知识资源的语义表示和出版模式研究
——以 Nanopublication 为例 *

吴思竹　李　峰　张智雄

摘　要　海量科学文献和数据为科学研究和交流带来了前所未有的巨大挑战，而传统出版物存在机器可读性差、缺乏知识关联性、不利于新的科学结论发现与传播等不足。本文以概念网络联盟(Concept Web Alliance)提出的纳米出版物模式(Nanopublication)为例，介绍面向大数据处理的新的知识资源语义表示、组织和出版模式，介绍其含义、核心模型、表示形式、构建方法，并从出版、知识组织、知识服务等多角度探讨其应用价值，以期为研究者了解知识资源的语义表达、组织和出版提供参考和帮助。图3。表1。参考文献18。

关键词　知识资源　纳米出版物　语义网　语义出版　语义表示
分类号　G350

Research on Semantic Representation and Publishing Schema of Knowledge Resource: Take Nanopublication as an Example

Wu Sizhu, Li Feng & Zhang Zhixiong

ABSTRACT　Vast amounts of scientific literature and data bring unprecedented challenges to scientific research and communication. Because traditional publications have shortcomings in machine readability and knowledge linkage, they are not beneficial to new scientific statements to be found and disseminated. Taking nanopublication as an example, which is proposed by Concept Web Alliance, this paper mainly introduces new schema of semantic representation, organization and publishing of knowledge resources related to processing of big data, and introduces the definition, core model, representation and realization of nanopublications. Then, it analyzes the signification and application value of nanopublication from different perspectives such as publishing, knowledge organization and knowledge services, so as to provide reference and help for researchers to understand the semantic representation, organization and publishing of knowledge resources. 3 figs. 1tab. 18 refs.

KEY WORDS　Knowledge resources.　Nanopublication.　Semantic Web.　Semantic publishing.　Semantic representation.

数字化技术发展推动科学研究成果的快速增长，海量科学文献和数据为科学研究和交流带来了前所未有的挑战。面对海量的文献数据，人们很难快速了解每篇文献的主题、观点，难以快速发现数据间的知识关联，发现和获得与特定数据相关的资源。语义网时代，借助自然语言处理技术、语义技术，大量科学团体开展了对科学文献进行语义内容标注、语义表示和再组织的工作，如英国皇家化学学会提出的 Prospect 项目[1]，Elsevier 的 Article of the Future 项目[2]、OpenMath 和 Content MathML 项目[3]，PLoS Neglected Tropical Diseases (PLoS NTDs)实施的 Semantic Enriching 计划[4]，LaTex 的

* 本文系国家自然科学基金项目“基于语言网络的文本主题中心度计算方法研究”(项目编号:61075047)的研究成果之一。

通讯作者:吴思竹，Email: wu.sizhu@imicams.ac.cn

SALT 项目[5]等。这些项目利用大量语义、自然语言处理等技术，基于领域本体、维基百科、实体数据库等，对科学文献进行语义标注和组织，构建有别于传统文献的新的知识资源语义表示模式，以减少在知识传播、学术交流过程中的障碍。概念网络联盟(Concept Web Alliance)提出的“Nanopublication”(纳米出版物)是科学文献语义出版的新模式，它从科学知识源头入手，利用计算机作为辅助工具，从文献和数据中抽取科学结论、科学事实或实验结果，建立带有语境、具有语义的表示模式，帮助人们进行科学情报和知识的发现、理解、交流、集成及共享。从知识组织的角度来看，它也是一种新的科学文献及数据的语义表示和组织模式。本文主要介绍纳米出版物模式的含义、核心模型、表示形式、构建方法，并分别从出版、知识组织和知识服务等多角度探讨和分析纳米出版物的研究意义和应用价值，为研究者了解前沿语义知识表示、组织和出版提供参考和帮助。

