

数字图书馆应用 如何选择开源软件

孙卫
国家科学技术信息研究所 顾问
北京万方数据技术研究院 总工程师
2009年3月 北京



明确利用开源软件的目的

- ❖ 研究数字图书馆软件
 - (1) 解剖开源软件，在架构、结合部的了解与研究
 - (2) 构造自己的数字图书馆架构、接口、标准、规范
 - (3) 开发数字图书馆软件并开源使用
- ❖ 学习数字图书馆软件
 - (1) 学习开源软件的操作、使用
 - (2) 掌握数字图书馆业务流程
- ❖ 使用数字图书馆软件
 - (1) 正式使用开源软件用于本单位的数字图书馆应用
 - (2) 业务流程、数据加载、操作维护



几个目的面对人员的差别

- ❖ 研究数字图书馆软件
以软件开发人员为主，业务人员为辅
- ❖ 学习数字图书馆软件
以业务人员了解操作、学习技术，熟练业务为主
- ❖ 使用数字图书馆软件
以图书馆业务需要为主，结合业务与软件人员



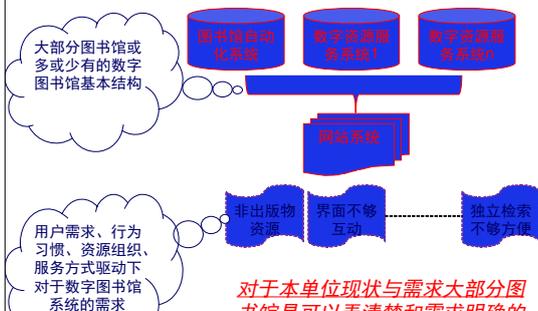
为使用数字图书馆软件而选择开放源代码

- ❖ 对于需求的了解
 - (1) 本单位现有系统状态（系统、系统接口、服务方式、性能、稳定性）
 - (2) 本单位现有（非系统）数据资源状态（数据类型、数据格式、数据结构、数据使用策略等）
 - (3) 新增系统部分与原有系统关系
 - (4) 业务流程
 - (5) 使用方法

对于本单位要利用开源软件构造数字图书馆的技术角度，要从软件工程的角度去准备

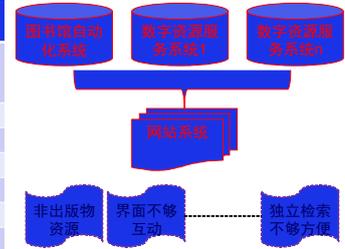


举例



举例

系统接口	接口传递函数	传递函数中的数据封装
自动化系统		
知网		
维普		
万方		
方正		
超星		
龙源		



绝大部分图书馆由于不清楚这两个部分，造成需求、流程、场景等，造成实现时不理想



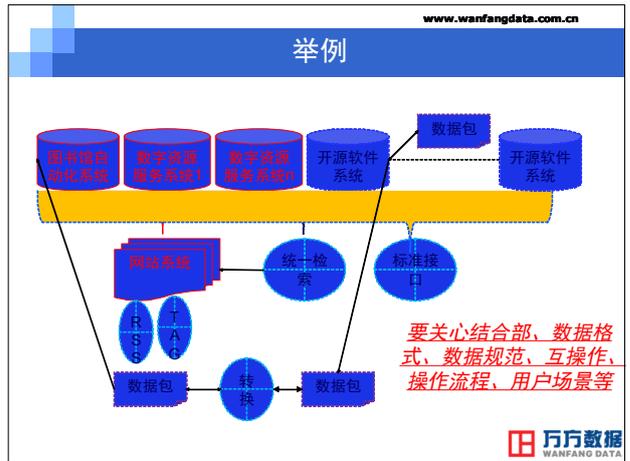
www.wanfangdata.com.cn

为使用数字图书馆软件而选择开放源代码

❖ 确定新的开源软件与原有系统的关系

- (1) 保留原有系统，新的开源软件作为子系统，通过网页功能划分进行链接、通过统一检索进行链接、通过标准接口进行链接（互操作方式）
- (2) 在资源层交换数据、在标准化接口下交换数据、在资源格式统一下组织数据、在传递函数规范下进行处理（交换方式）
- (3) 针对原有系统的服务多样性的改变
- (4) 替换原有系统

万方数据
WANFANG DATA



www.wanfangdata.com.cn

为使用数字图书馆软件而选择开放源代码

❖ 功能、性能、稳定性、人员能力

- (1) 根据需求明确了功能、流程
- (2) 根据使用者（图书馆人、用户）的使用习惯、服务类型需要的最低性能要求
- (3) 根据服务的重要性等规定的稳定性要求
- (4) 根据需求、流程、互操作、互交换等确定自己单位的人力资源能否满足选择开源软件？
- (5) 根据结合部、转换等确定自己单位的人力资源是否满足对这些软件进行二次开发？

万方数据
WANFANG DATA

www.wanfangdata.com.cn

举例

功能	流程	场景	接口	功能	数据量	操作响应时间	并发
提交数据			OA1 PMH	提交数据	100G	平均5秒	10
统一检索				统一检索	500G	平均2秒	1000

功能	稳定性	备份	恢复时间
提交数据	8小时	离线	10分钟
统一检索	7*24小时	在线	10分钟

OSS	基本要求	进一步要求	风险
DSPAC	安装、使用	网页设计	
FEDORA	安装、二次开发	集成其他OSS	

万方数据
WANFANG DATA

www.wanfangdata.com.cn

为使用数字图书馆软件而选择开放源代码

❖ 开源软件基本类型

- (1) 非行业应用目的，所有开源软件环境属于这类开源软件
操作系统、数据库等
- (2) 行业中共性应用目的，需要二次开发面向行业应用的开源软件
内容管理系统、检索系统、门户系统等
- (3) 行业专用软件，基本不需要开发就能使用开源软件
DSPACE等

