2011 年第 9 期（总第 9 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

2011 年 11 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

【专题报道】

英国长期保存自动化质量保证项目 AQuA 及其成果介绍

Andrew Jackson，Paul Wheatley 著　　齐燕　编译

AQuA 项目链接：http://wiki.opf-labs.org/display/AQuA/Home

AQuA 是自动化质量保证项目的简称，由 JISC 资助，利兹大学、英国约克大学、大英图书馆与 OPF 合作完成。

数字内容的质量保证如果完全依靠人工来完成，那结果将可能不可靠，而且藏品也会因为各种各样的质量问题而受到损害。较差的储存条件，会导致比特错误（bit-rot）从而引起进一步损害。技术陈旧也会带来附加的风险。检测、鉴定和修复这些问题，在传统的数字馆藏中是昂贵和费时的人工流程。在数字化或获取后及时地识别问题能够使之更加有效、节约更多成本。AQuA 项目运用各种现有的软件工具，以实现自动化的质量保证和评估。它的两次活动汇聚了数字资源长期保存的从业者，藏品管理者和技术专家们，展示了目前数字馆藏的各种问题，阐明其验证要求，并应用工具来实现保存和质量问题的自动化检测和鉴定。持续性的活动正在进行中，可能会有后续项目来检阅进展情况，并在不久的将来有所应用。

AQuA 致力于面对长期保存和质量问题的挑战，而长期保存和质量问题会时不时出现在我们数字藏品的以下各个过程中：

当我们创建数字化内容时，比如页面丢失、重复页、较差的聚焦/对比度。

当这些藏品被储存时，比如比特错误。

当这些藏品被处理或者从一种存储设备转移到另一种存储设备上时，比如当进程占据了全部的内存或磁盘空间。

当技术变化时，比如我们的标准和文件格式过时了。

其中许多问题都可以通过人工质量保证处理过程识别出来，使得损失或损害能够得以减轻。即使这样，人工质量保证方法也是劳动力非常密集的。自动化的检测方法有潜力提供周密的检测并能大幅度的降低成本。

AQuA 项目研究内容以及成果

1、藏品

该项目涉及的藏品由数据管理和长期保存从业者们提供，包括各种音频、图片、Word 文档、多媒体文件、PDF 文件等。

2、问题

下面是一些与上述藏品的长期保存和质量保证相关的已有的和潜在的问题。需要注意的是，一种问题会存在于多种藏品中，而且会有多种解决方案。

（1）问题——解决方案评价问题

（2）音频文件的问题：

审核音频批处理是否违反标准

隐蔽音频文件格式的识别和验证

数字音频文件的标准化

（3）图像文件的问题：

审核图像是否违反标准

BOPCRIS 问题——压缩和非压缩的 TIFF 混合

同一对象的多个扫描图像的去重

一次收集或作业内的重复图像

EAP 问题 1——损坏的 TIFF 图像

EAP 问题 2——不能在 Photoshop 或 Adobe Bridge 中打开的 TIFF 图像

EAP 问题 4——检测视觉错误

查找重复图像

历史影像藏品——检查图像版本

历史影像藏品——跨越时间或供应者的一致性

识别不同层次的旋转角度下相同的图像

报章刊登日期

从 TIFF 到 JPEG2000 的迁移过程中的质量保证

数字化网页上的质量问题

未知的 JPEG2000 的特点蕴含了对藏品质量，保护和访问的风险

使用 METS 数据报告分析结果

检验从 TIFF 转换来的 JPEG2000 文件，识别并追踪误差来源。

（4）元数据问题：

原生数字－－日志文件检查

原生数字－－元数据检验

EAP 问题 3——从音频、视频和图像文件的提取元数据

EAP 问题 5——识别空文件夹

EAP 问题 6——发现丢失或失序文件

数字音频文件元数据的提取

元数据和内容之间的不一致

未知原生数字文件的历史记录

（5）MS Office 文件的问题：

原生数字－－迁移成功

识别 MS Office 文档的内容

（6）多媒体问题

识别电子邮箱中的内容

使用者通过一些操作停止文件的播放

（7）OCR 的问题

BOPCRIS 问题——ABBYY“未知错误”

识别遗漏或重复的页面

OCR'ing 混合的内容（文字和非文字）

数字化页面中可能存在的质量问题

OCR 元数据的使用

（8）PDF 的问题

PDF 文档中嵌入的链接

PDF 文档中嵌入的对象

PDF 的未知的特性

3、解决方案

下面是针对上述问题的解决方案，是由 AQuA Mashup 活动的参与者开发出来的。

音频解决方案：

AQUAdio——对用户生成的音频源文件进行特征描述

音频审计脚本

图像解决方案：

捕获相关的配置文件，抽取并检查一致性

EAP 文件验证

识别压缩的 TIFF 文件，对其进行解压缩

利用哈希算法识别旋转的、重复的图像

技术文档

Java 图像块比较

JP2 标题分析

报纸的发行日期——解决方案

感知图像差异比较

ssdeep 软件包检测重复图像

tiff2RDF——可视化图像藏品的一致性

验证 TIFF 向 JPEG2000 格式的迁移

元数据解决方案：

对外部生成的内容进行特征描述

检查元数据和内容之间的一致性

利用 EAP 协议检查元数据是否符合要求

微软办公软件解决方案：

Lucene 索引中单词出现的频率分析

Apache 的 POI Office 文档分析仪

多媒体解决方案：

AQDC——文件比较

使用 FLVmeta 的 flvdump 诊断 FLV 问题

识别电子邮件信箱的内容——解决方案

OCR 解决方案：

比较同一源材料不同的格式（TIFF，JP2）下的 OCR 结果

OCR 比较

PDF 解决方案：

检测，抽取和分析 PDF 中的嵌入对象

PDF 表征工具

4、AQuA Mashup 工具列表（只列部分，全部列表详见下面链接）

工具名称	URL	描述 /用途	AQuA 解决方案的运用
File Utility	http://www.darwinsys.com/file/	开源的文件格式识别实用程序，用 C 语言编写，打包成 UNIX 发行版。 错误：http://bugs.gw.com/ http://bugs.debian.org/cgi-bin/pkgreport.cgi?package=file	
Sanselan	http://commons.apache.org/sanselan/	纯 Java 库读写各种图像格式，包括快速解析图像信息（大小，颜色空间，ICC 配置文件等）和元数据。	图像关系一致性的可视化工具
Extractor	http://planetarium.hki.uni-koeln.de/planets_cms/extractor	从一些如图像、音频等的文件格式中抽取技术性的元数据。	
JHOVE	http://hul.harvard.edu/jhove/	抽取文件属性并尝试利用格式规范进行检验。支持 AIFF ASCII GIF HTML JPEG JPEG2000 PDF TIFF UTF-8 WAVE XML 格式。	JP2 标题分析

JHOVE2	https://bitbucket.org/jhove2/main/wiki/Home	继承于 JHOVE。集成了 DROID。支持 ICC NetCDF SGML Shapefile TIFF UTF-8 WAVE XML。	
DROID	http://sourceforge.net/apps/mediawiki/droid/	基于文件内部“magic”签名或者文件扩展名来识别文件。记录不一致性。提供图形用户界面。	
ExifTool	http://www.sno.phy.queensu.ca/~phil/exiftool/	元数据/属性抽取和编辑工具，支持几十种格式，尤其以图像格式为重点。很可能是现有的最好的属性抽取工具，但奇怪的是，易被大多数的数字资源长期保存社区所忽视。	
PSTView Tool	http://pstviewtool.codeplex.com/ http://pstsdk.codeplex.com/	微软的开源项目，是一个 PST 浏览器和底层的 PST 访问库（C++）	
libPST	http://www.five-ten-sg.com/libpst/	PST 操纵/迁移库。请参阅 JvdK 的评论。	
Unpaper	http://unpaper.berlios.de/	针对后处理扫描。可以当场旋转，黑色标志等，可在诊断阶段使用。 http://unpaper.berlios.de/unpaper.html＃imagefiles	
b2x Translator	http://b2xtranslator.sourceforge.net/	一种从 DOC / PPT/ XLS 到 DOCX/ PPTX/ XLSX 格式转换工具，由微软的一个合作伙伴开发。	
JDeskew	http://www.jdeskew.com/	开源的 Java 纠偏库	
PeDALS	http://sourceforge.net/projects/pedalsemailextr/	将长期保存存档和 PeDALS 研究项目产生的信息用电子邮件发送给 XML 文件抽取器，以备数字长期保存之需。	
FITS	http://code.google.com/p/fits/	这个工具集可以识别、验证和提取各种文件格式的技术元数据。它集成了一些第三方的开源工具，包括 JHOVE、DROID、File 等，并对它们的输出结果进行规范化和联合，报告所有错误。	EAP 文件验证 识别压缩的 TIFF 文件，对其进行解压缩 tiff2RDF——可视化图像藏品的一致性
ODF Converter	http://odf-converter.sourceforge.net/	这个项目的目标是提供一个转换器以支持基于 ODF（OpenDocument）标准（目前是 ODF1.1）的应用程序和微软 OpenXML 的 Office 应用程序之间的互操作。 考虑到 Microsoft Word、Excel 和 PowerPoint 中的加载项，还提供了一个具有控制功能的转换器，支持进行批量转换。这些转换器也可以在某些情况下运行于服务器端。	

