

“第四届中日韩科技信息机构联合研讨会”概述* ——面向开放获取、数字标识及 实用型汉日双向机器翻译系统

□ 李颖 朱礼军 张钧胜 张婧 / 中国科学技术信息研究所 北京 100038
张秀梅 杨公亮 / 万方医学网 北京 100038

摘要: 文章首先简要介绍“第四届中日韩科技信息机构联合研讨会”的整体情况,其次,围绕会议设立的新议题,即开放获取、数字标识,以及实用型汉日双向机器翻译系统进行重点介绍,最后,简述会议专题讨论的议题以及第五届会议的计划。

关键词: 开放获取, 数字标识, 汉日与日汉机器翻译, ISTIC, KISTI, JST
DOI: 10.3772/j.issn.1673—2286.2013.11.007

1 会议概要

2013年9月9-10日,继第一、二、三届会议之后^[1-3],日本科学技术振兴机构(JST)、中国科学技术信息研究所(ISTIC)、韩国科技信息研究院(KISTI)在日本东京成功地召开了“第四届中日韩科技信息机构联合研讨会”。这是新一轮的交流盛会,本轮会议在交流内容与形式上都进行了新的尝试,是收获更多的会议。

除三机构之外,日本文部科学省(MEXT)官员、日本情报通讯研究机构(NICT)、日本国立信息学研究所(NII)、日本国立国会图书馆(NDL),以及北京万方数据有限公司的代表也参与了本届联席会。会议在JST现任理事大竹晓(见照片2左一,原文部科学省MEXT大臣官房审议官研究开发负责人)热忱的致辞后,参会代表进行了畅所欲言的发言与讨论。

本次会议的议题包括:(1) Evaluation and Analysis Based on STI; (2) Open Access; (3) Digital Identifiers; (4) e-journal; (5) New STI (科技信息) Services & Technologies; (6) Other。会议新设了半天专题讨论会:(1) Policy Oriented Session; (2) Technical Oriented Session and Service Oriented Session。详细信息可参见附录会议日程。

前三届会议,发言人以母语发言,通过专业口译者进行交流。为了直接传递各自的思想与主张,节省发表时间,本次交流以英语发言为主,并辅以口译。尽管大部分的中日韩三国代表还算不上英语的“达人”,但在共同兴趣与合作愿望的驱动下,交流的传递性达到了意想不到的良好效果,深化了代表们主动参与的意识,激情得以释怀。

有关会议内容,除了延续性的议题之外,新增了开放获取、数字标

识、汉日双向机器翻译等话题。同时,还新增了开放获取政策、信息技术与服务两个专题讨论。围绕这些新议题以及专题讨论,下文具体阐述。

2 新增会议议题

2.1 开放获取

(1) Akira Nishi's “Open Access Policy” (JST)

JST代表Akira Nishi发表的开放获取政策包括三部分内容:1)科技信息所处环境——全球开放获取的趋势;2)论文的开放获取——国外状况/日本国内状况;3)扩展开放获取——科研数据的开放获取,国外状况/日本国内状况。

1) 科技信息所处环境——全球开放获取的趋势

见图1全球开放获取状况。其中,欧美国家基金机构促进论文的开放获取已有十余年;科研数据的

*项目: 获得中国科学技术信息研究所科研项目预研资金“基于DITA标准的语义知识系统框架设计及应用研究”、“十二五”科技支撑计划项目科技知识组织体系共享服务平台建设(编号: 2011BAH10B03)的支持。

开放获取也在加速。

2) 论文的开放获取——国外状况/日本国内状况

就论文开放获取现状而言，欧美领导基金机构制定了相关政策，并采取了具体行动，见表1。

日本开放获取路线：2012年7月，科技审议会建议对学术成果实行开放获取^[4]；2012年12月，JSPS与JST联合举办GRC (Global Research Council, 全球研究委员会)^[5]亚太地区分会；2013年4月，作为日本第一个基金机构，JST制订了开放获取政策，并指定在CREST、Sakigake (战略创新研究推进事业) 项目中适用^[6]；2013年5月，在RCG的第二次年会上，一致同意对公共研究资金资助的研究论文实施开放获取行动方案。

作为基金机构，JST支持开放获取，其行动如下：

论文开放获取平台J-STAGE：J-STAGE是JST研发的论文传播系统，为日本最大的电子期刊平台之一。从1999年起，JST对一些学会论文的全文实施了开放获取^[7-8]。

3) 扩展开放获取——科研数据的开放获取



图1 全球开放获取状况

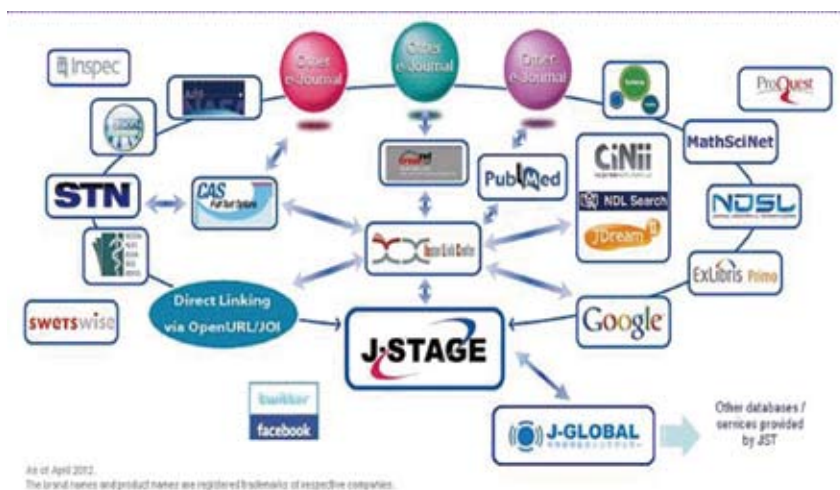


图2 J-STAGE定位

表1 论文开放获取出版的具体行动

机构名称	国名	开放获取（论文）的出版	
		启动时间	出版平台
NIH (National Institutes of Health)	美国	基金机构主导（义务） 2005起生效	PubMed Central
NSF (National Science Foundation)	美国	基金机构主导 2011起生效	出版社与学术团体， 出版费支持（作者付费模式）
Wellcome Trust	英国	基金机构主导 2006起生效	UK PubMed Central （结合作者付费模式）
Research Council UK	英国	基金机构主导 2005起生效	出版社与学术团体， 出版费支持（作者付费模式）
DFG德国科学基金会	德国	研究者主导，2003起试行	支持开放获取期刊

