB/OL]. [2016－12－21]. http://www.dlib.org/dlib/july07/littman/07littman.html.

17 McLeod R.. Risk assessment: using a risk-based approach to prioritise handheld digital information[EB/OL].[2015－05－22]. http://www.bl.uk/ipres2008/presentations_day1/20_McLeod.pdf.

18 CCSDS.Reference model for an open archival information system (OAIS) [EB/OL]. [2017－02－25]. http://www.ccsds.org/documents/650x0b1.pdf.

19 CCSDS.Audit and certification of trustworthy digital repositories [EB/OL]. [2017－02－03]. https://www.iso.org/obp/ui/#iso:std:iso:16363:ed－1:v1:en.

20 CCSDS. Requirements for bodies providing audit and certification of candidate trustworthy digital repositories[EB/OL]. [2017－03－03].https://www.iso.org/obp/ui/#iso:std:iso:16919:ed-1:v1:en.

21 PREMIS. Data dictionary for preservation metadata(Version 3.0) [EB/OL].[2017－03－19]. http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf.

22 同6.

作者单位:郑州大学信息管理学院,郑州,450001
郑州大学公共管理学院,郑州,450001

收稿日期:2017年4月5日

(转第111页)