基于科研人员本体的知识产出自动获取方法与技术研究*

卢利农 祝忠明 张旺强 李慧佳
（中国科学院国家科学图书馆兰州分馆/中国科学院资源环境科学信息中心 兰州 730000）

摘要：集成第三方系统中已有的知识产出元数据是机构知识库内容建设的重要途径，文章分析了常见的知识资源管理系统中元数据共享方式，并确定了三种内容采集策略。对采集到的知识产出元数据，结合科研人员本体等语义网技术尝试解决作者同名问题。最后，系统使用WOS提供的元数据共享接口，对整体方案进行了检验。测试结果表明本文基于科研人员本体的知识产出自动获取方法能够最大可能地从多种类型的资源管理系统中获取知识产出元数据，基于科研人员本体的作者唯一辨识也较好地解决了作者重名问题。
关键词：自动获取 人名消歧 科研人员本体 语义网
中图分类号：G252 G255.76

Research on Methods and Techniques of Automatic Knowledge Output Acquisition Based on Researcher Ontology

Lu linong Zhu Zhongming Zhang Wangqiang Li Huijia
（Lanzhou Branch of the National Science Library / Scientific Information Center for Resources and Environment, Chinese Academy of Sciences, Lanzhou 730000）

Abstract Extracting existed knowledge output metadata belonging to institute members from other systems is an important way for building institutional repositories. This article analyses the methods of metadata sharing of some common knowledge resource management systems, provides three metadata acquisition strategies. Attempts to solve the problem of author disambiguation combining with technologies of semantic web. Finally, the system collects metadata form WOS for testing. The result shows the automatic metadata acquisition method based on the researcher ontology can do the greatest possible to collect metadata form other systems, and which is a good solution for author disambiguation.
Keywords Automatic acquisition, Author disambiguation, Researcher ontology, Semantic web

1. 引言

长期以来，以保存与管理科研机构自身知识产出为宗旨的机构知识库（Institutional Repository，以下简称 IR），一直未能有效解决内容收集难与作者唯一标识问题。

*本文系中国科学院国家科学图书馆青年人才前沿领域基金项目“基于科研人员本体的学术产出自动获取方法与技术研究”（项目编号：Y200091001）的研究成果之一。

在 IR 的内容建设方面，机构已有的知识产出中很大一部分都已经被数字化，且分散保存在多个其他数字资源系统中，例如大型的数字出版系统、学科知识库、机构其他数字资源系统等。对这些知识产出的重新建设势必会造成人力、物力的浪费，更好的解决方案是加以重用。对外部系统中知识产出的复用，一般有人工采集与机器自动采集两种途径。人工采集相对难度高、花费时间多，当数据量大时，机器自动采集明显优于人工采集。

此外，IR 从多个外部系统汇集知识产出元数据时，面临着作者的唯一标识问题。由于不同系统名称规范、编码方式、数据格式等的不同，普遍存在着作者同名、同一作者多个名称的现象。当前，国内外已有一些项目对名称规范问题展开了研究。JISC 的 Names Project[1]尝试从已有的数据源中搜集名称方面的数据并自动产生相当规模的名称规范数据。ReasercherID[2]在全球范围内通过给每个注册用户分配一个唯一的标识符，以解决用户名称的冲突问题，并支持获取特定作者的引文信息。类似的还有国际标准组织 ISO 的 ISPI[3]，ORCID[4]以及 OpenID[5]等。已有的名称规范解决方案虽然一定程度起到了区分用户主体的作用，但缺少与科研人员其他背景信息的语义关联，在使用时过多地依赖于人工操作，无法实现对科研人员主体的机器自动推理、匹配，难以扩展。语义网技术的兴起，为解决这一问题提供了可能。

本文在研究知识产出自动获取方法与技术的基础上，结合科研人员本体等语义网技术，最终将通过机器采集获取到的知识产出与其作者主体间建立真正的对应关系。

2．知识产出资源存储系统及其采集方法分析

通过机器方式自动获取外部系统中的知识产出元数据时，根据源系统是否提供了程序访问接口可将其划分为两种类型，同时对应两种不同的采集方法。如果源系统提供了程序接口，采集程序可通过接口批量获取到格式化的数据集；如果没有，一般需要通过解析网页 HTML 源文件来获取元数据。