表2 全球科研数据的开放获取行动

机构名称	国名	开放获取（论文）的出版		科研数据开放获取
		启动时间	出版平台	
NIH	美国	基金机构主导（义务） 2005起生效	PubMed Central 出版社与学术团体，	2013年颁布了共享政策 （在某些CBI中发布）
NSF	美国	基金机构主导 2011起生效	出版费支持 （作者付费模式）	2011年颁布了共享政策 （研究者发布）
Wellcome Trust	英国	基金机构主导 2006起生效	UK PubMed Central（结合作者付费模式） 出版社与学术团体，	2007年颁布了共享政策
Research Council UK	英国	基金机构主导 2005起生效	出版费支持（作者付费模式）	颁布了共享政策 （发布年不详）
DFG	德国	研究者主导 2003起试行	支持开放获取期刊	2003年开始了原型试验

除了论文之外，科研数据开放获取也在增长，如表2右栏所示。

目前，全球都在努力推进科研数据开放获取。在发达国家，科研数据开放获取的手段不断增多，比如：美国、欧盟、英国及其澳大利亚等是主要的促进机构：

■ **RDA (Research Data Alliance)**：成立于2012年8月的国际机构，由NSF、欧盟iCORDI（科研数据基础设施国际合作）及澳大利亚ANDS资助。

■ **iCORDI**：创立于2012年9月的国际论坛，由欧盟委员会第七框架计划投资。

■ **WDS (World Data System)**：ICSU于2008年10月创立的旨在推进全球科学数据（库）开放的项目。

■ **CODATA (Committee on Data for Science and Technology)**：ICSU科技数据委员会1966年创立。

■ **GRC (Global Research Council)**：2012年5月成立，响应美国NSF的要求，由遍布世界各国的

代表性研究基金机构组成。

有关科研数据最新动态，2013年6月12日，在G8会议上，日本科技大臣与科学院负责人联合声明提出了科研数据的开放获取^[9]。

值得一提还有全球环境信息集成项目DIAS (Global environment information Integration program)。它将先进的信息科学技术与全球环境有关的各种科学技术相结合，有效整合地球观测数据、数值模型和社会经济数据，建立了数据基础设施，创造知识，从而引导全球环境问题的解决，并产出公共利益。

日本科研数据开放与共享途径：文部科学省、JST及NII的eRad-Read&Researchmap数据库集中收集全日本的研究者信息，但目前还没有跨部门应用，参见表3。

JST创新情况：生命科学领域科研数据的开放获取（国家生物科学数据中心，NBDC）：

为实现在生物科学领域研究者之间共享研究成果，促进社团数据库的整合，切实有效地推进研究，

NBDC提供该领域的数据库目录信息、横向检索、长期保存等服务。其运行的门户网站为“integbio.jp”。同时，进行跨部门合作（MEXT、MHLW、MAFF、METI），提供大约4个部的1300种数据库信息。

日本其他重要领域的数据开放：材料数据库（MatNavi）——建筑材料数据库。数据库组成为聚合物、无机材料、金属材料、扩散和超导材料等。大量的数据从出版的学术文献中抽取。

总之，JST在改变信息项目的方向：作为研究机构，尝试发布研发结果，从封闭模式向开放模式转变；整合日本的科技信息，构建关联网络；紧随开放获取的趋势，建立将科技信息与不同领域信息结合的体系，实现从信息的提供者向信息流通的运行者转变（changes from an information provider to information circulatory operator）。

(2) Tae-Sul Seo: 开放获取活动方面的新进展（KISTI）：

KISTI代表Tae-SulSeo发言

表3 日本科研数据开放获取情况

范畴	主要机构	现状	理想状况
资助信息 (研究项目、研究成果的报告等)	基金机构	各个基金机构独立公布数据	所有基金信息的跨部门应用 (研究者, 机构, 领域, 关键词, 时间...)
研究者 (被资助研究者)	研究者 (JST•NII)	集中于ReaD&Researchmap中 • 220,000公益机构研究者 • 大约涵盖70%的研究者	*有些在J-GLOBAL上发布 通过ReaD&Researchmap与所有的科技信息关联

的主要内容: 1) OA政策; 2) OA期刊; 3) 机构知识库 (Institutional Repository, IR)。

1) OA政策 政策研究:

■ 从2010到2011年的2年内, 研究开放获取, 以利于研讨政府资助成果的处置。调研韩国在公共获取研究成果方面的政策及立法机制; 调研国外公共获取政策; 分析包含研究者、大学、学术团体、数据库公司在内的所有干系人之间的利益关系; 准备起草韩国公共获取法案。

■ 然而, 此项研究没能持续, 理由是韩国开放获取意识的不成熟, 以及数据库公司反对立场的不妥协。

有关会议及论坛:

■ 自2010年起, KISTI每年都举办开放获取会议 (Open Access Korea, OAKconferences): 2010, 主题是科学与学术信息的开放获取。特邀发言为Max Plank Library、NII、Seoul Univ、KISTI; 2011年, 主题是开放获取的金色与绿色道路。特邀发言为中科院、NII和KISTI; 2012/13年, 参与韩国图书馆协会会议。

■ KISTI组织开放获取专家论坛: 有来自政府、研究机构、大学及数据库公司20~30个成员; 举办

开放获取方面的研讨会与讲座。

开放获取项目:

■ KISTI参与了SCOAP3项目: SCOAP3是通过重新定向订购费用来促进高能物理领域开放获取出版的联盟; KISTI代表韩国科学图书馆参与SCOAP3项目, 提升韩国高能物理期刊的获取性。

■ KISTI协助韩国的期刊出版社发行开放获取期刊: 为了适用于韩国的开放获取期刊, 开发生命周期型期刊出版模式与平台; 提供如何出版开放获取期刊的指引。

2) OA期刊

开放获取期刊的技术支持:

■ 构建XML格式全文: 包含25种同行评议的开放获取期刊中的2960篇文章; 基于PMC DTD 3.0描述全文; 开发管理工具、PDF文档转化为XML文档的半自动化工具。

