图书情报工作
第58卷 第5期 2014年3月

数字资源的持久标识符互操作参考模型构想

刘振

[摘 要] 认为国外持久标识符的研究存在研究对象过于具体、应用范围有限等不足，很大程度上是由于没有在一个整体框架指导下来开展研究所致，鉴于此，提出一个持久标识符互操作参考模型，模型主要组成部分是持久标识符域，包括登记机构、内容提供商以及解析器等，对三类数字资源实体(数字对象、作者、机构)和持久标识符之间的关系进行标准化，便于管理、表示和呈现数字资源，并可为实现正确的互操作解决方案和交换奠定基础。最后，指出本模型在安全性、可伸缩性等方面有待进一步完善和改良。

[关键词] 持久标识符 互操作 参考模型
[分类号] TP393
DOI: 10.13266/j.issn.0252-3116.2014.05.014

1 引 言

对数字对象(例如文章、数据集、图像或数据流)进行持久标识，可以使这些数字资源的定位和访问具有唯一性，把它们与相关的作者和其他实体(例如机构、项目或研究团体)相联系，以便使其得到可持续、可靠的发现、引用和重用。不同的领域使用着不同的持久标识标准，数字对象的持久标识系统有URN、DOI、PURL、ARK等，B. Bazzanella等[1]学者的调查研究表明，DOI、Handle System和URN是目前应用最广泛的系统，在欧洲范围内DOI占33%，Handle System占29%，URN占25%，而其他的系统(如PURL和ARK)只被少数人使用。有些作者识别符系统和项目最近几年也开始出现，例如AuthorClaim、Scopus Author ID、researcher ID、ORCID等，NISO机构识别工作组在信息供应链的环境下开始关注机构识别的标准制定。

在数字资源的生命周期中，根据数字资源管理机构的识别需要，一个数据对象能够被赋予多个持久标识符(persistent identifier，PI)，例如出版时被赋予一个DOI，提交给机构数据库被分配一个handle，各个持久标识符服务互相补充，就会出现部分重叠现象。因此，有必要在这些系统之间建立起一个互操作框架，满足不同领域、不同群体的识别服务需求，这对于保存、管理、访问和重用海量数据至关重要，也是整个信息社会的一个关键问题。

N. Paskin[2]为持久标识符互操作定义了3个层次：语法互操作性(系统处理语法字符串的能力，即使不同的系统采用了不同的语法模式，也可以识别出来)、语义互操作性(系统确定两个不同的标识符是否表示同一个指示物。如果不是，两者是否有关系)、交流互操作性(系统使用标识符合作和沟通的能力)。

PLIN项目[3]给出的定义是：在标识符管理系统的控制范围外对一个组件进行操作，这个组件就是可互操作的。这个操作必须在一个友好的界面下完成，如果这个组件不是互操作的，那么只有自己的标识符管理系统基础设施可以对其进行操作。如果操作可以通过开放协议(如Web服务)使用一个公共记录界面进行，那么它就是互操作的。

2 持久标识符互操作的相关项目研究

近年来，已有几个计划和项目开始解决PI互操作性的问题，提出在不同的环境下面对一些在标识符或元数据层次上的问题的解决方案。一些项目专门关注数字对象的PI互操作问题(例如PILIN)，而另一些项目则提出了作者标识符互操作问题(例如ORCID)。

2.1 ORCID

ORCID[4]即开放的研究者和参与者ID。这是一个有关学术交流的关键利益相关者(大学、资助机构、出版商、研究所)的项目。它的目标是建立一个所有研究人员的注册中心：持久、清晰和明确地记录作者和撰稿

* 本文系国家社会科学基金“数字资源长期保存技术的研究与实践”(项目编号：09FTQ005)和国家“十二五”科技支撑计划项目“面向外文科技文献信息的知识组织体系建设与应用示范”(项目编号：2011BAH10B00)研究成果之一。

[作者简介] 刘振，徐州工程学院管理学院讲师，中国科学院文献情报中心、中国科学院大学博士研究生，E-mail: liuzhen@mail.las.ac.cn。

收稿日期：2014-01-16 修回日期：2014-02-16 本文起止页码：82-85 本文责任编辑：刘远颖

表 1 持久标识符互操作的项目概况

项目名称	标识对象	涉及范围	资助单位
ORCID	作者	欧洲	私立
PersID	作者、数字对象	欧洲	SURF 基金
OKKAM	作者、数字对象	欧洲	欧盟委员会
PILIN	数字对象	澳大利亚	大学
CORES	元数据	欧洲	欧盟委员会
RIDIR	数字对象	欧洲	JISC
LATTES	作者	巴西全国委员会	政府