2.1 机器接口自动获取

目前，一些主流的数字资源系统或出版集团已经提供了知识产出元数据或全文的开放共享接口。WOS（Web of Science）的元数据共享 Web Service 接口，支持机构注册用户使用该接口获取 WOS 收录的本机构科研人员的知识产出[6]。BMC（BioMed Central）为知识库提供基于 SWORD 协议的知识产出自动存缴服务，科研机构首先在 BMC 网站注册登记， 之后 BMC 会自动将此科研机构在 BMC 最新出版的期刊文章提交到机构知识库中[7]。arXiv 为了便于机器访问网站元数据，提供了无限制、基于 Atom 的查询接口[8]。源系统的开放接口一般是基于一种或多种协议，常见的用于开放知识产出元数据接口的协议有 OAI-PHM、SWORD、SOAP、RESTful、Atom、SRU 等。不同的接口类型对应的数据返回格式不一，为了实现程序自动采集，需要针对各个源系统的接口开发专门的采集程序。由于不同系统往往使用的元数据描述框架不同，从源系统获取到数据集后，需要通过映射并转换为当前系统的标准元数据并保存。

这些开放接口已经被集成、应用到很多机构知识产出保存与管理系统中。例如，Eprints 支持通过 WOS、PubMed 等系统提供的接口获取知识产出数据[9]。中科院国家科学图书馆兰州分馆开发的机构知识库开源软件 CSpace[10]支持科研机构从 WOS 中自动获取本机构科研知识产出。中国科学院 OA 论文知识库[11]与 BMC 合作，BMC 会自动将其收录的中科院发表的文章提交到 OA 知识库中。

2.2 非机器接口自动获取

如果源系统没有提供机器接口，一般可通过以下两种方式来采集：一种是先通过爬虫程序取得目标源系统中知识产出元数据对应的 HTML 文档，使用抽取算法并结合自然语言处理、机器学习等技术抽取得到规范化编码的知识产出元数据。另一种是根据目标系统网页源文件的特定格式编写专门的获取与解析程序。第一种方式的优点是可以从多个数据源中快速、大批量地获取到知识产出数据，但准确性及查全率一般；第二种方式则正好相反。

对于非机器接口的知识产出元数据自动获取，具有代表性的项目是英国洛翰普顿大学自 2007 年开始，2009 年结束的 AIR（Automated Archiving for an Institutional Repository）[12]，该项目旨在解决机构知识库内容建设难的问题，实现了从机构网站下载包含本机构科研人员产出的网页文件，通过索引、知识抽取、元数据重新组织描述等步骤，将获取到的知识产出数据经科研人员辅助编辑确认后通过 SWORD 提交到机构知识库中。类似的还有 Google Scholar[13]、CiteSeer（又名 ResearchIndex）[14]，它们都支持从互联网上检索并抽取知识产出元数据。国内中科院软件所 Field Specialist Knowledge Collection 领域专家知识导航系统[15]可以从多个数字知识产出出版系统中自动获取科研人员的知识产出数据。

3 基于科研人员本体的学术产出自动获取总体设计

基于科研人员本体的学术产出自动获取主要由知识产出自动获取与语义化匹配存储两部分组成。前者负责从多个外部系统中批量获取并解析得到知识产出数据，后者对这些数据进行语义化描述，并实现作者名称消歧及作者与知识产出、机构、地址等资源间的匹配关联。系统整体框架的设计如下图所示：

