DOI:CNKI:11-2746/G2.20120619.1148.001

OpenCSDB:关联数据在科学数据库中的应用研究*

沈志宏　张晓林　黎建辉

摘　要　结合中国科学院“科学数据库”项目,以科学数据关联网络 OpenCSDB 为研究对象,分析关联数据在科学数据库中的应用需求、适用性、实施原则与框架、应用效果以及所面临的新的挑战。研究发现,关联数据机制由于其语义描述能力强、低成本、开放自治的特征,能够很好地满足科学数据库对开放访问机制在包容性、适应性、语义支持以及易推广性的要求,并在“科学数据库”项目内具有较好的实施可行性。尽管还存在着异构数据库的互操作、科学数据溯源、关联开放数据的访问控制、海量数据搜索排序等挑战,关联数据机制仍不失为科学数据开放访问机制的最佳选择。通过“十一五”应用效果证明,科学数据关联网络 OpenCSDB 促进了科学数据的共享,并将在“十二五”期间发挥更大的作用。图 4。参考文献 31。

关键词　关联数据　科学数据　数据记录　数据文件　数据发布

分类号　TP393

OpenCSDB:Application of Linked Data in Scientific Database

Shen Zhihong, Zhang Xiaolin & Li Jianhui

ABSTRACT　This paper introduces OpenCSDB as a solution of applying Linked Data in the Scientific Database, Chinese Academy of Sciences (CSDB) project. In this paper, we discuss the background, applicability, implementation principles, software architecture and application effects, and challenges for OpenCSDB. We found that Linked Data meets the needs of scientific databases, which is a low-cost, inclusive, adaptive, semantics-supported, easy-to-popularize, and open access mechanism, and it is feasible to apply Linked Data in CSDB. Although there are new challenges such as interoperation of heterogeneous data sources, provenance of scientific data, access control in Linked Open Data context, and ranking and search of massive scientific data, Linked Data can still be considered as the best choice for scientific data. As proved by the achievements of OpenCSDB in the “Eleventh Five-Year” program, OpenCSDB will promote the sharing of scientific data and play a greater role in the “Twelfth Five-Year” program. 4 figs. 31refs.

KEY WORDS　Linked Data.　Scientific data.　Data records.　Data files.　Data publishing.

1　问题与挑战

科研人员在长期的科研活动中,通过观测、探测、试验、调查等科学手段积累了大量的科学数据。以中国科学院“科学数据库”项目(Scientific DataBase, Chinese Academy of Sciences,简称 CSDB)[1]为例,“十一五”期间项目所涉及的 2 个参考型数据库、8 个主题数据库、4 个专题数据库、37 个专业数据库,积累了 200TB 以上的科学数据,其中在线数据达到 149.61TB[2]。这样的数据洪流给科学数据的管理与共享带来了巨

* 本文系中国科学院信息化专项“数据应用环境建设与服务”(项目编号:INFO-115-C01)和国家科技基础条件平台建设项目“基础科学数据共享网—理化天文空间生物”课题“标准规范及共享服务平台建设”(项目编号:BSDN2009-17)的研究成果之一。

通讯作者:沈志宏,Email:bluejoe@cnic.cn

大的机遇与挑战[3]。如何帮助科研人员及 e-Science 应用方便快捷地发现并消费这些分布在 Web 各个角落的数据,成为“科学数据库”项目关注的首要问题。

可以看到,在“十一五”期间,“科学数据库”项目的各建库单位遵循统一的界面风格、服务模式、用户认证等标准规范,通过自助开发的方式或者基于可视化数据管理与发布工具 Visual-DB[4]建立起了个体数据库的服务网站,形成了由 51 家数据服务网站组成的科学数据网站群,为最终的用户提供了 Web 化的数据共享。但我们同时认识到,目前这种共享还仅局限在 Web 页面信息的共享,实际上已有越来越多的应用程序开始关注科学数据(数据库描述、数据文件、数据记录等)的本身,而非它的 HTML 展现。因此,“科学数据库”项目需要找到一种面向科学数据(而非 Web 文档)的开放访问机制。根据科学数据以及“科学数据库”项目的特点,“科学数据库”开放访问机制需要具备以下特性:

(1)对多样科学数据的包容性。不同学科的科学数据呈现出不同的存储格式与组织方式,这种多样性要求“科学数据库”开放访问机制具有对不同科学数据格式的包容性。在科学数据库中,很大一部分数据以数据文件(data file)的形态存在于传统的文件系统中,而另外一部分数据则以科学数据记录(data record)的形态存在于不同类型的关系型数据库(RDBMS)中。其中数据文件的目录组织策略,以及数据记录的表格结构设计风格,往往取决于数据管理人员的技术背景和偏好。另一方面,科学数据文件的格式(如 FITS、DICOM、PDB、PDS、HDF、NetCDF、SDXF 格式等)[5],以及科学元数据格式(如 FGDC、ISO 19115、Darwin Core、MCM 等)[6]复杂多样,“科学数据库”开放访问机制需要包容这些描述格式。

