2014 年第 3 期（总第 36 期）

长期保存跟踪扫描

主办单位：中国科学院国家科学图书馆

2014 年 4 月

为传播科学知识，促进业界交流，特编译《长期保存跟踪扫描》，仅供个人学习、研究使用。

目 录

【信息扫描】 1

美国 2014 年个人数字存档会议介绍 1

2014 年数字监管创意挑战赛 2

【动态追踪】 3

2014 年 NDSA 创新奖现已开始提名 3

JISC 报告：《数据保存与分享的价值及影响》 4

【重要文献摘译】 4

《NDSA 数字资源长期保存级别：解读与使用》 4

4C 报告：《D4.3—数字保存中的经济决定性因素：质量与可信度》摘译 15

【技术与工具】 38

由 ARC 到 WARC 的可扩展性转化的几点思考 38

在 ARC 向 WARC 转换中如何处理重复数据删除的记录？ 43

ToMaR——如何让保存工具规模化 45

【信息扫描】

美国 2014 年个人数字存档会议介绍

2014 年个人数字存档会议于 4 月 10 日-11 日在印第安那州立图书馆举行。该会议旨在提高个人、公共机构和私营公司对个人数字内容进行创建、保存和持续使用的意识。一个关键的主题是图书馆、档案馆和其他文化遗产机构如何在我们的社区和特定的社区里支持个人数字存档。

为期两天的会议有多种多样的专题演讲，包括：当地社区的文献档案存档实践；处理数字存档的工具和技术；建立、管理及存档学术实践与家族历史档案情况的调查；在个人数字存档对不同类型用户益处的传播方面所面临的挑战。

通过以下的介绍可以快速了解该会议：

- 主讲嘉宾将从研究人员和个人数字信息创建者的视角探讨保存面临的挑战。来自印第安纳大学-普度大学印地安那波利斯分校信息学和计算学院的 Andrea Copeland 谈论了他对公共图书馆用户数字保存实践的研究。音乐历史学家和作家 Charles R. Cross 从一个传记作者的视角谈论了个人数字存档的价值。
- 许多机构缺乏充足的、实施个人数字记录保存的基础设施。许多报告是关于利用特定的工具和服务来进行个人数字信息保存的实践工作。其中一些服务需要收费，比如帮助个人建立他们自己的个人数字档案馆。其他的报告来自于图书馆馆员、档案工作者和研究人员，他们利用某些工具来帮助他们的机构管理个人数字记录。
- 在个人数字存档过程中，同样缺乏与接收、捐赠或法律规定及研究人员兴趣相关的知识与实用的保存策略。为了帮助理解其中的一些问题，来自不同领域的从业者、学者和个人将分享他们目前在个人数字存档主题方面的研究。此次会议首次邀请了当代建筑师和景观设计师来谈论他们开展的保存及存档工作。由专业人员组成的社区不经常在 PDA 会议中进行介绍，因此这次报告提供了一个分享他们面临独特挑战的机会。

编译自：

http://blogs.loc.gov/digitalpreservation/2014/03/things-to-know-about-personal-digital-archiving-2014/

（唐果媛编译，王敬 吴振新校对）

2014 年数字监管创意挑战赛

NDSA 创新工作小组推出了 2014 数字监管创意（DSII）挑战赛！

这是一个关于创意的挑战赛,即应对数字监管挑战的想法或创意。

NDSA 举办这一挑战赛，是为了鼓励社区去关注、思考并践行 2014 NDSA 所确定的研究重点。NDSA 整合了数十位专家的看法，目的是让资助者和决策者能深入洞察新兴技术趋势以及数字监管能力与重点发展领域之间的差距，该报告于去年夏天发布。

NDSA 提出了许多数字监管社区应考虑的重点研究领域。这些研究领域面临的共同挑战是缺乏足够可用的经验性实例来支撑这些研究领域。他们希望 DSII 挑战赛的推出及大家创意的分享可以开发一个用来解决 NDSA 研究重点的共享性实证基础平台。

以下是关于 DSII 挑战赛的一些细节

数字监管创意（DSII）挑战：提交一个创意

DSII 挑战的宗旨是鼓励和促进数字监管社区中的创新，特别是延长保存系统和媒体的寿命。该挑战赛面向来自工业、政府、学术界的任何人。

挑战赛申请要求:

提交的申请书必须明确提出与 2014NDSA 给出的研究重点相关的创意，研究重点包括：

- 应用研究的成本模型和审计模型
- 理解信息的等价性和重要性
- 信任度框架政策研究保存规模
- 数字保存的实证基础

提交的申请书需满足以下给出的最低要求：

- 申请书给出的创意能被在数字内容管理领域工作的技术和非技术专家理解，比如：信息技术专家、计算机科学家、保管人员、档案管理员和媒体专家；
- 申请书明确描述了提出想法的创新之处。

编译自：

http://blogs.loc.gov/digitalpreservation/2014/04/digital-stewardship-innovation-ideas-challenge-for-2014/

（唐果媛编译，王敬 吴振新校对）

【动态追踪】

2014 年 NDSA 创新奖现已开始提名

国家数字监管联盟（NDSA）创新工作小组宣布 2014NDSA 创新奖提名开始了。作为一个对数字保存有共同承诺的多元化成员组织，NDSA 明白创新与风险承担在开发及支持普遍而成功的数字保存活动中的重要性。这些奖项的设置就是 NDSA 承诺鼓励和表彰数字监管社区中创新的一个例子。

年度奖项是为了突出和表彰在数字保存领域提供创意与卓越贡献的有创造力的个人、项目、组织及未来管理者。该项目由来自 NDSA 创新工作小组成员组成的委员会管理。

去年的获奖者就是一个利用多样性和协作来支持数字监管社区的范例，其提供的创意是保存数字资料，并使数字资料具有可用性。欲知更多有关去年获奖者的信息，可参见去年获奖者的博客。

NDSA 创新奖侧重于识别以下一个或几个领域中的突出表现者：

- 在数字保存领域做出突出的、创新性贡献的个人；
- 目标或产出增加了对成功、可持续的数字保存监管或其所需流程的创造性、有意义的理解的项目；
- 采取创新的方法为数字保存社区提供支持和指导的机构；
- 未来的监管者，特别是学生，也包括教育工作者、培训人员和授课人员，采取创新的方法来推进数字保存理论和实践知识。

认识到创新的数字监管可以采取多种形式，因此获得这些奖项的候选资格被特意放大。任何个人或属于上述类别的任何实体都可以申请四个奖项中的任何一个提名。提名人应该是美国的人员和项目，或者是与美国人员有合作的国际项目。这给大家提供一个机会，来帮助突出和表彰那些为应对数字保存挑战而提出的新颖、承担风险及独出心裁的方法。

编译自：

http://blogs.loc.gov/digitalpreservation/2014/03/nominations-now-open-for-the-2014-ndsa-innovation-awards/

（唐果媛编译，王敬 吴振新校对）

JISC 报告：《数据保存与分享的价值及影响》

JISC 于4月2日发布一篇研究报告:《数据保存与分享的价值及影响（*The value and impact of data curation and sharing*）》。这篇报告汇集了对英国三个历史悠久的研究数据中心价值及影响的一系列独立调查结果，这三个数据中心分别是：经济与社会数据服务中心（ESDS）、考古数据服务中心（ADS）、英国大气数据中心（BADC）。

报告探索了数据中心的价值并试图让这种价值可测量化，继而思考价值及影响如何在未来得到最佳的监控与分析（就像服务的发展一样），并提出了一系列建议。这些建议包括保持对数据中心的支持、令使用的统计数据标准化（以更容易地进行对比与比较，并从最佳表现中学习经验），报告还建议进行进一步的工作来识别与评估更广泛的社会及英国经济的利益（这种利益随着数据中心的发展和成熟变得更为明显）。

全文参见：

http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

编译自：

http://www.jisc.ac.uk/publications/reports/2014/data-sharing-and-curation.aspx

（陈瑶编译，唐果媛 吴振新校对）

【重要文献摘译】

《NDSA 数字资源长期保存级别：解读与使用》

摘要

美国国家数字监管联盟（NDSA）把“数字资源长期保存的级别”定义为一套分层次的关于组织应该如何开始建立和加强数字资源长期保存活动的建议。这是一项进行中的工作，其目的不仅为刚开始考虑保存数字资产的机构，而且还为下一步计划加强目前数字资源长期保存系统和工作流的机构提供一套相对容易使用的指南。它允许机构对他们保管工作中获取的特殊资料进行保存级别的评估。它不能用于评估数字资源长期保存项目整体的稳定性，因为它不包括政策、人员和组织支持等。该指南把数字资源长期保存系统的核心分为五个功能区：存储和地理位置、文件固定性和数据完整性、信息安全、元数据以及文件格式。

本文介绍了各级别的含义，说明了NDSA项目的发展背景，描述了提出每一条指南的理由以及他们以这种方式被优先考虑的原因，对如何使用这些指南给出了建议，并把保存的级别与其他评估数字资源长期保存阶段的方法进行了对比。其他评估模型包括：Nancy McGovern和Anne Kenney提出的“数字资源长期保存的五个组织阶段”，Charles Dollar和Lori Ashley提出的“数字资源长期保存能力成熟度模型”，以及2012年的OCLC研究报告“跑之前必须会走：管理来自物理介质的原生数字内容的第一步”。最后，本文要求反馈工作进展，并概述未来的工作计划。

背景介绍

美国国家数字监管联盟（NDSA）由140多个组织组成，其使命是“为当代和后代子孙建立、维护及提高保存国家数字资源的能力”。NDSA最近制定出了数字资源长期保存级别。数字资源长期保存的级别是一套分层次的指南和实践，旨在五个不同功能区的四个渐进层次为保存数字内容提供清晰的基准说明。这些级别中的评论活动对于内容类型和技术来说是无关的，它们专注于特定的保存活动（不是组织要求的），其目的是为所有规模及资源层次的机构提供实施数字资源长期保存时可利用的实践蓝图。数字资源长期保存级别图表的首要目标是满足简单易得的实践需求，这些建议比传统的面向个人的数字存档建议要更具实质性，比可依赖数字存储库的认证需要要宽松和容易。本文详细介绍了数字资源长期保存级别在NDSA中的起源和发展，解释了它的目的和目标，评价了与数字资源长期保存相关的模型，展示并阐述了这些级别。本文包含了对使用这些级别和实施相关活动的建议。本文是为了鼓励更多的社区提供反馈，并支持级别的持续进化和改进。

数字资源长期保存级别的产生与发展的核心是支撑NDSA的合作精神。NDSA是一个由各种机构组成的联盟，包括大型的研究性大学和小型的文化遗产机构、非营利组织和商业合作伙伴。NDSA提供了一个理想环境，有利于开发资源帮助不同类型的个人和机构来启动或实施数字资源长期保存项目。NDSA中拥有不同的技能和专业知识的多样化成员致力于完成与数字监管领域相关的许多任务和职责。在构思和阐述数字资源长期保存级别的目标和最终形式时，这种多样性是非常宝贵的。值得注意的是，这个项目是第一个在NDSA范围内合作的项目，其成员来自NDSA的五个工作小组：内容小组、标准小组、基础设施小组、创新小组以及推广小组。该项目中的团队合作能反映NDSA成员的多样性，以及在协作和跨学科小组中工作的能力。

NDSA中许多具有不同背景的成员都认识到需要一个实际的、可操作的和可扩展的数字

资源长期保存指南，这个指南既方便那些在数字保存领域刚起步的机构，也方便那些全面实施保存项目的机构，基于此，该项目给出了数字资源长期保存级别的定义。通过对NDSA成员的非正式调查和对目前数字资源长期保存模型的研究，项目小组确定了一些数字资源长期保存级别的预期目标。团队希望这些级别能独立于特定的格式、内容类型和存储系统，从而加强其在跨领域中的可用性。团队希望数字资源长期保存的级别能在一定范围内实施起来，而且尽可能的在具体行动中简单实用。这些级别应该能够及时告知程序，以减轻数字内容的损失，同时也应能帮助预测保存领域的下一步走向，以及支持保存工作的战略规划和内部宣传。

项目团队着力打造一个活动矩阵，该矩阵很详细且有意义，但是不够简洁，因而不能在一张单页纸上呈现。这个活动矩阵的目标旨在增强图表知识的可访问性，并阐明与数字资源长期保存相关的一些眼花缭乱的活动。为了完成数字资源长期保存，以及提供让现有的项目变得更加健壮的建议，团队希望其产品能定义一个被社区认可的前提条件的最低水平。此外，团队还希望这些级别的语言是非技术性而通俗的术语，能让刚开始进行数字保存工作的机构感觉友好和避免混淆。同样，这些级别将采取非评判的方法，以帮助那些不确定如何开始数字资源长期保存项目的机构从中获得最大效用。从根本上讲，NDSA希望数字资源长期保存级别容易理解并被广泛使用。

数字资源长期保存级别的分层和矩阵方法具有多个层次和内容领域，并致力于灵活性，用户可以根据他们独特的需求和资源在不同的内容领域获取不同的级别。重要的是，为了达到实施的即时性，团队希望这些级别能专注于实践，而不是政策和工作流程。本着同样的精神，该项目认识到社区与使用驱动的发展是其核心特征，因此目前的级别图表被认为是“第一版 ”。该团队的最终目标是设计一个类似于数字监管本身的资源，该资源能随着时间不断适应和完善。NDSA成员和更广泛的社区持续、专注的支持是数字资源长期保存级别的持续发展的根本。