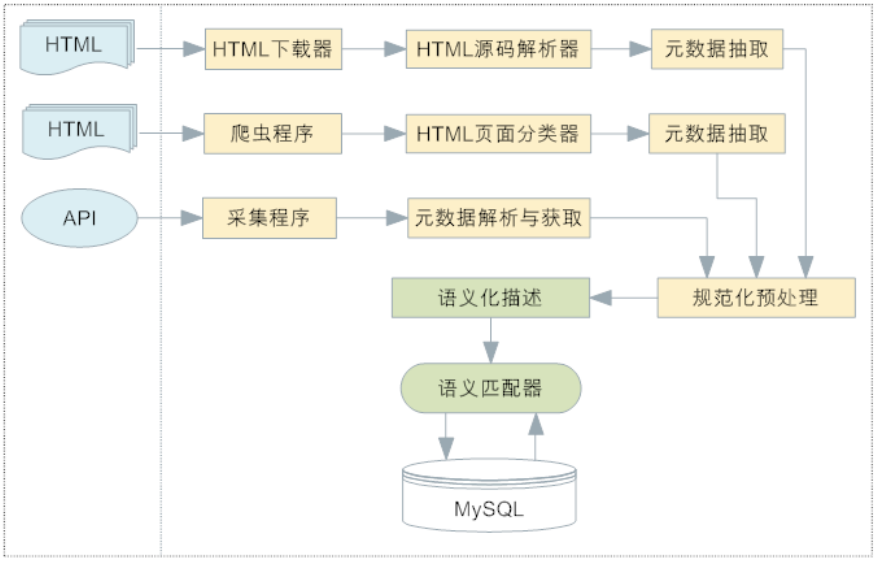

图 1 基于科研人员本体的学术产出自动获取整体框架

自动化采集包括以下三种方式。

第一种是针对目标系统没有提供程序接口，但每条知识产出对应元数据在各 Web 网页中有对固定的规范化 HTML DOM 编码。对于这类资源提供专门针对目标系统的 HTML 下载器与解析器来获取知识产出元数据。

第二种是使用爬虫程序来下载到目标系统的 Web 页面，通过页面分类器对原始文档进行分类，得到知识产出对应的文档，然后结合抽取算法从中得到知识产出元数据。这种方式也是针对没有提供机器接口的第三方系统，但与第一种采

集方式所面向系统不同的是，各知识产出元数据在不同的 Web 页面中没有使用固定的 HTML DOM 编码。

第三种是针对提供了程序 API 接口的第三方系统，根据目标系统共享元数据使用协议、认证方式、元数据编码组织方式等的不同，有针对性的编写专门采集与解析程序来获取知识产出元数据。

通过以上三种途径获取到知识产出元数据后，需要对不同来源的数据进行规范化预处理，包括特殊字符过滤、数据项的合并与拆解、元数据格式统一等。

最后是语义化编码与匹配存储，系统依据科研人员本体模型，将知识产出元数据转换为作者、机构、知识产出、地址等本体实例。在保存时，先判断系统中是否已存在同名作者，若存在，利用当前作者所属机构、地址、关键词、主题词、合著者等信息与已有数据进行语义匹配，确定是是否为同一作者。如果是同一作者，则不创建新的作者实例，将当前知识产出与已有的作者实例进行关联，否则，创建新的作者实例。

4. 科研人员本体的构建

现有的本体构建方法有很多，在对比分析几种常见的知识本体构建方法路线[16]基础上，这里我们采用七步法[17]构建科研人员本体。

首先，确定科研人员本体的目标与边界。构建科研人员本体的目标是通过刻画科研人员以及相关的知识产出、所在单位等资源的属性及资源相互之间的关系，实现对科研人员的唯一标识与名称规范，以及各资源实体间的精确语义关联。因此，科员人员本体的构建应结合这一目标展开，否则有可能会造成本体过于庞大且偏离需求。

复用已有本体可以使构建工作更加科学、快速、准确。围绕科研人员本体应用的需求，通过调研，我们选择复用与科研知识产出相关的 foaf[18]、bibo[19]、fabio[20]、prism[21]、vivo[22]等已有本体。

我们以人物概念作为整个本体模型的中心，通过对科研领域的各种人物实体进行抽象分析，选取与人物概念相关的各种概念，确定科研人员本体模型的核心概念集。如下图所示，科研人员核心概念集主要包括三部分，即科研人员、知识产出、机构。其中，科研人员属于某一个（或多个）科研单位、某一科研产出的作者为某一个（或多个）科研人员。

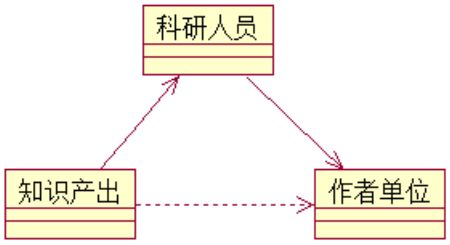

图 2 科研人员本体核心概念集关系

- 科研人员的核心概念包括：姓名、性别、出生日期、Email、别名、研究领域、ResearcherID、所在单位、知识产出等；
- 知识产出核心概念包括：作品的内容类型（如期刊论文、会议论文、专著、专利等）、 题名、发表日期、DOI、作者等；
- 作者单位核心概念包括：所属国家、名称、地址、研究部门、网址、成员等。

