面向科学交流的语义出版体系建设研究*

苏静
（陕西师范大学新闻与传播学院，西安 710119）

摘要：作为新兴的数字出版高级形式，语义出版有必要系统梳理正式交流和非正式交流的资源产出形态，并设计语义知识网络建设框架，以便有效满足科学交流进程中的信息诉求。在对比数字出版和语义出版实现流程的基础上，指出语义出版的实现流程应增加出版机构主动式内容采集过程，细化内容编辑与发布环节的增值运作，并在内容消费阶段着重关注用户反馈及其与作者、编辑之间的互动。基于此，根据语义出版的建设要求，提出从资源层、管理层、方法层和服务层4个层面构建适用于现有学术信息环境和科研人员需求的体系框架。其中，语义出版的知识服务效应需在多源化、规模化资源的基础上实现，深层次语义关联与推荐是语义出版体系建设的关键。

关键词：科学交流；语义出版；知识组织；关联数据
中图分类号：G237
DOI：10.3772/j.issn.1673-2286.2018.11.009

20世纪70年代，苏联情报学家А.И.米哈依洛夫[1]便指出，科学交流是科学研究中不可分割的一部分，是科学赖以存在和发展的基本机制。一方面，当前图书、期刊、报纸等传统型科学交流信息载体已无法有效满足科研用户的多层次需求；另一方面，微博、微信、社区、论坛、预印本系统、机构知识库、学科仓储等非正式交流方式应运而生，一定程度上挤压了传统科学交流渠道的生存空间，对原有的闭合式科学交流信息链造成冲击。同时，科研过程中的实验数据、视频音频、评述、讨论、补充性材料等科学资源大量涌现，碎片化内容、微传播内容也影响着传统出版形式的内容价值。语义出版作为新兴的数字出版高级形式，旨在满足科学交流进程中的信息诉求，有必要全面梳理正式交流和非正式交流的资源产出形态，以构建多源化、规模化资源基础上的语义知识网络，减轻科研人员在学术信息检索和利用方面的时间成本和智力成本，以期更好地发挥科学交流系统的整体功能。

由此，本文在对比数字出版和语义出版实现流程的基础上，指出语义出版流程的独特性，有助于语义出版建设主体理解出版流程融合或是再造时的重点环节。同时，根据语义出版的建设要求，提出适用于现有学术信息环境和科研人员需求的体系框架，具体从资源层、管理层、方法层和服务层4个层面构建，以保障语义出版体系有效、稳定和可持续地运行。

1 语义出版的实现流程

1.1 数字出版流程

传统出版是一种线型的内容资源生产、编校和传播的过程，是以著作权的权益让渡为基础，包括选题策划、组稿审稿、编辑加工、批量复制和发行等环节；编辑人员可根据策划活动结果选择合适的作者和作品，并将最终知识成果及其文化属性固化于图书、期刊等载体以进行交流和传承[2]。由此可见，传统出版流程是由选题、组稿、编辑、校对、装帧设计、出版发行等一系列环节组成的完整流程，其中，选题的策划、论证和组稿质量直接影响出版产品的出版效益，也是传统出版流程的侧重点；内容层面的描述局限于题名、出版者、出版时间、字数、定价等外部特征的揭示。

早期的数字出版流程是在传统出版的基础上，利用数字技术对已有出版内容资源进行数字化加工和传

*本研究得到国家社会科学基金重点项目“基于知识组织的图书馆资源发现服务体系研究”（编号：17ATQ002）资助。

播的过程，具体而言，是通过对数字内容产品的分类及编辑加工，进一步规范从内容转档、内容采编、内容管理到内容开发的数字化出版制作流程。现阶段的数字出版主要是基于XML（可扩展标记语言）解决版式和流式文件的转换，产生HTML、PDF、FLASH、EPUB、Umd等电子服务格式[3]，主要涵盖内容创作、内容编辑与发布、内容消费3个环节。其中，内容编辑与发布环节包括编辑层面的协同管理、内容标引与审校、版权服务管理的功能（见图1）。可知，数字出版流程侧重以计算机或是类似设备对出版内容资源的数字化，仍然属于一种先生产后销售的线型出版模式。