■ 运行开放获取期刊的门户网站: OAK中心、韩国开放获取期刊的门户。图形界面与各种服务, 利用XML全文, 可进行章节浏览与图表的独立窗口浏览等。网址为http://central.oak.go.kr。

3) 机构知识库IR

机构知识库的技术支持:

■ IR的扩展: 已扩展到25所大学、研究机构; 基于Dspace; 开发开源版软件。

■ 运行IR门户: OAK门户, 韩国IR门户; 利用OAI-PMH协议收割IR内容; 网址为http://www.oak.go.kr。

(3) Sang-Ho Lee's KISTI Digital human Project and Medical Repository (KISTI)

Sang-Ho Lee发言的题目是KISTI数字人类项目与医学知识库。

韩国人类信息的需求: 韩国人类信息缺乏, 而医疗、科学、生物医学、安全、法医、体育等多方面对人类信息的需求在增多。如果利用其他国家的人类信息, 这些信息与韩国有别。

在此背景下, KISTI实施了数字韩国项目。项目期间为2003年到现在。主要任务是数据的加工与发布。加工由Dept. of Anatomy, Catholic University Medical College (天主教大学医学院解剖学部) 担当, 发布与维护是KISTI。

数据内容包含: 男女各50具尸体的整套CT图像、合计100组男女3D骨骼图像、一组韩国男女平均骨骼3D图像、6组韩国男女3D皮肤图像、韩国人的手脚以及牙齿的微CT图像、骨头机理特性评测数据。网站是http://dk.kisti.re.kr。

医学数据知识库构建内容: 医学图像; 骨骼机理特性; 形状测定

值; 人类动作分析; 医疗教育视频; 人类仿真模型; 其他。

知识库元数据为DataCite元数据Ver. 3.0, 包括必要属性(5个字段), 即Identifier、Creator、Title、Publisher、Publication Year; 推荐和可选属性(13个字段), 即Subject、Contributor、Date、Language、Resource Type、Alternate Identifier、Related Identifier、Size、Format、Version、Rights、Description、GeoLocation; 附加属性, 即Sample Data、Measured Value Data、Property Data、Motion Data、Human Model Data、Disease Data。

KISTI构建的医学数据知识库包含所有数字韩国与可视化韩国数据的元数据目录, 将上述项目的元数据上载到医学数据知识库, 采用开放数据向公众开发, 使用P-CUBE^[10]作为医药数据编目和数据管理系统。

下一步计划:

■ 收集数据, 支持学会团体: 调查国家资金支持的医学研究项目; 收集可向公众开放的医学数据; 建立元数据并通过P-CUBE上载数据; 对那些可自主增加数据的知识库团队予以支持。

■ 与韩国NIH合作: 韩国NIH与KISTI的开放数据政策基本一致, 也是医学研究项目的基金机构。

2.2 数字标识符

(1) Takafumi KATO's Renewal of JaLC Project and Its Updates (JST)

JST代表Takafumi KATO介绍

的“JaLC项目与新进展”包含的内容有: 日本学术信息传播、JaLC项目、JaCL组织、业务模式及技术路线、进展。

JST的核心工作之一是促进科技信息的传播, 这与其支持创新的使命相一致。通过提供综合信息基础设施, 聚焦于数据库的易访问性, JST力图为日本以及全世界的创新作出贡献^[11,12]。

在文献数据库方面的考虑: JST构建与维护各种科学与学术数据库, 比如二次文献数据库、研究者与生物技术数据库; 这些数据库中, 最重要的是J-GLOBAL与J-STAGE; J-GLOBAL是将这些数据库互链的系统, 由于其高利用性, 深受研究者认可; 需要通过促进出版社的电子出版, 改善期刊的流通状况, 尤其是日语期刊。需要内容所有者的参与; 元数据的整合对网络获取的改善非常必要; 作为元数据, 必须考虑多字节语种、出版类型的兼容性。

在此背景下, JST启动了“Japan Link Center, 即JaLC”^[1], 来解决这些问题。

JST与其他大量内容的持有者参与了JaLC。JaLC是公共信息服务

整合的国家项目, 它是DOI的RA, 通过DOI来促进内容之间的链接, 拓宽传播, 并利用内容。见图3。

JaLC的功能: 对学术期刊、博士论文、纪要、报告等内容注册DOI; 通过DOI及PubMed ID, 在数据库之间提供互链(比如引用与被引); 提供DOI在日本的注册服务、各种类型数字内容全国范围内的解析服务、兼容多字节的元数据查询服务。通过JaCL, 促进DOI在日本的应用, 促进全球用户获取日本出版的学术信息, 并与IDF及其他DOI RA合作。

进展情况与挑战: 重复内容的处理; 机构知识库vs出版社、机构知识库vs机构知识库(具有多个作者的论文); 作者自有版本的论文处理; 科学数据或图书DOI注册; 接受作者ID的注册, 比如ORCID; 科学数据注册方面的问题: 范围、粒度; 元数据政策; 可扩展性等等。

(2) Takashi HORIE's Comprehensive Identification System - Improved Identification for Author Affiliation and Bibliographic Records (JST)