(2)对多样服务环境的适应性。“科学数据库”项目涉及中国科学院 50 多家建库单位,各单位在数据管理系统、网络环境、操作系统以及 IT 技术支撑人员的知识背景方面千差万别。以关系型数据为例:关系型数据往往借助于 RDBMS 来实现数据记录的存储、查询和获取服务,而 RDBMS 的数据访问协议往往因产品类型(如:Oracle、MySQL、SQL Server 等)以及版本的不同而存在很大差别,因此“科学数据库”开放访问机制需要对不同的数据库产品以及其他服务环境具有良好的屏蔽性。

(3)对语义及关联的表达能力。科学数据具有天然的语义,科学数据库中的每一个表格(如:birds 等)、每一个属性(如:title 等),以及不同概念之间的关系(如:“斑头雁”与“鸟”之间的从属关系,“CO2”与“二氧化碳”的同一关系,等等)都具有语义。再以青海湖数据库[7]中的某次考察记录为例,除了考察地点和考察对象数目之外,其考察对象(如:斑头雁)希望能够链接到动物数据库的某一条描述记录(包括名称、别名、鸟的照片等),参与该次考察的人员则链接到人员库的某条记录。“科学数据库”开放访问机制需要具有足够强的描述能力,表达出这种关联性,力求给用户以完整的数据展示。

(4)在实施层面上的易推广性。与文献资源不同,科学数据库目前还缺乏一套完善的数据出版、数据权益保护的机制,科学数据库资源往往由各个研究所的课题组分别掌握,如果在共享过程中忽视了数据的权益问题,会为数据共享带来较大的阻力,因此“科学数据库”开放访问机制要求具有非强制性、非集中性。另外该机制还要求足够轻量化、标准化,既能使数据在 Web 环境中的共享更为便利,又不至于造成过高的改造成本,不会影响到已有的 Web 架构及科学数据库格局。

综上所述,由于科学数据的格式和组织方式的多样性、服务环境与水平的差异性、数据的语义与关联性、权益保护的复杂性等特点,“科学数据库”项目需要一种包容的、普适的、支持语义及关联的、非集中式的、低成本的开放访问机制。

2 相关研究及关联数据

在资源描述技术中,XML 一直承担着重要角色。很多科学数据描述格式采用了 XML,如

MathML、CML、SMILES、EML 等。而 RDF(资源描述框架，Resource Description Framework)和 OWL(Web 本体语言，Web Ontology Language)的出现，更是增强了对资源描述的能力。以 RDF 为例，它采用 URI 来标识每一条资源，并采用“主语—谓语—宾语”三元组来表示每一条属性，这样的设计使得 RDF 天生成为 Web 上的描述语言。此外，通过 RDF(S)和 OWL，人们在科学数据领域通过构建本体来完成数据的组织。

在分布式数据访问技术上，与数字图书馆领域的 Z39.50、OAI-PMH 以及 OpenURL 等互操作协议占据主导地位不同，很多科学数据库往往会基于关系型数据库管理系统在有限的范围内(一般是局域网内)开放数据库的 SQL 查询接口，其底层的协议往往因为产品的不同(如：SQL Server 与 MySQL)而不同。部分数据库(如：SQL Server)提供了数据库的 Web Services 功能。然而，由于 Web Services 的重量级和复杂性，越来越多的研究关注于如何基于 HTTP + JSON/XML 来暴露数据库的内容。如：微软、谷歌等主流 IT 公司亦提出自己的开放数据协议(如 Odata[8]、Gdata[9]等)，并提供一系列的 API 和工具支持数据的访问与互操作。类似的努力还包括 SQL over HTTP(如：jtomyx，ChronicDB 等)以及 restSQL 等。在文件型科学数据交换与共享方面，FTP 和 WebDAV 作为标准化的共享协议仍占据着较重要的地位。但是，我们看到，这些接口仅仅适用于文件内容的开放访问，无法实现与其他上下文信息的完美集成。

在这种背景下，关联数据(Linked Data)的提出就具有重要的启示意义。互联网之父 Tim Berners-Lee 在 2006 年 7 月提出了“关联数据”的概念[10]，从技术框架角度来说[11]，关联数据是一组最佳实践的集合，它采用 RDF 数据模型，利用 URI(统一资源标识符)命名数据实体来发布和部署实例数据和类数据，从而可以通过 HTTP 协议揭示并获取这些数据，同时强调数据的相互关联、相互联系以及有益于人机理解的语境信息。