比较目前的模型

在提出NDSA的级别工作时，级别团队总结了现有的数字资源长期保存的工具和文档。2012年季春团队认为没有什么可以提供给保存专家，满足他们在初步开始保存工作或者在已有的保存工作基础上采取进一步措施时对于实践性的技术指南的需求。但是项目开始后，团队更加系统的评价了其它的工具并记录了NSDA级别适合的位置。目前许多的数字资源长期保存模型的目的都是给管理层而不是技术层的用户提供建议，同时，这些模型解决了全面的

数字化保存方案。相比之下，NDSA的级别图表假设的用户是数字资源长期保存的实践者，这些用户负责实际操作。NDSA的级别图表提供了一些可以逐步减少数字资料各种风险的活动，因此分析单元不是整个的数字资源长期保存项目，而是需要被保存的特殊材料。

2003年，当Nancy McGovern和Anne Kenney提出“数字资源长期保存的五个组织阶段”时，许多机构近乎束手无策，无法开启数字资源长期保存工作，因为他们在等待一个好的技术方案的出现。McGovern和Kenney提到，阻碍许多机构数字保存工作进展的主要障碍实际上是机构的准备状况而不是技术，因此他们在论文中划分了机构在维持数字保存工作发展的阶段。

在接下来的几年中，也有许多工作涉及到了机构对数字保存工作的支持，包括TRAC；近期讨论保存级别的其他成果都基于TRAC，并继承了以评估整体项目作为重点。比如Charles Dollar和Lori Ashley提出的“数字资源长期保存能力成熟度模型”使用了来自TRAC和其他相关成果的评估准则，这些准则对于数字保存项目发展提供了全面而严格的评估级别。

这些成果对于那些需要规划项目发展并为项目发展设计方案的管理者和行政人员来说是相当有用的。但是，它们几乎没有为从业者提供实用指南，以搞清楚在保管数字材料时应采取什么措施可以降低风险。2003年，McGovern和Kenney注意到，组织的准备对于开始一项数字保存工作来说是一个巨大的绊脚石，NDSA级别团队发现，各种类型的机构都努力想在2012年开始保存工作，但是又回到了关注保存领域的技术层面上。也许两种类型的指南文档能共同帮助机构从适度的第一步开始启动强健的数字保存项目。

级别团队不是唯一一个在2012年重新关注实践技术措施的团队。OCLC研究组也发表了类似主题的文章，“跑之前必须会走：管理来自物理介质的原生数字内容的第一步”。文章与NDSA的数字资源长期保存级别非常相似，事实上，文章中提出的技术建议与NDSA图表中的指南是相同的。但是，OCLC的文档给出的范围更狭窄。OCLC的研究范围限制在从物理介质上产生的数字内容，只解决了减少数字材料风险的第一步，而不是按照一个顺序逐步去减少数字材料的风险。

在对现有的和新兴的工具调查中，NDSA团队认为数字资源长期保存级别满足了在其他地方没有得到专门解决的需求。

数字资源长期保存的级别（第一版）

希望随着时间的推移，在收到额外的反馈、从实施各项建议的过程中获得经验以及实证

研究提供有关数据丢失的详细信息后，数字资源长期保存的级别能够进行更新。因此，“级别”将对每次更新进行版本管理。第一版见表1。

	级别一 （保护数据）	级别二 （了解数据）	级别三 （监控数据）	级别四 （修复数据）
存储和地理位置	•两个存储于不同地点的完整副本。 •对于来自不同介质媒体（光盘、硬盘驱动器、磁盘）上的数据，复制其数字内容并存入存储系统。	•至少保持三个完整副本。 •至少有一个副本在不同的地理位置。 •记录存储系统和存储介质，以及使用它们所需的信息。	•至少有一个副本放置在具有不同灾害威胁的地理位置。 •有一个监测存储系统和媒介退化的程序。	•至少三个副本都放置在具有不同灾害威胁的地理位置。 •拥有一个全面的整体规划，能够确保文件和元数据存储在当前可访问的媒介或系统上。
文件不变性和数据完整性	•如果内容附带有不变性信息，则在摄入阶段进行检查； •否则，创建不变性信息。	•检查所有摄入对象的不变性。 •在处理原始媒介时使用只读隔离保护器。 •对具有高风险的内容进行病毒检查	•定期检查文件的不变性。 •维护好不变性信息日志以备审计检查所需。 •能够检查出损坏的数据。 •对所有内容进行病毒检查。	•在特殊事件或活动发生后检查所有内容的不变性。 •能够替换或修复损坏的数据。 •确保没有人能对所有副本进行写操作
信息安全	•能识别谁对于哪个文档有读、写、移动、删除的权限。 •对独立文档拥有上述权限的用	•为文件内容设置访问限制	•维护文件操作日志，包括删除和保存操作	• 施行日志审计。

	户进行限制。			
元数据	•提供内容及其存储位置的清单。 •保证清单无重复记录和清单备份。	•存储管理型元数据。 •存储发生变化的元数据并记录发生的事件。	•存储标准的技术元数据和描述元数据。	•存储标准的保存元数据。
文件格式	•当创建数字文件时，鼓励使用一个有限的已知开放文件格式和编码的集合	•制订所采用文件格式的详细目录。	•监测文件格式过时问题。	•必要的话，进行格式迁移、仿真及类似的活动。

表1：数字资源长期保存的级别第一版

级别的一般结构：类别和层级

图表的整体结构是渐进的，第一级别的工作要么是第二至第四级别的先决条件，要么本身是需要最先完成的最紧迫的事情。这五大类别（存储和地理位置、文件不变性和数据完整性、信息安全、元数据以及文件格式）在项目初期就商定好了。考虑到对数字资源长期保存的技术和直接威胁，级别团队把这些领域确定为广义范围内的重点。在这方面，这些类别是级别团队中的主题专家对自己工作进行归类而形成的。这就是团队成员对他们降低风险与威胁的描述。

相对于其他工作，一些读者也许会问，为什么有关权利和政策的问题被排除在外。一开始，团队主要关注的是技术问题；目标是确定技术功能和特性，他们希望看到这些在某处出现并保障数字内容的长期访问，而不是维持这些保存活动的社会或法律结构。而且，该项目不是提供一个数字资源长期保存的计划，而是提供一个图表，以帮助对数字信息长期访问感兴趣的任何人评估他们在降低风险损失时所做的工作，并确定接下来为使他们的工作进入下一个级别而应采取的措施。

从广义上讲，保存工作从级别1到级别4每上升一个级别，那么保存工作就从保障少量保存的基本需求上升到跟踪数字内容并确保在未来更长的时间里能被访问的更广泛的需求。网格中的这五大类别在项目开始初期被商定时，每个级别（保护数据、了解数据、检测数据和

修复数据）的标签应涵盖的范围就存在争议。团队中的一些成员希望忽略这些标签，而严格按照所察觉的需降低损失的最大风险来组织文档。团队中另外一些成员认为标签可以从概念上帮助组织网格，并且能够帮助解释每个级别的总体目标。最终，分类的概念价值胜出，并作为图表的一部分。然而，非常值得注意的是应用到每个级别的标签只是粗糙的界定，而不是在给定的级别里究竟应该怎么去做的法令。任何情况下，当如何处理特殊活动的纯粹概念与团队认为需要首先解决的务实的行动存在冲突时，团队会站在务实的行动一方而不选择纯粹的概念。

级别和分级的详细解读

下文简要阐述了每个级别的每个特定特征的论证。

存储和地理位置

单元格中的第一个要素关注的是数字信息的存储。其下一个级别就是保存额外副本，这有助于规避因存储介质损坏和系统故障造成的丢失威胁。同样，接下来一个级别采用了额外的地理位置，来规避对存储系统的地区性威胁（如自然灾害和人为灾害）。在最基础的级别中，为了确保在未来能获得这些数字材料，所应采取的第一步就是创造副本。因此，这一要求是图表中的第一项。

除了在这个分类中的两个综合趋势，第一个级别需要从异质移动媒介（光盘、外置硬盘等）获得数据，然后存入存储系统。团队认为这是重要的第一步，因为这种异质的存储介质存在失败的风险并需要大量的人工工作来确保数据的完整性。此外，这个第一步还是保障本文其他级别所要求的保存活动的必要手段。“存储系统”的含义是含糊的，因为团队不希望过多关注任何特定的存储技术。给出“级别”文档中一整套级别文档要求的特性，存储系统通常可以被理解为全部使用旋转磁盘或使用旋转磁盘与磁带组合的近线系统或在线系统。

级别2、3和4有额外的要求，它们的重点是确保存储系统的寿命：首先需要记录系统，其次需要一个监测存储系统和媒介退化的程序，最后需要一个能够确保内容始终存储在当前可访问的媒介或系统上的全面规划。逐渐递增这些步骤的目的主要是能够比较好地开展一系列活动，并使得这些活动能够在前一级别的工作基础上变得越来越复杂。

文件不变性和数据完整性

数字资源长期保存最重要的组成部分是能够证明保存材料的不变性和完整性。这是数字资源长期保存的基本组成部分，但是对于许多机构来说，检查内容的不变性仍然是个挑战。该类别的目标是通过提供一系列的步骤，让组织开展强有力的行动以确保内容的不变性。

第一级别的建议比较简单：如果与内容一起提供了不变性信息（类似加密哈希函数MD5和SHA-1），则在摄入阶段进行检查，否则，创建不变性信息。对于机构来说这是必要的一步，以验证保存的内容是否是他们想要保存的内容。许多机构借助Bagger等工具或使用BagIt规范打包数字内容来完成该步骤。

下一级别增加了帮助进一步确保内容的完整性的补充活动。尤其需要注意的是，级别2需要检查所有摄入对象的不变性，级别3与4则更多的关注持续检查数字内容的需求。级别3和4的需求从对特殊存储媒介的质量和性能的信任转移到通过持续重复的检查数字内容来确保保存上。这样不但增加了额外的保障，对声明所管理的内容的不变性也更有信心。

信息安全

信息安全部分主要侧重于了解谁有权访问数字内容，谁可以对内容执行哪些操作并且强制执行这些访问限制。从最基本而简单的步骤开始，即确定谁可以对内容采取哪些操作。这是至关重要的，如果没有适当的程序来限制对于内容的操作，就必然存在有人误删数字内容的风险。在此基础上，级别2提升了访问限制。级别3建议保留操作日志，这有助于使机构采用的方法与存档最佳实践保持一致。级别4增加了审计这些操作日志的要求，这将有助于检查计划中及实际发生的操作。

像许多其他部分一样，这些级别主要是通过在每一个级别建立适当的内容来满足，这些内容是达到更高要求的先决条件，并相对于团队在其他类别中遇到的风险来调整到最小风险。

元数据

与元数据相关的问题在许多其他级别中出现过，这就决定了把元数据问题作为图表的单独一行是至关重要的。团队从广义上定义了元数据，包括：文件位置的库存信息，一套更广泛的管理元数据（比如创建的时间和方式，谁有权限访问），记录导致对象改变的变化元数据，技术和描述元数据以及保存元数据。

团队提出按顺序组织级别，这些级别给出了较高级别与较低级别所应具备最基本元数据的建议，元数据的附加层将使数据得到更好的保护，并使数据具有更好的识别性与访问性。值得一提的是，在大多数系统中，几乎所有的元数据（除了描述性元数据）都可以自动产生和处理，而不需要手动添加。

文件格式

数字对象是紧紧依赖于文件格式的结构和特性。图表的文件格式部分在公众审查过程中经历了最多的变化和修改。团队给出的建议考虑到了存在的可能性-即并非所有的格式都需要迁移或模拟，但是在级别4中，建议为需要迁移或模拟的任何格式主动开展保存措施。

第一级别只是简单建议，如果可能的话机构应鼓励使用有限的已知的开放文件格式。特别是在一个机构在数字化材料的过程中并对使用的格式有相当大的发言权的情况下。对这种工作，应该参考像联邦机构数字化准则这种权威资源作为附加的文件格式建议。应该注意到，在许多其他情况下，包括从异质有形媒介或网络存档中收集原生数字存档材料，把文件格式的任何变化强制作为一个基本的要求都是不可行的。

连续的级别从记录使用的格式并监测这些格式过时的问题开始，并最终进行格式迁移、支持仿真或考虑其他模式来确保保存的内容在未来能够使用和访问。团队在第四个级别中才基于格式过时的问题有所行动有以下几个原因。首先，过时是一个艰巨的问题，它需要通过基本的位保存和数据管理来解决。总之，如果不能打开一个文件，但是该文件仍属于某人所有。除此之外，有关迁移和仿真的问题是引起广泛而持续研究与发展的领域，很可能是当某个机构在解决与前三个级别相关的问题时，会在管理文件格式工作中出现实质性进展。最好首先按级别顺序开展工作，继而再讨论何时利用特定的方法来进行适用于特定目标的迁移和仿真。

指南的使用

级别团队设定的最初用例是作为机构在数字资源长期保存系统优化时的参考。虽然该指南比较新，但经证明它们不仅对这个目标有用，还在其他未知的方式下发生作用。本节中将描述保存级别的其他可能用途。

