

领域本体的设计与RDF数据的编码

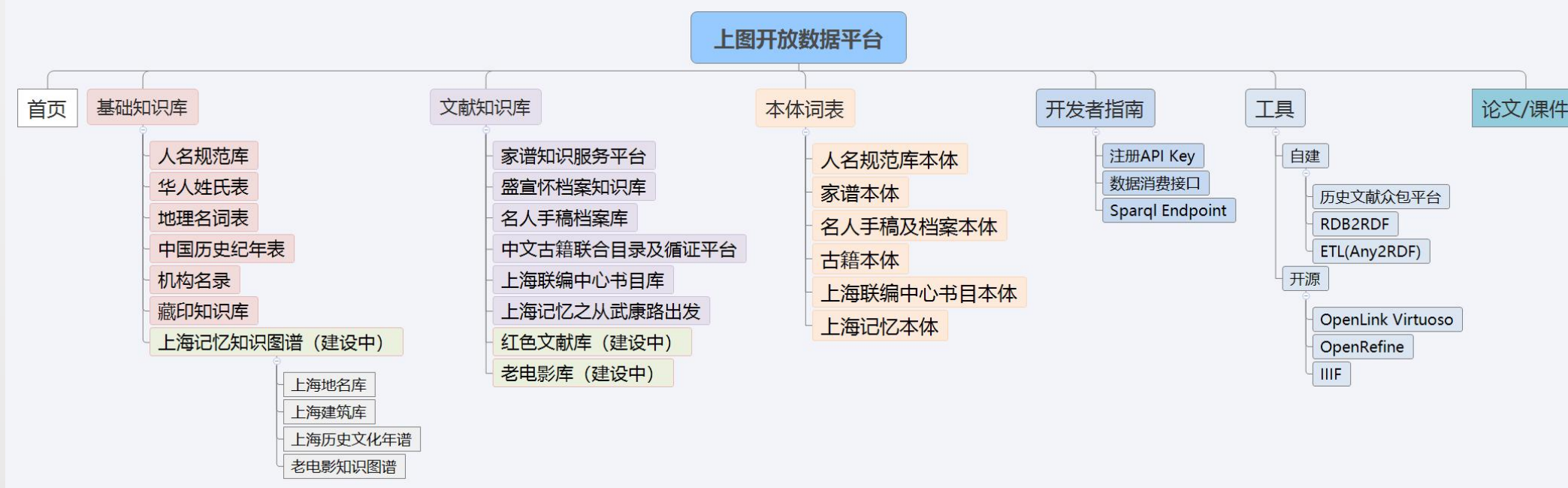
上海图书馆 夏翠娟

2018年12月15日 @华东师范大学

上海市图书馆学会2018年学术年会

上海图书馆人文数据开放平台

<http://data.library.sh.cn>



关联数据的四原则



关联数据的技术堆栈

资源标识和定位: HTTP URI

领域建模: W3C Web Ontology

数据抽象模型: RDF

数据编码格式: XML/Turtle/N3/Json-LD

数据存储: GraphDB (triple store)

数据查询: SPARQL

HTTP URI

IRI: 内部资源标识符

URI: 统一资源标识符 (isbn)

URL: 统一资源定位符

https://www.baidu.com/s?ie=utf-8&f=8&rsv_bp=0&rsv_idx=1&tn=baidu&wd=胡适

HTTP URI: 唯一标识、唯一定位、永久稳定

http://dbpedia.org/resource/Lu_Xun

本体 (Ontology)

是对领域知识进行抽象的、形式化的概念模型。



Europeana
Data Model

领域的
可共享的
概念模型

Classes				
foaf:Agent				
foaf:Person				
shl:Person				

shl:Person				
个人。				
Property	Subproperty of	Label	Used With	Expected Value
foaf:gender		性别	shl:Person	rdfs:Literal
rel:childOf		父	shl:Person	shl:Person

术语词表
类 (概念)
属性 (关系)

RDF Schema 1.1

SKOS

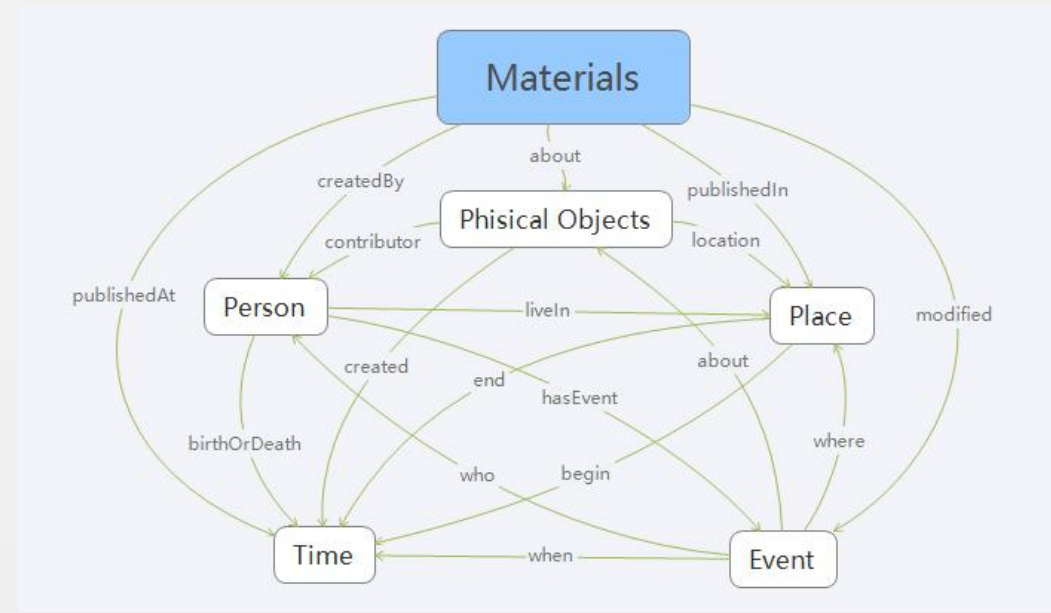
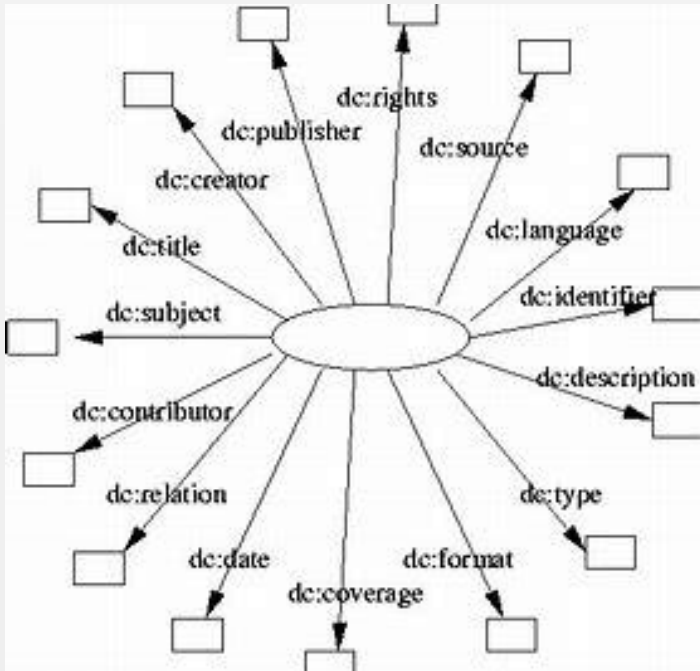
OWL

