情报工程 ISSN: 2095-915X
第1卷 第5期 68-76 2015年10月
TECHNOLOGY INTELLIGENCE ENGINEERING
ISSN: 2095-915X Vol.1 No.5 68-76 Oct 2015

doi:10.3772/j.issn.2095-915x.2015.05.009

基于 CKAN 的社会科学开放数据服务平台构建初探

余文婷，梁少博，吴丹
（武汉大学信息管理学院 武汉 430072）

摘要：运用开放知识基金会的开源软件 CKAN，根据 OAD 中的社会科学开放数据集，提出社会科学开放数据服务平台要素、关键功能及其实现方法，并结合 CKAN 的成功案例对平台功能进行具体描述。

关键词：开放科学数据，开放数据服务，平台模型，CKAN

A Preliminary Study on the Construction of Open Research Data Service Platform in Social Sciences Domain Based on CKAN

YU Wenting, LIANG Shaobo, WU Dan
(School of Information Management Wuhan 430072)

Abstract : This paper uses CKAN, the open source software developed by Open Knowledge Foundation, to put forward the key functions of an open social sciences data service platform and their realization. The social sciences datasets are chosen from OAD. Then some successful cases are given to demonstrate the functions.

Keywords : Open research data , Open data service , Platform model ,CKAN

基金项目：本文系国家万人计划青年拔尖人才计划项目“大数据环境下用户学术信息行为研究”的研究成果之一。
作者简介：余文婷，女，武汉大学信息管理学院硕士研究生。主要研究方向：信息组织与检索。E-mail：yuwenting1215@whu.edu.cn。通讯方式：15271904542。吴丹，女，博士，教授，博士生导师，武汉大学信息管理学院图书馆学系副主任。主要研究方向：信息组织与检索。E-mail：woodan@whu.edu.cn。通讯方式：13971252200。

1 引言

近年来，开放科学数据运动日益得到包括政府、学者、科研院所和图书馆等各方关注。开放科学数据是指通过收集、观察和创造的各种实验数据、观测数据、统计数据、仿真数据，表现为表格，数字、图像、多媒体或其他格式。可以是论文后附带的实验数据，也可以是独立的原始数据，包括对数据进行描述的元数据、数据集以及与数据相关的出版物[1]。同其他的开放存取运动一样，这些数据是免费的，且在获取、使用和重用上没有来自知识产权或其他机制的限制[2]。开放科学数据是前人研究的重要成果与经验的结晶，具有巨大科学与社会价值。如能对其进行创新性和个性化分析利用，就可以将开放数据转化成新的知识产出，更好地将所蕴含的内容充分挖掘与利用，加上没有外在的传播障碍与获取障碍，终将极大地提高知识生产、传播与利用的效率。另外，科研工作者也有着开放数据相关服务的需求。e-Science 和 e-Research 环境下，科学研究正在向数据密集 (data intensive) 型科研转变[3]，研究者们更加注重在海量数据中直接挖掘所需信息、知识和智慧。欧盟的一项调查显示，高达 91% 的科研工作者认为对已有数据的再分析十分重要[4]。所以，图书馆等信息机构应充分重视开放可续数据的价值和科研人员的需要，提供高质量的开放数据发现与利用服务，从而促进科研活动的效率，提高科学生产力。本文以 OAD[1] (Open Access Directory, 开放存取目录) 的社会科学开放数据为例，选取开放科学基金会[2](Open Knowledge Foundation, OKF) 开发的开放数据服务工具 CKAN 为开发平台，通过分析开放科学数据服务平台模型要素、提供的关键服务功能及实现方法，并结合已有平台，为图书馆支持开放数据发现、利用服务及平台实现提供借鉴与参考。

