文章编号:1005-6033(2013)04-0104-05 收稿日期:2012-09-13

图书馆数字资源应用统计系统的设计与实现*

高广尚

(广西师范大学图书馆,广西桂林,541004)

摘 要:通过对各类数字资源的研究分析,提出了一种统计终端用户访问数据的方法;针对系统结构迥异的数字资源站点,利用 REST 和 Ajax 的技术融合,设计并实现了数据收集操作一致的图书馆数字资源应用统计系统。该系统不仅能对各站点来的访问数据实时地、智能地加以分类、汇总与保存,且本身具备良好的安全性、扩展性及维护性,在一定程度上满足了将大量缺乏联系的数据转化成有用的图表信息的要求,进而为馆购决策提供了强有力的信息支持。

关键词:数字资源;应用统计系统;系统设计;REST;Ajax

中图分类号:G250.76 **文献标识码**:A

万维网是在因特网上以超文本为基础形成的信息网,终端用户通过它查阅因特网上的信息资源。然而,终端用户经由图书馆门户网站来访问感兴趣的电子期刊、电子书等数字资源,主要是通过被提供的远程站点链接和本馆镜像站点来链接以访问深层次的数字资源。如何很好地收集到终端用户访问这些数字资源的数据,以及如何高效、智能地整合这些分散的数据,并将数据运算与美观的图表完美地嫁接在一起,便成为本文所要解决的问题。鉴于此,本文提出了基于 REST 架构及 Ajax 技术的图书馆数字资源应用统计方案,该方案不仅不更改原有数字资源站点的系统架构,并且对数字资源站点数目没有规定,而且还能将收集来的数据依业务所要求的格式统一存放到相应的数据表中,以供更快地生成图表操作用。

1 相关的实践需求分析

图书馆机房中心的机柜里安置的是部署在本馆的、不同类型的镜像数据库(或数字资源),从本质上来说,它们是彼此间互不相关的"系统孤岛"。如何有效地集成终端用户对这些数字资源访问的数据(如访问次数、文件下载次数等),进而利用、分析这些数据,为日常的决策工作、服务规划的制定提供有效的支撑,这一直是业界关心的一大问题。

为此,文献[1]提出通过把终端用户访问不同类型数字资源的数据转化成 XML 元数据,最后再转成统一格式的 XML 元数据,以此来实现数据收集目的;文献[2][3][4]通过检查 Web 日志、浏览数字资源提供商提供的数据、利用网络代理和利用脚本自行开发等方式来收集用户的访问数据。诚然,对访问统计单个资源站点来说,这或许是一种好的数据收集方式,但随着系统设计的深入开展,如在创建各类图表中需要高效、自动和快速的整合数据,这些不统一、过程繁琐的数据收集方式会令后期实现系统时具有困难,因为收集到的数据分散在各自的数字资源站点所在的服务器上,这种异域性无助于数据的自动化、智能化聚集操作时所要求的灵活性、方便性,且不易扩展要收集的数据源。但无论采用何种技术手段将这些数字资源站点的访问数据集成为有机整体,对图书馆来说,扩展并重用现有基础设施可能更符合图书馆业务日益增长的需求。

考虑到数字资源本身是基于 Web 来部署的站点,最重要的是,本文提出的图书馆数字资源应用统计方案也是基于 Web 部署的,因此,要想系统地整合这些数字资源的终端用户访问数据,就要预先对数字资源加以分类,如将数字资源分类为镜像资源或远程资源。以广西师范大学图书馆为例,数字资源主要分为三大类:

1.1 镜像数字资源

通常这种资源对外网没有依赖性,部署在本馆机房中心,主要是为了加快终端用户对资源的访问速度。为统计终端用户对这类资源站点的访问信息,以下三方面值得特别关注:一是以匿名方式监控并收集站内链接的点击流数据。二是对这些数据信息处理的基础是需要强壮的网络,因为只有这一块稳定后,才能实现后续的数据汇聚操作。三是采用易扩展的 Web 服务架构来进行规划、设计。只有这三部分紧密结合,数字资源应用统计的目标才能最终得以实现。该类数字资源的应用统计是本文讨论的重点。

