[文章编号] 1004-325X(2010)01-0004-05

中国政府公开信息整合服务平台的现状与未来

The Status Quo and Future of the Chinese Government Public Information Online

梁蕙玮 王志庚 （国家图书馆 北京 100081）

[摘 要] 国家图书馆建设的中国政府公开信息整合服务平台开通半年来，其基本建设情况、功能实现及服务状况等都有较好、较快的发展。然而，发展的同时在机制、技术、标准规范等方面也存在一些问题。对政府公开信息整合服务平台开展联盟化发展、多方位合作、多类型整合和多方式服务，可以促进平台的建设进程，加强政府信息的整合服务工作。

[关键词] 政府公开信息 国家图书馆 中国政府公开信息整合服务平台

[中图分类号] G253 [文献标识码] B

[Abstract] Since the openness of the Chinese Government Public Information Online half a year ago, the platform has had better and faster development in the basic construction situation, function realization and service condition. However, there are some problems in the mechanism, technology, standard and specification, and so on. The author proposes that the Chinese Government Public Information Online should implement the federated development, multiple cooperation, multi-type integration and multi-mode service, in order to promote the construction of the platform and enhance integration services of the government information.

[Key words] Government public information；National Library of China; Chinese Government Public Information Online

1 引 言

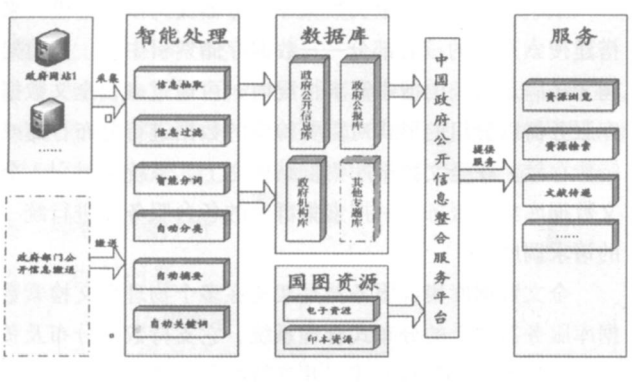
政府信息公开是提高政府科学执政、民主执政、依法执政能力，构建社会主义和谐社会的必然要求。《中华人民共和国政府信息公开条例》[1]（以下简称《条例》）实施1年以来，各地政府在政府信息公开专栏、政府信息公开目录建设等方面做了大量的工作。但要进一步提升我国政府公开信息的服务水平，必须充分重视对海量政府公开信息进行科学的组织与整合，为公民提供统一的开放的信息服务平台，使公众可以方便、快捷、一站式地获取政府公开信息。

国家图书馆（以下简称我馆）作为国家的重要文化基础设施，一直致力于推动我国公众文化信息共享事业的发展，是公众获取信息的重要窗口；作为从事信息资源组织与管理的专业机构，对于政府公开信息的整合亦责无旁贷。为此，国家图书馆对政府信息的整合进行了深入的研究，并于2009年4月30日推出了我国首个政府信息整合平台——“中国政府公开信息整合服务平台”（以下简称平台）(http://govinfo.nlc.gov.cn/)。平台的开通是国家图书馆为公众服务、为政府服务的一项重要举措，开创了图书馆对政府信息资源整合与利用的先河，更是国家图书馆创新服务手段、提高服务水平的重要体现。

2 平台介绍

“中国政府公开信息整合服务平台”的建设目标是根据《条例》赋予的职责，全面采集并整合我国各级政府公开信息，构建一个方便、快捷的政府公开信息整合服务门户，使用户能够一站式地发现并获取政府公开信息资源及得到相关服务。该平台不仅能成为公众获取政府信息的窗口、政府部门公开信息的重要渠道，同时也将为各级公共图书馆依法开展政府信息服务提供基础资源，成为图书馆开展所有政府信息服务的基础资源平台，使图书馆成为中国政府公开信息资源的保存者、整合者、传播者。平台的整体框架见图1。