2 基于 CKAN 的社会科学开放数据服务平台构建

2.1 CKAN 简介

CKAN[5] 是开放知识基金会开发的一款用于发布、查找和利用数据的开源门户平台工具。专为各类有着开放数据并使其可利用需求的数据发布者而设计。CKAN 目前已被超过 40 个数据仓储使用，包括地方、国家等各级政府机构和国际组织等用它来发布官方和社团数据，如英国的 data.gov.uk 和欧盟的 publicdata.eu[6]。利用 CKAN，可以提供成熟的开源数据管理解决方法，如每个数据集都有自己的界面，并包含丰富的元数据集，使得数据成为有价值的易检资源。具体说来，它包括以下 7 个特点：具有易用网络界面和强大 API 的完整目录系统；细粒度获取控制；和诸如 Drupal 和 WordPress 等第三方内容管理系统 (Content Management System, CMS) 的紧密整合；集成数据存储和完整数据 API；数据可视化及分析；方便通过普通查询建立新例子的联合结构；以及支持部门或者小组管理自身数据发布的工作流 (workflow) 功能，涉及发布、查找、保存、管理、交互和扩展等整个开放数据生命周期[7]。为了更好地支持开放数据服务，CKAN 是一款免费软件，利用不受其他机制限制。不过需特别注意的是，使用前需要获得所要发布的数据和元数据的使用和发布许可[8]。

1 OAD 是 wiki 形式的开放数据仓储和数据集列表，由 OA 团体自由维护。（http://oad.simmons.edu/oadwiki/Data_repositories.）
2 Open Knowledge Foundation 是全球开放知识活动领导者，促进其利用和价值实现。(http://okfn.org/.)