用途：识别保存社区中在哪些方面存在普遍共识

为了获得级别的第一版本，参与级别制定的NDSA成员对此进行了大量的讨论和辩论。基于这些讨论产生并修改了很多版本的“级别”。在这个工作过程中，“级别”演变成一个产品，虽然仍有一些条款没有得到一致支持，但总体来说，团队是支持该指南的。“级别”被发布到网上后，新一轮的讨论和辩论接踵而至，包括来自世界各地的从业者，涉及到的重要主题包括：验证文件格式的用处，是否需要在摄入对象时就全力实施格式标准化，以及需要在不同灾害威胁的地方存放副本的数量。达成共识的建议被纳入了版本一。预计随着时间的推移，“级别”将会在其他论坛获得评论和辩论，所达成的共识也会继续纳入版本中，那么“级别”就将继续反映社区的数字资源长期保存的最佳实践。

用途：内容创建者和贡献者的培训与发展指南

通过仔细阅读“级别”图表可以发现内容创建者的活动和工作（比如使用开放格式和编码）与能提供数字内容保存服务的级别之间存在直接关系。例如，如果内容创建者提供一些关于内容的描述性元数据，那么内容就可能获得级别3的服务。一个机构可以直接把该图表作为一个教育工具，或把图表中的信息转化为指南以展示内容创建者如何更好的参与数字内容保存。

用途：验证本地保存指南

有人对“级别”的第一版给出了反馈：“这正是我们在NPS中所需要的，这样当谈到保存建议时，我们可以向从业者和管理者展示我们并没有胡编”。因为在“级别”图表中展示的指南是由数字资源长期保存从业者提出来的，他们提出的全部或部分建议很可能会与机构内部给出的意见交叉。当这种情况发生时，从业者可以把”级别”图表作为证据，以证明他们的意见是与更广泛的保存社区的想法和实践保持同步的。

用途：为第三方保存服务提供商制定要求

“级别”图表为保存定义了一些核心的最低需求。当从外部公司或机构征求或谈判保存服务时，这些指南可以从内容持有者的角度重新表达成需求。例如某个RFP可以规定必须实施在文件不变性和数据完整性类别中所描述的所有活动。

用途：评估是否符合保存的最佳实践，并确定要提升的重点领域

“级别”图表看起来似乎很简单，但实际上它可以支持多种类型的评估。评估的单元是灵活的。它可以用来评估整个保存信息库或信息库中的某个环节（比如存储）的保存能力。或者它也可以用来评估所接收的特定集合或内容的数据流的保存程度。引用的图表内容也可以不同。级别1这一列可单独用于参考被推荐的第一步。或者可以参考单独一行以深入探讨某一特定领域（如，只是“存储和地理位置”）。另外整个图表可以用来做一个全面的评估。

与其他一些评估模型不同的是，使用“级别”的评估结果不可能是一个单一的分数，比如级别2。图表是由五个不同的功能区域（元数据等）组成的，在给定的实现中他们不一定是相关的。例如，某个机构可能发现他们的信息库的信息安全处于级别2，而元数据处于级别3。此外，在一个功能区域中，并不是所有的级别都建立在之前的级别基础上。某个机构会发现它符合级别3的元数据而不是级别2。最后，许多单元格都包含多种指南。一某个机构可能会发现在文件不变性和数据完整性中只是部分符合级别2的指南。由于这些原因，最好把”级别”看作渐进的阶段或服务层次，而不是“分数”。为了符合最佳实践，它们可以被

用来识别广泛的领域，以改善、识别卓越的服务领域并精确定位特别需要加强的领域。此外，他们可以用来证明项目大规模提升的效果并随时间追踪其进展。

“级别”团队希望了解其他人是如何使用“级别”来评估他们的保存库的。一种方法已经被证明在其中一个团队成员机构中是有用的，具体描述见下文。在该机构中，一个大型信息库的提升项目（在这里称为X项目）正在进行中。每一个功能区域都是按照级别1到级别4的顺序来检验的。对于每个单元格，五种值中的其中一个被写入汇总表中：

PASS-意味着已经开始实施这些活动

PASS(在X项目后得到改善)-意味着已经开始实施这些活动，但是在项目X完成之后将有一个更好的实施

PASS（在项目X后）-意味着在项目X完成之后，将实施这些活动

INCOMPLETE-意味着正在实施的活动不能完全令人满意

FAIL-意味着还没开始实施这些活动

每个单元格对应一个值，并且为了使得图案有明显的视觉效果，还为每个单元格涂上了颜色，基于此，汇总表提供了一个强大的可视化功能，不仅能呈现存储库如何与这套最佳实践比较的可视化，还能呈现提升项目在进展过程中效果的可视化。通常，在很大程度上，保存系统的增强是在“幕后”进行的，所以像汇总表这样的可视化功能可以通过沟通组织价值来帮助机构调整成本与工作。数字保存级别可以服务于其它目的的相关信息及级别的实际使用案例将会是有用的反馈。

反馈和未来的工作

正如在本文中提到的，“级别”是在一个协作的环境中开发的，并且仍然是一项进展中的工作。“级别”团队为了提高文档质量而诚邀各种意见和建议。本文和相关的2013文档展示了“级别”团队的机制，进一步传播了“级别”，并邀请该领域的专家进行审查以帮助改进文档。

除了修改“级别”图表，项目团队也在计划其他几种类型的未来工作。根据之前的反馈，团队计划合并图表中使用术语的定义，以及合并计划或执行每个步骤时的可用资源。团队正考虑如何在文档自身中使用（如上所述）级别。团队还希望提供一个在线图表版本，这样可以使用户深入到每个级别的每个单元格，以访问与该主题相关的定义和资源。同样也欢迎关于这些计划的反馈。

编译自：

http://blogs.loc.gov/digitalpreservation/2014/04/protect-your-data-file-fixity-and-data-integrity/

http://digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf?loclr=blogsig

（唐果媛编译，王敬 吴振新校对）

4C 报告：《D4.3—数字保存中的经济决定性因素：质量与可信度》摘译

执行摘要

4C项目组提交的这篇报告是用于分析间接经济决定因素（这些间接经济决定因素在“D4.1——*A prioritised assessment of the indirect economic determinants of digital curation*”已作分析）的一系列案例研究的一部分。这篇报告讨论了基于标准的质量保障（此质量保障是通过实际的审核与认证与可信度联系在一起的）的成本及效益。

目前，在数字保存中用于确保质量而应用有关标准的主要方法集中于处理过程和操作的级别上。数字存档的过程、工作流、信息安全和服务质量的质量评估是自我评估和正式审核的对象。质量与可信度在数字保存中几乎成为了同义词，并一同并入审核可信赖的数字存储库的实践中。

目前至少有五种方法用于评估数字存储库的可信度，从自我评估（如DRAMBORA, DSA）到正式审核及通过外部审核（如TRAC, DIN 31644, ISO 16363）。这些方法主要用于两方面：一、提高数字存储库在利益相关集团中的可信度并最终提高机构组织的名声；二、提高所承担的保存活动的质量。可信度和名声可以通过审核结果（已成为质量评估的同义词）的公共性（透明性）得以提高。

审核的概念已经成为数字存储库中主要内容的一部分，但接受审核的潜在动机是不同的。绝大多数接受审核的机构会基于以下一些因素生成一个商业案例：

- 改善业务流程；
- 满足合同义务；
- 提供一个关于质量与可靠性的公开且易于理解的声明。

现已证实用于数字保存过程中质量保障的成本费用难以衡量或计算，这是由于质量是由多个重叠的成本因素直接或间接决定的，而成本存在于各种长期的活动中。难以通过调查问

卷来获得有关审核与认证的成本数据，这是因为这些需要的数据无法获取或是无法公开。这些问题一旦完全显现出来，我们就会重新考量所使用的方法并关注更为定性的方法。我们采访了一些数字存储库的高级经理和数字保存领域的专家，这些经理知道如何开展某种形式的审核与认证过程。我们也成立了一个专门的团体来与代表资助机构（承担数据归档的责任）的顾问委员会成员深度讨论我们初步的调查结果。

对数据的收集和分析使得成本费用变得明确起来，具体的成本费用在最初的审核工作中只占很少的部分，效益也从为解决审核问题而进行的简单探索和准备工作中开始积累。保存过程的许多方面都涉及到审核，例如记录管理的改进、商业过程分析——与其他理想化地商业成果相交，但却常常无法确定这些工作只是单纯为了审核所做的投资。

1 介绍

数字保存是确保数字对象保持可用性。质量是包括有用性、客观性和完整性的一个术语。在数字保存中，质量保障能提高数字对象随着时间的推移而依然能重新利用的几率。可信赖的数字存储库的使命是：无论现在还是未来都能向指定的社区提供可靠且能长期获取的数字资源。在数字存储库中，信任不仅指信任存储库所使用的保存方法，而且还包括信任广泛的组织问题，如资金基础、政策框架、员工培训、存在可传递的技能等等。一个可信赖的数字存储库可以经受对其系统和组织的威胁和风险。质量与可信度在数字保存中几乎成为了同义词，并一同并入审核可信赖的数字存储库的实践中。这篇报告从成本与效益的角度分析了质量和可信度的问题，用以阐明和详述有关审核与认证数字存储库的商业案例。

1.1 报告的目标

这篇报告是作为4C项目组早期的一篇报告（“D4.1—*A prioritised assessment of the indirect economic determinants of digital curation*”）的后续报告，D4.1这篇报告讨论了作为通用管理工具的间接经济决定因素应用于任何机构中以助于确保可持续性的数字保存。在4C项目组建议的15个作为间接经济决定因素中，可信度排在风险之后位列第二。这篇报告详细介绍了在数字保存中质量保障的经济模型，目的是找出基于标准的质量保障方法的主要成本因素与带来的效益。审核是基于标准的质量保障方法的一个实例。这项任务的成果是展示数字保存的成本如何包括对信任的考量，这是与已存在的成本模型有很大差异的一个地方。我们所采用的方法是通过提供审核与认证之间关系的概括来设置阶段，然后在讨论成本之前解决主要的益处（以及一些问题）。

这篇报告的目标受众是高级存储库经理和存储库的资助者。

1.2 报告的范围

这篇报告的预期讨论范围是做为案例研究报告，研究由经过审核与认证过程而 成为“可信赖的存储库”或类似机构的有关开销、成本、知识投入及最终效益。基于这个目的，这篇报告提供了一个有关计算审核与认证的成本和效益的方法的初步大纲（见第4部分）。同时，这篇报告还包括了一些有关经济可持续性参考模型的建议（见5.2部分）。对于能获得的审核及认证成本相关数据的分析还在进行之中，在4.4部分解释了相关原因，但主要是因为基于ISO 16363 （*Audit and Certification of Trustworthy Digital Repositories*）的存储库认证还没有达到成熟应用的水平（我们期望随着时间的推移及这个项目和报告的发展，在不久以后能达到这一水平）。

国际标准组织（IOS）于2012年2月发布了ISO 16363。对应的ISO/DIS 16919（Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories）草案却延期发布了，尽管PTAB（Primary Trustworthy Digital Repository Authorisation Body，曾与两个ISO附属机构（CASCO和IAF）起草了ISO 16363和ISO/DIS 16919）在对可信赖的数字存储库进行审核与认证过程中扮演了各自的角色。直到这篇报告讨论结束之时，ISO/DIS 16919草案相关的版本才得以完成并发布，没有哪项违背ISO 16363标准的审核能够实施。

我们已经收集、分析了不同的成本信息，这使我们得出了一系列关于成本、效益评估的最高水平的尝试性结论（就像把它们应用于数字存储库和数字保存活动的审核与认证中一样），也能通过一些来自4C项目其他任务和活动阶段的结论来支持一些丰富的交互活动并向路线图工作（WP5）提供有宝贵的投入。

我们访问了高级数字存储库员工、档案学家、研究资助者（即那些负责数据保存的资助者）以及在数字保存领域内的专家（这些专家证实了一些假设并了解我们的结论）。

的确，在这篇报告中，有关数字存储库的审核清单对《RLG/OCLC：*Trusted Digital Repositories: Attributes and Responsibilities* (2002)》做了一些引用。自这篇报告产生开始，结合其持续的适用性有助于表明存储库在发展成为一个小型的符合新信任标准的收集机构之前所经历的过程。该报告的结论部分反映了在一个缓慢发展的社区环境下这些标准在上下文信息中的成熟度，以支持在这些标准所存在的一个相当短的时期内使用它们。然而，需要指出的是，这篇报告没有提供任何有关审核过程的有效性或有用性的评估方法。很明显，必须

审查获得 TDR-status 的相对成本和效益，并应该独立于 TDR 标准所作的评估，这是为了确保实行和维持这些标准的成本与效益之间独立、公正的评估关系。

2 数字保存中质量与可信度的介绍

在机构中存在不同的标准用于保障数字保存过程及产品的质量，在这些标准中，ISO9000质量管理系统标准系列是最著名的。应用这些标准能为机构提供相应的方法来控制保存活动、服务、产品的质量。使用质量标准经常对提高机构声誉有益。

机构活动、服务和产品的质量已经与数字保存社区和TDR中与“信任”的概念相关联。对保存机构活动质量的核查能够提高机构的信誉度，但是信任本身应该通过可量化的标准量具控制来产生。