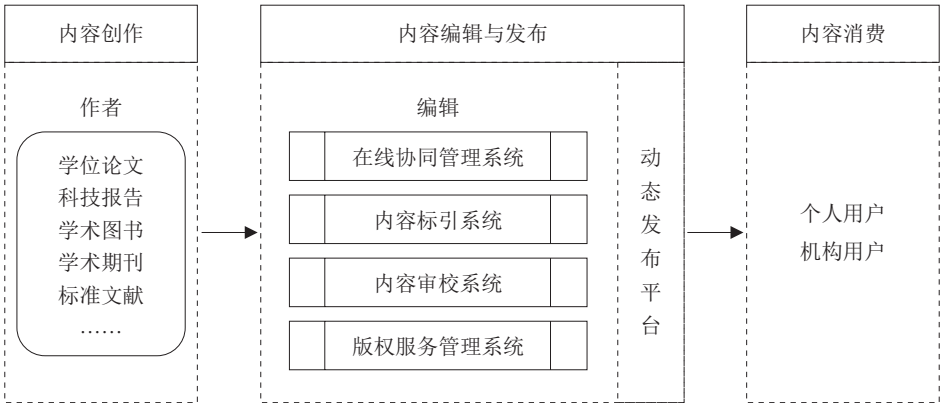

图1　数字出版实现流程

目前，数字出版的内容编辑加工阶段已经基本实现外部内容特征的揭示、章节结构、图表、视频、公式等内部内容的模块化抽取，以及基于字符串匹配的文本标注和关联，其结果是大规模数据集的集成整合和依据一定属性对内容进行分门别类地展示。如检索某一主题的相关文献，页面显示结果除文献列表外，还会提供出版日期、学科分类、语种、作者、机构、基金、文献载体、文献来源等内容特征的分类选项。但是，依据分类选项而被划分的下一级数据结果仍然存在数据规模较大、数据质量参差不齐、数据相关性模糊的问题。针对于此，科研用户往往需要基于自身知识结构，通过人工判断和逐层点击跳转至目标信息界面，这不仅无法实现节约用户时间的目的，反而会极大地干扰和分散用户的思维逻辑。

究其原因，现有的数字出版实现流程大致面临3个关键问题：①缺乏内容资源的多源关联，表现为内容载体的形态较为局限，仅包括期刊、图书、学位论文等，用户无法获取专利、科技报告、标准等相关主题文献，也无法访问作者的研究工具和数据；②缺乏内容资源的深度加工，仅对内容资源进行字面匹配，难以洞察和挖掘隐性语义关系；③缺乏内容资源的语义推荐，未实现在某一类别内按照一定学科/领域/自定义规则对内容相关性和内容质量进行评价、筛选和排序的高级功能。

1.2　语义出版流程

相较于数字出版流程，语义出版的实现流程有必要围绕科研用户需求和行为特征，以加强语境理解、提高阅读效率为目标，增加出版机构主动式内容采集过程，细化内容编辑与发布环节的增值运作，并在内容消费阶段着重关注用户反馈及其与作者、编辑之间的互动（见图2）。换言之，语义出版是以内容和用户需求为核心的出版行为，属于先生产采集，再加工重组，后销售的双向、互动型出版模式。

具体来看，一是内容采集资源和内容创作资源构成语义出版的资源基础，使其不只局限于拥有版权属性的本地资源仓储，扩大了数字出版的关联对象，为知识化服务奠定了坚实的数字资源基础。二是语义出版的实现流程创新了学术资源组织与发布方式，使其更加侧重于内容的结构化加工、语义化关联、知识化挖掘和动态化重组与发布，包括从海量内容资源中抽取知识单元，并进行语义化标引、关联、分析和评价，进而形成机器可读的规范化表示方式，以可视化、交互式的在线表现形式对外呈现出来，充分盘活出版内容资源的知识属性，提升内容资源的检索、聚类和应用的能力。三是在现有学术环境驱动下的语义出版，不仅仅是“生产-传播-消费”的线性过程，编辑、作者和读者的交互频率明显增强。借助多方主体共同完成的知识选择、复制和