者的学术交流。ORCID 将从研究人员、参与者机构或其他第三方得到信息。这些概要文件包含生物/书目信息。研究人员能够创建、编辑和维护一个 ORCID ID 和免费资料，并能控制自己的 ORCID 概要文件数据的隐私配置设定。该基础设施能和所有 PI 相关项目实现互操作和这些项目相互利用和补充。ORCID 得到全球 250 多个组织的支持，因这些组织看到了这样一个基础设施的存在价值：大学将能够轻松地作为一个整体来整理他们的科研成果，资助机构将能够跟踪他们的投资结果，出版商将能够丰富他们的投稿工作流。ORCID 的应用不仅限于作者：最终这个新兴数据基础设施将建立起人、角色和数据集之间的唯一联系。

2.2 PersID

PersID 项目[5] 旨在支持持久访问学术和文化信息。PersID 的持久标识符使用了 URN：NBN 方案，目的是在 NBN 全国范围内创建一个全球的解析器基础设施。它构建在已经广泛使用的技术和标准——IETF RFC3188 上。该项目由不同的组织共同参与，这些组织负责大量的出版物、文化材料和研究数据的长期保存。还有 10 个项目合作单位，主要是各国国家图书馆。PersID 政策要求合作单位要确保标识符所表示的数字资源的长期保存。PersID 是 SURF 基金会资助的一个项目。

2.3 OKKAM

OKKAM 项目[6] 为了实现实体的全球标识符系统的重用，创建了一个名为 ENS（实体名称系统）的基础设施。ENS 允许给一个给定的实体分配一个全局标识符，把这个实体和其他的可替代标识符联系起来。基本想法是相同的实体可以在不同的上下文中用不同的标识符进行标识。OKKAM 创建了自己的全球标识符，还有一个用来消歧的简介文件和标识符相关联。简介文件的部分内容是其他系统创建的标识符集合。这样，给定一个 OKKAM ID 或任何指代的其他 ID，可以收集一个指代的入口点；事实上，解析工作一般由维护指代本身的系统来完成，而 OKKAM 提供了一个全球 OKKAM 标识符的解析器，用来返回指代的概要文件。

2.4 PILIN

PILIN（PIs 和连接基础设施）项目[3] 是一个由 ARROW 和南昆士兰大学领导的国家级项目，它于 2007 年结束。PILIN 项目的目的是建立一个基于 CNRI 处理技术的可持续的共享标识符管理基础设施，以便保障可持续发展的全局标识符，使得标识符和相关服务能够持久存在。项目的主要目的是满足在澳大利亚电子学习、电子研究和电子科研中普遍存在的可持续的标识符基础实施的专门需求，来处理不同的电子科研环境下产生和存储的大量数字资料。

2.5 CORES

CORES 项目[7] 是一个关于元数据互操作性的欧洲项目。该项目的核心目的是鼓励共享元数据语义。通过建立一个达成共识的声明元数据术语的语义的数据模型，该项目旨在使现有标准在一个集成的、机器可以理解的语义 Web 环境中一起工作。为实现这一目标，项目启动了一个互操作性标准论坛，汇集重要标准化活动的关键人员，来讨论标准之间互操作性的实用性。项目的主要成果之一是一个基于共同模型的注册中心，用来声明和共享元数据模式。

2.6 RIDIR

RIDIR（数据仓储资源标识符互操作）项目[8] 是一个联合信息系统委员会数据仓储和保存计划组资助下的项目。其正在调查 PI 的需求、优势和使用情况，以便提高不同类型的数字存储库之间的互操作性。RIDIR 项目的主要目标是充分了解标识符团体的需求，然后建立一个完全有效运行的示范系统，并提高对 PI 互操作性问题的认识。这个项目不是关于共享的 PI 服务，而是专注于 PI 的使用。作为示范，他们设计了一个查找丢失资源的服务系统，当一个 PI 失效时，该系统允许用户把这个 PI 对应的资源重定向到新的正确的位置。该位置也可以由数据仓储管理员指定。

2.7 LATTES

LATTES[9] 是巴西国家科学技术发展委员会（CNPq）推动的政府级项目。项目建立了一个课程信息系统（LATTES CV 系统），这个系统是在 CNPq 和拥有发布科技信息的数据库和网站的机构之间达成的一系列协议的基础上建成的。系统的目的是收集参与科学和技术发展的所有机构人员的相关信息。该系统主要用于以下 3 个方面：①奖学金或研究支持的候选人的能力评价；②顾问、委员会和顾问团体的成员的挑选；③巴西研究生以及研究成果的评价。