作为关联数据的一项运动，2007 年 5 月，W3C 的关联开放数据项目(Linking Open Data，LOD)正式启动。在 LOD 项目启动后的三年中，越来越多的数据拥有者将他们的数据以关联数据的形式发布到 Web 上。截至 2011 年 9 月，LOD 已收录 295 个数据集，其中包含 310 亿条 RDF 三元组，以及 5 亿条 RDF 链接[12]。

在图书馆领域，瑞典联合目录(LIBRIS)是全球第一个将书目数据发布成关联数据的联合目录。继 LIBRIS 之后，至少有 5 个国际、国家级书目数据、规范数据开放了关联数据服务。2010 年 5 月，W3C 设立了图书馆关联数据孵化器小组[13]，藉此将图书馆领域以及其他领域中那些关注语义网活动中关联数据的人们组织起来，以增强 Web 上图书馆的全球互操作性。另外，一些知名的机构如英国广播公司(BBC)、纽约时报、路透社、百思买等，纷纷采用关联数据发布多媒体、新闻等数据。

在科学数据领域，Linked Data 已经作为一种开放数据的标准化机制渗透到各个学科。如 Linked Life Data[14]整合了 UniPort、PubMed、Entrez Gene 等 20 多个数据源，提供集成检索和浏览服务。NCBO Resource Index[15]应用通过利用 200 多个现有本体中的知识，为用户提供了生物医药资源的浏览功能。Diseasome Map[16]应用整合了不同生命科学的数据源，生成由已知的疾病和基因联系关联的疾病和疾病基因网络。

国内在关联数据方面也在进行着一系列的跟踪与研究，目前这些研究单位主要集中在数字图书馆领域(如：上海图书馆数字图书馆[17-18]、中国科学院国家科学图书馆[19]、中国科学技术信息研究所[20]等)。2010 年 8 月 23 日，上海市普陀区图书馆举行了“2010 图书馆前沿技术论坛：关联数据与书目数据的未来”专题会议。目前看来，关联数据还没有引起国内数据库领域足够的关注，关联数据在国内尚没有形成有影响力的或成熟的应用，基本处于起步探索阶段。

3　科学数据关联网络 OpenCSDB

3.1　关联数据适用于科学数据库

关联数据制订了关于内容对象的四项基本

准则[10]：

①使用 URI 来标识事物；

②使用 HTTP URI 使人们可以访问到这些标识；

③当有人访问到标识时，提供有用的信息；

④尽可能提供关联的 URI，以使人们可以发现更多的事物。

在具体实现关联数据的过程中，准则③往往会具体化为提供资源的 RDF 描述，准则④中的关联 URI 则通过 RDF Link 来体现。基于以上准则我们看出关联数据适用于科学数据库的几点特性：

(1)描述能力上，关联数据采用了 RDF，增强了对不同形态科学数据(包括数据文件的元数据、数据记录、数据值等)语义化的描述能力。另外，关联数据提倡同时发布数据之间的关联(RDF Link)，类似于网页之间的超链接。这种描述机制保障了科学数据内容的完整性，可以充分发挥科学数据的联合效应。通过 RDF Link，把各种完全自治的“数据孤岛”连接起来，形成了一个全面的知识库，从而为上层数据应用(如：数据集成检索、数据融合)提供了丰富的数据源。

(2)实施成本上，关联数据完全架构于目前的 Web 体系之上，这对科学数据库来说几乎意味着零成本升级。“科学数据库”项目在“十五”、“十一五”期间已经构建了足够好的 Web 环境，包括域名体系、Web 服务器、应用服务器等。在关联数据时代，这些环境可以继续使用，数据发布人员要做的就是将原来的 HTML 发布工作转换成数据发布(RDF 格式)工作。当然，在这个过程中，正如网页发布者需要熟悉网站制作工具 Frontpage 或者 DreamWeaver 一样，数据发布者需要借助一些工具，如数据发布工具 D2R、Pubby、Triplify 等。

(3)实施方式上，关联数据机制打消了科学数据所有者的三个顾虑。其一，关联数据更多的是一种发布机制，它定义了一种有别于原始数据的物理存储的中间格式，这一特性可以打消数据所有者对数据流失或者数据从采集到加工的流程会被外界打乱的顾虑。其二，关联数据的关联机制巧妙地规避了一些复杂的数据权益纷争问题，有利于科学数据的健康发展和信息的传播利用。例如：从特定植物中提取的化学成分，通过药物合成进行药物研制，这里面会牵涉到不同的数据库，如植物库、有机化学库、药物库，但它们之间仅仅是通过数据链接关联起来，仍分属于不同的数据拥有者。其三，关联数据秉承了 Web 的 AAA(Anyone can say Anything Anywhere)理念，它提倡每个人发布自己的数据，并鼓励同时加入与外部资源之间的链接(正如 WWW 用户在自己的主页中加入其他网页的链接一样)。在这个开放环境中，每个发布者并没有被强制采用一个集中的数据存储中心，或者一套统一的数据表达模型。

