Publishing the Paul Kendel Fonoroff Collection for Chinese Film Studies as Linked Data: An Architecture of Distributed Data Services

Haiqing Lin

C.V. Starr East Asian Library

University of California,Berkeley

PRRLA 2019 MEETING

Korea University

September 1-4, 2019.

Outlines

- Background
- Why linked data? An architecture
- Design aspects of ontology, defined classes and singleton properties
- Query and services, SPARQL endpoint, relationships and data integration
- Challenges

Background

In 2016, UC Berkeley C.V. Starr East Asian Library has acquired a significant Chinese film studies collection. The collection contains over 70,000 items. The highlight of the collection includes 436 old film periodicals in 5910 issues and a large amount of film ephemera, i.e. 4195 film posters, 21233 lobby cards in 2,194 sets, 3332 theatre flyers, 9,214 photographic negatives & slides, 4,145 stills & publicity photos.

To make these valuable resources available to researchers, we are proposing a collaboration project with Shanghai Library to develop a Chinese film knowledge base based on two great valuable Chinese film collections, Shanghai Library film collection and our collection.

Example of materials: Film Posters

1957?

导演: 刘国权

1957

导演: 石挥

1957

导演: 留铿

1958

导演: 黄佐临

1958

导演: 王家乙

1959

导演: 郑君里

1959

编导: 查列, 张永枚等

1959

导演: 李俊

1959

Scripts, Dialogues

星星月亮太阳 "Sun, Moon and Star" (对白本)

空中小姐：对白本
晶文

倩女幽魂：對白本

给我一个吻: 剧务本
李翰祥; 编剧：程刚

夜半歌声 (对白本)
主要演员: 乐蒂，赵雷等

火烧红莲寺：第三集
监制: 庄明枢; 编剧: 凌云; 导演: 王天林; 助导: 石慧

挺进！红卫兵：征求意见稿
编剧: 岳桦

Flyers


街頭巷尾
張伐
黃宗英
章曼蘋
張雁
潘子農編導
女教師跳出教育圈
獲得勞工神聖代價
★打破階級戀愛
★協倡自力更生
男教員踏起三輪車
解決飲食男女問題
三輪車夫
公園拍拖
上流人物
大動公憤
中山影院廣告
中央電影企業
國語輕鬆鉅片

Complexities of the Collection

The collection contains a large amount of various materials including monographs, periodicals, still images, posters, lobby cards, and other prints, and objects.

- Enormous quantity, more than 70000 items
- Various formats and types
- Covers entire live cycle of film, form production to exhibition.

Linked Data, Definitions

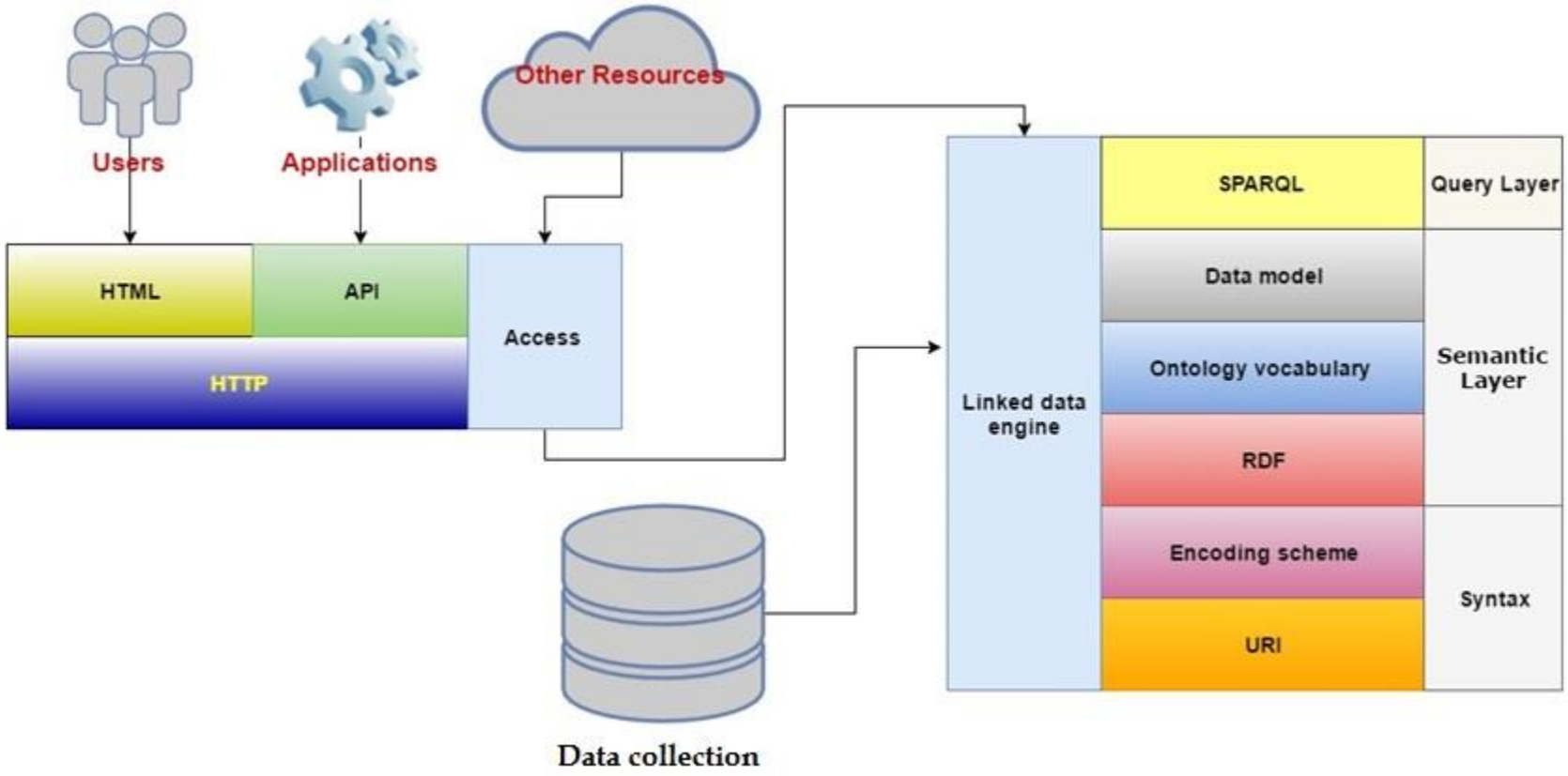
- Old definition in Wikipedia:

  a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data , information , and knowledge on the Semantic Web using URIs and RDF.

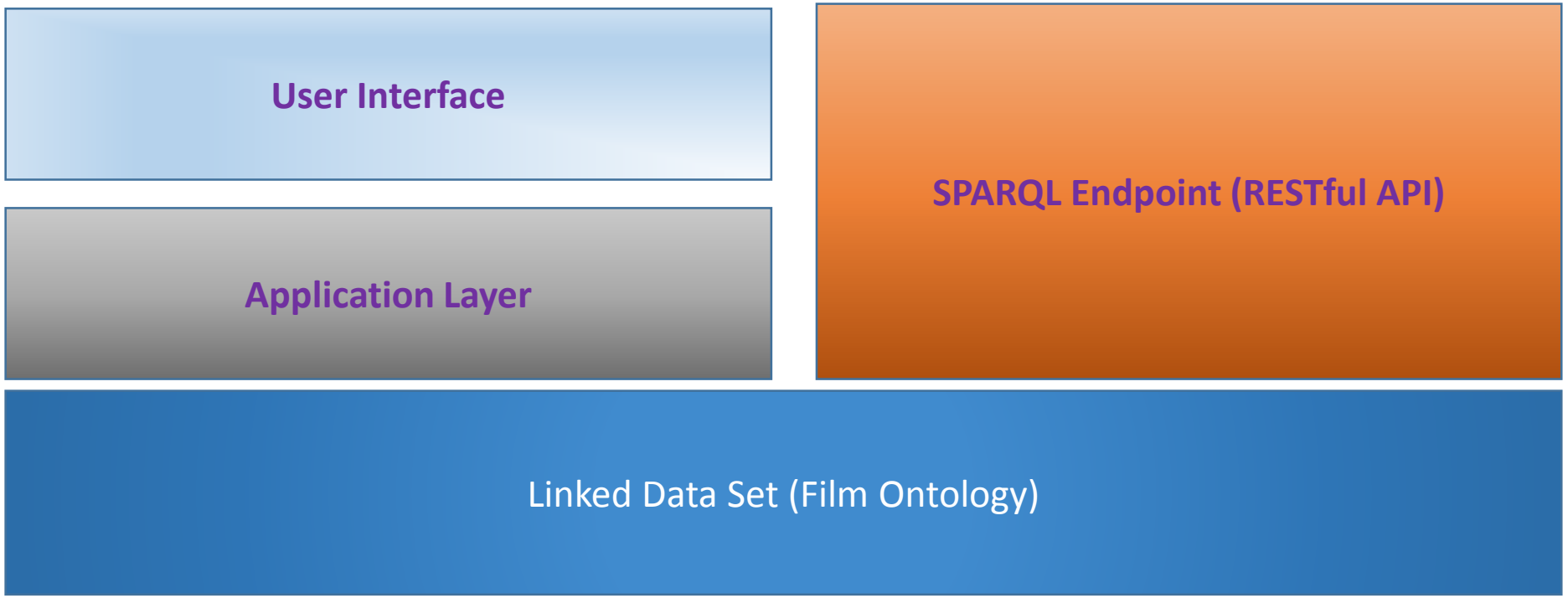
- New definition in Wikipedia:

  In computing, linked data (often capitalized as Linked Data) is structured data which is interlinked with other data so it becomes more useful through semantic queries.[1] It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only **for human readers**, it extends them to share information in a way that can be **read automatically by computers**.