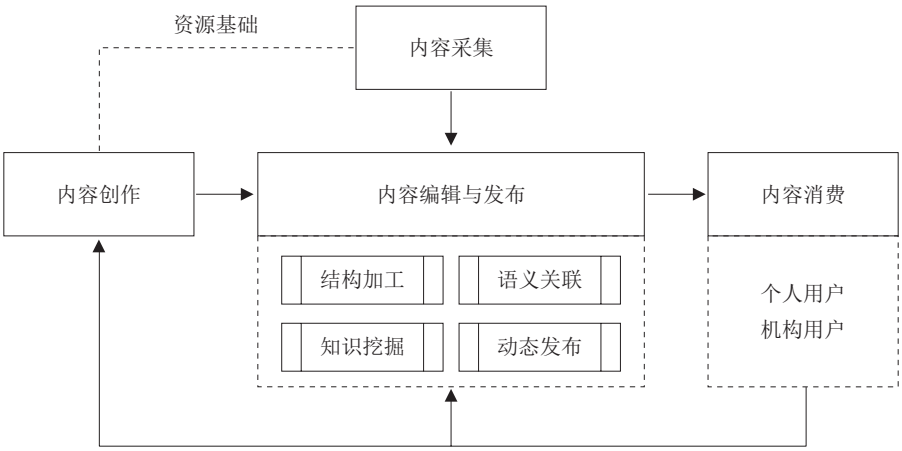

图2 语义出版实现流程

消费环节，编辑和作者能够及时收到用户反馈信息，编辑可以快速调整选题方案和设计知识服务产品，作者在与目标用户的动态交互中深化研究重难点和扩展研究思维，促使语义出版的知识价值呈现螺旋式上升。同时，传统式规模化批量制作的知识生产方式俨然已经不适用于现有时代，需要与大众参与方式相融合，如让用户参与语义标注等环节的构建与更新。

需要注意的是，出版不再仅是出版社的行为，所有从事信息产业的内容提供者都可成为出版者[4]。语义出版的建设主体既包括传统出版企业，又包括信息服务机构、信息集成商和信息技术提供商等。其中，传统出版企业自身拥有极为丰富和具有特定特征的内容资源优势，信息技术提供商可通过语义技术开发并优化内容资源的采集、处理和用户服务方式，图书馆等信息服务机构可利用用户使用日志分析用户行为特征和完善用户服务手段，信息集成商则在内容资源和用户资源方面占据优势。

2 面向科学交流的语义出版体系建设要求和框架设计

2.1 基本要求

目前，以出版机构的数据资源来看，可以被称为小规模、零散式、异构化数据。其中，小规模是指数据存量不大、增量不大、实时性不强；零散式是指数据来源没有标准化通道，数据存储和管理则散布在不同系统和部门；异构化是指数据存储方式、管理方式、数据结构、语义表示和知识内容本身等问题的不统一。因此，语义出版的体系框架，应按照“统一数据标准、统一业务流程、统一信息服务、统一组织工具”的要求构建，利用媒介融合、立足优质内容、基于用户定位，实质性推动内容生产向实时生产、数据化生产、用户参与生产的方向转变，形成在文献高度增值利用和知识发现驱动下的语义出版内容传播系统。

2.2 体系框架

语义出版体系框架是基础性、工程化的建设方案，可适用于一篇论文或一本图书，但要形成语义出版的知识服务效应，需要在资源规模化、多源化的集成基础上实现。其中，深层次语义关联与推荐是语义出版体系建设的关键。语义出版的语义关联与推荐，对内需要提升知识组织能力，对外需要知识呈现和管理能力，这既包括对语义出版对象集的质量评价、遴选、确定和采集，也要设计和应用统一的标准和知识体系对语义出版对象集进行知识抽取、知识表示和知识关联，完成语义出版内容资源的标引、管理、整合和展现，以智能技术实现知识资源的动态构建与扩展，还需提供对知识关联结果进行深层次识别、评价、筛选和排序的解决方案，并且基于用户行为和自身需求，以软件系统为媒介提供内容交互性强、精准度高的语义出版产品及其知识服务，以加强知识的易获得性和可利用性。语义出版的体系框架见图3。