形式化的
机器可理解

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix bf: <http://bibframe.org/vocab/> .
bf:Work a rdfs:Class ;
    rdfs:label "作品" ;
    rdfs:subClassOf bf:Resource ;
    rdfs:comment "家谱作品。" .
```

元数据 vs. 本体

		元数据	本体
相同点		对信息的结构化描述	
不同点	定义及构成	“关于数据的数据。” 元数据方案（元素集）、置标方案、元数据记录	“领域的可共享的概念及概念间关系的形式化定义。” 模型、词表（类、属性）、实体/实例对象
	描述对象	文献	内容
	描述重点	文献的外在特征	内容对象的特征及其相关关系
	词表定义	元素	类、属性
	标准规范	MARC DC	FRBR BIBFRAME FOAF schema.org
	编码	RDF XML	RDFs OWL
	存储与查询	RDB SQL	RDF Store Graph DB SPARQL
联系		元数据元素可作为本体的属性，本体可看做“元”元数据。	



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	今山东省新
	(2)	古邑名	历史地名					春秋	今江西省修
艾比湖	(1)	湖名	现今地名				库尔湖、布尔哈齐湖		新疆维吾尔自治区
艾城镇	(1)	乡镇名	现今地名						江西省永修县
艾丁湖	(1)	湖名	现今地名				觉洛院		新疆维吾尔自治区
艾泽	(1)								
艾家镇	(1)	城镇名	现今地名						湖北省宜昌市
艾兰湖	(1)	湖名							
艾力西湖	(1)	乡镇名	现今地名						新疆维吾尔自治区
艾陵	(1)	古邑名	历史地名					春秋	今山东省新
艾陵湖	(1)	古湖名	历史地名						今江苏省溧水
艾门扎拉	(1)	区片名	现今地名						内蒙古自治区
艾山	(1)	山名	现今地名						山东省胶州市
艾塘	(1)	古地名	历史地名						今江苏省溧水
艾塘	(1)	古地名	历史地名						今江苏省溧水

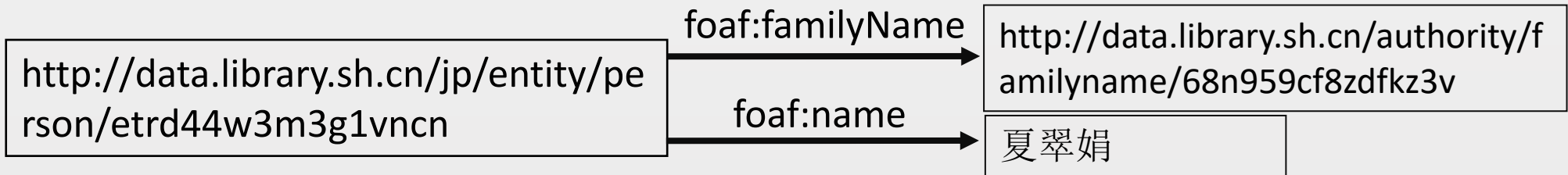


当我们在谈论RDF的时候，我们在谈论什么？

主体(subject) 谓词(predict) 客体(object)



抽象模型
Abstract Model



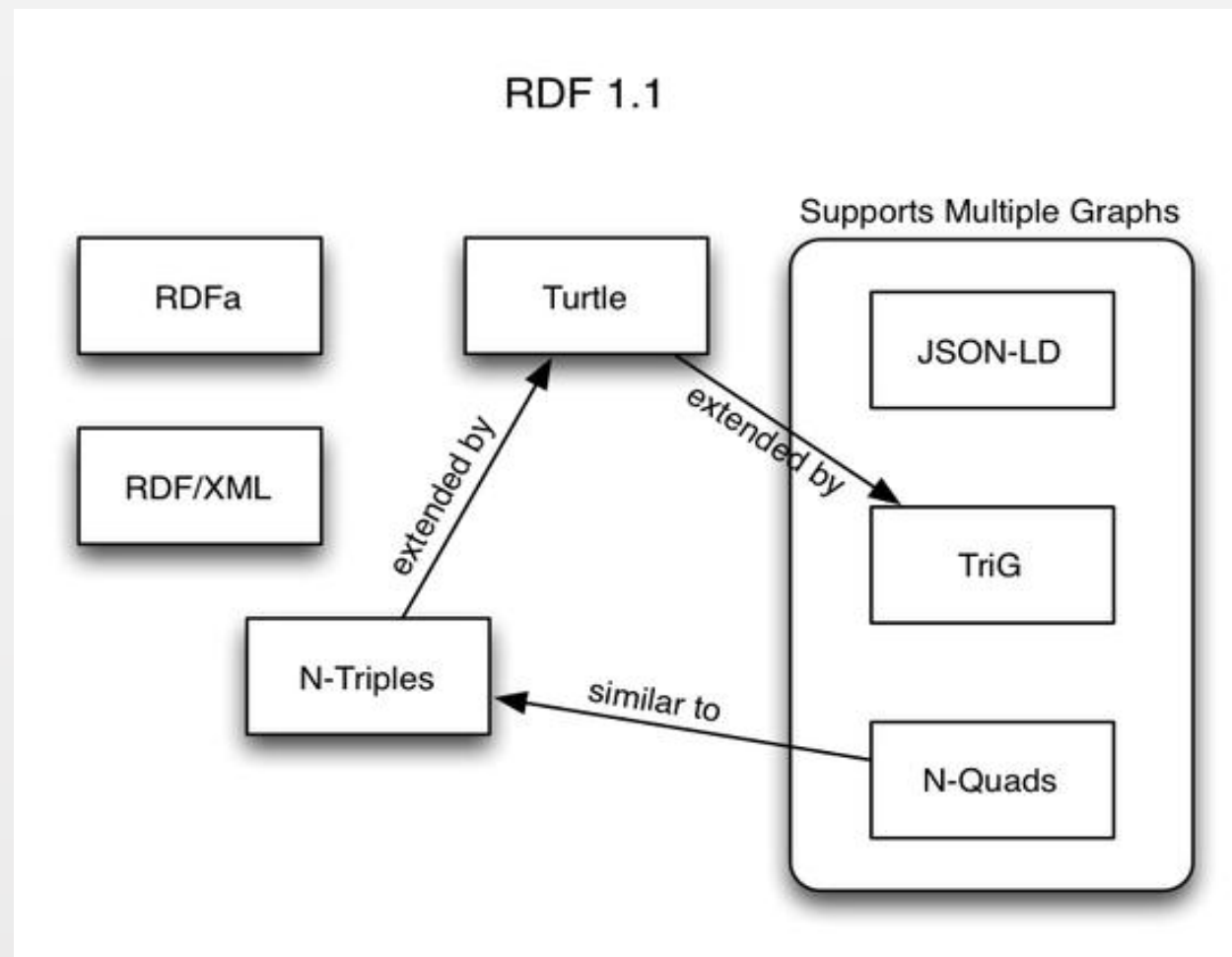
数据实例
Data Instance

```
@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.
```

数据编码
序列化格式
Serialization

RDF

序列化格式



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

RDF

序列化格式—

RDF/XML

EXAMPLE 3

Complete description of all graph paths

```
<rdf:Description rdf:about="http://www.w3.org/TR/rdf-syntax-grammar">
  <ex:editor>
    <rdf:Description>
      <ex:homePage>
        <rdf:Description rdf:about="http://purl.org/net/dajobe/">
        </rdf:Description>
      </ex:homePage>
    </rdf:Description>
  </ex:editor>
</rdf:Description>

<rdf:Description rdf:about="http://www.w3.org/TR/rdf-syntax-grammar">
  <ex:editor>
    <rdf:Description>
      <ex:fullName>Dave Beckett</ex:fullName>
    </rdf:Description>
  </ex:editor>
</rdf:Description>

<rdf:Description rdf:about="http://www.w3.org/TR/rdf-syntax-grammar">
  <dc:title>RDF 1.1 XML Syntax</dc:title>
</rdf:Description>
```

RDF

序列化格

式—Turtle

EXAMPLE 1

```
@base <http://example.org/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix rel: <http://www.perceive.net/schemas/relationship/> .

<#green-goblin>
  rel:enemyOf <#spiderman> ;
  a foaf:Person ;    # in the context of the Marvel universe
  foaf:name "Green Goblin" .