Linked Data as a Solution



Architecture

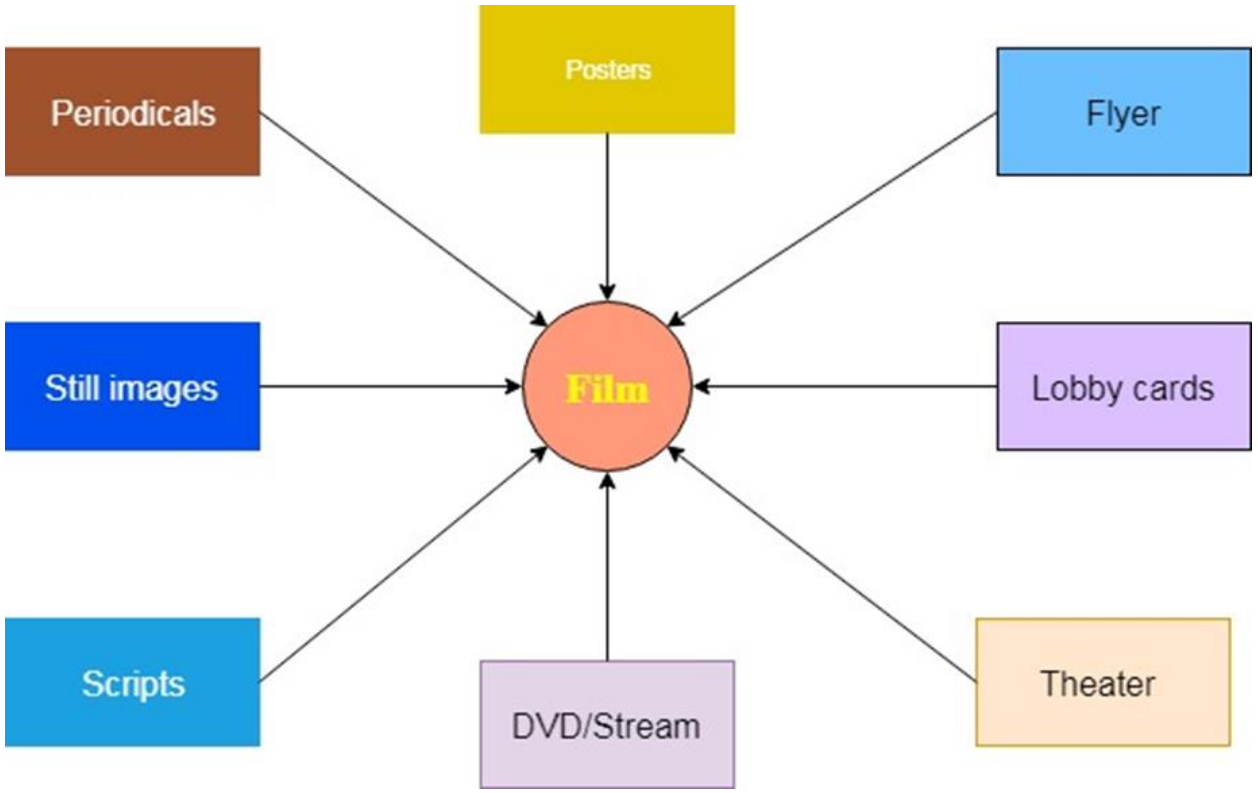


Virtuoso as Triple Store

- Virtuoso Universal Server is a middleware and database engine hybrid that combines the functionality of a traditional Relational database management system (RDBMS), Object-relational database (ORDBMS), virtual database, RDF, XML, free-text, web application server and file server functionality in a single system. The software has been developed by OpenLink Software. (Wikipedia)

Data Model

- Film-centered approach
- Film represented by film materials
- Focused on relationships among film entities
- Life-cycle modeling of films