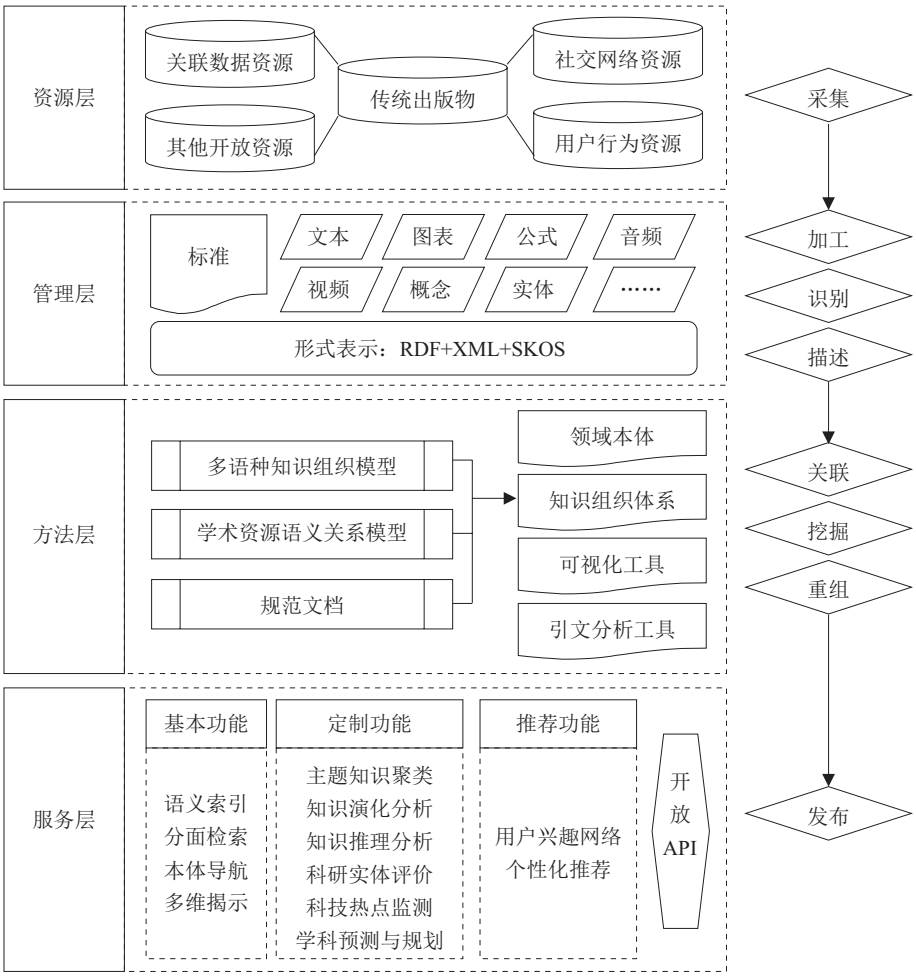

图3　面向科学交流的语义出版体系框架

3　面向科学交流的语义出版体系内容解析

3.1　资源层

语义出版具有高度关联性，打破了文字和图片、表格、数据、工具、软件的桎梏，消解了学术期刊、图书、科技报告、会议论文、光盘等信息载体的形式化。为适应“数据密集型科学研究”，语义出版需将资源对象扩展到视频、音频、实验数据等原始数据，带来传统出版形态与开放出版、社交出版[5]等新兴出版形态的在线融合，帮助科研用户对知识的相关性、影响力、质量和可信性做出较为准确的判断。因此，语义出版系统架构的基础层需要着重围绕传统出版物资源，联合采集和存储关联数据资源、社交网络资源、用户行为数据和其他开放资源，共同推动数字出版向高级的语义出版及其知识服务转型。

3.1.1　传统出版物

传统出版物包括图书、期刊、科技报告、地方志、工具书、标准、法律法规、专利、统计年鉴等，属于精英生产内容，即具有严格的内容评价与筛选机制，数据结构完整统一，内容表达符合语法规范，基本不存在异构和混乱数据，具有较高的知识价值含量。因此，利用已有的存量出版资源是建设领域本体、开发专业知识库的基础，是出版机构向知识服务提供商转型的发力点。

3.1.2　关联数据资源

2006年，万维网联盟提出关联数据的概念，强调数据的相互关联和便于人机理解的语境信息，强调构建具有结构化和富含语义的数据网络，强调在语义网络发布、共享、链接各种数据集、信息及知识，其主要作用是为本地数据建立外部关联，形成多种数据混搭建立的新数据

集，以助于语义挖掘和推理实现知识的发现[6-7]。具体而言，关联数据是指一种在Web上以结构化数据发布的推荐形式，其基本原则包括：一是使用URI作为事物的名称；二是以HTTP/URI协议请求获取事物；三是当有人查找URI时，需使用推荐的标准（RDF、SPARQL）提供有用信息；四是应包含其他事物的URI链接，以便关联发现更多事物[8-9]。