1.2 自建特色数据库数字资源

同样部署在本馆机房,其数据收集方式与镜像数字资源类似。它集中了高校独有的数字化特色文献资源,体现在:一是以某重点学科或某特定专题、或具有交叉学科和前沿学科、或能体现高等教育特色的资源;二是具有一定的地域和历史人文特色,

*基金项目:广西人文社会科学发展研究中心资助项目(No:ZX2011044)。

或与地方的政治、经济和文化发展密切相关的资源；三是具有他馆、他校所不具备或只有少数馆具备的特色馆藏，或散在各处、难以被利用的资源等。

1.3　远程数字资源

由于它是以远程 IP 地址（或结合用户名/密码的组合形式）来访问，所以针对这类站点所进行的访问数据收集，只能在图书馆门户网站提供的远程站点链接处进行，即不能对远程资源站点内的页面进行链接监控，这与镜像数字资源的数据收集操作不同。

它们在图书馆门户网站中的显示情况如图 1 所示，图中还给出了该网页背后的链接元素（a 元素）代码，通过链接元素中的 href 值能容易判断出此链接是远程链接还是镜像链接。

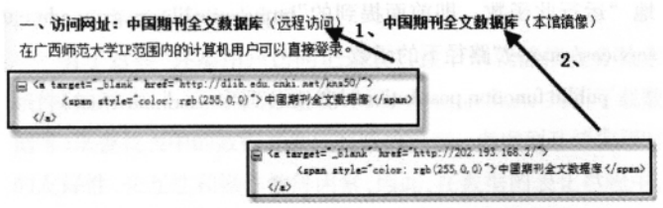

图 1　图书馆门户网站提供的远程访问和本馆镜像链接

2　系统设计与实现

2.1　系统设计

系统设计是实现所有需求的创造性过程，是在着手编写代码前，把系统分解成几个易管控的子系统或模块的过程[5]。在 Internet/Intranet 领域里，B/S 结构在现代软件系统的开发过程中被优先采用，数字资源站点本身就是由 B/S 结构搭建而成。该结构最大的优点就是可以在任何地方用浏览器操作而不用安装任何专门的软件，只要有一台能上网的电脑就能使用，客户端零维护。当然，软件发布和维护的工作不是自动停止了，而是转移到了 Web 服务器端。本文提出的图书馆数字资源应用统计系统是围绕 B/S 结构并基于 REST 软件架构及 Ajax 技术来设计的[6]，如图 2 所示，其主要涉及以下 3 个子模块。

第一是数据收集。正如前文所分析的那样，针对镜像资源和远程资源的数据收集工作，其所采用的方式会有所不同，但技术实现原理本质上是一样的，即链接元素动态都绑定相应的事件函数。考虑到本系统是部署在与图书馆门户网站同域的服务器内的情况，因此，统计域外的镜像资源的访问可以检查其站内众多页面上的众多链接（如统计下载次数），而统计域内的远程资源的访问只能检查图书馆提供的远程资源链接（如统计访问次数），但它们都让包含有正在访问的数字资源“标志码”数据传送到 REST 架构服务端，这是区别于其他方法的关键所在。

第二是数据入库。在 REST 架构服务端，系统能根据传送来的“标志码”把数据写入对应的数据表中。如此一来，既满足了不同“标志码”对应不同数据表的系统可扩展性要求，及数据要集中存放到同一数据库的要求，又避免了由一个表来在线保存累积的大量访问数据（毫无疑问，这在一定程度上会提升查找速度），甚至，还保障了数据入库前的操作安全性、有效性。

第三是数据显示。对收集到的相关数据，基于业务逻辑的检索条件系统智能地、实时地显示出有价值的信息图表。

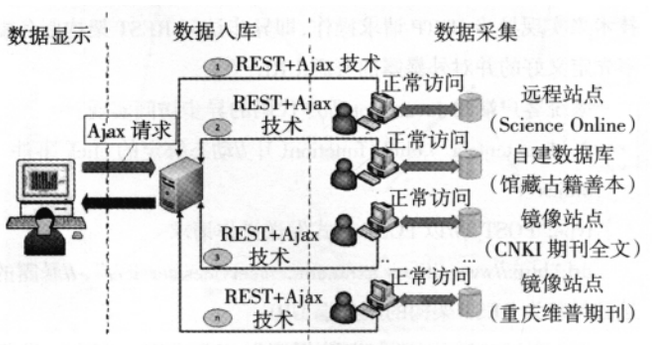

图 2　图书馆数字资源应用统计系统的系统架构

2.2　模块实现

在各个模块被划分好后，接下来就要对各模块内、模块间的交互进行详细的设计。对数据收集模块，利用 JavaScript 浏览器脚本语言实现对终端用户行为数据的收集；对数据入库模块，利用 PHP 语言来对传送来的数据执行有效性判断，并执行数据入库操作；对数据显示模块，利用 ExtJS 程序来动态生成图表界面。