2002 年， 当 RLG 与 OCLC 发布报告：《*Trusted Digital Repositories: Attributes and Responsibilities*》 时，数字保存被认为是一项复杂的任务，并认为需要高技能的员工、昂贵的计算机基础设施和跨学科的知识。TDR的特性试图定义可持续性的数字存档：它能提供大规模、多样性的数字保存资源，并由长期数字保存授权的国家级组织建立。TDR的特性之一是：按照普遍接受的惯例和标准设计它的系统，以确保对所存储的数字资源的持续管理、获取及安全保障。

ISO16363描述其质量指标如下：

“指标是由经验推导出的，是具有有效性的一致措施。当一起评估时，指标能用于判断存储库整体适用性，确信它已提供一个保存环境，这与OAIS的目标是一致的。分开来说，单个的指标或尺度能用于识别存储库功能的缺陷和面临的衰退。”

通过参照OAIS参考模型（现已作为 ISO 14721:2003 标准模型）的推测性、遵从性的要求，保存活动与服务的质量已成为可信赖的数字保存机构的重要概念之一。对一些利益相关方来说，一个能使其数字对象在很长一段时间内保持可用性的数字保存机构明确能提供有质量的服务，因此，能够继续依靠它来保存数字资源。对于其他的利益相关方来说，保存活动的质量或效率对于存储库信任来说是必要的，而这种质量和效率能通过审核与认证的质量标准来核实。

有时，在数字保存社区，信任既是目标本身，存储库会努力通过认证来提高自身的信誉。这篇报告基于这样一种理解，即数字保存各个环节的质量是信赖的先决条件。如果应用标准的结果能对关心机构信任度的利益相关方公开的话，基于标准的方法对于存储库的设计、运

转和不断改善就是一种正式的、重视质量的方法。因此，TDR审核与认证的做法被当做是基于标准的质量保证的一种方法。

2.1 建立可信度

数字资源长期保存的核心挑战是随着时间的推移，保证数字对象的真实性与可理解性的能力。由于很多因素会导致风险存在，包括：老化的存储媒介、过时的底层系统与软件、技术和组织基础设施的改变、有害或错误的人类活动。如果一个机构能控制足够的技术和组织资源、能根据自身的目标进行保存活动、能通过透明性与证据来论证建立可信度的关键措施，那么它就能被信任并成功承担数字资源的长期保存任务。后面的几个特征至关重要：闭门造车的审核与认证是一个徽章俱乐部，而并不是一个有着常见做法的开放社区。

对数字保存机构信任度的声明是非常容易做到的，但是却很难公正客观地证明。在过去的12年里，数字存储库的工作集中于一系列与OAIS相一致的松散的声明，尽管这与参考模型不符。在十年的时间中，审核与认证成为建立机构可信度的主流方法。这在1996年的报告“Task Force on Archiving of Digital Information (CPA/RLG 1996)”中有相关的行动呼吁：

“数字保存基础设施的关键部分是存在足够多的能存储、迁移及提供数字保存资源访问的可信赖的机构。数字保存机构的认证过程对建立有关数字保存资源的发展愿景的整体信任是非常必要的。”

这项行动由RLG和NARA领导的国际数字保存认证工作小组发起，并于2007年制定发布了数字存储库的认证清单（“Trustworthy Repositories Audit and Certification (TRAC) checklist ”(OCLC/RLG 2007)）。

类似的倡议有德国数字资源长期保存关于专业知识的网络（nestor），是2004年开始建立的一个认证可信赖的保存机构的工作组。基于TRAC核查清单的草案版，nestor小组把工作集中于识别机构的特征及价值，这些与评估现有或计划中的数字对象存储库是相关的。用于审核数字保存机构的Nestor标准(nestor 2006) 于2006年发布，于2008年更新 (nestor 2008)。从nestor项目的总结能得出，可信度标准已转化为德国国家标准，新版的准则已作为一项国家标准：DIN 31644:2012（“*Information und Dokumentation – Kriterien für vertrauenswürdige digitale Langzeitarchive*”）发布。

上面两篇文章为国际工作小组开发一系列新标准用于对数字存储库的完整核查与认证奠定了基础，并促使在2012年形成了一项支持OAIS参考模型的ISO标准：ISO16363：2012（*Audit and certification of trustworthy digital repositories*）。这个工作小组还制定了一个附加

的标准（*Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories* (ISO/DIS 16919)），这项标准会提供一种规范的准则来判断哪一个机构对数字存储库提供审核与认证，并在一个更为宽泛的 ISO 17021 标准框架中，描述了合格性评估的整个审核过程。

TRAC、ISO 16363（TRAC的主要后继者）、DIN 31644等核查清单或指标提供了审核的基础，这些核查清单或指标作为可信赖存储库应该遵守的质量标准。它们为各种规模的数字存储库提供了展示存储库遵守质量与一致性以及尊重数据完整性的方向，并提供对于存储库所管理信息的长期保存及获取的承诺。这些指标主要来源于实践及对OAIS参考模型（ISO 14721）的参考，因此，需要把质量的概念与OAIS模型的原则联系起来。

按照存储库运营所涉及的活动的水平和类型，标准被分成不同的分组，例如（ISO 16363）：

- 组织基础设施，它所强调的问题有：管理与组织的结构、人员配置、过程可说明性、政策框架、财务可持续性、合同、许可证、债务；
- 数字对象管理，即评估所收集的数字内容、创建保存信息包（AIP）、保存计划、AIP的实际保存以及信息（如元数据）与获取的管理；
- 技术基础设施和安全风险管理。

其他两个项目在数字存储库中建立和评估工作质量时遵循自我评估的路线并开发了“柔性”的方法。“Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)(Hofman et al., 2007)”这篇报告提出了把数字保存描述为一种风险管理活动的方法，在这个方法中，数字资源保存者的任务是把那些阻碍数字对象的真实性与可理解性的不确定性和威胁进行理性解释，并把它们转化为可管理的风险。DRAMBORA的线上工具能帮助存储库记录风险并把这些风险作为一种促进信任的透明的方式发布。

荷兰的数据保存与网络服务（DANS）最早提出了DSA（*Data Seal of Approval*）方法，DSA包含了16条指导方针用于帮助数据保存机构建立可信赖的研究数据数字存储库。DSA把保存机构的评估分为两个阶段：一是机构根据DSA指导方针进行自我评估；二是通过一些DSA国际评估团队进行同行评审。评审者会告知理事会保存机构是否遵从了这些指导方针以及DSA印章能否授予给它们。DSA并没有进行实地考察，它依赖于能获取的公开文件信息以及公共性质的自我评估报表，这就导致印章（Seal）成为一种确保同行评估过程中可信度的方法。

除了这五种方法能建立保存机构的可信度，还存在一个三层的审核与认证架构（欧洲框架2010）：

- 当保存机构通过一系列的自我审核与从另外一个机构（早前接受过DSA的审核与认证）得到一份公开的同行评估声明获得DSA认证时，那么这个机构就能获得基本的认证；
- 当已获得基本认证的保存机构经过了一个基于ISO 16363或DIN 31644的结构性的外部审核与公开的自我审核之后，那么这个机构就能获得进一步的认证；
- 当保存机构既获得基本认证，又获得了基于ISO 16363或DIN 31644完整的外部审核与认证时，那么这个机构就能获得最终认证。

在这三个层次中，当前只有DSA的认证是可操作的，到目前为止，有24个机构接受了印章（Seal）的审核。基于ISO 16363的认证只有在ISO/DIS 16919完成及公布时才能开始。德国nestor网络已开始实行一个*"a nestor Seal for Trustworthy Digital Archives"*架构，它会在在成功取得扩展认证之后发布，且是收费的。美国CRL已开始基于TRAC核查清单对数字存储库进行审核，并已经发布四个基于TRAC核查清单的可信的数字存储库凭证。

对于数字存储库的设计与运营的讨论也有其他一些国际标准。ISO/TR 17068:2012 *"Information and documentation — Records management — Trusted third party repository for digital records"* 建立了可靠的数字记录管理所需要的信任标准，并说明了对可信的第三方数字存储库（TTPR）的服务、系统及管理的要求。TTPR被认为是代表授权机构维护数字资源与对保存资源提供认证及公证服务的组织。这份技术报告描述了TTPR的服务、系统及管理要求，TTPR一致性审核与认证不以ISO 17068为依据。

ISO 14641-1:2012电子归档--第一部分：关于数字资源保存信息系统的设计与运营最初是依据法国国家标准，介绍了在保存机构中用于管理文件的数字信息系统实施的方法与技术。它主要介绍了系统设计的标准和运营过程的规范，以确保在保存期限内文件的易读性、完整性与可追溯性。这个标准描述了一种基本方法，用于评估作为第三方的服务提供方遵从标准的要求，列出对服务提供方的标准服务水平协议的内容。

除了这些专门为从事数字保存的机构而设计的标准与自我评估方法之外，遵从其他的被广泛接受、且能保证保存机构的质量的标准对数字存储库也是有益的。例如，质量管理系统的标准系列（ISO 9000 标准系列），它提出了质量保证的各组成要素，并建立了一个对质量持续性管理的系统；信息安全管理体系标准（ISO 27000系列）试图对信息资产进行控制；用于风险管理的ISO 31000标准系列；用于数字资源的ISO 30300管理系统。所有这些管理系

统的标准与所有数字存储库密切相关，但在谈论数字资源的可用性时却缺乏建立可信度的明确目标（在数字保存社区已经进行探讨）或没有考虑长期保存。

反过来也是如此，与数字保存的需求紧密结合的这些标准必然会缩小他们的关注点，从而降低它们通用的适用性。如果信息安全是存储机构使命的关键内容时，ISO 16363最后部分可能更适用于数字保存本身的需求，可能不会获得像ISO 27000系列标准相同的机构信誉度。

当标准出现并被采纳时，最佳实践以及合规证据的约定层次无法立即提供。无论标准的严密性和深度如何，它都会在满足其要求的应用与社区指南中得到发展。理想化的标准总是处在不断地改进的状态中，并通过审核的过程取得管理中一致性的理解。在保存社区中，包括利益相关方和审核人员，通常应在保存早期采用可信度标准，而且需要考虑现有的质量标准。

因此，（当前）建立可信度并不只是满足任何单一标准的要求。对一个保存机构来说，可信度的某些要素或许能通过一项标准来展示，但是正如这项工作的两位作者早前表达的那样：

“当记忆机构保存其他机构产生的资源时，他们仍然需要采用其他社区创建的标准。数字保存是一个具有包容性的领域，就标准本身而言，不能仅依靠他们自身的标准。学着把不同领域的标准拼凑在一起，与识别哪些是适合本地运用的标准，是每位数字保存专家必须掌握的技能。”

此外，我们确信采用标准并不只是名义上的工作，它还需要与质量的概念联系在一起。

2.2 数字保存中质量的信任层级

可信度是一个具有深远的意义的概念，它与存储库的内外部都有很强的联系。首先，存储机构的管理层、员工、资金筹集人与合作伙伴必须对他们的努力能满足正式的要求与期望感到满意。同样，信息生产者、保存者及用户很关心对提供维护、保存和传播服务的保存机构能否获得类似的保证能力。

原来的报告“*Trusted Digital Repositories: Attributes and Responsibilities*”基于以下三方面对建立信任进行了讨论：

1) 保存机构如何获得目标社区的信任；
2) 保存机构如何信任第三方提供者；
3) 用户如何信任由保存机构提供给他们的文件。

对于可信度标准，接下来的工作集中在数字保存机构中开展数字保存活动的操作层面上（如ISO 16363 和 DIN 31644的数字对象管理）。如今，在一个典型的数字存储库的信任网络中，利益相关方可以分为四种类型：

1) 数据生产者：如出版商、记录产生机构、捐赠者，关注数字资源的保存与获取服务的可信度以及数据质量保障。

2) 数据用户：如读者、研究者、机构等等，对来自保存机构的服务与数据的信任度和确保两者的高质量很关注。

3) 保存机构所有者与（或）投资者：如政府、股东及作为数字保存机构一部分的机构等等，常常对保存机构的整个服务的信任度以及确保它们的服务和产品有足够的质量很关注。

4) 数据版权所有者：如由数据保存机构建立的数据集的主题、在数字图书馆中图书的作者、作为公共记录的联合创作人的公民，他们对保存机构及数字保存的服务的信任度以及确保他们的数据得到最高质量标准的管理感兴趣。

尽管在可信赖的存储库的环境下还没有阐述服务的概念，但是把数字保存的可信度当做一种分层概念是非常有用的，这种分层把可信度分为从宏观层面的可信度到微观层面的可信度。灌输着信任度的质量措施或许取决于利益相关方，并且它可能会因为利益相关方的不同而不同。对于数据用户来说，一个可信的、可用的和有内涵的数字对象是熟练的并能直接使用的；从投资者的层面来看，难以承担单个数字资源的质量，在众所周知的开发应用的良好实践中的证据，以及为流程与风险实施的产品管理，反而更流行。因此，质量必须以不同的层次存在,且必须以不同的方式得到证明和认可，所以，质量可以在程序上存在合格或不合格的水平或者达到一个最低标准，它能是一个机器可执行的指标或是一个人为能操作（这符合质量专家的共识）的阀值（带来的技能和角色）。只有这些不同的要素组合在一起才能表明一个机构的可信度。