通过以上分析看出，关联数据完全适用于科学数据库，在“科学数据库”项目内具有较好的实施可行性。与 WWW 的蓬勃发展历程类似，一旦开放数据成为一种良好的风气，大大小小的科学数据拥有者就会积极效仿。这种建立在独立自治的本地环境上的开放化机制，有利于科学数据的共享。OpenCSDB 在这样的背景下应运而生，提出面向所有的科学数据库资源构建科学数据关联网络的构想。

3.2 OpenCSDB 构建准则

对照关联数据的基本准则，OpenCSDB 的构建可遵循如下准则：

①在 OpenCSDB 中，每一个数据库都拥有唯一的 URI；

②在 OpenCSDB 中，每一份数据(数据记录或者数据文件)都拥有唯一的 URI；

③当访问 HTTP URI 时，数据服务器会根据请求中“Accept”的参数值分别返回数据的 HTML 页面或者 RDF 描述信息(关联数据的内容协商机制[21])；

④每个数据库会公布所采用的 RDF 词表[22]，以便消费程序理解数据的格式；

⑤除数据本身的属性之外，不同数据之间的语义关联会被揭示出来；

⑥每个数据库开放针对数据资源描述的 SPARQL 查询接口。

作为例子，图 1 标示出青海湖数据库中某次环湖调查记录的 RDF 表示（准则③），主语 http://qinghailake. csdb. cn/wod/resource/record/investigation/889-2 为符合数据记录命名约定的 HTTP URI（准则②），谓语 investigation：location、investigation：object 等由青海湖数据库的 RDF 词表（准则④）所定义。

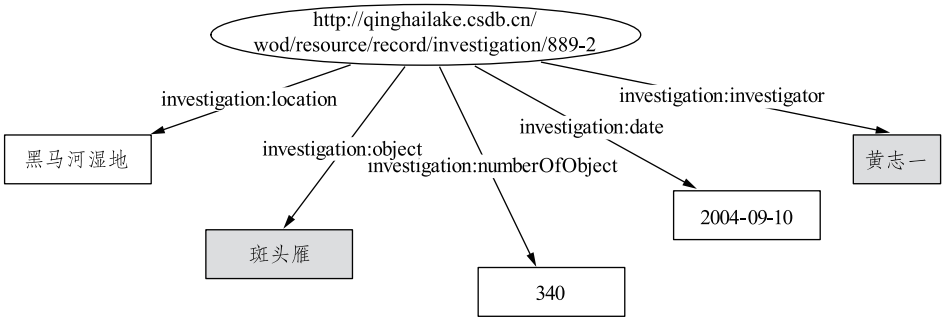

图 1　科学数据关联网络中的 RDF 表达

准则⑤要求揭示数据之间的关联，图 2 标示出一个更为复杂的例子，其中表示了环湖调查记录与调查对象、调查人员的关联关系。通过链接看出，调查对象指向动物数据库中的“斑头雁”（准则②），而调查人员则指向人员库中的“黄志一”（准则②）。

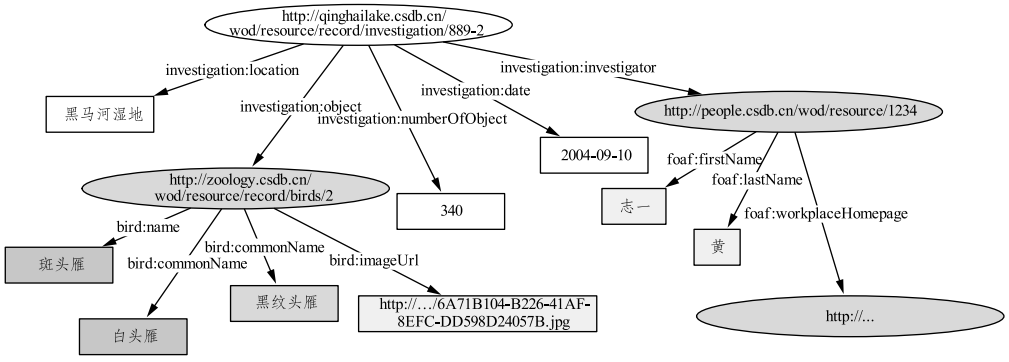

图 2　科学数据关联网络中数据关联的 RDF 表达

OpenCSDB 针对数据记录和数据文件分别定义了 URI 命名约定，数据记录的 URI 设计如下：

< baseURI > /resource/record/ < recordId >

其中：baseURI 表示数据库网站的服务地址（如：http://qinghailake. csdb. cn/wod），recordId 表示记录的完整 ID，它一般由 < tableName > / < itemIdValue > 组成，tableName 表示数据记录来源于哪张表，itemIdValue 表示该条数据的主键值。