JST的Takashi HORIE发言题

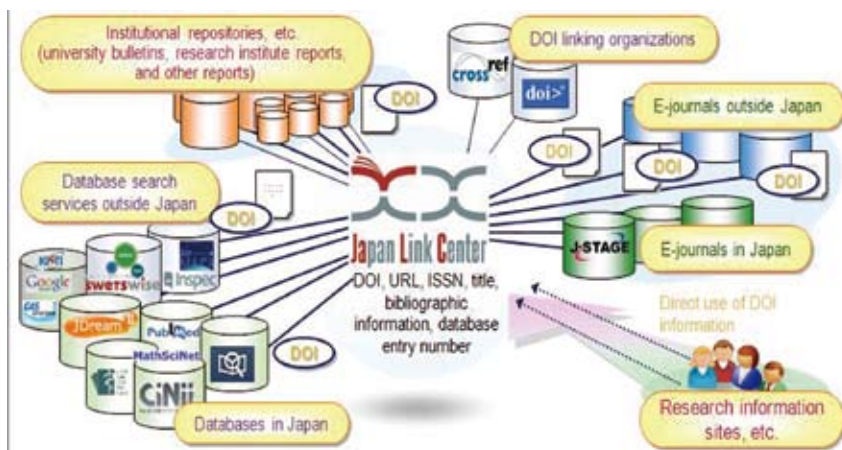


图3 JaLC定位

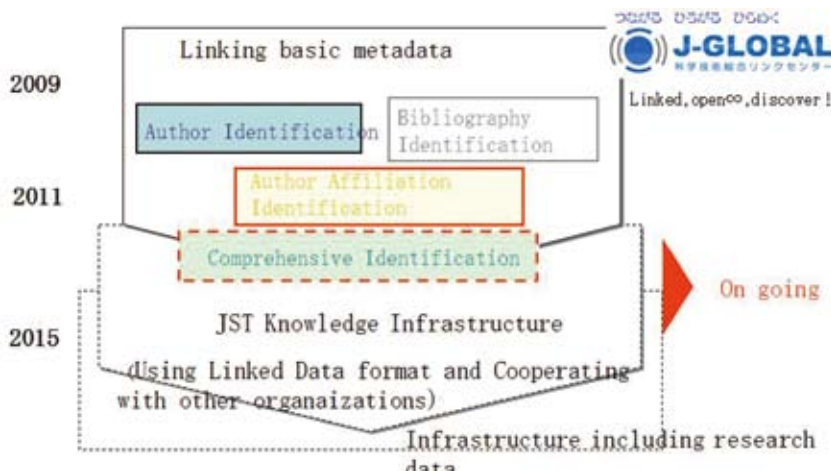


图4 综合标识体系的发展路线

paper search with the author ID in J-GLOBAL		paper search with keywords		
Author ID 200901100457227208	985	-		Different people from different institution with the same name came up.
北沢宏一	545	986		"KITAZAWA ○○" "△△△ K" also came up
KITAZAWA K	404	13,644		
北澤宏一	30	35		"KITAZAWA ○○" "△△△ Koichi" also came up
KITAZAWA Koichi	6	799		

papers by Kitazawa Koichi, executive director of JST are shown.
Results include his papers at Tokyo University.

图5 作者标识必要性

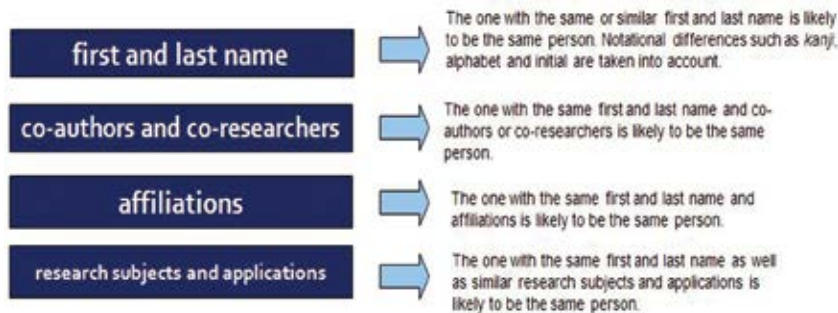


图6 作者标识所需数据的整合方法

目的是综合标识系统——促进作者机构与二次文献记录的标识。

参见图4, 在促进科技信息传播

方面, JST提供的J-GLOBAL系统中, 对科技基础元数据进行了组织与关联, 以利于产业创新。该系统

的目标是“JST知识基础设施”。它更新频繁, 是更为先进和对公众开放的系统, 用于国家与企业的战略决策。为此需要进行如下工作:

首先构建作者标识系统。各类科技作者, 包含专利的申请人, 即使他们具有相同的姓名或有不同的标记方式, 都要与其他作者加以区别。见图5。

第二个项目是作者机构标识系统。通过机构历史数据等, 即使持有相同名称的机构、同一机构具有不同的标记(名称)、或者名称有变更, 也可以唯一标识(同定)。目前, JST正在开发综合标识系统, 也就是作者与机构标识系统的组合, 它将提供更加先进的标识符。本发言主要介绍综合标识系统。

目前, 科技数据应用于国家和企业战略决策时, 主要任务之一是分析技术趋势。该方法着重于研究人员和从属机构。它将从某一领域、从属某一研究机构的某一人的数据进行整合与可视化, 并基于研究论文与专利文献, 而其中使用了不同的标记。例如, 日本研究者的名字可以采用相同读音的汉字、不同字符或仅包含首字母等方式。如果用普通的关键词进行检索, 具有完全或部分相同姓名的不同研究者都会命中, 导致结果中含有不需要的数据。作者标识系统要实现只包含必要的信息, 去除噪音。

图6展示了作者标识系统中所用数据的整合方法: 具有相同或相似姓与名的人可能为同一个人。标记的不同, 比如汉字、字母及首字母等, 都要考虑; 具有相同的姓与名以及合作者或共同研究者可能是同一人; 具有相同的姓与名以及机构可能为同一人; 具有相同的姓与名以及相近的研究题目与专利申请可