<#spiderman>
  rel:enemyOf <#green-goblin> ;
  a foaf:Person ;
  foaf:name "Spiderman", "Человек-паук"@ru .
```

RDF 序列化格 式—RDFa

本网页资源是一个由
schema.org定义的博客
[] rdfs:type schema:Blog

EXAMPLE 1: Example of an HTML+RDFa 1.1 document

```
<!DOCTYPE html>
<html lang="en">
  <head>
    <title>Example Document</title>
  </head>
  <body vocab="http://schema.org/">
    <p typeof="Blog">
      Welcome to my <a property="url" href="http://example.org/">blog</a>.
    </p>
  </body>
</html>
```

EXAMPLE 2: Turtle output of Example Document

```
[] a <http://schema.org/Blog>;
    <http://schema.org/url> <http://example.org/> .
```

RDF 序列化格式—JSON-LD

```
{
  "@context":
  {
    "xsd": "http://www.w3.org/2001/XMLSchema#",
    "foaf": "http://xmlns.com/foaf/0.1/",
    "foaf:homepage": { "@type": "@id" },
    "picture": { "@id": "foaf:depiction", "@type": "@id" }
  },
  "@id": "http://me.markus-lanthaler.com/",
  "@type": "foaf:Person",
  "foaf:name": "Markus Lanthaler",
  "foaf:homepage": "http://www.markus-lanthaler.com/",
  "picture": "http://twitter.com/account/profile_image/markuslanthaler"
}
```

如何设计一个本体？

本体形式化 05

以机器可读的格式编码。并在网上公开发布。

术语的复用 03

注意属性的Domain和Range，在复用时不要改变原有术语的外延和内涵。

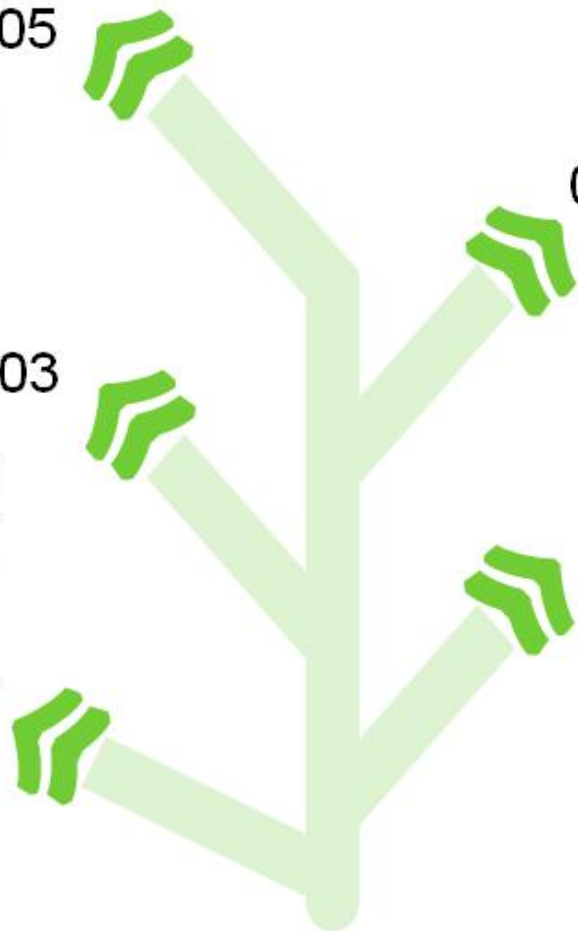
实体关系分析 01

词表设计 04

类、属性、关系的定义。尽量复用，必要时自定义。关键在于定义好属性的Domain和Range，一个类同时继承不同的类。

参考模型/词表调研 02

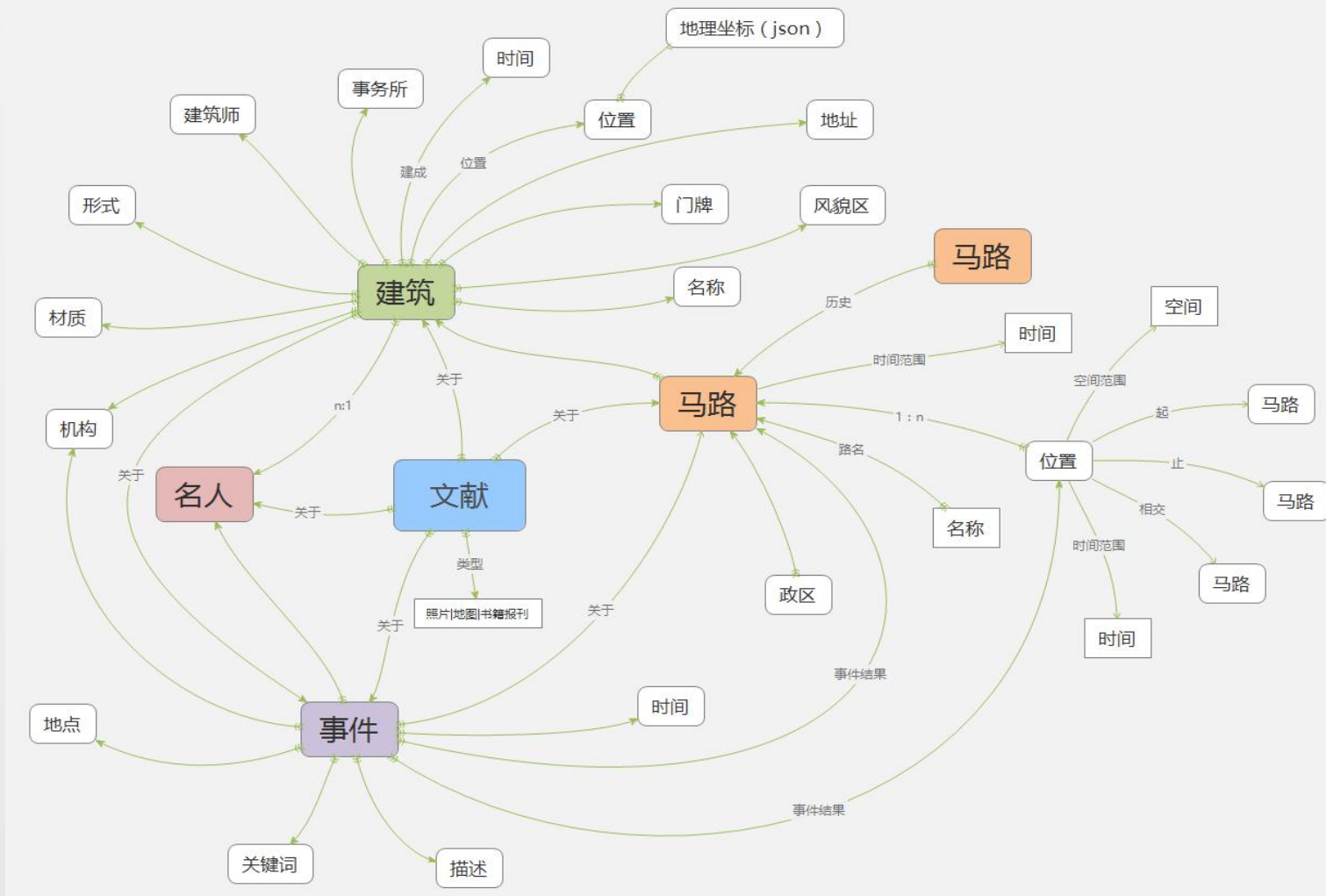
调研与需求领域相关的模型和词表有哪些？



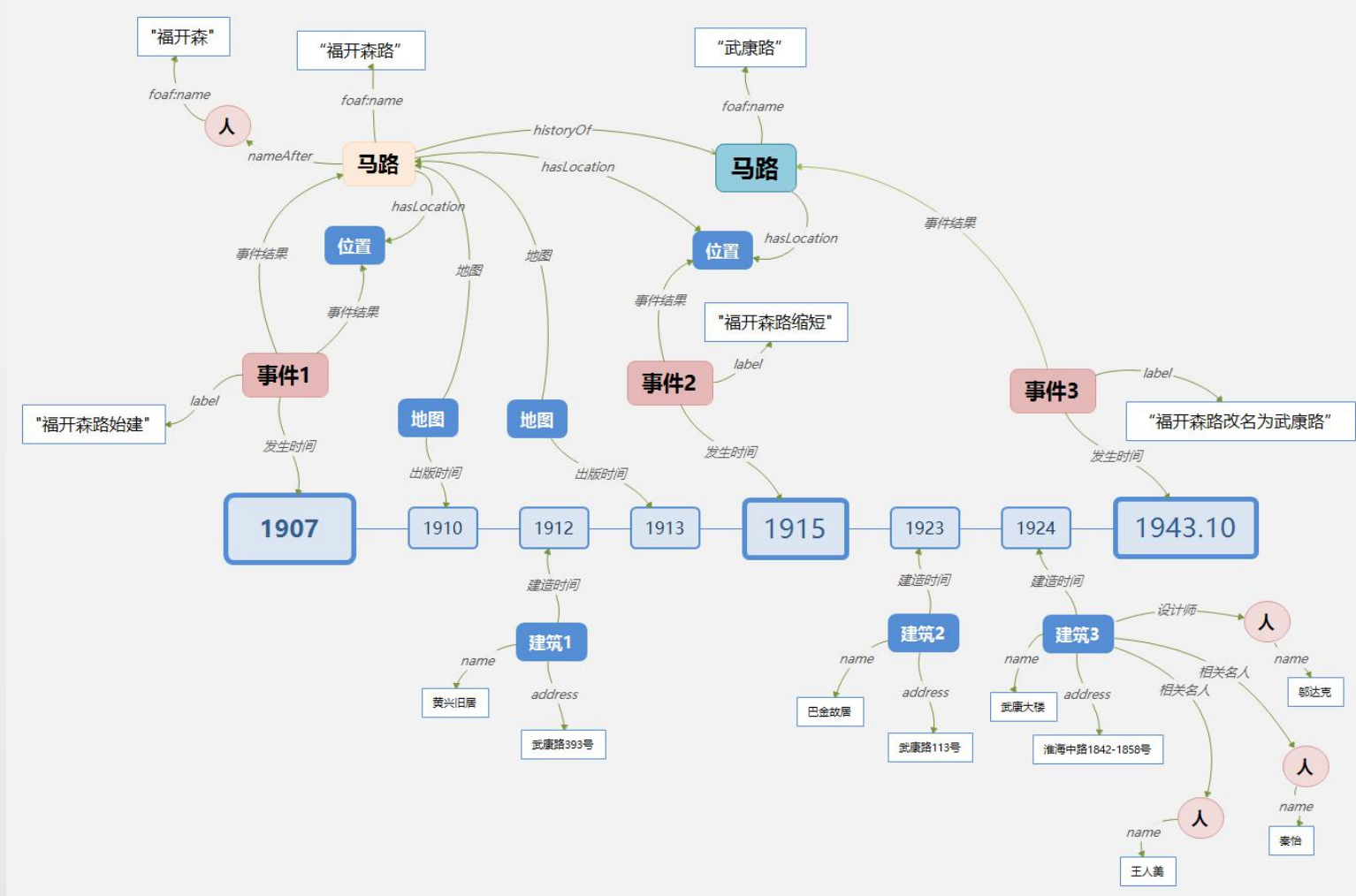
步骤1：调研——需求、数据结构、已有的领域本体模型和词表，设计抽象模型

1907-1915：福开森路
起：徐家汇路 终：徐家汇路
相交：宝昌路
