CASHL 特藏++平台调研与框架设计

北京大学图书馆

王铮 曾丽军 周春霞 梁南燕 张慧丽 吴亚平

引言：

目前，已有武汉大学图书馆、中山大学图书馆等十所高校图书馆着手特藏文献的“特藏++”项目，但是各馆特藏资源揭示平台并不一致。随着特藏++资源的揭示，将有更多的服务馆加入该项目，CASHL 管理中心有必要建立一个统一的“CASHL 特藏++平台”，把特藏资源有序、统一揭示，利于读者检索使用，也有利于服务馆直接利用 CASHL 项目成果平台进行资源加工添加，无需重新设计平台，避免重复建设。

本文主要研究 CASHL“特藏++”项目平台调研与框架设计。

目录

1， 大型特藏的文献学特征
2， 对“特藏++”大型特藏揭示项目的分析与研究
3， CASHL 特藏++平台调研与框架设计
4， 总结与展望

1， CASHL 大型特藏的文献学特征

在人文社会科学教学和科研中，特藏文献被公认为极具科研价值与收藏价值的珍贵文献，尤其是大型特藏文献，多为第一手的原始档案资料，但受其价格昂贵的限制，诸多高校图书馆无力购买收藏。

为了满足全国人文社科科研人员的研究需求，也为了弥补高校图书馆收藏的空白，CASHL 于 2008 年度开始大批购入特藏文献，保存于一个高校图书馆中，其他高校馆可通过 CASHL 文献传递的方式获取。大型特藏学科集中,有相对完整的专题.在国内至少高校范围内,具备相对的唯一性,也是没有必要在国内买多个复本的,系统性和完整性,需要在一个地方收藏的,无法拆分的,平时经费很难采购的文献,能够成为文专图书建设标志性收藏的,能够揭示,报道并为全国服务,原则上各中心馆需要提供推荐学者信息 及荐购建议.

目前 CASHL 主页上大型特藏栏目下展示的大型特藏文献，主要分布在北大，武大，复旦，川大，中山等大学图书馆，涉及图书、缩微资料、数据库等不同介质。涵盖多个人文学科，现有大型特藏 187 种，其中历史考古最多，共 105 种,哲学马列 13 种，法学 5 种，社会学 11 种，语言文字学 9 种，区域研究 13 种，文学艺术 6 种，政治军事 9 种，图书馆学 1 种。

现开世览文主页上大型特藏页面，由简介，题名字顺浏览，学科列表浏览，及题名检索

几部分组成。可检索点非常有限，读者仅能对大型特藏的总名称进行检索，无法深入检索。其次，揭示深度不够，读者只能按题名字顺和学科分类浏览，题名字顺浏览大多为读者检索时使用，学科浏览分类较粗，一些特藏资源很可能与多个学科相关，这点体现的不足。读者选定某项大型特藏后，可见到题名，出版社，资源类型，学科类别及馆藏址，目次等信息，其中大多数的目次信息是出版社提供的图片，只能逐页逐条细览，无法检索。

例如：北京大学图书馆特藏部收藏有 17 种 CASHL 大型特藏，其中数据库一套，可由网页链接进入。纸质书型文献共两套，其中一套没有目次信息。缩微格式的文献 14 种，其中 3 套没有目次，其余均为图片格式的目次。读者只有通过文献题名搜索及浏览，如需查看目次级别的信息，需查看出版商提供的简单的目次页图片，而多数文献目次信息都揭示得不深，甚至完全没有目次，读者查阅文献非常不便，据特藏部门使用统计，校内外读者对大型特藏文献请求量使用量偏低。

图表 1 北大馆藏历史类缩微平片在 CASHL 平台的揭示现状

2，对“特藏++”大型特藏揭示项目的分析与研究

近几年来，关注到大型特藏文献揭示深度不够，检索不便，利用率偏低等现状，CASHL 中心推出了“特藏++”项目，希望能够对这些特藏文献进行深度揭示，提高利用率。由收藏大型特藏的 CASHL 中心馆，组织有教师科研人员参加的团队对特藏进行内容深度挖掘，在尊重知识产权前提下，以小型数据库等各形式在线发布运行。

目前有武大，厦大，中山大学，吉大，北师大，东北师大，兰大，南京大学，南开，华东师大等高校图书馆积极参与“特藏++”项目，着手文献的进行深度揭示。目前已完成项目 6 项，在建 4 项。按资源类型分，大型丛书 4 项，缩微交卷或平片 6 项。