下面部分展示了信任的三个核心方面：组织、商业（业务）功能以及所保存的资源（数据）。

2.2.1 信任数字保存机构

承担数字资源长期保存的机构应该展示其可持续性和（或）业务连续性。存储机构一般都有几十年（或数百年）的合法经营授权与保存商业活动。数字保存机构并没有长期可追溯性的记录（最老的数字保存机构也没有超过半个世纪的历史），这些机构不仅需要证明它们

的保存能力，还需要证明维持保存服务的能力。这就延伸到了比技术解决方案更广泛的方面，包括更广泛的组织问题：如法律的合法性、政策的充分性、和能力的可用性。在商业领域，企业风险管理和业务的持续性（或应急性）的计划方法经常用于满足利益相关方的持续性的期望。然而，用于上述的标准方法（如用于风险管理的ISO 31000标准系列、用于商业持续管理的ISO 22301 和 ISO 22313标准系列）无法满足保存的具体需求。数字存储库的审核方法从以下几方面详细说明了机构与保存功能和评估机构的可行性之间的联系：

- 反映长期保存承诺的任务或使命声明；
- 继任计划、应急计划和（或）托管安排；
- 关于存储库运转的法律和监管环境的认识；
- 反映及支持保存使命的政策；
- 为实现机构使命而需要明确的职责分工，及为实现这些职责而存在的具有足够技能和丰富经验的职员；
- 有效率的保存政策；
- 致力于支持保存的决策和活动的透明度和问责制;
- 机构的财务可持续性；
- 对合同义务和外包服务的有效管理。

从质量保障的角度来看，这些问题能有助于确定作为一个整体嵌入到机构中的数字保存（商业功能或服务）是如何实现运作的完整性与可持续性。4C项目的一项单独可交付成果（D4.2）正在探索机构内部的可持续性数字保存模型。在保存机构如何保存数字资源及如何融合保存功能与机构其他的商业功能方面有很大的差异。对于评估这些商业模型的适用性，不存在普遍适用的指标；因此，在这里，审核工具就指来源于该存储库的外部利益相关方（或指定的用户群体）的需求。而且这些概念不完全兼容任何分布式存储库的情况，即多个机构始终作为一个组织在一个水平进行商业的运作。在这里，可以把宏观层面的信任当做是顾客与服务提供者之间的一个流程或交互，这种信任是基于给定用户的期望的。在“透明度”在这种交互中带来有利作用的同时时，由于“透明度”能使“信任”更加公开化，因此没有必要让机构将“信任”转化为更高”级别”的综合声誉（因为声誉是基于信任的多种诠释）。

2.2.2 信任数字保存的商业功能

数字对象如果要获得循环价值，就需对其进行管理。目前对数字对象的管理主要在数字存储库中进行，它模仿记忆机构对集中收集的资源中值得保存的资源对象的传统管理方法。

现已提出大量流程描述、模式及图表来支持保存活动与广泛的存储库流程之间的集成，如，DELOS 与 DL.org 项目 (DELOS 2007; DL.org 2011)开发的数字图书馆参考模型；InterPARES 项目的保存链模型(InterPARES 2007)；英国数字保存中心保存生命周期模型（DCC 2009）；但是 2001 年最先发布的开放档案信息系统模型（OAIS，ISO 14721）完成了其作为一个通用参考模型所应发挥的预期作用。TRAC、ISO16363 及 DIN 31644 都直接依赖于 OAIS 的定义、概念及功能，而 DSA 的设计了涵盖 OAIS 的功能实体及角色的 16 条指导方针。

正如 TRAC 核查清单所述：“TRAC 的评价指标应当被用来评判存储库的总体适应性，存储库应提供一个与 OAIS 目标一致的可靠保存环境。TRAC 的个别指标（或度量）可被用来确定存储库功能可能存在的缺点或即将发生的退化。”

审核清单中的指标遵循完全覆盖操作级别的数字生命周期的存档阶段。实际上，这些指标是数字保存流程与运营的质量标准，透过 OAIS 参考模型可以一探究竟，并且要确保存储库顺利实现数字保存的能力。

数字存储库自我评估方法（DRAMBORA, DSA）并不依赖于存储库运营的固定模型，它们的标准被设计用来测量相关活动目的的适应性。恰当的质量水平会在存储库考虑其机构环境、存储内容类型、存储目的及保存的目标客户后由存储库自身决定。支持数字保存的信息安全与技术框架的可靠性在所有审核清单中都有深入的阐述。大部分指标与信息安全标准(如 ISO 27002)相同， 但最重要的是这些指标包含了在非数字保存标准中不重要的方面。

在这个层面的恰当的质量水平可以通过宏观层面的机构环境定义。微观层面（如下）产出的质量有效性可能意味着质量，因此可获得中等程度信任度，但是并不能确保这种信任的准确性。有公共定义的流程的透明度、过程描述及质量评价将支持有效的审核成果和整个保存社区向最佳实践的转变。

2.2.3 信任所保存的数据

用户必须能够信任数字存储库提供的数字对象。存储库的职能是在存储媒介、文件格式或对象的其它类型的表现形式可能发生变化的整个有效的数字保存生命周期内保障维护数字对象的重要属性。TDR 报告（RLG/OCLC 2002）讨论的问题有：用户是如何确定他所接收的文件正是其所需要的、如何确定某个文件正是以前存储在某个数字存储库中的。

最近关注的焦点是建立一个更加清晰的数据起源，以保障经过整合、加工、转换或迁移的数字对象保持它们的完整性。起源信息包含但并不仅有不变性核查。起源性信息越来越多

的被用作衡量数据质量的一个指标，这不仅体现在保存活动连续步骤的细节维护中，而且还体现在所用方法、软件及校准的观念上的细节维护上。这被认为是一个棘手的问题，因为这涉及到源自生命周期预存储阶段的具体流程及值的保留，但这对具有较长使用生命周期的数据来说是极其重要的，如地球观测数据；在哪里实施了复杂的处理过程、什么时候需要识别软件的不同版本；或者为了研究完整性及重现性的目的。

一个通用的质量标准不太可能从用户角度以评估数据质量的精确标准出现，但是这些用户通常与数据有着最为直接的联系，他们也是针对质量做就事论事、数据级别决策的最大群体。确保对数字存储库所采用保存行动的足够的元数据的维护，以提供数据来源审核跟踪方面还有很长的路要走。

2.3 质量与信任度的探索之总结

数字存储库的规模可大可小，它存储着来源于文化遗产、研究、政府或商业机构的许多各种不同的资源；它们有着不同的机构环境、运行状况、技术架构及机构存储库。定义一个用于确定所有不同数字存储库是否能被各种不同的利益相关方信赖的通用标准是一个巨大的挑战，利益相关方当前提出的解决方案是制定一个易于接受的自我评估与审查的普适性质量标准。存储库审核方法的潜在概念是：如果某个存储库符合通用标准的最小集合（尽可能与所有的存储库相关）并能证明具有完成这些指定功能的能力，那么这个存储库就是可信赖的。

一个客观的基准性的机制是判定一个良好数字存储库的前提条件。作为通用参考模型的 OAIS 模型提供了一个标准检查程序，通过参照该模型，可以衡量或比较不同的审核方法。如果审核结果被普遍承认，那么这样的通用参考基准就会变得至关重要。这些开发中的存储库审核工具面临的主要挑战是要提供一个相当真实的体系，用于判断哪里具备一致与成功。机构已经对把质量标准分成几个层级逐渐达到一个共识：机构生存能力、数字存储库活动、技术框架与数据起源。然而对审核过程本身的透明性没有达成共识。

数字保存社区的组成机构各不相同——存储机构与研究机构、政府机构、商业机构以及服务提供商共同组成了数字保存社区。保存社区已经制定了许多用于工作流控制的标准，但是不可能完全统一整个社区所有的保存工作流。数字保存社区商定的标准类型（可被称为自愿遵从性标准）主要适用于改善工作流程（在广泛的意义上）。然而，很难强制使用统一的推荐性标准。但是没有必要期望所有事物的一致性，特别是如果数字保存活动的主要目的是

确保所保存的资源在未来的某些定义点上的重复使用，而未来的保存定义及需求的特殊目的则会因机构的不同而不同。

数字保存作为一个领域，在制定标准方面的成熟是一个持续进行的过程，在此过程中采用了审核标准并发现了审核标准在社区活动中的作用。OAIS 标准提供了一个通用的参考模型（但不是一个完整的流程或技术实现标准），存储机构可以以此为基准管理它们的活动。在此之后，一些以信任度为导向的评价标准正与同行评议及审核流程相继制定。随着标准被更多机构接受，进而发展到顶峰，各种不同的利益相关者在如何应用这些标准来适应他们的时间方面都将占有一席之地。

不论是经过五年发展周期的 ISO16363 标准，还是那些被很快开发的非正式的方法，这些第一代审核及认证标准都不断的发展。随着时间的推移，审核实践将给每个标准的指标分配适当的值，同时开发一致性证据的通用方法也将不断发展。这些通用方法本身的特性将支持数字保存实践方法的一致性，反过来也可以降低认证的成本。

这些不同的信任度标准与上面所列出的更广泛的标准框架共存。随着时间的推移，提供资源的网站与其它受审核方将为信任度指标的应用提供更有效的指导。这个过程也会识别社区相关的反应，如标准在多大程度上满足了它们的需求、什么级别的认证是合适的、如何比较信任标准、如何与广泛的质量评估标准交互等等。

机构应用标准的成熟度曲线从测试与基准（benchmarking）开始，然后从风险管理移向质量管理，最后在一个具有学习能力的机构中达到最高点。数字保存的标准仍处于曲线的开始位置，它关注的重点是保存任务执行情况的基准并开始着眼于保存的风险管理。一个成熟并且能够适应改变的机构，它更关注流程的效率，而不是产品（系统）控制及互操作能力。在从质量控制转向质量保证再转向质量管理的过程中时，所面临的最大的挑战是人员与技能的管理而不是技术与工作流程。要重塑一个机构，最关键的是员工，而非所使用的技术。

相同的成熟度曲线似乎也影响了更好地理解这些活动成本。因为经费缺乏，实施数字保存的大部分机构乐于引进这些信任度的方法，并将其作为一种研究活动而不仅仅把它作为一种信任度活动。显然，随着数字保存变得越来越主流，会更加直接地识别对培养信任度及提高质量的成本费用，同时也能将它们与标准的机构的成本费用区分开。

最后，我们需要再次重复：对保存机构活动质量的正式核查会提高一个存储库的声誉，但是信任度本身应通过可测量的标准化控制产生。声誉似乎是信任度的代名词，但是两者的

关系难以区分。机构可借助其它因素来提高声誉，这些因素通常与质量无关并且通常是象征性的；信任度仅能通过基于证据的、独立的审核活动来完全获得。

3 质量、信任度及其带来的利益

对数字存储库质量保障的投资不仅有助于数字保存服务，也有助于该组织的整体发展（参见 D4.1）。运用质量管理原则需要以客户为中心，提升高级管理人员的积极性和参与度，并支持处理方法及持续改进（ISO 2012）。引入质量管理并进行评估或审核所产生的副产品文档能促进意向的沟通、增强行动的一致性和可追溯性，该文档还能增强业务的连续性，特别是在由于工作人员更替所导致知识转移传递时。

通过在活动中效仿最佳做法并采用标准技术，可以实现或改进服务质量。然而，为了使质量能被测量，需要依据既定的标准进行评估。评估或审核过程的结果可以通过认证来传达及验证。组织申请认证的原因很多，如认证可以：

- 成为合同性或监管规定；
- 需要用来满足客户需要；
- 在风险管理计划的范围内；
- 通过为其管理系统发展设置一个明确的目标来激励员工。

在现实中使用数字存储库审核方法有两个主要驱动因素：第一，可以在其利益相关者群体提升资源库可信度，并使机构最终享有较高声誉；第二，审核过程提升了质量。审核结果（即质量评估）的公布（透明化）也被认为提高了机构的信任度和声誉。

从质量和信任度中所得益处— — 结论

10 年前，ERPANET 项目讨论了进行数字存储库审核的原因，并预测他们在不久的将来会有明确的需求（ERPANET2004）：

“那么，什么是数字资源长期保存审计的价值呢？传统保存记忆组织的长期经验，以及他们的责任感让他们愿意信任彼此，因此很少进行审计。那么，如果有些事情发生了改变，又会如何呢？事实上，数字资源比传统基于纸质的文件脆弱得多，有可能是新的数字顺序使人们和组织对其使用的有关技术感到不安，人们怀疑其方法和程序是否充足和有效，以及保存组织是否可以保证他们所负责的数字对象的真实性和长期性。”

现在审核这个概念逐渐成为数字存储库的主流观点，而且进行审核的潜在原因有很多。大多数从事审核工作的机构会进行以下工作：

- 改进工作流程;
- 满足合同义务;
- 提供关于质量和可靠性的公开合理声明。

4 质量成本

4.1 介绍

数字资源长期保存成本模型并没有明确包括与信任度或质量有关的活动成本。4C 项目组数字资源长期保存成本模型（D3.1 成本模型的评价及需求与差距分析）的差距性分析表明：成本模型缺乏衡量数字保存活动及服务质量的能力会容易形成一种“显著差距”，而通过审核和认证可以弥补这个差距。

