上海图书馆　开放数据应用开发竞赛　2018

上海图书馆开放数据——模式、内容及结构

上海图书馆高级工程师　夏翠娟

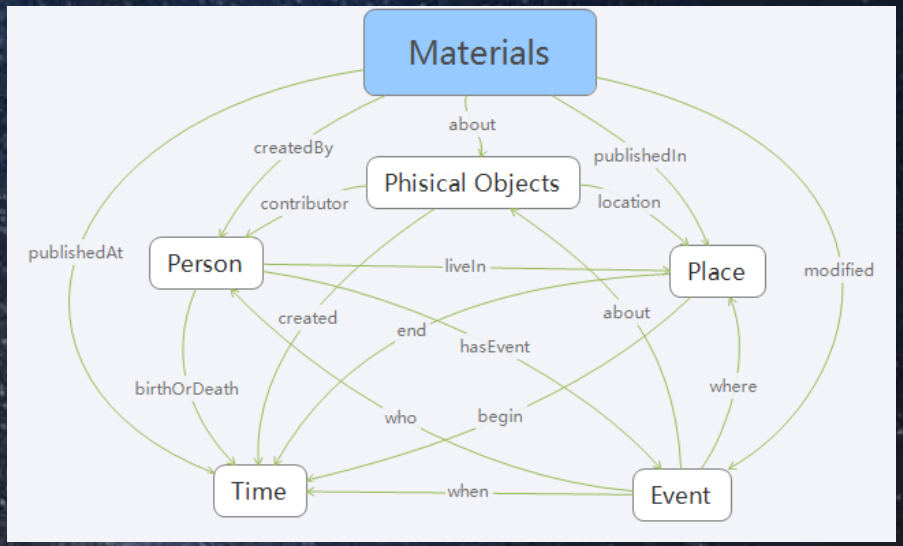
2018. 5. 23

背景—数字人文

从数字化到数据化
从文献服务到知识服务

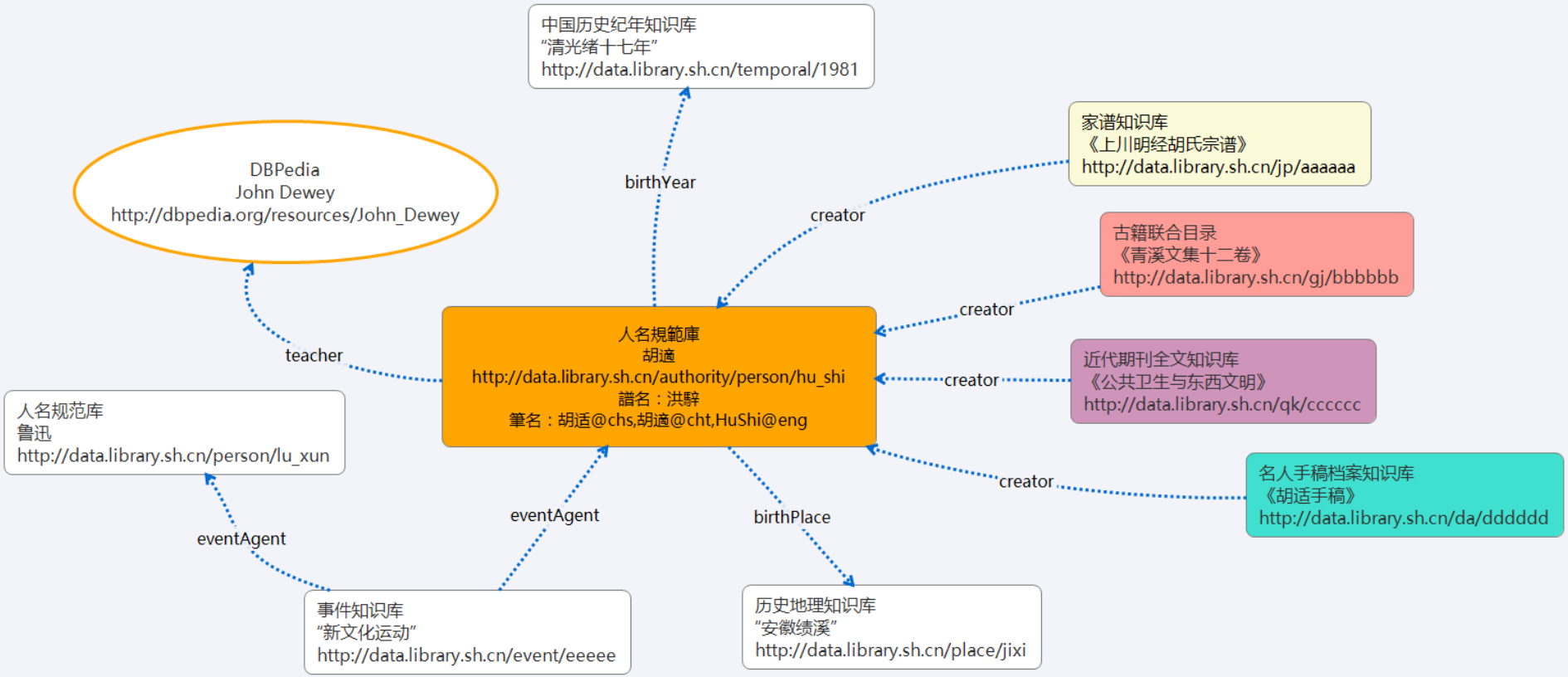
知识编码的形式化——机器可读
知识单元的细粒度化——机器可计算
知识表示的语义化——机器可理解
知识组织的关联化——机器可推理
知识增长的自动化——机器可自学习

内容组织框架



建设人、地、时、事、物基本知识库和文献知识库
成为面向人文研究的数据基础设施的一部分

举例


中国历史纪年知识库
"清光绪十七年"
http://data.library.sh.cn/temporal/1981
DBPedia
John Dewey
http://dbpedia.org/resources/John_Dewey
birthYear
家谱知识库
《上川明经胡氏宗谱》
http://data.library.sh.cn/jp/aaaaaa
creator
古籍联合目录
《青溪文集十二卷》
http://data.library.sh.cn/gj/bbbbbb
creator
人名規範庫
胡適
http://data.library.sh.cn/authority/person/hu_shi
譜名：洪騂
筆名：胡适@chs,胡適@cht,HuShi@eng
teacher
creator
近代期刊全文知识库
《公共卫生与东西文明》
http://data.library.sh.cn/qk/cccccc
人名规范库
鲁迅
http://data.library.sh.cn/person/lu_xun
creator
名人手稿档案知识库
《胡适手稿》
http://data.library.sh.cn/da/dddddd
eventAgent
eventAgent
birthPlace
事件知识库
"新文化运动"
http://data.library.sh.cn/event/eeeee
历史地理知识库
"安徽绩溪"
http://data.library.sh.cn/place/jixi