2.2.1　数据收集模块的实现

与统计远程资源的访问相比，统计镜像资源的访问要复杂得多，下文重点讨论如何对镜像资源的访问情况进行统计。由于图书馆的镜像资源可能有几十个，其中包含资源链接的页面，数量巨大。如图 3 所示，该页面显示终端用户正在访问“朱镕基答记者问”文档，其中包含两个重要的“CAJ 下载”和“PDF 下载”链接，当它们被点击时，就表示终端用户想下载此文件。更为复杂的是，这些镜像资源本身是由不同的程序语言开发实现的（如 Java、.Net 和 PHP 语言等）。如何在不影响终端用户使用的情况下，高效、有序地实现监控网页上匹配的链接元素绑定一个事件处理函数的目标，是收集数据模块部分应先解决的问题，同时也是实现 REST 架构客户端的前提。很显然，对分布在数十个镜像资源中的众多链接元素，不易一个个通过静态绑定事件处理函数来实现 REST 架构客户端，更何况有些链接元素是动态产生的。在这种情况下，利用 jQuery 的动态绑定技术就能轻松解决这个问题，它提供的 bind 和 live 方法就能容易地对页面上的链接元素进行动态绑定[7]，这种技术不仅不更改原有镜像资源的页面结构，而且还解决了因语言开发的不同而造成终端用户访问数据难统一整合的问题。

图 3　包含文档链接的页面

在链接被动态绑定后，比如已动态绑定图 3 中“PDF 下载”链接，当终端用户点击该链接元素时，会瞬间触发两个操作：一是正常访问某资源的 HTTP 请求，即下载该文档；二是发起针对

REST 架构服务端新的 HTTP 请求，新的 HTTP 请求操作是在动态绑定的事件处理函数中实现的。为不希望因新的 HTTP 请求操作导致终端用户浏览器挂起的假死现象出现，这里采用了 Ajax 技术来实现异步 HTTP 请求操作，即异步访问 REST 架构服务端事先定义好的并对外暴露的资源 URL[8,9]。

系统客户端的 Javascript 程序代码的异步访问实现：

```
$('#contents a').click(function(){//动态绑定的 click 事件
$.ajax({
type:"POST",//以 POST 方式发送操作请求
url:"http://www.library.gxnu.edu.cn/services/service/", //暴露的且已封装为 REST 架构的服务端 URL
data:({database: "中文数据库",subDatabase:"CNKI 数据库",local: "是",resourceName: "朱镕基答记者问.pdf"}),//通过 JSON 对象传送数据
cash: false,
async: true //异步发送 HTTP 请求
})
})
```

在上述代码中，data 属性值表示在客户端收集到的用户访问数据，其中 resourceName 字段中的值"朱镕基答记者问.pdf"是动态获取到的，url 属性值表示向外暴露的 REST 架构服务端 URL，这个 URL 是整个数据收集操作的核心，同时也是 REST 架构的精髓部分，通过此 URL 数据才能使数据统一在线保存在 REST 架构服务端的数据库中。最后，页面上特定链接动态绑定事件发生后，用户在点击该链接的同时就实现了数据的传递操作，这是在不影响用户使用的前提下静默进行的。对于其他镜像资源站点的数据收集，依照此方法嵌入上述 Javascript 代码，并依据预先编制好的分类表（如表 1 所示）来更改相应的 database，subDatabase 属性值，就实现了对多个不同数字资源的用户访问数据进行收集的功能。

表 1　数字资源的分类及编号

数字资源大类	数字资源大类编号	数字资源小类	数字资源小类编号	是否镜像
中文数据库	1	CNKI 期刊	1	1
中文数据库	1	读秀学术搜索	2	1
…	…	…	…	…
外文数据库	2	IEL	1	2
外文数据库	2	Elsevier SD	2	2
…	…	…	…	…
试用数据库	3	北大法意数据库	1	2
…	…	…	…	…