1 纳米出版物的含义及发展

“Nanopublication”不是表示与“纳米”或“纳米领域”相关的出版物，其中“Nano”表示的是“Smallness，Mini”的含义，它是一种具有科学意义、机器可读的、最小的可出版信息单元[6]。纳米出版物主要以概念或实体作为基本元素，描述科学文献中的科学结论、科学事实或大量实验数据中的实验结果，并提供唯一标识，描述结论的出处、原文作者、纳米出版物的创建者等背景及语境信息，是科学文献在细粒度上的语义表示、组织和出版形式。

纳米出版物的发展时间不长，最早由概念网络联盟于2009 年提出。概念网络联盟作为处理大数据挑战的思想库，由 70 多位从事数据和语义网相关领域研究的专家建立，是一个非营利组织，主要致力于开放环境下大量学术及专业数据间的互操作和出版的研究与实践[7]。他们提出纳米出版物的主要目标是用于解决大数据的语义知识关联。随后，Barend Mons 和 Jan Velterop 基于开放标准提出纳米出版物模型[8]。2010 年 Paul Groth 对纳米出版物的结构进行了剖析。2011 年 Open PHACTS (Open Pharmacological Concept Triple Store) 项目启动，建立了开放药物学空间(Open Pharmacological Space，OPS)[9]，目标是利用不同来源的信息资源集成药物数据，提供基于集成数据的工具和服务，以支持药物发现研究，它采用纳米出版物作为数据通用方式，该项目也是当前纳米出版物的主要应用实践。在 2011 年第一届语义出版工作会议上，Amanda Clare 描述了纳米出版物的创建和使用实例，建模了科学结论、科学事实等知识资源在纳米出版物创建者、使用者和计算机间的动态循环过程，为科研工作者创建了用作实验室电子化记录本的 MediaWiki 语义注释工具，以纳米出版物形式标注和表示，记录科学实验的结论或事实，用于知识的检索、重用和集成[10]。2012 年 Open PHACTS 推出了纳米出版物的指南并构建了 Nanopub. org 网站，提供纳米出版物构建的相关信息和实例。历经不到四年的发展，纳米出版物的研究和应用方向逐渐明确，并通过多机构、团体合作的方式开展了相关尝试和探索。

2 纳米出版物模型及其构建

2.1 纳米出版物核心模型

纳米出版物是科学文献的一种新的出版模式，也是新的语义知识表示和组织模式。Groth 在 2011 年给出了纳米出版物的结构详解[11]，Open PHACTS 项目在纳米出版物指南中进一步确立了纳米出版物的核心模型(见图 1)。

纳米出版物模型由五部分构成，包括结论、结论出处、支持信息、完整性密钥和唯一标识 ID。其中，结论的出处和支持信息提供了结论产生的背景和语境信息。

(1)结论(Assertion)：作者得出的科学事实、实验结果或结论，是最小的思想单元。实验数据中的统计 P 值和其他有效性指标也可记录在结论中。

(2)出处(Provenance)：结论的出处和起源，包括结论是如何形成的，在什么时间得出的，由谁得出的，谁拥有其版权，纳米出版物的创建者、发布者等。

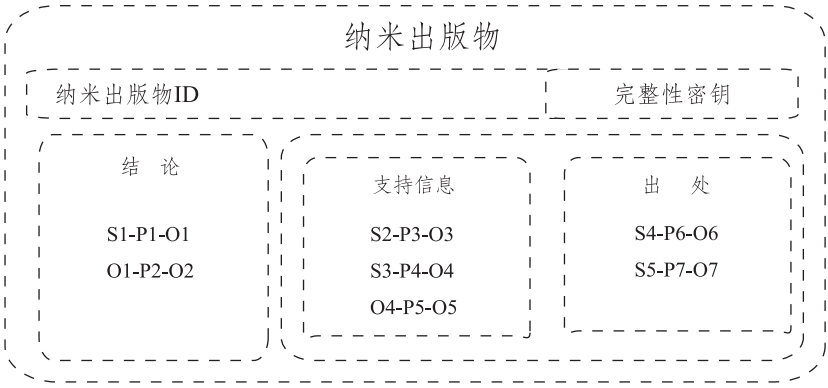

图1 纳米出版物核心模型[12]

