中国科技资源导刊 ISSN 1674－1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2009 年 3 月 第 41 卷第 2 期 42－46 ISSN1674－1544 Vol. 41 No. 2 42－46,Mar. 2009

三种元数据整合方法的比较

孙 卫
（中国科学技术信息研究所,北京 100038）

摘 要:本文介绍了掌握知识的 6 个纬度方法与数据、信息、知识转换的相互关系,分析了在现实社会中,资源产生的过程和类型,提出针对稳定的资源,用户可以按照各自使用习惯进行资源整合。在数学上和结构上,文章给出了元搜索的匹配关系,并认为在元搜索理论中,元检索结果整合和元整合检索两个方法已成为如今资源整合中实验与尝试的方法。

关键词:元搜索;资源整合;资源共享;数字图书馆;元数据
中图分类号： TP391. 1; G250. 76 文献标识码： A DOI: 10. 3772/j. issn. 1674－1544. 2009. 02. 008

1 引 言

数据、信息、知识是逐步发展和形成的。事实上,它们是一种互补的过程。前人的知识、经验、教训都是人们进行下一步创造的源泉。一旦新的知识产生，立即就变成了自己和别人下一次创造的源泉——数据和信息。所以,从这个角度来看,提供给研究人员的数据、信息、知识必须是经过整合或者融合的。

在人们学习掌握知识的过程中，都要经历 6 个阶段,即记忆阶段、理解阶段、应用阶段、分析阶段、评估阶段和创作阶段。而在这 6 个阶段中,不同阶段对科技资源的需求也是不同的。

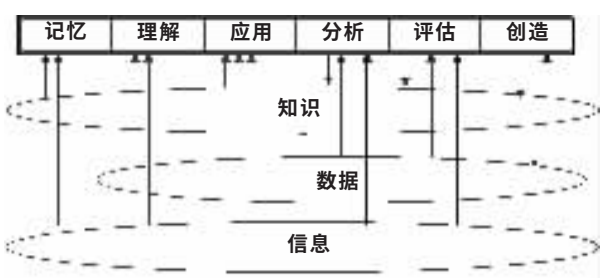

图 1 学习掌握知识的 6 个阶段

在图 1 中,可以看到学习与掌握知识中的 6 个阶段，而信息、数据、知识是这 6 个阶段中必不可少的素材或资源。数据是应用知识、分析知识、评估知识所需要的素材，新创造的知识也会变成别人知识进化的数据或者知识。信息则是各个掌握阶段所必需的。

实际上，在知识创造的过程中，利用扎实的基础知识（通用知识）发现别人没有发现的规律（隐性或者显性的学术知识），然后对知识、技术进行改进、优化,从而创造出新的知识(理论)、产品(实用新型)、外观等。还有一种创造是依赖于实验环境。在实验环境中,积累多种实验的数据,从中发现一些规律，即便是不能上升为理论或者不能用理论来说明,但这些实践结果也是一种成果。二者必居其一。

为了讨论方便，本文把数据、信息、知识统称为资源。

2 资源的划分与管理体系

我国大部分资源产生与管理的实际关联关系如图 2 所示。

图 2 展现了在中国现实社会中最常见的 5 个相对稳定资源（出版物与学术作品、档案、专利、产品、电子政务信息资源）产生的事实关系,以及不同的管理机构与体系。中图分类法：针对正式

