# 基于碎片重组的动态数字出版模型研究*

温有奎
（北京万方软件股份有限公司，北京 100038）

摘要：大数据具有数量大而密度低的特点，正在加剧人们对知识获取的困境。传统的知识出版业以本或以篇为单位的出版方式成为制约人们有效检索和利用知识的瓶颈。针对这一问题，文章提出基于碎片重组的动态数字出版模型，首先研究了知识碎片产生的社会背景和对科学研究所带来的科学价值，以及技术发展环境对当前社会化知识碎片阅读的推动作用；其次讨论了动态组合的数字出版模型，以及实现动态数字出版中所考虑的关键技术。本研究不仅发展了知识组织的理论，还推进了知识的深度检索、小粒度知识获取和高效利用的方法，大大提高了知识出版与传播的效率。

关键词：知识碎片；动态组合；数字出版
中图分类号：G237
DOI：10.3772/j.issn.1673—2286.2014.04.001

## 1 引言

大数据的到来，并没有改变人类被信息淹没却又知识贫乏的困境。大数据具有数量大而密度低的特点，正在加剧人们对知识获取的迷茫。传统的科技文献信息检索为人们获取知识提供了科学手段，但随着科技文献数量的日益剧增，研究内容的深度日益加深，跨领域的直接有用知识的寻找变得难以胜任。这两种现象反映了当前知识的生产数量在剧增和研究内容的深度同时在剧增。这两个剧增对传统的知识组织与传播的方式提出了挑战。至2012年末，非结构化数据占有比例达到整个数据量的75%以上[1]。数字出版加快了知识传播的速度，但这种以本、以篇为出版单位的数字出版方式却并没有解决让读者获取直接有用知识的问题。为解决上述问题，人们在不断寻找新的信息载体与传播手段，寻找各种新的出版形态[2]。数字出版被看成是一种新的、有前途的出版形态。如何将传统的出版标准形式、流程和模式转换为标准的数字化、结构化和规范化表达进行了大量的研究，而基于XML（eXtensible Markup Language）的开放式电子文档标准是解决文档有效表达的必要前提。目前，国外有关的开放式电子文档标准有很多：EPub、SCORM[3]、S1000D和NewsML。本文认为海量、非结构化的科学文献知识碎片化是影响多模态数字出版发展的关键问题之一，我们提出科学文献内容知识碎片化组织与按需动态关联重新组合的出版模式，发挥数字出版的多元化知识表示的优势，解决传统科技文献以本、篇为出版单位带来的知识难以有效利用的瓶颈问题；创立一种新的科学知识碎片化存储与按需动态聚合的数字出版模式，以推进科学知识的多模态利用。

## 2 碎片动态知识数字出版的挑战

### 2.1 碎片知识的科学价值

早在20世纪中叶科学家就在积极地探讨科学知识分裂现象，寻找直接挖掘所需要知识的方法，但一直没有很好的解决方案。20世纪60年代，美国情报学家Swanson教授对科学知识碎片（fragmentation of science knowledge）理论提出新的看法[4]：（1）客观知识总量与人类吸收能力存在巨大的差距；（2）跨学科的信息传递变得更加困难；（3）跨学科间存在潜在未被发现的关联，首次提出并验证了利用文献间存在知识碎片的推理发现新知识的方法。2000年Swanson

* 本研究得到“十二五”国家科技支撑计划“科技文献动态数字出版技术研发与应用示范”（编号：2012BAH90F00）资助。

为此荣获ASIST（American Society for Information Science and Technology）最高荣誉奖[5]。继Swanson方法之后有许多研究人员提出了很多改进方案，但由于知识碎片未能从文献中分离出来，依靠人工方法寻找和识别文献之间的互补性知识碎片，识别效率非常低而难以推广。20世纪70年代后期，美国情报学家弗拉基米尔•斯拉麦卡教授也曾看到了知识碎片的价值，提出从文献单元深化到文献中的数据、公式、事实、结论等最小的独立的“数据元”的思想。进入21世纪，我国情报学家徐如镜研究员提出了“知识元”的概念，指出知识的控制单位长期原则还停留在文献这一级上，而人对知识的需求一般不是以文献为单位的[6]。2002年，清华光盘股份有限公司开始进行了知识元的研究和商业化试验，为知识的深度挖掘和有效传播开创了先例。

