欧洲数字资源平台(Europeana)核心技术浅析

摘 要 本文主要从战略规划、知识版权、关联数据模型(EDM)、语义融合、云计算等方面分析了数字图书馆建设成功的案例Europeana的核心技术。将国内外的数字图书馆的核心技术与Europeana数字图书馆进行比较，从中分析Europeana成功的关键因素。

关键词 Europeana;欧洲数字资源平台；EDM；开放互联；语义融合

1 引言

随着互联网产业的快速发展，人们都希望可以更便捷地获取想要的信息。与此同时，各国的文化资源共享机构也希望能够更好地为人们提供信息搜索、浏览以及查询等服务[1]。近年来，世界各国针对文化资源整合问题，开展了一些线上的数字化文化资源服务平台的建设。如欧洲数字图书馆Europeana，世界数字图书馆WDL( World Digital Library)，美国数字公共图书馆DPLA( Digital Public Library of America) ，法国国家图书馆Gallica，澳大利亚国家图书馆Trove，荷兰国家图书馆Delpher等。在这些数字文化资源展示平台中，建设成果较为显著的就是Europeana，其无论在战略规划、数据模型、云服务以及语义融合等方面都取得了不错的进展。Europeana是数字文化资源服务平台建设中比较成功的一个项目，其将欧洲的图书馆、档案馆以及博物馆等具有文化历史资源的机构整合到一个统一的平台，初步实现了人们一站式的搜索和浏览服务[2]。

Europeana的诞生因素主要有三个方面：一是20世纪初，文化产业渐渐进入人们的视线，而欧洲的各大图书馆、档案馆以及博物馆等文化机构拥有其它地区数量以及质量都无法匹敌的文献馆藏，将这些馆藏数字化对文化产业的推动具有非常重要的作用。二是随着互联网的发展，人们对于线上平台的使用开始逐渐地增加，将文献资源数字化，对文化的共享和传承具有很重要的意义，同时也是对欧洲现有文化的一种保护[3]。三是2004年12月，全球最大的搜索引擎公司Google宣布要在几年时间内建成拥有1500万数字化图书的全球最大的数字图书馆，一支对美国霸权统治嗤之以鼻的法国人，认为这是美国文化入侵的方式，为了抵制文化入侵，欧盟各国提议共同构建一个欧洲数字图书馆，于是2005年欧洲19个国家联合创建欧洲数字图书馆(European Digital Library)。随着加入的国家文化机构越来越多，最后更名为Europeana。

Europeana的初衷是提供一个整合了全欧洲具有文化资源的开放统一的线上平台，为人们了解欧洲文化提供一个一站式的浏览与搜索服务平台，抵制外来文化的冲击，实现欧洲优秀文化的传承与发展。Europeana项目建设的初期主要是纸质资源的数字化，2007年数字化的资料已达十万多册，继而2008年11月Europeana平台(http ://www.europeana.eu )正式对外开放，对欧洲的文学、电影、建筑等感兴趣的人们能通过平台访问多达200万册的数字资料[4]。2011到2015，Europeana平台致力于资料内容整合，实现人们不受时空限制的访问文献资料，开辟新的方法来民众参与Europeana平台的建设，为欧洲文化知识的传播和继承贡献更大的力量[5]。更在2015年，资料内容数据达到4800万余条，将欧洲的图书馆、档案馆以及博物馆等文化资源机构融合在一起。Europeana的成功案例也是世界上线上资料整合平台的标杆，为我们国家数字图书馆建设指明了道路和方向。

[4]

2 相关工作

2.1 国内数字资源平台与核心技术

国图 花瓣 视觉中国

中国国家图书馆数字化文津搜索于2012年开始正式运转，成为了国内数字资源分享平台。国家数字图书馆通过一期、二期工程将馆藏文献资料数字化，再整合公共文献资源以及大量的网络资源，包括期刊、报纸、论文、文档、文献等类型，国家数字图书馆的目标是建成全球最大的中文文献发现平台。花瓣网是一家”类Pinterest”网站，主要为用户提供图像采集工具以及用户将自己喜爱的图片收藏。同时，也是一个小型的社交网站，用户可以上传分享自己喜爱的图片以供别人浏览下载，也可以浏览其他用户收藏的有趣内容，结交朋友，发现新知。花瓣网的信念是”重新链接/收藏/整理/分享你所感兴趣的事物”。2000年视觉中国创立，是中国最大的视觉服务平台。视觉中国的三大业务核心为“视觉内容与服务”、“视觉社区”和“视觉数字娱乐”。整合了其为客户提供高质量的图片、音乐等一站式服务。平台技术基于云计算、大数据以及最新的存储技术，是中国最大的互联网版权交易平台。