在确定核心概念之后，使用本体建模工具 Protégé 建立了科研人员本体模型，核心类关系如下图所示：

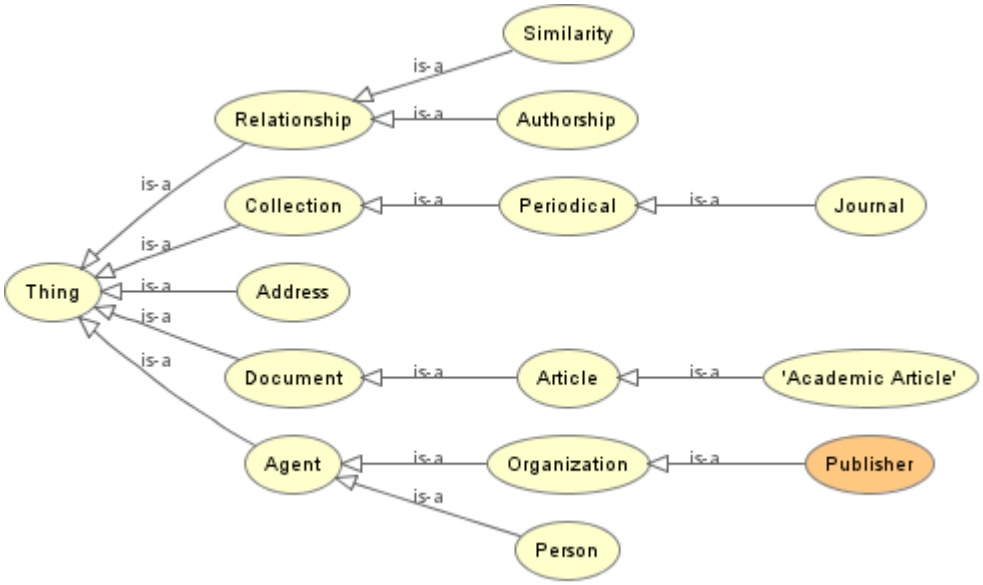

图 3 科研人员本体核心类图

其中，复用的 foaf:Person、foaf:Organization 分别表示作者和机构。作者与机构之间的关联通过 vivo:currentMemberOf、vivo:hasCurrentMember 两个逆反对象属性实现。作者与作者间的合作关系用对称属性 vivo:hasCollaborator 表示，机构之间的上下级关系使用逆反属性 vivo:hasSubOrganization 和 vivo:subOrganizationWithin 表示。

对于知识产出，本文暂时只考虑了期刊论文，并使用 bibo:AcademicArticle 表示。为了保存作者的排名，文章与作者间的关联通过 vivo:Authorship 作为中介，由 vivo:Authorship 的 vivo:authorRank 数据属性记录文章的作者排名。某一篇文章与其作者进行关联时，每个作者对应一个 vivo:Authorship 实例，两者间通过 vivo:authorInAuthorship 与 vivo:linkedAuthor 两个逆反属性关联。vivo:Authorship 实例与文章通过 vivo:linkedInformationResource 与 vivo:informationResourceInAuthorship 两个逆反属性关联。

文章所属期刊对应 bibo:Journal，文章与期刊间通过 vivo:hasPublicationVenue、vivo:publicationVenueFor 两个逆反属性进行关联。

机构的地址信息使用单独的类 vivo:Address 表示，对应的数据属性有详细地址、省、市、邮编等。机构与地址间通过 hasAddress、locationFor 两个逆反属性进行关联。

Similarity 类表示两个同名作者间的相似度，其数据属性 similarityValue 记录具体的相似值。相似度与作者间通过逆反属性 hasSimilarity 与 similarityWith 实现关联。

科研人员知识产出本体的主要概念及属性如下表所示。

表 1 科研人员知识产出本体的主要概念及属性表

（复用）概念	属性		
	名称	（复用）元数据	类型
foaf:Person	姓	foaf: lastName	DataProperty
	名	foaf: firstName	DataProperty