Film-centered

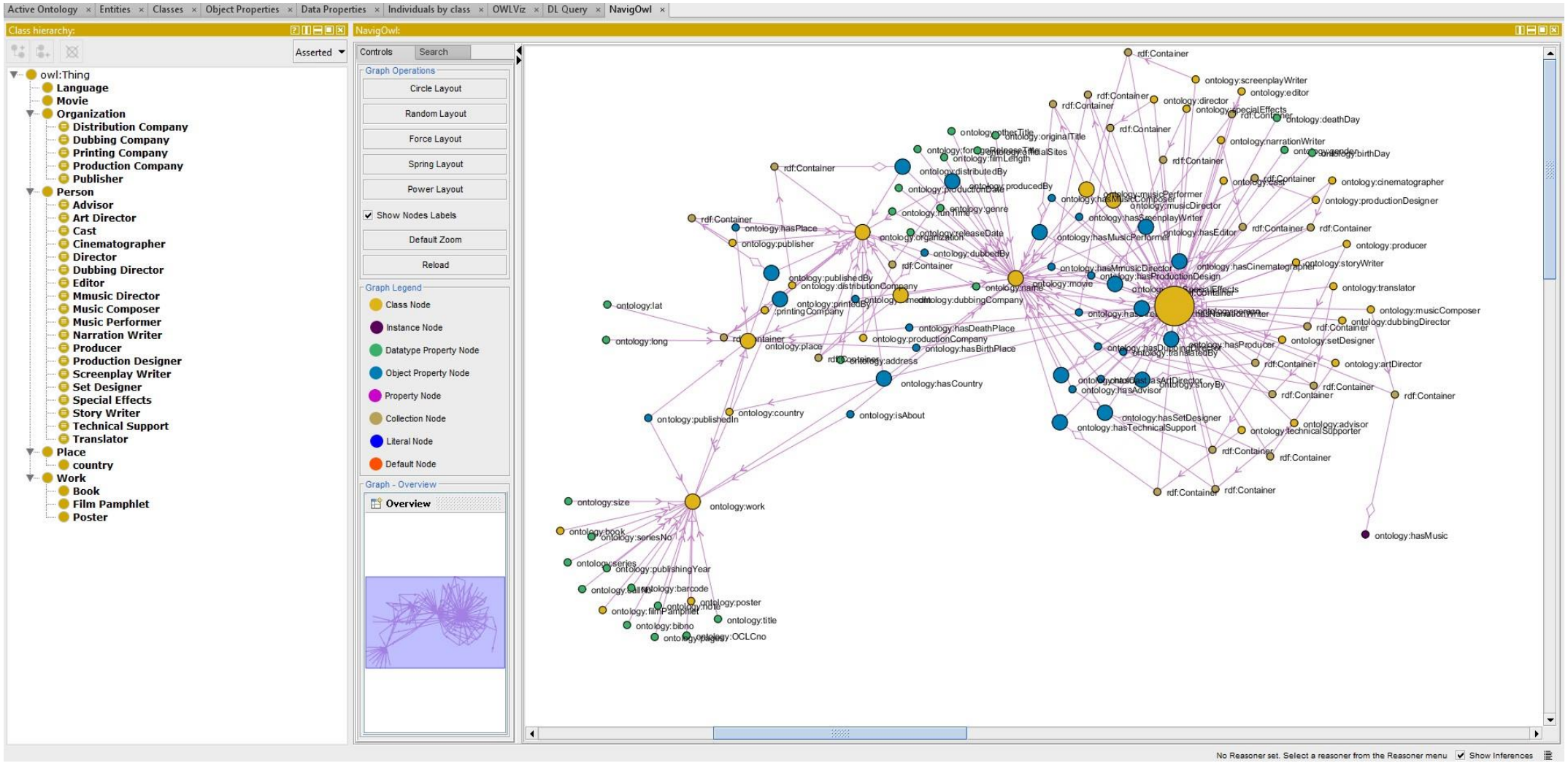
- RDF: subject predicate object

A film is always a subject when describing something related to a film .

Film Ontology

- Ontology is an unified framework to describe the resources and formalize relationships among film entities as well as film resources, not just the film itself. It is a foundation of query and inference algorithms of the system.
- The ontology will also be used to achieve semantic interoperability with other collections. (OWL:SameAs)

Protégé View



Generate RDF triples

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1		9 title	bibno	OCLCno	callNo	barcode	printedBy	note	size	isAbout	originalTit	originalTit	originalTit	foreignRe	hasDirector	dubbingD	hasP
2	POS02354	'91童黨							20 x 30 cm	MO003654	'91童黨	'91童党	Gamgs '91		曹建南		鄭順
3	POS02355	'93 似夢年華							21 x 31 cm	MO003655	'93 似夢年	'93 似梦年华					
4	POS02356	'93 似夢年華							21 x 31 cm	MO003655	'93 似夢年	'93 似梦年华					
5	POS00004	13号地区							103 x 74 cm	MO001349	13號地區	13号地区			陈家林		
6	POS00005	1918年苦难的历程(第二部)				C111343211		内容简	53 x 38 cm	MO001350	1918年苦難的歷程(第二部)			1918年苦难的历程(第二部)			
7	POS00006	2 x 2=5				C111345079			54 x 39 cm	MO001351	2 x 2=5			2 x 2=5			
8	POS00008	306号案件		1E+09	2/13/2(	C118638860			78 x 54 cm	MO001352	306號案件			36号案件	A·雷巴闥夫		
9	POS00007	306号案件	b23388842	1E+09	Cabine	C111344297			77 x 53 cm	MO001352	306號案件			36号案件	A·雷巴闥夫		
10	POS02357	3個受傷的警察							40 x 61 cm	MO003664	3個受傷的	3个受伤的	The Log				
11	POS02358	3個茶煲1個佬							40 x 61 cm	MO003665	3個茶煲1個佬		Multiplici	3个茶煲1	夏萊藍米斯		
12	POS02359	405謀殺案							31 x 40 cm	MO003666	405謀殺案	405谋杀案			沈耀庭		
13	POS02360	405謀殺案							21 x 30 cm	MO003666	405謀殺案	405谋杀案			沈耀庭		
14	POS02361	405謀殺案							21 x 30 cm	MO003666	405謀殺案	405谋杀案			沈耀庭		
15	POS00009	45号地区			2/13/2(	C111359031/	国营天津人民印刷厂		104 x 77 cm, 77 x 53	MO001354	45號地區			45号地区	ю·维申斯基		
16	POS00010	45号地区				C111540817	国营天津人民印刷厂·		104 x 78 cm	MO001354	45號地區			45号地区	ю·维申斯基		
17	POS02362	4個婚禮一個葬禮							40 x 61 cm	MO003669	4個婚禮一個葬禮		Four Wed	4个婚礼一	米克里维		
18	POS02363	4面夏娃							40 x 61 cm	MO003670	4面夏娃	4面夏娃	4 Faces of Eve		甘國亮, 林		
19	POS00011	51号兵站				C111347758			78 x 53 cm	MO001356	51號兵站	51号兵站			刘琼		
20	POS02364	666魔鬼復活							26 x 40 cm	MO003671	666魔鬼復	666魔鬼复	Satan Returns		林偉倫		蔡忠
21	POS02366	666魔鬼復活							20 x 30 cm	MO003671	666魔鬼復	666魔鬼复	Satan Returns		林偉倫		蔡忠
22	POS02365	666魔鬼復活							27 x 40 cm	MO003671	666魔鬼復	666魔鬼复	Satan Returns		林偉倫		蔡忠
23	POS02367	666魔鬼復活							20 x 30 cm	MO003671	666魔鬼復	666魔鬼复	Satan Returns		林偉倫		蔡忠
24	POS02368	7金剛							13 x 19 cm	MO003675	7金剛	7金刚	Wonder Seven		程小東		劉倩
25	POS02369	91家有囍事							40 x 61 cm	MO003676	91家有囍	91家有囍	All's Well End's Well		張堅庭		
26	POS02370	92黑玫瑰与黑玫瑰							30 x 20 cm	MO003677	92黑玫瑰	92黑玫瑰	92 Legendary La Ros		劉鎮偉		
27	POS02371	97古惑仔戰無不勝							40 x 61 cm	MO003678	97古惑仔	97古惑仔	Young and Dangerou		劉偉強		張敏
28	POS02372	97古惑仔戰無不勝							27 x 40 cm	MO003678	97古惑仔	97古惑仔	Young and Dangerou		劉偉強		張敏
29	POS02378	B計劃							25 x 39 cm	MO003685	B計劃	B计划	Extreme Crisis		羅禮賢		