ODF Toolkit	http://odftoolkit.org/	包括一个验证器。主要是 Java，也有些.NET 代码。	
PDFtk	http://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/		
pdf2xml	http://sourceforge.net/projects/pdf2xml/	见：http://discerning.com/hacks/docutils/pdf2xml/readme.html	
PDFSSA4MET	http://code.google.com/p/pdfssa4met/	为元数据的抽取和标记而进行的 PDF 结构和句法分析。PDFSSA4MET 试图在 XML 内容的结构和句法分析的基础上进行元数据的抽取和标记。	
pdftohtml	http://pdftohtml.sourceforge.net/		
pstoedit	http://www.pstoedit.net/pstoedit		
Multivalent	http://multivalent.sourceforge.net/		
JODConverter	http://www.artofsolving.com/opensource/jodconverter	JODConverter，基于 Java 的 OpenDocument 格式转换器，可对不同的 Office 格式的文件进行转换。它充分利用了 OpenOffice.org 所提供的当前最好的 OpenDocument 格式和微软 Office 格式的导入/导出过滤器	
pdf2svg	http://www.cityinthesky.co.uk/opensource/pdf2svg		
Email Preservation Parser	http://siarchives.si.edu/cerp/parserdownload.htm		
pHash	http://www.phash.org/	开源的感知哈希库。一个感知哈希是多媒体文件的指纹，可由来自内容的各种功能推导而来。加密哈希函数会产生雪崩效应，即输入存在微小的变化导致在输出时的剧变。而感知哈希函数则与加密哈希函数不同，功能间相互隔离，输入变化不会影响其他相似功能。参见：http://stackoverflow.com/questions/596262/image-fingerprint-to-compare-similarity-of-many-images	利用哈希算法识别旋转的、重复的图像
Fiji	http://pacific.mpi-cbg.de/wiki/	Fiji 是一个图像处理包。它既可以被描述为一个 ImageJ 与 Java、Java 3D 的分配器，也可以描述为集成到关联菜单结构中的各种插件。另见：http://fly.mpi-cbg.de/~saalfeld/Projects/javasift.html , http://rsb.info.nih.gov/ij/plugins/mssim-index.html	

getID3()	http://getid3.sourceforge.net/	getID3()是一个 PHP 库，从 MP3 及其他多媒体文件格式中抽取有用的信息。	AQUAdio——对用户生成的音频源文件进行特征描述
The GIMP	http://www.gimp.org	GIMP 是 GNU 图像处理程序。	识别压缩的 TIFF 文件，对其进行解压缩
Taverna	http://www.taverna.org.uk/	这是一个开源的工作流管理系统。它有一套工具，用于设计和执行科学的工作流。	
Cue	https://github.com/jdf/cue.language	一个小型的 Java 库，用于进行简单的文本分析——计数字符串，识别语言，并消除停用词。 用于 futureArch 的一个很简单的词云生成器	
Apache Tika	http://tika.apache.org/	Apache Tika ™工具包使用现有的解析器库来检测和抽取各种文件的元数据和结构化文本内容。	对外部生成的内容进行特征描述 AQDC——文件比较
Apache Lucene	http://lucene.apache.org/java/docs/index.html	Apache Lucene（TM）是一个高性能、全功能文本搜索引擎库，完全用 Java 编写。它是一种适用于几乎任何需要全文检索、特别是跨平台的应用程序的技术。	对外部生成的内容进行特征描述 Lucene 索引中单词出现的频率分析
Java Image Comparison	http://mindmeat.blogspot.com/2008/07/java-image-comparison.html	基于块块比较方法对图像重复/差异进行基本比较	Java 图像块比较
BWF MetaEdit	http://bwfmetaedit.sourceforge.net/	抽取特定文件的元数据（采样率，采样比特率）。	做好音频审计脚本
jHears	http://jhears.org/	音频指纹（这也依赖于 SoX）。需要有客户端和服务器端双方软件。	做好音频审计脚本
Kakadu	http://www.kakadusoftware.com/	JPEG2000 软件框架	Jp2 标题分析
ssdeep	http://ssdeep.sourceforge.net/	ssdeep 是一个用于计算上下文触发分段哈希（CTPH）的程序。 CTPH 也称为模糊哈希，可以匹配同源输入。这些输入中同一字节序列有相同顺序，虽然在这些序列之间的字节可能会在内容和长度都不同。 ssdeep 使用一个滚动的哈希算法，因此该文件的变化将只会导致在 CTPH 签名中的局部变化。	ssdeep 软件包检测重复图像
pdiff: Perceptual Image Difference utility	http://pdiff.sourceforge.net/	图像比较/区分工具	感知图像差异比较
ImageMagick	http://www.imagemagick.org	位图图像软件套件	tiff2RDF ——可视化图像藏品的一致性
OpenJPE	http://www.ope	OpenJPEG 图书馆是一个开放源码的 JPEG2000 编解码器，	验证 TIFF 向 JPEG2000

G	njpeg.org/	是用 C 语言编写的。它早先是为了 JPEG2000 使用的推广而开发，JPEG2000 是来自 JPEG 新的静态图像压缩标准。	格式的迁移 比较同一源材料不同的格式（TIFF，JP2）下的 OCR 结果
Apache POI	http://poi.apache.org/	操纵各种基于 Office Open XML（OOXML）标准的文档格式和微软的 OLE 2 复合文档格式（OLE2）的 Java API。	Apache 的 POI Office 文档分析器
tesseract-ocr	http://code.google.com/p/tesseract-ocr/	OCR 工具	比较同一源材料不同的格式（TIFF，JP2）下的 OCR 结果
PDFbox	http://pdfbox.apache.org/	面向 PDF 文档的创建，操纵和内容抽取的 Java PDF 库	检测，抽取和分析 PDF 中的嵌入对象 PDF 表征工具
itext	http://itextpdf.com/	具有操纵，内容抽取和创建功能的 PDF 库	PDF 表征工具

结语

AQuA 项目和活动已经结束，但这项工作还将继续作为 OPF 和 AQuA 新兴社区的一部分。希望保持项目中培养出的动力，并鼓励数字资源长期保存社区之间有更多的基层的合作。DPC 和 OPF 正在举办一场使用 AQuA 格式的新活动，同时计划在 2012 年开展 AQuA 跟进活动，目的在于回顾 AQuA 参与者的工作所产生的影响，并考虑 OPF 和 JISC 如何能够更好地支持数字资源长期保存社区。

编译自：

http://www.openplanetsfoundation.org/community/opf-events/hackathon-practical-tools-digital-preservation

http://wiki.opf-labs.org/display/AQuA/Home

（刘晓敏　吴振新校对）

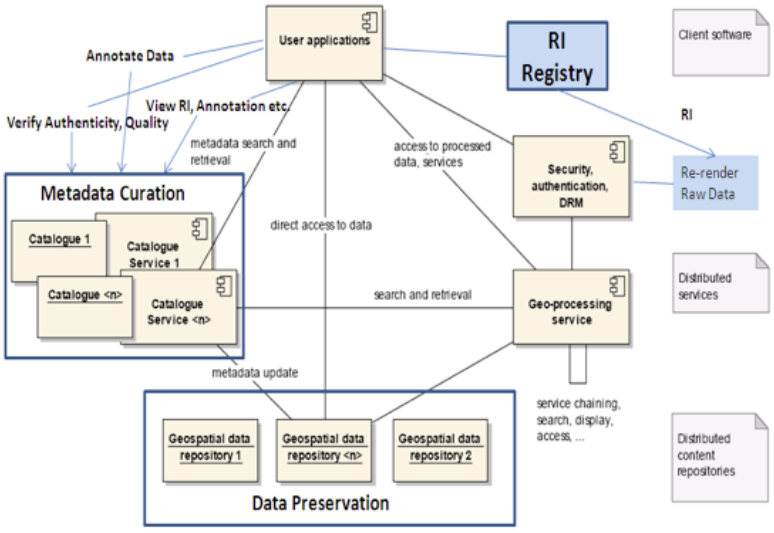
【重要文献摘译】

空间数据基础设施的长期保存：一个元数据框架和地理门户网站的实现

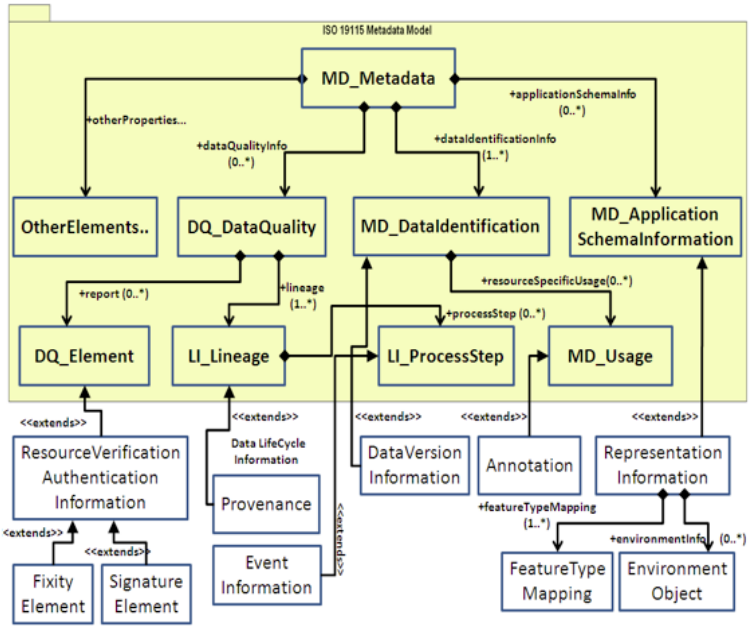
Arif Shaon，Andrew Woolf 著　　么媛媛编译

人们对环境问题越来越多的关注和过去十年间以指数级增长的计算能力使得地理空间信息社区产出了日益浩繁和多样化的环境数据集。如何对这些已经通过统一且可互操作的空间数据基础设施（SDI）公开揭示的环境数据实施长期保存还没有得到解决，但在需要持续访问当前和历史数据的一些实际应用中，比如说监测气候变化，数据的长期保存就显得尤为重要了。这篇文章从一个有保存意识的空间数据基础设施的角度，研究了确保可持续获取环境数据所包含的要求。文中的研究和模型开发以 INSPIRE 为样本，并以建立一个地理信息门户网站的形式提出了一个解决办法，同时该门户网站还介绍了 ISO19115 元数据标准在长期保存中的应用。