2.2.2　数据入库模块实现

当各数字资源站点的众多链接被访问时，从各数字资源站点传来的 data 属性值不同的 JSON 对象数据，不请自来地传送到了图 2 中的 REST 架构服务端，在收到客户端以 POST 方式发送的 HTTP 请求后，REST 架构服务端都会理解为"创建一个新资源"（这种智能化的理解正是基于 HTTP 协议的 REST 架构的精髓），即依据 JSON 对象中的 database 字段值将 JSON 对象中的数据统一存储到服务端的相应数据表中，也就是记录一次该终端用户的访问行为，这些表就是"终端用户访问记录表"。为了表述的需要，在表 2 中记录了 2 个终端用户的访问情况，不难看出表中数据其实就是传递过来的 JSON 对象中的数据。

表 2　终端用户访问记录表

id	ip	database	subDatabase	resourceName	local	Datetime
1	202.193.168.11	中文数据库	CNKI 数据库	朱镕基答记者问.pdf	是	2011/11/22
2	202.193.168.12	中文数据库	人大复印资料	论当前国际形势.pdf	是	2011/11/22

系统服务端的 PHP 程序代码实现：

```
//一收到以 POST 方式发送的新 HTTP 请求，服务端会"智能地"运行此函数，即前面提到的"http://www.library.gxnu.edu.cn/services/service/"路径下的函数
public function postAction() ?
{
/*
```

为满足系统的安全性、动态扩展性及可维护性需求，程序在这里通过获取客户端传来的值来判断客户端请求规则的有效性。如通过$this-> getRequest() -> getPost('database') 的值来调用相应的"终端用户访问记录表"供程序保存数据用，假设接收到的值是"外文数据库"，程序就调用相应的"外文数据库"的"终端用户访问记录表"对象[10]，其他情况类推。以下给出的是访问"中文数据库"这一大类的情况。

```
*/
$ip=$_SERVER["REMOTE_ADDR"];//获取客户端的 IP 地址
$Item=new Gxsd_Model_ZhongWenVisitedDetails(); //调用相应的"中文数据库"的"终端用户访问记录表"对象
$Item->fromArray($this->getRequest()->getPost()); //获取客户端传送来的 JSON 对象数据
$Item->visit_date=date("Y-m-d H:i:s");//记录操作发生的时间
$Item->visit_addr=$ip; //记录访问终端用户的 IP 地址
$Item->save(); //向数据表增加一笔终端用户访问记录
}
```

随着时间的推移，这些"终端用户访问记录表"就会逐渐产生大量数据集，甚至产生大量的冗余数据。考虑到这种状况会在日后出现，从数据精简及查询速度优化的角度出发，这里仅让 JSON 对象中的 database，subDatabase 和 local 字段值取整数值来表示各数字资源的大类和小类类别及是否属镜像资源类，并且这些值对应表 1 中预先设定的整数值。通过这种方式，客户端传来的优化后的 JSON 对象数据就会被有效地归档并在线保存于数据表中，如表 3 所示。当然，我们还能依业务要求收集其他"感兴趣"的数据，如增加对该数字资源所属的"学科领域"数据的收

集等，且这种额外的增加操作并不会对本文提出的系统架构产生影响。

表 3　优化后的终端用户访问记录表

	id	database	subDatabase	resourceName	local	visit_date	visit_addr
	2336	1	1	试论图书馆数字资源整合.pdf	1	2012-04-07 11:33:17	172.30.49.54
	2339	1	1	高校图书馆数字资源评价指标体系研究.pdf	1	2012-04-07 11:34:38	172.23.6.232
	2341	1	1	谈高校馆藏外文电子资源的利用.pdf	1	2012-04-07 11:35:58	172.20.32.74
	2347	1	1	高校图书馆电子资源体系的构建战略.pdf	1	2012-04-07 11:38:46	202.193.170.72
	2393	1	1	电子资源使用统计的现状分析.pdf	1	2012-04-07 12:10:28	172.19.25.247
	2452	1	1	CALIS引进电子资源地区分布研究.pdf	1	2012-04-07 13:02:14	172.19.76.83