2.2国外数字资源平台与核心技术

数字典藏(台湾) 中华书局 数字人文组织联盟(ADHO)

2008年8月澳大利亚国家图书馆建设Trove平台，目的是为了构建一个澳大利亚统一的数字资源平台。同Europeana一样整合图书馆、档案馆等文化机构的档案、资料、音乐以及文献，以便人们更便利的浏览检索澳大利亚文化资源，同时也是对其文化资源的保护。美国数字公共图书馆(DPLA)于2013年4月正式为大众提供服务，其致力于整合美国所有的图书馆、档案馆等文化资源共享机构的资源，建成开放的文化资源共享平台。该平台的理念是公共服务，像公共图书馆一样公平、自由地为人们提供服务。世界数字图书馆(WDL)2005开始筹划建设，于2009年4月正式上线为人们提供查询浏览联合国成员国档案馆、图书馆等资料的服务。为联合国各国文化友好交流提供了支持，为需要文化资源服务的个人、研究团体等提供了便捷的帮助，通过加强各成员国的合作，打破国家间无法轻易逾越的障碍。

2.3 Europeana平台(简介 技术罗列)

2004年12月，全球最大的搜索引擎公司Google对外宣布，计划建成世界上最大的数字图书馆，一直和美国不对付的法国人为了反击所谓美国文化入侵，提出了建立欧洲数字图书馆的设想，一直提出欧洲委员会全票通过，然后Europeana平台建设正式提上日程。2008年Europeana上线以来，凭借其快速的发展引起了世界的注意。

Europeana整合欧洲的图书馆、档案馆以及博物馆等文化资源机构的资料，将之汇集于一个平台。由于欧洲各国文化机构资源储存的元数据格式不一致，在资源收集方面造成很大的困难。Europeana小组参考RDF等规范后，将原先基于DC标准的ESE数据模型升级为EDM模型，创建了拥有自己特定属性和类的EDM模型，解决了元数据格式问题，成功将不同格式的元数据进行融合。再根据OCLC的本体相似度计算方法，将融合的元数据投在同一空间度量下，计算元数据之间的关系，以RDF资源链接来表示元数据之间的映射并将之存储，实现文化资源语义融合。通过HTTP URIs方式结合LOD云与外部开放的数据集建立映射关系，实现Europeana跨域资源语义融合。以分布式语义资源搜索，提供关联数据的SPARQL端点查询的方式进行关联数据的发布，并提供开放API，实现Europeana的开放互联。

另外，Europeana平台成功与它的战略规划也有很大的关系，从2008年平台发布以来，针对每个时期的工作重点制定合适的计划，制定的两个五年计划是Europeana走向全世界的关键。Europeana Cloud分为计算云和存储云的设计既满足Europeana内容提供商低成本资源存储的设备也满足科研人员开展探索研究新空间的需求。

3 Europeana核心技术

3.1 ESE

欧洲各国图书馆、档案馆和博物馆等文化机构存储的资源采用不同的元数据方案，要对不同标准的元数据进行有效地组织和管理，需要建立统一的数据规则。为此，Europeana根据自身数据特点，在2008年11月设计了Europeana Semantic Element(ESE)的元数据方案，它是基于Dublin Core的一种扁平式数据结构，添加了欧洲特定字段，如dataProvider或uri。

作为Europeana互操作性问题的基本解决方案，ESE存在许多问题。首先，它是一个为元数据字段使用简单字符串值的平面模型。这会妨碍收录入Europeana的项目正确链接到其他对象（例如一系列画像）或由复杂资源表示的上下文实体，例如具有多种命名变体的创造者或比另一个更具体的概念。此外，ESE聚合在同一个记录元数据字段中，可应用于不同的实体，从而打破了一对一的原则，并造成极大的混淆：例如，某些提供商使用权限或日期相关的字段描述他们持有物体信息，但其他人也会使用相同的字段发送有关这些物品的数据。