该文章首先介绍了欧洲委员会的 INSPIRE 指令的具体要求和持续获取环境数据的难度及重要性，阐述了保存环境信息所面临的一些挑战；然后开始建立开放存档信息系统（OAIS）的参考模型，该模型包含内容信息、保存描述信息（PDI）、表示信息（RI）、包装信息、指定的社区/知识库；之后提出了一个具有保存意识的空间数据基础设施，如下图（A Preservation-aware SDI）；



文章就 ISO19115 元数据模型在长期保存领域的应用进行了介绍（如下图 A preservation profile of ISO 19115 Metadata Model）。



然后介绍了如何实现这种空间数据基础设施的方法，提供了确保数据质量、数据完整性和数据长期保存所需要的信息，并使这种基础设施得到有效的开发和利用。

在文章的最后，作者对空间数据长期保存的现状进行了总结，并指出未来发展方向：通过统一且可互操作的空间数据基础设施 SDI 来公开的环境数据的长期保存在 INSPIRE 指令中尚未得到解决，但如果在实际应用中需要持续访问当前和历史的数据，长期保存尤为重要。

今后的工作将需要把重点放在通过 SDI 提供的数据资料库的高效且彼此协作的保存解决方案的实施上。这里尤其应该包括源数据集和其相应的一个或多个功能视图之间的映射，这些映射可能存在不同程度的复杂性，它们使那些功能视图在将来能得到准确的响应。新兴的欧洲空间科学（ESA）长期数据保存（LTDP）计划预计会在这方面做出重大贡献，其目标是在整个欧洲制定一个协调一致的空间数据档案长期保存方案。该文所做的工作或许可以作为欧洲空间科学计划的一个探索性活动。

编译自：http://www.dlib.org/dlib/september11/shaon/09shaon.html

（齐燕　吴振新校对）

如何开发一个数据计划和数据共享计划

Sarah Jones 著　　刘晓敏编译

1. 为什么要开发一个数据计划

管理和共享数据的好处包括以下几点：

(1) 当你需要使用数据的时候，你可以找到并理解你的数据；

(2) 当项目人员离职或新研究人员加入的时候，可以确保数据的可持续性；

(3) 可以避免不必要的重复，例如：重复收集或数据回修；

(4) 出版物中潜藏的数据可以得到维护以对结果进行验证；

(5) 数据共享可以促进研究和合作；

(6) 让研究变得可视化，提升研究的影响力；

(7) 其他研究者可以引用你的数据，这意味着你得到了承认；

2. 研究基金会想要什么？

已有诸多的英国基金会发布了数据政策来倡导对数据的管理和共享。有些基金会还要求数据管理和共享计划应该作为一项授权应用来提交。基金会期望数据计划能够勾勒出数据是如何被创造的，如何管理，以及如何进行共享和保存，以更好地评判这个过程中需要应用的限制条件。因此，数据计划绝对是一个好机会，一来可以展现你的优秀的实践意识，二来可以给基金会一颗定心丸，让基金会看到你的提议是紧跟数据政策步伐的。

DCC 通过对英国各基金会关于数据管理和共享计划的指导方针进行整理后认为，基金会特别建议大家根据自身具体情况，拓宽需考虑的主题和问题领域。对字数的要求和页数的限制将会进一步加强，因此我们的计划需要更简洁明了。基金会希望一份简洁的摘要能作为“支持案例”的一部分提交，或者在申请表的指定位置中填写。千万要避免重复的内容以及与数据管理共享毫不相干的内容。

此外，DDC 还为研究人员提供了 DMP 在线工具，供研究人员根据基金会的要求创建数据管理和共享计划。DMP 在线工具的结构基于 DCC 数据管理计划清单，其中定义了需要包含的所有原理。这些内容包括：

(1) 数据类型、格式、标准以及捕获方式；

(2) 道德规范和知识产权；

(3) 数据的访问、共享以及再次利用；

(4) 短期保存和数据管理；

(5) 长期保存；

(6) 资源提供。

3. 关于如何规划的建议

(1) 咨询与协作：认真分析不同的选项，通过征求意见决定哪一项最符合你的环境。对于技术方面，这一点尤其适用。因为技术的使用往往会影响到项目的规划、专业知识、所需的应用工具以及获取和分析数据的方法。这些建议可以通过身边的同事、图书馆、IT 支持机构、咨询机构、伦理委员会、数据贮存库等获得。

(2) 使用已有的支持：有的问题已经有人在你之前解决过了，因此你可以在一个已有的模型上构建你的规划。数据管理基础设施数量正在不断增长，尤其是在机构内部。IT 服务运行在已建立好的备份机制上，你的研究团队或部门因此也有可借鉴的本

地政策和流程。此外，多学科数据中心、知识库以及结构化数据库也是获取支持的途径。

(3) 证明你的决议：基金会对于你具体使用的文件格式、标准或方法并不会有特别的要求。你需要自己做出选择并描述你做出的选择是最适合当前的环境、规则以及未来用户的。同样地，对于数据共享的限制，你也需要呈现一个有说服力的案例。

(4) 准备好落实你的规划：基金会希望看到你理解他们的要求并且已经做好切实的计划。对于规划工作的描述应该力求清晰和可行，规划制定者也会有足够的信心确信你们能够充分理解并且传达提议。清晰的角色定义和责任会为你的规划加分，因此规划中应该清楚地交代谁将要做什么，怎么做，什么时候做。

4. 数据管理共享计划的内容

(1) 数据类型、格式、标准和捕获方法

- 描述并证明你的选择：你应该详细说明你将创建什么数据并解释为什么你选择这种数据格式、标准和方法。时刻要牢记在心，选择的好坏将会对你的数据保存产生正面或负面的影响。
- 使用业内普遍认可的数据格式去捕获数据往往是不错的选择。使用标准的或广泛应用的数据格式将会让数据更具互操作性。开放或非专利的格式是更可取的，这样可以免去诸多不必要的麻烦。如果你的数据将要被保存至储存库，那么就应该选择特定的数据格式。
- 文档及元数据：让你的数据更容易被其他的程序理解和发现。获取关于数据如何被创建以及为什么被创建的上下文信息是非常重要的。元数据就是这个大文档的一个子集，可以对数据进行详细描述。目前已经有许多标准元数据可以用来对数据进行描述。关于相关的元数据标准，图书馆员、数据保管机构或者你的同事都是你可以咨询的对象。
- 认真评估，决策透明：这可以展示你优秀的实践能力，也可以表现你为了计划的制定，已经做足了功课。

(2) 道德准则和知识产权

- 用实例为共享限制做辩护：需要对一切限制做出解释，例如限制期限或者受限访问。对于大家普遍认为公共基金支持的研究数据应该尽快地开放，我们也应该确保能够恰到好处地做出解释。
- 所有涉及人类数据或材料的研究都应该遵循正规的道德准则。必要的话，我们应该制定出保护研究成员的详细步骤，例如匿名数据。这样可以确保我们能够在数据共享的问题和期望之间做出平衡。有许多大学道德委员会提供同意书和服务的样本，如英国的数据保管中心就在这一领域提供出色的指导和帮助。此外，也应该留意相关的法律法规，如《数据保护法》。

- 数据所有权应该事先声明，必要情况下计划书之前应该附上协商许可。如果是通过购买或同意相关许可来重用第三方数据，就应该在数据保存和共享过程中留意相关的条件限制。JISC 在版权、IPR 和相关法规（《数据保护法》、《信息自由法》）方面可以为我们提供诸多建议。此外，也可以从大学图书馆的相关专家、数据管理和研究办公室获得相关的支持。

(3) 数据访问、共享和再利用

- 参与和规划数据再利用：这可以帮助我们了解我们的目标用户及其需求，以此采用合适的数据满足其需求。数据中心也会要求你的数据符合最低标准以确保数据能够为其他研究者理解和再利用。
- 为数据访问提供具体说明：让基金会清楚地明白数据在什么地方、什么时候通过何种方式可以被获取到。DDC 可以就如何授权许可数据提供指导，以清楚说明谁出于何种目的可以使用数据。基金会还经常会对数据发布的时间做出要求，例如什么时候该发布在出版物上。如果你不能做到这些或需要做出使用限制，那么就应该阐明你已经为克服这些困难做出了巨大的努力。
- 使用已有的基础架构：尤其是在选择一个合适的数据库、数据中心或贮存库的时候。如果你不确定哪一种服务适合你，那么可以查询 DataCite、BioCentral 以及 DCC 上的贮存库清单。如果你的数据存在访问限制，那么应该选择可靠的数据服务商。