图 1　平台整体框架图



首先是资源的获取。对于资源的获取，目前我们采用机器自动采集的方式，将各政府网站上的相关信息采集到我馆。同时我们也考虑下一步与政府部门开展合作，通过政府部门定期提交资源的方式获取政府公开信息；采集到我馆的信息经过信息过滤、信息抽取、自动分类等智能处理后，按我们设计的数据库结构自动地生成政府公开信息库、政府公报库、政府机构库（今后还可以考虑在公众比较关注的热点领域创建一些专题数据库），并且我们还将这 3 个数据库的内容与我馆印本文献信息、网络采集信息资源进行整合，对外提供服务（如信息浏览、信息检索等）。另外，还可以通过文献传递服务提供印刷版政府信息的复制等。

对于该平台，我馆采用了边建设边服务的策略。目前，该平台已完成了中央政府及其组成机构、各省及省会城市的上百家人民政府网站上政府公开信息栏目下资源的采集与整合，形成政府信息、政府公报和政府机构三大部分内容，其信息量超过 40 万条，收录时间跨度已超过 10 年，同时与国家图书馆的馆藏资源进行了整合，此外还收集整理政府机构 3 000 余家，为公众的查询提供服务。

3　平台功能实现

“ 中国政府公开信息整合服务平台 ”采用先进的系统构建方法、智能化及人性化的信息服务与检索方式。其设计目标是要建立一个安全、稳定、准确、及时、全面的政府公开信息整合服务系统，并且整个系统在总体设计上遵循开放、可扩展、经济、安全的原则，从而使整个系统结构合理、技术先进、易于扩展，既能满足当前的业务数据处理要求，又符合长期发展的需要。系统软件架构如图 2。

在应用功能层，主要设计了项目所需的各个应用系统或功能模块，包括网络信息采集系统、数据加工系统、信息发布系统、资源检索系统等，各系统的技术实现如下：

3.1　网络信息采集

在本项目中，为了完成系统的网络信息采集任务，针对采集网站数量多、信息海量的特点，我们采用了分布式体系结构以实现高速网页采集，具体的技术应用包括信息智能化采集，以实现各采集工作站任务均衡、各网站信息的更新；使用了采集任务集中控制，多台采集工作站分布采集的方式，实现可扩展的系统；采用多线程并发采集和控制，将采集模块分别安装在不同的采集工作站上，实现多采集工作站协同工作的模式，从而支持对大量网站的实时采集。

3.2　信息分析加工

3.2.1　信息自动分析和标引

为了满足本系统的应用，采集到的网页等信息对象必须经过以下智能化处理：正文内容提取——通过结构分析的方法确定信息对象的正文、图片及表格内容，自动剔除广告、导航信息等与主体信息无关的信息；格式自动转换——自动将 HTML 等格式文件转换为 TXT 文件，以方便再加工；属性自动标引——分析出信息对象的名称、文号、发布机构、分类等属性，分析并标注这些属性信息（元数据自动提取）；内码自动转换——将网页等信息对象中可能会包含的多种中文内码（如繁体 Big5，简体 GB2312、GBK，Unicode 等）转换成统一的中文内码，以便统一管理。

3.2.2　汉语分词

汉语分词系统是实现实体抽取标引与中文智能检索的基础，也是实现全文数据库和其他模块功能的重要基础。该系统内嵌汉语自动分词系统和多种分词词典（包括默认分词词典、附加分词词典、停用词典、附加停用词典、稀疏元组词典和单字词典），可实现规则与统计相结合的分词技术；可以准确识别人名、地名、组织结构名等信息；可以提供词性标注信息；系统同时内嵌分词歧义规则库，可以有效解决大部分的切分歧义。

3.2.3　信息自动过滤

自动过滤包括除噪和内容过滤两部分，除噪是指对网页无关内容进行过滤处理，如剔除广告、频道导航、版权信息等噪声信息，为后续的智能化处理、建立查询索引及纯文本保存提供干净的内容；内容过滤是识别和过滤各种有害文本信息（如色情、反动、封建迷信、商业垃圾邮件等），从而摆脱有害信息的侵扰。