2006年起，多领域的参与者将数据发布为关联数据并相互关联，形成关联数据集云图（Linking Open Data）。据笔者统计，截至2018年10月31日，关联数据类型包括跨学科（cross domain）、地理科学（geography）、政府数据（government）、生命科学（life sciences）、语言学（linguistics）、媒体（media）、出版物（publications）、社交媒体（social networking）和用户生成（user generated）9大类，共有1 229个关联数据集合被发布[10]。其中，出版物关联数据集合数量为147条，仅占关联数据总数的11.96%，并且与其他关联数据集合的入链数和出链数的最高值为32和55，属于中等偏下水平，说明出版物在关联数据发布方面还有较大的提升和发展空间。此外，按照被链接次数统计，DBpedia、NCI Thesaurus（国家癌症研究所词表库）、SNOMED Clinical Terms（系统临床医学术语集）、Medical Subject Headings（医学主题词表）、NIFSTD（神经科学信息框架标准本体）等已成为各类关联数据集相连的基础资源。

关联数据具有较强的数据整合和重用功能，可以有效实现出版内容资源组织与语义网的融合，在未来的知识服务中必然发挥重要作用，而我国出版业界对于关联数据建设方面的重视程度不足，这将限制语义出版数据集规模效应的发挥。因此，有必要在语义出版体系框架内引入关联数据的概念，一方面，可通过将MARC、本地XML格式的原数据结构转换为关联数据，采用开放的标准，结合部分扩展元素（如schema.org、FOAF、DC等），形成通用结构的数据并以stylesheet输出，以提升本地资源的对外显示度和被链接的可能性；另一方面，着重关注和遴选适用于本地资源的关联数据集合，尤其是已被业界认可的、链接度较高的关联数据集合，以扩充语义出版知识资源的语义容量。

3.1.3 社交数据

2016年6月，美国陆军部发布《2016—2045年新兴科技趋势——领先预测综合报告》[11]，认为在未来的30年内，社交媒体将会给人们带来可以创造出各自微型文化群体的能力。目前，从中国数字内容产业的整体发展趋势来看[12]，在内容创造、内容互动、内容分享和内容消费各个阶段的社交用户数量均呈现规模化特征，分别为1.4亿、2.7亿、3.7亿和5.1亿，社交用户生成的作品数量更是加速增长。由此，依托社交媒体形成的去中心化的学术网络结构，会对基于正式交流渠道的传统学术交流体系产生深远影响，需要引入“赞”“评论”等即时性较强的社交媒体数据，以补充传统出版内容生产流程复杂、周期过长而产生的非实时性评价数据。

3.1.4 其他开放资源

其他开放资源主要包括数据仓储、政府统计数据与新闻公告、研究报告等。目前，学科常用的数据仓储包括GenBank（基因数据）、Dryad（综合学科）、PANGAEA（地球科学）、Knowledge Network for Biocomplexity（KNB）（生态和环境科学）、National Biological Information infrastructure（生物科学）、DataBasin（空间科学）、DataONE（跨学科）、PaleoBiology Database（古生物科学）、Protein Data Bank（PDB）、the Universal Protein Resource（UniProt）（序列和注释数据）、INSPIRE（空间科学）。此外，开放知识基金会（Open Knowledge Foundation）是2004年在英国剑桥成立的一家非营利性机构。它专注于在数字时代推进各种形式的开放数据和开放内容，旗下的旗舰级开源软件项目CKAN，是世界顶级的开源数据门户解决方案，已经被美国政府数据开放门户网站（data.gov）、英国政府数据开放门户网站（data.gov.uk）、欧盟开放数据平台等诸多国家/组织的政府机构用于建设数据门户。该平台也可作为开放资源的关联对象之一。以我国而言，国家科技管理信息系统、国家自然科学网、国家社会科学网、中国科学院的科学数据共享平台也可成为语义出版的重要数据来源。

3.2 管理层

管理层的核心是基于标准的规范化加工、识别、描述，以实现文本、图表、公式、音频、视频、概念、实体等对象的抽取与结构化集成，最终以“RDF+XML+SKOS”进行语义表示。以形成结构化、数据化、