ESE的问题在于其平坦的建模方法，它不允许在Web上嵌入到外部资源的链接以及在更专业化，细粒度的模型方面的不可扩展性。然而，在迈向关联开放数据之前，要克服的主要障碍是欧洲语义元素（ESE）格式及其基础简单且强大的数据模型。ESE的创造和使用作为欧洲首个早期原型的先决条件是不可避免的，但与此同时，它的局限性已经变得可见。

3.2 EDM

Europeana为了避免ESE扁平式数据结构的一些缺点，参照RDF、METS等规范对ESE数据模型升级，设计了Europeana data model（简称EDM），EDM在不改变原来ESE字段的前提下，又加入了EDM本身特有的类和属性，能够对不同元数据的资源进行统一的规范，EDM在外部资源连接的嵌入和可扩展性都大大加强，与现有的其它数据模型(EAD、MARC)更好的融合，实现元数据语义互操作。

EDM主要结构由三个核心类组成，数据资源聚合类(Europeana aggregation)、数字文化资源对象类(ProvidedCHO)及其对应的资源数字表现形式类(NonInformation resource)。数据资源聚合类将数据对象、网络资源以及情景关系通过聚合关系聚成一个数据集合。用事件、主体、地点、自然实体、控制词表和时间范围等字段来描述对象的情景信息，共同构成资源数字表现形式类。



图1 EDM核心类层次结构

EDM主要属性间有四种关系：一是数据聚合关系，由资源呈现形式、数据来源网页以及聚合结构关系组成；二是资源对象间的相似关系，包括衍生、后继、包含等；三是数据资源情景关系，包括事件发生时段、事件发生场所、事件发生主题、对象相似关系、邻接、资源共现以及包含链接等；四是主题对象间的关联关系，由主题、类型、实现、关联关系、对象主题开始以及对象主题结束等组成。



图2 EDM主要属性关系

3.3 开放互联 语义融合

ESE元数据模型数字化文化资源，元数据会有多种格式(如TEI、LIDO、EAD、MARC)，格式不统一对语义关联数据网络构建带来诸多不便，Europeana通过EDM模型元数据映射规则将不同格式的元数据与本体间建立一一映射构成符合EDM描述规则的数据集，通过扩展样式表转换语言(XSTL)数据集转换成契合RDF/XML语言格式描述的元数据集，成功将不同格式的元数据进行语义融合。

Europeana利用OCLC的本体相似度计算方法，用空间向量来度量语义融合之后的元数据关系，关联关系结合SKOS、GEO、VIAF等受控词表映射加以形式化语言描述，以RDF形式存储映射关系，通过RDF语义链接间的关系描述数字文化资源间的关联关系，实现数字文化资源间的语义融合。

Europeana资源用URI标识和定位资源或者资源的链接，对领域资源关联数据集中有URI标识的资源对象语义标注，通过HTTP URIs方式结合LOD云与外部开放的数据集建立映射关系，实现Europeana跨域资源语义融合。

Europeana以RDF形式存储资源以及资源链接，关联数据的发布必采用基于RDF方式存储的发布形式，由于Europeana资源数据是以RDF三元组格式存储，采用分布式语义资源搜索，提供关联数据的SPARQL端点查询。

Europeana除了提供SPARQL端点查询外，提供了4个应用程序编程接口(API), 分别为REST-API、Annotations API、OAI-PMH及Linked Open Data，Europeana基于这些API接口开发了158个APP供关联数据开放使用。



图3 Europeana语义融合

3.4 知识版权

版权问题一直是数字图书馆建设不可避免的问题，尤其Europeana整合欧洲各国文化遗产资源，原本在著作人、用户以及文化服务机构之间的知识产权平衡很容易被打破，Europeana结合自身特点采用积极的策略应对版权问题。

Europeana首先收集公共领域的文化资源数据，这些资源大多为文学、历史书籍、建筑资料，没有版权归属问题，将这些纸质资源首先进行数字化处理，在一定程度上对这些文化资源保护。为了确保对欧洲文化文献感兴趣的人们能正常的检索和预览，在2010年发布《公有领域宪章》( Public Domain Charter)，随后又制定了《公有领域使用指南》( Public Domain Usage Guidelines)。