基于知识元的知识组织也得到了数字化转型中的出版社的借鉴和响应。数字化出版社开始思考和探索以知识片段为单位的动态内容提供，围绕内容资源的知识片段分解、标引、检索和增值服务。读秀突破了文献单元检索和浏览的知识获取瓶颈，实现了图书章节物理碎片化技术，开创了真正意义上的文献内容深度揭示功能。这种把图书以章节为基础进行物理拆分、重新整合，提供以页为单位的文本资料数据库关联系统，为跨越传统的图书书名检索、图书目录检索向图书的全文内容知识点检索和浏览树立了榜样，也赢得了市场。

### 2.2 知识碎片化产生的原因

传播学对知识碎片化的研究文献出现在上世纪80年代。我国传播学学者认为知识碎片化产生的原因是社会阶层的多元裂化，并导致消费者细分、媒介小众化[7]。第三届《人民日报》读者评报活动调查结果显示，网络（54.12%）、报纸（46.32%）和电视（43.83%）是受访读者最重要的三种信息渠道。被调查阅读习惯的结果是38.35%的人习惯于“先看标题，如果感兴趣就往下看”，另有32.99%的人会“挑喜欢的版面或栏目看”，“从头到尾仔细看”的不到15%。人们的需求已经从获取“丰富信息”向获取“更多有效信息”转变。

Elsevier副总裁、技术服务研究与发展实验室负责人Allen博士认为[8]，随着学术信息在线搜索与获取的日益普及，学术出版不可避免地趋向于“在线与互联”，这就要求现代期刊（Journal 3.0）在内容出版方面必须具备片段化、知识化、语义化、可视化等特征，即：学术信息的发布应采用在被“人理解”的同时也要被“计算机理解”，信息的传播技术应采用“一次编辑、多渠道出版”的传播方式。

引起碎片化出版发展的原因有三个，一是需求，二是价格，三是效率。首先，碎片化来自读者的选择性需求。专业读者出于研究或是论文写作的需要，对知识的查阅、引用和更新是其主要目的，而读者真正感兴趣的可能只是整本、整篇信息中的一章、一节，甚至是一个片段。其次，以往读者只为了其中一部分有用的信息而支付整本书费用的方式，无疑增加了读者的负担。再次，专业出版社采取碎片化销售模式不仅可为读者提供更为精确的碎片内容，还可使碎片多次复用以及按需组合销售，极大地压缩成本。读者按照所需章节或片段的流量或字数支付费用，会产生不可估量的效率。

### 2.3 碎片动态知识数字出版的挑战

目前阅读方式的极大变化对静态的图书、期刊等知识传播方式提出了严重挑战，学术文献服务商不仅从出版商那里购买数据，更有可能与出版商联合出版。作者和出版商不仅可以整本出版，还可以以碎片知识数字方式动态出版，碎片知识以动态数字方式排版、存储、重组、联合出版[9]。动态碎片化数字出版方式大大节约人们的阅读时间，有效提高人们对知识获取和创新的速度，这将成为知识服务的新市场。读秀在数字图书阅读的初级市场上抓住了重要机遇、赢得了巨大的文献阅读市场。新的动态碎片知识数字出版在手机知识点阅读、多媒体阅读、多维度阅读市场的前景会更加广阔、潜力更大，用户更喜欢。科技文献动态数字出版内容版式分离、跨媒体数字资产管理、内容碎片化管理与动态关联、按需重组与内容复用、多出版形态数字产品同步生成、多渠道数字出版发布、多终端适配与移动阅读等关键技术将大大推进科技文献动态知识服务应用市场。