(3)支持信息(Supporting):纳米出版物的支持信息提供的是所描述结论的背景或语境信息,如结论 Protein A interact with Protein B,它的支持信息可能包括两种蛋白质具体存在于哪类物种(人或鼠)体内,实验条件是怎么样的,实验地点在哪里等。此外,支持信息还可以记录对所描述结论的引用情况、对结论持有赞同或批判态度的人或观点,同行评议等信息。

(4)完整性密钥(Integrity Key):确保结论作者的身份认证,用户可以通过完整性密钥确定该结论是哪位作者得出的,同时密钥可用于进行纳米出版物的版本控制。

(5)纳米出版物唯一标识 ID:每个纳米出版物都是唯一的,具有唯一标识 URI。

在 Nanopub. org 网站给出的纳米出版物的模型与上述模型有细微区别,主要是对出处和支持信息进行了重新划分,认为出处分为两个部分,包括支持信息和属性。其中,属性(Attribution)用于描述结论的作者、创建时间、创建机构等,而支持信息描述背景信息,二者在描述和表达方式上没有太大区别。

2.2 纳米出版物的表示与生成

纳米出版物以概念为单元,通过 RDF 三元组生成命名图进行形式表达。RDF 是万维网联盟(W3C)提出的一组标记语言的技术标准,能够更丰富地描述和表达网络资源的内容与结构[13]。命名图是对 RDF 的扩展,可以为一个给定的 RDF 图分配 URI。对于某个特定的图,命名图支持跟踪它的出处和上下文定义。纳米出版物被映射为命名图,每个纳米出版物描述一个唯一结论,其构成的最小单元是概念。结论被表示为一个三元组,三元组是由分别表示主语(Subject)、谓语(Predicate)和宾语(Object)的三个无歧义的概念或实体构成一个独立的命名图,如:Acyclovir(主) - > is used to treat (谓) - >herpes zoster(宾)。它具有唯一的、可识别的 URI。结论的出处通过一个或多个三元组表示,三元组中的主语是当前出版物描述结论的命名图的 URI,宾语可以是结论的作者、结论的原始文献、期刊、数据库、专利、创建时间等。谓语描述主语和宾语之间的关系,如 authoredBy、createdBy 等。支持信息同样由一个或多个三元组构成。主语同样是当前描述结论的命名图的 URI,宾语用于描述结论的所属机构、涉及的细胞类型、采用的实验方法、实验条件等,谓语描述主语和宾语间的关系。图片、图表、数据等信息也可以通过添加支持信息链接的方式表示。最简单的纳米出版物可能仅包含两个命名图,一个是结论本身,另一个是该结论的出处。

目前,很多领域,如生物、医学、农业等,构建了大量的规范词表、本体及开放资源,其中大量高质量、无歧义的概念可应用于纳米出版物的构建。Open PHACTS 项目也强烈建议重用现有遵循开放关联数据原则建立的本体、URI 和数据模型。纳米