2.2 社会科学开放数据服务平台要素

根据已有服务平台构建经验，本文从数据、技术以及服务主体和功能三个方面构建社会科学开放数据服务平台，如图 1 所示。

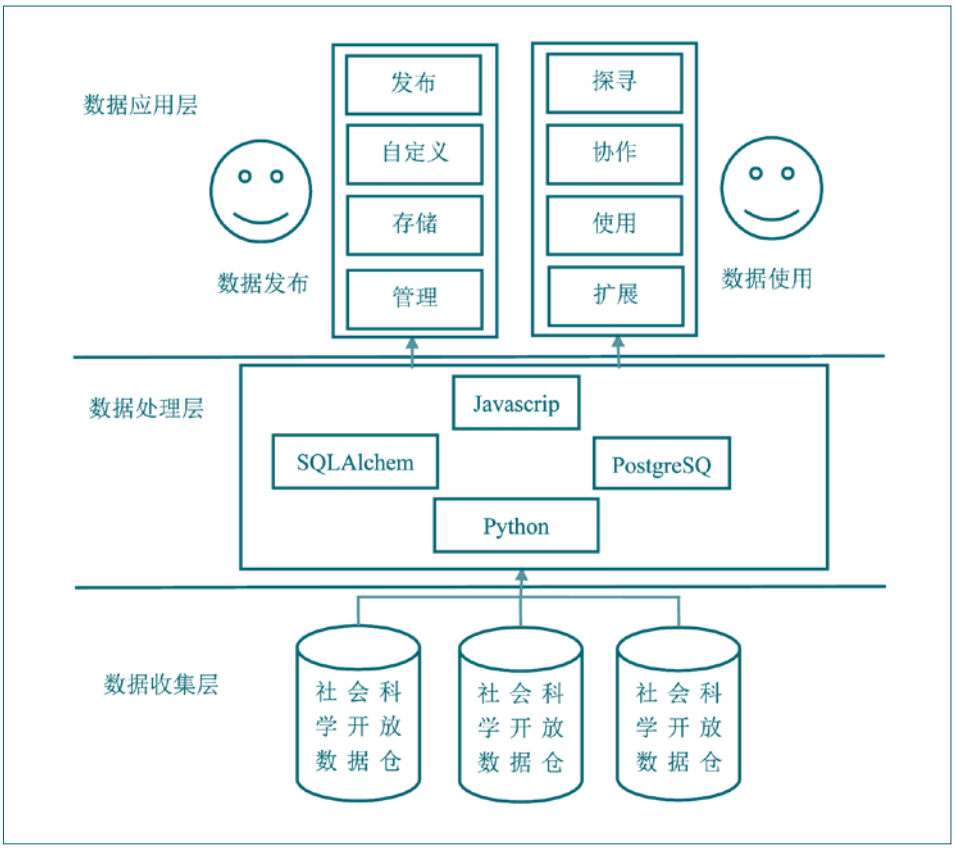

图 1 社会科学开放数据服务平台模型

2.2.1 社会科学开放数据集

社会科学是一门以社会现象为研究对象，解释社会生活本质和发展规律的学科。其复杂性、主观性决定了在研究方法和研究结果上与自然科学的差异。本文选取 OAD 的社会科学开放数据仓储作为数据对象。截至 2014 年 1 月 12 日，OAD 收录 111 个数据仓储，包含社会科学数据仓储 8 个，除因网络问题不能访问“Digital

表 1 OAD 中社会科学开放数据仓储数据基本情况

数据仓储	数据内容类型	数据接口	数据许可协议
Association of Religion Data Archives	调查问卷、投票、报告、地图等	检索 API	需申请
Australian Social Science Data Archive	问卷、投票、统计表格、图像、音频视频、文本资料等	即将支持 OAI-PMH	自定的许可协议
CESSDA Data Portal	文档、调查问卷等	检索 API	需申请
Economic and Social Science Data Service	视频、网络研讨会、调查、长期研究、案例研究和社交媒介等	检索 API	需注册或实名，少部分开放存取
ICPSR	调查问卷、文档、案例研究等	OAI-PMH、检索 API	CC
National Archive of Criminal Justice Data	问卷调查、政府数据、庭审案例、长期调查等	OAI-PMH、检索 API	需注册或实名
Roper Center for Public Opinion Research	调查问卷、投票等	检索 API	需注册

Repositories E-Science Network”外，实际调查数为 7 个。如前文所述，使用 CKAN 必须先获得所需数据和元数据的使用和发布许可。表 1 所示是这 7 个数据仓储的数据类型、数据接口和数据许可协议情况。

从表 1 可以看出，虽然社会科学开放数据并不像其他开放数据那样支持 CC 等开放数据许可协议，从获取程序上看比较麻烦，但是最终可以获得数据集的使用许可。

2.2.2 技术

由于服务平台基于 CKAN 工具，从技术方面与 CKAN 一致。故本节主要介绍 CKAN 采用的主要技术。

总体上，CKAN 在终端使用的是 Python，在前端使用 Javascript。因此也决定了 CKAN 的开发环境。相对而言，CKAN 的安装在 Ubuntu12.04 64 位系统下运行较为方便，下载安装包即可。同时也支持其他操作系统，不过在安装前需要配置好 Python、PostgreSQL、libpq、pip、virtualenv 等工具 [9]。作为数据处理极其重要的部分，它将 Pylons 网络架构和 SQLAlchemy 作为对象关系映射 (ORM) 模型，而数据库引擎使用对象 – 关系型数据管理系统 PostgreSQL，支持数据类型和接口类型丰富，功能齐全且使用自由。就交互访问而言，其搜索通过 Solr 提供接口。从个性化角度来看，CKAN 采用模块化结构，方便使用人员根据自身需求提供额外的功能，如数据收割或数据上传等，提高了平台的可扩展性。另外，在数据使用方面，CKAN 利用内部模型存储不同记录的元数据，并且通过网络界面让用户能浏览和检索这些元数据。为了提高数据重用性，加强开放数据的传播与共享，CKAN 也提供强大的 API，这样第三方应用和服务能在此基础上建立新的平台。需强调的是，CKAN 遵循 Affero GNU GPL v3.0 开放许可。