1915-1943：福开森路
起： 终：霞飞路
相交：白赛仲路
1943-：武康路
起：华山路 终：淮海中路

1907年：福开森路始建
1915年：福开森路缩短
1943年：福开森路改名为武康路



步骤2：用真实数据来验证与完善抽象模型



步骤3：设计术语词表，定义类和属性

上海记忆 (武康路+电影+红色旅游) Applica...vocab-20181102		上海记忆 (武康路+电影+红色旅游) Applica...cab-20181102-xcj.xlsx * ×	
B32	道路	B4	题名
16 shl:Music	A		
17 shl:Audio	B		
18 shl:Photo	C		
19 shl:Text	D		
20 foaf:Image	E		
21 prov:Activity	F		
22 shl:Event	题名		
23 shl:Name	本体 (属性)		
24 shl:Place	property		
25 geo:SpatialThing	label		
26 shl:Location	domain		
27 shl:Temporal	range		
28 shl:Theater	usedwith		
29 prov:Activity	expectedValue		
30 shl:Event	scheme		
31 shl:Architecture	rdfo:label		
32 shl:Road	dc:title		
33 shl:Block	dc:creator		
34 shl:Lane	shl:donator		
35 shl:PhysicalCulturalHeritage	dc:type		
36 shl:Drama	dc:language		
	dc:date		
	dc:format		
	foaf:name		
	shl:name		
	foaf:img		
	shl:abbreviateName		
	dbpedia:director		
	dbpedia:producer		
	shl:writer		
	shl:screenWriter		
	shl:supervisor		
	shl:actor		
	shl:mainActor		

步骤4：形式化编码

RDF Schema Alignment

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from y placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute

Base URI: <http://www.umac.mo/> [edit](#)

RDF Skeleton

[RDF Preview](#)

Available Prefixes: bf shl rel rdfs foaf [+ add prefix](#) [manage prefixes](#)

Class URI

[x](#) rdfs:Class

[add rdf:type](#)



[x](#) >rdfs:label→



Label cell

[x](#) >rdfs:subClassOf→



SubClassOf URI

[+](#) ...

[add rdf:type](#)

[x](#) >rdfs:comment→



Comment cell

shmemory_ontology_chenss_2018-11-13.ttl