Europeana版权小组秉着先急后缓的原则，制定了一系列版权保护原则，如孤本作品原则、高级原则以及绝版作品原则等。高级原则主要针对版权所有者和文化资源共享机构。对版权拥有者而言，应当尊重其版权及其应该受到保护的权利，文化资源共享机构无论是对其作品数字化还是传播使用，都必须征得版权所有者的同意。对文化资源共享机构而言，在举行文化资源交流活动必须取得法律认可，对有版权的作品整理收集时，一定要获得版权保护人的许可。所谓孤本作品就是版权拥有者不能确定或者找不到的作品。无论是文化资源共享机构还是个人在使用孤本作品之前，需要进行以找到版权所有人为目的合理的检索，再使用孤本作品之前，需要在找版权所有人方面付出一定的努力。孤本作品原则鼓励Europeana的成员制定相应的策略，若在使用孤本作品之前找到其版权所有者，则要付以相应的报酬，然后其作品就可以正常的传播使用。绝版作品在与版权所有人谈好利益分配取得许可之后，图书馆等文化机构可以对其付费注册用户提供服务，也可以对其他机构进行接口共享。

3.5 云计算与云服务

Europeana从2008年11月上线以来，Europeana用户可以通过其门户网站访问历史、文学、建筑等文化资源。在Europeana内容提供商需要代价更低的资源存储设备、Europeana科研人员需要新的空间继续展开科研探索以及其在2015年数字化资源要达到3000万以上的伟大目标的背景下，Europeana于2013年启动了Europeana Cloud计划。云计划从6个维度满足内容提供商和科研人员的需求。Europeana Cloud计划通过工作分解结构(Work Breakdown Structure)将工作项目分成7个Work Package，分别为需求评估、基础设施建设、服务模式开辟、元数据创建、可持续化发展、网络平台建设以及Europeana项目管理。



图4 Europeana Cloud系统架构

Europeana Cloud项目包括两种云，即计算云和存储云。计算云由前端系统和后端系统组成，数据处理验证服务、元数据及内容服务、数据注释、标识符服务以及通知服务共同构成前端系统，前端系统主要负责和用户进行交互，又有功能性服务端说法，向外提供客户端API。由异步消息传送服务、登录服务、许可/授权服务共同构成后端系统，后端系统主要是对前台数据逻辑判断、控制，虽然没有直接服务于用户，但也是计算云中不可或缺的一部分。存储云采用NoSQL数据库和分布式文件系统，主要为计算云部分提供数据存储功能。

Europeana根据市场主流制定云服务策略，确定了五类云服务对象，即对文化历史感兴趣的一般用户、正在接受教育的学生、科学研究的工作者、从事文化遗产工作的专业用户以及专家研究员[6]。

3.6 战略规划

Europeana数字图书馆建设的成功与其制定的一系列战略规划密不可分。自2008年Europeana正式对外开放以来， Europeana就制定了“2011-2015年战略”和“远景2020”两个五年计划。后者在Europeana发展迅速的基础上将工作的重心慢慢转移到资源内容的质量和市场的需求上，来Europeana平台对其用户更加友好。前者正好是Europeana服务战略转变的时期，Europeana致力于跨域合作、信息共享，慢慢地将资源的集成方式开始向分布式转变，以此来增加资源的种类和数量。其特色是充分体现用户至上理念，从资源的可用性、易用性、可重用性以及易嵌入性四个方面来保障Europeana用户的权益和利益。

4 分析与对比

4.1国内外数字资源平台核心技术(指标)

资源数量一直是数字化服务平台服务能力的指标之一，反映了数字平台资源的多样性和丰富性。从表1得知，拥有数字资源最多的平台是Trove，其资源总量达到5.4亿件。中国国家数字图书馆的文津搜索也达到了3.6亿件左右，仅次于Trove平台。WDL、DPLA和Europeana岁在资源总量上比Trove和文津搜索少，Europeana资源总量5400万也是比较全的数字化资源平台。采集文献资源的语种也是度量数字化平台资源全面性的指标，其中当属DPLA收集的语种最多，达到400多种。WDL和Europeana在语种方面分别达到134种和37种。其中Europeana平台为了消除语言的隔阂还提供了语言的翻译功能。文津搜索和Trove等在资源收集语种方面相对较少，但也在慢慢发展。