	邮箱	vivo:email	DataProperty
	所属机构	vivo:currentMemberOf	ObjectProperty
	合著者	vivo:hasCollaborator	ObjectProperty
	作者相似度	hasSimilarity	ObjectProperty
	关联作者关系	vivo:authorInAuthorship	ObjectProperty
foaf:Organization	机构名称	foaf:name	DataProperty
	机构别名	anothrName	DataProperty
	地址	hasAddress	ObjectProperty
	机构成员	vivo:hasCurrentMember	ObjectProperty
	子机构	vivo:hasSubOrganization	ObjectProperty
	上级机构	vivo:subOrganizationWithin	ObjectProperty
vivo:Authorship	作者次序	vivo:authorRank	DataProperty
	所属知识产出	vivo:linkedInformationResource	ObjectProperty
	关联作者	vivo:linkedAuthor	ObjectProperty
bibo:AcademicArticle	标题	dcterms:title	DataProperty
	摘要	bibo:abstract	DataProperty
	关键词	prism:keyword	DataProperty
	所属期刊	vivo:hasPublicationVenue	ObjectProperty
	关联作者关系	vivo:informationResourceInAuthorship	ObjectProperty
bibo:Journal	期刊名	dcterms:title	DataProperty
	发表时间	dcterms:issued	DataProperty
	ISSN	bibo:issn	DataProperty
	包含的文章	vivo:publicationVenueFor	ObjectProperty
vivo:Address	地址	vivo:address1	DataProperty
	国家	vivo:addressCountry	DataProperty
	省	vivo:addressState	DataProperty
	市	vivo:addressCity	DataProperty
	对应机构	locationFor	ObjectProperty
Similarity	相似度	similarityValue	DataProperty
	关联相似作者	similarityWith	ObjectProperty

5. 基于科研人员本体的学术产出自动获取实现

下面以 WOS 数据库为例介绍知识产出的自动获取与作者语义匹配的实现。WOS 提供了 Web Service 接口，支持科研机构通过查询的方式获取机构的知识产出元数据，返回数据格式为 XML。系统使用 Java+Jena+SparQL+MySQL 实现，语义化编码与作者匹配使用 Jena 与 SparQL，存储通过 Jena 的 SDB 保存到 MySQL 数据库。