作者简介:孙卫(1957－),男,电子技术与计算机专业高级工程师,研究方向是知识内容组织、挖掘、处理与相关技术结合。

收稿日期:2008 年 9 月 10 日。

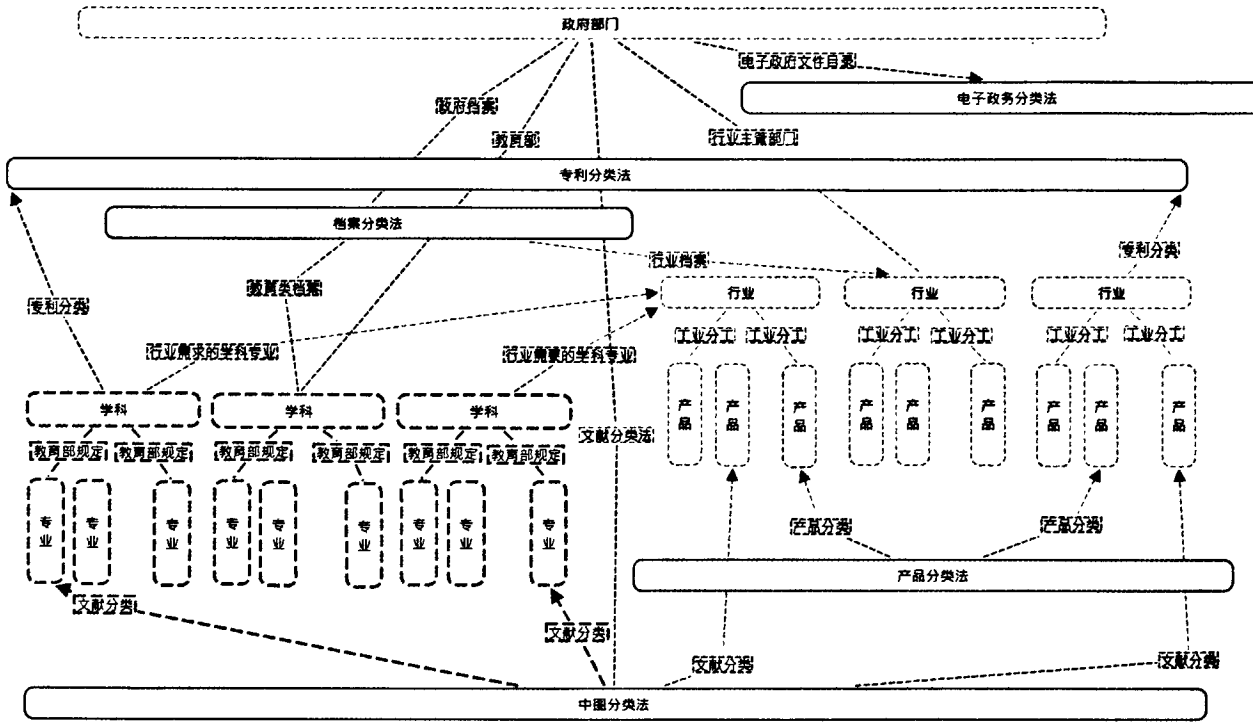

图 2　主要资源产生与管理示意图

出版物和学术作品的管理；产品分类法：针对商品交换与流通的管理；专利分类法：针对发明、实用新型、外观设计等知识产权的管理；档案分类法：针对各个行业的过程文件的管理；电子政务目录法：针对全国各级政府文件的管理。还有教育专业分类法，这属于专业分类，对于不同行业可以使用很多相同的专业的分类。对于现实社会的人，更熟悉的是所从事行业、所属专业的知识体系，而不是用于文献、档案、专利、政务、产品等的资源管理体系。

在网络化与数字化的时代，这 5 类资源和对应的管理方法，已经在各自的专业范畴发挥了巨大的作用，但是，使用者只能分门别类地进入这些不同的资源系统中去。而对于实际研究的需要，却存在资源产生与管理的区分与混合、专业与行业的区分与混合、自然知识体系与管理模式的区分与混合等。

由于存在这些区别与混合，在资源服务组织的时候便遇到了很大的挑战，既要满足资源原有产生与使用的群组合现实需要，也要实现不同资源互相跨越和多组合地提高信息服务发现、利用率的需求。于是，笔者根据数字图书馆和科学数据的理论研究和实践，提出以元数据技术为核心的资源整合方法。

3　资源整合的方法

下面我们先定义一组集合：

中图分类——LC、产品分类——PC、国际标准专利分类——IPC、档案分类——AC、电子政务分类——GC、领域集合——DC。

根据上述集合可以发现，集合 DC 属于 LC、PC、IPC、AC、GC，即

$$DC \in \{LC\ \ PC\ \ IPC\ \ AC\ \ GC\}$$

所以，我们在处理这个 DC 的时候，需要与 LC、PC、IPC、AC、GC 等建立交换资源空间 CC。

由此可见，DC 是通过 CC 与 LC、PC、IPC、AC、GC 相关的。

$$DC = CC\{LC\ \ PC\ \ IPC\ \ AC\ \ GC\}$$

那么，这个交换空间如何实现？笔者认为，在 LC、PC、IPC、AC、GC 要遵守各自的建立标准，因为这些部分是稳定的，是长期积累下来的，需要在 LC、PC、IPC、AC、GC 中增加复分的部分。因为很多事务同属于多个大类是经常发生的事实，而归类管理单一化是没有遵循事实客观规律的。