出版物的创建首先从科学文本中提取关键结论或从科学实验中提取实验结果，通过现有可控、规范词表或领域本体将结论映射为消歧概念，用三元组表示。而后，提取结论相关的语境及出处信息进行概念映射，生成 RDF 命名图。构建中要标注纳米出版物的版本信息。图 2 为一个纳米出版物的实例[12]，使用了 DC(Dublin Core Metadata)、Uniport、OBO(Open Biological and Biomedical Ontologies)、SWAN(Semantic Web Applications in Neuromedicine)等本体和数据库创建。其中结论表示为一个三元组，描述人类 PDE5 基因属于一系列 GO(Gene Ontology)本体的分类。出处信息包括六个三元组，描述纳米出版物当前版本、之前的版本、创建时间、创建者、版权所有者和引用信息。支持信息包括一个三元组，描述基因表达的物种分类。

```
@prefix : <http://www.example.org/mynanopub/>.
@prefix ex: <http://www.example.org/>.
@prefix np: <http://www.nanopub.org/nschema#>.
@prefix dct: <http://purl.org/dc/terms/>.
@prefix go: <http://purl.obolibrary.org/obo/>.
@prefix up: <http://purl.uniprot.org/core/> .
@prefix pav: <http://swan.mindinformatics.org/ontologies/1.2/pav/>
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
{
:nanopub1 np:hasAssertion :G1;
np:hasProvenance :G2;
np:hasSupporting :G3.
:G1 a np:Assertion.
:G2 a np:Provenance.
:G3 a np:Supporting.
}
:G1 {
<http://purl.uniprot.org/uniprot/O76074>
up:classifiedWith go:GO_0000287, go:GO_0005737, go:GO_0007165,
go:GO_0008270, go:GO_0009187, go:GO_0030553.
}
:G2 {
:nanopub1 pav:versionNumber "1.1"
:nanopub1 pav:previousVersion "1.0".
:nanopub1 dct:created "2009-09-03"^^xsd:date.
:nanopub1 dct:creator ex:JohnSmith.
:nanopub1 dct:rightsHolder ex:SomeOrganization.
:nanopub1 up:citation <http://bio2rdf.org/medline:99320215>.
}
:G3 {
:nanopub1 up:organism <http://bio2rdf.org/taxon:9606>.
}
```

图 2　纳米出版物实例

3　多角度分析

3.1　从出版角度看纳米出版物

纳米出版物作为新的语义出版模式，与传统出版物间既存在联系，也存在区别，本文分析如下：

(1)纳米出版物与传统出版物没有本质区别，它是一种新的语义出版形式，而不是传统出版物的替代品，不会影响传统出版物的出版。

(2)传统出版物主要用于人工阅读和理解，计

算机很难识别和读取。而纳米出版物通过为文献或数据库中的科学结论添加语义标注,更适于人、机阅读和理解。

(3)纳米出版物具有比传统出版物更小的粒度,深入概念、实体层面,能够对文献或数据库中的科学事实、科学结论、实验结果进行描述和揭示。它可以从未出版的文献或数据库中提取结论和背景及语境,创建新的出版物,也可以对已出版的文献进行二次加工,生成新的出版模式。

(4)传统出版物是对科学结论完整的、带有全文的论述,具有丰富的论点、论据、上下文背景,包含纳米出版物要表达的信息内容,可作为纳米出版物结论出处的资源链接。纳米出版物是传统出版物外部特征和内容特征的提炼,可溯源,提高了结论来源文献、来源期刊或来源数据库的可见度和被认知度。

(5)传统出版物的结构是静态、线性的,而纳米出版物结构具有动态性,可分割,能集成与其描述结论相关的不同来源、不同类型的资源。

表 1 传统出版物和纳米出版物特点对比

	传统出版物	纳米出版物
语义编码	无	有
内容	全文(详细)	结论 + 情境(精简)
形式	文字表述	概念三元组集合
格式	PDF、Doc、Txt 等	RDF、XML
最小单元	词	概念
结构形态	静态、单一线性	动态连接性
可分割性	无	有
机器可读性	弱	强
链接	无	有
集成多种资源	无	有
适于大数据处理	否	是

3.2 从知识组织角度看纳米出版物

从知识组织角度看,纳米出版物是一种新的知识资源语义表示和组织模式,将信息资源由宏观的期刊、文献层次深化到微观概念的细粒度层次。其对原有文献的外部特征(如作者、出版地、参考文献等)或内容(结论、实验结果、支持结论的相关背景等)进行语义标注、提取和组织,使文本中的重要信息和知识模块化、语义化,使计算机可以快速读取、处理和复用。纳米出版物实现与外部资源的关联,提供结论来源文献或数据库,以及相关的图、表等各种类型的资源链接,将与结论相关的不同来源和不同类型的资源组织、聚合。通过对大量纳米出版物的汇聚,可将科学结论、实验结果、科学假说及其相关数据、来源进行整合集成,组织为大的语义化知识网络,基于此进行分析、挖掘、判断和推理,使研究者获得重要结论的评价、新的科学研究线索并发现新的知识关联。