如武汉大学图书馆对馆藏大型丛书《日本外交文书》目次检索平台的开发，其主要的方法是：将 djvu 图片格式的目次内容利用 OCR 识别技术转化为 Excel 文件格式，以目次中的时间、事件、人物为检索词，形成目次索引检索系统。并增加了与 CASHL ILL 接口的关联，

在最终显示结果中集成了 CASHL 文献传递功能。

中山大学图书馆将馆藏《卫理公会传教士信件》缩微胶卷转化为 3 万张数字化图片，根据缩微胶卷附带的人名索引，建立了完整规范，深度挖掘内容的元数据，包括：人名、地区、时期、胶卷顺序号等，合计著录约 6000 条。并以开源平台 Piwigo 搭建特色库等等。

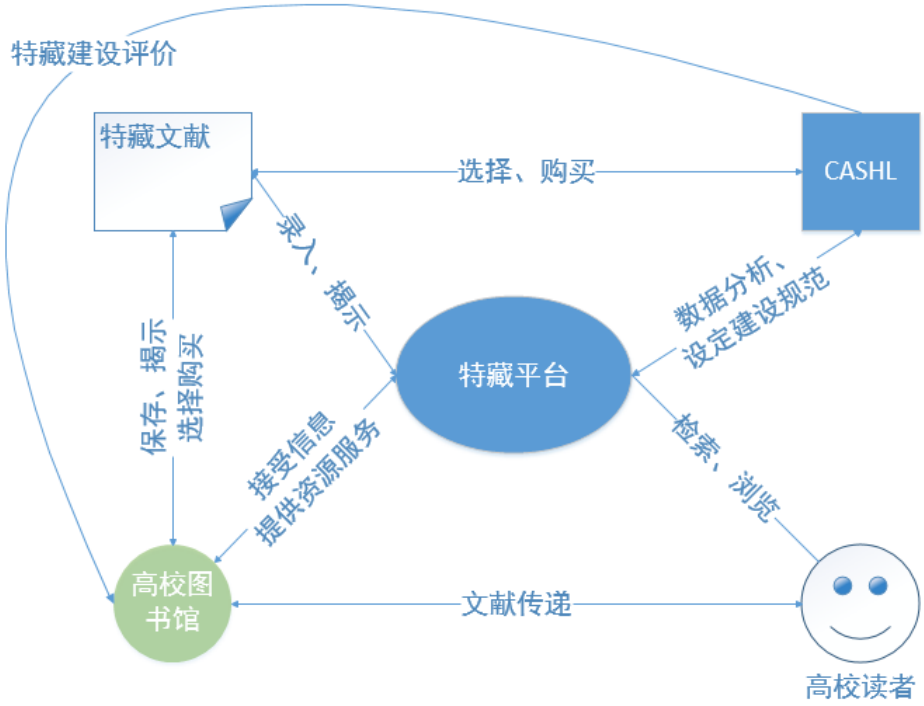
东北师范大学为《希腊罗马作家作品集》丛书开发平台，为用户提供检索功能和浏览功能，并与 CASHL 文献传递系统挂接，以便用户在需要时能够直接申请文献传递。其主要方法是：将这 415 卷丛书的书目信息按照“序号、著者、书名、编者、出版年、系统号、其他馆藏地”各项内容录入 EXCEL 表格，并对每本书的封面页、题名页、版权页和目录页进行了扫描，形成图片格式文件，将 EXCEL 和扫描图片导入 MySQL 数据库中并关联，设置检索字段，调用 CASHL 馆际互借系统接口，实现文献传递请求链接。

厦门大学图书馆对《Foreign Office files, United States of America. Series two, Vietnam》缩微胶片进行了数字化加工，并通过人工标引的方式进行了深度揭示，选用 MySQL 作为底层数据库，Loris 作为图片文件库，选用开源内容管理平台 Drupal 作为主要平台框架，自主开发 Microfile_images 模块，实现在 Drupal 平台调用 Loris 库中的图片。经加工整理后《美国外交部档案-越南》共 205 卷，1908 件档案，共计 181896 幅数字化图片全部进入深度揭示平台。平台可根据档案编号或者档案题名进行检索，全部档案图片均可在线浏览高清图片，按需自行放大缩小，浏览效果与直接在缩微胶片阅读器上的效果相同。