采集程序首先获取并解析得到知识产出元数据，结合科研人员本体，将元数据转化为本体的实例，下图是一条知识产出数据对应语义描述的关键片断示意。

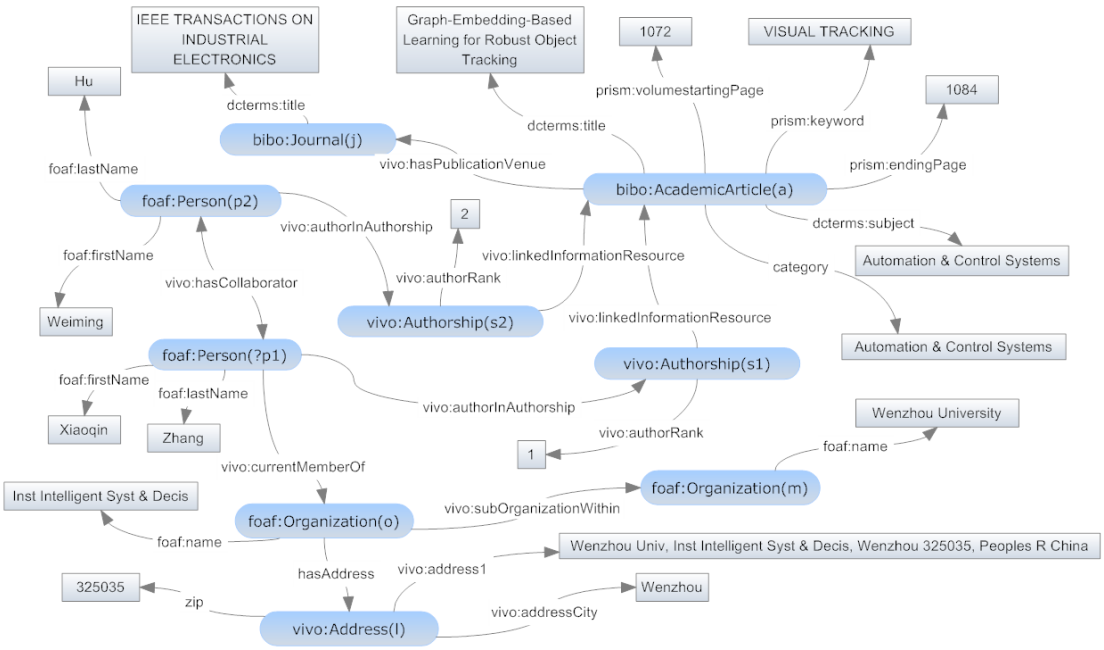

图 4 科研人员本体实例示意

由于从多个外部系统采集知识产出元数据时，各系统使用的元数据框架不一、元数据格式不一、数据质量不一，所以，本系统结合科研人员本体等语义网技术，使用模糊匹配的方法来实现作者的唯一辨识。具体方法是：在向系统中新增一条知识产出数据时，首先查询系统中是否已有同名作者，如果有，则对作者相似度变量 s，对作者所属机构、机构地址、作者发表文章的关键词、主题词，文章所属学科分类以及合作作者相关信息等项进行匹配并赋不同的权值，当各参数项相似度总和 s 达到某一值时，确定当前作者与此同名作者为同一作者实体。其中 s 值的变化是在分析原始数据与程序实际测试的基础上确定的。具体匹配流程如下图所示：