语义化结果为目标，对原始内容资源进行细颗粒度加工工作，支持知识单元加工与管理过程中通用标准的应用，完成结构化、半结构化与非结构化数据、文档的存储，形成多个XML数据库、关系数据库等，为下一步构建专业词库体系、专业内容分类体系、知识关联网状体系等创新型知识网络奠定基础。其中，知识单元是管理层的核心概念，它包括两个方面：一是文章、篇、章、节、段落等；二是如概念、原理、图表、数据的知识元，有助于后期通过知识元的语义逻辑关系构建知识网络[13]。知识元具有极好的扩展性，在分类和索引数据中较为有用，由知识元链接形成的知识网络，一方面通过知识元间的隐含逻辑关系和语义关联，可以较好地揭示概念对象间复杂丰富的语义关系；另一方面借助与更多知识领域达成的良好互动，能够及时展现某一学科领域中信息吸收与知识扩散的发展演变，有利于潜在知识的发现和深度挖掘。因此，厘清知识元关系，加强知识元解释至关重要，这就要求语义出版体系框架内的管理层通过对数字内容进行多元化资源管理，实现资源碎片化加工、标引标注、主题词创建等技术处理，对知识单元的修改、标引、超链、备注、标签、关联等进行专业化编辑加工，对文字、图标、公式、表格进行矢量化、深层次、准确地标引，从而确保信息提取的精确性，满足分类存储和数据挖掘的需要。

3.3 方法层

方法层的主要任务是通过领域（行业）本体的构建，借助多语种知识组织模型、学术资源语义关系模型、规范文档等类型的知识组织体系，以及可视化分析工具和引文分析工具，实现知识单元的自动关联、挖掘与动态重组。

同时，方法层又可以理解为语义层、逻辑层和评价层，具有知识计算、知识地图和知识评价的功能。具体来看，一是根据知识组织体系和领域本体完成知识库和知识网络的构建，达到语义唯一性、互操作、关联揭示和富含一定逻辑推理关系的目标[14]，揭示结果可以是面向某一项目、机构、地区、学科、人物、主题的知识系统；二是根据关联权重进行推荐计算、评价与智能排序，含筛选功能，有助于进行个性化推荐；三是根据用户定制需求完成浏览界面互动设计和渲染，主要完成文章、段落、图表、数据、附件资料等对象的交互性设计，达到信息可视化和交互化的目的。

在此过程中，语义出版建设主体需要持续研究URI、researchID等规范标识应用于知识单元的语义关联与映射方式，研究从出版内容数据中挖掘关联知识、分类知识、聚类知识、预测知识、时间序列知识等的知识发现理论和方法，研究语义出版内容包含的文本、声音、图像、视频等富媒体数据的组织方法、技术和工具，研究知识的标引、关联和重组技术与工具，研究知识服务创新模式和知识应用方法。

3.4 服务层

服务层是用来对语义出版服务产品进行功能展示和在线发布的途径，具有用户服务、管理和知识产权保护的基础功能。在服务功能方面，一是需具有开放的理念与平台嵌入式接口，强化数据开放服务模块，可提供OAI-PMH接口，支持第三方在遵循使用许可协议，使语义出版服务产品能在知识产权保护条件下，可以自由灵活地嵌入多种信息发布系统或应用环境，同时，支持多种属性内容资源标识符标准的注册、登记，以满足数据共享、集成与融汇的需求；二是具有语义索引、分面检索、本体导航、多维语义揭示的基本功能，如通过概念级别的扩检与缩检，实现不同颗粒度的智能查询；三是具有主题知识聚类、知识演化分析、知识推理分析、科研实体评价、科技热点监测、学科预测与规划的定制功能，也可为期刊编辑部识别核心作者和潜在作者群，为研究人员识别科研合作对象，了解同类别的高被引核心期刊，为研究机构识别科研合作对象，更好了解同类别机构，为管理部门遴选专家，进行科研评估参考；四是关注用户反馈，借助COUNTER statistics和CrossRef等系统开展基于用户兴趣与行为的个性化推荐服务，以体现服务方式的差异性。

同时，服务层应加大关注协同创新发展，一方面优化用户参与和反馈机制，允许用户添加语义标注的行为，及时收集用户知识需求重点；另一方面，面向研究人员、工程技术人员及管理人员构建学术研究、技术革新、产品发明、决策支持等的协同研究和创新平台。此外，语义出版的版权环境，也是促进语义出版健康、稳定发展的必要基础，需要加强版权保护技术处理和其他版权保护形式，加快技术创新和标准制定，为版权保护提供有效的技术手段，达到进行数字内容资源版权保护的目的，切实保障著作权人合法权益和出版机构的合法利益。