这些高校深入细致的工作，为大型特藏文献的深度揭示做出了有益的尝试。综合研究了这些高校馆给出的课题报告，可以看出深度揭示工作的关键是获取更深层次的详细的目次信息，和提取可检索的元数据。纸本书类，包括大部头丛书，刊等，基本可正常编目，可以揭示到最详细的目次，深度揭示难度不大。而缩微类大型特藏多是以图片格式存储的文献资料，不能直接获取文字信息，需要利用如 OCR 识别等技术手段将其转化为可检索的文字，获得目次信息，提取元数据，而后录入 Excel 表，以此为基础选择合适的平台或独立搭建平台，做成独立的数据库。数据库格式的大型特藏文献本身即是独立的可检索库了。

随着“特藏++”项目的逐年开展，将有更多的服务馆加入，虽然思路大体如上，但各馆选用的元数据格式，目次深度，搭建数据库选用的平台等等并不一致，CASHL 大型特藏文献只能以网页链接的形式展示。这对读者而言仍存在很大的不便。CASHL 管理中心有必要建立一个统一的“CASHL 特藏++平台”，把特藏资源重新整合，有序、统一揭示，利于读者检索使用，服务馆也可直接利用此平台进行统一化，标准化的大型特藏文献的深度揭示工作，获得 CASHL 相关技术支持，互相交换工作经验，免去各自选取元数据标准，设计平台等困难工作，避免重复建设。

CASHL 特藏++平台是连接与 CASHL 中心，高校图书馆，高校读者和大型特藏文献直接相关，其中，CASHL 中心设计，制定标准，建设特藏++平台，同时也可获得平台提供的数据分析与各方意见反馈，这对于逐步完善平台有重要的意义。高校图书馆是平台后台的数据提供者和平台前台的服务维护者，高校图书馆选择购买，保存特藏文献，并对特藏文献做深度的揭示，提取元数据并录入平台后台系统，并通过平台前台与读者联系，提供馆际互借与文献传递服务。读者可以在平台前台检索，浏览和提出馆际互借与文献传递申请，并可与其它读者，图书馆，CASHL 中心等相互交流沟通，同事读者的使用数据也是平台数据分析的基础，也是大型特藏资源建设评价的重要参考数据。CASHL 特藏++平台还可以开发各种新颖的交流互动，宣推等模块，充分发挥上述四者间的纽带作用，以上即平台设计的大体思路。



图表 2　CASHL 特藏平台功能预想

3，CASHL 特藏++平台调研与框架设计

平台框架设计：此平台分为后台与前台两部分，读者界面为前台，CASHL 中心与各高校馆业务在后台完成。后台为前台提供数据支持。

图表 3　CASHL 平台的前台与后台

后台主要包括元数据导入系统，特色库建设模板，多维度应用数据库，电子化文献存储，数据搜集与分析系统。

元数据导入系统：

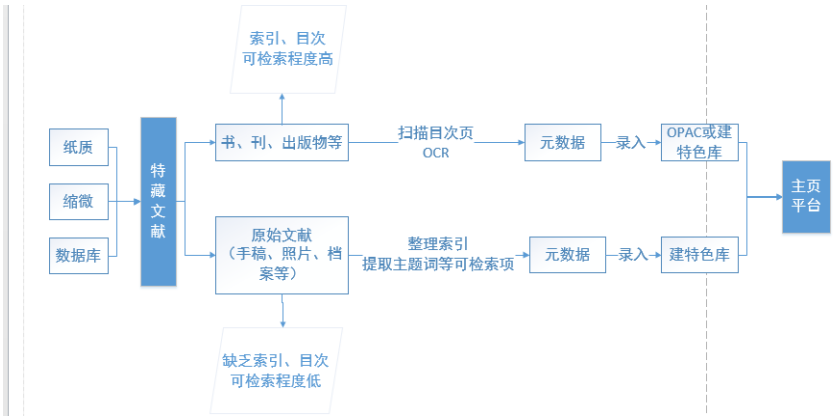
因为平台是面向 CASHL 全国成员馆的师生开放的，而当初购买大型特藏的时候是如何规定文献使用人数是要弄清楚的，尤其是数据库类型和缩微类的大型特藏文献，是否允许将缩微或纸本书等整理为数据库并开放给全国读者？而每一个大型特藏文献的元数据等信息的公开都必须符合订购时的合同规定及相关法律法规，必须具体深入的研究版权法等法律法规及每一个订购合同，这是这项课题进行下去的前提，CASHL 及高校馆相关人员必须认真对待，在全部深入了解研究之后，才能制定下一步的统一元数据标准。也可考虑 CASHL 另建一与此相关的课题或负责人专门负责这方面的事务，为之后的进行提供版权和法律保障。这里不详细展开了。