进展

中文古籍聯合目錄及循證平台
Chinese Ancient Books Union Catalogue and Evidence-based Platform
http://gj.library.sh.cn

上海历史文化时空再造
——从武康路出发
http://wkl.library.sh.cn

上海图书馆
名人手稿档案库
http://sg.library.sh.cn

上海市文献联合编目中心
关联书目数据发布平台
http://bib.library.sh.cn

家譜 知識服務平台
Genealogy knowledge service platform
http://jiapu.library.sh.cn

盛宣怀档案知識庫
http://sd.library.sh.cn

规范控制

Sparql Endpoint+ API +HTTP URI

http://names.library.sh.cn

地 时 人名规范库 马路 建筑 事件

http://data.library.sh.cn as **Data Infrastructure**

开放数据竞赛2018

中文古籍聯合目錄及循證平台
Chinese Ancient Books Union Catalogue and Evidence-based Platform

http://gj.library.sh.cn

上海市文献联合编目中心
关联书目数据发布平台

http://bib.library.sh.cn

盛宣怀档案知識庫

http://sd.library.sh.cn

家譜 知識服務平台
Genealogy knowledge service platform

http://jiapu.library.sh.cn

上海图书馆
名人手稿档案库

http://sg.library.sh.cn

规范
控制

地

人名
规范库

时

开放数据的内容及数量

文献知识库

家谱：60296

盛档：178844

手稿：70239

古籍：作品364853，版本1147445，注释53955

上海市文献联编中心书目数据：1466173 种

基础知识库

地
1855

人名
规范库

时
秦～民国

古籍开放数据内容

联合目录
(1400+ 家机构1147445种)

上图馆藏古籍目录
(138022种)

历代古籍目录
（12部53955种）

中文古籍聯合目錄及循證平台
Chinese Ancient Books Union Catalogue and Evidence-based Platform

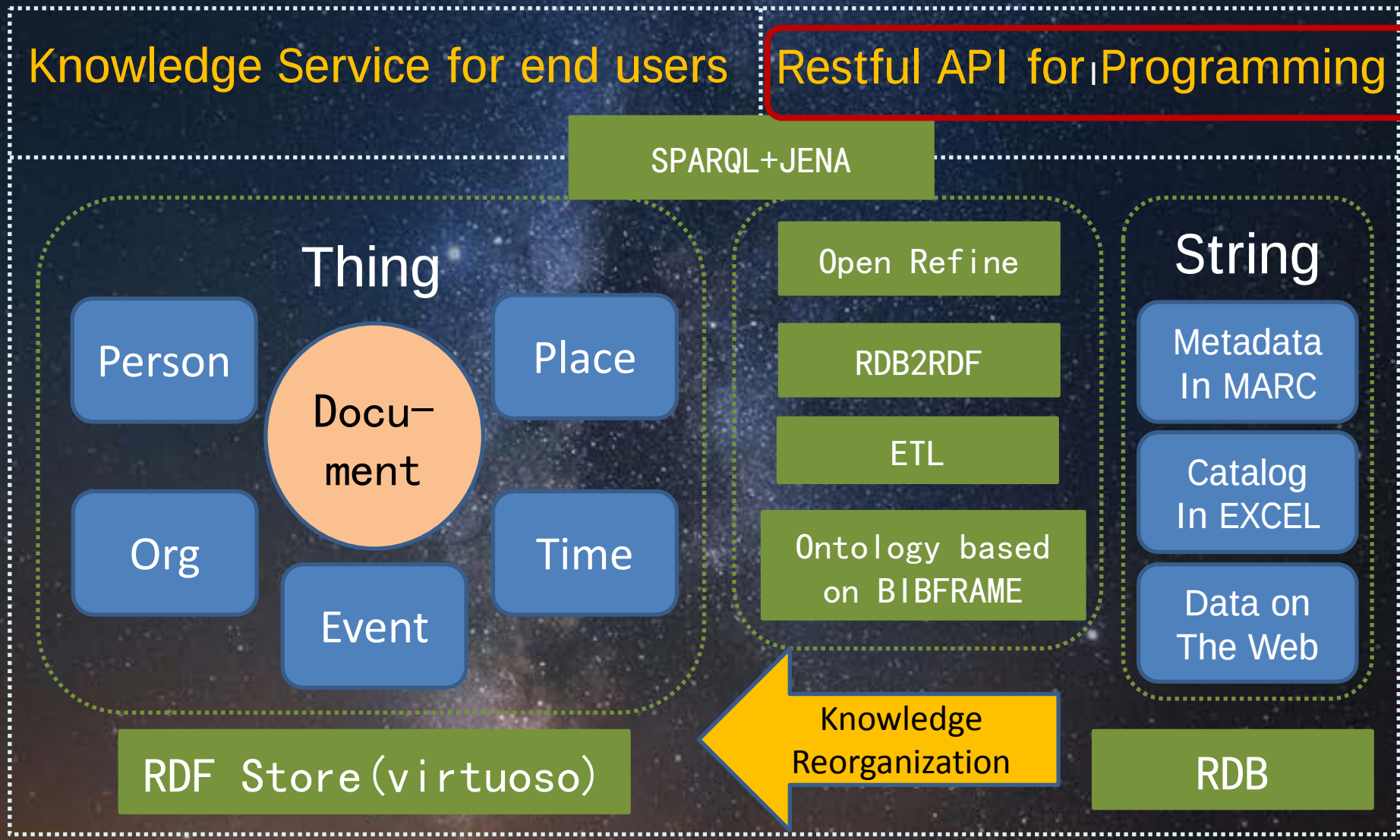
古籍开放数据内容

- 史志目录
- 官修目录
- 私家目录
- 藏书楼目录
- 现代联合目录
- 上图馆藏目录

表 5　14 种不同来源的古籍目录数据融合情况

数据融合前			数据融合后		
类别	题名	书目种数	版本数量	注释数量	作品数量
史志目录	汉书・艺文志	614		614	290 005
	隋书・经籍志	3 168		3 168	
	旧唐书・经籍志	2 974		2 974	
	新唐书・艺文志	5 247		5 247	
	宋史・艺文志	9 188		9 188	
	明史・艺文志	3 673		3 673	
	清史稿・艺文志	7 176		7 176	
官修目录	崇文总目	3 389		3 389	
	四库全书总目	10 249		10 249	
私家目录	郘亭知见传本书目	3 681		3 681	
联合目录	中国古籍善本书目	约 6 万	96 728		
	中国古籍总目	约 20 万	407 581		
机构藏书目录	柏克莱加州大学东亚图书馆中文古籍善本书志	802	802	802	
	上海图书馆古籍数据库（善本）	11 125	11 125		

技术框架


Knowledge Service for end users
Restful API for Programming
SPARQL+JENA
Thing
Person
Docu-
ment
Place
Org
Time
Event
Open Refine
RDB2RDF
ETL
Ontology based
on BIBFRAME
String
Metadata
In MARC
Catalog
In EXCEL
Data on
The Web
Knowledge
Reorganization
RDF Store(virtuoso)
RDB

