

DCMI 2013听会点滴

□ 曾蕾 / 美国肯特州立大学 美国肯特 44242

摘要: 文章分享作者参加都柏林核心元数据组织 (Dublin Core Metadata Initiative, DCMI) 2013年国际会议的几点体会, 重点在于采用规范数据支持数据实时混搭的两种不同方式、元数据属性映射的两种不同水平和表现方法, 以及数字图书馆数据模型的三种不同实现方案。

关键词: 都柏林元数据, DCMI, 关联数据, 规范数据, 元数据映射, 数据模型

DOI: 10.3772/j.issn.1673—2286.2013.12.002

2013年都柏林核心元数据组织 (Dublin Core Metadata Initiative, 简称DCMI) 国际元数据会议9月2日—6日在葡萄牙里斯本召开。今年的会议与iPRES2013 (International Conference on Preservation of Digital Objects, 数字对象保存国际会议) 共同举办, 与会者能选择任何一方的专题会议参加, 主旨发言也特别强调数据保存与元数据的关系。值得注意

的是今年DCMI首次实行与其他组织共同举办专场和学习班, 例如RDF词汇长期保存和管理 (Long-term Preservation and Governance of RDF Vocabularies) 专场是W3C (万维网联盟, World Wide Web Consortium) 协办的, CAMP-4-DATA网络基础构架及元数据协议 (Cyber-infrastructure & Metadata Protocols) 学习研讨班是研究数据联盟RDA (Research

Data Alliance) 协办的。从元数据方面的参加人员来看, 依然保持那种让人激动和向往创新的气氛, 不论是DCMI元老们 (Karen Coyle、Tom Baker、Gordon Dunsire、Diane Hillman、Jane Greenburg、Liddy Neville、Shigeo Sugimoto、秦健、陈雪华), 还是后起之秀 (Antoine Isaac、Kai Eckert), 或是新一代研究生 (李恺、Tsunagu Honma、Biligsaikhan Batjargal、

Mariana Curado Malta), 他们的文章和发言都引起听众极大的兴趣和踊跃提问, 整个会议内容实际、面向问题, 同时也让人很开眼界。秦健和李恺的关于科学元数据标准的可移植性的文章采用了大量数据, 而且提出了一系列衡量标准, 水平很高, 是四个最佳论文候选者之一, 在此希望二位作者能自己写成中文给大家介绍。



DCMI 2013会议学术委员会主席Kai Eckert发言

这次上海图书

馆的几位没能参会，本人临时受命将听会的几点体会在此与大家交流，但是因为事先没有准备写报道，会议期间也在忙专业组的工作，所以不能为大家提供全面和比较正式的报告。这里介绍的是在元数据项目中不同方法的比较，这是在听会过程中得到的点滴启示，结合本人已经知道的一些方法来进行一点讨论，供大家参考。

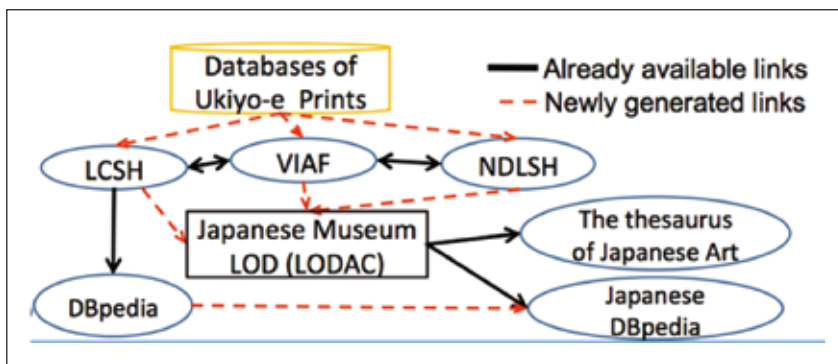


图1 通过规范数据联接多种资源^[1]

1 采用规范数据支持数据实时混搭的两种不同方式

日本立命馆大学的Biligsai Khan Batjargal介绍了他们采用关联数据驱动的动态联接来提供跨语种访问多种人文数据库的方法^[1]。立命馆大学的研究团队在关联数据资源、人名规范数据、主题词规范数据的基础上建立了一个联邦情报检索系统原型。以联接各种日本浮世绘版画（17世纪到20世纪）的资源为例，鉴于数据库可能来自不同的博物馆和数字图书馆，多语种、且没有统一的名称规范，在跨语种检索过程中要先针对用户输入的人物名称进行实时规范处理。用户输入名称后，系统及时将之纳入规范文档进行查找，一旦选定名称（例如歌川重，Hiroshige Utagawa），系统马上根据其国际虚拟规范文档VIAF的URI、英文DBpedia URI和日文DBpedia URI获取各种数据（关于歌川重本人及其作品等）进行实时混搭并提供给用户。我们知道国际虚拟规范文档VIAF、美国国会图书馆规范主题（LCSH）及名称文档，以及日本国家图书馆规范主题（NDLSH）及名称文档相互