元数据的统一格式需由 CASHL 深入研究大型特藏文献的实际情况，综合研究决定。比较简明的 DC 格式元数据可以作为这次平台后台基础元数据库的初步统一标准。即：题名日期 Date，创建者 Creator，主题 Subject，出版者 Publisher，类型 Type，描述 Description，其他责任者 Contributor，格式 Format，来源 Source，权限 Rights，标识符 Identifier，语种 Language，关联 Relation，覆盖范围 Coverage 等。

根据文献载体的实际情况，及考虑到上述的合同，法规等允许条件下，各馆尽可能的深度挖掘揭示文献信息，可考虑有读者，专家等参与这项工作，深度揭示以全文电子化可检索为最终追求，此外，甚至广泛收取国内外引用涉及此文献的学术论文，出版商给的信息等各个方面信息，都可以加入供读者参考。

纸本书类大型特藏，可以由高校馆的编目或相关人员通过编目或出版社提供的相关元数据信息，整理为较为正规完整的 MARK 数据或统一格式元数据，逐步深度揭示，也可建立统一的纸本书目的数据库。数据库格式的文献，整个需要将此数据库元数据与平台的基础元数据库做好对接，提取数据库元数据，按标准元数据格式整理，导入元数据库之中。缩微类文献的揭示工作已有一些高校馆在做，方法思路入前所述，本平台中，CASHL 应提供相应的文字识别技术支持，各馆之间也可相互分享成功经验。最终将提取的标准元数据录入系统。

一些原始文献，如手稿，照片，档案等，缺乏索引，目次，CASHL 应指导各馆人工整理编辑索引，尽可能提取可检索的信息，导入元数据库，并设置一些各种文献类型的特色库模板，根据文献的客观实际建立独立的特色库。



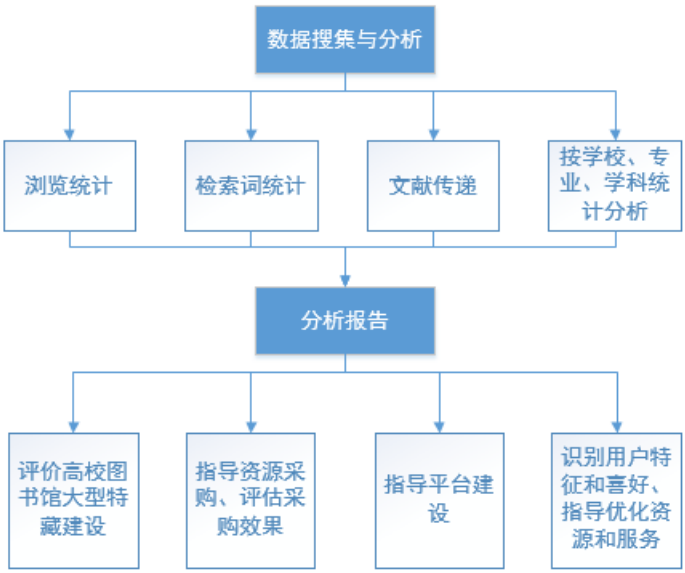
图表 4 CASHL 特藏平台后台的元数据导入

除上述纸版书，刊等文献可整理为完备的书目数据库外，还可以根据不同的知识维度，在元数据库的基础上建立应用型数据库，如地理维度上，收取所有元数据中的地理地域信息，建立地理信息数据库，在时间维度上，收取所有元数据中的年代，时间信息，建立历史年表数据库，此外还可以考虑语种，人名等等各种应用型数据库，这些库对平台前台的检索，浏览等模块提供数据支持。此外，还可以建立一个引用过这些大型特藏文献的论文库，这对深

度揭示这些文献，使读者更好的理解这些文献的价值都很有意义。

平台后台除了提供元数据库存储空间，还应有逐步转为电子化的文献存储空间，支持各馆将缩微等文献转为数据库并存储在平台上，还要有馆际文献传递上传文献的存储空间，这些传递的文献可以按大型特藏名称建库保存，不断积累至全部电子化。

数据收集与分析系统主要收集前台数据，包括读者注册信息，浏览统计，检索词，文献传递统计，学科统计等等，以此基础上，进行数据分析挖掘，对高校图书馆大型特藏资源建设与揭示程度，平台建设等做分析评估，还可识别用户特征喜好等，进行分析，推送，指导资源建设与优化资源服务。