2.2.3 服务主体及功能

社会科学开放数据服务平台的服务主体既包括数据发布者，也包括数据使用者，而且针对二者需求不同，所实现的功能也有所不同。

(1) 数据发布者

数据发布者包括地方或国家级调查机构、科研院所及个人等其他类型数据提供者。所涉及的功能主要有 4 种：①发布，即通过指导流程或者通过 API 从其他目录导入；②个性化，即添加自己的元数据字段、主题和商标；③存储，即在 CKAN 内或在外部网站上保存数据；④管理，即提供全面控制、可回溯的版本历史、INSPIRE/RDF 支持及用户分析。

(2) 数据使用者

数据使用者包括社会上一切个人、组织，如科研工作者、记者、非政府组织及普通公民。所涉及的功能主要也有 4 种：①探寻，即能通过网络前端或 API 搜索、添加、编辑、描述、标签和分组数据集；②协作，即提供用户资料、通告版、社会网络整合和评论等功能；③使用，即支持元数据和数据 API、数据预览和可视化；④扩展，即提供用于扩展的完整文档。

2.3 关键服务的实现

2.3.1 数据存储

数据存储主要有两种方式，数据上传和从其他数据集导入。数据能够通过 resource_create() 和 resource_update() 两个命令执行 API 功能，上传新数据文档。从其他数据集导入的功能则主要通过数据存储扩展实现。通过扩展可以提供特定的临时数据库来存储 CKAN 中的结构化数据，能从文

3 现名为 UK Data Services(http://ukdataservice.ac.uk/.)

档中抽取数据并保存在数据存储中。如利用数据推送功能，能自动向数据仓储添加数据。例如，利用 Python 内部请求，创建一个新 CKAN 资源并且上传文档的方法如下所示：

```
import requests
requests.post('http://0.0.0.0:5000/api/action/resource_create',
data={"package_id":"my_dataset}",
headers={"X-CKAN-API-Key": "21a47217-6d7b-49c5-88f9-72ebd5a4d4bb"},
files=[('upload', file('/path/to/file/to/upload.csv'))])
```

2.3.2 数据关联

CKAN 能支持关联数据及 RDF，主要通过完整和功能化的 CKAN 数据集框架和关联数据格式的映射实现。在 CKAN 中可以使用都柏林核心元数据 (Dublin Core, DC)、DCAT、VoID 和 SCOVO 等方法描述数据集。目前 DC 和 DCAT 使用较多，以下是其关联映射的例子：

```
<rdf:RDF xmlns:foaf="http://xmlns.com/foaf/0.1/" xmlns:owl="http://www.w3.org/2002/07/owl#"
 xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dcat="http://www.w3.org/ns/dcat#"
 xmlns:dct="http://purl.org/dc/terms/">
<dcat:Dataset rdf:about="http://127.0.0.1:5000/dataset/worldwide-shark-attacks">
<owl:sameAs rdf:resource="urn:uuid:424bdc8c-038d-4b44-8f1d-01227e920b69"></owl:sameAs>
 <dct:description>Shark attacks worldwide</dct:description>
 <dcat:keyword>sharks</dcat:keyword>
 <dcat:keyword>worldwide</dcat:keyword>
 <foaf:homepage rdf:resource="http://127.0.0.1:5000/dataset/worldwide-shark-attacks"></foaf:homepage>
 <rdfs:label>worldwide-shark-attacks</rdfs:label>
 <dct:identifier>worldwide-shark-attacks</dct:identifier>
 <dct:title>Worldwide Shark Attacks</dct:title>
  <dcat:distribution>
   <dcat:Distribution>
    <dcat:accessURL rdf:resource="https://api.scraperwiki.com/api/1.0/datastore/sqlite?format=csv&amp;name=worldwide_shark_attacks&amp;query=select+*+from+`Europe`&amp;apikey="></dcat:accessURL>
   </dcat:Distribution>
  </dcat:distribution>
  <dcat:distribution>
   <dcat:Distribution>
    <dcat:accessURL rdf:resource="https://api.scraperwiki.com/api/1.0/datastore/sqlite?format=csv&amp;name=worldwide_shark_attacks&amp;query=select+*+from+`Australia`&amp;apikey="></dcat:accessURL>
   </dcat:Distribution>
  </dcat:distribution>
  <dct:creator>
   <rdf:Description>
    <foaf:name>Ross</foaf:name>
    <foaf:mbox rdf:resource="mailto:ross.jones@okfn.org"></foaf:mbox>
   </rdf:Description>
  </dct:creator>
  <dct:contributor>
```

```
        <rdf:Description>
          <foaf:name>Ross</foaf:name>
            <foaf:mbox rdf:resource="mailto:ross.
jones@okfn.org"></foaf:mbox>
        </rdf:Description>
      </dct:contributor>
       <dct:rights rdf:resource="http://www.
opendefinition.org/licenses/odc-pddl"></
dct:rights>
      </dcat:Dataset>
    </rdf:RDF>
```