(4) 短期数据存储和数据管理

- 明确数据管理支持：描述出在机构内部可利用的资源以及出于安全考虑所需的其他技术和资源。如果本地的支持就已经能够满足需求，那么就应该遵循相关的要求。如果需要外部的安全支持，则需要根据预算来做出相应的选择。此外，还需要针对不同的任务确定负责人。
- 实用性考虑：研究者是否属于同地协作？是否需要基础设施来协调远程安全访问？分布式数据如何进行质量监控？文件命名规则和版本控制应用对于跟踪开发进度，尤其是几个人协同工作时，将会大有帮助。
- 使用适度层级数据管理：基金会往往期望看到每天的数据管理都能符合预期的目的。对此，可以采用分级服务或以下两种方式的结合：
    - 相比于通过许可拥有的二手数据，亲自收集的敏感数据对安全的要求会更高。考虑一下数据是如何安全地进行转换的：对数据进行编码或使用安全的在线存储。如果使用的是在线存储，那么你应该了解你的数据存储于何处并且能够合法获取利用。
- 对唯一性数据的备份往往要比二手数据的备份更重要。越重要的数据往往越常被访问，也应该越频繁地进行备份。具有自动备份功能的文件管理服务往往可以为你节省大量的时间和精力。这种服务也可以同便携存储或云计算集合在一起来满足特定

的需求。

(5) 长期保存

- 选择具有长期保存价值的数据：数据共享和保存并非在所有的案例中都是可行的。DDC 提供了一本数据评估的指导手册，提供了选择重要数据的一些实用性策略。决定什么数据具备长期保存价值并将这些数据转换为特定的格式加以保存是相当耗费时间的一个过程，因此应该做好重要资源的合理分配。
- 保护图表背后的数据：RCUK 的基金会普遍认为，研究结果的发布会包括获取数据的方法。即使你的大多数数据没有固定的存放地点，对于支持研究的数据也应该被抽取出来，以机器可读的形式存放于可为人们访问的地方。
- 确保数据的可访问性：无论你采用何种途径，务必确保你的数据是始终可访问的。如果数据存放于数据中心，则需要与数据中心的工作人员做好沟通，从他们那里获取到合适可行的数据保存方法。此外，许多大学也开始提供越来越多的基础设施来支持数据的管理，我们也可以从那里获取适合使用的服务。

(6) 资源

- 成本描述和评价：如果需要购买存储、外部资源服务（如备份服务和保存服务）或者数据管理支持服务，那么应该在计划中对这些服务的费用进行事先的描述和评估。当已有内部资源可用时，应该确保你所提出的要求是经过详细讨论并获批的。这对于连接外部资源以及规划实施过程中角色和责任的分配也是有帮助的。
- 不要低估所需的人为努力：文档的创建以及让你的数据对于其他研究者而言可理解是一件很耗费时间的事，因此应该客观评价在数据共享和保存过程中所需的付出。UKDA 提供了一个工具包以帮助研究人员评估管理和共享社会科学数据所需的成本。
- 有效使用公共基金：《RCUK 关于数据政策的准则》提出要合理利用公共基金来支持研究数据的管理和共享，而且不但要合理，还需要高效和成本节约。

原文链接：http://www.dcc.ac.uk/webfm_send/480

编译自：http://www.dcc.ac.uk/resources/how-guides/develop-data-plan

（么媛媛 吴振新校对）

《分布式数字资源长期保存指南》摘译之七

第七章 PLNs 的缓存和网络管理

Matt Schultz, Bill Robbins 著 何莉娜编译

概述

如前所述，一个 PLN 要么依赖于 LOCKSS 项目对其提供网络管理和支持，要么作为一个独立实体进行本地管理运作。本章首先简要介绍组织机构选择某种方案的技术和管理方面的原因，之后提供这两种情况下的缓存和网络管理中最佳实践的技术概述。最后，本章推荐了一些有效的交流工具和策略，它们或许能够辅助 PLNs 进行缓存/网络的管理活动。

本地管理 PLNS 和斯坦福管理 PLNS

很明显，所有 PLNs 都采用 LOCKSS 软件来构建其长期保存技术基础设施。LOCKSS 代码是开源的，但并不仅仅依赖开源编程社区维持其软件开发和维护活动。相反，LOCKSS 提出了一种可持续性模型，即：要求使用 LOCKSS 软件的机构和希望得到斯坦福大学团队支持的机构，向 LOCKSS 联盟交纳年度会员费。这些费用可以支持 LOCKSS 核心团队持续地更新和改善 LOCKSS 软件。所有的 PLN 会员，就像是一个公共 LOCKSS 网络的成员，共享着与 LOCKSS 联盟的关系。PLNs 可能采取两种方法中的其中一种来管理和监控其网络：他们可能组成一个由斯坦福管理的 PLNs，或组成一个由本地管理的 PLNs。

由斯坦福管理的 PLNs 依赖于 LOCKSS 项目的中央基础设施，包括题名数据库、插件仓储库、插件开发和缓存管理。PLNs 职员负责内容识别和内容准备。由斯坦福管理的 PLN 利用 LOCKSS 中央程序工作流来为其仓储库测试、配置适合的插件，并使其内容通过题名数据库可以获取。LOCKSS 项目组也可以帮助 PLN 成员完成类似增加或移除缓存的任务。

本地管理式的 PLN 也采用 LOCKSS 软件，但是在所有其它方面则选择自行管理各自网络基础设施，包括维护其独立 LOCKSS 题名数据库形式的配置环境、自行开发程序来配置缓存、提供一个或多个插件库等。他们必须已经在添加或修改插件规程上达成一致意见，并且需要建设监测基础设施来保证网络/缓存的正常运行。

这两种创建和运行 PLN 的方法都已经构建了一些强健高效的长期保存网络。

创建和维护 PLN

合作构建保存网络的机构的共同目标决定了创建和维护 PLN 的最优方案。根据第三章介绍的 PLN 技术方面的考虑和第四章介绍的组织方面的考虑，确定 PLN 依赖 LOCKSS 项目基础设施的程度，或者决定选择本地管理式的 PLN 结构等，都需要综合技术和组织两方面的决策因素。

在前面章节描述的案例中，那些预期其成员机构在维护技术基础设施上有困难的 PLNs 已经选择了依赖 LOCKSS 项目的基础设施。对于那些想要建立一个不依赖于 LOCKSS 项目组并且在中央和个体成员站点都有足够的技术专家的 PLN 来说，他们能在网络配置和运行方式上拥有更大的灵活性。

推荐硬件

如前所述，一个 PLN 至少由七个保存站点组成，每个站点都运行 LOCKSS 软件并通过网络配置相互连接。由于 LOCKSS 软件本身具有很高的冗余度，而且控制着网络中的所有保存活动，所以实际上，无论是委托管理还是本地管理的情况，在其保存站点上运行的软件

都是非常基础、廉价的。

LOCKSS 团队建议：PLN 中的每个缓存，要么是一台低成本 PC，利用 Live CD 运行其 LOCKSS 软件；要么是一个低成本的 UNIX 或基于 Linux 的服务器，运行 LOCKSS 后台软件的安装包。

第三章 PLNs 的技术考虑中提到，PLNs 已经能够处理与硬件购买和系统配置需求有关的决策，这些决策必须是建立在维护其网络中缓存一致性的基础上。最重要的因素是缓存的 CPU 性能要与磁盘性能相匹配，因为 LOCKSS 后台程序要不断地对缓存内的内容的完整性和正确性进行推敲、公认和轮询。缓存中的内容越多、磁盘越大，CPU 的处理速度就需要越快，以保证 LOCKSS 后台程序能够顺利执行各种审计活动。当然，只要网络中的缓存可以和其它缓存有效交流，并且没有受到 CPU 速度和磁盘空间不相匹配的影响，保存网络就可按其意愿自主选择任何硬件（缓存中的或网络层面中的）。

LOCKSS 软件安装

选择实时 CD 还是 Linux (RPM)安装包来安装 LOCKSS 软件在很大程度上需要从技术和组织两方面来考虑。实时 CD 安装不仅包括 LOCKSS 后台程序，更需要有预配置的、高安全性的操作系统。这种类型的安装对于那些寻求方便其成员建立和更新缓存的 PLNs 来说是理想的方式（或途径），因为这使个体成员的技术投入缩减到最小。

而对于那些想要独立于 LOCKSS 项目的 PLNs 来说，基于 RPM 的 Linux 安装包可能更加合适。这使得像 MetaArchive Cooperative 此类的 PLNs 在为其保存网络做硬件选择、安全配置、内容监管工具的开发/执行方面具备了较大的灵活性。

一旦缓存中的内容达到了一定的数量，实时 CD（OpenBSD）的执行效果就相对不够理想了，因为 Linux 系统可以比 OpenBSD 更有效地处理大型的磁盘阵列。如之前第三章讨论的，LOCKSS 团队打算近期将实时 CD 转为虚拟机软件。

测试 LOCKSS 后台程序

LOCKSS 中心同时对斯坦福管理的 PLNs 和本地管理的 PLNs 的后台程序负责，定期地发布相应的更新程序。对于使用实时 CD(OpenBSD)的 PLNs，有一种软件自动更新系统可以自动地、安全地升级缓存。当 LOCKSS 中心发布新程序时（大约每六周发布一次），该软件不需要人工干预就可自动安装。每年 LOCKSS 团队会发布两次新的实时 CD，缓存站点管理员只需要将新的程序拷贝到磁盘或驱动器上重启即可。对于在 Linux 上使用 LOCKSS 软件并且自行管理其题名数据库实例的 PLNs，更新必须手动完成。目前的最佳实践都建议，由指定成员机构或中央系统管理者先在测试网络上对新发布的程序进行测试，然后再安装到缓存设备上，以确保其兼容性。

测试网络

测试网络是一系列独立的、配置了 LOCKSS 的缓存，用于测试而非生产目的。它可以运行一些灾难场景进而来测试缓存的恢复性能。它使得 PLN 能够测试内容的收割，确保