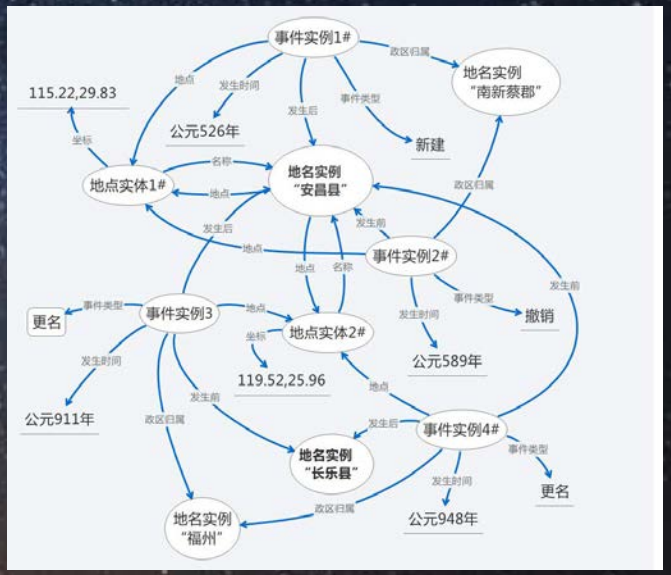
核心技术

RDF（资源描述框架）

Ontology（知识本体）

Graph Data Store （图数据存储）

地名	序号	属性	古今属性	简称	曾用名	现名	别名	时代	位置
艾	（1）	古邑名	历史地名					春秋	今山东省
	（2）	古邑名	历史地名					春秋	今江西省
艾比湖	（1）	湖名	现今地名				库尔湖、布尔哈齐湖		新疆维吾
艾城镇	（1）	乡镇名	现今地名						江西省永
艾丁湖	（1）	湖名	现今地名				觉洛浣		新疆维吾
艾浑	（1）								
艾家镇	（1）	城镇名	现今地名						湖北省宜
艾兰湖	（1）	湖名							
艾力西湖	（1）	乡镇名	现今地名						新疆维吾
艾陵	（1）	古邑名	历史地名					春秋	今山东省
艾陵湖	（1）	古湖名	历史地名						今江苏省
艾门扎拉	（1）	区片名	现今地名						内蒙古自
艾山	（1）	山名	现今地名						山东省胶
艾塘	（1）	古地名	历史地名						今江苏省



上图开放数据技术——LOD

什么是关联数据

用HTTP URI作为一切事物的名称

当访问HTTP URI时提供RDF数据

尽可能多地描述事物间的关系并使机器可理解

原则一、三

用HTTP URI作为一切事物的名称

古籍：

http://data.library.sh.cn/gj/resource/work/4rapf3lohdqcicr6

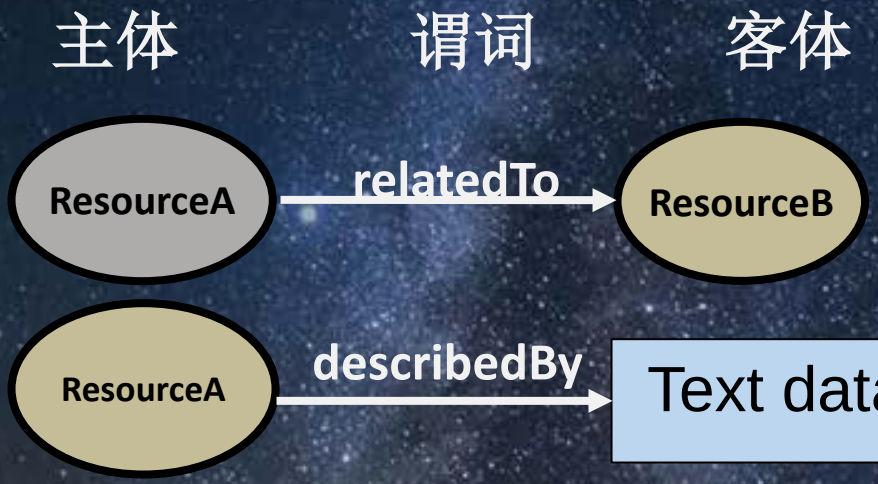
人： http://data.library.sh.cn/entity/person/ezpburgzs3jcaszi