2.3.3 数据查看

CKAN 资源界面能对不同资源类型的数据进行预览。一般而言，由于资源类型不同，其预览实现方法也不同，如图像能够直接植入查看，非结构化或者普通文本文档需要上传到 iframe 中，更为复杂的数据需要使用自定义部件。所以 CKAN 根据资源类型，提供了不同的查看方法。如 XML、JSON、普通文本和 PDF 的查看必须首先在 CKAN 结构文档里将 text_preview 扩展添加到 ckan.plugins。而远程资源除了需要添加 resource_proxy 扩展到 ckan.plugins 外，如需对资源进行代理，还要使用 ckanext.resourceproxy.plugin.get_proxified_resource_url() 命令替换用于上传文件的 URL。

2.3.4 统计扩展

CKAN 的统计扩展能分析发布者的数据库并提供网站的图表分析，包括数据集总数、每周数据校订、最常编辑数据集、热门标签和拥有最多数据集的用户等。在 CKAN 结构文档中将统计数据添加到 ckan.plugins 只需使用 ckan.plugins = stats 命令。其效果如图 2 所示。

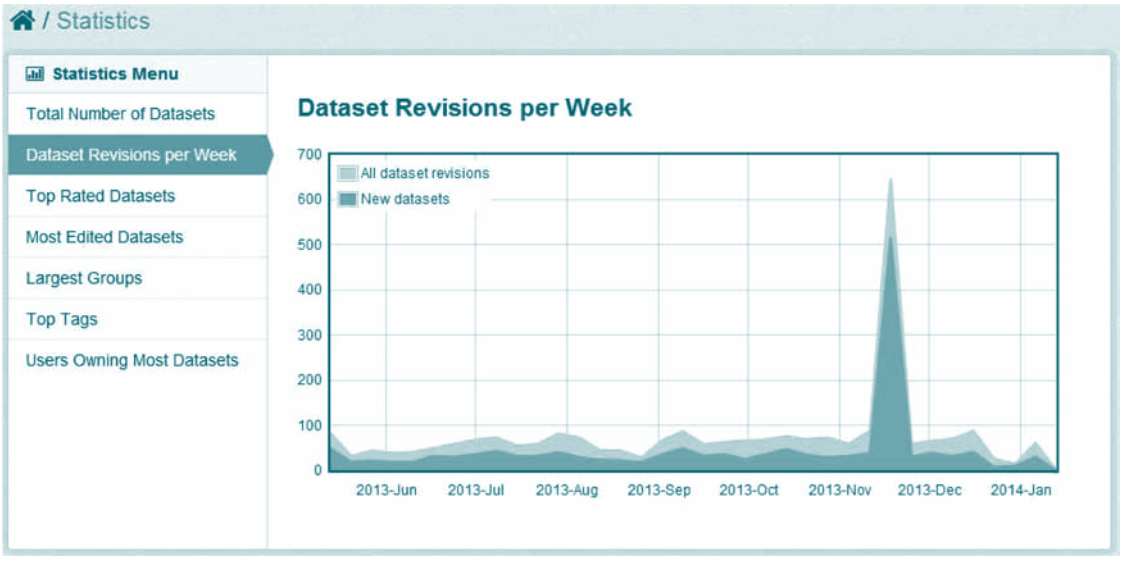

图 2　CKAN 统计功能效果图

3 社会科学开放数据服务平台实例

结合社会科学开放数据以统计调查报告为主的特点，本文选取日本经济产业省的 METI 和欧盟的 PublicData.eu 作为实例，对基于 CKAN 的相关开放数据服务平台功能进行说明。