纳米出版物作为知识资源的表示和组织模式也将推动学术交流模式的变革。科学创新的发展在于对之前科研成果或资源的发现、集成和复用,纳米出版模式对科学文献或数据库中的重要内容进行提炼、重组、语义化,不仅揭示了科学文献中的结论或重要实验结果,还提升了科研文献的内在价值,通过快速的计算机处理,将传统的小范围、封闭的以文献为中心的学术交流方式逐步改变为以共用、共享,以概念、知识为中心的学术交流方式。Mons 描述了它在未来学术交流中的核心作用(见图 3)[14]。每个纳米出版物包括一个 UUID,通过它

可以使纳米出版物进行互操作，使计算机可读，成为一个可访问的独立资源。纳米出版物可用于引文计量、专家评估、数据集成等方面，适于作为知识资源长期保存和语义互操作的基础方式，在学术交流中有助于提高知识的传播速度，缩短学术交流周期，促进科学研究信息化（e-Science）的快速发展。

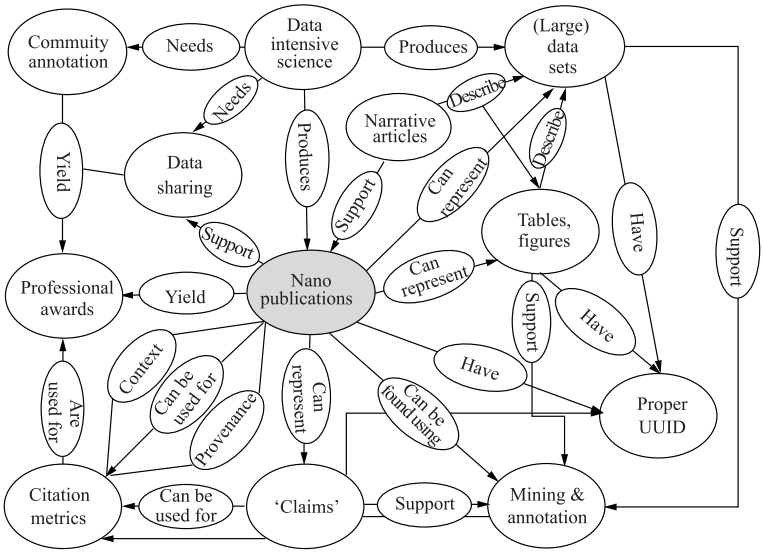

图 3　纳米出版物在学术交流中的核心作用

3.3　从知识服务角度看纳米出版物

语义技术的快速发展为面向知识资源的语义知识组织、检索、存储、传播、展示等知识服务提供方式带来了机遇和挑战。纳米出版物是为了满足面向大规模数据处理的需求而提出的，通过语义标注将更多的科学结论、科学事实或实验结果转换成可发布的知识内容，实现知识资源的自动整合、链接和互操作。在开展知识资源的自动搜集、分析和语义检索、过滤时，可以直接找到提出某一观点、结论或实验数据的相关信息或资源而不是通过整篇文献去获得，粒度更小，层次更深，可以提高知识的共用、共享程度。图书馆及情报研究机构作为知识服务的提供方，应思考如何有效利用这种新模式的，借助语义技术和自然语言处理等相关技术使之能为科研用户提供满足其需求的有效的知识服务。利用纳米出版物对结论、实验结果及相关背景的语义标注，进行研究热点和文献关联资源的推荐服务；通过结合可视化技术，提供文献结构的揭示服务；集成海量科学文献数据资源，开展深度知识语义检索、语义挖掘服务；帮助科研用户开展“战略性阅读”，提供追踪新的科学研究线索、新的科学发展趋势和新的知识发现服务，为科研工作者提供深度知识推理和分析计算服务，提供定量与定性相结合的引文计量和研究成果评价服务，从而促进图书馆和科研情报机构从信息服务到知识服务模式的快速转变，提高面向科研用户的知识服务能力，帮助用户提高科研效率，将面临的挑战转变为适应语义及大数据时代发展的新契机。