2.8 总体分析

目前，多数研究项目主要致力于解决某些局部范围

的持久标识符互操作的问题，或者是研究对象的特定性，或者是适用范围的限制性，多数是零散的、不完整的，例如 ORCID 主要面向学术交流，CORES 研究的是元数据互操作性，PILIN 主要满足澳大利亚的电子科研的需要，LATTES 主要服务于巴西科技机构。目前最普遍的、跨领域的方法是开放文档，没有采用开放文档的机构或组织会采用其他的解决方案，由于还没有一个唯一的全球标识符被统一采用，只能在目前的各个不同的持久标识符系统中寻求一个互操作性方案。互操作服务是在复杂的社会和组织环境下来实现，如何在各个不同的持久标识符系统的范围内识别基本的概念、属性以及概念之间的关系，以便实现跨系统的交流，这需要一个整体框架即互操作参考模型，该参考模型要定义一个公共的语义，即一个公共的概念表示，不涉及任何的持久标识符互操作的标准、系统或具体技术实现，该参考模型是设计互操作服务的基本前提，用来支持互操作服务的开发以及有关活动，突破各自持久标识符系统的边界，为各个不同的系统的信息交流建立一个公共层，以便产生新的跨系统的互操作服务，确保数字资源被可靠持续地访问，有利于实现资源的长期保存。

3 持久标识符互操作参考模型构想

可尝试定义一个框架，来建立持久标识符系统之间互操作的条件和环境。持久标识符互操作参考模型应该能提供一个公共的、顶层的、不同的 PI 系统之间的框架，便于管理、表示和呈现数字资源，在不同持久标识系统下的数字资源的交流、重用和集成，实现不同的 PI 系统之间的互操作，支持设计和开发新的互操作服务。持久标识符的主要应用对象有数字对象、作者和机构 3 类，所以模型要对这 3 类实体和它们的持久标识符之间的关系标准化，为实现正确的互操作解决方案和交换奠定基础，系统设计者可以使用这个抽象的参考模型，作为模板，设计基于它的不同的互操作技术方案和服务，见图 1。

在参考模型中，最终用户、数据提供者、搜索引擎、目录服务等各种终端应用通过调用各个互操作服务的 API 接口，完成不同的 PI 系统之间的交流和沟通。互操作服务的提供者就是参考模型中最重要的组成部分——PID(persistent identifier domain) ，PID 是用户和服务提供商的系统，负责为各种类型的相关实体(数字对象、作者和机构) 分配 PI，分为数字对象 PID、作者 PID 和机构组织 PID 3 类，对于任何一个数字资源，除了一些描述性的元数据以外，持久标识符域应该声明

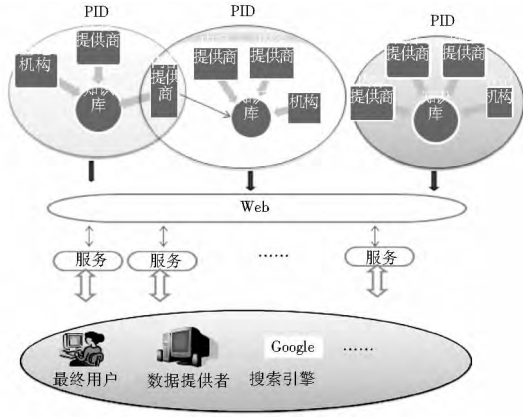

图 1 持久标识符互操作参考模型

一个已经存在的持久标识符以及与域内其他对象的关系。PID 主要包括: ①登记机构: 负责管理 PI 的分配和登记，为登记者声明和维护自己的元数据提供必要的基础设施，确保 PI 的持久性以及与被标识的数字资源的关联。这些系统提供的服务包括质量保证、措施保障等基本需求，例如，PI 在一个领域内的独特性或相关数据的正确更新。②内容提供商: 使用 PI 存储、管理和保存数字对象的机构。③解析器: 能够识别 PI，在 PI 和关于对象和当前位置的信息之间建立起关联的系统，解析器在互联网上能被访问到，它可以把一个 PI 解析成一个单独的对象(网页或文件) ，或者对象和元数据，或者多个对象(同一个对象的不同格式或者内容形式) 。例如对于同一篇文章，它的 PDF 版本和 HTML 版本在 DOI 领域内能够被看作是同一个对象的等同的表现形式，而在 NBN 领域内要被分配不同的 PI，一个数字资源根据长期保存的需要，进行一定转换，在一些域内会被看作同一个数字对象，而另一些域就被视为是不同的数字对象，这些都需要解析器提供可靠、准确的服务，进行有效的识别和判断。解析器和持久标识符之间还要具有较低的耦合性，也就是持久标识符的句法中不能包含解析器的 URL 地址，这样即使解析器的 URL 地址发生变更，也不会影响到 PI，这就相应地增强了标识符的持久性。在每一个 PID 内部，可以有不同的方法和框架在注册机构、认证机构、域解析器、数字仓储管理员和内容提供者、数字长期保存管理员等系统里的不同的组件中共享角色、共担责任，用户可以自由地选择最佳的解决方案，控制和设置一个数字资源的整个生命周期及对其管理的一系列条件、规则、限制，例如访问、重用、引用等。