3.1 METI

Open DATA METI 成立于 2013 年 1 月，用于检验日本经济产业省的开放数据政策 [10]。所包括的数据包括白皮书 (年度报告)、财政情况、统计数据等。提供的服务主要有数据检索、浏览，

链接，BBS 和可视化案例分析等。其数据检索界面，可按资源类型检索，并提供标题、更新时间等排序。每一条记录都从作者、维护者、版本、创建者、频率、发布者、发布日期和标签等方面描述元数据，并链接相关文档数据资源。并且在 METI 平台上还利用关联数据，整合了本地和国家政府机构及海外重要财政金融机构的网站。

METI 比较有特色的服务是 BBS 和可视化案例分析。为了方便数据使用者对数据进行交流，METI 设置了 BBS 功能，支持发表帖子，并且能对帖子进行分类、搜索、评论，还能点击相应的按钮支持或反对，表明自己的态度。具体见面如图 3 所示。

图 3　METI BBS 功能界面

METI 的另一大特色是可视化案例分析。以 2011 财年经济产业省政策成本数据的可视化为例，主要有两种可视化方式，即树图和 Buble 树。图 4(a) 和 (b) 分别显示的是成本数据的树图和 Buble 图，其中图 4(a) 用不同颜色表示不同的构成内容，色块的大小与数值比例大小一致，使人一

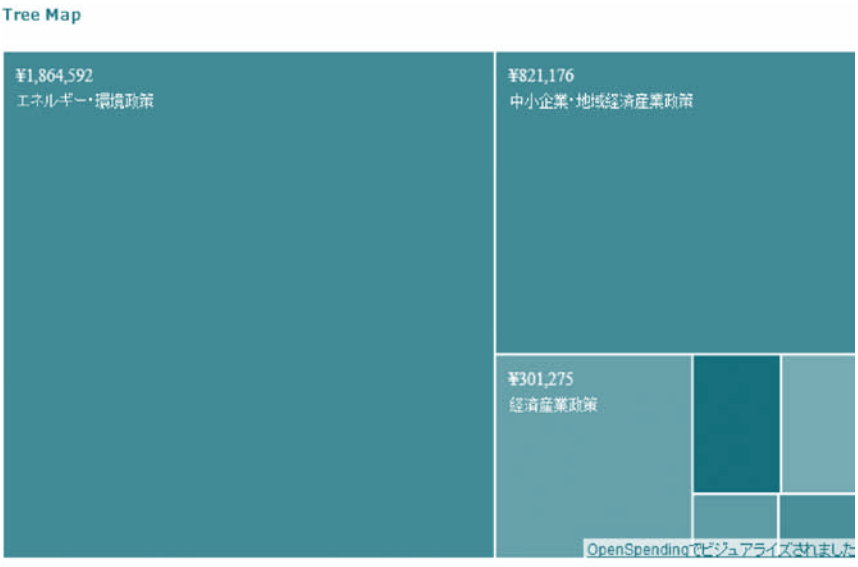

图 4(a)