对应的，每一个数据文件，其描述信息对应的 URI 设计如下：

< baseURI > /resource/file/ < repositoryName > / < fileId >

其中：repositoryName 代表存储位置的名称，fileId 为每个文件的唯一 ID。

OpenCSDB 提倡发布描述数据模型的 RDF 词表（准则④），同样的，一个数据库的 RDF 词

表具有如下 URI 的名字空间：

<baseURI>/vocab/

在词表设计上，OpenCSDB 提倡采用 Web 上已有的、流行的 RDF 词汇（类名、属性名等），如：采用 FOAF 和 vcard 词表来定义人员信息，采用 dc:title 描述物种名称，等等。

4 OpenCSDB 软件框架

根据不同的功能和用户来分，OpenCSDB 的软件框架主要包括 3 个层次，即构建层、管理层和应用层（见图 3）。

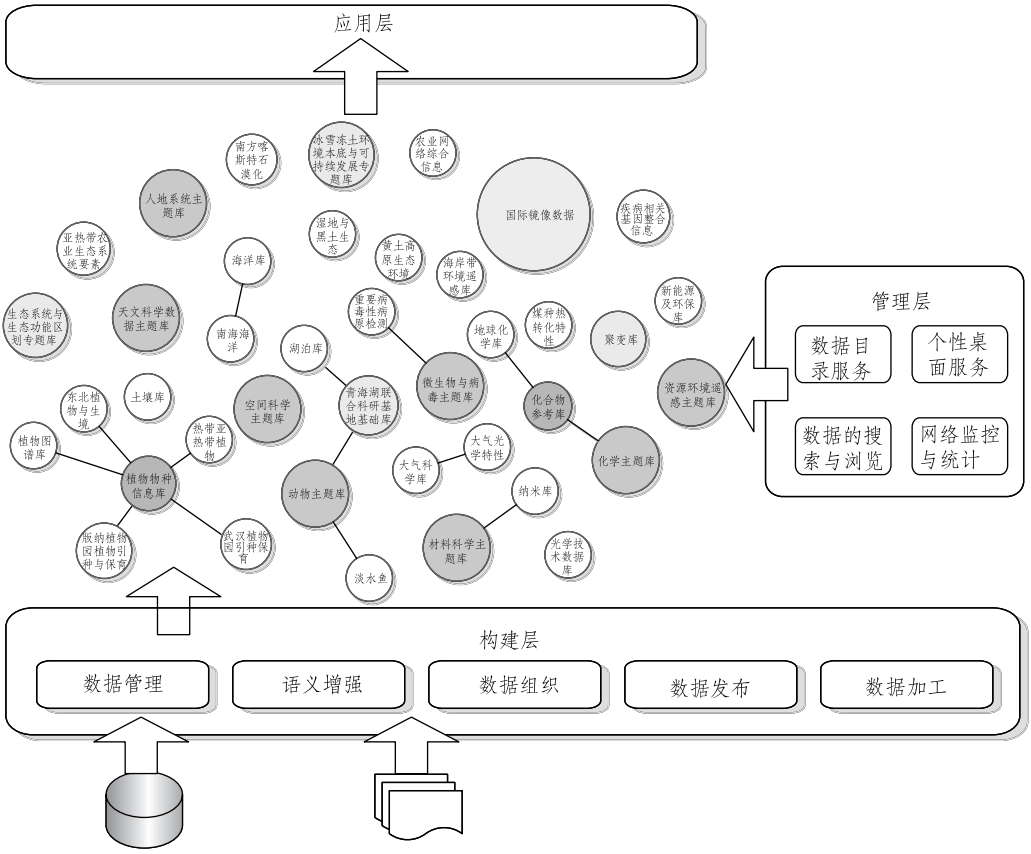

图 3 OpenCSDB 的软件框架

OpenCSDB 构建层主要完成数据网络的构建与编织，主要功能包括：

①数据管理：完成非结构化数据的结构化描述，以及不同类型的数据（数据记录、数据文件）的本地化管理等；

②语义增强：包括数据文件的元数据抽取，基于模板的元数据录入，数据库描述，数据标签（tagging）等；

③数据组织：数据库的分类与目录组织，URI 设计，词表设计等；

④数据发布：D2R 在线映射（E-R 模型到 RDF 数据模型的映射），数据的静态化发布，属性值（如：化学结构式）的 RDF 化，基于规则的关联构建等；

⑤数据加工：基于个体数据库的主题数据库构建，跨数据库的关联发现，科学数据与科技文献的关联等。

除了软件工具、平台及中间之外，构建层还包括一些规范，主要涉及数据库的核心元数据

标准、数据库与数据的 URI 标识机制、学科数据库的数据标准等。

OpenCSDB 管理层主要完成数据网络的管理和服务，提供的服务包括：

①数据目录服务：建立科学数据在线目录，提供科学数据库、数据库的注册与导航目录服务，提供 RDF 词表的注册服务等；

②科学数据的搜索与浏览：提供 OpenCSDB 搜索引擎，提供数据资源的浏览器（基于 B/S 的，或者基于 C/S 的）等；

③个性桌面服务：类似 EndNote 的数据 reference 管理，科学数据评价评论等；

④网络监控与统计：提供科学数据的资源量统计，数据访问与更新情况的统计，科学数据的评估，科学数据库组织可视化等。