4 结语

面向科学交流的语义出版体系建设，应围绕科研用户行为和需求，通过资源购买、共享协议签订、数据交换等方式拥有数据的知识产权，提高数据采集、存储、管理和运用能力，支持数字文献资源的战略保存管理与二次开发利用，加强出版内容资源、产品主题知识库、用户数据库的建设，服务功能层面则应支持可视化分析、排序、智能推荐、分享等。同时，提供开放性的API数据接口，保障数据资源在一定范围内的互通互享。在具体操作中，针对多源基础资源采集与整合的难题，语义出版建设主体须从整体实际情况出发，对内容、渠道、技术、资本、产品、人才等内外部资源进行统筹协调，以提升各个环节及整体协作的效率，加快语义化转型步伐。此外，还应做好具有实践性、前瞻性的顶层设计，开拓融合发展思路，提升数字出版内容质量和产品技术应用深度，拓展内容服务范畴，加强人才队伍建设。

参考文献

[1] А. И. 米哈依洛夫. 科学交流与情报学 [M]. 北京：科学技术文献出版社，1980：5-10.

[2] 王勇安，张雅君. 论出版产业融合发展的战略思维 [J]. 出版发行研究，2016（4）：14-18.

[3] 黄孝章，张志林，陈丹. 数字出版产业发展模式研究 [M]. 北京：知识产权出版社，2012：42.

[4] 史领空. 数字时代的出版 [J]. 编辑学刊，2000（4）：11-15.

[5] TUTEN T L，SOLOMON M R. Social Media Marketing [M]. 2nd ed. Los Angeles：Sage，2015.

[6] HEATH T，BIZER C. Linked Data：Evolving the Web into a Global Data Space [M]. San Rafael：Morgan & Claypool，2011.

[7] 萨蕾. 数字图书馆元数据基础 [M]. 北京：中央编译出版社，2015：25-30.

[8] BERNERS-LEE T. Linked Data-Design Issues [EB/OL]. [2018-07-02]. http://www.w3.org/DesignIssues/LinkedData.html.

[9] BIZER C. Expert Report on Linking Data [R/OL]. [2018-07-02]. http://151.1.219.218/b43d3f37-bd5d-4144-9779-b27a0ca3d1d5.pdf.

[10] 关联数据云（LOD Cloud）[EB/OL]. [2018-10-18]. https://lod-cloud.net/versions/2018-10-31/lod-cloud.png.

[11] Office of the Deputy Assistant Secretary of the Army（Research & Technology）. Emerging Science and Technology Trends：2016—2045——A Synthesis of Leading Forecasts Report [R/OL]. [2018-10-05]. http://www.defenseinnovationmarketplace.mil/ resources/2016_SciTechReport_16June2016.pdf.

[12] 腾讯研究院：2016中国数字内容产业全景解读 [EB/OL]. [2018-10-22]. http://www.alibuybuy.com/posts/90054. html.

[13] 曾建勋. 知识链接的构建方式研究 [J]. 图书情报工作，2010，54（12）：32-35，77.

[14] 许鑫，江燕青，翟姗姗. 面向语义出版的学术期刊数字资源聚合研究 [J]. 图书情报工作，2016，60（17）：122-129.

作者简介

苏静，女，1988年生，博士，讲师，研究方向：数字出版与知识组织，E-mail：owensujing@163.com。

Research on the Construction of Semantic Publishing for Scientific Communication

SU Jing

(School of Journalism and Communication, Shaanxi Normal University, Xi'an 710119, China)

Abstract: As an emerging advanced form of digital publishing, semantic publishing is necessary to systematically sort out the resource output form of formal communication and informal communication and its semantic network construction framework in order to effectively meet the information demands in the process of scientific communication. On the basis of comparing the implementation process of digital publishing and semantic publishing, this article points out that the implementation process of semantic publishing should increase the active content collection process of publishing institutions, refine the value-added operation of content editing and publishing, and pay attention to user feedback and interaction with authors and editors during the content consumption phase. Based on this, according to the construction requirements of semantic publishing, it proposes to construct an institutional framework suitable for the existing academic information environment and scientific research personnel from the four levels of resource layer, management layer, method layer and service layer. Among them, the knowledge service effect of semantic publishing needs to be realized on the basis of multi-source and large-scale resources. Simultaneously deep semantic association and recommendation is the key to the construction of semantic publishing system.

Keywords: Scientific Communication; Semantic Publishing; Knowledge Organization; Linked Data

（收稿日期：2018-11-02）