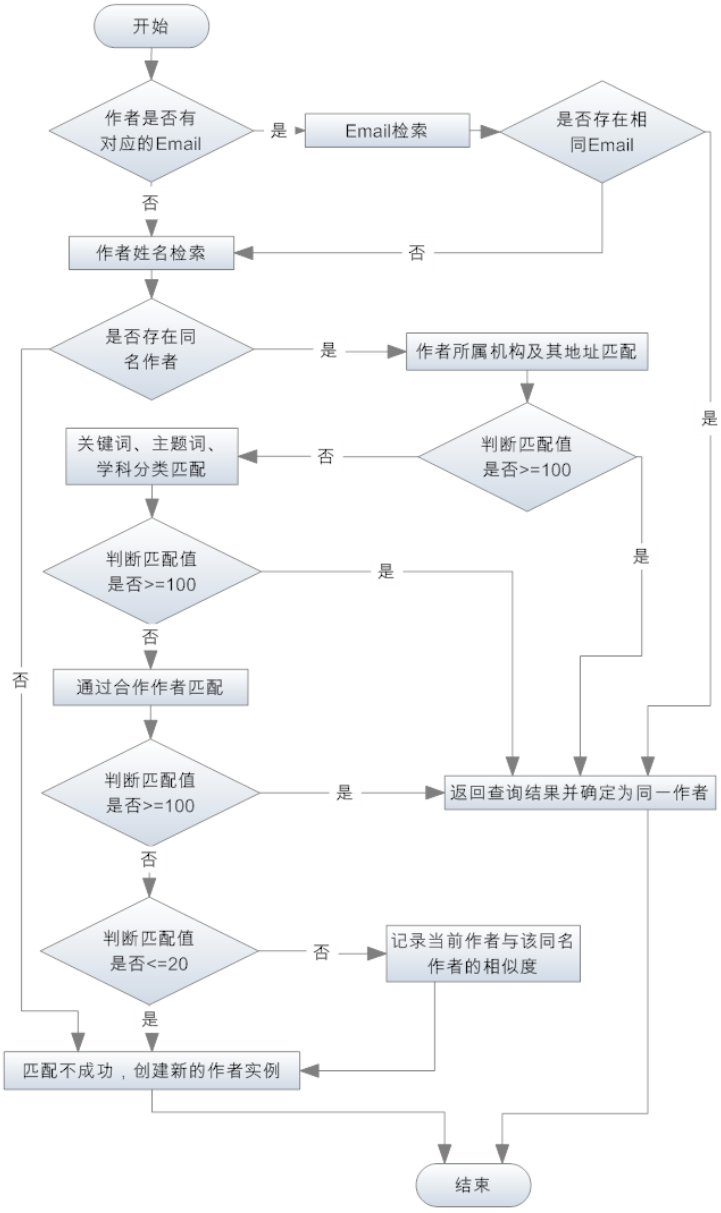

图 5 匹配流程图

整个匹配过程包括以下六步：

第一步，若当前作者有 Email，将此 Email 与系统中作者实例的数据属性 vivo:email 进行匹配，如果匹配成功，转至结果（1）。

第二步，若当前作者没有 Email 或上一步匹配不成功，则检索系统中是否有同名作者。如果存在同名作者，对作者所属机构项进行匹配，匹配成功 s 值增加 20；然后，判断是否存在上级机构，如果存在且匹配成功，s 值增加 10；最后获取作者所属机构对应的地址，对于完整地址（vivo:address1），考虑到每篇文章中相同地理位置描述的不统一，使用字符串模糊匹配算法计算相似度（取值为 0~1），当地址相似度>=0.8 时，s 值增加 60，当地址>=0.6 时，s 值增加 30。此外，当省、市匹配成功时，s 值各增加 10。

第三步，判断 s 值是否大于等于 100，如果是转至结果（1），否则，判断其值是否大于等于 50，如果不是，跳至结果（2），如果是，获取系统已存在同名

作者发表文章的关键词、主题词、文章所属学科分类，与当前作者对应项进行匹配，每次成功匹配，s 值增加 10。

第四步，判断 s 值是否大于等于 100，如果是转至结果（1），否则，如果待添加作者在当前知识产出中的合著作者已匹配成功，判断数据库中同名作者与此合著作者是否有合作关系，如果有，s 值增加 30。判断 s 值是否大于等于 100，如果是则转至结果（1），否则转至结果（2）。

结果（1）：同名作者匹配成功，认为当前作者与系统中的同名作者为同一作者实体，在原有作者实例基础上，关联当前知识产出相关的其他实例对象。

结果（2）：同名作者匹配不成功，为当前作者创建新的作者实例。判断 s 值是否大于 20，如果是创建 Similarity 类实例，并与两个作者进行关联，将 s 值保存到实例的 similarityValue 数据属性，以便管理员手动验证同名作者是否为同一作者。

系统测试使用“单位名称=Chinese Acad Sci”获取到作者单位为中科院的 2376 条知识产出数据，共计 16253 个作者，同名作者 6117 个，成功合并作者 3998 个。下表是成功合并的前 10 个作者信息：

表 2 Top 10 合并作者信息

作者名称	同名数	合并数	作者单位地址
Li Yan	18	10	Chinese Acad Sci, Kunming Inst Bot, State Key Lab Phytochem & Plant Resources West Ch, Kunming 650204, Peoples R China
Wang Robert	7	7	Chinese Acad Sci, Inst Elect, Dept Space Microwave Remote Sensing Syst, Beijing 100190, Peoples R China
Brierley Gary	7	7	Univ Auckland, Sch Environm, Auckland 1, New Zealand
Santosh M.	8	7	Kochi Univ, Fac Sci, Div Interdisciplinary Sci, Kochi 7808520, Japan
Li Can	7	7	Chinese Acad Sci, Dalian Inst Chem Phys, State Key Lab Catalysis, Dalian 116023, Peoples R China
Yin Yulong	12	7	Chinese Acad Sci, Res Ctr Hlth Breeding Livestock & Poultry, Key Lab Agroecol Proc Subtrop Reg, Hunan Engn & Res Ctr Anim & Poultry Sci,Inst Subt, Changsha 410125, Hunan, Peoples R China
Ge Feng	7	6	Chinese Acad Sci, Inst Zool, State Key Lab Integrated Management Pest Insects, Beijing 100101, Peoples R China
Jiang Tianzi	6	6	Chinese Acad Sci, Brainnetome Ctr, Inst Automat, Beijing 100190, Peoples R China
Tian Jie	7	6	Xidian Univ, Sch Life Sci & Technol, Life Sci Res Ctr, Xian 710071, Shaanxi, Peoples R China
Wu Guoyao	9	5	Texas A&M Univ, Dept Anim Sci, College Stn, TX 77843 USA

通过对此 10 个对应同名作者（包括匹配成功与不成功）数据的人工检查统计，准确率 P（Precision）与召回率 R（Recall）分别为 94.5%、91.2%。