云图是管理层科学数据库组织可视化的一种方式。参考 LOD 云图[23]，“十一五”科学数据库形成的云图[24]，如图 4 所示。

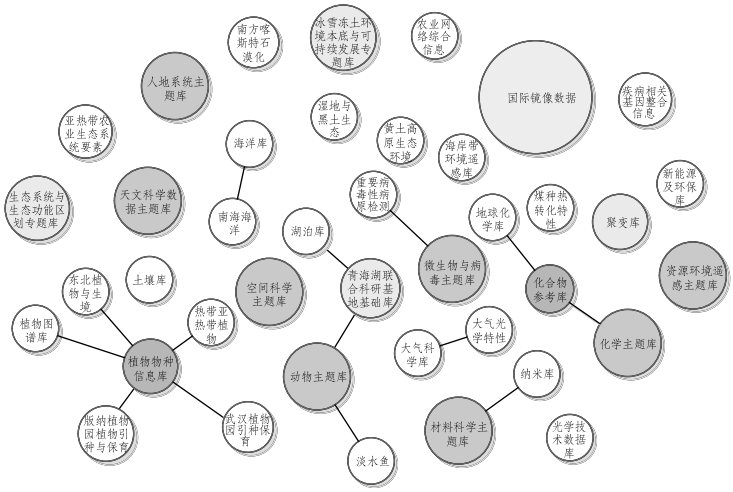

图 4　科学数据关联网络数据库云图

在 OpenCSDB 云图中，每个圆圈代表一个数据库，它们之间的连线标明它们之间的关联（可以看出，云图中关联的集中地是一些参考型数据库，如：植物物种信息库、化合物参考库等）。每个圆圈的颜色、位置和半径都具有其特殊的含义，如：圆圈的半径对应于数据库中数据项的数目，圆圈的颜色对应于数据库的学科领域（天文、空间、物理、生物）等。

最上层即 OpenCSDB 的应用层，该层主要面向特定的需求，开发特定的 e-Science 应用。在该层，OpenCSDB 主要提供数据的访问接口规范和完善的 API。

5　OpenCSDB 的应用情况

“十一五”后期，在“科学数据库”项目中，OpenCSDB 软件框架的雏形已基本建立，部分内容已经得到实现。

规范方面，“科学数据库”项目中已建立“科学数据库核心元数据标准（2.0 版）”[25]，以及科学数据库的“专业库、主题库、参考型库”的组织体系，制定了个体数据库的程序化访问接口，完成了 528 个数据库及子库 URI 的制定，并制定了科学数据库中数据记录和数据文件的 URI 命名规范。

软件方面，通过在可视化数据管理与发布工具 VisualDB 2.0 的基础上嵌入 VDB - WOD 中间件，实现了对数据记录和数据文件的 RDF 化发布；最新版本的 VisualDB 已经增加了语义增强功能，包括元数据抽取、元数据标注、制定或导入分类体系、建立数据目录等；基于 VDB - WOD 接口，开发了科学数据搜索引擎 voovle [26]，

提供了基于关键字的模糊、精确匹配检索功能，并通过关联发现工具 voovle-LDT 根据指定的规则生成新的跨数据库的关联；研发了科学数据资源与服务注册系统 RSR[27]，完成了数据库核心元数据的在线注册和收集，并无缝集成到 voovle；研发了在线资源量访问统计系统 RESSTAT，完成了 VDB－WOD 服务接口的监控统计，对数据项的数目和资源存储量实现在线统计，实现了数据记录和数据文件的数量对比，并给出可视化统计报告。

从应用效果看，“十一五”期间，通过 VisualDB，OpenCSDB 共收录涉及中国科学院 40 多家研究所的 52 个科学数据库，RSR 提供了 528 个数据库及子库的元数据注入服务，voovle 提供了包括 37 家建库单位的 124 个数据库共 564 万条科学数据的语义搜索服务，RESSTAT 也提供了对近 150TB 的数据库在线资源的统计服务。