## 3 动态组合的数字出版模型

### 3.1 动态数字出版流程

在多介质跨媒体的数字时代，以纸介质出版物为核心的编、印、发的传统出版流程，已成为制约出版行业发展的障碍，已无法满足内容组织和服务过程中作者

远程协同写作、读者需求个性化定制和智能识别、编辑自动化等需求。因此，打破传统出版流程和概念的约束，建立一个基于内容对象的、协同工作的、“一次制作、多元发布”的动态数字出版流程成为数字出版行业的关键问题。

传统出版的流程主要是围绕作者的作品、编辑整理、三审三校、排版、印刷、发行、零售进行的，为此流程服务的关键技术必须保证内容结构、版式风格、文件格式不能分离。传统出版的内容与结构是一种固定模式，即把单一的文字、静态图像组合成作品变成出版物；将音乐作品变成音频出版物；将电影电视剧作品变成视频出版物。

因此，为了实现动态数字出版，首先必须解决传统出版的内容结构、版式风格、文件格式不能分离的关键问题。动态数字出版的关键还是内容，但动态数字出版内容结构与表现方式分离，只有到使用者选择时才确定表现方式，也就是内容结构、版式风格、文件格式是分离的，而不是传统出版的以版式为基础的变形。这样可以将内容从原来的种、册、件、篇、章、节到更小的片段内容按需重组。其次，按照多样性终端，将文件格式转变到动态检测终端后的适应格式，再以适合的格式文件发行。典型的动态数字出版流程如图1所示。

动态数字出版流程主要分为选题策划、编辑加工、内容管理、发布服务四个环节，从环节划分来说与传统出版流程有一定相似之处，但是在每个环节内的具体工作内容和特点，已经有了很大区别。动态出版流程的最主要特点就是利用互联网云服务的广泛性实时性、海量数据收集与处理能力、基于XML的内容版式分离和再现技术，来实现出版内容的结构化、碎片化、扩展性、自动多样性，从而为读者用户提供更加方便、快捷、廉价、智能的信息获取与知识服务。

## 3.2 数字内容碎片化组织模式

（1）数字出版物内容组织规范

目前用于描述数字出版物组织结构方式主要有三种：第一种是基于文档的描述方式；第二种是基于HTML的描述方法；第三种是基于XML的描述方法。基于文档的描述方式最常用的是PDF、WORD等格式，其组织方式是线性的，且组织结构和版式信息的描述具有专用性，在重构数字对象和个性化信息服务方面存在一定的难度，无法满足个性化阅读需求，不能进行跨平台的数据交换，也不能提供非线性的网状导航机制和立体的表现形态。基于HTML的描述方式虽然可以通过嵌入与超链接机制将线性阅读方式改为立体阅读，但也无法满足个性化数字出版的多维度信息检索和网状导航需求，再加上HTML本身的特点以及不具备跨平台间数据交换的缺点，已逐步被第三种方式——基于XML的方式所替代，其中应用较广的有两种：OPF和METS。