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:flm="http://cnfilmstudies.online/ontology/"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<flm:poster rdf:about="http://cnfilmstudies.online/filmdata/POS00001">
<flm:isAbout rdf:resource="http://cnfilmstudies.online/filmdata/MO001343" />
<flm:title>"特快"列车</flm:title>
<flm:size>53 x 34 cm</flm:size>
</flm:poster>
<flm:movie rdf:about="http://cnfilmstudies.online/filmdata/MO001343">
<flm:originaltitlePOS00001 xml:lang="zh">特快列車</flm:originaltitlePOS00001>
<flm:originaltitlePOS00001 xml:lang="zh">特快列车</flm:originaltitlePOS00001>
m:distributedByPOS00001 rdf:resource="http://cnfilmstudies.online/filmdata/ORG00938" />
producedByPOS00001 rdf:resource="http://cnfilmstudies.online/filmdata/ORG00192" />
movie>
escription rdf:about="http://cnfilmstudies.online/ontology/originaltitlePOS00001" >
s:subPropertyOf rdf:resource="http://cnfilmstudies.online/ontology/originaltitle" />
m:hasSource rdf:resource="http://cnfilmstudies.online/filmdata/POS00001"/>
/rdf:Description>
<rdf:Description rdf:about="http://cnfilmstudies.online/ontology/distributedByPOS00001" >
<rdfs:subPropertyOf rdf:resource="http://cnfilmstudies.online/ontology/distributedBy" />
<flm:hasSource rdf:resource="http://cnfilmstudies.online/filmdata/POS00001"/>
</rdf:Description>
<rdf:Description rdf:about="http://cnfilmstudies.online/ontology/producedByPOS00001" >
<rdfs:subPropertyOf rdf:resource="http://cnfilmstudies.online/ontology/producedBy" />
<flm:hasSource rdf:resource="http://cnfilmstudies.online/filmdata/POS00001"/>
</rdf:Description>
<flm:poster rdf:about="http://cnfilmstudies.online/filmdata/POS00002">
<flm:isAbout rdf:resource="http://cnfilmstudies.online/filmdata/MO001343" />
<flm:title>"特快"列车</flm:title>
<flm:size>73 x 53 cm</flm:size>
</flm:poster>
<flm:movie rdf:about="http://cnfilmstudies.online/filmdata/MO001343">
<flm:originaltitlePOS00002 xml:lang="zh">特快列車</flm:originaltitlePOS00002>
<flm:originaltitlePOS00002 xml:lang="zh">特快列车</flm:originaltitlePOS00002>
<flm:distributedByPOS00002 rdf:resource="http://cnfilmstudies.online/filmdata/ORG00938" />
<flm:producedByPOS00002 rdf:resource="http://cnfilmstudies.online/filmdata/ORG00192" />
</flm:movie>
```

Size of the data set

Materials

Poster	Flyer	Script	Pamphlet	Theaters (not finished)
4180	5	154	1342	41

Objects

Film	Person	Organization	Country
3980	9840	1850	56

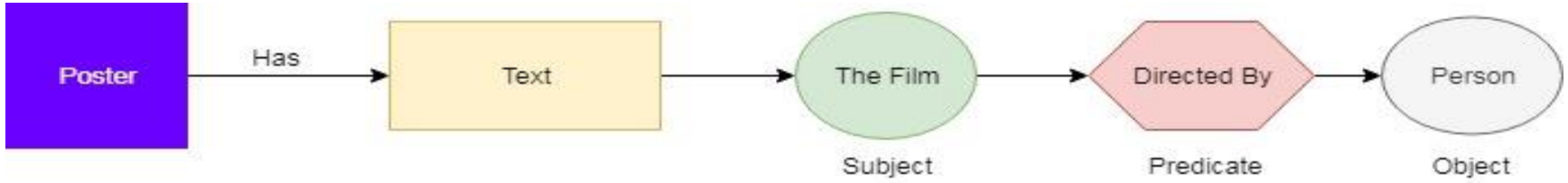
Triples

Triples
251609

Provenance

- Representing complicated sentences like

  “The text on the poster said ‘XXXX directed by YYY’ ”



五朵金花 导演 王家乙

Singleton property

Subject	Predicate	Object
BobDylan	isMarriedTo	SaraLownds
BobDylan	isMarriedTo#1	SaraLownds
isMarriedTo#1	rdf:singletonPropertyOf	isMarriedTo
isMarriedTo#1	hasStart	1965-11-22
isMarriedTo#2	hasEnd	1973-12-03

Source: Don’t Like RDF Reification? Making Statements about Statements Using Singleton Property by Vinh Nguyen, Olivier Bodenreider, and Amit Sheth, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4350149/ on 10/23/2018 Subject

Singleton Property Presentation

- How to represent

五朵金花 导演 王家乙

```
<http://cnfilmstudies.online/filmdata/MO001551>
      a                    flm:movie ;
      flm:hasDirectorPOS00203    <http://cnfilmstudies.online/filmdata/PN00003356> ;
      flm:originaltitlePOS00203  "五朵金花"@zh .