建立 CC 有两个方法：应用结果交换法和原始元素重组交换法。

3.1　应用结果交换法

假定：

LC、PC、IPC、AC、GC、CC 有各自的应用系统，分别为 ALC、APC、AIPC、AAC、AGC 和 Acc。在各自的应用系统中具备检索功能和链接对象功能。检索结果计为 S，链接对象计为 L。那么 Acc 为：

$$A_{CC(S,L)} = \sum \left\{ \begin{array}{l} (S_{LC}, L_{LC}) = F(A_{LC}) \\ (S_{PC}, L_{PC}) = F(A_{PC}) \\ (S_{IPC}, L_{IPC}) = F(A_{IPC}) \\ (S_{AC}, L_{AC}) = F(A_{AC}) \\ (S_{GC}, L_{GC}) = F(A_{GC}) \end{array} \right. \quad (1)$$

公式（1）表示了不同的检索结果集合抽取最小公共集合的整合原理。

根据公式（1），可以建立多种检索结果的最小公共集合整合方法。在数字图书馆领域，这个整合方法已经被广泛采用。在系统体系结构上，不需要对原来的 ALC、APC、AIPC、AAC、AGC 应用系统做改造，只需要针对各个应用系统的 S 函数做匹配，在各个应用系统检索结果返回后，保留原有的 L 函数。在所有的 S 返回后，对最小共集（各个应用系统元数据解释含义相同的元素）进行排序，结果给出，含 L 关系。

此法存在以下问题：①无法针对使用的需要给出最合理的元素集合；②同类资源的重复判断整合困难，只能多次重复排列。

此法的基本要求是：各种应用系统有检索功能，可以是 B/S 结构也可以是命令行结构；各种应用系统检索结果数据结构存在最小共集；CC 系统可以把用户检索命令分别转换成各个应用系统的检索表达式，可以接收并解析各个应用系统的检索结果数据、数据包，可以根据最小共集原理，整理接收到的所有数据，进行排序、整理、显示。

此法的高级要求[1]是 可以利用 Z39.50、ISO10160/10161 等标准进行机构应用系统互操作；可以转换成 Z39.88 的标准进行链接；可以转换成 XML 文件的格式[2]进行交换。

在美国 NISO 标准委员会关于元搜索[3]的标准建议中，也强调了这个类型。元搜索可以上升成为数学表达关系并得到广泛的应用。

3.2　原始元素重组交换法

假定：

LC、PC、IPC、AC、GC、CC 有各自的元数据元素集 M1 到 Mn，并且这些元素都是遵循某一类标准规范的。那么 CC 的索引集 Index cc 为：

$$\text{Index CC} = \text{Index}\left\{ \begin{array}{l} CC(M1) \\ \vdots \\ CC(Mn) \end{array} \right\} \quad (2)$$

这里的

$$\left\{ \begin{array}{l} CC(M1) \\ \vdots \\ CC(Mn) \end{array} \right\} \in \left\{ \begin{array}{l} LC(M1, Mn) \\ PC(M1, Mn) \\ IPC(M1, Mn) \\ AC(M1, Mn) \\ GC(M1, Mn) \end{array} \right.$$

公式（2）表示了元重构与索引重构集合整合原理。

根据公式（2），首先是不改变 LC、PC、IPC、AC、GC 的元数据元素的结构，保证原始资源层的稳定。根据需要重新建立 CC 的元数据的元素，大部分元素直接来源于原始稳定数据元素，然后根据这个重建的元数据元素建立新的索引集。

重构 CC 的原则是基于多种数据结构，适应使用者共性的习惯，从而保证最大比例的用户利用这个重构的 CC 层，并发现和获得所需要的资源；重构的 CC 层必须能引导到对应引出的数据结构层、对应的记录号层；在抽取索引时，重构的 CC 层要方便使用者进行检索、排序等。

2004 年以后，国内的清华同方知网在元重组上进行全新的构造。在图书馆研究的 FRBR 也是一种重构的标准。Exlibris 公司的 Primo 产品[4]就是一个元整合以后的应用产品。国家发改委等 12 部委开发的基于 GIS 的资料信息交换系统，也是重构这个数据元素层的，是基于 ISO TC211 标准进行的。