图 2　系统软件架构图

3.2.4 自动分类

自动分类是指利用计算机，根据文献内容进行类别划分。本项目基于系统的分类模块来实现对政府信息从主题、题材、机构等多个维度的分类标引。自动分类功能支持基于语义规则的自动分类（机检分类）和基于统计原理（基于内容）的自动分类两种方法。用户可以自由维护分类词表，人工添加或修改规则。词表大小没有限制，规则分类支持多条件的与、或、非关系，具有设定词频数功能，并提供方便的规则定义界面。

3.2.5 自动排重

自动排重需要使用相似性检索技术。相似性检索是指对于给定样本文献，在文献数据集合中查找出与之内容相似的文献的技术。相似性检索技术需要在文献数字化表示（比如空间向量模型VSM）的基础上，通过计算文献之间的相似程度（向量之间的距离）给出文献之间的相关度指标。实践表明，相似性检索技术可以达到很好的网络内容自动排重、相关文章推荐效果。相似性检索的算法主要是基于特征词的提取和倒排索引技术，在效率上能达到百万级资料库的秒级响应速度。

3.3 信息存储与管理

系统采用Oracle数据库作为原始信息存储，并且为实现高效的信息整合存储与全文检索，系统引入了全文数据库系统，提供基于多种索引模式和知识词典的全文检索，并提供自然语言检索和相似性检索等全方位智能检索。全文数据库同时支持结构化数据和非结构化数据的存储管理，并且实现了Native-XML数据库功能，具备强大的结构化、非结构化和半结构化信息的处理和检索能力。它是整个搜索引擎的数据仓储中心，也是整个搜索引擎的检索动力核心，同时，结合全文检索网关实现完整的数据存储。

3.4 信息发布与检索服务

在本系统中通过全文检索网关来接入关系数据库，同步其中的数据到全文数据库中建立索引，依靠全文数据库服务器系统强大的检索功能和高效的检索性能来为上层的检索应用系统提供核心的检索动力支撑。在应用层基于全文数据库系统实现后台数据库中信息的对外发布及检索交互、结果表现等功能，为上层提供包括门户构建、政府信息资源搜索等信息服务。

为了使检索系统具备强大的检索性能及高度的稳定可靠性，本方案采用了2台全文数据库服务器及1台集群服务器构成检索的集群，并且该集群架构中的全文数据库采用数据镜像方式，数据可在2台服务器上相互镜像存储，每台服务器可存储另一台服务器的部分或全部索引数据。

3.5 搜索引擎的架构扩展模式

搜索引擎系统部署以后，因需要索引和提供服务的信息会随着时间的推移和应用的需求而不断增加。数据和用户数量的不断增长，会对搜索引擎系统的负载能力和扩展能力提出更高的要求。

本系统中采用的全文数据库服务器支持以集群模式来搭建搜索系统的核心部分——数据存储索引中心。当前架构无法满足增长的搜索需求情况时，可通过多台全文数据库服务器以分组的形式对需要检索的数据进行分布存储或镜像存储。在全文数据库集群结构之上，系统通过引入全文数据库集群服务器对搜索集群中的各台服务器进行统一的请求调度。

全文数据库集群服务器是架构在多个物理全文检索数据库服务器之上的分布式管理系统，它支持数据分布及负载均衡两种基本分布方式，并支持两种方式的组合运用。

4 平台服务状况

平台向公众开放服务后，受到了各方人士的关注。我们对近半年来的使用情况进行了统计，力图根据公众的实际需求对平台的发展进行调整，为公众提供更好的服务。统计包括如下指标：