（2）碎片标引与索引技术

对数字出版产品进行碎片标引与索引是对文献知

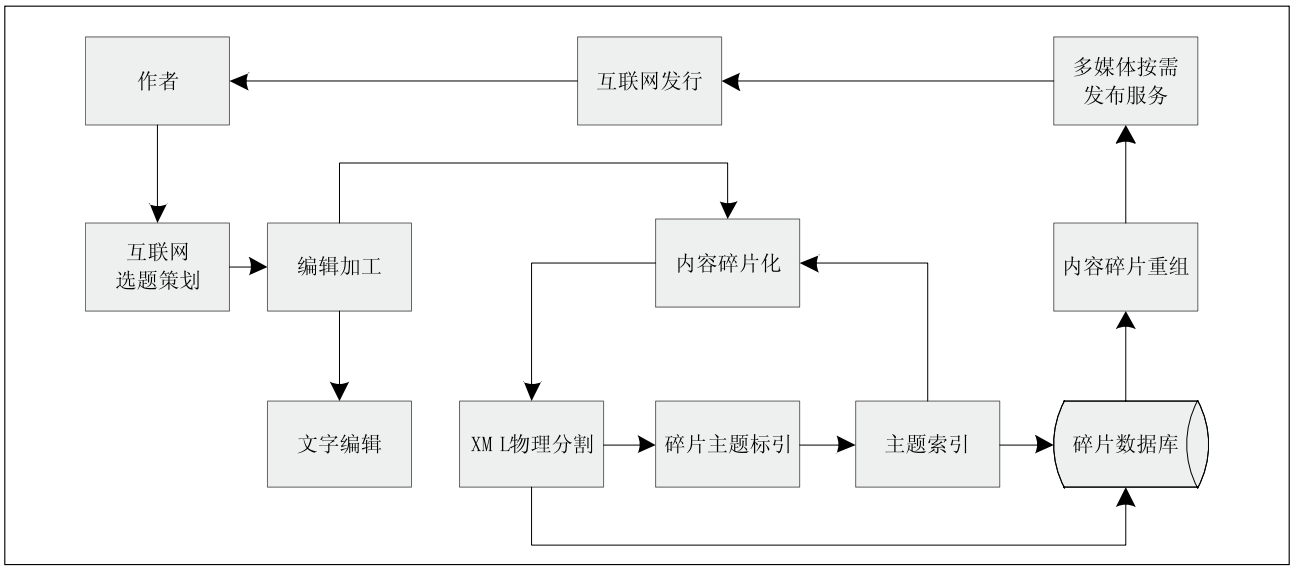


图1 动态数字出版流程

识组织理论的发展。与数字文献出版的元数据加工不同，除了对整本或整篇内容进行元数据标注外，碎片标引还要对数字文献各个章节的知识更详细地分别单独标引和索引。经过标引和索引后的碎片知识更容易被读者获取和利用，其生命周期要比整本书的更长、更有效。数字内容碎片化组织需要考虑几个问题：

①维持传统出版内容，保存作者稿件、终审稿件、终排文件，并转换终排文件按照种、册、件、篇、章、节模式进行组织；

②将形成的篇章节内容按学科、中图分类、主题等方式分类；

③将形成的分类按照某一学科、某一方向、某一行业的知识形成知识体系；

④将知识领域再拆分成不同方向的知识单元，知识单元拆分成知识点，最后拆分成主题词、关键词；

⑤通过关键词间语义关系将知识点进行动态关联，形成网状互联关系；

⑥将内容按需重组及多出版形态同步生成技术实现动态出版。

### 3.3 内容按需重组的多出版形态

（1）样式和模板技术

传统数字出版，自动化排版引擎采用数字内容与版面样式相分离的设计思想，在后期完成结构化或半结构化的数字内容与版面样式的组合，并对排版结果进行智能化校正。

对于动态数字出版，在进行内容按需重组时，完全可以借鉴其数字内容和样式模板分离的设计思想。对于不同出版形态的数字产品制作，可以预定义对应的样式模板，通过样式模板与数字内容的关联抽取，实现对应数字产品的自动生成。

（2）满足多出版形态需求的数据格式

由于数字出版形态的多样化，包括纸版印刷、网站发布、光盘出版、多终端移动阅读等。对于不同出版形态，其内容展现设备，如PC机、手持阅读器、PDA、手机等显示屏幕大小不一，所以需要研究输出内容自适应技术。该技术需要考虑在不同的发布渠道下，如何充分发挥不同终端设备的展示优势，从而将制作的内容更恰当地展示给读者，还原显示电子图书的规范版式，并能方便阅读。

（3）可扩展的多渠道输出技术

动态数字内容出版需要研究可扩展的多渠道输出技术，支持不同出版形态的输出结果，以适应包括纸质出版、电子书出版、移动终端出版在内的多渠道出版发布的需要。可扩展的多渠道输出技术是连接数字资产管理系统数字内容资源和多渠道发布平台的桥梁，由数据分发管理平台、多渠道发布平台、多终端支持接口组成，如图2所示。