图2 联合国粮农组织OpenAgris实时数据混搭结果截图^[2]

之间以及它们与DBpedia之间都有映射（图1中实线），其已发布成关联数据的值（value）可能已在不少数据集里被采用，通过它们进而又能与日本博物馆和其他浮世绘版画数据库进行联接，这样查回来的数据自然成为混搭的效果，给用户提供更多相关资源。

值得注意的是这种方式不同于联合国粮农组织等已经采用的方法。以粮农组织的OpenAGRIS为例，在显示被查到的书目数据时，OpenAGRIS会自动将文献内容相关的事实（例如某种植物在世界的

种植地段地图、属于濒临灭绝的动植物数据、某国家的基本档案等）以及相关的文献（例如《自然》杂志上的有关文章）同时混搭出来，使得每条书目数据都成为一个小型的知识网络，让读者感到有触类旁通的效果（图2）^[2]。

粮农组织的做法是利用其书目数据中的主题关键词，以其为出发点，通过在AGROVOC词表中已经匹配好的映射词的URI去“抓”其他含有同样主题URI的文献资源的元数据。这样做的前提是先有受控词汇之间的映射。粮农组织花了很

多人力将词表与十多个词表类表做了映射。在主题词表中这种做法是很合适的,但对于人名机构名称,这种做法很难达到类似的效果。首先是量大,名称规范的词条量比主题规范的词条量要大很多倍;第二是增加很快,每天都有新人成为作者或者成为新闻人物,提前将已有规范文档匹配起来不现实;第三是多语种。由此可见日本研究者提出的这个方法的可行性和合理性。

2 元数据属性映射的两种不同水平和表现方法

我们知道在试图合并数据或者转换数据时,要先对数据格式的可互操作性作出评价并采用映射对照表来列举对应的元素和字段(例如DC的“creator”对应Marc的1xx字段a子字段)。长期以来人们采用“crosswalk”(直译为人行横道,意译是对应表)这个词来表现这种映射方法和结果。在DC2013年会上,Diane Hillmann、Gordon Dunsire和Jon Phipps三员大将特别指出了对应表的局限^[3]。他们提出应该提倡用“map”(地图)以便看出映射的情况和断层情况。所谓映射“地图”是两个或两个以上的RDF元素(类、属性、概念)通过本体属性(RDFS、OWL)的连接。有很多映射关系,人当然能判断出其匹配程度,但是机器就不行了,所以有必要用机器能看懂的方式准确表达出来。图3是一个具体的地图。举最中间的一条属性为例,最下面是RDA的本体类“剧本作家(Screenwriter)”,是RDA的“作者(Author)”的子类,“作者”的上位类为RDA的“创作者(Creator)”类。这个类或许可

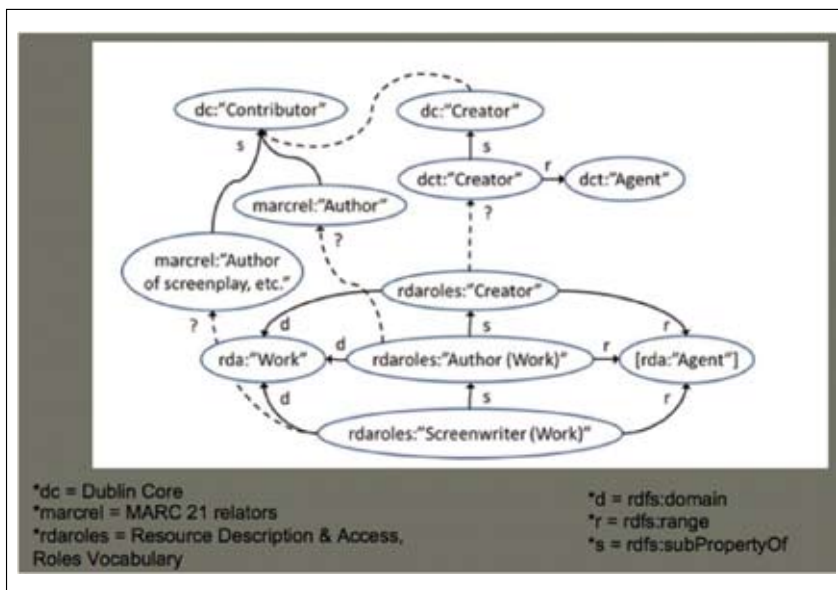


图3 采用自下而上的匹配方法得出的RDF元素映射地图^[3]