地： http://data.library.sh.cn/entity/place/tk5s4pej6linq9tr

时： ttp://data.library.sh.cn/authority/temporal/4alljneqiivh5691

原则三——

当访问HTTP URI时提供RDF数据

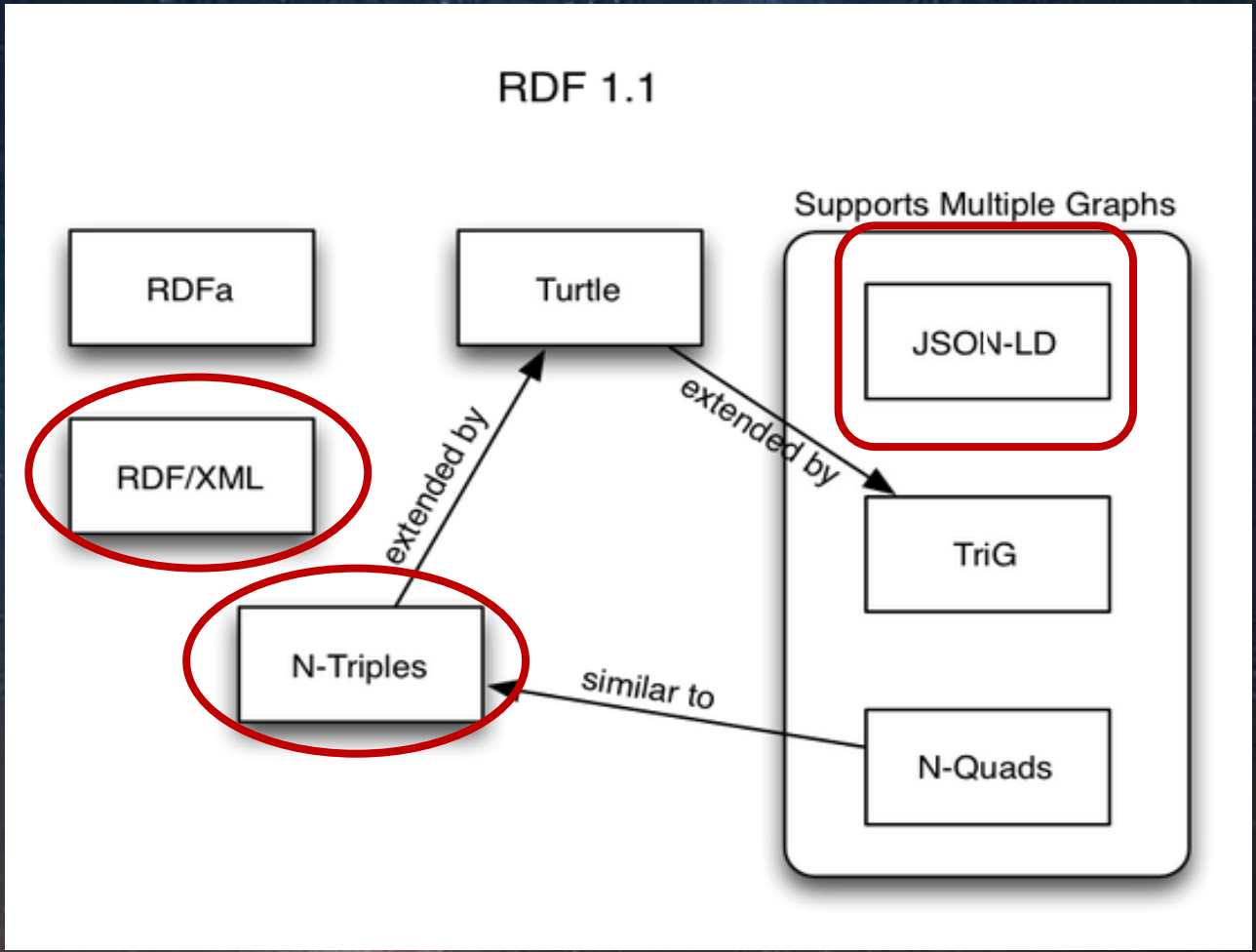


抽象模型
Abstract Model

```
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
<http://data.library.sh.cn/jp/entity/person/etrd44w3m3g1vncn> a
shl:Person;
        foaf:familyName
<http://data.library.sh.cn/authority/familyname/68n959cf8zdfkz3v>;
        foaf:name “夏翠娟”@zh_cn.
```

序列化格式
Serialization

RDF序列化格式



http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

http://data.library.sh.cn/entity/person/ezpburgzs3jcaszi.xml
http://data.library.sh.cn/entity/person/ezpburgzs3jcaszi.json
http://data.library.sh.cn/entity/person/ezpburgzs3jcaszi.ntriples

names.library.sh.cn/mrgf/service/work/persons?uri=http://data.library.sh.cn/entity/person/ezpburgzs3jcaszi&dataType=1

上海圖書館 人名規范庫

人物 地点 输入姓名/异名/民族/性别/职衔搜索 高级搜索

王士祯 RDF | NT | XML | JSON

1634.9.17 - 1711.6.26

国籍: | 籍贯: | 民族: | 性别: 男

异名

子真 阮亭 王士禛 王士禛 王士正 王士祯 王渔洋 豫孙

作品

师友诗传录 浯溪考 渔洋山人古诗选 带经堂全集七编 香祖笔记 北征记
蜀道驿程记 登燕子矶记 王渔洋遗书 古钵集选 粤行三志 东西二汉水辩 纪恩录
冬心斋摭言枕祕 [查看更多]

职衔

刑部尚书

上图名人手稿数据库 上图古籍数据库 国图规范档

王士祯（1634年9月17日—1711年6月26日），原名王士禛，字子真，一字贻上，号阮亭，又号渔洋山人，世称王渔洋，谥文简。山东新城（今桓台县）人，常自称济南人。清顺治十五年（1658）进士，康熙四十三年（1704）官至刑部尚书，颇有政声。清初杰出诗人、文学家，继钱谦益之后主盟诗坛，与朱彝尊并称"南朱北王"。诗论创"神韵"说，于后世影响深远。早年诗作清丽澄淡，中年转为苍劲。擅长各体，尤工七绝。好为笔记，有《池北偶谈》、《古夫于亭杂录》、《香祖笔记》等。

关系

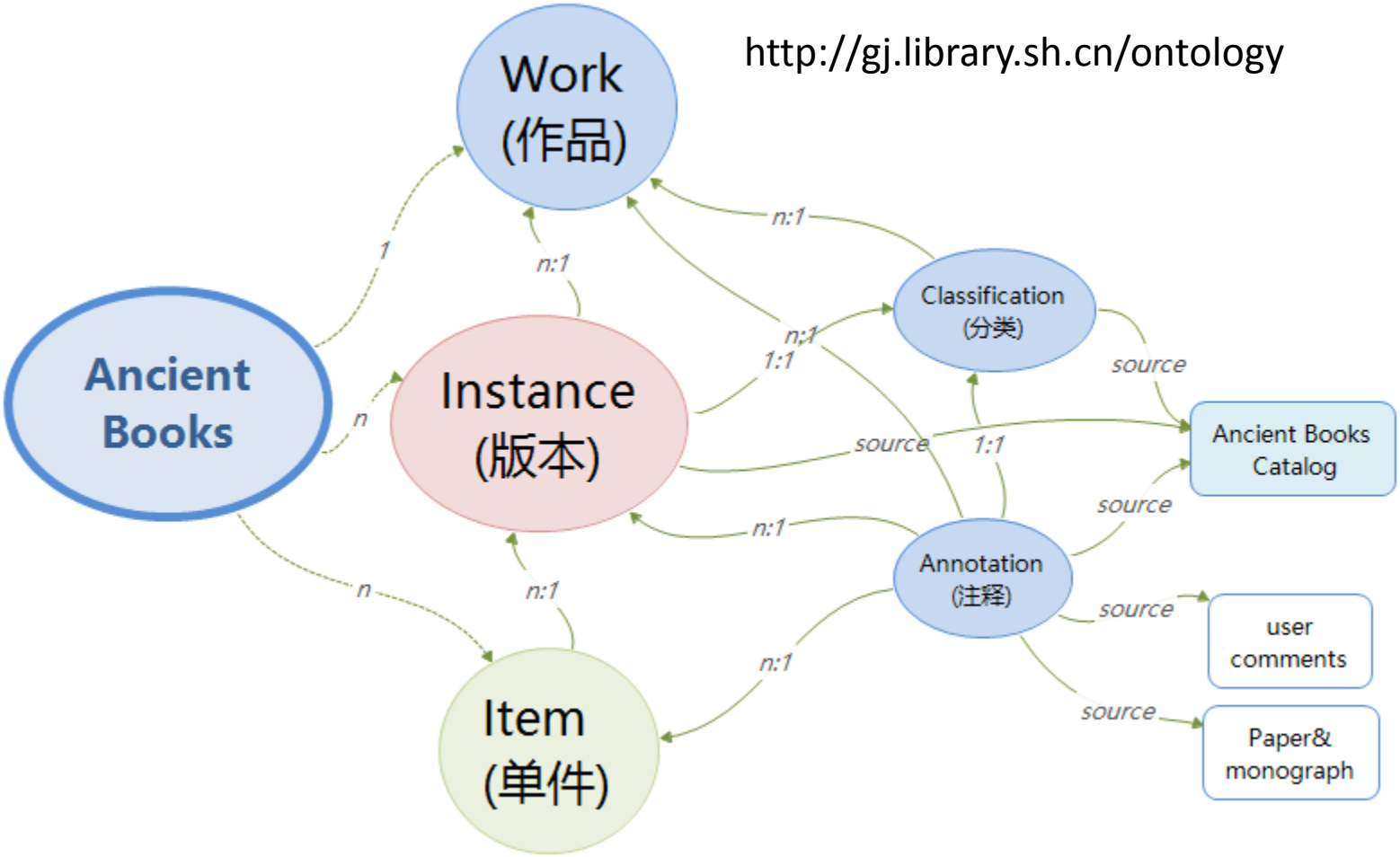
王启涑(长子) 王启访(第三子) 王端(长女)
王宫(第三女) 王婉(第二女) 王启浑(次子)
王启汧(第四子) 王与敕(父) 孙氏(母亲)
张万钟(岳父) 张氏(妻) 陈氏(妾)
王象晋(祖父) 王士禄(兄) 王士祜(兄)