```
77
78 shl:PhysicalCulturalHeritage a rdfs:Class ;
79     rdfs:subClassOf shl:Resource ;
80     rdfs:subClassOf crm:PhysicalThing ;
81     rdfs:label "文化遗产" .
82
83 shl:Drama a rdfs:Class ;
84     rdfs:subClassOf dbpont:Drama ;
85     rdfs:subClassOf shl:Resource ;
86     rdfs:label "戏剧" .
87
88 shl:ArtSchool a rdfs:Class ;
89     rdfs:subClassOf shl:Resource ;
90     rdfs:label "艺术流派" .
91
92 shl:ArtPeriod a rdfs:Class ;
93     rdfs:subClassOf shl:Resource ;
94     rdfs:label "艺术时期" .
95
96 rdfs:label a rdf:Property;
97     rdfs:label "标签" ;
98     rdfs:domain rdfs:Resource ;
99     rdfs:range rdfs:Literal .
100
101 dc:title a rdf:Property;
102     rdfs:label "题名" ;
103     rdfs:domain shl:Resource ;
104     rdfs:range rdfs:Literal .
105
106 dc:creator a rdf:Property;
```

步骤5：发布到网上

data.library.sh.cn

gj.library.sh.cn/ontology/view/list

Classes

选择下面的一个类名并跳转到一个

bf:Agent

bf:Category

bf:Classification

bf:Collection

bf:Contribution

bf:Identifier

bf:Instance

bf:Item

bf:Organization

bf:Person

Properties

选择下面的一个属性名并跳转到一个

bf:agent