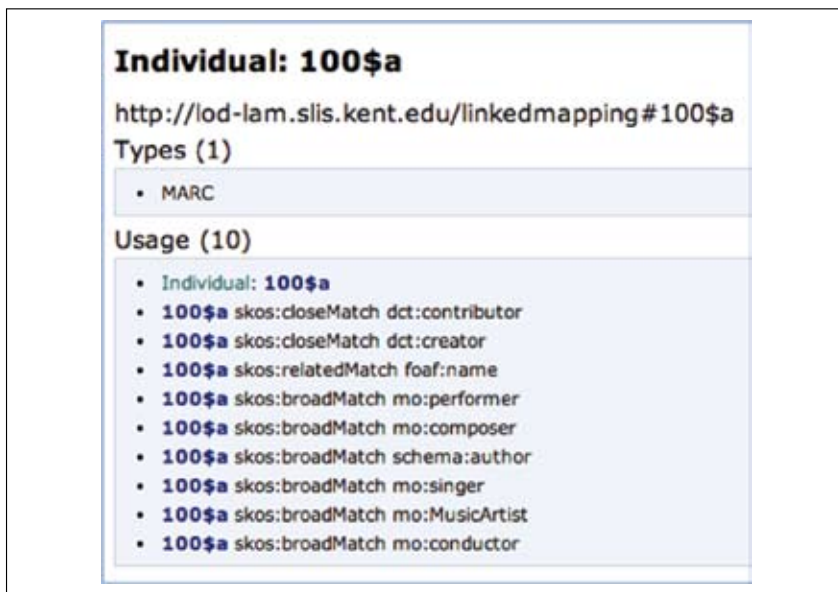


图4 MARC 100字段a子字段与DCT、音乐本体(MO)、schema.org、FOAF有关属性的映射^[4]

以与DCT的“创作者(Creator)”相对应。但是看看左边那一系列匹配,出现了一系列不能确定等同映射的情况。最后归到最上面两个DC元素(Contributor和Creator)时,在这两者之间可以看出是有很大的鸿沟的。

总的来说,对应表是不同元数据格式之间的转换,只能反映看到

非常有限的映射。而映射地图则显示出不同元数据元素之间的上下、平行、重合、鸿沟等位置,使用户对这些不同元素间的映射关系一目了然,也便于机器去理解。

实际上我们可能在研究和实际工作中早已采用了类似这种地图的方法,在匹配中引进SKOS的不同的匹配程度的表述,包括准确匹

配、广义匹配、狭义匹配、相关匹配、交叉匹配等,只不过大家一直采用“对应表”这个词。这里有一段截屏(图4)^[4],是本人的关联数据团队做的映射表中的一个小部分,图中展示MARC 100字段a子字段与DCT、音乐本体(MO)、schema.org之间属性的映射,比如最后一行说明:MARC 100字段a子字段(责任者)与音乐本体里的“指挥mo:conductor”的匹配关系是广义匹配(broadMatch),不是对等(exact)匹配。

3 数字图书馆数据模型的三种不同实现方案

另外体会到的一个“不同之处”是数据模型的实现方式。在一个专题专场中,Karen Coyle和Tom Baker特意提出了元数据应用纲要(Application Profile)作为OWL本体的可取代形式的观点。这一场中欧洲数字图书馆Europeana的Antoine Issac的发言最给人以启示,他从Europeana的数据模型与对数据应证(validation)功能的要求的关系来看,讨论到底是做成什么好:XML schema、OWL本体,还是做成应用纲要(application profile)的形式来实现^[5]。Europeana已经有来自2200个博物馆、档案馆、图书馆以及其他机构共2900万的对象元数据,其数据模型需要达到的目的包括:(1)区分“对象”(绘画、书籍、电影等)

与它们的数字表现;(2)区分对象及其元数据记录;(3)允许同一个对象的多个记录,包括潜在的有矛盾的陈述;(4)支持由对象组成的对象;(5)支持上下文的资源,包括受控词表里的概念。对于大量的来自多方的数据,数据模型应该能进行监测以保证进来的数据满足Europeana的基本数据功能并达到一定的质量。