本模型还提供了一个共享的概念基础设施——互操作知识库，定义该领域内实体之间的基本关系，由于

知识库明确地表示了这些关系，也就创建了一个可供访问的知识层，在此基础上建立起互操作服务。数字对象之间、数字对象和人之间或机构和他们的 PI 之间都存在一些可能的关系，当一个实体进入互操作知识库时，PID 必须提供这些关系。对于任何数字对象，PID 必须声明已有的 PI（例如 DOI、NBN 等）、在该领域内跟其他对象的关系和 PID 已知的人或机构的任何的 PI。在这个框架下的每个可信 PID 独立产生的知识可以用共同语义和格式呈现在网络上，第三方可以通过不同的方式利用这些知识，实现互操作服务。这样，在通过互操作框架提供的一个共同的界面上，这些内容是可见的，并由可信 PID 进行增添，用户能够建立访问所有的域的服务和使用这些内容，哪怕这些内容来自不同的 PID。

如果一个机构给他们的资源采用了两个 PI 系统（DOI 和 NBN），对于同一个关系声明有 DOI PID 和 NBN PID，因此互操作知识库就可以提供某些重叠区，这个重叠区可以作为一个桥梁，实现不同 PID 之间的沟通和交流，可以使一个服务发现新的关系，在数字资源上进行推理。例如，一个提供一个给定作者的全部出版物列表的服务能够利用作者 PI 和出版物 PI 之间的关系网络，从不同的 PID 进行知识聚合和匹配。

另外，其他服务也可以专门设计通过使用不同的技术（例如元数据推理），去抽取其他种类的实体和 PI 之间的关系（例如在 PID 中没有明确指出的关系），并给 PIDs 提供这些信息（例如以概率形式的关系），PID 可以利用这些更新其框架里的显式关系。

4 结 语

通过不同的途径访问使用不同标识符系统的数字资源，可提高数字资源保存的可靠性和灵活性，持久标识符互操作性参考模型，为此提供了相应的保障，奠定了数字资源长期保存的基础。但该持久标识符互操作参考模型还有待于进一步细化和深入探究，很多问题亟待进一步解决，例如互操作参考模型的安全性、可伸缩性等。这些不仅仅是技术的问题，还涉及数字资源的整个生命周期内的管理、职责等方面，例如当内容提供商在 PID 的管理范围和责任改变时，如何能够灵活地提供可持续性的互操作，这也是互操作参考模型设计时需要思考的问题。

参考文献：

[1] Bazzanella B, Palpanas T, Stoermer H. Towards a general entity representation model [C/OL]. [2013 - 11 - 12]. http://disi.unitn.it/~themis/publications/swap08.pdf.

[2] Paskin N. Digital object identifier (DOI) system [J]. Encyclopedia of Library and Information Sciences, 2010(3): 1586 - 1592.

[3] PIs Linking Infrastrucure (PILIN) project [EB/OL]. [2013 - 11 - 12]. http://www.pilin.net.au/.

[4] ORCID [EB/OL]. [2013 - 11 - 12]. http://www.orcid.org.

[5] PersID project [EB/OL]. [2013 - 11 - 12]. http://www.persid.org/.

[6] OKKAM [EB/OL]. [2013 - 11 - 12]. http://www.okkam.org/.

[7] Makx Dekkers and Thomas Baker CORES project, Standards Interoperability Forum Resolution on Metadata Element Identifiers [EB/OL]. [2013 - 11 - 12]. http://www.cores - eu.net/interoperability/cores - resolution/.

[8] Resourcing Identifier Interoperability for Repositories (RIDIR) project [EB/OL]. [2013 - 11 - 12]. http://www.jisc.ac.uk/whatwedo/programmes/reppres/ridir.aspx.

[9] Lattes Platform. [EB/OL]. [2013 - 11 - 12]. http://lattes.cnpq.br/.

Conception of Persistent Identifier Interoperability Reference Model for Digital Resources

Liu Zhen

Xuzhou Institute of Technology, Xuzhou 221008

National Science Library, Chinese Academy of Sciences, Beijing 100190

University of Chinese Academy of Sciences, Beijing 100190

[Abstract] This paper investigates related projects of persistent identifier overseas. It finds some limitations, such as too specific research objects, limited applying scope, which are largely due to lack of a general framework. Thus, it proposes a persistent identifier interoperability reference model which is mainly composed of persistent identifier domains including registration agency, content provider and resolver. The model makes relationships between entities (digital objects, authors, institutions) and their persistent identifiers standardized, which is convenient to manage express and present digital resources, to lay ground for correct interoperability solution and exchange. Lastly, it points out the model's needs of refinement and improvement in security, scalability, etc.

[Keywords] persistent identifier interoperability reference model