唯一已知包含对信任成本的隐式理解的成本模型是数字存档成本模型（CMDA），这个模型假设所审核的组织拥有符合数据认证标准(DSA)所列的 16 条准则的值得信任的数字存储库（TDR）。

在最近的两份报告中，APARSEN 项目基于 ISO 16363 标准，提出了数字资源长期保存成本模型的基准调查（APARSEN 2013A 和 2013B）。该报告将该模型涉及的活动映射到标准规范的活动中，即映射为三个主要部分——“组织架构”、“数字对象管理”、“基础设施与安全风险管理”，以及各部分的子部分。该报告得出结论，确保质量效益所需成本还是一个未定义的领域，仍需要进一步的研究（APARSEN2013B，第 34 页）：

“该成本模型仅仅描述活动与成本的关系，并没有覆盖全部业务情况，如范围、约束、假设、好处益处和成功措施、质量管理、方案评估、资金和风险价值。在成本模型发展过程中，假设保存基本原理和获益将被独立开发，而且蕴含在成本模型的活动中。”

对工作和服务的质量改进往往与较高的成本相关，或者至少与重大的启动投资有关。在数字存储库环境中，这种投资主要依靠因此带来的潜在好处（例如，得到利益相关者的更高信任以及其它间接引发长期投资的结果）。由于价格标签或 ROI（投资回报率）会随组织及其业务情况变化而变化，所以把它们应用于基于标准的质量测量中，基本是不可能实现的。然而，在 4C 项目中，对于质量保障这个长期过程，其第一阶段的经济因素识别应开始于审核和认证费用的计算。

4.2 数据收集

设计一个简短的问卷来记录存储库审核和认证成本的信息。该问卷设计得越通用（即非“标准”制定）越好、越细致（即把成本与“低水平”活动联系在一起）越好。该问卷是从一个客观角度设计的，它不是对已知的或现有的数据进行审查，而是对最有可能加入成本模型中的关键活动进行检查：

问卷的第一部分旨在记录审核的组织环境，并了解可能影响这项工作成本的关键因素。问卷的第二部分试图收集开展和实施审核过程的时间的信息以及这些活动直接成本的信息。问卷最后部分是关于事后审核处理和与维护所获得证书的成本相关的问题。

在 4C 项目合作伙伴组织的问卷调查测试中，明显能发现收集该调查所需的审核成本信息是非常耗时的，并且涉及组织内部的几位专家。尽管如此，还是决定联系大约四十家已成为 4C 任务成员的机构（因为他们已经经历过审核），并要求他们参加调查。所有作出回应（只有 7 家）的机构认为，该问卷提问的形式是正确的，也试图挖掘正确的信息，但时间是有限的，并且没提供一些他们觉得对组织作出建设性比较的必要详细信息。大部分受访者表示可能需要两到三个月来收集所有必要的数据。这超出了本报告的交付期限，所以我们将无法进行实证分析。

一旦这些问题完全显现出来，在后期的过程中，我们需要重新考虑我们的方法，并侧重于开发更为定性的方法。我们采访了一批高级数字存储库管理者，他们都已经经历了审核和认证过程，并且是数字保存领域的专家。该调查表可作为采访的模板。我们还成立了一个专题小组，与代表资助机构的咨询委员会委员（负责数据档案）一起深度讨论初步调查结果。讨论结果将在下节阐述。我们的审核成本研究采用了定性数据，并论证了一些我们的结论。

4.3 审核和认证成本指标

收集审核成本数据的调查问卷反映了开展审核和授予证书过程中的主要成本指标。基于这种结构，下面主要讨论的领域是从审核及认证是如何影响被审计机构的经济的角度顺序展开的。

4.3.1 组织的类型和范围

无论是国家档案馆、国家图书馆、多学科资源库、学科基础信息库、机构库、商业服务供应商，组织的类型将在很多方面影响审核和认证的成本。组织的类型可能是进行审核和认证活动的主要动力。因组织的法律、规章制度或授权不同，可能需要在固定时间间隔内进行审核。在更大的机构内，数字存储库的规模也可能影响审核费用。

在对创新障碍低的较小规模的组织或在有某种形式合同以执行认证程序的组织中，审核和认证活动更容易被采用。例如，数据生产者机构需要具有某种形式的安全授权，有时还需要用法律来进行某种形式（信息）安全的认证。相反，有法律授权的保存记录组织通常不需要进行任何形式的信任认证。

很多参与讨论这些问题的机构也证实了采用多种自我评估/审核过程的工作对后续活动的成本有显著影响。ICPSR 于 2005 年 6 月运用 TRAC 清单进行过自我评估，并于 2009 年用 DSA 的自我评价，它发现这次评估成本低于上次。不只是总成本的降低，而且证据的准备与后续活动的比例完全颠倒了。

4.3.2 审核/认证的类型

上述审核和认证的三个层次（见第 2.1 节和 3.2 节）——自我评估作为基础认证、经核实或同行评议的自我评估作为扩展认证、基于外部专家审核的正式认证，它们都有不同级别的相关成本。

尽管自我评估可能需要外部专家协助讨论，但它通常作为一项内部工作来执行。自我评估的结果往往可以记录并可以通过利用传统的内部手段和决策机制得以实现。一旦确认要将自我评估结果进行外部验证，组织将花费更多的精力记录评估及准备文档来规范机构存储库工作。确保政策和活动透明度的成本将随认证级别的提高一起增长。

正式审核的成本将包括聘请外部审核人员进行评估的成本，在某些情况下，也包括证书颁发的费用。在某些情况下，也会聘请拥有审核经验的顾问来为审核过程做准备，这可能对成本产生显著的影响（顾问一天的费用可能比内部员工一天的费用多，但由于他们经验丰富，通常效率更高。但是，内部人员一般都是固定的机构费用，而顾问更可能是一笔额外的费用。）。

重复审核或更新认证的花费通常比第一次评估要低得多。换句话说，组织的准备、以往的评估和认证经历可以显著降低成本。

4.3.3 审核的范围

建立数字存储库的质量管理体系和可信度的审核工作涉及整个组织。DIN 31644 明确指出：评估涵盖了组织和技术两方面的问题，它不可能仅仅评估数字存档的一部分（例如，只有归档存储）(nestor 2013, p. 3)。对于自我评估、风险评估或安全审核，可能要确定一个较窄的范围，例如，单个储存库功能（如摄取）、单个组件系统（如存储和备份解决方案），或

是一项特殊的服务（如电子期刊的访问服务）。容易理解的是，较窄的范围节省了参与评估和准备文档的工作人员的时间，但可能无法揭示在全局或整个工作流程中存在的问题。

4.3.4 时间和货币成本

根据 APARSEN 项目在 TDR 的评估中所作出的努力，成本活动可以细分为以下主要方面：

- 准备证据
- 评估与标准的一致性程度
- 如果没有，则需创建新证据
- 改变（预审核）不符合标准的政策和程序

在调查问卷中，对审核费用进行了更细致的分类：

- 资源：
  - 内部职员
  - 外部职员（顾问）
  - 设备
  - 硬件
  - 软件
- 准备审核阶段，职员需要参与的活动：
  - 证据/文件的准备
  - 创建资产清单
  - 与员工面谈
  - 评估与标准的一致性程度
  - 如果没有，则需创建新证据
  - 风险评估
  - 改变（预审核）不符合标准的政策和程序
  - 执行政策和程序
- 在实际审核认证的过程中，职员需要参与的活动：
  - 分析证据与产生证据
  - 面谈和讨论
  - 现场审核及认证的支持

- 实地考察后的后续活动

- 审核/认证后的活动：
  - 调整政策和实践，包括为审核而维护质量管理系统
  - 维护和更新证据基础和文档以支持后续审核
  - 任何额外费用

这些费用类别还不知应该归于审核或应用标准的哪种类型。结合审核工作的种类和范围，它们应该提供一种适当的工具来规划和预算审核工作。组织的范围：工作人员的总数、法律要求、指定的用户社区都会影响审核成本的级别以及基于成本的生产活动能力。来自记录管理的支持程度及机构其它的支持活动或基础设施服务都能对审核费用产生显著影响。

4.4 审核与认证成本的评估

分析有效的审核与认证的成本数据仍然是一项处于进展中的工作，有三个原因使作者还不能达到他们分析的精确程度。第一个原因是缺乏公开可用的成本信息；第二个原因是缺乏有效成本信息之间的比较，这主要是由于获取到的一些敏感性的信息已被证明不可能在我们最初所期望的粒度级别上呈现成本信息。第三个原因是截至撰写该报告时，ISO 16363 标准仍然没有达到我们所期望的应用成熟度。因此，作为 FP7 APARSEN 项目一部分执行的 ISO 16363 审核，该标准实现的成本与效益的证据完全是作为测试审核，它关注的是审核人员处理过程的测试部署而不是被审核方视角下的测试部署，此时，活动成本的捕获就不是测试关注的重点。

我们的问卷调查表试图收集实施及完成审核过程所需的时间以及这些活动的直接成本的信息。

4.4.1 与资源相关的成本

第一个成本是采购机构审计所需的审核标准。在某些情况下（如 ISO 27000，ISO 9000），付费的工具集在指导及支持审核与自我评估的准备过程中是有效的。雇佣外部专家及咨询人员需要的费用通常比购买用户与指导手册所花费用高的多，尽管这些专家可以为机构提供专门定制的建议。也有针对促进 DRAMBORA、26 DSA27 及 TRAC.28 进行自我评估的在线工具集。

受访的存储库（只有一个例外）对新设备的投资看得没那么重要。对于信息安全的审核来说，为了物理安全，必须购买一些与安全相关的额外软件或设备。某个存储库报告称：为

了满足 OAIS 参考模型的要求而开发存储库管理软件是一项重要成本，然而，即使开发出的功能已存在于产品研发路线图上，简单地加快开发也会即时地产生了相关的开发成本。用于共同创造与分享文档的平台与工具（如 wikis）虽然可以免费获得，但也会产生人员开发成本。

受访的机构都没有雇佣新员工直接支持审核过程，但是一些专职的员工可以花费大部分的工作时间（50%-100%）主持审核项目。这种取代现有员工正常工作职责的做法使总成本的计算更加复杂化。某个机构只有在它们的自我评估与 TRAC 标准不符之后才雇佣一个数字保存主管。

4.4.2 员工成本

与审核相关的主要费用是员工成本，而其中还依赖于使所收集的数据具有可比性的难度。不同国家的员工成本差别很大，所需员工的角色可能涵盖了高级管理人员及行政管理人员，很难获得一个清晰的答案。同一机构做着相同工作的两个员工，可能会因工作时间的不同而其工资也会相差超过 10%。

审核期（包括准备阶段）是一个相当长的时间周期。员工参与审核阶段的时间长短不一，平均共计 2-3 个月。下面给出了一些已完成的自我评估与审核的可信存储库例子。这些给人深刻印象的数据没有考虑任何早期的自我评估或审核，而且由于面谈，因此不可能保证数据的精确性，如当存储库经理在超过两个小时的访谈中被两次问到对某一活动的看法时，就出现了误差（如“三个星期”变成了“我做了三个月，其他三人又做了三个月”。）

机构	审核类型	过程持续时间	员工总工作量
Portico	TRAC 认证	10 个月	
HathiTrust	TRAC 认证	14 个月	
Chronopolis	TRAC 认证	14 个月	
Scolars Portal	TRAC 认证	6 个月	
DANS	ISO 16363 试点审核		3 PM
CINES	ISO 16363 试点审核		3 PM
UKDA	ISO 16363 试点审核		2 PM
DNB	DIN 31644 试点审核		1.5 PM

表 1——审核工作与持续时间

UKDA 最初为 ISO 16363 实验性（或测试）审计提交的报告称：

“两个核心员工负责流程，而负责主要相关领域的员工负责标准的审查。在审核过程中，我们估计在外部审计团队到来之前，将花费六十个人工工作日，之后还需要十个人工工作日。准备工作包括以下情形：档案馆持续进行的定期检查、控制文件的更新以及即将到来的针对 ISO 27001 审计所需材料的修订，但这也许是一个较高的估计。然而，如果没有前期针对 ISO 27001 的认证或一些其它包含存档细节风险评估的审计流程，我们估计该流程会更加耗时。”

而且，UKDA 已经针对 DSA 进行了自我评估。德国国家图书馆总结的关于 DIN31644 实验性审核的员工费用如下：“DNB 的几名工作人员在测试审核中的工作强度各不相同。DNB 没有一个专门的数字保存单位，但是从事数字保存工作的员工分布在 IT 部门的一些部门里。认证过程的准备大多由一个单位进行协调与执行。其它单位的员工则以参加会议的形式评论及讨论，并将信息提交给外部审核（加上一些准备及重复的时间）。部门的负责人监督这一过程。总共花费了 212.5 个小时用于准备及实行审核。这相当于 1.51 个人花费数月的时间”

4.4.3 认证的成本费用

许多认证会产生颁发证书的管理费用，或也包括外部审核人员固定时间内实地视察所产生的费用。可信数字存储库的自我评估(DSA, DRAMBORA)则不产生相关的费用。可信数字存档 nestor 印章目前花费了 500 欧元。这包含自我评估结果的复查及使用 nestor Seal（如，在一个机构的网站上）宣传的权限。