4　现有应用研究

目前，已有一些项目利用纳米出版物开展应用研究，Open PHACTS 项目是纳米出版物的主要研究项目，由欧洲的创新药物计划（Innovative Medicines Initiative, IMI）资助，目标是实现一个开放的药物学空间，它将纳米出版物作为用于表示实验数据和科学结论的标准格式[15]。Queralt-Rosinach 利用现有本体及可控词汇集，如语义科学集成本体

(SemantiScience Integrated Ontology,ISO)、美国国家癌症研究所的"国家癌症元词表"(NCI Metathesaurus)、都柏林核心元数据(Dublin Core Metadata,DC)等,将基因疾病数据库 DisGeNET 的数据由关系数据库转换为纳米出版模式的 RDF 数据[16]。Chichester 为蛋白质知识平台 neXtPro 中的关联数据添加语义注释,将其表示为 RDF/XML 格式,使用纳米出版物模式描述关于特定蛋白质异形体转录后修饰的结论及其属性[17]。生成的 neXtProt 关联数据集包含了超过 90,000 个纳米出版物的图结构。通过纳米出版模式可以反映结论相关的不同的数据来源和数据质量,通过它也可以进一步进行数据挖掘、查询检索及进行计算机推理。

EMTO Nanopub 是哲学事实集成的基础建设项目,它将纳米出版模式作为存储和复用早期现代哲学史和相邻学科知识资源的标准格式[18],探索纳米出版物在人文领域的价值和应用,并请对"语义标记语言"不了解或缺乏相关知识的哲学家参与进来。该项目主要关注三种类型的哲学事实,包括历史性事实、哲学性事实和关联性事实,并尝试对事实表达的纳米出版模式进行可视化呈现。

5 结论

纳米出版物本身提供以概念为单元的科学结论,增加了科学生产的深度,而大量纳米出版物通过链接及聚类等方法进行集成,能够形成大的语义知识网,进行不同数据资源的共享和新的知识发现,扩大了科学知识的广度。这种知识资源的新的语义表示和出版方式将改变出版商和科学研究者的数据资源发布、使用、传播方式,也将对用户获取知识的方式产生影响。现有研究将纳米出版物作为一种知识资源的表示和存储模式标准,在较小范围进行应用尝试,还没有开展进一步大规模数据处理的应用实践。纳米出版物等语义表示及出版模式的研究还在发展和探索中,还存在一些问题亟待解决。

(1)虽然已有大量本体、可控词表、语义维基等,但其发展并不平衡。生物医学领域的资源建设较为深入和成熟,而其他领域里,规范、可用的开放资源还不足以支持纳米出版物在广泛的学科领域开展,成为通用的标准。

(2)纳米出版物的生产及发布涉及很多语义网及不同资源使用的相关标准和规范,虽然实施中提倡参考和使用已有的标准,但兼容性和适用性问题仍需进一步思考。

(3)在实际应用中,抽取文献或数据中的结论或实验结果的准确性、选择用于概念或实体映射的本体和可控词表的权威性、结论或数据结果的概念和实体表达的准确性也有必要进一步评估和确定。

参考文献

[1] RSC semantic publishing[EB/OL]. [2013-03-27]. http://www.rsc.org/Publishing/Journals/ProjectProspect/index.asp.

[2] Schemm Y. Experience the article of the future[EB/OL]. [2013-06-26]. http://www.elsevier.com/reviewers/reviewers-update/archive/issue-4/experience-the-article-of-the-future.