6. 总结

在知识产出资源管理系统中，如何快速、有效地从外部系统中集成已有的资

源，以及如何将获取到的知识产出与其作者实现唯一、准确的关联，是包括 IR 在内的很多知识资源管理系统共同面临的难题。本文在分析外部数据源系统元数据共享方式的基础上，根据各类系统的特点，确定了相适宜的采集方案，以减少信息资源内容重复建设带来的人力、物力支出。结合基于科研人员本体的语义网相关技术，对采集到的知识产出数据进行语义化编码转换与匹配存储。实验结果表明，本文的方法较好的实现了对同名作者的唯一辨识，相比传统的人工匹配，大大减轻了系统管理人员的负担，且具有较高的准确率与查全率。对同名作者的匹配，既支持将匹配度较高的作者自动合并，同时对匹配度较低作者间相似值进行保存，为管理人员人工识别提供接口。

与此同时，系统仍有很多不足之处。例如：如何保证使用爬虫程序采集多个数据源的知识产出元数据时都取得比较好的采集效果；目前，系统的作者语义匹配主要解决了作者同名的情况，当元数据来自多个数据源系统时，往往面临同一作者有多种形式名称的问题；此外，当数据量逐渐增大时，会面临效率问题，需要结合传统技术与语义网技术做进一步的优化。这些问题有待进一步的研究并解决。

参考文献：

[1] Names Project.[EB/OL].[2013-08-13].http://names.mimas.ac.uk.
[2] ResearcherID.[EB/OL].[2013-09-20]. http://www.researcherid.com/
[3] ISO TC 46/SC 9 N 429, Outline for ISO Standard ISPI[S/OL].[2013-07-05].http://www.collectionscanada.gc.ca/iso/tc46sc9/docs/sc9n429.pdf.
[4] ORCID 主页[EB/OL]. [2012-4-12].http://www.orcid.org/.
[5] OpenID[EB/OL].[2009-10-8]. http://openid.net.
[6] Web of Science.[EB/OL].[2013-9-25].http://ip-science.thomsonreuters.com/info/terms-ws/.
[7] BMC.[EB/OL].[2013-9-25].http://www.biomedcentral.com/libraries/aad.
[8] arXiv API.[EB/OL].[2013-09-18].http://arxiv.org/help/api/index.
[9] Eprints.[EB/OL].[2013-9-25].http://www.eprints.org.
[10] 中国科学院开源软件 CSpace.[EB/OL].[2013-10-06].http://cspace.org.cn/cspace.
[11]中国科学院 OA 论文知识库.[EB/OL].[2013-10-06].http://casoar.irgrid.ac.cn/.
[12]AIR.[EB/OL].[2013-9-25].http://clg.wlv.ac.uk/projects/AIR/.
[13]Google Scholar.[EB/OL].[2013-8-15].http://scholar.google.com.
[14]ResearchIndex.[EB/OL].[2013-8-15].http://www.researchindex.com.
[15]中科院软件所 Field Specialist Knowledge Collection 系统.[EB/OL].[2011-9-1]. http://124.16.136.213/mediawiki/default.php.
[16]李景,孟连生.构建知识本体方法体系的比较研究.[J].现代图书情报技术,2004(7):17-22.(Li Jing,Meng Liansheng. Comparison of Seven Approaches in Constructing Ontology.[J]. New Technology of Library and Information Service, 2004(7):17-22).
[17] Natalya F.Noy, Deborah L.McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology[EB/OL].http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html.
[18]FOAF.[EB/OL].[2013-9-25].The Friend of a Friend (FOAF) project. http://www.foaf-project.org/.
[19]BIBO.[EB/OL].[2013-9-25].Bibliographic Ontology Website. http://bibliontology.com/.
[20]FABIO.[EB/OL].[2013-9-25].http://www.essepuntato.it/lode/http://purl.org/spar/Fabio.
[21]PRISM.[EB/OL].[2013-9-20].http://prismstandard.org/namespaces/1.2/basic/.
[22]VIVO.[EB/OL].[2013-09-16].http://vivoweb.org/.

作者简介： 卢利农（1985-），男，中科院国家科学图书馆兰州分馆馆员；祝忠明（1968-），男，中科院国家科学图书馆兰州分馆研究员；张旺强（1985-），男，中科院国家科学图书馆兰州分馆馆员；李慧佳（1984-），女，中科院国家科学图书馆兰州分馆馆员。