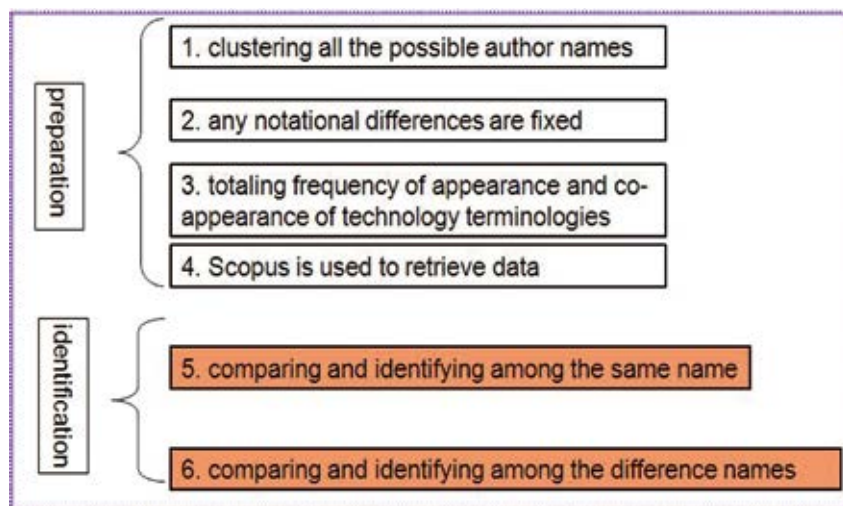


图7 作者标识流程

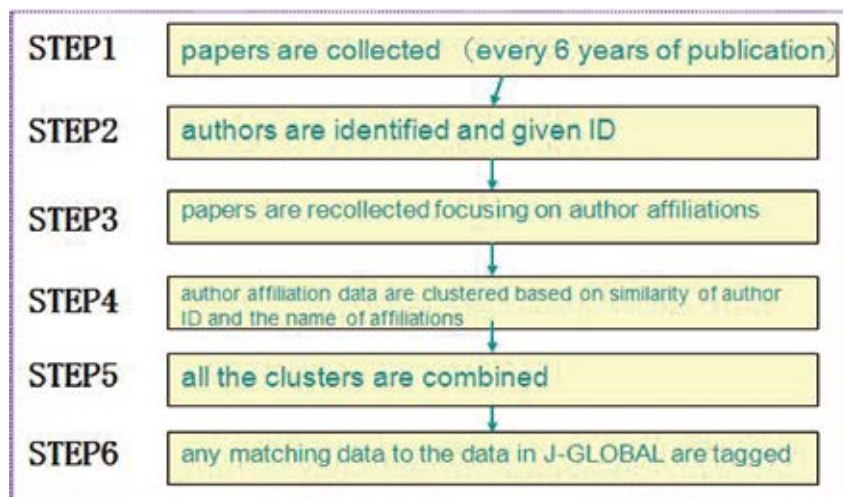


图8 作者机构标识的原型机制

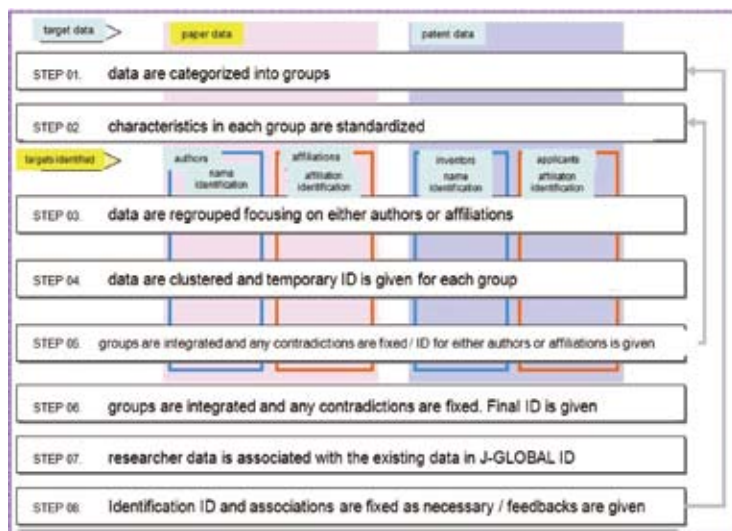


图9 综合标识系统的整合机制

能是同一人。

JST的作者标识系统考虑了不同的姓名标记形式，合作者也包括在处理过程中。除了姓名，他们的机构、研究题目和专利申请等也考虑在其中，这些有助于更准确地标识。作者标识处理过程如图7所示。

作者标识系统的精度为98.6%，召回率90.8%。以下阐述作者标识的过程，它大致分为两步：准备和标识。准备，构建一个大的群簇cluster，包含了所有可能的作者，它们都可能是需要检索的人。确定其姓名或机构的所有不同标记，比如，不同的字母或缩写等。接下来确定科技术语的出现或共现频率。在收集E-mail或参考文献之际，利用Scopus；标识，该过程分为两步，第一步是查找日语和英语的同一人。作者和机构名称，要考虑术语及合作者。同样的步骤再用于专利申请。最后，在两个数据库间进行标识。

作者机构标识的原型机制如图8所示。步骤一，收集每6年的论文；步骤二，给组中的作者作标识，赋予ID；步骤三，聚焦于作者机构再收集论文；步骤四，聚类作者机构数据，基于作者ID和机构名称相似度进行标识；步骤五，对过去6年所有群簇进行组合；最后，对所有与J-GLOBAL数据相匹配的数据给出标签。作者机构标识系统的精度为95.3%、召回率95.0%。

综合标识系统的开发。JST的开发团队认为，作者标识与机构标识整合的综合标识系统会更有效和有用。作者机构标识系统应类似地采用作者名称、术语和分类范畴。需要利用任何可获得的数据和手工修正。作者机构标识系统的结果应当用于作者标识系统以获得较好

的精度。整合机制如图9所示。

综合标识系统机制：目标是论文与专利数据库，整合作者和作者机构系统的共通部分，不同部分分别处理。步骤1和2是整合处理。论文与专利按照年代被分成不同的组，并对这些组中的数据进行规范。步骤3~5为各个不同处理。规范化的数据按照作者或机构被再次分组。然后被聚类、整合，确定所有的冲突。整合论文群簇和专利群簇，确定冲突。最后，赋予ID，研究者数据将与J-GLOBAL ID的既有数据关联。需要时，适当进行手工修正。

本年度末JST将实现的目标为：论文与专利数据库中作者标识，加权平均精度大于98%、加权平均召回率大于90%；论文与专利数据库中机构标识，加权平均精度大于95%、加权平均召回率大于95%。

(3) Seok-Hyoung Lee's Author Identification and Author Name Authority Control Project of KISTI (KISTI)

Seok-Hyoung Lee发言题目是KISTI作者标识与作者名称规范控制项目。主要内容包含：1) 引言；2) 作者标识的战略；3) 规范控制系统。

1) 引言

作者标识：作者标识是现实世界中，出现在学术信息中的作者与其他人的对应，从强化信息检索精度的角度，作者标识是一个重要的问题。主要工作是作者消歧：区别相同姓名的不同作者；解决同一人的名称不一致问题。