页面浏览量：统计实际被点击的网页数量，“页面浏览量”往往被用来衡量网站内容的受欢迎程度和被访问情况。

唯一访问者数：是指访问网站的IP数量。

网粘度：是指在某指定时间内所有用户每次访问网站所用时间的平均值。

回访数：是指非第一次访问网站的用户数量。

访问深度：指用户每次访问网站时被请求的网页的数目。

统计结果如表1。

表1 中国政府公开信息整合服务平台部分数据统计表

指标	5月	6月	7月	8月	9月	10月
页面浏览量(页)	25 761	138 737	698 516	1 438 326	163 992	91 782
唯一访问者数(个)	1 953	7 314	8 367	9 700	11 153	8 131
网粘度（分钟）	12'2"	18'15"	16'34"	18'49"	15'17"	17'36"
回访数（个）	未统计	6 223	4 271	4 328	4 770	3 816
访问深度（页）（≥20页）	未统计	40.11%	39.08%	36.85%	35.77%	32.66%

注：2009年10月1-8日没有提供服务。

由表1可以看出，平台的页面浏览量从最初5月的25 761页骤增到8月的1 438 326页，尽管9、10月页面浏览量有所下降，但半年来平台页面总访问量已达到256万页；平台的唯一访问者数从5月刚开通时的1 953人，猛增到6月的7 314人，其后的几个月也在稳定地持续增长（10月尽管表面看上去有所下降，实际上是因为国庆节8天没有提供服务）；平台的网粘度基本上维持在12分到18分之间；每月的回访数都保持在4 000以上；用户每次访问20个页面以上的已经占了访问总量的32%-40%。

以上统计数据反映出在为公众提供政府信息的服务中，平台起到了一定的作用。另外，我们也对访问者所属地区进行了统计分析：到目前为止，访问者遍布国内各地（除

西藏外），同时也有来自美国、韩国、日本、澳大利亚等30多个其他国家的用户对平台进行了访问，这也反映出平台受众之广、影响力之大。

当然，由于平台的建设刚刚起步，有些工作还未开展，尤其是对外的合作与宣传还远远不够，这也造成了公众对平台的知晓度不高，故平台还未充分发挥它应有的价值。

5 平台问题分析

5.1 机制问题

5.1.1 缺乏法律保障

《条例》规定了政府信息应当在图书馆公开，但并没有对政府本身进行强制性的规定。图书馆作为政府信息公开法定的服务主体，它所提供的公开信息资源来源于政府机关，但在目前，图书馆对于政府信息的获取还处于被动地去各个政府网站上抓取的状态，这对政府信息的整合服务很不利。图书馆能否成为一个合格的信息公开服务主体，在很大程度上取决于图书馆能否与政府机关形成一个双赢的协调机制，这个机制主要来源于法律制度的保障。

5.1.2 缺少规模化建设

目前只有国家图书馆对国家层面上的政府公开信息进行开发，但是对于全国的政府公开信息的整合与服务，不是一家图书馆能完成的。任何一家图书馆对政府信息资源的组织服务，都很难照顾到其层级或其他地方的特殊问题。我们在平台的开发和建设中也是困难重重。

5.2 技术问题

平台的建设涉及到资源的自动采集、自动分类标引及资源的保存与服务等多方面的技术，许多方面都处于探索研究阶段，会遇到多种问题：

5.2.1 资源采集的问题

尽管在先期我们定义只采集政府信息公开栏目下的内容，但是由于该栏目本身的情况参差不齐，有的政府把正式的公文公报法律法规整合到信息公开栏目；而有的政府虽然设置了信息公开栏目，但是仅放了一些目录而没有内容，或是放了一些动态新闻；还有的直接就链到了网站的其他内容，这就给信息的采集造成了很大的困难，因为机器很难区分哪些是要采集的信息，哪些是无用的信息。所以在后期，对数据修改所花费的时间比前期建设的时间还要长许多倍。到目前为止，还有一些有问题的数据混杂在其中。

5.2.2 资源分类的问题

目前平台的主题分类仅为一级分类，即22个类目。对于数十万条的政府信息来说，基本上每一类数据都可以有几百页以上的结果，这样的分类是远远不能满足用户的需求的；对于检索技巧不是很高的用户来说，要想通过浏览的方式找到相关的资源其难度也是很高的，因此我们还需要对分类进行细化，进行二级、三级甚至四级分类。而且现在发布的平台仅保留主题分类的方式，但对于政府信息还可以从其他角度进行分类查找，比如说信息的类型，诸如公报、法律法规、统计数据或是动态信息。在这方面，我们做了初步的尝试，但是由于政府信息本身的多样性及海量性，目前自动分类的效果还很难达到理想的状态。