LOCKSS 后台程序可以从某一指定成员的托管内容中成功地收割资源。它也提供了在投入生产网络前对插件和为资源内容创建的显示页面的精确性的测试。

LOCKSS 团队维护一个测试网络用于测试后台程序的更新，确保提交的插件在投入 PLN 的生产网络使用前即能测试其在 PLN 成员站点上成功收割资源。本地管理的 PLNs 可能需要建立和维护其自身的测试网络，该网络可以由单个缓存或由两到三个缓存组成的小型网络构成。

为网络配置缓存

长期保存网络中的每个缓存都必须配置好，确保与网络中的其它缓存、中央管理服务器进行通信交流，中央管理服务器中包括诸如题名数据库、插件仓储库以及监管网络的工具。本部分介绍了最佳实践，概述了配置斯坦福管理的以及本地管理的缓存的步骤。

缓存通信配置

建立任何 PLN 时，仅需做少量的基本配置即可保证网络中所有缓存间的成功通信。这些设置主要包括：缓存 IP 地址以及完全合格的域名；包含本地工作站 IP 地址的子网络，这些工作站需要访问 LOCKSS 缓存的用户界面；邮件交换服务器信息；本地系统管理员的邮件地址；LOCKSS 后台程序 Web 界面的用户名和密码等。

前两条信息是最重要的，因为它们用于 LOCKSS 后台程序识别出参与网络共享题名数据库的缓存和本地工作站，以执行其保存功能和提供用户反馈。在斯坦福管理的 PLN 中，配置缓存包括选择与 LOCKSS 推荐兼容的硬件，收集上述信息并在软件安装和初始化过程中跟踪具体指令来提供相应信息。该过程对所有系统管理者来说都是简单明了的，不需要掌握任何详细的 UNIX 操作系统知识。在本地管理的 PLN 中，根据其选择的硬件和操作系统的不同，其启动程序也各不相同。配置过程要么像斯坦福管理的网络一样简单，要么可能需要更多的本地专家，这取决于集中提供安装的自动化水平，以及缓存设备的操作系统和硬件与 LOCKSS 支持平台的偏差量。

最佳实践建议不要对任何成员内容的空间设置硬性的限制，这可以使其在做内容摄取决策时具有较大的灵活性。在 MetaArchive Cooperative 案例中，使用相同硬件的缓存都被配置了相同的文件系统结构。新缓存加入到网络中时，中心职员为成员机构提供一个启动文档。启动文档像 LOCKSS 安装程序一样运行，它们需要相同的本地安装设置。启动文档运行的结果就是给缓存配备了安全性能增强的 LINUX 而不是 OpenBSD，同时还有利用 SSL 加密通信 RPM 管理包和一个 LOCKSS 后台程序。

安全的缓存通信

已经设计出的 LOCKSS 缓存内部交流协议(LCAP)，即使在开放网络中也可以对抗攻击，PLNs 可以通过使用 SSL 进一步保护网络安全。在这种情况下，所有网络通信都是加密的、授权的。加密使得任何数据对外界都不可见，甚至网络设施本身（如路由器）也一样。授权提供了缓存身份的安全保证——通信只能在拥有身份证明密钥的缓存间进行。这种要求需要

PLN 管理者做一些额外的事情，包括创建密钥并安全地将其分发给每个参与站点，并且随着站点的增加或减少更新密钥（详见下文的最佳通信实践）。

内容服务器配置

当一台缓存被合理配置能够与其它缓存和中央管理服务器进行通信，它就可以使用 LOCKSS 后台程序从拥有资源的网络服务器上开始爬行和摄取内容（对摄入长期保存网络的内容的遴选和准备的讨论详见第五章：内容筛选、准备和管理）。网络服务器管理者需要确保 PLN 中的所有缓存都有权访问预定的内容。如果内容是可公开访问的，网络服务器管理者无需采取任何行动；但是如果内容是访问受限的，那么必须开放访问。为了使 LOCKSS 后台程序可以访问内容，本地系统管理员需要将所有缓存的 IP 地址加入到其网络安全设置，如网络服务器认可列表和防火墙中。

两种类型的 PLN 都必须具备有效和一致的程序使得所有 PLN 成员都知晓这些 IP 地址。如果没有以恰当的方式执行，没有考虑成员机构的网络服务器管理者，那么会有部分甚至所有缓存都将无法获取内容，LOCKSS 后台程序也将无法摄取、更新或保存内容。

MetaArchive Cooperative 已经通过维护一个最新的 IP 列表将这一过程流水化了，同时提供给网络服务器管理者 Apache 配置文档以帮助其自动更新进程。

网络管理

当缓存成功涉入 LOCKSS 系统的保存业务后，必须执行许多操作以确保适当地维护整个保存网络。本部分详细介绍了斯坦福管理的 PLNs 和本地管理的 PLNs 两种情况下的这类操作。

维护网络配置

中央管理服务器上的题名数据库中包含一些 XML 格式的参数，这些参数为 LOCKSS 后台程序提供三条重要信息：（1）作为签名文档插件的位置；（2）存储单元的位置和定义；（3）参与网络（测试和成品的）的缓存的 IP 地址列表。注意：这是可选的。可以使用一个种子列表，当新缓存加入到网络中时不需要为其更新列表。

在斯坦福管理的和本地管理的网络中，题名数据库和插件仓储库都是经网络控制的。一经启动，题名数据库的 URL 就会传递给每个缓存上的 LOCKSS 后台程序。后台程序从题名数据库中抽取插件仓储库的 URL。PLN 可能将题名数据库和插件设置在同一个网络服务器上；也可能由某台网络服务器单独持有题名数据库，而插件仓储库则遍布在由其成员机构维护的多台网络服务器上。当可以开放获取插件仓储库时，限制对题名数据库的访问是明智的。尽管缓存不容易受到牵连，但也没有必要暴露其 IP 地址。依赖 LOCKSS 团队管理其题名数据库的 PLN 必须与 LOCKSS 成员协同工作以确保这些信息得以恰当配置。

本地管理的 PLN 必须开发自己的程序来确保题名数据库和插件仓储库能够得到适当地维护，因为缺少更新信息会影响 LOCKSS 后台程序执行大量程序时的性能。主要的关注点在于，开发良好的程序可以确保：（1）加入或离开网络的缓存都有适当的责任（可选）；

（2）新插件或插件的变更能够很容易且一致地传送到插件仓储库中；（3）新的可获取的内容/AU 定义成为题名数据库的一部分，使得网络缓存可以发现新内容。

网络监管工具

PLNs 有许多监管工具，包括 LOCKSS 后台程序用户接口、缓存管理器、计划任务工作的执行。

LOCKSS 后台程序用户接口

LOCKSS 后台程序用户接口是由 LOCKSS 提供的基于网络的工具。本地系统管理员和中心职员可以登录到某一缓存的网络界面上查看系统运行时间、可用的和已用的磁盘空间，已保存在缓存上的存储单元的状态、大小、包含的文件以及最新爬取的信息。界面同样也会提供有关正在进行的和已完成的轮询和调查的信息。

LOCKSS 后台程序会将他们的行动记录在日志文件中，通过用户接口也可以获取到这些日志文件。简而言之，该接口提供了详细的信息，可以作为检验个体存储单元和整个缓存完整性的主要工具。

最重要的是，当 LOCKSS 后台程序无法爬取某一存储单元的内容时，用户接口可以提供有关爬取问题的详细信息。标准的网络管理最佳实践认为，持续不断地警示 PLN 员工多加注意这类问题，有利于快速且一致地发现并恢复故障，从而有助于维护网络拥有一个较高的轮询和调查成功率。

缓存管理器

缓存管理器是由 LOCKSS 团队和 MetaArchive Cooperative 共同开发的一个开源的、基于网络的工具。它查询个体缓存上安装的 LOCKSS 后台程序，收集有关 AUs 保存在什么地方、网络的整体状态、缓存存储容量等信息。

缓存管理器通过与中央和本地系统（或缓存）管理员的通信，可以告诉他们成功参与到 LOCKSS 后台程序轮询和调查进程的缓存数量，以及哪个缓存可能会超出存储极限而导致爬行程序出现中断。

计划任务信息

PLN 可以执行检索脚本，来发现那些适合其保存优先权的网络活动的特定信息。例如，计划任务工作（自动化的 Linux 指令或 shell 脚本）可以被指定在特定的时间和日期自动运行。本地管理的 PLNs 可以安装计划任务工作，使其提供预定的详细报告，这些报告内容主要是关于 LOCKSS 后台程序的轮询结果、缓存上存储单元的状态、管理服务器资源和目录的变更等信息。下述的样例信息可以安排在这样的报告中，MetaArchive Cooperative 通过这些机制的使用对其进行了论证：

完整的调查：保证有足够数量的缓存达成投票请求协议的成功率；投票的完成——以量的形式表示。

无法定人数错误：保证足够数量的缓存参与投票的失败率——以量的形式表示。

Fetch 错误：LOCKSS 后台程序爬取和再爬取 AU 的失败率——识别有问题的 AU。

不可提取许可页面错误：爬行程序取得资源显示或许可页面的错误率——识别有问题的 AU。

低重复警告：在一定数量的缓存间 AU 没有得到有效分发——识别有问题的 AU。

管理和有效利用工具

工具可以识别出问题，但是最终还是需要由工作人员采取措施解决这些问题。同样，不管网络使用什么工具都需要对其进行维护。

LOCKSS 团队负责维护由斯坦福管理的 PLNs。本地管理的 PLN 必须确定由谁负责启动和维护网络监管工具，指定可靠的软件工程师解决监管工具发现的问题。由试图访问内容的缓存所报告的爬取问题将可能需要由本地网络服务器管理员来解决；无法爬取内容的插件需要由插件开发者来处理；没有响应的缓存需要由本地缓存管理员重启。在本地管理的 PLN 中，这些解决措施都需要得到有效的指定和优先考虑。