bf:category

bf:categoryType

bf:categoryValue

bf:classification

bf:collection

bf:contributor

bf:creator

bf:dimensions

bf:edition

bf:extent

Class View

Classes

foaf:Agent

foaf:Person

shl:Person

org:Organization

shl:Agent

geo:SpatialThing

oa:Annotation

rdfs:Resource

schema:Thing

skos:Concept

time:TemporalEntity

shl:Person

个人。

Property	Subproperty of	Label	Used With	Expected Value
foaf:gender		性别	shl:Person	rdfs:Literal
rel:childOf		父	shl:Person	shl:Person
rel:parentOf		父	shl:Person	shl:Person
rel:spouseOf		配偶	shl:Person	shl:Person
shl:birthday		生于	shl:Person	rdfs:Literal
shl:courtesyName		字	shl:Person	rdfs:Literal
shl:deathday		卒于	shl:Person	rdfs:Literal
shl:genealogyName		谱名	shl:Person	rdfs:Literal
shl:generationCharacter		字辈	shl:Person	rdfs:Literal
shl:orderOfSeniority		行	shl:Person	rdfs:Literal
shl:posthumousTitle		谥号	shl:Person	rdfs:Literal
shl:pseudonym		号	shl:Person	rdfs:Literal

藏印

More...

家谱

More...

人名

More...

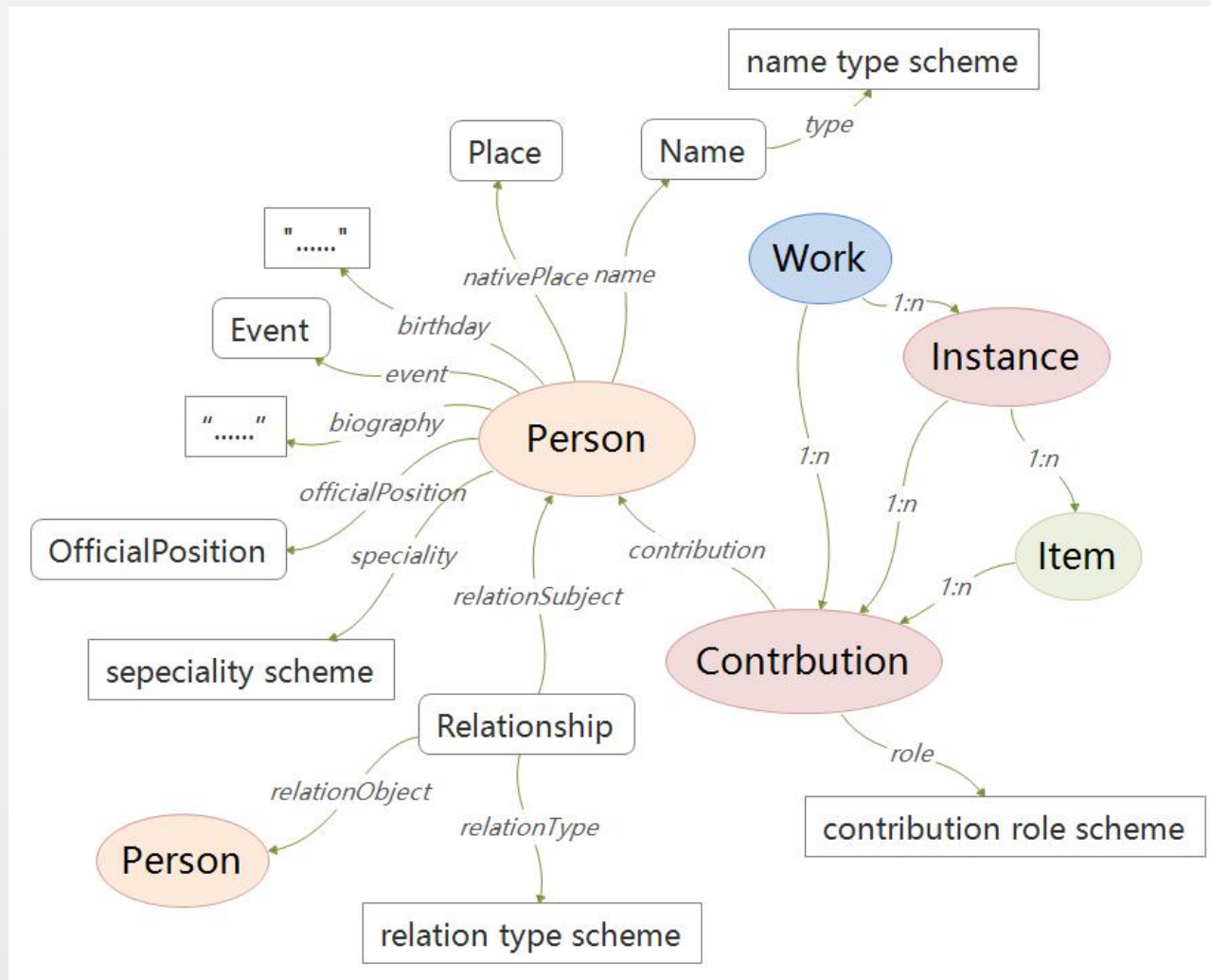
org:memberOf

pmb:titlePage

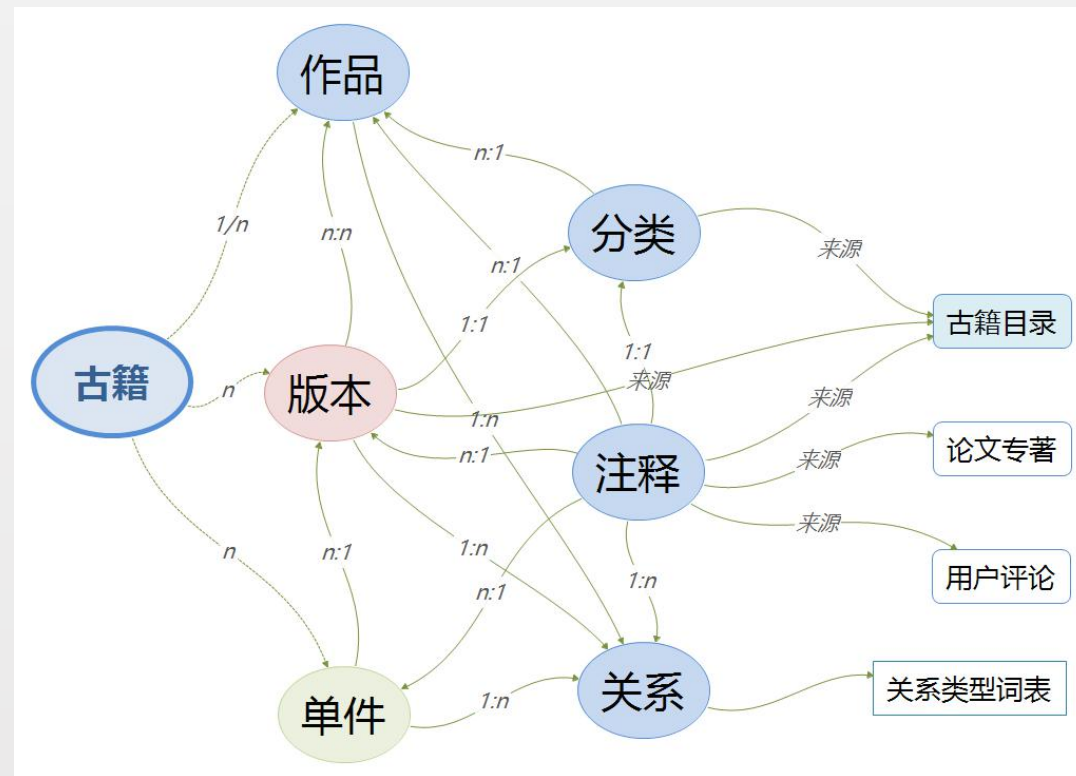