<http://cnfilmstudies.online/ontology/originaltitlePOS00203>
      flm:hasSource       <http://cnfilmstudies.online/filmdata/POS00203> ;
      rdfs:subPropertyOf  <http://cnfilmstudies.online/ontology/originaltitle> .

<http://cnfilmstudies.online/filmdata/POS00203>
      a           flm:poster ;
      flm:isAbout  <http://cnfilmstudies.online/filmdata/MO001551> .

<http://cnfilmstudies.online/ontology/hasDirectorPOS00203>
      flm:hasSource       <http://cnfilmstudies.online/filmdata/POS00203> ;
      rdfs:subPropertyOf  <http://cnfilmstudies.online/ontology/hasDirector> .
```

Defined Classes

- Director ≡ person ∩ ∃ hasDirector.movie

<owl:Class rdf:about="http://cnfilmstudies.online/ontology/director">

<rdfs:label xml:lang="en">Director</rdfs:label>

<rdfs:comment xml:lang="en">The person with the overall responsibility for all creative and technical aspects of making a film</rdfs:comment>

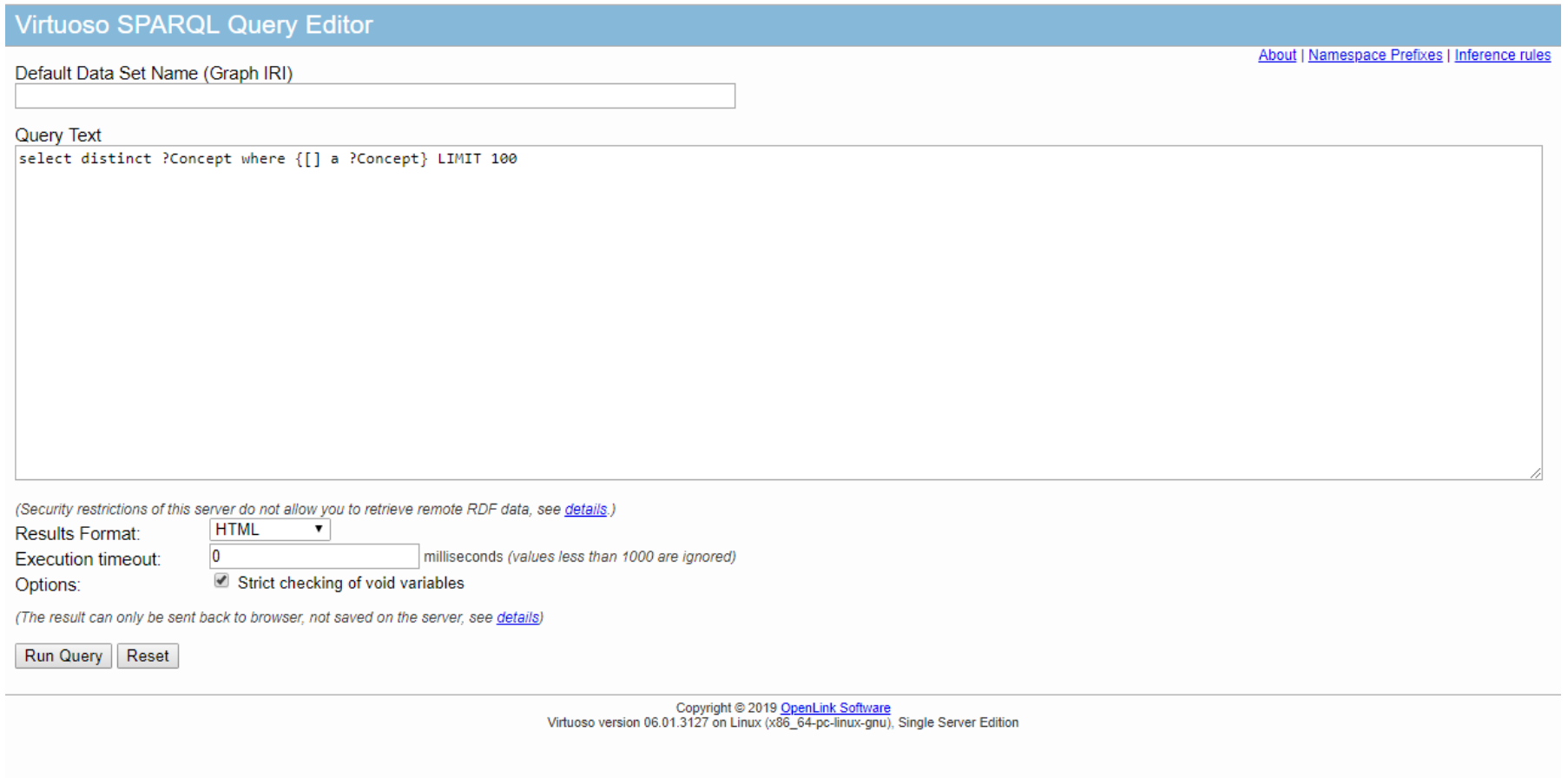
<rdfs:subClassOf rdf:resource="http://cnfilmstudies.online/ontology/person"/>

<owl:equivalentClass><owl:Restriction><owl:someValuesFrom rdf:resource="http://cnfilmstudies.online/ontology/person"/>

<owl:onProperty rdf:resource="http://cnfilmstudies.online/ontology/hasDirector"/></owl:Restriction></owl:equivalentClass>

</owl:Class>

SPARQL Endpoint



Access Sparql Endpoint via Excel

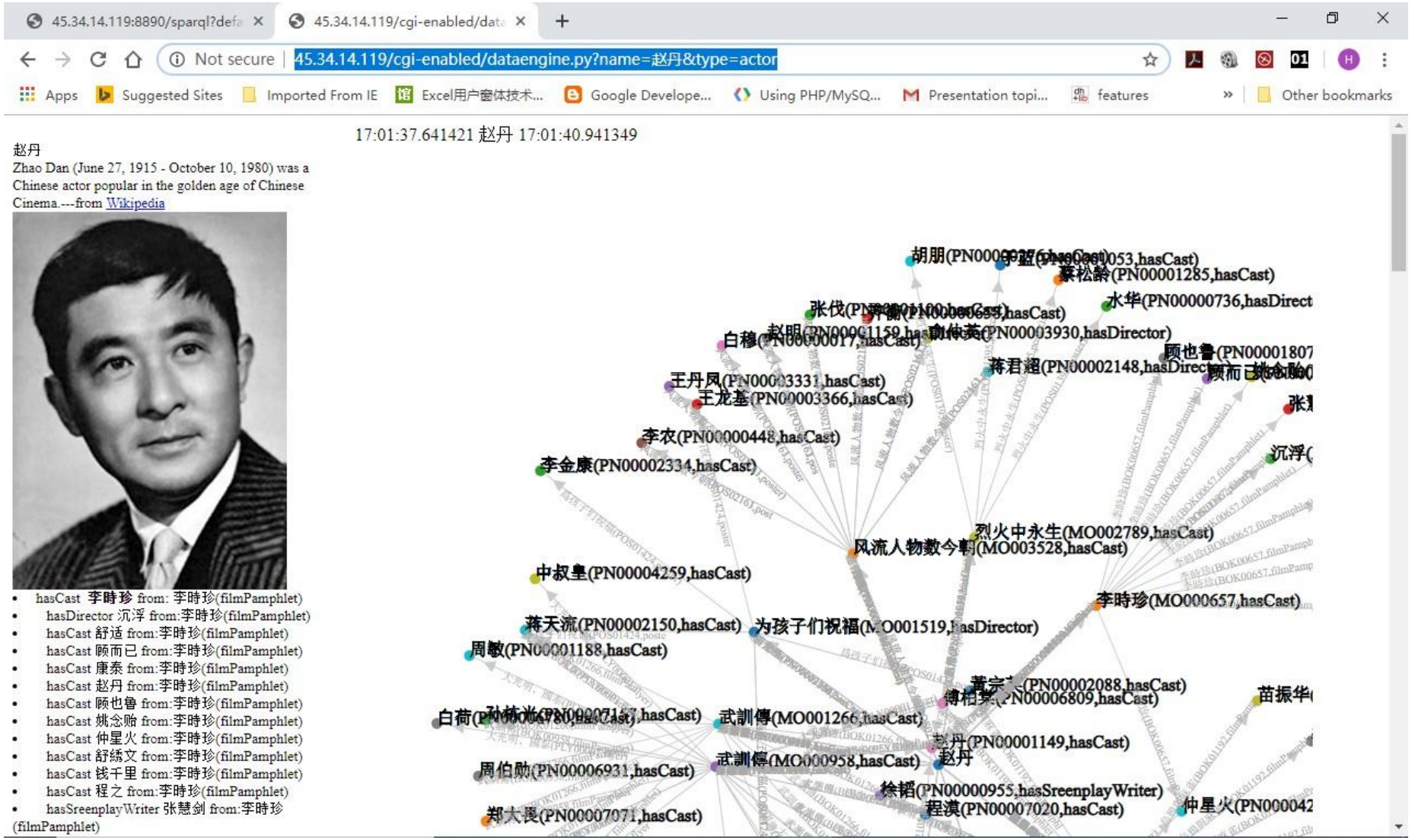
```
Function sparql()
Dim sparqlUrl As String
    sparqlUrl = “http://45.34.14.119:8890/sparql?default-graph-uri=&query=%0D%0Aselect+distinct+%3FConcept+where+%7B%5B%5D+a+%3FConcept%7D+LIMIT+100&format=text%2Fhtml&timeout=0&debug=on”