管理服务器

题名数据库和插件仓储库都被托管在某一个管理服务器上。不管 PLN 是使用缓存管理器还是其它同样位于该服务器上的网络监管工具，最好将其置于不可广泛访问的环境中。通过网络访问题名数据库应当限定在 PLN 网络缓存和 PLN 成员范围内。例如，MetaArchive Cooperative 已经选择将缓存管理器、题名数据库、中央插件仓储库托管给一个单独服务器，该服务器近期刚刚迁移到亚马逊 EC2 云上。基本没有用户访问过该服务器。

本地管理的 PLNs 需要考虑在哪里放置这些工具、配置文档以及更新组件的相关文档。而 LOCKSS 团队会代替由斯坦福管理的 PLNs 考虑这些注意事项。

任何网络监管工具都需要查询 LOCKSS 缓存的状态。因此缓存必须允许来自控制网络监管软件的服务器的访问。PLN 缓存的系统管理员需要确保其本地网络设置允许访问 LOCKSS 后台程序的用户接口端口。LOCKSS 后台程序使用的端口是其本地配置的一部分。在斯坦福管理的 PLNs 中，该位置由 LOCKSS 团队设置。本地管理的 PLN 应当在所有缓存中使用统一的端口号。

本地管理的 PLN 最佳案例建议，管理服务器不应受任何单一成员站点控制，因为这会使得网络依赖于该成员的持续参与。相反，它应当在 PLN 的中心位置接受管理。如前文提到的 MetaArchive Cooperative，已经使用了亚马逊 EC2 的云服务器。

管理服务器备份

PLN 应当确保中央管理组件没有被绑定在单独一家成员机构上。这也引出了备份策略的问题。对本地管理的 PLN 较有吸引力的两种备份策略是：在云环境中配置一个备份服务器；与某个独立的保存网络（PLN 或其它形式）合作建立备份服务器。

作为本地管理的 PLN，MetaArchive Cooperative 正在通过使用亚马逊 EC2 和与圣地亚哥超级计算机中心 NDIIPP 资助的 Chronopolis 项目（主管运营 SRB（Storage Resource Broker）

基础设施的分布式数字资源长期保存网络）合作，来践行这两种策略。

交流最佳实践

显然，分布式数字资源长期保存要求在众多位置和成员间进行交流和合作。成功建立和维护一个 PLN，需要及时地与成员机构和中心职员协调，他们可能会有助于网络的监管和维护。

文档

记录建立和维护 PLN 时的开发和决策制定，对于不论是技术团体还是管理机构来说都是至关重要的，因为数字资源长期保存解决方案的成功在一定程度上依赖于它们的可获取性、透明性和可说明性。下面有一些为 PLN 成员和有兴趣的公众提供的管理文档的工作复本和完整版本的建议：

Wikis：维基软件（例如 MediaWiki）允许在文档开发方面进行合作，可通过配置进行内部和公开的出版。

公共网站：网站可以为 PLN 提供发布权威文档和指南文档的平台。

内容管理系统：强健的和模块化的网站开发软件（如 Drupal）可以提供管理 PLN 重要信息的多层级的用户许可。

注意事项

能否确保合作站点和任何中央管理位置的技术和管理职员能够及时地接收到网络中的问题警告关系着 PLN 的成败。影响组织机构的注意事项必须能够及时地、随时地、安全地得到公示，以巩固决策制定和网络政策的实施。

确保及时、安全的通信的建议如下：

专用邮件列表：专用邮件列表可由某一合作伙伴站点管理，或者如果性价比划算，可委托给外包服务；专用邮件列表应当专注于 PLN 组织的某一项工作（如技术讨论组、管理讨论组、甚至可能是一般的通信讨论组等）。

标签系统：项目管理和故障/问题追踪系统（如 Trac）可以使合作伙伴站点的技术职员能够请求反馈，请求中心职员优先考虑某些问题并及时回应。

会议电话：合作伙伴站点的中心职员和技术及管理职员定期安排的电话有助于对 PLN 范围内重要问题的考虑。

加密密钥：允许对在公共网络上传送的电子文档和信息进行加密和解密。PLN 参与成员需要发送诸如用户名和密码之类的敏感机密信息时，密钥可能就会派上用场。加密密钥最好由人工亲自分配和发放，或者是烧制在写保护的媒介上然后通过邮局邮件的方式寄送。

会议和研讨会

为确保 PLN 成功催化其成员机构的集合知识和专门知识，面对面会议将是有益的：

年会：为合作伙伴机构的管理阶层提供了参与重大决策的机会，同时随着 PLN 的发展不断完善保存计划。

研讨会：这有助于 PLN 成员间的技术知识和专业知识的传递，考虑焦点问题的解决方案，巩固最佳实践。

结论

选择依赖 LOCKSS 或单独构建 PLN，是区分两种 PLN 类型的关键因素，而不同的 PLN 类型又会导致在管理需求/选择上的差异。但是，每种类型都会牵涉到一系列的在缓存和网络管理层面上的考虑和责任。这些都需要力图管理 PLN 的机构进行慎重权衡和考虑。

编译自：http://www.metaarchive.org/GDDP

（齐燕 吴振新校对）

【技术与工具】

Terrier, IR Platform v3.5

Terrier Team 著　　刘晓敏　编译

Terrier 是由格拉斯哥大学计算机科学学院开发的一款基于 Java 的开源软件。作为一个灵活度高、高效的开源搜索引擎，可方便地部署于大规模文档集之上。Terrier 的索引和检索功能采用了当前的最新技术，可为大规模检索应用的开发和评估提供一个理想的平台。作为一个开源平台，通过提供诸如 TREC 和 CLEF 等标准测试集，Terrier 可以为文本检索领域的实验研究提供综合、灵活和透明的平台，从而让检索研究得以更容易地进行。目前 Terrier 已经更新至版本 3.5，可从 Terrier 官方网站获取该软件源码。

功能特征

一般功能

支持对普通桌面文件格式的索引，同时也支持对常用 TREC 研究集（如：TREC CDs 1-5，WT2G，WT10G，GOV，GOV2，Blogs06，Blog08，ClueWeb09）的索引；

支持多种文档权重模型，如基于随机的多参数自由分割权重模型、Okapi BM25 和语言模型；

支持传统检索语言，包括短语检索、基于标签的词项检索；

支持大规模文档集的全文索引，使用 Hadoop MapReduce 分布式索引技术，可在集中架构环境下对至少 500 万篇文档进行处理；

模块化，开放索引和查询 API，可方便对自己的应用和研究进行扩展；

开放源码；

跨平台支持，可在 Windows、Mac OS X、Linux 和 Unix 环境下运行；

索引功能

基于标记文档集（如 TREC 测试集）的即开即用索引方法；

针对多种文件格式的即开即用索引技术，可处理诸如 HTML、PDF、Microsoft Word、Excel、 PowerPoint 等类型文档；

支持 Hadoop MapReduce 环境下的分布式索引；

支持对字段信息的索引，如某一 HTML 标签中的词项频率；

支持位置信息索引；

支持多种文档编码和多语种检索；

支持词法分析的变形；

支持对查询偏颇报告（query-biased summarisation）的索引；

通过 HTTP 协议对提取文档进行索引，让内部网络更容易被检索到；

对索引进行高比率压缩；

对文件进行高比率压缩以支持高效查询扩展；

高速度的单次索引和基于 MapReduce 的快速索引；

支持多种词干合并技术，包括对欧系语言的滚雪球（Snowball Stemmer）词干合并技术。

检索功能

支持桌面检索、命令行检索和基于 Web 的用户界面检索；

提供标准的查询工具和查询扩展机制（伪正例反馈）；

可应用于各种交互式应用，例如嵌入桌面检索中或在研究和试验过程中进行批量设置；

提供多种标准的文档权重模型，如超过 126 种 Divergence From Randomness (DFR) 文档排序模型，以及诸如 Okapi BM25，语言模型以及 TF-IDF 权重模型。此外还包括两种新的第二代 DFR 权重模型和 JsKLs、SqrA_M 模型。 这些模型在测试集中因展示了优秀的性能而无需任何的参数调整或训练。