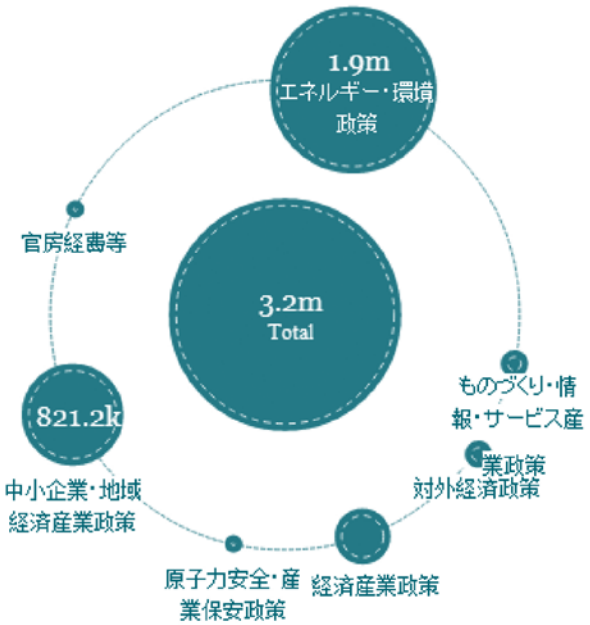

图 4(b)

图 4 METI 数据可视化分析实例

目了然。并且可以点击具体色块查看其成本组成情况，该色块将被分解为更多的小色块，表示更细致的数据情况，直至最小统计单位。图 4(b) 用不同圆圈表示不同统计内容，圆圈大小同数值比例大小一致，也可点击某一圆圈查看该项的具体组成情况，同样用圆圈表示。

3.2 PublicData.eu

PublicData.eu[11] 是一个跨欧洲数据目录和整合模型研究原型系统，成立于 2011 年 6 月，它也是 LOD2[4] 的一部分，运用了 LOD2 关联数据栈技术。通过利用 CKAN 的终端，PublicData.eu 可以检索欧洲包括国家、地区以及官方和团体 25 个目录所含 32538 个数据集的元数据。PublicData.eu 的特点体现在 3 个方面 [12]。首先，系统里包含多语言元数据，用于过滤和多语言描述；其次，它提供应用和想法 (Applications and Ideas) 功能，将在开放数据挑战比赛 (Open Data Challenge competition) 中开发出的相关成果整合在一个目录中，并提供截屏和标签，以便突出在普通和特殊数据集中可用数据的价值；另外，能够通过地图总览整个欧洲的数据，表明哪个国家在开放政府信息方面做的最好，如图 5 所示，颜色越深，表示该国开放数据数量越多，点击该国家后可以直接显示该国所有数据集的检索界面。