Set sparqlReq = CreateObject("MSXML2.XMLHTTP")
    With sparqlReq
        .Open "GET", sparqlUrl, False
        .Send
    End With

sparql=sparqlReq.ResponseText
```

Architecture of Application Layer

Application Layer Example



SPARQL Template for relationships

```
PREFIX flm:<http://cnfilmstudies.online/ontology/>
PREFIX shl:<http://www.library.sh.cn/ontology/>
select distinct ?qsstitle ?qsst ?c ?fm ?qs ?bn ?ws ?wsstitle ?wsst where {{?a a flm:person.
                ?b a flm:person.
                ?c a flm:movie.
                ?c ?q ?a.
                ?c ?w ?b.
                ?a flm:name "赵丹"@zh.
                ?b flm:name ?bn.
                ?q rdfs:subPropertyOf ?qs.
                ?q flm:hasSource ?qss.
                ?qss a ?qsst.
                ?qss flm:title ?qsstitle.
                ?w rdfs:subPropertyOf ?ws.
                ?w flm:hasSource ?wss.
                ?wss a ?wsst.
                ?wss flm:title ?wsstitle.
                } OPTIONAL {
                  ?c ?t ?fm.
                  ?t rdfs:subPropertyOf flm:originaltitle}
•                   }
```

SPARQL Template for Defined Classes

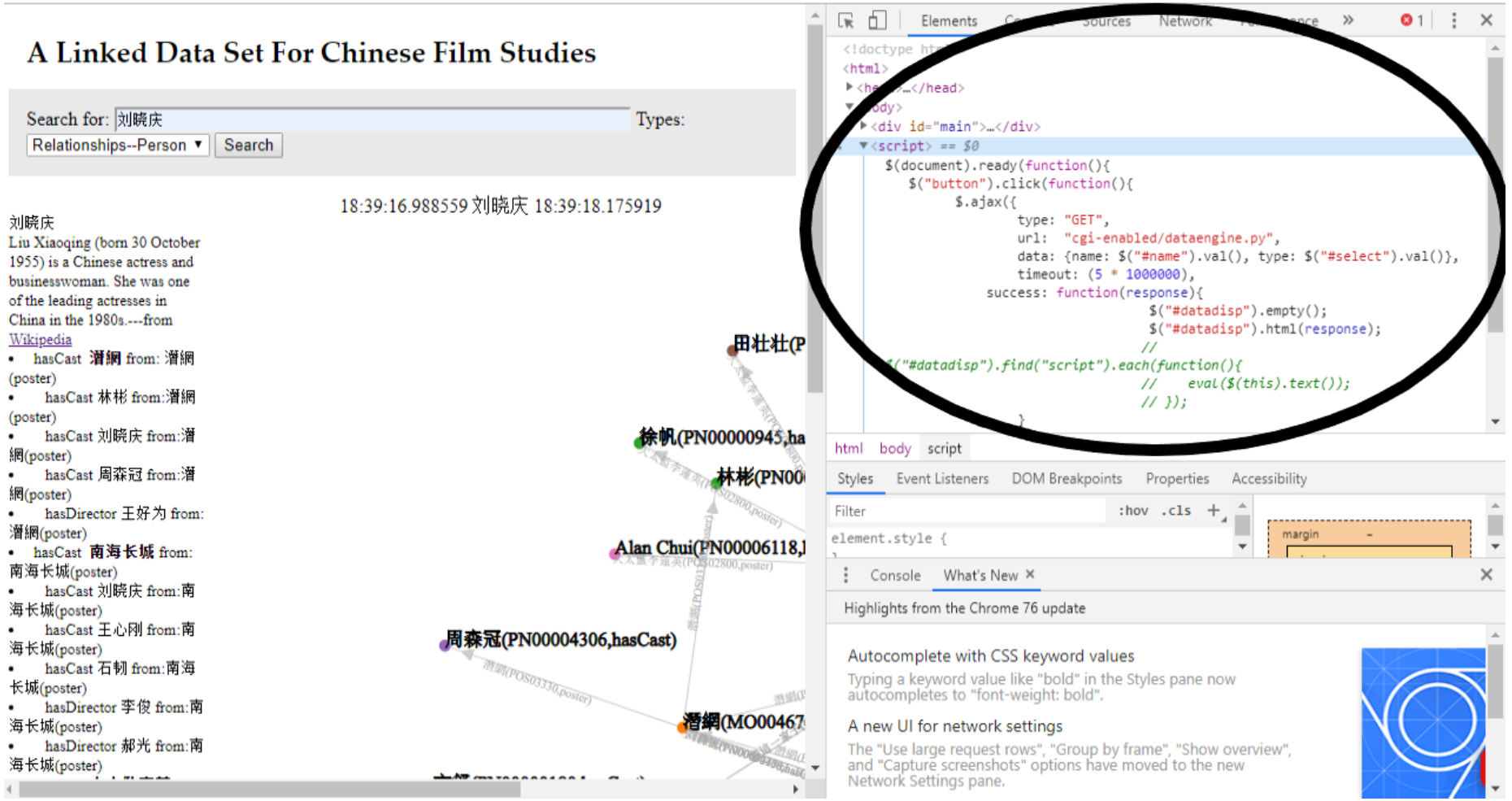
```
prefix flm:<http://cnfilmstudies.online/ontology/>
select distinct ?cn ?c where {?a a flm:movie .
                    ?a ?q ?c.
                    ?q rdfs:subPropertyOf ?u.
                    ?class owl:equivalentClass ?r.
                    ?r owl:onProperty ?u.
                    ?class rdfs:label ?cn}  LIMIT 100
```

SPARQL Template for Data Integration : Federated Query

- Ability to take a query based on information from many different sources.

```
prefix shl: <http://www.library.sh.cn/ontology/>
prefix flm:<http://cnfilmstudies.online/ontology/>
select distinct * where{?a shl:actor ?b .
                        ?b flm:name "赵丹"@zh.
                        ?a foaf:name ?t. SERVICE <http://data1.library.sh.cn:8890/sparql> {?x dc:title ?t}
. SERVICE <http://dbpedia.org/sparql> {?w rdfs:label "赵丹"@zh.
                                       ?w a <http://dbpedia.org/ontology/Artist>.
                                       ?w rdfs:comment ?comm. filter(lang(?comm)="zh")}}
```

As a Service: an AJAX example



User Interface

Working with researchers

- Challenges
- Products vs. Platform
- Inputs from researchers

Thank PRRLA for The Karl Lo Award,

and

Thank you all

- http://45.34.14.119/cgi-enabled/dataengine.py?name=%E8%B5%B5%E4%B8%B9&type=person
- http://45.34.14.119/cgi-enabled/dataengine.py?name=%E8%B5%B5%E4%B8%B9&type=actor
- http://45.34.14.119/cgi-enabled/dataengine.py?name=刘晓庆|姜文&type=actor