重构 CC 需要在交换元素层中进行，所以这个层要遵循一定的标准，目前大部分的重构都是基于都柏林元数据标准的。

此法的优点是可以结合使用共同特点构造元数据元素，可以最大限度地适应使用者的习惯。但

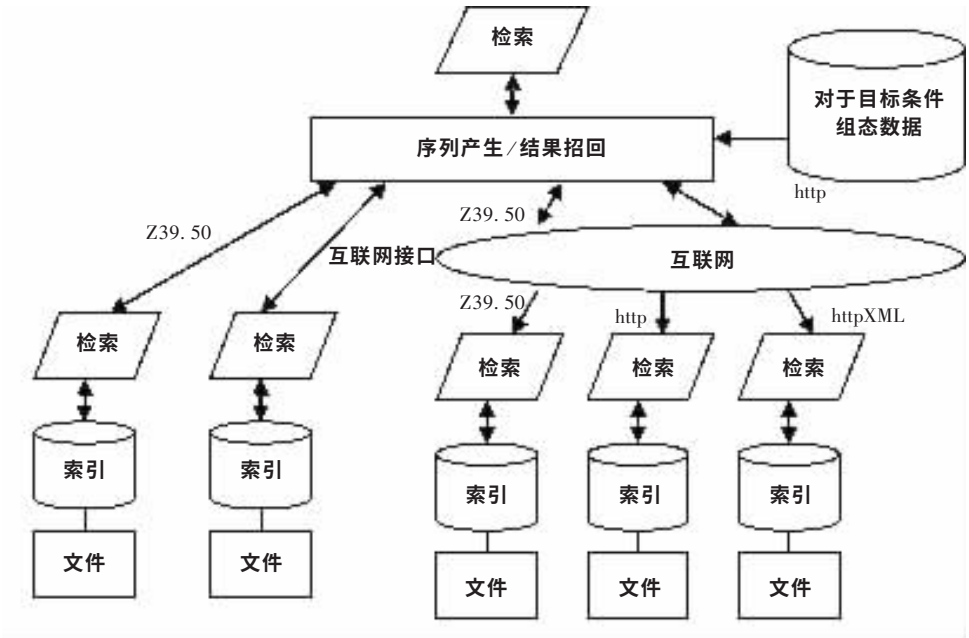

图 3　整合系统 :元搜索/门户解决示意

最大的问题是当没有资源的原始元数据时这种整合是不可能进行的。同时,都柏林元数据描述更复杂的元数据时也存在较大的限制。

3.3　混合法

在应用结果交换法和原始元素重组交换法的具体使用过程中，都有一定的约束性，所以，对于没有资源元数据的,采用应用结果交换法 ;有资源元数据的则采用原始元素重组交换法,再建立 Acc 组合到其中,形成一个混合的整合方式。

图 3 是基于检索结果进行整合的结构示意图[5]。整合的远程互操作的接口是 HTTP 和 Z39.50,在互联网上网站已经具备了检索的功能。在本地互操作的接口可以是专用的，也可以是符合 Z39.50 的标准。在国际对于基于检索结果集进行整合的系统称之为检索门户。这个方法与 Google 的检索最大的不同,就是不在本地建立元数据层和索引层。那么，使用者可以根据自己的需要选取有效的和对于本身有用的资源。Google 无法在用户需要的专业类型资源上进行进一步的选择，而这种 Acc 的模型是一个非常成熟的架构模型，用户可以挑选需要的专业资源类型，组配检索结果的数量等。所以对于使用者，这种类型的整合可以提高检索的效率。在国内数字图书馆中，这种模型被称之为统一检索、单一检索或智能检索等。

图 4 是基于重构索引进行整合的结构示意图[6]。这种方法的关键在于在目标系统文档中建立了元数据要素和重建索引层。国外对这个系统的定义是本地中央索引解决方案，属于典型的元整合结构。Google 则属于这个模型。

混合结构就是把图 4 作为图 3 的一个子系统集合在一起。

综上所述，元搜索实际上是在元检索结果集的整合和元整合利用的基础上实现的。

4　结　语

在对不同行业主管的信息平台需要进行整合时，最好采用检索结果集最小共集的整合模式，这个方法的投入很小，不会打破原来行业的应用体系。在不同资源供应服务商的产品中这个模式是最常采用的一个方法,是非常简单有效的。