4.2 与Europeana平台对比(列表)

WDL、EUropeana、DPLA等各大平台都取得了较为突出的成功，各平台都有专门的组织机构负责平台的运营，并且WDL、Trove、DPLA、Europeana、文津搜索等平台都与各自区域内各大图书馆、档案馆等文化机构建立了密切的合作，保证了平台文化资源的供应。花瓣网以及视觉中国等虽然大部分与个人机构建立合作，但在如今互联网发展迅速的时代，个人机构的信息资源储备量也达到了前所未有的水平，所以与个人机构合作也是平台资源融合的途径之一。为了数字平台更好地共享文化资源，服务对文化文献感兴趣的用户，各大平台大都提供多种语言支持，例如视觉中国、文津搜索提供中英文支持，而WDL更是支持了阿拉伯语、中文、英文、法文、西班牙文、俄文、葡萄牙文七种语言，欧洲区域性数字图书馆Europeana更是支持英文、法文等23国欧盟官方语言。Trove和DPLA已经意识到支持语言种类对打破分享平台语言障碍的巨大作用，但发展相对比较慢。元数据采集是平台建设根本，不同的平台会根据资源的种类以及元数据来拟定不同的采集方式，WDL、花瓣网、视觉中国等采用资源元数据提交上传是最直接、廉价的方式，但像WDL这种接受实体数字资源上传的方式，对数据的格式规范方面是一个问题。DPLA只是采集元数据，数字资源的特征可能会展示不完全。Europeana采用的方式较为规范，其收割用EDM数据模型统一格式后的元数据，在数据规范一致性方面有了保障。(详见表1)

表1 Europeana与各平台比较

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 平台 | 组织机构 | 合作范围 | 合作机构类型 | 支持语言 | 资源采集方式 | 资源总量 |
| WDL | 美国国会图书馆项目组 | 面向全球 | 图书馆、档案馆 | 阿拉伯语、中文、英文、法文、西班牙文、俄文、葡萄牙文 | 采集元数据，同时接受实体数字资源上传 | 84万件 |
| Trove | 澳大利亚国家图书馆 | 面向澳大利亚 | 图书馆、档案馆 | 英文 | 收割元数据 | 5.4亿件 |
| DPLA | 美国指导委员会 | 面向美国 | 图书馆、档案馆、博物馆 | 英文 | 采集元数据 | 1600万件 |
| 花瓣网 | 个人机构 | 面向中国 | 个人机构 | 中文 | 采集元数据和提交元数据 | 不清楚 |
| 视觉中国 | 个人机构 | 面向全球 | 个人机构 | 中文、英文 | 采集元数据和提交元数据 | 不清楚 |
| 文津搜索 | 国家图书馆 | 面向全球 | 图书馆以及文化机构 | 中文、英文 | 提交元数据 | 3.6亿件 |
| Europeana | Europeana基金会 | 面向欧洲 | 图书馆、档案馆、博物馆 | 英文、法文等23国欧盟官方语言 | 收割统一格式元数据 | 5400万件 |

5 小结

本文介绍数字图书馆建设成功的案例Europeana平台。通过简单介绍Europeana统一数据源格式的EDM模型、关联开放语义网络、知识版权等方面，以便于为人们了解Europeana数字图书馆平台。然后通过与国内外现有的数字化资源平台从资源总量、语言支持等方面来做了简单的对比。分析Europeana资源收集方式等建设方面的优点。

6 参考文献

[1]王娜. 欧盟数字图书馆Europeana建设研究及启示[J]. 图书馆学刊, 2017(10):135-138.

[2]汪静. Europeana发展现状及启示[J]. 数字图书馆论坛, 2017(3):46-53.

[3]肖德玉. 创建一个公共的文化空间——欧洲数字图书馆项目[J]. 数字图书馆论坛, 2007(9):63-67.

[4]刘剑英. 欧洲数字图书馆“Europeana”现已在线开放[J]. 晋图学刊, 2009(2):80-80.

[5]王莉亚. Europeana 2011—2015战略规划[J]. 图书情报工作动态, 2011(3):1-9.

[6]Purday J. Think culture: Europeana.eu from concept to construction[J]. Bibliothek Forschung Und Praxis, 2009, 33(2):170-180.