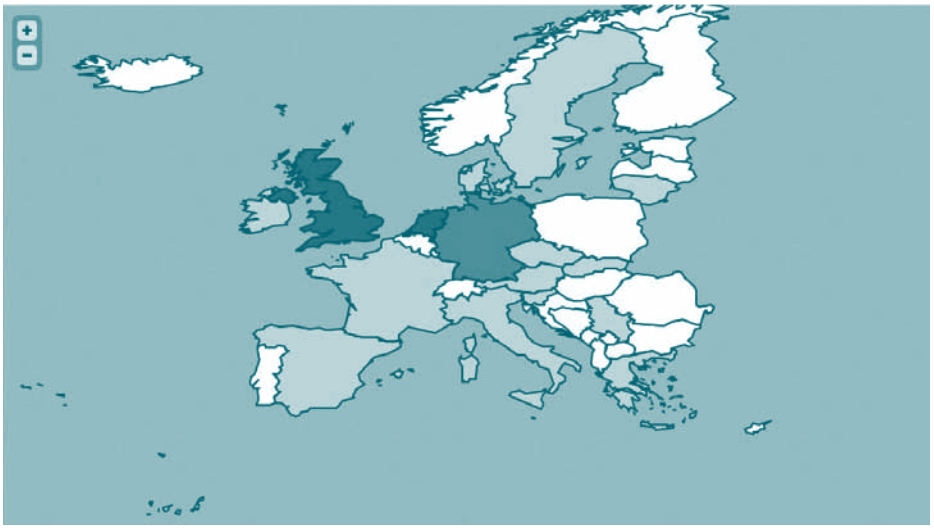

图 5 PublicData.eu 地图浏览界面

4 LOD2 是欧洲的一项关联数据项目，目的是从数据关联中创造新的知识。(http://lod2.eu/Welcome.html.)

4 结语

正如开放知识基金会（Open Knowledge Foundation）所认为，开放数据意味着更好的科学 [13]。开放科学数据反映了科学研究领域的新要求和趋势，开放数据运动方兴未艾，各类开放数据服务平台也在不断涌现，尤其是社会科学领域。如何提高开放科学数据的发现、利用与重用，最大程度发挥其价值是图书馆等信息机构所需应对的课题，为图书馆的发展提出挑战，但同时也为图书馆提供机遇，促进图书馆，特别是研究型图书馆开放创新服务平台的发展 [14]。本文利用成熟的 CKAN 平台开发社会科学开放数据服务平台，提出服务平台要素包括服务平台要素包括数据、技术、服务主体和功能，针对数据发布者和数据利用者提供了不同的功能，并提出了关键服务的实现方法。此外，还对基于 CKAN 的社会科学开放数据成功案例 METI 和 PublicData.eu 进行了功能分析。通过这些实际的案例可以发现，服务平台应该注意数据的可视化，不管是数据呈现还是数据分析方面，这样更为直观。另一方面，也要注重数据使用者的表达需求，为他们提供一些思想展示和交流的平台。但是本文由于操作系统限制，并没有实现所述服务平台模型，不过通过理论分析和实例描述，为数字图书馆在 e-science 环境下的开放服务及平台设计提供参考。可以预见，开放科学数据关联将极大地提高科学生产力，促进未来的科学发展朝着智慧化发展。

参考文献

[1] 黄永文，张建勇，黄金霞等 . 国外开放科学数据研究综述 [J]. 现代图书情报技术，2013(5): 21-27.

[2] OKF Open Science Working Group[EB/OL] [2014-1-6]. http://science.okfn.org/.

[3] Hey T, TanSley S, Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery[EB/OL] [2014-1-12]. http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_complete_lr.pdf.

[4] PARSE. Insight. INSIGHT into issues of Permanent Access to the Records of Science in Europe [EB/OL] [2014-1-12]. http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf.

[5] CKAN[EB/OL] [2014-1-12]. http://ckan.org/.

[6] CKAN. About[EB/OL] [2014-1-12]. http://ckan.org/about/.

[7] CKAN. Feature Tour[EB/OL] [2014-1-12]. http://ckan.org/features/.

[8] CKAN. About CKAN[EB/OL] [2014-1-12]. http://ckan.org/developers/about-ckan/.

[9] Installing CKAN[EB/OL] [2014-1-12]. http://docs.ckan.org/en/latest/installing.html.

[10] Open Data METI[EB/OL] [2014-1-12]. http://datameti.go.jp/?lang=ja.

[11] PublicData.eu[EB/OL] [2014-1-12]. http://publicdata.eu/.

[12] CKAN. Case studies[EB/OL] [2014-1-12]. http://ckan.org/case-studies/publicdata-eu/.

[13] Molloy J C. The Open Knowledge Foundation: Open Data Means Better Science[J/OL] (2011-12-6) PLOS BIOLOGY, [2014-1-12] 2011, 9(12). http://www.plosbiology.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pbio.1001195&representation=PDF.

[14] 张晓林 . 开放获取、开放知识、开放创新推动开放知识服务模式——3O 会聚与研究图书馆范式再转变 [J]. 现代图书情报技术，2013(2):1-10.