对于远程数字资源访问次数的统计来说，如前文所分析的那样，系统只能在如图 1 所显示的远程链接处进行点击次数的统计。很显然，这无法收集到远程数字资源站内下载的次数等数据信息，也就无法进行相应的应用统计，在“终端用户访问记录表”中的体现就是 resourceName 字段没有数据。

2.2.3　数据显示模块的实现

有了这些“终端用户访问记录表”，接下来就能依业务检索条件（如对 database 字段分组统计，对 resourceName 判断下载数据等）来查找表中的数据并生成直观的图表。考虑到系统使用时的友好性、交互性和操作性等因素，因此，在数据图表化过程中，系统利用 ExtJS 程序并基于 Ajax 技术来获取“终端用户访问记录表”中的数据来生成无刷新的、互动的信息图表。这里获取访问数据的方法与上文谈及的文献中提出的方法不同，因为这里的数据都已自动地汇聚在同一数据库中，且并不零乱地分散在其他域中的某个地方。借助图表我们能更形象、直观地对比不同时段访问不同数据库的信息，甚至查看单个数据库的下载情况以及其他更多类似数据。图 4 是“中文数据库”这一大类中各子库的下载次数统计图，从图中易得知“CNKI 期刊”数字资源在 4 月份的下载量远比同类的其他数字资源的下载量大；图 5 是广西师范大学图书馆各数据库在 2011 年度总的访问情况统计图，从图中可以清晰看出最上面那条折线表示“中文数据库”的各月访问量。该库在 2011 年的 3 月份访问量达到 49 879 次，这反映出终端用户对“中文数据库”的使用频繁，最下面的那条折线表示“开放资源”的使用趋势，可以看出其总体用量非常低，这说明我们收集的“开放资源”并没有引起用户强烈的兴趣。当然，依业务需要，系统还可以按季度、月度、IP、学科领域和关注度等参数绘制出更详细的数据信息统计图。

图 4　中文数据库中各子库的下载次数统计

3　结语

本文设计了图书馆数字资源应用统计系统，通过系统生成的

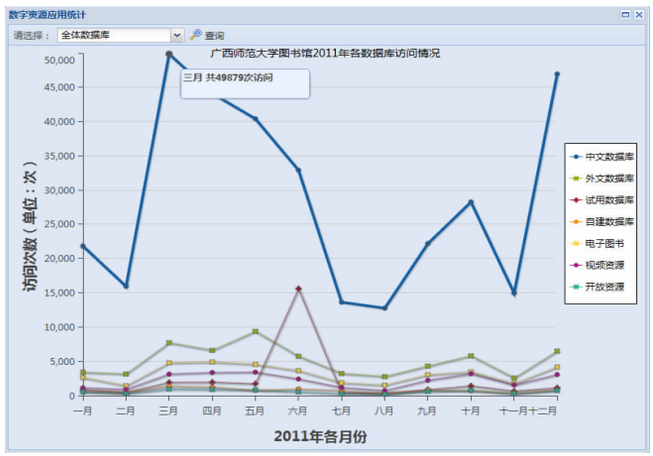

图 5　广西师范大学图书馆各数据库在 2011 年度总的访问情况统计

直观图表，我们能很好地了解各种数字资源的使用趋势，借助这些趋势图表来传达存在于数据中的基本信息，以向馆领导或工作人员提供跨业务、跨系统的主题数据分析及决策支持。该系统从根本上解决了图书馆数字资源应用数据难以分析利用的问题。在当今基于 Web 的数据收集过程中，该系统的设计无疑对业界的数据收集工作具有一定程度上的参考、研究价值。

参考文献

[1]　谢靖，马自卫.基于 WebService 的数字资源集成与服务平台的研究与实现 [J] .现代图书情报技术，2008 (11) :7-12.

[2]　徐革. 大学图书馆电子资源利用统计数据的获得模式评析 [J] .大学图书馆学报，2007 (1) :54-58.

[3]　陈陶，夏立娟，马克芬.图书馆电子资源利用统计与分析 [J] .图书情报工作，2005 (4) :92-95.

[4]　马建霞. 图书馆数字资源访问统计研究 [J] . 图书馆杂志，2005 (8) :25-29.

[5]　Pfleeger S L. Software Engineering Theory and Practice [M] . [S.l.] : Prentice Hall, 2010: 141-290.