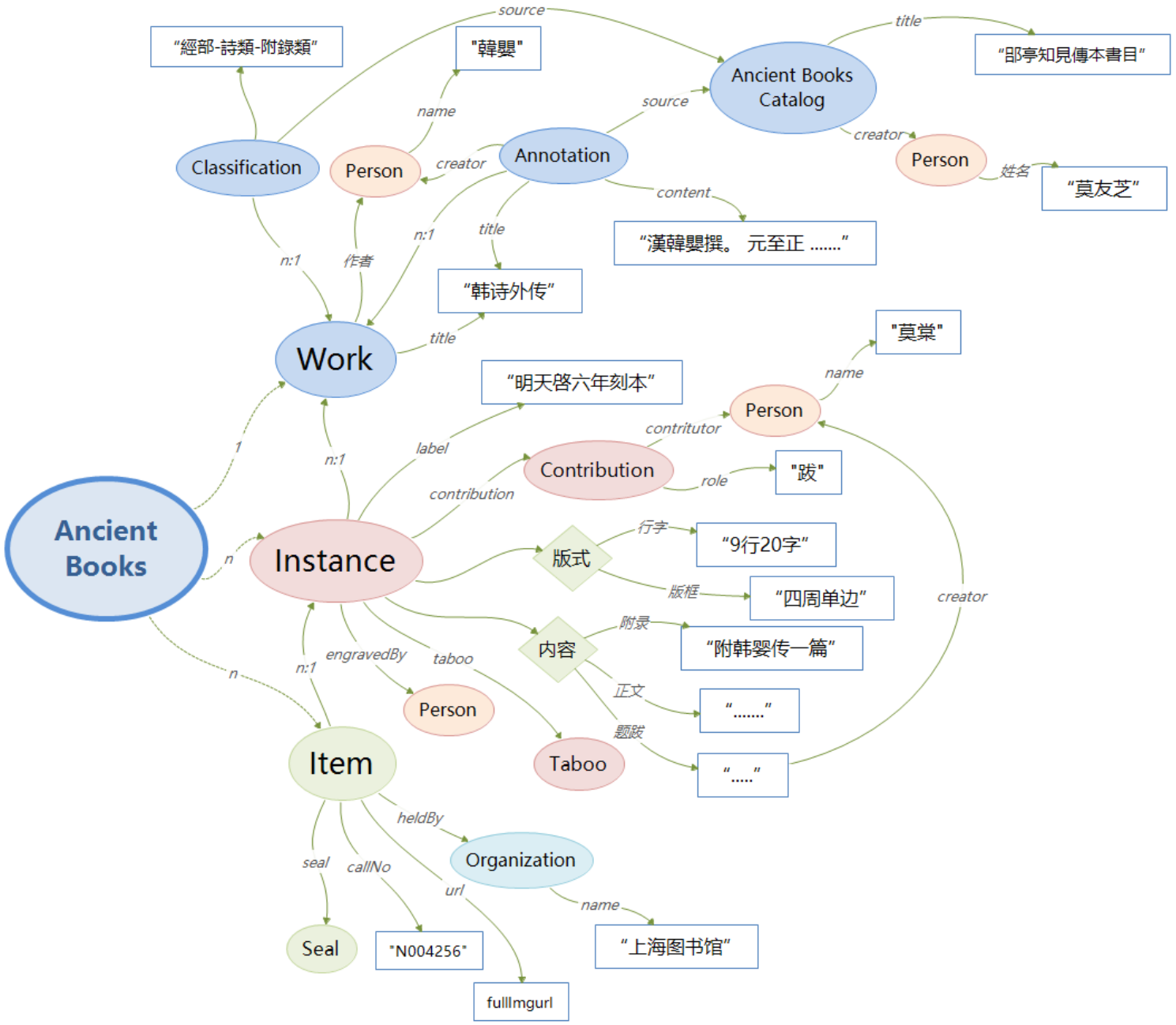
原则四

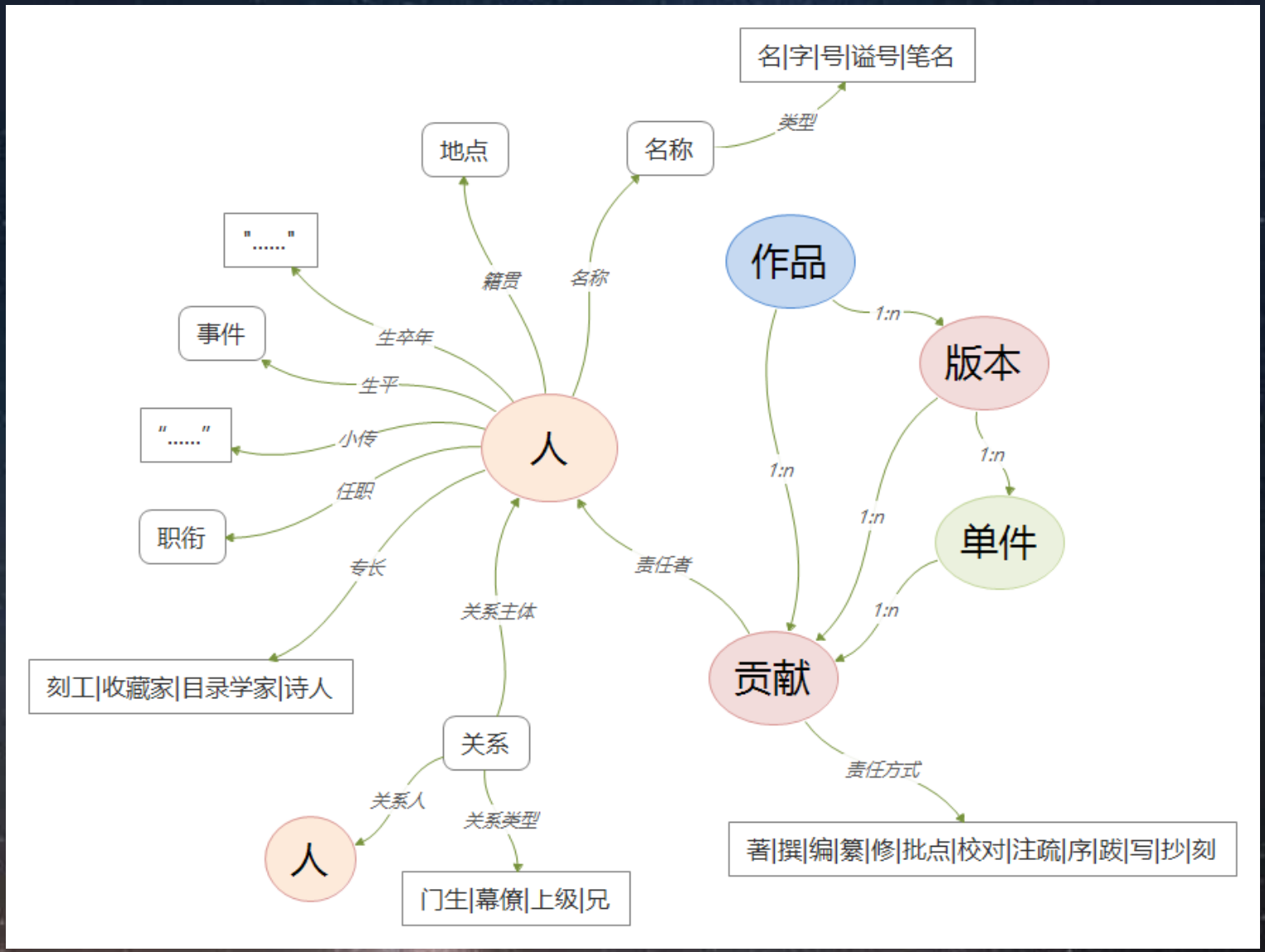
尽可能地描述事物间关系
并使机器可理解

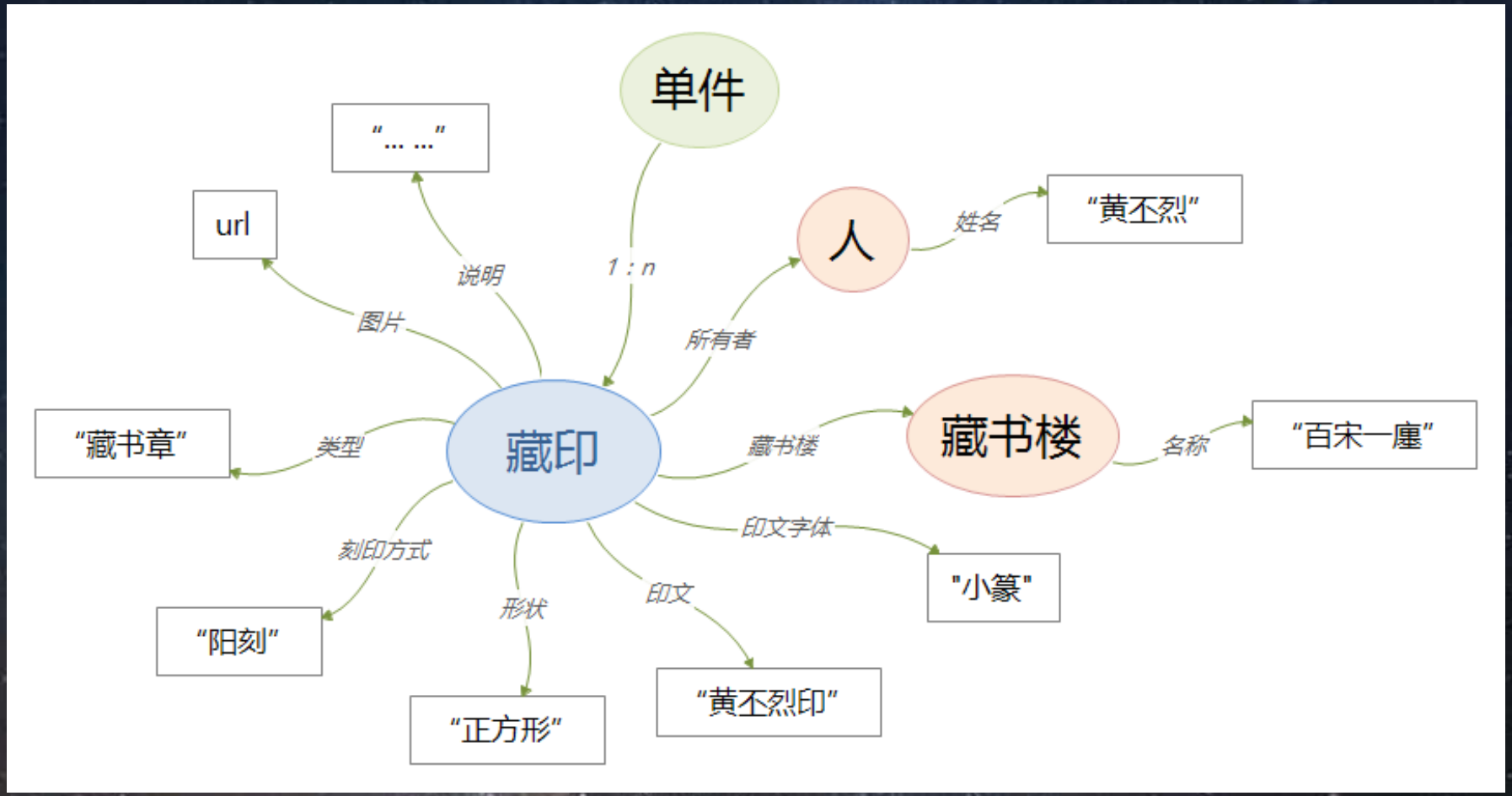
本体（Ontology）

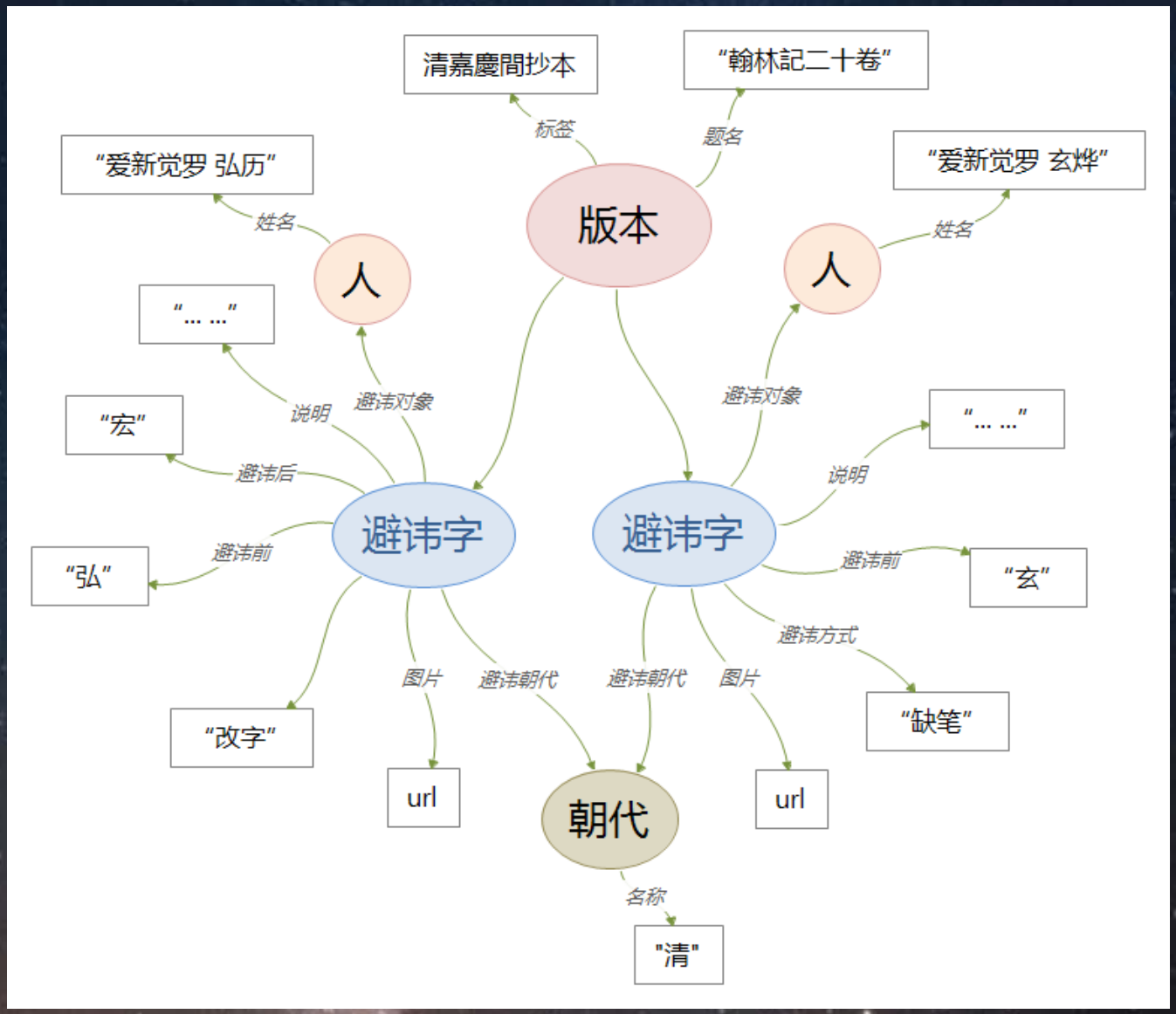
古籍数据模型：本体


http://gj.library.sh.cn/ontology
Work
(作品)
Ancient
Books
Instance
(版本)
Item
(单件)
Classification
(分类)
Annotation
(注释)
Ancient Books
Catalog
user
comments
Paper&
monograph
1
n
n
n:1
n:1
n:1
n:1
n:1
n:1
1:1
1:1
source
source
source
source
source


"經部-詩類-附錄類"