图表 5　平台的数据收集与分析

以读者为服务对象的平台前台应界面友好，主要包括检索系统，浏览模块，宣传推广模块，互动模块等。

检索系统，可考虑分为三个模块，即一站式搜索模块，高级搜索模块，和数据库搜索模块。

一站式搜索最为便捷，对应基础库全部元数据为检索点。检索结果列表如能考虑加入对结果的相关度进行智能排序，或联想推荐则更好。另外可以加入按年代，地域，语种等分类，便于读者查找。每一条结果应该是最深入的目次一级，点击可直接进入链接（馆际互借链接）同事标有被点击次数等来自读者的数据。还可设置读者标签，即读者对此目次做批注。

高级检索可设标准格式元数据检索项之间的逻辑关系。执行检索时，先在高级检索界面上的“Search For”后的检索对话框中输入第一个检索词，接着在“Index”下面的字段下拉列表框中选择所需检索的字段；其次，在左侧的逻辑算符列表框中选择所需的逻辑算符，然后输入第二个检索词，并作字段选择，按上述方法一次输入整个检索式。通常，系统不允许超过两个“OR”运算符的检索表达式。以上步骤完成后，可点击“Start Search”图文框，系统即进行检索运行。

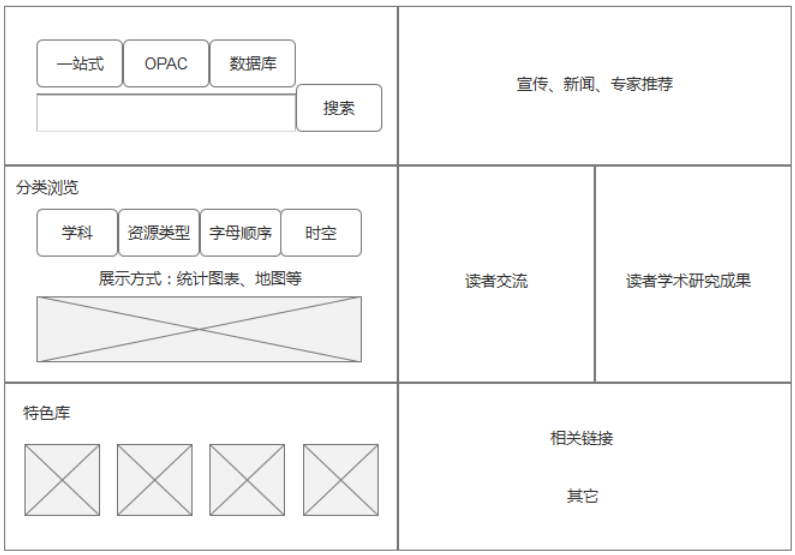
此外还可加上数据库模块，即数据库类型或各馆自己建设为数据库型的，可以在此检索，然后直接进入数据库继续查询。现在 CASHL 大型特藏平台上只有一个检索，及特藏文献名称检索，这个可以归于一站式或高级检索中的一项。

检索结果界面可以附加点击量，年代等排序，及相关学科，语种，地域等信息，方便读者从大量检索结果中找到自己所要的文献。

浏览模块设计可以按照不同维度浏览大型特藏文献，如学科，语种，年代，地域等等。与后台的多维度应用数据库相对应。一些特色数据库也可在此列出展示。

宣传推广模块为大型特藏文献做宣推及解读，引导读者使用文献。包括专家推荐意见，专家解读，新购文献介绍等等。

读者交互模块用于读者留言，评价讨论文献，纠错，推荐以及读者与参考大型特藏文献所做的学术研究成果展示等等。也可采用微信微博等媒体方式，方便与读者沟通。



图表 6 平台前台示意图

平台构想的实现计划可分为以下几步：

首先，对从 2015-2017 年的 10 个特藏++揭示项目成果进行深度研究.重点包括采购协议及版权,呈现方式,访问限制,元数据结构等.其次,整合前述成果,确定平台需求.形成平台需求报告。再次，与技术人员合作或外包,搭建平台,与文献传递系统无缝衔接,丰富宣推模块,打造交流平台，以大型特藏文献为核心，集结人文学科与计算机技术等学科的专家学者，向数字人文发展。

4， 总结与展望

CASHL 特藏++平台的建设，对于促进各馆深度揭示 CASHL 大型特藏文献提供技术保障与平台支持，整合资源，利于读者检索文献，有重要意义。今后应进一步调研读者与图书馆的需求，参考国内外相关特藏文献平台建设，逐步总结各馆深度揭示经验，收集更多读者和高校馆的使用反馈，边服务边建设，逐步完善平台建设，逐步积累文献电子化，争取文献的最有效利用。CASHL 也应争取更多资金投入及更合理的资源采购，以达到更大的学术影响力。