支持同义词替换、+/-运算符、短语查询、邻近检索以及字段检索；

除 Rocchio 查询扩展以外，还为自动查询扩展提供无参数 DFR 词项权重模型；

通过一系列组件提供灵活的词项处理过程，例如停用词去除以及词干还原。

实验功能

支持当前可用的所有 TREC 测试集；

通过简易的脚本编写或使用批处理形式提供多种权重模型来评估各种参数设置；

通过使用带有 TREC ad-hoc 和已知项目的检索结果的嵌入式评估工具来产生各种准确率和召回率测度。

组件交互

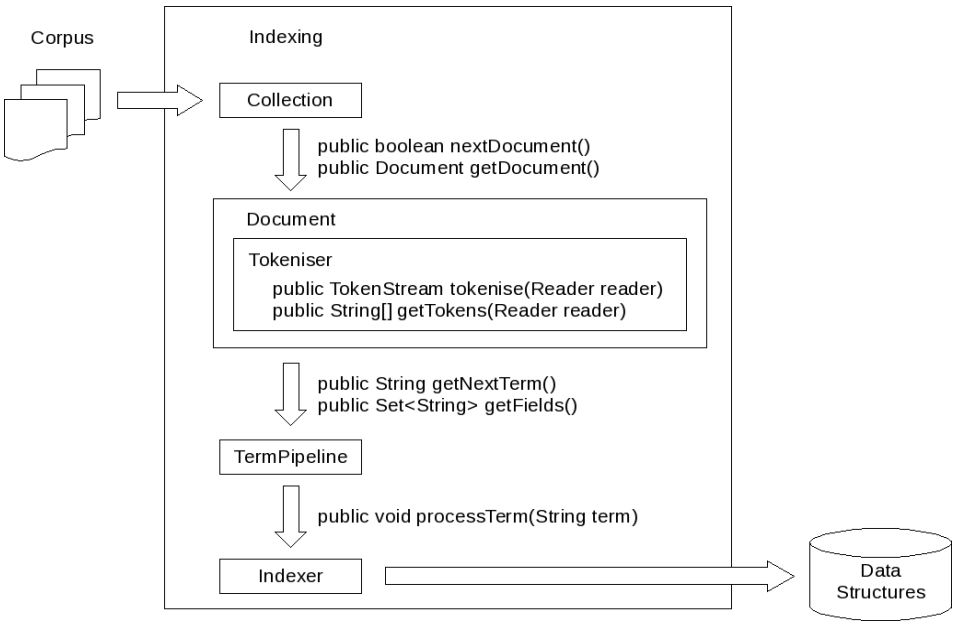
索引阶段的组件交互

语料库（Corpus）将会以集合对象的形式展现。原生文本数据以文档对象的形式展现。

文档则由词法分析器的一个实例提供，通过词法分析器可以将文本分割成单个索引标记。

索引器（Indexer）负责管理索引进程。索引器对集合中的文档进行迭代操作，通过 TermPipeline 组件发送每一次发现的词项。

TermPipeline 可以对待索引的词项实行继续转换或移除操作。TermPipeline 链的一个例子是：termpipelines=Stopwords,PorterStemmer。即先使用停用词表将文档中的词项移除，然后对其他词项应用 Porter 的词干还原算法（PorterStemmer）。



索引阶段组件交互

一旦经过 TermPipeline 的处理，词项就将被聚合，并通过词项所对应的文档创建者（如 DirectIndex、DocumentIndex、 Lexicon 和 InvertedIndex）应用不同的数据结构；

对于单遍索引，将采用不同的顺序来编写数据结构。倒排文档记录表在内存中进行构建，当内存即将耗尽的时候再从内存中调出，以索引块的形式存入硬盘。一旦集合索引完毕，所有的索引块被合并为一个倒排索引和词典。

检索阶段的组件交互

某一项应用程序，如 Terrier 的桌面应用或 TrecTerrier 应用，向 Terrier 框架提交一个查询；

一开始，查询将会被解析，同时将会触发一个查询对象的实例；

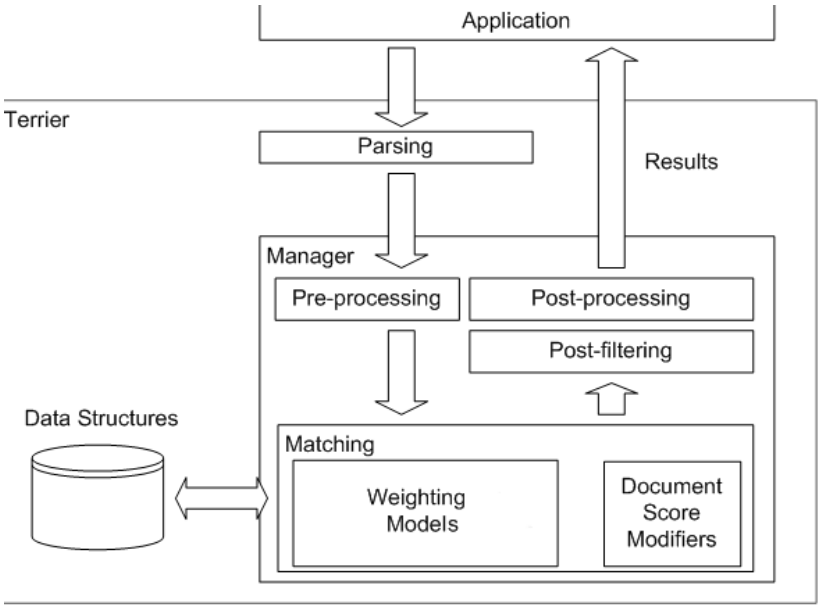
随后查询将会被提交至管理组件（Manager component）。管理组件将会通过配置好的 TermPipeline 组件对查询进行预处理；

预处理结束之后，查询将会被提交给匹配组件（Matching component）。匹配组件负责对合适的权重模型（Weighting Model）和文档评分改进组件（Document Score Modifiers）进行初始化。一旦这些组件都被实例化之后，将进行文档与查询之间的匹配计算；

随后，进行后处理（Post Processing）和后过滤（Post Filtering）操作。在后处理阶段，

可能会对结果集进行稍许的修改，例如通过对查询进行扩展之后再调用匹配组件来生成和改进文档排名。后过滤阶段是一个简化阶段，可以对文档进行包含和排除操作。这对于交互性应用而言尤其合适，因为用户总是想要进一步限制检索到的文档的范围。

最终结果集返回至应用组件（Application component）。



检索阶段组建交互

组件描述

索引组件描述

名称	**描述**
集合	该组件通过使用 Terrier 将最重要的概念囊括到索引中。一个集合就是一组文档。详情参见 org.terrier.indexing.Collection
文档	该组件将一个文档中的概念囊括其中。需要对文档中的词项进行迭代处理。详情参见 org.terrier.indexing.Document
词法分析器	由文档对象调用，将连续性文本（如一个句子）分割成一组单词并插入到索引中。详情参见 org.terrier.indexing.tokenisation
TermPipeline	对词项处理过程中的一个组件中的概念进行建模。实现这一接口的类主要有词干还原算法、停用词移除或首字母缩略词扩展。详情参见 org.terrier.terms.TermPipeline
索引器	该组件负责管理索引进程，对 TermPipeline 和生成器进行实例化。详情参见 org.terrier.indexing.Indexer
生成器	生成器负责将索引写入磁盘。详情参见 org.terrier.structures.indexing

	package

数据结构描述

名称	描述
BitFile	使用了伽马和一元编码的高压缩比率 I/O 层。详情参见 org.terrier.compression
直接索引	直接索引存储出现在每一个文档中词项的标识符以及词项频率。可用于查询自动扩展，同时也可用于用户活动描述。详情参见 org.terrier.structures.DirectIndex
文档索引	文档索引中存储关于文档的详细信息，如文档长度、标识符以及索引中指向对应条目的指针。详情参见 org.terrier.structures.DocumentIndex
倒排索引	倒排索引中存储倒排记录表，例如文档的标识符以及对应的词项频率。此外，还可以存储词项在文档中的位置信息。详情参见 org.terrier.structures.InvertedIndex
词典	词典中存储的是词汇集合以及对应的文档和词项频率。详情参见 org.terrier.structures.Lexicon
元索引	元索引中存储关于文档的额外信息，例如其独特的文本标识符或 URL。详情参见 org.terrier.structures.MetaIndex

检索组件描述

名称	描述
管理组件	该组件负责处理和协调一个查询的主要高级操作。包括： 预处理（Term Pipeline、发现控制、词项聚合）； 匹配 后处理 后过滤 详情参见 org.terrier.querying.Manager
匹配组件	匹配组件主要负责决定哪一篇文档与一个特定的查询相匹配，并计算文档与查询的匹配程度的得分。详情参见 org.terrier.matching.Matching
查询组件	查询组件对查询进行建模，包括子查询和查询词项。 详情参见 org.terrier.querying.parser.Query
权重模型	权重模型可以展现检索模型，该模型可用来评判文档中某一词项的重要

	程度。详情参见 org.terrier.matching.models.WeightingModel
文档评分改进组件	负责根据查询对文档得分进行改进。 详情参见 org.terrier.matching.dsms package

应用描述

名称	**描述**
Trec Terrier	可以支持对 TREC 测试集进行索引和检索的应用。 详情参见 TrecTerrier
Desktop Terrier	可支持对本地用户内容进行索引和检索的应用。 详情参见 org.terrier.applications.desktop package
HTTP Terrier	支持通过浏览器进行文档检索的应用。 详情参见 src/webapps/results.jsp 或 relevant documentation

编译自：http://terrier.org/

（么媛媛 吴振新校对）

【动态追踪】

HathiTrust 相关动态跟踪

A．HathiTrust全文数据库现可通过EDS进行检索

EBSCO 出版部的 EDS（EBSCO Discovery Service™，EBSCO 发现服务）的基本目标是为终端用户提供优质的馆藏目录，EBSCO 出版部和 HathiTrust 最近签署的合作协议允许在 EDS 中对 HathiTrust 的内容进行全文搜索，即向 EDS 用户们公开 HathiTrust 的所有收藏，这极大地扩展了他们的搜索结果，同时也最大限度地利用了馆藏资源。

HathiTrust所收藏的数字资源来自超过50家的主流研究机构和图书馆。HathiTrust数字图书馆为了长期保存这些海量的数字馆藏，而将其合作机构的资源整合在一起。该数字图书馆很大一部分馆藏是与Google图书协议生产的数字化书籍，另外还有来自HathiTrust的一些合作伙伴，如大学出版社和个人图书馆馆藏的数字化内容。

HathiTrust巨大的数字化仓储深入地扩展了检索用户的搜索体验。尽管HathiTrust也提供了独立的搜索服务，但通过EDS，用户们很快就能从HathiTrust的超过850万册数字图书中进行搜索，还包括950万册图书，500万条书名和25万条丛书名，而这些仅作为EDS的部分内容。