对由行业牵头并提供经费的各个机构的资源进行整合时，最好采用重构索引集的整合方式，这个方法有利于保留最有价值的字段帮助使用者进行发现，也不破坏原有的应用体系，只是需要重新构造一个层，这个层最好是利用技术手段进行处理,而不要采用重新加工的方法。清华同方知网工程在 2004 年就使用重构和技术并举的方法进行了整合，以发改委牵头的基于地理信息为基础的 12 个部委资源的整合是通过技术方法整合各个要素后形成的。

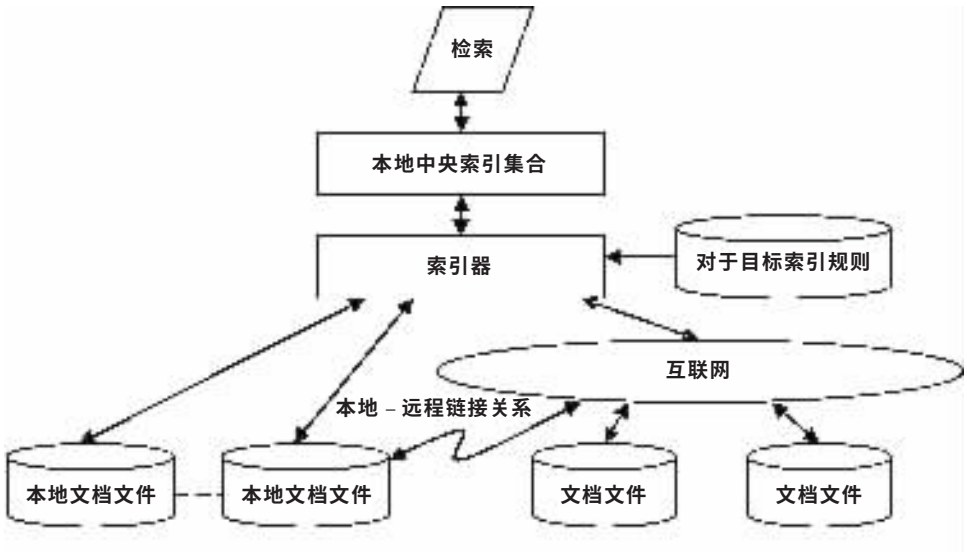

图 4　整合系统 :本地中央集中索引解决示意

在这个领域中，下一步的发展主要是对应用系统包括命令集、应用网关、数据封装等的进一步规范和标准化，逐步在元数据层实现规范化，在应用系统上有利于互操作。

参考文献

[1] Daniel J. Crichton ,Information Architecture for Space Data Systems[EB/OL]. [2008 - 08 - 29]. http: //cwe. ccsds. org/sea/docs/SEA - IA/Meeting% 20Materials/2005/2005% 20Fall% 20Meeting/Fall2005 - IAWG - Discussion - Slides. ppt.

[2] Exlibris. Aleph XML Services[EB/OL]. [2009 - 03 - 11]. http // nlso. kavi. com/news/sevents/niso/past/D2D - 06 - Groves. pdf.

[3] NISO. RP - 2006 - 02 [EB/OL] . [2008 - 08 - 29] . http: //www. niso. org/publications/rp/RP - 2006 - 02. pdf.

[4] Exlibris. Primo [EB/OL]. [2008 - 08 - 29]. http: //www. exlibrisgroup. com/category/PrimoOverview.

[5] Eric Sieverts. Google and/or/not Database[EB/OL]. [2008 - 08 - 29] . http: //conference. ub. uni - bielefeld. de/archiv/2002/lectures/Sieverts. ppt.

[6] Eric Sieverts. Google and/or/not Database[EB/OL]. [2008 - 08 - 29] . http: // conference. ub. uni - bielefeld. de/archiv/ 2002/lectures/Sieverts.

Comparison of Three Methods for Metadata Integration on Knowledge Study Process

Sun Wei

(Institute of Scientific and Technical Information of China ,Beijing 100038)

Abstract: There are six methods to acquire knowladge, and data, information and knowledge are the conversion of mutual relations. This paper analyses the process and type of resources producing in the real society. On stabilizing resources, users can integrate resources in accordance with their respective and customary use. In mathematics and structure, the paper gives the matching relation of meta search. In meta search theory, the integration of meta search results and the meta integrated search have become today's experimental and trial methods in resources integration.

Keywords: meta search, resources integration, resource shared, digital library, metadata