[6]　Song J Y, Wei J, Wan S C, et al.Extending Interactive Web Services for Improving Presentation Level Integration in Web Portals [J] . Journal of Computer Science and Technology, 2006, 21 (4) :620-630.

[7]　Bear Bibeault Y K. jQuery in Action [M] . [S.l.] : Manning Publications Co, 2010: 235-278.

[8]　Richardson L, Ruby S. Restful web services [M] . [S.l.] : O'Reilly Media, 2007: 143-166.

[9]　Vaswani V. Zend framework: A beginner's guide [M] . [S.l.] : McGraw-Hill, 2010: 332-342.

[10]　Wage J H, Vesterinen K. Doctrine ORM for PHP [M] . [S.l.] : Sensio SA, 2009: 112-113.

（实习编辑 杨　昆）

第一作者简介：高广尚，男，1978 年 11 月生，2009 年毕业于桂林电子科技大学（硕士），助理工程师，广西师范大学图书馆，广西壮族自治区桂林市育才路 15 号，541004.

文章编号:1005-6033(2013)04-0108-03 收稿日期:2012-09-11

数字图书馆信息推荐引擎初探

王 旭,马慧娟

(中国电子科技集团公司第 38 所,安徽合肥,230088)

摘 要:现有的搜索技术忽略了用户的个性差异,使用户被大量无关信息干扰而无法在图书馆资源中准确选择其需要的信息资源。信息推荐引擎以用户需求为导向,主动为用户推荐其可能感兴趣的信息资源,大大缓解了信息快速增长带来的负面影响。对数字图书馆信息推荐引擎进行了探讨,简要介绍了数字图书馆信息推荐引擎的分类和工作机制。

关键词:数字图书馆;信息推荐引擎;工作机制

中图分类号:G250.76 **文献标识码**:A

随着数字图书馆内数字信息资源的丰富以及各种信息工具的普及,人们似乎能够更迅速更方便地获取目标信息,可事实并非如此。越来越多的信息用户发现信息资源丰富了、信息获取工具更先进了,同时新的困难也出现了。人们无法快速地从搜索到的海量结果中甄别真正所需的信息。随着信息过载问题的尖锐化,数字图书馆研究人员开始意识到,单纯的搜索引擎并不能完全满足用户的信息查找需求,人们需要一种更完善的、更个性化的服务来缓和数字图书馆信息过载导致的信息迷失问题,这种新型的信息服务就是信息推荐。信息推荐这种服务在实际应用中是通过推荐引擎来实现的。下文从分类、机制两个方面着重对推荐引擎进行介绍。

1 推荐引擎

数字图书馆中存储着上亿条信息资源,按照某一关键词或检索式进行检索时可能会出现成百上千条结果,用户很难通过屏幕就了解到所有的检索结果,更无法了解信息与自身需求的匹配度。因此数字图书馆用户需要一种新型的信息推荐服务工具来帮助其查询可能感兴趣或满意的信息资源。推荐引擎就是实现这种服务的具体工具,推荐引擎按照数据挖掘和过滤技术,通过挖掘用户属性和行为等源数据,分析其需求偏好并据此对相关的信息资源进行过滤,为用户提供其感兴趣的资源原文或链接地址。数字图书馆推荐引擎工作机制见图 1。

2 数字图书馆推荐引擎分类

数字图书馆推荐引擎按照不同标准可分类,本节将数字图书馆推荐引擎按照推荐结果的个性化程度、源数据的种类、推荐模型的建立方式等方面分类介绍。

2.1 按照推荐结果的个性化程度分类

按照推荐结果的个性化程度可将数字图书馆推荐引擎分为大众化推荐引擎和个性化推荐引擎。前者指推荐引擎推荐结果

The Design and Implementation of the Application Statistics System for Library's Digital Resources

GAO Guang-shang

ABSTRACT: This paper puts forward a method for statistical analysis on end users' access data through studying and analyzing various types of digital resources, and in the light of digital resource sites with different system architecture and by using the technical integration of REST and Ajax, designs and implements the application statistics system for library's digital resources with consistent operation in data collection, which can not only make real-time and intelligent classification, summary and preservation of access data from each site, but also meet the requirements to some degree for translating the abundant unrelated data into useful charts information because of the good security, scalability and maintainability of the system, and thus provide strong information support for library's purchasing decisions.

KEY WORDS: digital resources; application statistics system; system design; REST; Ajax