[3] OpenMath and MathML[EB/OL]. [2013-03-27]. http://www.openmath.org/projects/esprit/final/node6.htm.

[4] Shotton D, Portwin K, Klyne G, et al. Adventures in semantic publishing: Exemplar semantic enhancements of a research article[J]. PLoS Computational Biology,2009,5(4):1-17.

[5] SALT semantically annotated LaTex[EB/OL]. [2012-12-01]. http://salt.semanticauthoring.org/.

[6] What is a Nanopublication[EB/OL]. [2013-3-27]. http://nanopub.org/wordpress/?page_id=65.

[7] Concept web alliance[EB/OL]. [2013-03-27]. http://www.nbic.nl/about-nbic/affiliated-organisations/cwa/introduction/.

[8] Mons B, Velterop J. Nano-publication in the e-science era[EB/OL]. [2013-03-27]. http://www.surf.nl/sitecol-

lectiondocuments/nano-publication%20-%20mons%20-%20velterop.pdf.

[9] OPS[EB/OL].[2013-03-27]. http://www.openphacts.org/index.php?option=com_content&view=article&id=46&Itemid=53.

[10] Clare A, Croset S, Grabmueller C, et al. Exploring the generation and integration of publishable scientific facts using the concept of Nano-publications[C/OL]// Castro A G, Lange C, Sandhaus E, et al. Proceedings of the 1st Workshop on Semantic Publishing 2011. Hersonissos, Crete, Greece, May 30, 2011: Semantic Publishing 2011[2013-06-27]. http://ceur-ws.org/Vol-721/paper-02.pdf.

[11] Groth P, Gibson A, Stickler P. The anatomy of a Nanopublication[J]. Information Services and Use, 2010, 30(1): 51-56.

[12] The open PHACTS Nanopublication guidelines[EB/OL].[2013-03-27]. http://www.nanopub.org/guidelines/recent.

[13] Resource description framework (RDF)[EB/OL].[2013-03-27]. http://www.w3.org/RDF/.

[14] Mons B, Haagen H V, Chichester C. The value of data[J]. Commentary, Nature Genetics, 2011(43): 281-283.

[15] Williams A J, Harland L, Groth P, et al. Open PHACTS: Semantic interoperability for drug discovery[J/OL]. Drug Discovery Today, 2012, 17(21-22): 1188-1198[2013-3-27]. http://www.sciencedirect.com/science?_ob=MiamiImageURL&_cid=271275&_user=10&_pii=S1359644612001936&_check=y&_origin=article&_zone=toolbar&_coverDate=30-Nov-2012&view=c&originContentFamily=serial&wchp=dGLzVlV-zSkWb&md5=9b50f74c925682e8dd209a6fa0e9f97c&pid=1-s2.0-S1359644612001936-main.pdf.

[16] Queralt-Rosinach N, Furlong L I. DisGeNET: From MySQL to Nanopublication, modelling gene-disease associations for the semantic Web[EB/OL].[2013-03-27]. http://ceur-ws.org/Vol-952/paper_46.pdf.

[17] Chichester C, Karch O, Gaudet P, et al. Converting neXtProt into linked data and Nanopublications[EB/OL].[2013-03-27]. http://www.semantic-web-journal.net/system/files/swj461.pdf.

[18] He?brüggen-Walter S. EMTO Nanopub: An infrastructure for collecting doxographic facts[EB/OL].[2013-06-27]. http://emto.tumblr.com/post/27837095978/emto-nanopub-an-infrastructure-for-collecting.

吴思竹 中国医学科学院医学信息研究所助理研究员，博士。
通讯地址：北京市朝阳区雅宝路3号118室。邮编：100020。

李 峰 北京师范大学图书馆馆员。
通讯地址：北京市新街口外大街19号北京师范大学图书馆。邮编：100875。

张智雄 中国科学院文献情报中心研究馆员。
通讯地址：北京市北四环西路33号。邮编：100190。

（收稿日期：2012-09-24；修回日期：2013-05-03）