发布 TRAC 证书的费用由 CRL 的认证顾问小组收取，顾问小组“保证认证过程兼顾整个 CRL 社区（包括馆藏建设、保存及信息技术的领导者）的利益”。

依据 ISO 16363 与 DIN 31644 的外部审核也将以收费为基础，但是目前还没有可用的做法。

基于 ISO 9001 成本质量管理系统与 ISO 27001 信息安全管理系统的认证费用通常取决于外部审核人员现场审核所花费的天数，但是对于中型规模的机构来说，可能要花费 50,000 欧元以内。

4.5 审核与认证的成本费用：金钱的价值

我们的数据收集与分析明确表明详细的成本几乎不是初始审核的工作内容，效益的产生来源于解决审核问题时简单的探索与准备。涉及到审核的许多流程（如记录管理改进、商业

流程分析，与其它期望的商业产出发生了重叠），通常不能把这种工作单纯地确定为审核投资。

实践经验以及与存储库管理人员的讨论说明每项活动都有着不同程度的费用成本，这主要由以下因素决定：

- 应用的标准
- 机构的成熟度
- 机构的规模与结构
- 政策的控制以及正式的管理结构的级别
- 第三方的信任度
- 外包的程度
- 所需机构之间的交互性水平
- 合同义务
- 应用标准的基本原理

5 总结

该报告是第四阶段的工作：间接经济决定因素（在“D4.1——*A prioritised assessment of the indirect economic determinants of digital curation*”中已作分析）中案例研究的一部分。这份报告讨论了审核和认证实践过程中与信任度有关并基于标准质量保证的成本和效益。两份后续报告将关注作为成本决定因素的风险评估与商业模型（关注即将发布的 D4.4 与 D4.5）。

随着数字保存逐渐发展为一门学科，对该领域标准应用的关注已经从产品质量转向保存流程的质量，并开始逐渐地向员工能力及技能的质量转变。一般而言，1994 年前后，数字保存的发展主要关注产品、软件与解决方案质量的评估，效率及与设备相关的成本一般说来是可直接计算的，如，可以计划与预算技术成本。2014 年，数字保存领域关注的是数字存档流程、工作流、信息安全与服务方面的质量评估。质量是多维度的或者说是非常具有机构特殊性的，如它依赖机构的商业模式以及数字保存是如何嵌入到机构的核心业务中的。流程质量保证的相关投资，即使是按照定义，也很难计算和检测，这是由于质量是直接或间接地由多重成本因素决定的。本报告仅选取了长期质量保证流程中的一个成本因素——通过审核进行质量评估及通过认证给予数字存储库可信赖的地位——并且已经证明从相关活动中直接抽取出成本的困难程度。

随着数字保存中标准应用成熟度的不断提高，数字保存未来 10 年的发展将会与技能质量及存储库员工的能力评估相关。2014 年的能力级别仍处于一个抽象的水平，并且在数字保存领域的定义是相当不明确的，作为数字保存领域的一个质量保证方法，在未来几年内它会有更好的发展势头。

2.3 节指出了质量标准制定与通过审核标准的定期检查及更新而进行数字保存控制的模式。与建立领域质量有关的标准包括通过建立的一个正式的框架来维护不同审核之间的质量（由于其固有的性质，只能评估某个时间段内捕捉到的状态）。国际标准化组织介绍了“管理系统标准”（MSS）的概念，它提供了一个在设置与运营一个管理系统时可遵守的模型。

MSS 在许多期望持续提高质量的领域都有定义，如质量管理、信息安全、风险管理、记录管理、环境管理等。数字保存已经能够从应用这些方法的机构中获益，因为这些方法间接地提高了质量并增加了数字保存过程的透明度。然而，这些方法将不可能也不会像当前可信赖的数字存储库审核清单那样对数字保存活动提供直接的控制与基准。对于提供确保持续的质量提高机制的数字存储库审核标准来说，如果未来要对它进行正式的认证与信任，就需要建立一个管理系统。数字存储库基于标准的现有质量保证的差距如下图 1 所示，它比较了存储库审核与信息安全管理的审核流程：

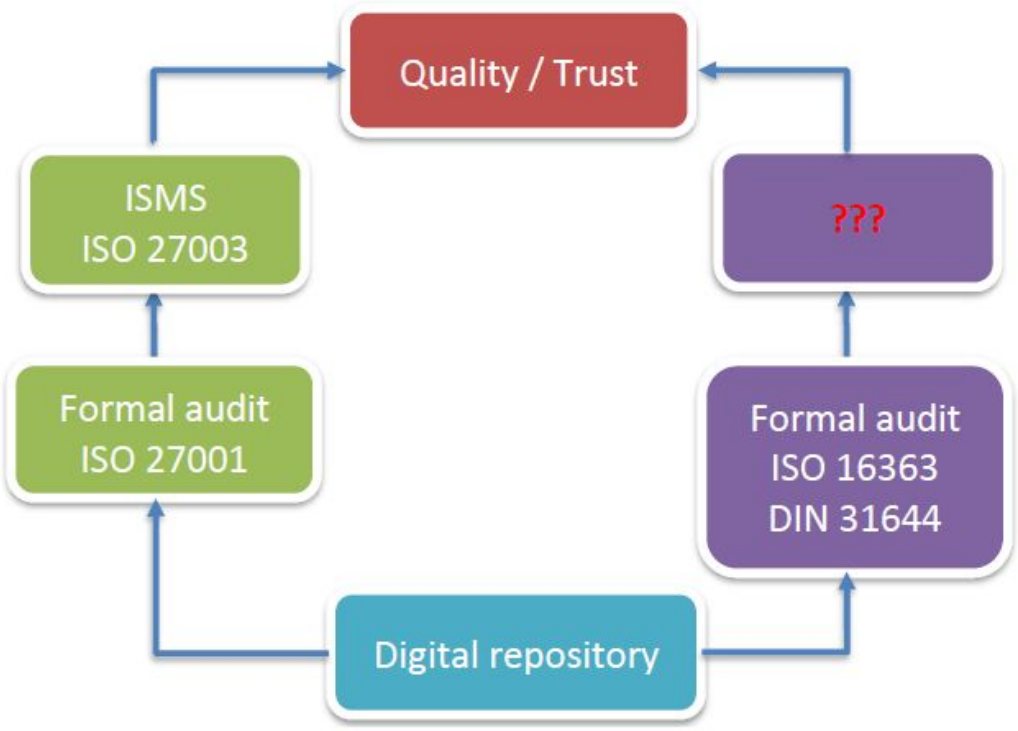


图 1 数字存储库审核标准与信息安全管理系统标准的比较

在数字保存中，定义一个质量管理系统将允许通过流程或服务的水平控制来更好地管理保存活动的成本，并且提供了一种机制来保证保存活动投资的可持续性。

来自于数字保存社区、关于审核与寻求认证证明的效益的报告在现阶段的主要动机是非常务实的，它提高了声誉并且获得了市场上的竞争性优势。通过提高效率及工作质量而获得经济利益不是存储库管理者的主要目的。

编译自：

http://www.4cproject.eu/community-resources/outputs-and-deliverables/d4-3-quality-and-trustworthiness-as-economic-determinants-in-digital-curation

（多人翻译，多人校对）

【技术与工具】

由 ARC 到 WARC 的可扩展性转化的几点思考

SCAPE 项目正在制定侧重长期保存的大型数据集处理解决方案。应用领域之一就是网络信息存档，长期保存直接关系到不同的任务领域，如信息收割、存储与获取。

网络存档信息常常包括数 TB 大小的大型数据，最大型的是网页存档，其依据自身情况存储了3640亿个网页，占了大约有10PB 的存储量。IIPC（International Internet Preservation Consortium）及其全球40多个成员展示了各个机构之间不同的关注焦点，每个机构都把焦点集中在他们感兴趣的不同的内容（类型）材料上。

这就取决于国际社区联盟及个体成员机构能确保网络存档内容在将来能正确地获取及展示。做好这项工作的挑战在于处理好网络内容的本质，就像是网络自身一样：使用不同的格式发布文本及多媒体信息、使用丰富多样的标准和编程语言、利用数据库驱动的网站及强烈相互关联的功能提供交互式的用户体验、从多源头获取社交媒体信息等等。这就是说除了庞大的数据规模，数据还具有异构性和复杂性，这就向数据保存者、开发者及长期保存专家提出了重大的挑战。

IIPC 解决上述问题的方案主题之一是如何对网络存档内容进行真正的长期保存。最初，信息内容是由网络存档以 ARC 格式存储，这种格式是为了把多个网络资源聚合在一起，它能选择性地压缩为一个容器文件。但是这种格式并不是用于长期保存最理想的格式，例如，它缺乏支持以标准格式添加上下文信息的特性。基于此，产生了作为一个 ISO 标准的 WARC

格式，它提供了许多额外的特性，特别是它能以自包含的方式把获取的信息与其相关的元数据保存下来。

网络存档的特征之一是一边的信息内容处在不断变化时，另一边的内容却保持静止不变。为了保存这些变化，利用爬虫工具按一定的频率获取网页。如果每次访问都存储同样的内容，那么就会产生冗余的信息，从而不能有效地利用存档空间。由于这个原因，Netarchive Suite（最初由皇家图书馆及国家和大学图书馆发起，后来也得到了其他图书馆的使用）提供了一种称之为“重复数据删除”的机制，用以检测已取回的信息内容并因此引用现有的有效负载内容。被引用的信息内容实际上是从爬虫日志文件中获得的，这就意味着一旦爬虫日志文件不在了，就无法获得任何引用内容的信息。例如，为了显示有多种图像的单个网页，wayback machine 需要了解从哪里找到可能分散在各种 ARC 容器文件中的信息。一个索引文件，如 CDX 格式的索引文档，要包括必要的信息，为了建立这个索引，需要在索引构造过程中包括 ARC 文件与爬虫日志文件。

从长期保存的角度来看，这种依赖是有问题的。ARC 容器文档不是自描述的，它们以一种非标准化的方式依赖于执行性数据（如由爬虫软件生成的日志文件）。网络存档运营者与开发者知道从哪里获取这些信息，而且这种依赖性需要提供很好的记录文档。但这存在着丢失信息的风险，这些信息对显示及访问内容来说是必要的。这是许多机构考虑把 ARC 格式的文档转化为 WARC 格式文档的原因之一。但是，把最初看似简单的格式转换迅速变成一个有复杂需求的项目并不容易实现。

在 SCAPE 项目中，在我们看来有几个值得关注的地方：

1) 从 ARC 格式到 WARC 格式的转换特别适用于处理大型数据集，因此解决方案必须能提供一个高效、可靠且可扩展的转换过程。这就要求必须有扩展的能力，即意味着它应该增加处理能力（通过使用一个适当大小的计算集群,使机构在给定的时间内完成文档格式转换）。
2) 读、写大型数据集需要很大的成本。有时，数据必须先转移到(远程)集群。因此，它应该有可能方便地附着在用于从内容中提取额外的元数据的其他进程中。
3) 从一种文档格式转换为另一种文档格式有信息丢失的风险。如计算负载散列以及对对应的 ARC 与 WARC 实例进行内容比较，或者在 Wayback machine 中对转换内容的子集进行呈现测试等，这些质量保证措施在这一方面来说都是可行的方法。
4) 解决 ARC 容器文件对任何外部信息实体的依赖是一个必然要求。因此解决方法不

能只观察 ARC 与 WARC 之间一对一的映射，还应该包括在转换过程中的上下文信息。

关于这个活动的首要步骤是为上面提到的第一个方面找出正确的方法。

在 SCAPE 项目中，Hadoop 框架是被称为 SCAPE 的平台的重要组成部分。Hadoop 的核心职责是把处理任务有效的分配给计算集群中可利用的单元。

利用 SCAPE 项目的软件开发成果，可以在实施一个解决方案时有不同的选择。第一种选择是利用 SCAPE 平台的一个模块——ToMaR，该模块是一个可很容易地在计算集群中实现简单分布命令行应用处理的 Map/Reduce java 应用，如下述：ARC2WARC-TOMAR。第二种选择是在 ARC 格式的定制阅读器和 WARC 格式的定制写入器中使用 Map/Reduce 应用，因此 Hadoop 框架能够直接解决这些网络存档的文件格式问题（如下述：ARC2WARC-HDP）。

通过一个实验来测试这两种方法的表现，主要问题是与使用潜在命令行的执行工具 ToMaR 相比，本地的 Map/Reduce 实践表现是否有一个显著的性能优势。

这个优势之所以应该“显著”的原因是选择 ARC2WARC-HDP 有一个重大的局限性：为了达到基于本地 Map/Reduce 实现的转换，就需要使用一个网络存档记录的 Hadoop 表现方式，这是读出 ARC 文档记录与写入 WARC 文档记录之间的中介表现方式。当使用字节数组字段来存储网络存档资源的负载内容时，由于字节数组的整型长度与整型值相近，所以大约达到2GB 时就有理论上的限制。实际上，若取决于硬件设置或集群配置，负载内容大小的限制可能会低得多。

这种限制会随着需要大型负荷内容记录的替代解决方案出现。“小型”记录与“大型”记录之间的差别可能会增加应用的复杂度，特别是在转换过程中不同容器文档之间需要涉及到上下文信息的时候。