万方数据
WANFANG DATA

www.wanfangdata.com.cn

为使用数字图书馆软件而选择开放源代码

❖ 需要用软件工程进行二次开发的开源软件人员的基本要求

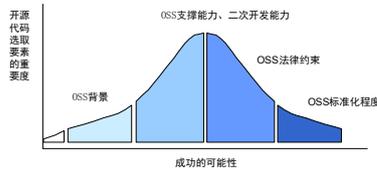
- (1) 要具备相当能力的软件工程师
- (2) 对于源代码的结构熟悉，对于数据的安装熟悉
- (3) 对于某一类应用的需求稳定，对于基本的开发具备能力
- (4) 要获得企业的服务和社区的支持
- (5) 对于性能、稳定性改善有较高的经验

用软件工程师的角度要求利用基础开源软件进行数字图书馆应用设计的软件开发人员

万方数据
WANFANG DATA

OSS的可适用性、稳定性关键点

- ❖ 开源软件本身的设计背景（流程、功能、场景、使用对象）
- ❖ 开源软件提供者对于开源软件的支持程度和能力
- ❖ 要应用开源软件的机构的自身能力
- ❖ 开源软件的标准化能力



OSS支撑能力

- ❖ oss社区
 - 大部分开源软件包是在研究与发展之后，依靠企业进行支撑的。
 - 企业在对开源软件的资金支持、发行版本的支持、社区的可持续发展的支持是支撑的重要要素。
- ❖ 支撑的能力
 - oss讨论社区（对于问题的讨论）。
 - oss版本的更新计划，实际版本变更情况。
 - 开源开发环境的更新与版本变更。
 - 案例

<http://www.apache.org/>

The Apache Software Foundation

<http://www.dspace.org/>

二次开发能力

- ❖ OSS的消理解能力
 - 对于要采用的oss很好的在流程、功能、操作层面的理解
 - 对于oss的核心、接口、传递函数的理解
 - 对于结构化oss的源代码的理解，特别是对于关键函数、关键限制的理解
- ❖ OSS的二次开发
 - 开发环境、oss针对二次的结合点
 - 互操作、互交换、页面层面的开发
 - oss嵌入oss的开发

举例

- ❖ Lucene 检索软件 **能进行二次开发的人员的基本要求**
某公司招聘Lucene开发工程师的要求：
职位描述：
 1. 充分理解产品需求，进行产品分析，搭建系统框架，保证公司产品产品的质量；
 2. 完成公司产品的开发工作。职位要求：
 1. 大学本科以上学历，两年以上Java开发的经验，熟练掌握Eclipse开发环境；
 2. 熟练掌握数据结构、常用算法
 3. 熟悉 Lucene 程序开发，一年以上的 Lucene 项目开发经验；
 4. 能熟练的编写软件开发文档和良好的编码基础和习惯；
 5. 有良好的沟通能力；

举例（续）

- ❖ Lucene 检索软件
 - (1) 改进中文切分词——替换原来的部分——对于检索性能的影响？
 - (2) 内存索引如何使用——对于索引和检索的影响是什么？
 - (3) 并行索引的解决？
 - (4) 大字符集？
 - (5) 大并发？ **尽可能选择产品化的OSS，如果选择的OSS都要进行二次开发，是有相当大的难度。**
 - (6) 超大数据集？

OSS的标准化

- ❖ W3C的标准
很容易满足，因为软件一般从开发环境产生的。绝对不依赖开发环境开发的，不容易满足W3C的标准。
- ❖ 图书馆的标准
和图书馆合作开发的，遵循图书馆的标准。
为了图书馆应用为目的开发的，遵循图书馆的标准。
- ❖ 非标准化
 - (1) 数据转换程序，使得交换的数据标准化、规范化
 - (2) 利用API进行互操作的转换

OSS原始开发目的

- ❖ IOSSPL (*Integrated Open Source System for Public Libraries*)
- ❖ DSPACE
- ❖ Lucene
- ❖ MySQL
- ○ ○ ○ ○ ○

重点关注什么？

- ❖ OSS开放软件产品——基本不开发可以直接使用
- ❖ 开放源代码软件——一定需要开发才能使用
- ❖ 小的各类型开源软件，使用在客户端为主
- ❖ 关注服务器端的OSS，在功能、性能、稳定性、二次开发可掌握程度、数据加工的工作量（难度、完整性、周期、质量）
- ❖ 关注同行、使用OSS的问题讨论、少走弯路
- ❖ 完善文档（一般的OSS文档简练，图书馆使用时要完善各种文档，使开源软件变成可管理、可维护）
- ❖ 区分软件工程师和软件工程工程师的差别

馆长们关心什么？

- ❖ 尽可能听取同行关于OSS应用的经验与教训，判断OSS在本单位使用的基本条件是否具备
- ❖ 在有条件的情况下，尽可能与软件专业更熟悉的人合作进行OSS应用
- ❖ 不要对于OSS的性能、稳定性有过高的期望
- ❖ 在使用OSS时，系统集成时对知识产权更多的关注，数据结构、数据源、应用开发接口、传递函数等是开发时容易产生知识产权问题的要点
- ❖ 在进行数字图书馆招标项目上时，需要在招标书上，明确工程使用的OSS名称与数量，防止应标单位用开源软件，但是在计价时按照自己开发计价

谢谢！

孙卫 13801158729
SUNWEI@GMAIL.COM