source
"韓嬰"
Ancient Books Catalog
title
"郘亭知見傳本書目"
name
source
Classification
Person
creator
Annotation
creator
Person
姓名
"莫友芝"
content
"漢韓嬰撰。 元至正 ....."
n:1
作者
n:1
title
"韩诗外传"
title
Work
"明天啓六年刻本"
label
Person
name
"莫棠"
contritutor
1
n:1
Contribution
role
"跋"
contribution
Ancient Books
n
Instance
版式
行字
"9行20字"
版框
"四周单边"
creator
内容
附录
"附韩婴传一篇"
正文
"....."
题跋
"....."
n
n:1
engravedBy
taboo
Person
Item
Taboo
heldBy
Organization
seal
callNo
url
name
Seal
"N004256"
"上海图书馆"
fullImgurl


名|字|号|谥号|笔名
类型
名称
地点
"……"
籍贯
名称
作品
1:n
事件
生卒年
版本
生平
"……"
小传
人
1:n
1:n
任职
1:n
单件
职衔
专长
责任者
1:n
关系主体
贡献
刻工|收藏家|目录学家|诗人
关系
责任方式
关系人
关系类型
人
著|撰|编|纂|修|批点|校对|注疏|序|跋|写|抄|刻
门生|幕僚|上级|兄


单件
"... ..."
url
说明
1 : n
图片
人
姓名
"黄丕烈"
所有者
"藏书章"
类型
藏印
藏书楼
藏书楼
名称
"百宋一廛"
印文字体
"小篆"
刻印方式
"阳刻"
形状
"正方形"
印文
"黄丕烈印"


清嘉慶間抄本
“翰林記二十卷”
标签
题名
“爱新觉罗 弘历”
版本
“爱新觉罗 玄烨”
姓名
姓名
人
人
“… …”
“… …”
说明
避讳对象
避讳对象
“宏”
说明
避讳后
避讳字
避讳字
避讳前
避讳前
“弘”
“玄”
避讳方式
图片
避讳朝代
避讳朝代
图片
“缺笔”
“改字”
url
朝代
url
名称
"清"