5.2.3 数据质量的问题

数据质量的问题主要体现在数据准确性上。数据准确性问题在机构信息中最为明显，主要是因为机构信息中动态的信息比较多，如政府机构的人事变动，原网站有可能将发生变动的人的网页撤除了，当我们再次采集时，无法采到新的信息，这就很难对已经采集并发布的原始网页进行更新，从而导致页面出现部分信息不准确的现象。另外，平台在数据的时效性和全面性方面也还有所欠缺。

5.2.4 资源保存的问题

在资源的保存上，我们采用两种方式，一种是保存纯文本，提供检索与浏览的常规服务；另一种是保存网页的原貌，以应对原始网页消失的问题。但目前，对于一些特殊格式的网页，我们不能原汁原味地保存网页的全貌。

5.3 标准规范问题

标准规范是信息资源一致性及平台扩展的基本保证，应围绕信息采集、组织、分类、保存、发布与使用等信息生命周期各环节建立相应的规范与标准。但在本平台的建设中，标准规范的建设还不完善，除元数据标准、分类标准外，其他环节的标准规范还有所欠缺。这主要是因为平台建设处于初创阶段，还有许多标准规范的建设有待尽快开展。

6 平台发展规划

6.1 联盟化发展

目前，国家图书馆所建设的平台，采集整合了中央政府及其组成机构和省人民政府的信息。这个平台只是我馆所构想的中国政府公开信息整合服务平台的一小部分。我馆所设想的平台宜走联盟化发展的道路，即由国家图书馆牵头，联合国内各省、市公共图书馆及部分重点区县图书馆成立政府信息整合服务联盟，共同打造一个可以让各个图书馆共同参与建设的大规模集中式的政府信息整合服务平台，通过该平台各个联盟成员馆可以采集整合各自行政区域的政府信息，实现分层建设、共建共享，同时还可以实现个性化展示和统一展示的完美结合，为公众提供更完善的政府信息服务。

6.2 多方位合作

政府信息的整合与服务仅靠图书馆的热情是远远不够的，除在公共图书馆界开展合作共建、走联盟化发展的道路外，还需要和政界、学界、法律界及各类从事信息检索的单位合作。一方面可以争取政策上的支持，甚至是法律上的保障，从而确立图书馆在政府信息公开中的地位，为

平台的进一步发展创造条件；另一方面也可将相关单位的研究成果及经验纳入到平台的建设中，从而进一步推进平台标准化、规范化建设。

6.3 多类型整合

目前，平台以规范性的文件为主，像公文、公报、法律法规，但还有大量的其他类型的政府信息没有涉及，如各种统计数据、电子政务项目，另外还有大量的“泛”政府信息，也就是公开目录未涉及的相关内容，如政府组织的各类会议和公共活动的相关报道、政府官员出席活动的讲话或者政府官员的博客。如果对这些信息进行深度的挖掘，做到全方位的整合，将能给公众展示某一事件的全貌，或给政府机构决策提供参考。如国务院发布一个条例后，会产生相关的新闻快讯、条例解读，这种条例流转可以衍生成不同的信息，并且条例的颁布还能在社会上产生不同的影响、导致各种事件的发生，进而又产生新的条例，等等。将这些信息进行分析，挖掘信息之间的关联关系，实现基于语义分析的政府信息关联，做到多类型政府信息资源的深度整合与服务将是政府信息整合的一个重要发展方向。

6.4 多方式服务

政府信息整合服务的展现方式是可以多种多样的，网站只是其中之一，还可以通过触摸屏、手机、电视等不同的方式向用户提供服务。目前，国家图书馆已经实现了网站和触屏的服务，都受到了用户的好评，后续国家图书馆还将考虑通过手机或数字电视的方式为用户提供服务，让用户可以随时随地地获取到政府公开信息。