6 挑战与展望

由于科学数据库自身的质量问题和关联数据的局限性，OpenCSDB 在实施过程中遇到了新的挑战，主要包括以下方面：

（1）关联数据暴露出数据之间的语义一致性问题。科学数据库的异构现象普遍存在于不同来源和不同机构建立的数据库之间，严重影响了科学数据的共享[28]。由于现有科学数据库各单位分头负责建设，前期缺乏相应的规范，而领域知识表达又具有复杂性，无论是元数据标准规范还是元数据参考模型在结合领域具体化时都会有很大难度，再加上缺乏成熟的技术规范和工具，数据管理人员很难建立起成熟的领域概念模型。因此，不同建库单位的科学数据依然采用自治的办法，采用不同的标识系统、不同的分类体系、不同的词表、不同的数据模型，由此带来的同名异义、异名同义、同体异构等问题给科学数据的语义一致性表达带来较大的障碍。

（2）科学数据发布的静止性与科学数据处理的动态性存在着冲突。科学数据的产生、传输、加工与应用是一个持续的、不断迭代的、变化的过程，与其相关的上下文环境（如：处理程序等）也处于不断变化的状态。相反，关联数据化的资源描述往往会被认为是一个静态的切面（一条数据的描述内容不应该被频繁改变，否则应用程序会很容易崩溃），这就暴露出关联数据对科学数据共享的局限性。Sean Bechhofer 等提出关联数据环境下科学数据的 7-Re 标准[29]，即 Reusable、Repurposeable、Repeatable、Reproducible、Replayable、Referenceable 和 Revealable。要想达到 7-Re 标准，“科学数据库”项目还需要更多的努力。

（3）数据访问控制问题。关联数据在用户身份认证、数据访问控制方面没有做更多的标准化工作（至少目前还没有标准的方法可推荐使用），因此数据访问过程中的统一认证和权限控制还必须由应用服务器自己来实现，这样大大限制了不同系统之间互操作性。OpenCSDB 尽管提出了数据网络构建的 6 条准则，但其中提供的 SPARQL 服务往往只针对指定地址开放。如何采取一套适用于关联数据的、标准化的机制实现到数据库、类、属性等不同粒度的访问控制，是一个急需解决的问题。

（4）海量数据搜索与排序问题。OpenCSDB 中包含了海量的科学数据，由于科学数据特有的格式多样、属性丰富、结构化、互相关联的特征，传统的搜索引擎已无法满足 OpenCSDB 的需求。因此需要研究基于关联数据网络的搜索引擎技术，通过 sitemap. xml 协议或者 SPARQL 协议，完成海量数据的抓取，并研究针对科学数据文档（Scienfitic Data Document）的 Ranking 机制，优化科学数据的索引和检索，从而实现科学数据搜索服务的优化。

总的来说，尽管存在着种种困难和挑战，关联数据仍不失为一套标准化的、实用可行的开放访问机制。不同于以往的 Web of Document，基于关联数据机制，我们完全可以在“十二五”期间建立并完善科学数据的数据网络，在这个 Web 中，每一个数据库和每一份数据都可以被开放访问，并返回语义化的内容。纵观关联数据日益蓬勃的应用前景[30]，以及近年来图书出

版界的“科学数据热”(数据出版、数据引用、数据溯源等)[31],我们相信,随着 OpenCSDB 的进一步推广与应用,它将会在科学数据的共享中发挥更大的作用。

参考文献:

[1] 中国科学院数据应用环境[EB/OL]. [2011-12-31]. http://www.csdb.cn(Scientific database platform, Chinese Academy of Sciences[EB/OL]. [2011-12-31]. http://www.csdb.cn.)
[2] 科学数据库资源量在线统计系统[EB/OL]. [2011-12-31]. http://resstat.csdb.cn. (Resource statistics system, CSDB[EB/OL]. [2011-12-31]. http://resstat.csdb.cn.)
[3] Gray J, Liu DT, Nieto-Santisteban M, et al. Scientific data management in the coming decade[J]. ACM SIGMOD Record 2005, 34(4).
[4] Shen Zhihong, Li Jianhui, Li Chengzan, et al. VisualDB: Managing and publishing scientific data on the web [C]//Procedings of International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, 2011.
[5] Day M. DCC Digital Curation Manual: Instalment on Metadata[EB/OL]. [2012-8-15]. http://www.era.lib.ed.ac.uk/handle1842/3321.
[6] Scientific data formats. [EB/OL]. [2012-01-31]. https://nf.apac.edu.au/facilities/software/IDL/docs-6.2/sdf.pdf.
[7] 青海湖联合科研基地数据库[EB/OL]. [2012-01-31]. http://www.qinghailake.csdb.cn. (Basic database of joint research center of Chinese academy of sciences and Qinghai Lake national nature reserve[EB/OL]. [2012-1-31]. http://www.qinghailake.csdb.cn.)
[8] Open data protocol (OData)[EB/OL]. [2011-12-31]. http://www.odata.org.
[9] Google data protocol - Google Code[EB/OL]. [2011-12-31]. http://code.google.com/apis/gdata.
[10] Tim Berners-Lee. Linked data-design issues[EB/OL]. [2009-02-18]. http://www.w3.org/DesignIssues/LinkedData.html.
[11] Linked data FAQ[EB/OL]. http://structureddynamics.com/linked_data.html.
[12] W3C. LinkingOpenData[EB/OL]. [2010-07-06]. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.
[13] W3C library linked data incubator group[EB/OL]. http://www.w3.org/2005/Incubator/lld.
[14] Momtchev V, Peychev D, Primov T, et al. Expanding the pathway and interaction knowledge in linked life data [C]//Proceedings of International Semantic Web Challenge, 2009.
[15] Jonquet C, LePendu P, Falconer S, et al. NCBO resource index: Ontology-based search and mining of biomedical resources[C]//Proceedings of Web Semantics: Science, Services and Agents on the World Wide Web, 2011.
[16] Diseasome Map: Explore the human disease network. Dataset, interactive map and printable poster of gene-disease relationships[EB/OL]. http://diseasome.eu/map.html.
[17] 刘炜. 关联数据:意义及其实现[EB/OL]. [2010-07-06]. http://www.kevenlw.name/archives/1435. (Liu Wei. Linked Data: Meaning and implementation[EB/OL]. [2010-07-06]. http://www.kevenlw.name/archives/1435.)
[18] 刘炜. 数据的万维网[EB/OL]. http://www.kevenlw.name/archives/1185. (Liu Wei. The web of data[EB/OL]. http://www.kevenlw.name/archives/1185.)
[19] 黄永文. 关联数据驱动的 Web 应用研究[J],图书馆杂志,2010(7). (Huang Yongwen. Research on linked data-driven web applications[J], Library Journal, 2010(7).)
[20] 白海燕. 基于关联数据的书目组织深度序化初探[C]// 2010图书馆前沿技术论坛. 上海:2010-08. (Bai Haiyan. Ordering deep of information organization based on linked data[C]//Proceedings of Library Frontier technology Forum 2010. Shanghai; 2010-8.)
[21] Heath T, Hausenblas M, Bizer C, et al. How to publish linked data on the web[C]//Proceedings of LDOW2008.
[22] Best practice recipes for publishing RDF vocabularies [EB/OL]. [2012-01-31]. http://www.w3.org/TR/swbp-vocab-pub.
[23] LOD data set cloud diagram[EB/OL]. [2012-01-31]. http://richard.cyganiak.de/2007/10/lod/lod-

datasets_2010 - 09 - 22. html.
[24] 沈志宏,胡良霖,侯艳飞,等. Linked Data 在科学数据库中的应用探讨[C]//科学数据库与信息技术论文集(第十一集),2012. (Shen Zhihong, Hu Lianglin, Hou Yanfei, et al. Application of linked data in scientific database project: An open discussion [C]//Proceedings of Scientific Database and Information Technology, 2012.)
[25] 科学数据库核心元数据标准2.0[EB/OL]. [2012 - 01 - 03]. http://standards.csdb.cn. (Scientific database core metadata v2.0[EB/OL]. [2012 - 01 - 03]. http://standards.csdb.cn.)
[26] Shen Zhihong, Hou Yanfei, Li Jianhui, et al. Voovle: A linked data search engine for scientific data [C]// Proceedings of International Conference on Fuzzy Systems and Knowledge Discovery, 2012.
[27] 科学数据资源与服务注册系统[EB/OL]. [2012 - 01 - 03]. http://rsr.csdb.cn. (Resources and services registry, CSDB[EB/OL]. [2012 - 01 - 03]. http://rsr.csdb.cn.)
[28] 陈维明. 科学数据的个体识别和跨学科集成[C]//科学数据库与信息技术论文集(第十一集),2012. (Chen Weiming. On the Individual Identification and Interdisciplinary Integration for the Scientific Data[C]// Proceedings of Scientific Database and Information Technology, 2012.)
[29] Bechhofer S, Buchan I, De Roure D, et al. Why linked data is not enough for scientists[J]. Future Generation Computer Systems, 2011.
[30] Hausenblas M. Linked data applications[J]. First Community Draft, DERI, 2009.
[31] De Schutter E. Data publishing and scientific journals: The future of the scientific paper in a world of shared data [J]. Neuroinformatics, 2010, 8(3).

沈志宏 中国科学院计算机网络信息中心高级工程师。
通讯地址:北京海淀区中关村南四街4号。邮编:100190。
张晓林 中国科学院国家科学图书馆馆长,教授,博士生导师。
通讯地址:北京西四环北路33号国家科学图书馆。邮编:100190。
黎建辉 中国科学院计算机网络信息中心,正高级工程师,博士生导师。
通讯地址:北京海淀区中关村南四街4号。邮编:100190。

(收稿日期:2012-04-30;最后修回日期:2012-05-22)