Antoine Issac和团队将数据模型先做成OWL本体,但是其监测的功能不强。他们又将数据模型用XML schema的形式来实现,这样做虽然能解决监测的问题,但是局限于XML格式,而且回到以元数据整条“记录”为基础的老的方法,不是RDF强调的三元组“陈述”为基础的方法。当外面数据进来时,又成了内部格式与外部格式要匹配的问题。XML schema一级的限制导致大量的重复声明,而且很难贯彻限定条件。那么现在这种数据模型是

不是可以按照应用纲要的方法来实现呢?从内容上来看,其数据模型已经融合了几种元数据标准的元素,每个元素也都有限定条件,看来符合应用纲要的要求,但是还要能与SPARQL关联数据查询语句的限定条件相吻合。下面他们还会做更多的测试,但是基本上已经决定放弃OWL本体和XML Schema的办法。

以上是几点令人思考的问题和方法,因为觉得很受启发,就在此作为重点讨论了。DCMI已经将所有发言资料公布在网上,包括培训课程、主旨演讲、会议论文的文章全文以及发言PPT,都可在<http://dcevents.dublincore.org/IntConf/dc-2013/schedConf/presentations>处查到^[6],是十分珍贵的资料。

今年DCMI还有一件大事,就是它脱离了原来在新加坡注册管理的单位。经过挑选,最后和ASIS&T(原来名称为“美国信息科学与技术学会,The American



DCMI 2013会议现场的学生表演

Society for Information Science and Technology”，现在正式改名为“信息科学与技术学会，Association for Information Science and Technology”，缩写不变）达成协议，成为ASIS&T的一个项目，前景非常乐观。DCMI顾问委员会也开始更加结构和功能化，未来一年间将成立若干执行委员会，分管管理、教育、开发、应用等方面的主题和活动。今年DCMI也正式开始实行会员制，已经有很多单位和个人成为DCMI成员。2014年DCMI年会在美国得克萨斯的奥斯丁召开，希望能见到更多来自中国的同行。



曾蕾和秦健（左，美国雪城大学）、陈雪华（中，台湾大学）在DCMI 2013会议

参考文献

- [1] BATJARGAL B, KUYAMA T, KIMURA F, MAEDAA. Linked Data driven dynamic Web services for providing multilingual access to diverse Japanese humanities databases [C/OL]//DC2013: Linking to the Future. International Conference on Dublin Core and Metadata Applications, Sept. 2-6, 2013. Lisbon, Lisbon, Portugal. <http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/150/129>.
- [2] OpenAgris. Agricultural Information Management Standards (AIMS). FAO of the UN [OL]. [2013-10-24]. <http://aims.fao.org/openagris>.
- [3] HILLMANN D I, DUNSIRE G, PHIPPS J. Maps and gaps: strategies for vocabulary design and development [C/OL]//DC2013: Linking to the Future. International Conference on Dublin Core and Metadata Applications, Sept. 2-6, 2013. Lisbon, Lisbon, Portugal. <http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/185/133>.
- [4] Linked Open Data- Library Archives Museum Research Group, Kent State University. Crosswalks from MARC to Digital Collections and Music Ontology [OL]. [2013-10-24]. http://lod-lam.slis.kent.edu/07/classes/MARC___442367101.html.
- [5] ISAAC A. Validation of Europeana data: application profile, OWL ontology, or else? [C/OL]//DC2013: Linking to the Future. International Conference on Dublin Core and Metadata Applications, Sept. 2-6, 2013. Lisbon, Lisbon, Portugal. <http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/139/159>.
- [6] DC2013: Linking to the Future. International Conference on Dublin Core and Metadata Applications, Sept. 2-6, 2013. Lisbon, Lisbon, Portugal. Presentations and Authors [OL]. [2013-10-24]. <http://dcevents.dublincore.org/IntConf/dc-2013/schedConf/presentations>.

作者简介

曾蕾 (Marcia Lei Zeng)，教授，DCMI首位咨询委员会主席。研究方向：知识组织系统、元数据、关联数据、语义技术。
个人主页：<http://marciazeng.slis.kent.edu/>。E-mail: mzeng@kent.edu

Notes from DCMI 2013 Conference

Marcia Lei Zeng / Kent State University, Kent, OH 44242, USA

Abstract: The author shares her experience of participating in the 2013 DCMI (Dublin Core Metadata Initiative) International Conference on Dublin Core and Metadata Applications, with the emphases on some of the issues and processes, including two different methodologies used in mashups of Linked Data supported by controlled vocabularies and authority files, two different levels of metadata property mapping and their representations, and three different approaches for implementing the data model in a large digital library.

Keywords: Dublin Core Metadata, DCMI, Linked data, Authority data, Metadata mapping, Data model

(收稿日期：2013-11-11)