数据分发管理平台主要包括模板引擎、任务调度、数据格式解析、多渠道输出引擎。其中：

①模板解析引擎：主要解决数字内容的提取与动态重组；

②任务调度引擎：主要解决多出版形态数字产品的自动和同步生成；

③数据格式解析引擎：提供对于电子书的格式解析、版式适应生成和流式阅读支持；

④多渠道输出引擎：提供面向不同出版形态和发

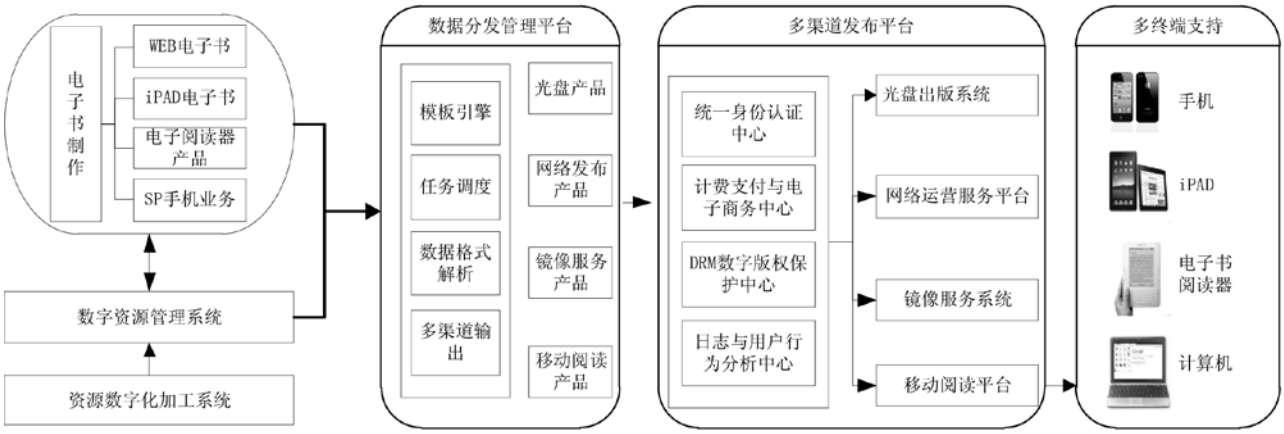


图2 数字资产管理系统

布平台的数字产品提供和输出。

## 4 动态数字出版关键技术

### 4.1 基于XML的内容版式分离技术

资源描述框架的理念已广泛应用于美国及欧洲等国家的数字出版与数字图书馆的建设中。本系列标准的研制将以资源描述框架为基础，建立一套适用于中国数字内容资源对象存储、复用与交换的新闻出版行业标准，使出版单位资源加工有据可依，使数字资源存储格式统一，实现数字内容的复用与交换，改变出版单位各自独立建立自用加工标准，全社会、全行业无法资源共享的现状。

内容版式分离技术在字处理软件和排版软件中均有应用。字处理软件包括Word、WPS等，其中以Word应用最为广泛；通过对Word2007和2010两个版本软件的相关分析、格式提取，利用XML结构化标引技术，实现软件中内容、内容结构、版式结构的分离，分离后的内容形成可重组和复用的资源，为资源积累、动态发布等环节做准备。

目前流行的排版软件如Indesign、方正排版、Word等，由于分属不同的公司，其使用的核心技术、版式规范各不相同，相互间无法实现有效转换，不利于数字出版多形态的生成和发布，也无法高效完成数字内容的按需重组。所以，必须要对三个主流软件产品的文件进行内容和版式的分离，才能做到根据内容进行碎片化，根据需要决定版式和格式。我们利用XML内容的中间文件作为三者的同步文件，这个技术的突破可以极大地提高中国出版技术自动化的水平。