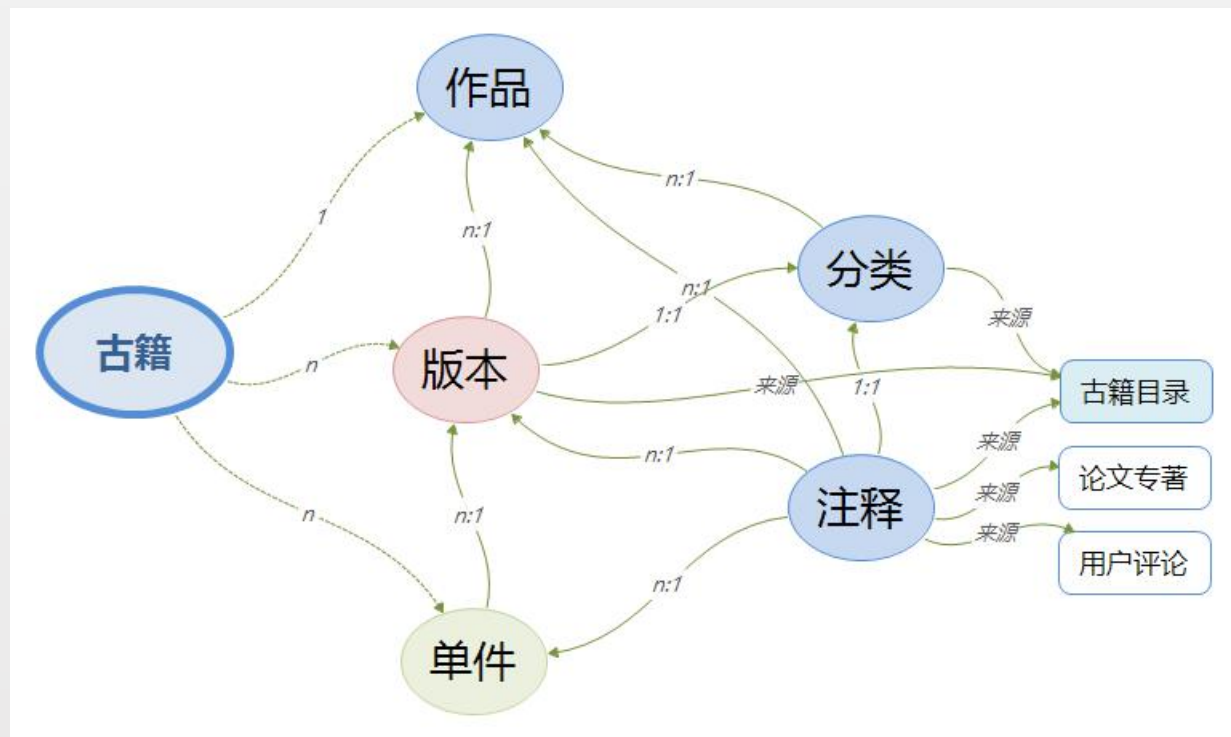
shl:monarch

难点

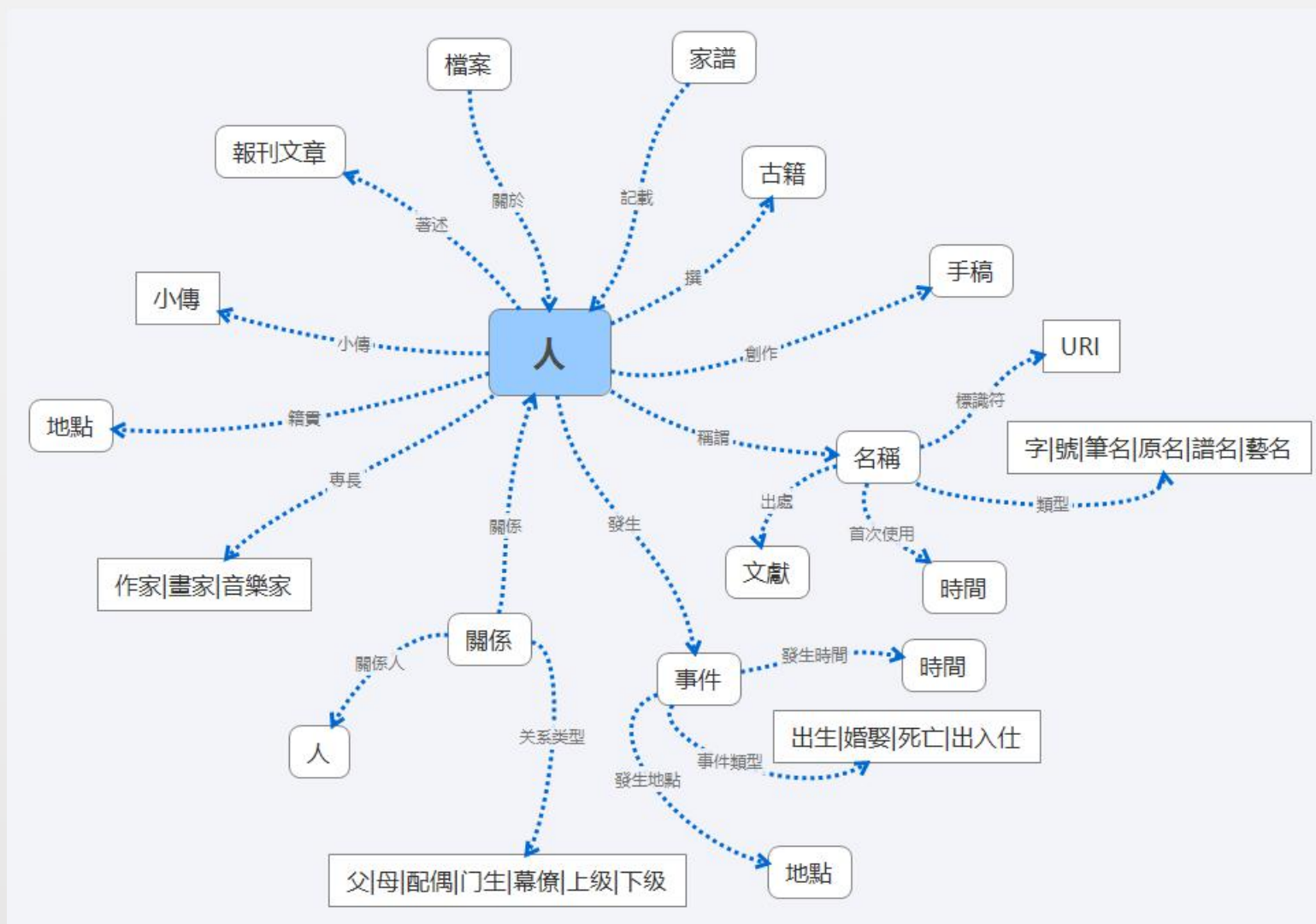
- 何时需要定义一个类?
- 对象属性还是数据属性?
(eg: 家谱姓氏)
- 复用、继承还是自定义?
- 如何理解属性的domain和range?
- 取值词表还是自由词?



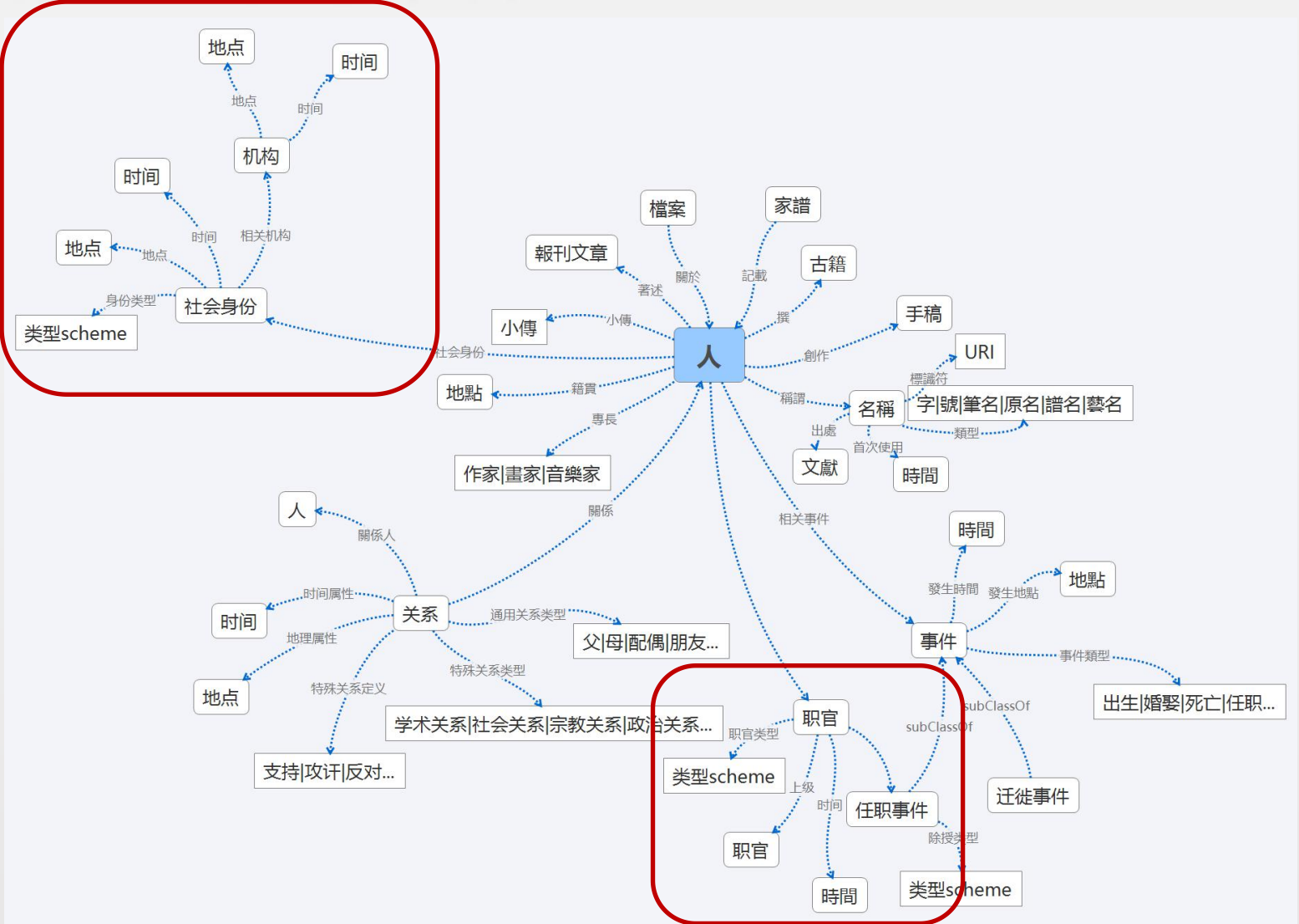
如何扩展已有的本体？



如何扩展已有的本体?



如何扩展已有的本体？



RDF数据的编码

- URI设计规范

URI是关联数据的第一原则；

也是“Linked Data”与“Knowledge Graph”的主要区别之一

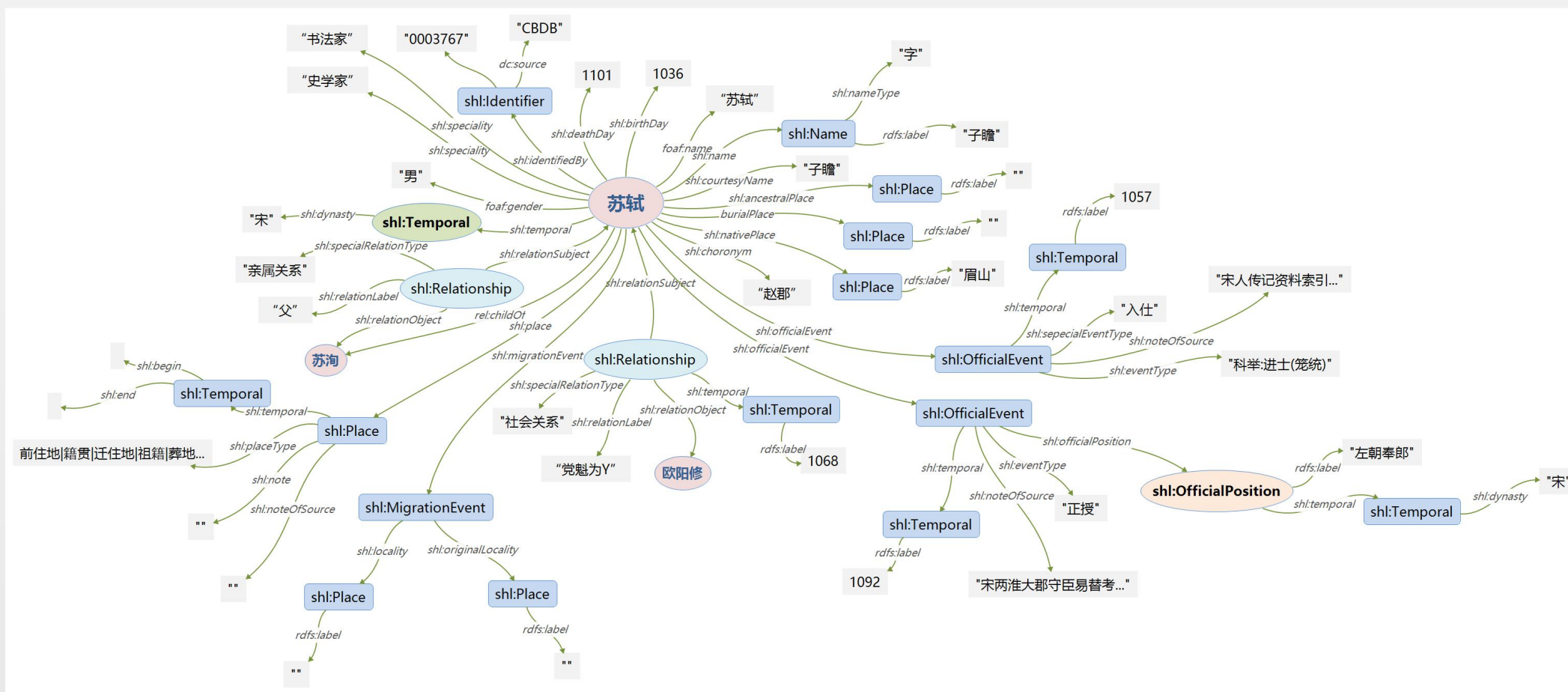
表 8 上海图书馆家谱 URI 设计模式

类型	设计模式	举例
本体	http:// domain /ontology/[class][property]	http://www. library. sh. cn/ontology/ 或: http://ontology. library. sh. cn/
规范数据	http:// domain /authority/ termtypel / term}	http://data. library. sh. cn/authority/CCT/A (中 国分类主题词表) http://data. library. sh. cn/authority/LGJ/000 (刘 国钧分类法) http://data. library. sh. cn/authority/FDC/jing (四部分类法) http://data. library. sh. cn/authority/familynamel / ding (姓氏)
实体	http:// domain /entity/ conceptl / IRI}	http://data. library. sh. cn/jp/entity/person/ 6ad6f4b81 (人) http://data. library. sh. cn/entity/organization/ 6ad6f4b81dadf (机构) http://data. library. sh. cn/entity/place/ dadfadfdafdd (地)