用于执行转换的实现是一种概念证明性的工具，意味着在这个阶段不打算把它们用于产品转换。这就意味着有以下限制：

1) 前面提到过，由于使用内存呈现网络存档记录，ARC2WARC-HDP 存在文件大小的限制，在实验中用到的数据集中最大的 ARC 文档大概有300MB，因此，记录负载的内容可以很容易以字节数组字段来存储。
2) 虽然异常情况会被捕捉和记录下来，但是没有收集处理错误或任何分析结果。由于把焦点集中在评估性能上，因此没有考虑任何具体的记录处理细节。
3) 当前的实现既没有质量保障也不包含上下文信息，而这些在 ARC 格式向 WARC 格

式转化时是被认为是很重要的方面。

实现的基础是用于读取网络存档 ARC 容器文档及遍历记录的 Java 网络存档工具包

作为一个用于读取数据的处理范例，这种实现包括了 Apache Tika（作为一种选择的方式来识别负荷内容）。因此，所有 Hadoop 工作的执行都会测试识别的负荷内容。

上面已经谈到，ARC2WARC-HDP 应用是以一个以如下命令行开始的 Map/Reduce 应用实现的：

hadoop jar arc2warc-migration-hdp-1.0-jar-with-dependencies.jar \-i hdfs:///user/input/directory -o hdfs:///user/output/directory

ARC2WARC-TOMAR 工作流使用一条 Java 实现命令行，并由 ToMaR 执行。一个 bash 脚本用来准备 ToMaR 所需的输入，另一个 bash 脚本用于执行 ToMaR 的 Hadoop 工作，这种工作流的组合实现就是“Taverna”工作流。

在 ToMaR Hadoop 过程中，行动开始需要有所谓的“工具规格”，它详细说明了输入与产出，这需要执行 java 命令：

```
<?xml version="1.0" encoding="utf-8" ?>
<tool xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://scape-project.eu/tool tool-1.0_draft.xsd"
    xmlns="http://scape-project.eu/tool"
    xmlns:xlink="http://www.w3.org/1999/xlink" schemaVersion="1.0" name="bash">
  <operations>
    <operation name="migrate">
      <description>ARC to WARC migration using arc2warc-migration-cli</description>
      <command>
        java -jar /usr/local/java/arc2warc-migration-cli-1.0-jar-with-dependencies.jar -i ${input} -o ${output}
      </command>
      <inputs>
        <input name="input" required="true">
          <description>Reference to input file</description>
        </input>
```

```
        </inputs>
        <outputs>
          <output name="output" required="true">
            <description>Reference to output file</description>
          </output>
        </outputs>
      </operation>
    </operations>
</tool>
```

所有命令允许使用“-p”标记，以便 Apache Tika 对负荷内容进行确认。

在实验中使用到的聚簇有一个控制器（Master）以及五个工作器（Slaves）。主节点有两个2.40GHz 频率及24GB 内存的四核 CPU（8物理/16超线程核心）。从属节点有一个2.53GHz 频率及16GB 内存的四核 CPU（4物理/8超线程核心）。关于 Hadoop 的配置，每个机器有五个处理核用于 Map 任务，两个处理核用于 Reduce 任务，一个处理核保留给操作系统。总共有25个处理核用于 Map 任务，10个处理核用于 Reduce 任务。

实验在执行的过程中使用了不同规格的数据集，一个使用了1000个 ARC 文档（9158GB），一个使用了4924个 ARC 文档（44547GB），最终的实验使用了9856个 ARC 文档（89021GB）。

结论总结如下表：

		Obj./hour (num)	***Throughput (GB/min)***	***Avg.time/item (s)***
	Baseline	*834*	*1,2727*	*4,32*
Map/Reduce	*1000 ARC files*	4592	7,0089	0,78
	4924 ARC files	4645	7,0042	0,77
ToMaR	*1000 ARC files*	4250	6,4875	0,85
	4924 ARC files	4320	6,5143	0,83
	9856 ARC files	8300	12,4941	0,43
	Baseline	*545*	*0,8321*	*6,60*

Map/Reduce w.				
Tika	*1000 ARC files*	2761	4,2139	1,30
	4924 ARC files	2813	4,2419	1,28
ToMaR w. Tika	*1000 ARC files*	3318	5,0645	1,09
	4924 ARC files	2813	4,2419	1,28
	9856 ARC files	7007	10,5474	0,51

基准线的值由执行单独的 java 应用（使用 JWAT 把数据内容和元数据从一个容器转移到另一个）决定。它作为分布式处理的参考点在聚簇和服务的工作节点上执行。

观察这些数据能发现，与本地的 java 应用处理聚簇相比，聚簇处理对所有的 Hadoop 工作的表现都有显著的提升，这并不让人感到意外，因为这正是分布式处理的目的。当使用一个大型的数据集（在比较4924篇 ARC 文档与9856篇 ARC 文档的产出时变得很明显）时，生产能力发生了显著的增长。在 ARC2WARC-HDP 与 ARC2WARC-TOMAR 两种不同的方法之间只存在些许不同，在上面提到的 Map/Reduce 实现警告中，把 ToMaR 作为产品使用时更有潜力的候选方法。最后，这个图表表明在使用 Apache Tika 时，处理时间提高了50%。

为了展望未来工作及接下来的讨论点，为创建自包含的 WAR 文档而解决上下文信息的依赖性（为建立自包含的 WAR 文档）是未来应该研究的地方。

最后需要说明的是，这里呈现出来的概念证明实现方式离用于实际的工作流还有很长一段距离。在网络存档社区中正在进行的讨论是：在存储机构中解决这样一个项目是否有意义。确保路径追溯工具的向后兼容性及安全保存上下文信息是一个可取的方法。

编译自：

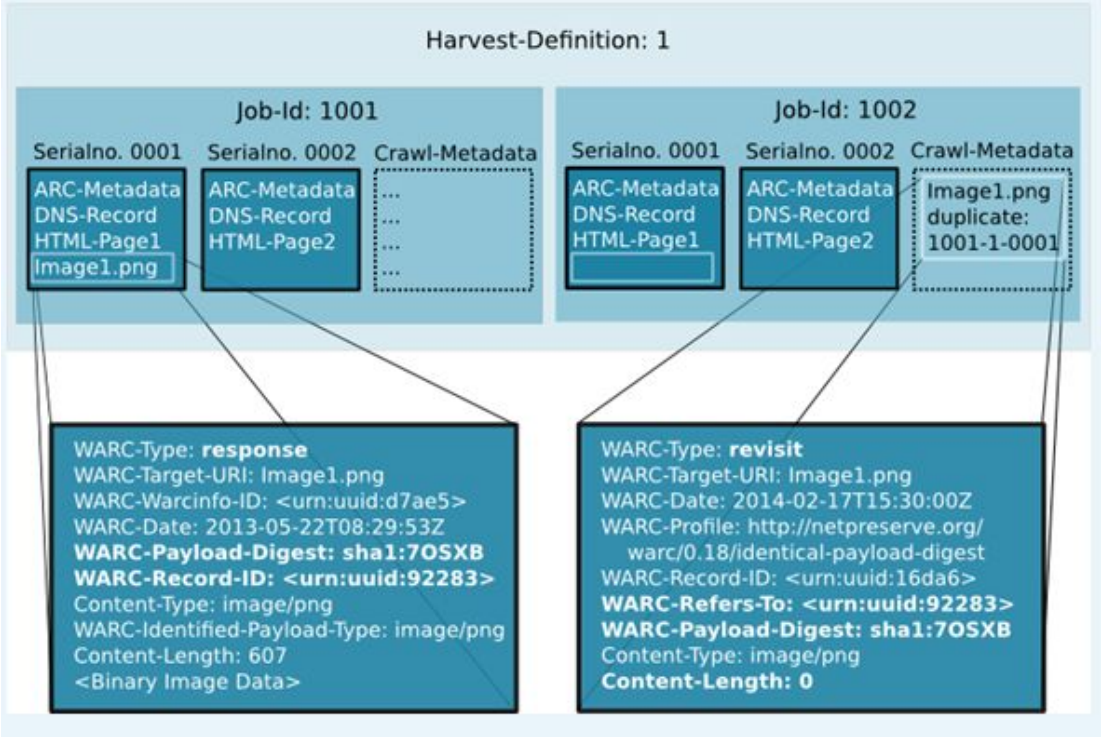
http://openplanetsfoundation.org/blogs/2014-03-07-some-reflections-scalable-arc-warc-migration

（陈瑶编译，唐果媛 吴振新校对）

在 ARC 向 WARC 转换中如何处理重复数据删除的记录？

“重复数据删除”是指网络爬虫访问网站时参考之前已经存储的相同信息的机制，其主要目的是为了避免冗余地存储信息内容，从而减少所需的存储空间。

Netarchive Suite 用一个网络爬虫的模块来描述“重复数据删除”，并附着在多层次的获取定义上。下面的图表大致勾勒出了最主要的信息项及它们之间的依赖性。



这个示例显示了两个后续工作作为相同的获取定义的一部分被执行。基于配置参数，例如，就 ARC 文档所需的大小，每个爬虫工作都会创建一个或多个 ARC 容器文件与相应的爬虫元数据文件。在上面的例子中，第一个爬虫工作（1001）产生了两个 ARC 文档，每个文档都包含了 ARC 元数据及一个 DNS 记录和一个 HTML 页面。此外，第一个 ARC 文档包含了一个 PNG 图像文件，该图像文件在 HTML 文档中被引用。第二个爬虫工作（1002）产生了相同的内容，除了 PNG 图像文件不包含在第一个 ARC 文档中，它只在使用符号（工作-id）-（获取-id）-（系列编号）作为一个“删除重复数据”标识在爬虫元数据中时才被提及。

问题在于：在爬虫元数据文件中真的需要删除重复数据信息吗？如果一个索引（如 CDX 索引）涵盖了所有的 ARC 容器文档，我们（或 wayback machine）能知道如何定位文件，在这种情况下，删除重复数据信息就是过时的。我们会失去爬虫工作的一部分信息（删除重复数据信息真正产生时候的信息），这就关系到爬虫工作的信息完整性（由于外部依赖性不再清晰）。因此，下面是一个符合 WARC 标准方法、用以保存这些信息的方法。

原始 ARC 文档的每个内容记录在 WARC 文档中都会转化为一个对应的记录，就像上图中左下方的盒子所示的那样。任何请求/回应元数据都能作为一个标题字组添加到记录负荷中或作为一个与回应记录有关的单独的元数据记录。

在爬虫元数据文档中能获取的删除重复数据信息项会转化为重新访问的记录，它作为一个独立的 WARC 文档（每个爬虫元数据文档都会有）显示在上图中的右下方。负荷摘要必须相等，并应该声明应用记录的完整性是得到核实的。WARC 引用的性质指的是 WARC 记录包含了记录负荷，此外，内容长度为0指的是记录负荷不在当前记录，应该被存在于其他地方。

编译自：

http://www.openplanetsfoundation.org/blogs/2014-03-24-arc-warc-migration-how-deal-de-duplicated-records

（陈瑶编译，唐果媛 吴振新校对）

ToMaR——如何让保存工具规模化

在没有任何写分布式程序的线索时，用户如何把一个习惯的命令行工具应用到 Hadoop 集群的大量的文件中？这时候就需要用到 ToMaR。

ToMaR 能做什么？

当运行 Hadoop 集群时，ToMaR 只需用户实施简单的三个步骤就能让用户的保存工具在处理数以千计的文件时像本地的单目标 Java MapReduce 应用一样效率。ToMaR 把命令行工具封装到一个 Hadoop MapReduce 工作中，它会在 Hadoop 集群上同时执行所有工作节点的命令。如果用户想通过 ToMaR 使用这些工具，那就需要事先在每个集群节点上安装它。用户需要做的是：

1、指定工具，这样 ToMaR 才能使用 SCAPE 工具规范模式理解它；

2、详细列举控制文件中的每个输入文件的工具调用的参数；

3、运行 ToMaR。

经过 MapReduce，控制文件中参数描述的列表会被分裂并被分配给每个节点部分。例如，ToMaR 可能已经从控制文件中创建了10行的分裂点，然后每个节点会逐行解析,每次都会调用参数指定的工具。

优势

1、通过指令与工具物理调用之间的明显映射能很容易地接受外部工具；

2、使用 SCAPE Toolspec 和现有的 Toolspecs，以及简单的关键字与复杂的命令行模式相关联表现出来的优势；

3、没有编程技巧的要求，只需要设置控制文件即可。

总结

当处理大型文件时（如在文件格式迁移或描述任务的上下文中），一个独立的服务器通常无法在可行时间内提供足够的吞吐量来处理数据。ToMaR 以可伸缩的方式提供了一个简单、灵活的解决方法在 Hadoop MapReduce 集群中运行保存工具。

ToMaR 提供了一种在 Hadoop 分布式环境中使用已存在的命令行工具的可能，这非常类似于桌面电脑的操作。通过使用 SCAPE 工具规范文档，ToMaR 允许用户把复杂的命令行模式和简单的关键词联合在一起，这就可以为计算机集群的执行提供引用。ToMaR 是一个一般的 MapReduce 应用，使用它不需要任何编程技术。

编译自：

http://www.openplanetsfoundation.org/blogs/2014-03-14-tomar-how-let-your-preservation-tools-scale

（陈瑶编译，唐果媛 吴振新校对）