KISTI的作者标识，目的：提高基于作者名称的学术信息检索精度；范围：韩国出版学术期刊中出现的作者。其中，韩国人占95%，中国

人、日本人、东南亚、中亚、英美等占5%。还要考虑在国际期刊中出现的韩国人。

考虑事项：

- ◆ 大规模联机数据库
 - ◇ 韩国出版的超过1,000,000篇的论文
 - ◇ 国际期刊中大约100,000,000篇论文（韩国作者为100,000篇）
 - ◇ 大约需要处理3,500,000作者，标识每一作者
 - ◆ 成本
 - ◇ 回溯数据的快速与准确处理
 - ◇ 自动标识处理的限度
 - ◇ 新增论文作者信息的更新
 - ◆ 研讨
 - ◇ 作者标识的战略
 - ◇ 规范数据的管理系统——规范控制系统
 - ◇ 应用
- 2) 作者标识系统的战略

目标：快速、准确地解决在大规模联机数据库中出现的具有完全相同名称的不同人及其同一人的不一致名称的问题。

作者标识的重要属性：

- ◇ 作者名称字符串author's name string
- ◇ 作者机构affiliations of author
- ◇ e-mail地址e-mail address
- ◇ 合著者名称与机构co-author's name and affiliations
- ◇ 作者关键词keywords by authors
- ◇ 期刊名称journal titles
- ◇ 主题分类subject classifications
- ◇ 出版年published year
- ◇ 合著者顺序co-author's order

作者标识系统的流程如图10所示。

3) 规范控制系统

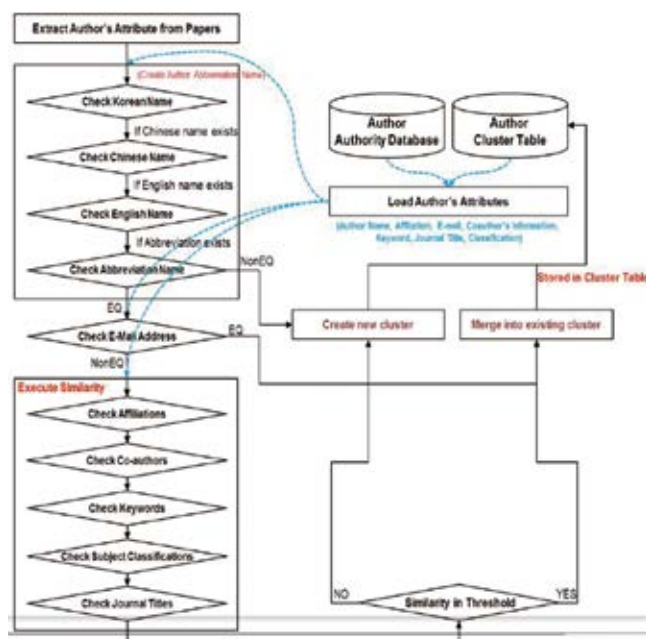


图10 作者标识的流程

参见图11规范控制过程。

KISTI规范控制系统的功能如下(图12):

◇ 验证作者簇群

→2-phase验证(构建者, 管理员)

→论文验证、Web检索(利用机构网页、Web门户)

◇ 从作者簇群到规范数据的转换

→FRAD(规范数据的功能需求)概念模式的应用

→采用MADS(Metadata Authority Data Scheme) 2.0描述

◇ 检索与规范数据的更新

(4) ZHANG Xiumei's From Information Service to Intelligence Service—Practice of Wanfang Med Online (Wanfang Data)

作为科技信息行业唯一的企业代表, 万方医学网张秀梅经理的发言题目是“从信息服务到情报服务——万方医学网的实践”。包含的内容主要有: 万方医学网的发展路线(图13); 万方医学网体系结构(图14); 万方医学网的现在——基于DOI的应用(图15与16)。

2.3 实用型汉日双向机器翻译

(1) Kenichi SUMIMOTO's Promoting Science and Technology Exchange using Machine Translation (JST) [13,14]

见图17, 英语科技信息一直处于统治地位。然而, 其他语种论文的增加也不容忽视。中文论文快速成长、日语论文的份额在非英语的语种中处于前列。

机器翻译项目的背景:

■ 牛顿说过: 如果我看得更

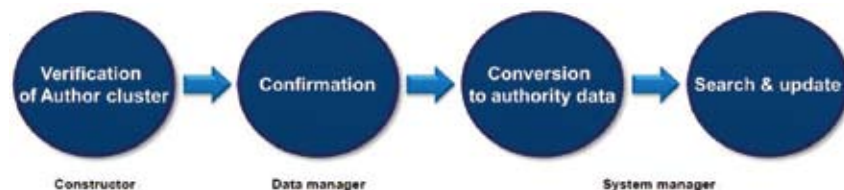
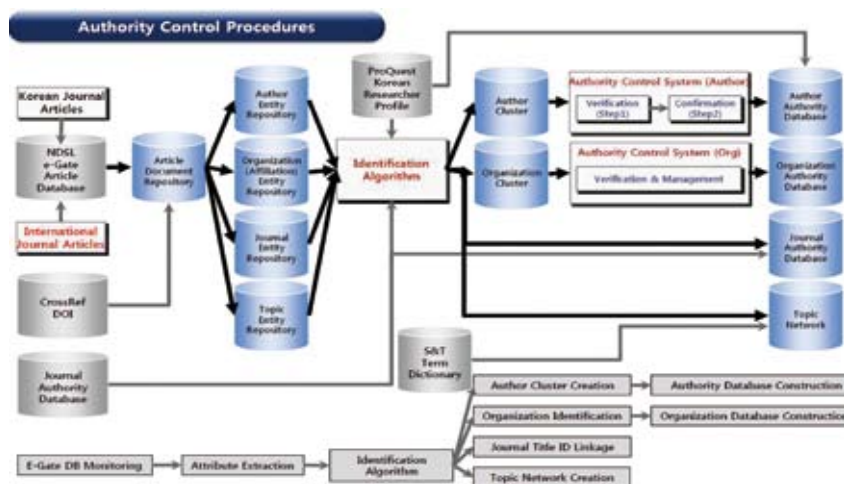
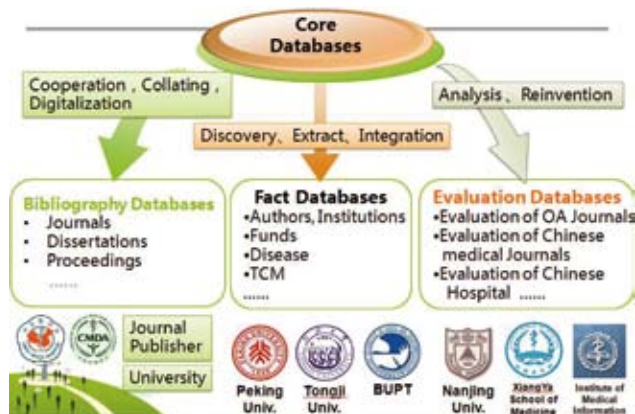