### 4.2 多媒体碎片化内容的管理及复用技术

数字化背景下，大规模内容生产成为可能，同时也出现了规模化的内容消费需求，而内容融合的大趋势使动态数字出版应用示范平台上集成了包括文字、声音、图片、图像在内的各种形态的多媒体、碎片化内容。数据内容资源在内容和形式上越来越丰富，这就要求研究和开发多媒体碎片化内容管理及复用技术。同时，数字内容的复合出版、碎片化内容的立体使用也成为必然趋势，即数字内容同时在广电、报纸、书籍中使用的局面。如何针对不同的载体需要，对原始内容要素进行标准化、数字化的加工和存储；碎片规则确定以后，计算如何提取、如何保存、如何进行标引和知识组织、如何进行动态重组是需要事先进行复用规则的约定，在这个约定下进行管理和利用，也是本研究需考虑的问题。

碎片化解决以后，复用与重组是动态数字出版的关键技术之一。传统的内容管理可以管理碎片化的内容，但是无法管理碎片化内容的复用和重组规则，特别是动态的重组，需要实现申请请求、组合、输出等一系列标准化的动态重构。

对于碎片化的内容、整体化各种格式化的文件，在存储过程中如何检查、管理等对于出版机构是一个挑战，基本不可能把数万的文件一个一个打开，检查是否完好，必须要有一个检查海量文件存储后是否损坏的方法，然后对于损坏的部分进行备份修复，建立海量文件特征管理，以便于检查、管理、修复，这是目前数字内容检查及复用技术中的关键。

### 4.3 多出版形态同步生成技术

对于出版社来说，数字内容资源经过碎片化处理，可以满足其重用、按需出版和个性化服务的需求，这些内容资源通过数字资产管理系统进行统一管理和输出使用。

出版社实现内容资源数字化的最终目的还是为了满足其出版、发布、服务“一次制作、多元发布、多次服务、按需出版”的需要，因此，需要研究内容按需重组及多出版形态同步生成技术，来满足其对于数字资产管理系统管理内容资源的动态重组，并根据不同出版形态封装生成相应的数字产品，通过多渠道发布系统进行数字内容的出版发布。也就是说，在数字资产管理系统与多渠道发布系统之间，还有一个桥梁，这就是数据分发管理系统。

对于内容按需重组及多出版形态同步生成来说，需要重点研究样式和模板技术、满足多出版形态需求的数据格式、可扩展的多渠道输出技术等。

### 4.4 内容动态重组及按需出版平台技术

平台总体技术框架路线按业务流程、功能及特点，分为相对独立的三个层次：数据服务层、数据管理层和数据获取层。平台总体技术框架路线如图3所示。其中数据服务层主要包括多渠道数字出版服务系统、移动

阅读系统；数据管理层主要包括数字资源管理系统、数据验证管理模块、海量数据特征处理等模块；数据获取层主要包括：在线出版编纂系统、作者、编辑、专家标引工具、基于互联网的科技符号、图形的复杂编辑工具等。

向内容复用的跨媒体科技文献数字资源管理平台主要实现对于出版社数字内容资源，包括书报刊、篇章节、知识点、音视频、动画、图片等多媒体资源的集中加工处理、资源管理和数字内容输出服务。

数字资源管理平台分为内容存储层、通用组件层、内容整理层、逻辑内容库层及内容展现层。

（1）内容存储层：将各类数字内容存放入统一内容管理平台，其后通过内容碎片化处理，把内容按章节、图片等进行分割，并在分割后进行语义化标注，将处理后的结果存入碎片内容存储平台。

（2）通用组件层：系统将对内容的描述信息（属性标签）进行统一管理，并管理各类内容间的关联信息，同时，系统将为管理的内容提供全文搜索引擎，对全部内容进行统一检索。

（3）内容整理层：提供了语义引擎，帮助加工人员对数字内容进行标注；同时提供了内容标注工具，该工具帮助分割PDF文档为章节与图片，并为切割后的碎片化内容添加语义标签；内容检索系统提供了对不同层次内容的检索能力，并将检索到的内容按权重排序。编辑个人空间提供了编辑与作者积累和管理个人内容的工具。