为了使这些全文能够在EDS检索，重新生成了这些记录并推送到用户的检索前端，使用户能够发现更多的数字集合。

编译自：

http://www2.ebsco.com/EN-US/NEWSCENTER/Pages/ViewArticle.aspx?QSID=492

B．HathiTrust全文索引将被集成至OCLC服务器

OCLC与HathiTrust已签署一份允许OCLC将HathiTrust全文索引整合编入OCLC服务的协议，OCLC成员馆及各馆用户将能更加方便快捷地通过WorldCat检索到这一重要数字集合中的资源。这个全文索引的整合完成之后，用户们将不仅仅能搜索到著书目录信息，还能在这些合作馆藏中进行全文检索。

作为一个国家级顶尖研究型图书馆的数字仓储，HathiTrust数字图书馆把来自各个合作机构的数量庞大的资源整合在一起。HathiTrust为这些图书馆提供了一个存储和访问其数字内容的方法，不管是扫描图书、特殊馆藏还是原生数字资源。这些资源的数字化最大限度地为科研、教学与学习的创新应用提供了机会。

编译自：http://www.oclc.org/us/en/news/releases/2011/201150.htm

C．提供对孤本的访问服务

杜克大学、康奈尔大学、埃默里大学、约翰霍普金斯大学和加州大学在八月份宣布，其机构用户将有权访问在 HathiTrust 中的一些孤本书籍，这些孤本的副本被收藏在这些大学的图书馆中。HathiTrust 已给出约 160 册准孤本作品的名单并资助密歇根大学进行为其确定年代的试点工作。HathiTrust 的孤本总量是未知的，但 HathiTrust 的执行理事乔恩·威尔金估计孤本数量占总藏书量的比例可能高达 50%。目前发现的准孤本作品名单已被列在公共联机目录上，如果信息发布后 90 天内没有任何人能证明自己是其版权人的话，它们将被确定为孤本。随着 HathiTrust 对孤本的鉴定从试点阶段向正式实施阶段过渡，其范围预计将扩大到多个拥有类似现有版权审查管理系统的机构。

编译自：http://www.hathitrust.org/updates_august2011

D．HathiTrust 研究中心获资助调查非消耗性的研究

HathiTrust研究中心（HTRC）从Alfred P. Sloan基金会获得60万美元的资助，进行以众多数字馆藏为主要对象的非消耗性研究（non-consumptive research）的首次调查。非消耗性的研究涉及到对一本或多本书的计算分析，研究人员往往没有能力重组所有藏书，他们不去阅读资料，而是用专门的算法来分析一个庞大数据集的文本。Sloan的资助将有助于确保这些工作可以在一个安全的环境中进行。

这笔资金将使HTRC继续围绕利用HathiTrust数字资料进行的非消耗性研究进行调研。该团队期望在项目结束时已经建有泛在的计算机网络基础设施，以此来成功地证明非消耗性研究能在非故意的恶性用户算法条件下安全地进行。

在某些情况下，研究人员运用的算法式所有权归属于HTRC，所以HTRC需要研究用户、

算法和数据的安全性要求，所有这些都在运用SEASR（Software Environment for the Advancement of Scholarly Research）中的一整套算法的背景下。其他情况下，研究人员有自己算法的所有权并提交这些算法式以便运用，Sloan基金将被用来构建一个“数据概要框架”原型系统让被资助者在庞大的信息集合中自由地实验各种新算法，但需存在一个“相信但需证明”的技术机制来确认其是否遵守了非消耗性研究的政策。

不考虑材料的实际内容，研究人员们用他们自己的复杂算法可以分析很大规模的数据集，从不断重复同样几个词语的简单情况，到复杂的语言学的结构或者是某一时间和空间范围内词语运用的演化发展，甚至是人口统计分类等。HathiTrust知识库拥有超过950万册数字资源，其中大概27%，即约250万册都属于公共领域并且能够用于非消耗性研究。实施非消耗性研究的模型建立在相信但需核实原则的基础上，研究人员默认是可以信任的并且拥有开展创新性研究的自由，但需要有相应的机制来保证其良好的行为和遵守规定。

编译自：http://www.dlib.org/dlib/september11/09inbrief.html

（么媛媛编译，齐燕 吴振新校对）

DuraSpace 推出开源云服务

DuraCloud 对学术、科学和文化遗产的长期保存实践推动了人们对云技术的认识。

致力于保存世界学术、科学和文化遗产的非营利组织 DuraSpace 于 2011 年 11 月 1 日宣布其托管的云服务 DuraCloud 正式上线。作为唯一一种允许组织机构通过多个云服务提供商存档其内容的托管式软件服务，DuraCloud 能保证那些不可替代的文档、影像和视频的可获取性。全国最负盛名的一些机构，包括麻省理工学院、哥伦比亚大学、西北大学和莱斯大学都已签署了使用托管式云服务来保护数字资源的协议。

关于数字图书馆、学术机构、博物馆和其他知识管理单位研究成果的存档，DuraCloud 提供了：

- 通过一个统一的接口实现多个云服务供应商间的内容复制和同步；
- 用户可使用嵌入 DuraCloud 平台的一系列应用程序来更好地处理数据；
- 将数据分布保存到任何连接互联网的设备中；
- 对数字文件进行安全存储，并对其内容进行周期性的“健康检查”以确保其原始完整性；
- 提供一个简单易用且功能强大的控制面板，来管理云中所有内容；
- 有一个致力于云技术持续发展的开源社区，同时提供向云端转换的协助和支持。

DuraSpace 的部分资金由美国国会图书馆通过其国家数字信息基础设施和保存计划（NDIIPP）来提供，致力于开源社区发展和云技术进步的研究。目前的开源项目包括 DSpace 和 Fedora。

DuraCloud 既是一个免费的开源项目也是一个收费的订阅服务，它能让任何希望对其内

容进行归档和保存的组织使用云服务，而不必受限于单独一家云服务提供商。

编译自：**http://duraspace.org/duraspace-launches-open-source-cloud-service**

（么媛媛编译，齐燕 吴振新校对）

【信息扫描】

POCOS 项目：应对复杂可视数字资源长期保存的挑战

复杂可视资源长期保存的专题讨论会（POCOS）是由英国联合信息系统委员会（监委会）资助的一个新项目。POCOS 项目(http://www.pocos.org)通过在英国的一些特定地区开展一系列的三个专题讨论会来应对其在复杂可视数字资源长期保存过程中所面临的巨大挑战。POCOS 主要关注复杂的资源类型和其各自的数字化长期保存问题的三个相互关联的领域：可视化和模拟、软件艺术、游戏环境和虚拟世界。这三个领域的研讨会突出了复杂资源长期保存中协同作用的重要性，并能同时考虑到每个领域不同的数字资源长期保存理论和实践。POCOS 研讨会汇集了在这些领域的顶尖的研究人员和从业人员，并邀请他们展示各自的研究成果，确定一些尚未解决的关键问题，并规划出日后关于保存复杂资料和环境的研究议程。

Planets 项目和 KEEP 项目最近的研究揭示了与复杂可视资源长期保存相关的很多实际问题，尽管并不是很棘手。POCOS 意识到，为了不断进步，必须要吸引和激励更多的数字资源长期保存团体参与进来并发挥作用。其中一个重要方面就是要通过综合分析至今为止的多个独立项目所取得的研究成果来准确地表述现有的技术水平，并明确指出仍需努力的领域。

POCOS 成员包括来自学术界、商界、存储机构的杰出的研究人员、开发人员和从业人员，他们在数字资源长期保存研究领域中有着国际公认的优秀才能。POCOS 还有很多优秀的合作者。

目前为止，POCOS 已经成功举办了首次可视化& 模拟研讨会（相关会议资料见 http://www.pocos.org/index.php/publications/webcasts）。第三场关于游戏环境的会议将于 2012 年 1 月在加地夫举行。

编译自：http://www.dlib.org/dlib/september11/09inbrief.html

http://www.pocos.org/index.php/pocos-symposia

（么媛媛编译，齐燕 吴振新校对）

第七届国际数字管理会议

2011 年 12 月 5 日至 7 日，第七届国际数字管理会议将在英国布里斯托尔召开。数字保管工作包括管理、维持、长期保存，贯穿数字信息整个生命周期的价值增值，减少对数字资源长期价值的威胁，减少数字资源退化的风险并提高其对研究和教学的效用。

本次会议将会选择一组知名专家做开幕报告和闭幕演讲。NoTosh 创始人和首席执行官伊万 麦金托什将发表关于“公共数据的机会”的演讲。《公共科学图书馆 计算生物学》的首席主编菲利普 E 布恩教授将以一个“2020 年开放数据驱动的学术交流”的研究开始第二天的会议。会议闭幕式的主题演讲——“模型、科学、开放”——将会由微软亚洲研究院计算科学的专家史蒂芬 爱莫特教授给出具体内容。

会议计划还包括全体受邀专家一起在某个下午进行相互交流，“社区空间”的海报宣传，以及有关个人基因组学的介绍、一些非正式会议和一次研讨会。

会议的理论方向包括对涵盖数字保管的所有研究和知识的同行评议论文的展示和讨论；实践方向包括来自不同机构、研究领域或地区的实践经验的展示。

详细内容参见：http://www.dcc.ac.uk/news/key-speakers-announced-idcc11

编译自：

http://www.dpconline.org/newsroom/whats-new/752-whats-new-issue-37-september-2011

（齐燕编译，刘晓敏 吴振新校对）