数据驱动的古籍循证研究

Data Driven Evidence-based Research for Chinese Ancient Books

构建古籍循证的证据链

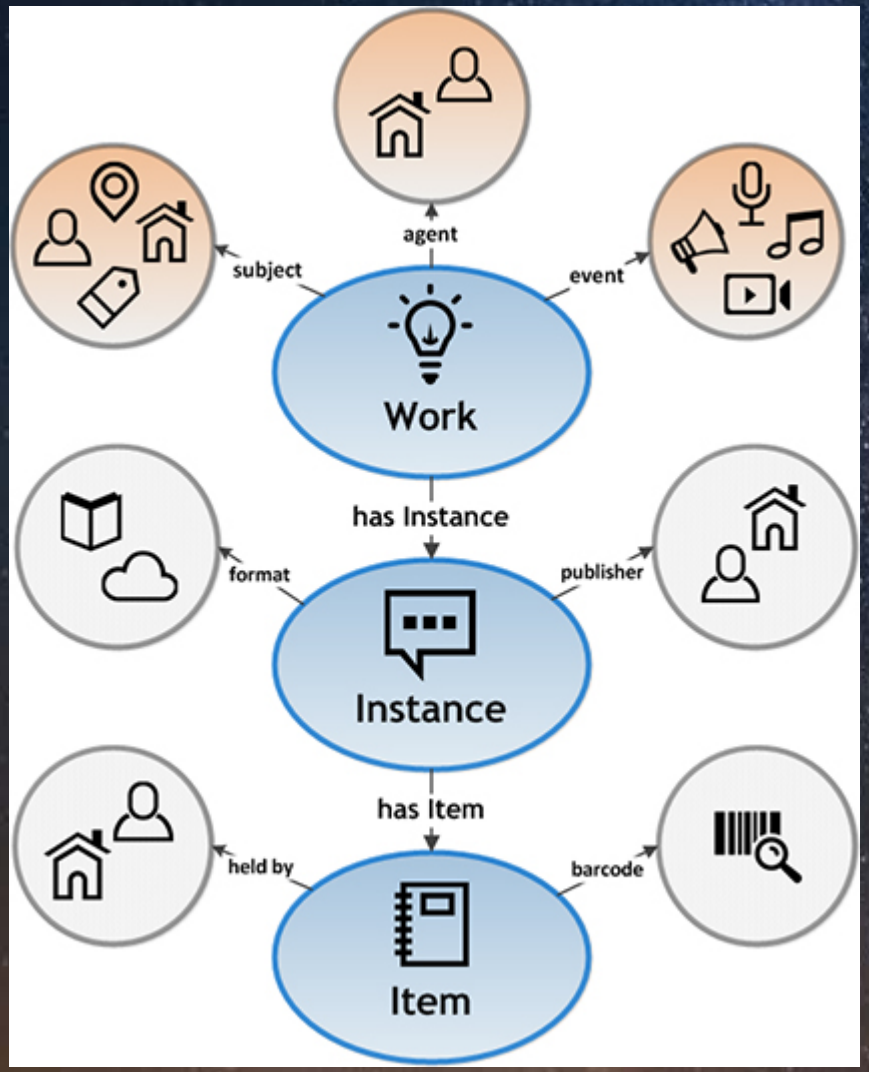
Build Evidence Chains for Chinese Ancient Books

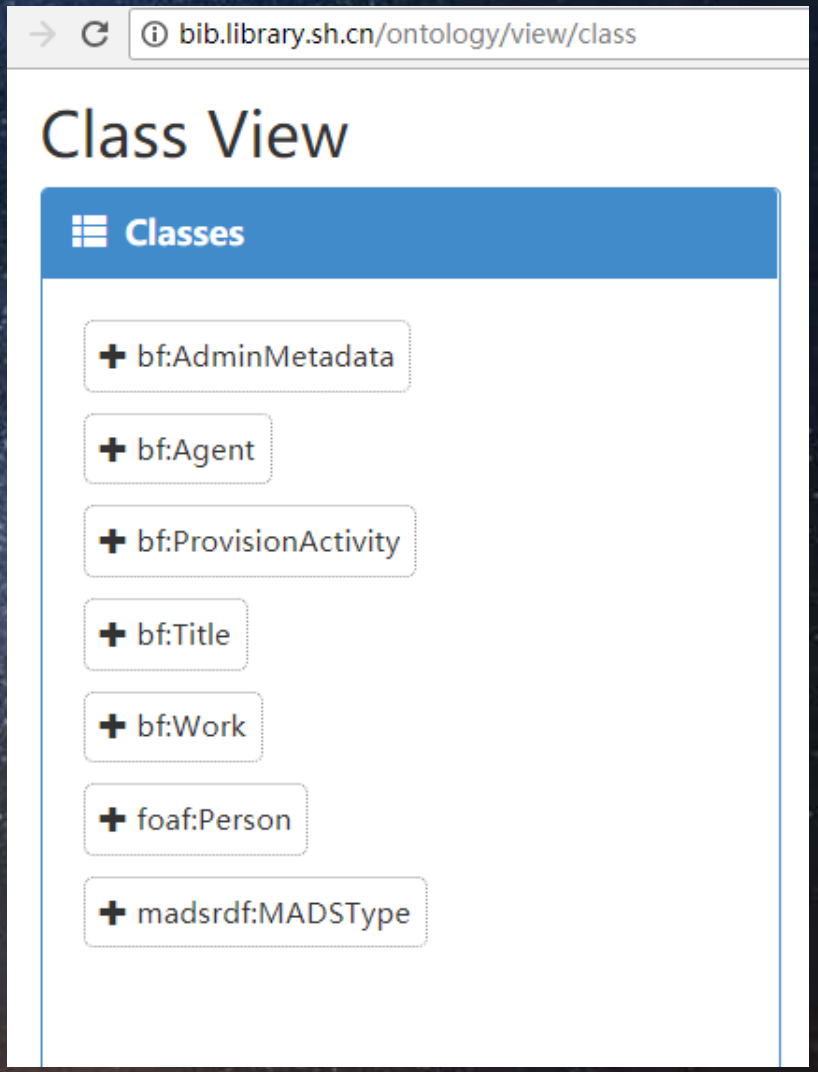
- 物理证据 (Phisical Evidence)
- 历史证据 (Historical Evidence)
- 内容证据 (Content Evidence)
- 关联证据 (Relational Evidence)

夏翠娟，林海青，刘炜. 面向循证实践的中文古籍数据模型研究与设计. 中国图书馆学报，2017（6）

联编中心书目数据模型：BIBFRAME2.0

http://bib.library.sh.cn/ontology





数据开放方式

中文古籍聯合目錄及循證平台
Chinese Ancient Books Union Catalogue and Evidence-based Platform

上海市文献联合编目中心
关联书目数据发布平台

家譜 知識服務平台
Genealogy knowledge service platform

上海图书馆
名人手稿档案库

盛宣怀档案知識庫

机构名录

地理名词表

中国历史纪年表

人名规范库

http://data1.library.sh.cn

SPARQL Endpoint

Restful API

HTTP URI

数据消费接口

JSO
N-LD

返回数据

开放数据接口调用流程

1.申请API Key
data.library.sh.cn

2.阅读接口说明文档
data.library.sh.cn

接口调用

3.1. SPARQL Endpoint

特点：
Web 调用，跨网域搜索，要求精通SPARQL语言。
网址：
http://data1.library.sh.cn:8890/sparql
http://data1.library.sh.cn:8892/sparql

3.2. Restful API

特点：Web调用，跨语言，门槛低。
方法：
http://data.library.sh.cn/sg/person/data?[参数1]&[参数2]&[参数3]?key=YourAPIKey

3.3.HTTP URI内容协商

特点：Web调用，跨语言，访问实体的HTTP URI即可获得其结构化数据。
方法：需先通过调用前两种接口获得实体的HTTP URI

返回数据格式：

JSON-LD

W3C推出的用于关联数据消费的JSON格式

一起来，更精彩！

Be together, Be better!

祝各位取得好成绩！

cjxia@libnet.sh.cn