（4）内容展现层：数字内容经过整理后，会形成各种逻辑内容库，如原始素材库、图片库、文章库、音视频库等，这些内容库既可以在出版社内部使用以加快各类内容编辑进度，也可作为增值服务平台向外部销售。

## 5 结语

早在20世纪中叶，科学家就在积极地探讨科学知识分裂现象，寻找挖掘所需知识的方法，但一直没有很好的解决方案。20世纪60年代，情报学家对科学知识碎片理论提出新的看法并进行了新的尝试，这一思想引起了出版业的关注，但由于技术实现上的问题未能得到全面的实验。大数据的到来，又一次加剧了人们对知

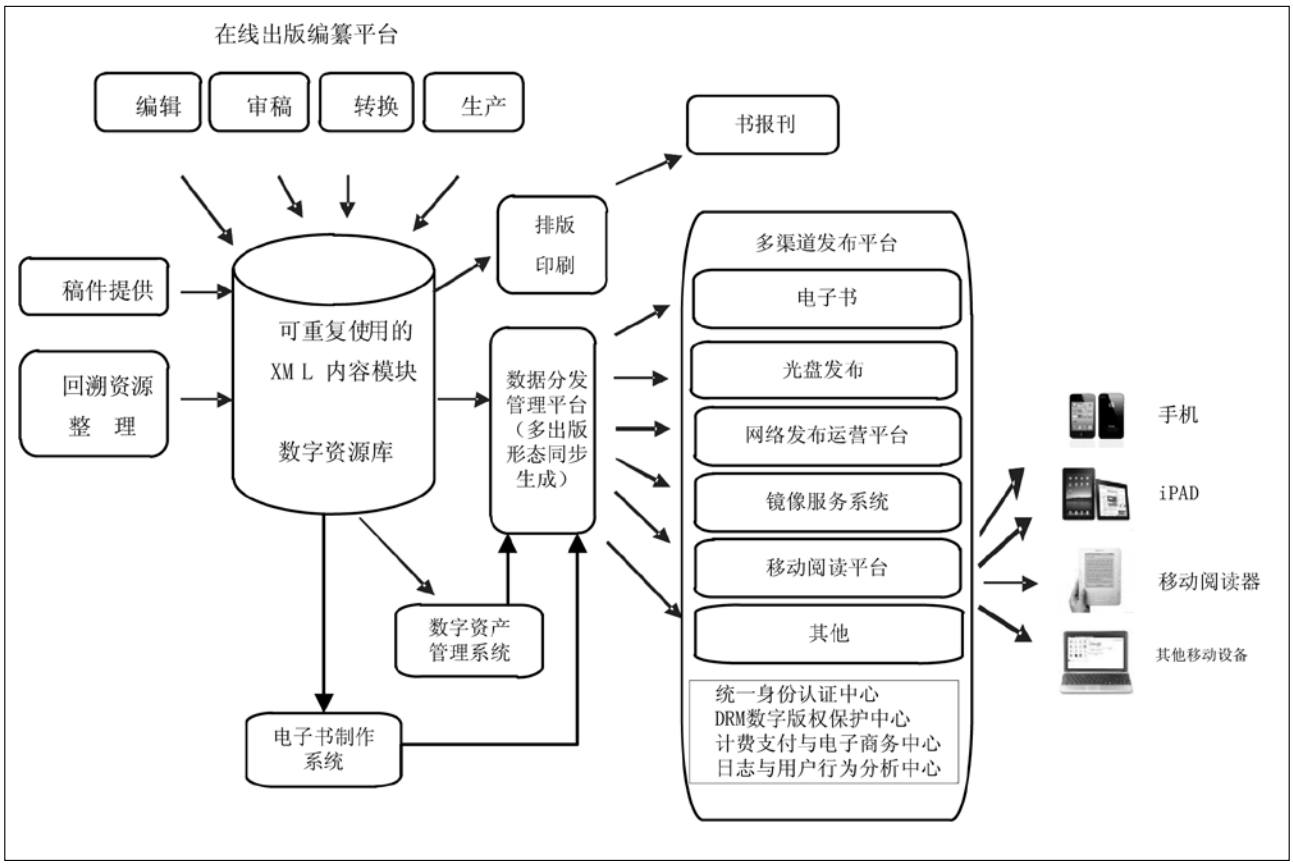


图3 平台总体技术框架路线图

识获取的困境。传统的知识出版业以本或以篇为单位的出版方式成为制约人们有效检索和利用知识的瓶颈问题，数字出版技术得到了广泛的重视和研究。数字出版兴起的强大的社会背景在于人们对知识传播粒度的更小要求，和能借助于最小粒度的知识片段进行知识的发现。因此，基于碎片重组的动态数字出版就成为知识传播领域的关键问题和研究目标。本文对这一问题的研究只是基本概念和模型的宏观研究，后续的具体技术研究将会深入进行，相信本文的研究对知识传播的发展会带来推进作用。

## 参考文献


[1] 李国杰,程学旗.大数据研究:未来科技及经济社会发展的重大战略领域:大数据的研究现状与科学思考[J].中国科学院院刊,2012,27(6):647-657.

[2] 高蝴蝶,张志林.基于知识元的内容组织对数字出版的启示[J].北京印刷学院学报,2009,17(5):33-36.

[3] scorm [EB/OL]. [2014-03-21]. http://baike.baidu.com/view/834676.htm.

[4] 马明,武夷山. Don R. Swanson的情报学学术成就的方法论意义与启示[J].情报学报,2003(3):261-266.

[5] SWANSON D R. On the Fragmentation of Knowledge, the Connection Explosion, and Assembling Other People's Ideas [EB/OL]. [2014-03-21]. http://www.asis.org/Bulletin/Mar-01/swanson.html.

[6] 温有奎,徐国华,赖伯年,等.知识元挖掘[M].西安:西安电子科技大学出版社,2005.

[7] 黄升民,杨雪睿.碎片化:品牌传播与大众传媒新趋势[J].现代传播,2005(6):6-12.

[8] HEY T, TANSLEY S, TOLLE K. The Found Paradigm: Data-Intensive Scientific Discovery [EB/OL]. [2014-03-21]. http://www.doc88.com/p-777870574601.html.

[9] 温有奎,温浩.中国学术搜索市场面对知识挖掘的挑战[J].情报学报,2013,32(12):1288-1294.


## 作者简介


温有奎，男，1951年生，管理学博士，教授，北京万方软件股份有限公司，研究方向：智能搜索引擎、文本知识挖掘。E-mail：wykui123@126.com。


## Dynamic Digital Publishing Model Based on Fragmentation Recombinant


WEN YouKui

（Beijing Wanfang Software Co., Ltd., Beijing 100038, China）



Abstract: The large number of low-density characteristics of big data, are exacerbating the plight of our knowledge acquisition. In the traditional knowledge publishing industry, this publishing way in articles and books as units are becoming the bottleneck of our knowledge retrieval and use. To solve this problem, this paper proposes a dynamic digital publishing model of fragment reassembling. Firstly, it studies the social background and scientific value of knowledge fragmentation, and the role in promoting the current social fragmentation reading brought by environmental technologies. Thus, it discusses a dynamic combination of digital publishing model, as well as key technologies to achieve dynamic digital publishing consideration. Our approach is not only to develop the theory of knowledge organization, but also to promote the depth of knowledge retrieval, small particle size and efficient use of knowledge acquisition methods, and to greatly improve the efficiency of the publication and dissemination of knowledge.



Keywords: Knowledge fragments; Dynamic combination; Digital publishing

（收稿日期：2014-04-01）