资源	http:// domain /resource/ conceptl / IRI}	http://data. library. sh. cn/jp/resource/work/ 6ad6f4b81 (创造性作品)
通用取值词表	http:// domain /vocab/ termtypel / term}	http://data. library. sh. cn/vocab/binding/jing - zhe - zhuang (经折装) http://data. library. sh. cn/vocab/edition/gao - ben (稿本)
家谱取值词表	http:// domain /vocab/jp/ termtypel / term}	http://data. library. sh. cn/jp/vocab/ancestor/shi - qian - zu (始迁祖)

RDF数据的编码

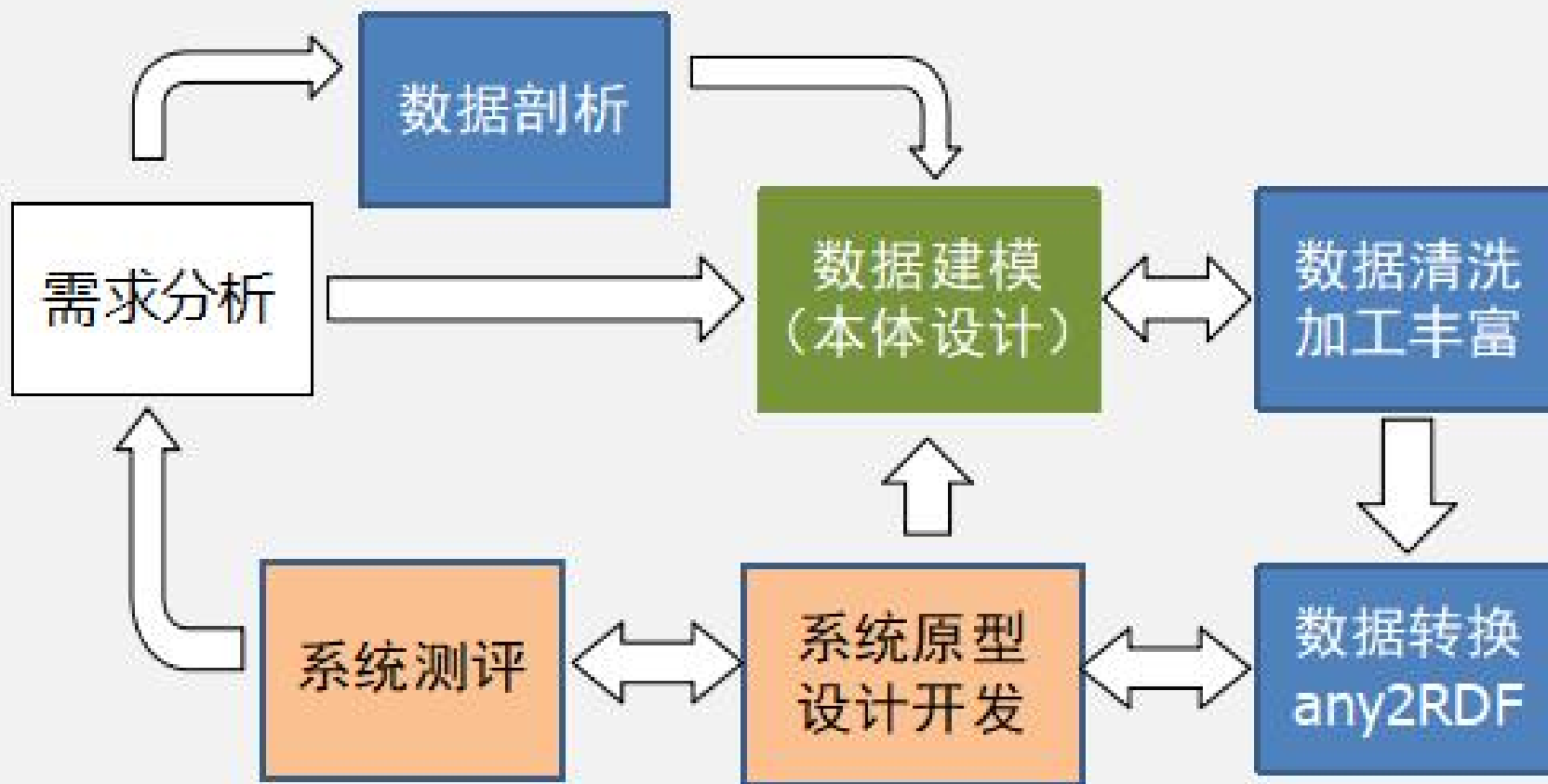
- 当range是一个类时，取值是URI还是空节点？
- 当range是一个父类时，如何取值？(bf:creator bf:Agent shl:Person)
- 多值，属性重复 (eg: 文献类型、专长)
- 一个trick：对象属性和数据属性同时存在（允许数据冗余）
 dc:creator rdfs:range **Literal**, bf:creator rdfs:range **shl:Person**;
 rel:friendOf, shl:Relationship--relationType --" 朋友"
- 定义本体与元数据的映射
 以名人手稿为例

用一个实例来验证RDF数据的编码



设计本体所需的知识储备

- 领域知识
- 用户需求
- 数据结构
- 系统需求
- 已有词表
- RDF编码
- SPARQL



本体是现实世界与数据世界之间的桥梁

数据建模的终极目的是以面向对象的思维尽可能地模拟现实世界

同时又能将错综复杂的现实世界用有序的数据表达出来

并使机器可以计算和理解

从而可以挖掘知识

最终获得智慧

臻至人机合一的化境

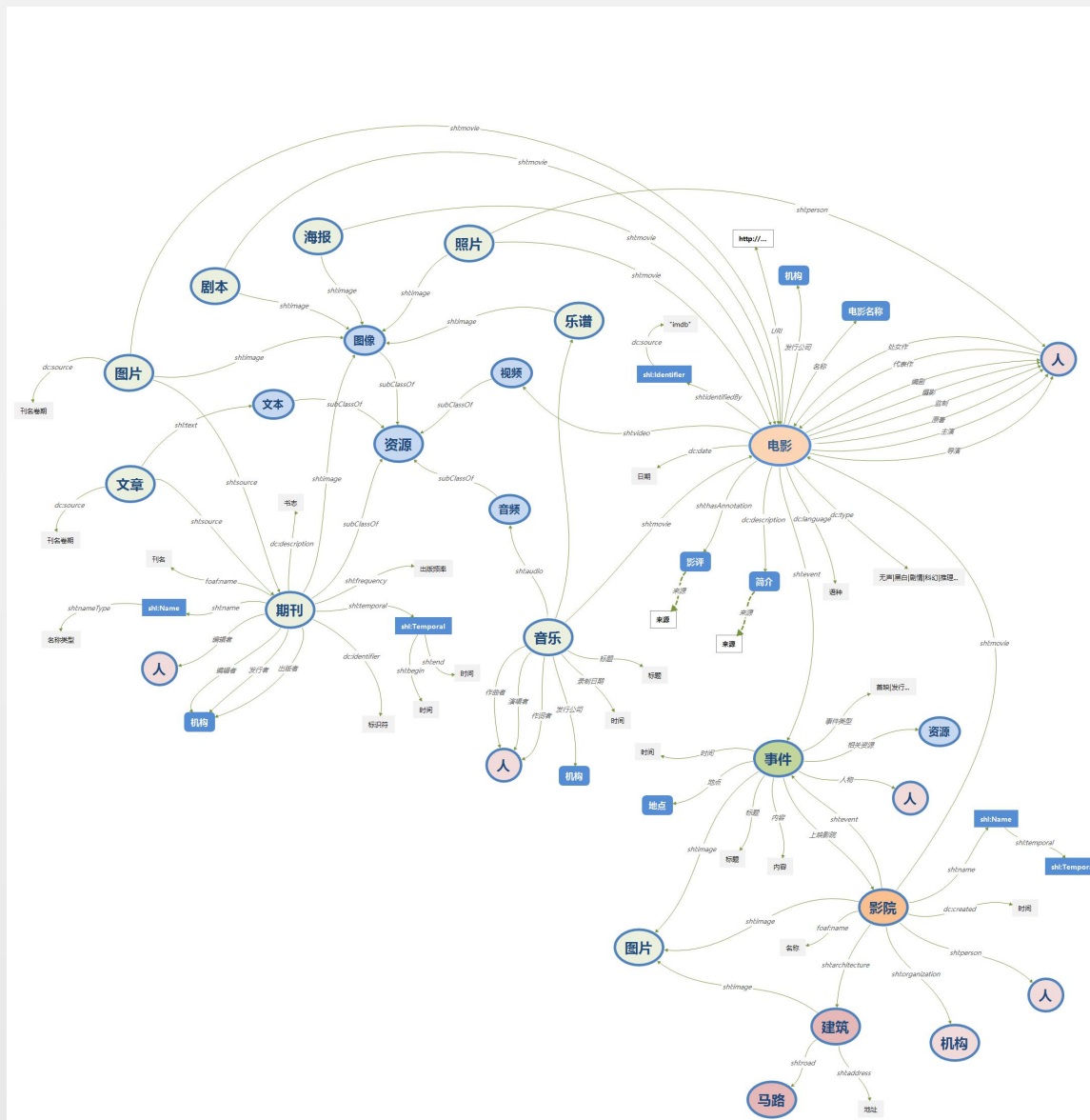
电影本体模型

分离内容和表现形式

电影不等于视频

音乐不等于音频

老照片不等于图像





谢谢！

xtykc@yeah.net