图12 规范控制系统的功能



DOI Application on WFM Online now

Over 75% Chinese Journal

Papers on WFM online

already using DOI

DOI Registered paper

500,000+

DOI Linked paper

4,500,000+



图15 万方医学网的现在

Future plan of DOI Application on WFM Online



图16 万方医学网的未来

远,是因为站在巨人的肩膀上

科技发展依赖于研究成果的共享

科研工作需要克服语言障碍,来直接获取有价值的外文文献

机器翻译的重要性在增高,中日实用化机器翻译重要性凸显

机器翻译项目的成果将极大地促进科技的进步

项目计划见图18,本年度的目标:中日数据词典的构建。从关键词列表中收集双语术语,采用受限隐马尔可夫模型和中间语言。

如图19词典构建原理所示,英语为中间语言,需要建立日-英资源。首先利用CiNii(NII的学术信息导航系统),获得有关数据:

利用论文、图书及杂志的数据库检索服务

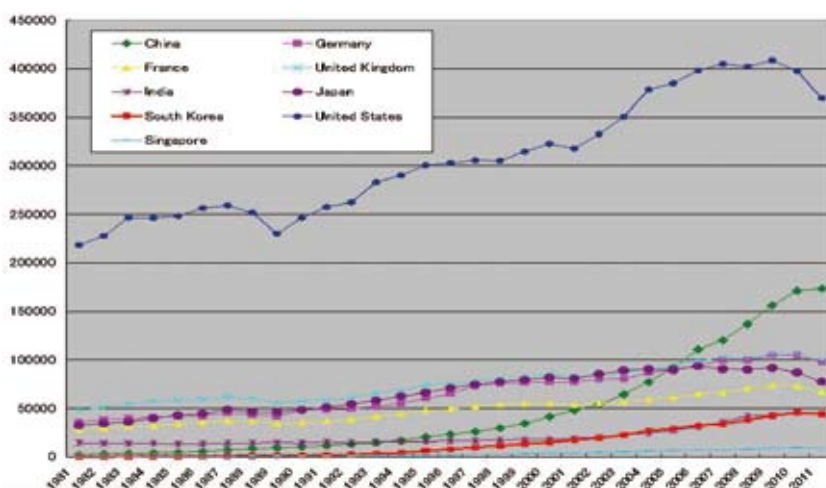


图17 世界研究论文[*基于Web of Science的数据]

从关键词一览中抽取有关字段

应用JST科学技术语词典

手工修正

应用LCAS(中国科学院文献

情报中心)数据:

LCAS向JST免费提供中文文献数据库

目前为921K文献(since 2007)

	2013	2014	2015	2016	2017
Enhancement of Chinese-Japanese science and technology terminology dictionary	Target: 1million				
Acquisition of parallel corpora		Target: 5million			
Improvement of Parsing Accuracy			Target: over 90%		
Development of Translation engine					
Practical Use		(Keywords)			(Translation System)

图18 项目计划

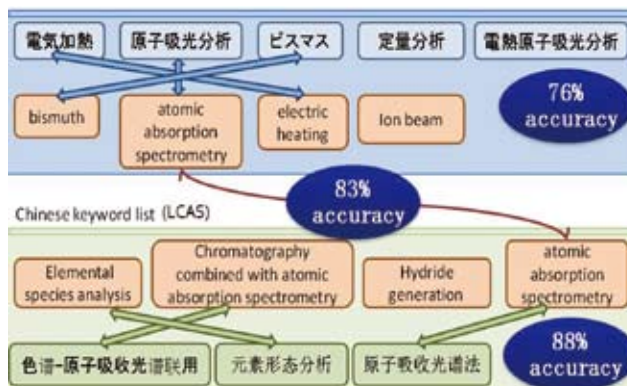


图19 词典构建原理

■ 每一文献记录中包含

- ◇ journal name
- ◇ keyword list
- ◇ title
- ◇ abstract

■ 大约80%的文档有中英文,其他仅为中文或英文

■ JST为所有文档的文摘建立了日文翻译,可用于平行语料

开发实用化机器翻译系统需要做的工作:

■ 提高对齐质量,需要考虑

◇ 基于实例的翻译系统对对齐错误非常敏感

◇ 功能词、虚词如日语情态表达式容易引起对齐误差

■ 中文句法分析错误的鲁棒性

◇ 中文句法分析困难极大(词切分、词性标注与依存分析)

◇ 需要鲁棒的方法处理句法分析错误

■ 构建一个“生态系统”

◇ 国家、机构与人之间的合作

◇ 共享知识、资源与技术

◇ 推出通用的服务,获得反

馈,改进产品

(2) ZHANG Junsheng's Towards Multilingual Scientific and Technical Information Service: Machine Translation and Cross-lingual Information Retrieval (ISTIC)

张钧胜博士介绍了ISTIC机器翻译与跨语言检索工作的使命:为科技信息服务提供技术支持;担负的任务:支持多语言科技信息检索、支持多语言科技信息的理解、支持多语言科技信息的翻译。目标是实现多语言科技信息的服务。目前,ISTIC多语言科技信息服务有:机器翻译、跨语言信息检索。

正在进行的工作:

☆ 日中机器翻译与跨语言信息检索

■ 构建日中翻译引擎

◆ 结合统计与实例机器翻译技术

◆ 建立实用化的日中跨语言检索系统

◆ 在几个研究研究采用更多的双语术语

■ 基于NSTL数据图书馆系统进行开发

☆ 多语言科技叙词表

■ 语种:日-英-中

■ 条目:大约60,000条;大约250,000个关系。

3 专题讨论

有关OA政策分组讨论的主题:变化与新发展,在科技信息领域和其相关领域适应新政府的管理;科技信息领域的开放获取。

有关技术与服务的分组讨论的主题:科技信息的标识系统;RDF与Linked Open Data;本体。

结语:第五届会议规划

1) 举办时间/场所

2014年,中国。主办方:ISTIC。

2) 举办方式

基本沿袭第四届会议的方式,并考虑有所创新。希望通过邀请国内相关机构的参与,扩大合作的深度和广度。

参考文献

- [1] 李颖,刘静波.基于DOI RAs机制的科学数据管理与共享:“第一届中日韩DOI项目合作讨论会”概述[J].数字图书馆论坛,2013(8).
- [2] 李颖,徐硕,朱礼军,等.面向技术创新服务与DOI应用的国际合作:“第三届中日韩科技信息机构联合研讨会”概述[J].数字图书馆论坛,2012(11):51-60.
- [3] 李颖,梁冰,乔晓东.“第二届中日韩科技信息机构联合会议”综述[J].数字图书馆论坛,2011(6).
- [4] 文部科学省科技与学术审议会[EB/OL]. [2013-09-12]. http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu4/toushin/attach/1323861.htm.
- [5] Open and shared access to scientific information [EB/OL]. [2013-09-12]. http://www.jst.go.jp/report/2012/121211_4_e.html.
- [6] オープンアクセスに関するJSTの方針[EB/OL]. [2013-09-12]. http://www.jst.go.jp/pr/intro/pdf/policy_openaccess.pdf.
- [7] 李颖,等.“科技期刊开放存取服务国际研讨会”综述[J].数字图书馆论坛,2010(6).
- [8] 李颖,等.日本学术信息的OA出版和机构仓储共建项目的概述及启示[J].数字图书馆论坛,2009(11).
- [9] G8科技大臣声明[EB/OL]. [2013-09-12]. <http://www.g8.utoronto.ca/science/130613-science.html>.
- [10] 刘静波,李颖.科学数据整合与管理开放平台P-CUBE[J].数字图书馆论坛,2013(8).
- [11] 李颖,等.基于DOI系统的中日韩三国之间科技信息共享[J].中国科技资源导刊,2012(6).
- [12] 李颖,乔晓东,深泽信之.日本科学技术振兴机构纵览及其科技信息流通促进体系[J].数字图书馆论坛,2011(3).
- [13] 李颖,于薇,张钧胜.“中日两国机器翻译技术合作研讨会”综述[J].数字图书馆论坛,2011(12).
- [14] 李颖,吴琳.从跨语言信息检索论日本国日中·中日机器翻译研发前沿[J].数字图书馆论坛,2008(9).

作者简介

李颖, 信息系统专业博士。研究方向: 知识工程, 知识服务系统。E-mail: liying@istic.ac.cn

Open Access and Digital Identifiers – Review for "the 4th China-Japan-Korea Joint Seminar"

Li Ying, Zhu Lijun, Zhang Junsheng, Zhang Jing / Institute of Scientific and Technical Information of China, Beijing, 100038

Zhang Xiumei, Yang Gongliang / Wanfang Med Online, Beijing, 100038

Abstract: First, this article briefly introduces "the 4th Japan-China-Korea S&T Information Institutes Joint Seminar" generally. Then, focusing the newly set up agendas, namely, open access, digital identification, as well as the Practical Chinese-Japanese Machine Translation System are introduced. Finally, the article gives the topics of Parallel Session, as well as the plan of the 5th Seminar.

Keywords: Open access, Digital identification, Chinese-Japanese Machine Translation, ISTIC, KISTI, JST

(收稿日期: 2013-10-09)

附录: 会议日程

Agenda 1: Evaluation and Analysis Based on STI

- 1) JST: Knowledge Flows in the J-Global foresight: Selected Indicators (Mari JIBU)
- 2) KISTI: Improving Data Quality for the Construction of National R&D Information in NTIS (Chulsu Lim)
- 3) JST: An initiative to share and use JST-funded research information (Tutomu KUROSAWA)



1 会议标识照片

Agenda 2: Open Access (includes Research Data)

- 1) JST: Open Access Policy (Akira NISHI)
- 2) KISTI: KISTI's Open Access Activities Update (Tae-Sul Seo)
- 3) KISTI: KISTI's Digital human Project and Medical Repository (Sang-Ho Lee)



2 JST大竹晓理事致辞

Agenda 3: Digital Identifiers (and Formats)

- 1) JST: Renewal of JaLC Project and its Updates (Takafumi KATO)
- 2) JST: Comprehensive Identification System - Improved Identification for Author Affiliation and Bibliographic Records (Takashi HORIE)
- 3) KISTI: Author Identification and Author Name Authority Control Project of KISTI (Seok-Hyoung Lee)
- 4) WanfangData (ISTIC): From Information Service to Intelligence Service—Practice of Wanfang Med Online (ZHANG Xiumei)

Agenda 4: e-journal

- 1) JST: The Future Role of J-STAGE (Kota AOYAMA)
- 2) KISTI: Life-cycle Journal Publishing Service (Tae-SulSeo)

Agenda 5: New STI Services & Technologies

- 1) JST: Promoting Science and Technology Exchange using Machine Translation (Kenichi SUMIMOTO) (Toshiaki NAKAZAWA)
- 2) ISTIC: Towards Multilingual Scientific and Technical Information Service: Machine Translation and Cross-lingual Information Retrieval (ZHANG Junsheng)
- 3) ISTIC: The Application of service virtualization technology in science and technology information services (ZHANG Jing)
- 4) ISTIC: Science & Technology Information Service for Technology Innovation (ZHU Lijun)
- 5) JST: Read & Research Map (Miho HORIUCHI)
- 6) JST: Ontology development in National Bioscience Database Center (Tatsuya KUSHIDA)

Agenda 6: Other

- 1) JST: ASTP (Asia Science & Technology Portal) Updates (Junko SHIRAISHI)
- Parallel Session (Round-Table Style)
Policy Oriented Session

Topics

1. Changes and new developments in response to the new governmental administrations in the area of STI and its related fields
 2. Open Access in STI (with emphasis on R&D data)
 3. Additional topics (to be agreed upon among the participants)
- Technical Oriented Session and Service Oriented Session

Topics

1. Identification in STI
2. RDF & LOD
3. ONTOLOGY



3 KISTI代表团长致辞



4 会场鸟瞰



5 日式工作餐照片



6 ISTIC/KISTI代表团长会间讨论