7 结 语

中国政府公开信息整合服务平台的建设是国家图书馆依照《条例》在政府信息领域开展整合服务的探索与尝试，这对于政府机构、公共图书馆乃至公众来说都是一个新生的事物。如何能以平台为基础，为公众、政府、图书馆提供政府信息服务还有很多问题需要研究，为此，国家图书馆还将继续努力，不仅联合全国各级公共图书馆结成“全国图书馆政府公开信息服务联盟”，同时还要加强和政府部门的联系，增进同专家学者的交流，团结一切可以团结的力量，共同做好政府信息的整合服务工作。

参考文献：

[1] 中华人民共和国政府信息公开条例[EB/OL].[2009-10-28].http://www.gov.cn/zwgk/2007-04/24/content_592937.htm.

[**作者简介**]

梁蕙玮　女，1974年生，情报学硕士，工作于国家图书馆数字资源部，副研究馆员，主要研究领域：数字资源整合与服务，已发表文章10余篇。

王志庚　男，1973年生，国家图书馆数字资源部(中文互联网信息资源保存保护中心)主任，副研究馆员，研究领域为学术出版与传播、电子出版与电子资源管理、图书馆自动化和信息服务，研究重点是元数据与数字资源长期保存，已出版专著2部，发表论文17篇，参与国家自然科学基金课题研究1项、国家社科基金项目1项、国家图书馆科研课题6项。

[收稿日期：2009-12-04]

(上接第3页)共建实现公共图书馆在政府公开信息的整合开发方面的统筹协调发展，以实现对各级政府信息资源的收集、整理、保存、开发、利用并服务于公众。国家图书馆倡议建设“全国图书馆政府信息整合服务联盟”，其主要目的也在于通过共建共享打造联合舰队，采取分工合作的模式提高公共图书馆系统的社会影响力，创造新的社会价值。

政府信息整合服务是一项长期而艰巨的任务和使命，全国公共图书馆应该联合起来，秉承为政府服务、为公众服务的宗旨，以《条例》颁布实施为契机，以积极的态度、职业的精神、专业化的操作充分参与到政府信息公开和服务工作中去。通过不断创新服务搭建好政府与公众之间的桥梁，使公共图书馆成为公众获取政府信息的重要窗口，将公共图书馆系统建设成可信赖的、权威的中国政府公开信息的收藏者、整合者和传播者，成为政府信息资源的增值服务者。

参考文献：

[1] 中华人民共和国政府信息公开条例[EB/OL].[2009-09-28]. http://www.gov.cn/zwgk/2007-04/24/content_592937.htm.

[2] 邹 勇.政府公开信息有效组织的实现方式研究[J].情报资料工作,2009(1).

[3] 王 钰.浅议政府公开信息的采集、保存与管理[J].档案与建设,2009(5).

[4] 周勇娟,陈艳红.从网站角度探讨我国公共图书馆的政府信息公开[J].情报资料工作,2009(2).

[5] 锐 进.浅议政府信息资源的增值开发利用[J].国有资产管理,2007(7).

[6] 王守炳.试论政府信息资源管理[J].重庆行政,2002(1).

[7] 黎延凯.政府信息公开环境下的政府信息资源管理[J].科技促进发展,2009(5).

[8] 杨和焰.论网络时代的政府公共信息管理与服务[J].兰州学刊,2006(9).

[**作者简介**]

王志庚　男，1973年生，国家图书馆数字资源部(中文互联网信息资源保存保护中心)主任，副研究馆员，研究领域为学术出版与传播、电子出版与电子资源管理、图书馆自动化和信息服务，研究重点是元数据与数字资源长期保存，已出版专著2部，发表论文17篇，参与国家自然科学基金课题研究1项、国家社科基金项目1项、国家图书馆科研课题6项。

陈　力　男，1958年生，研究馆员，博士生导师，国家图书馆副馆长，中国图书馆学会副理事长。

[收稿日期：2009-12-01]