关联数据应用：知识发现

SinoPedia

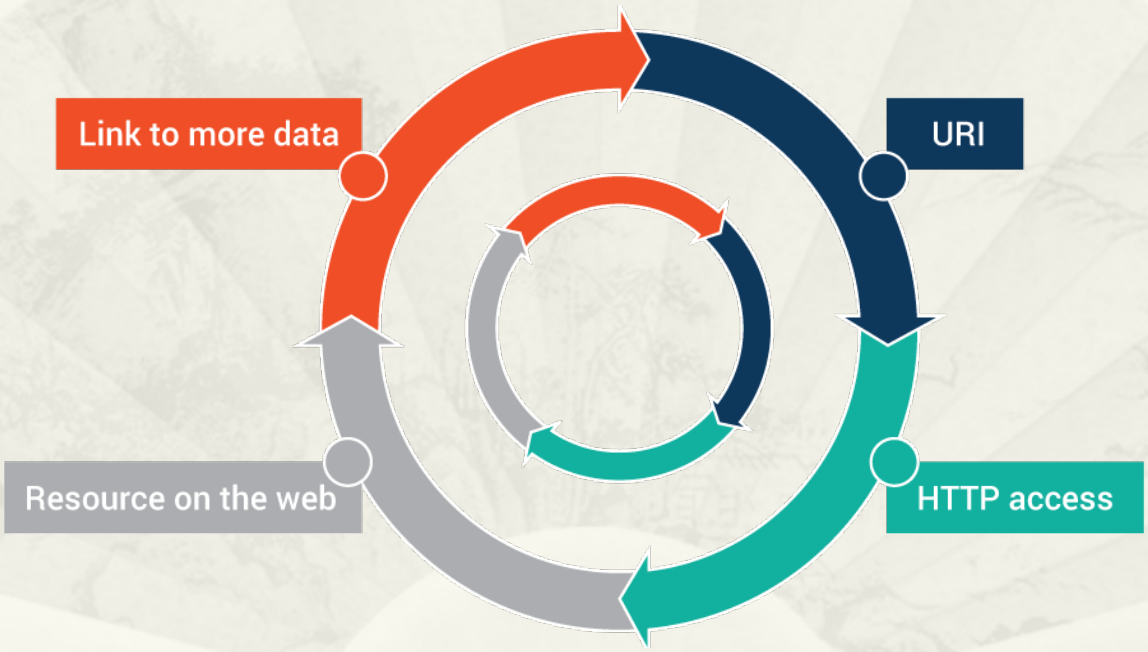
陈 涛
上海图书馆
2018-05-23

提纲

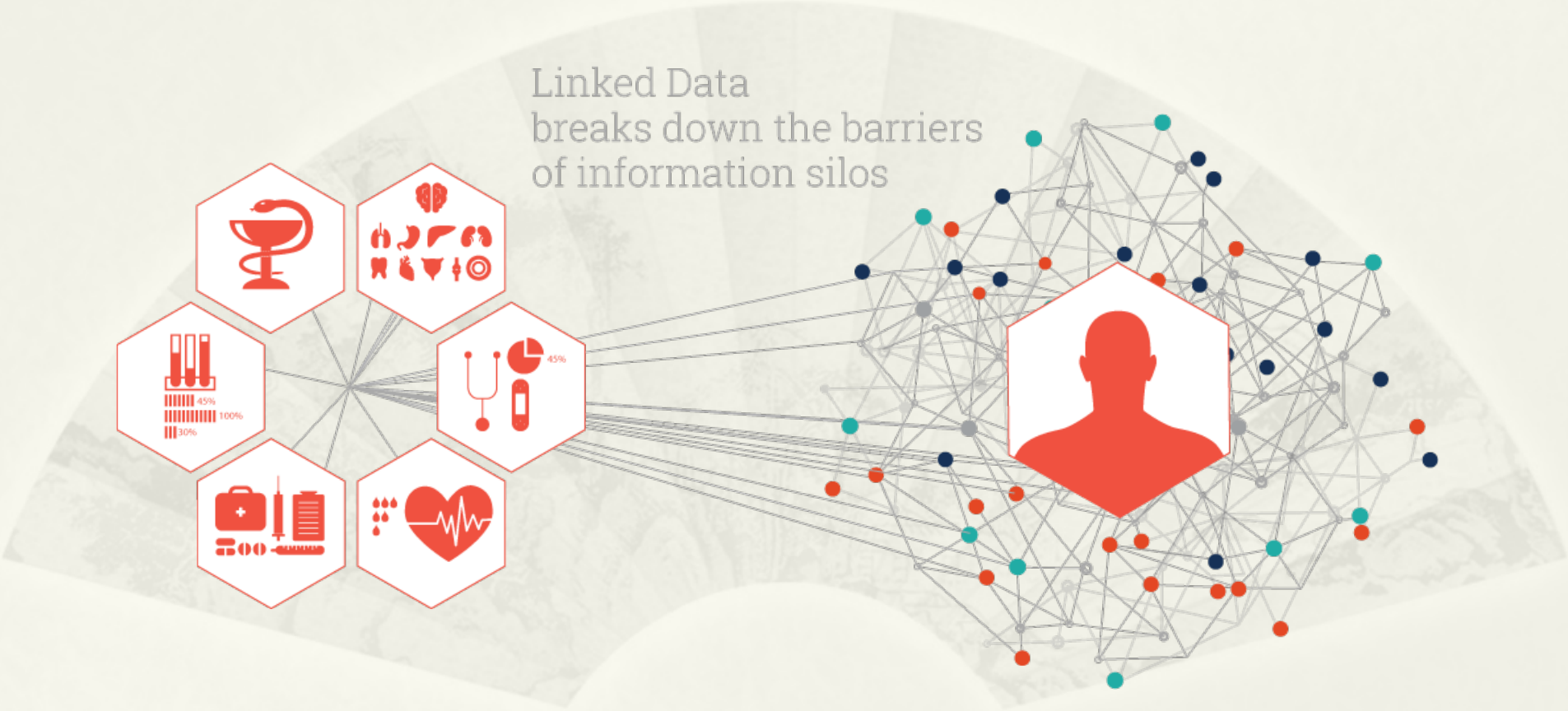
1. 关联数据
2. SinoPedia知识库
3. 如何使用关联数据
4. 知识图谱与知识发现

关联数据

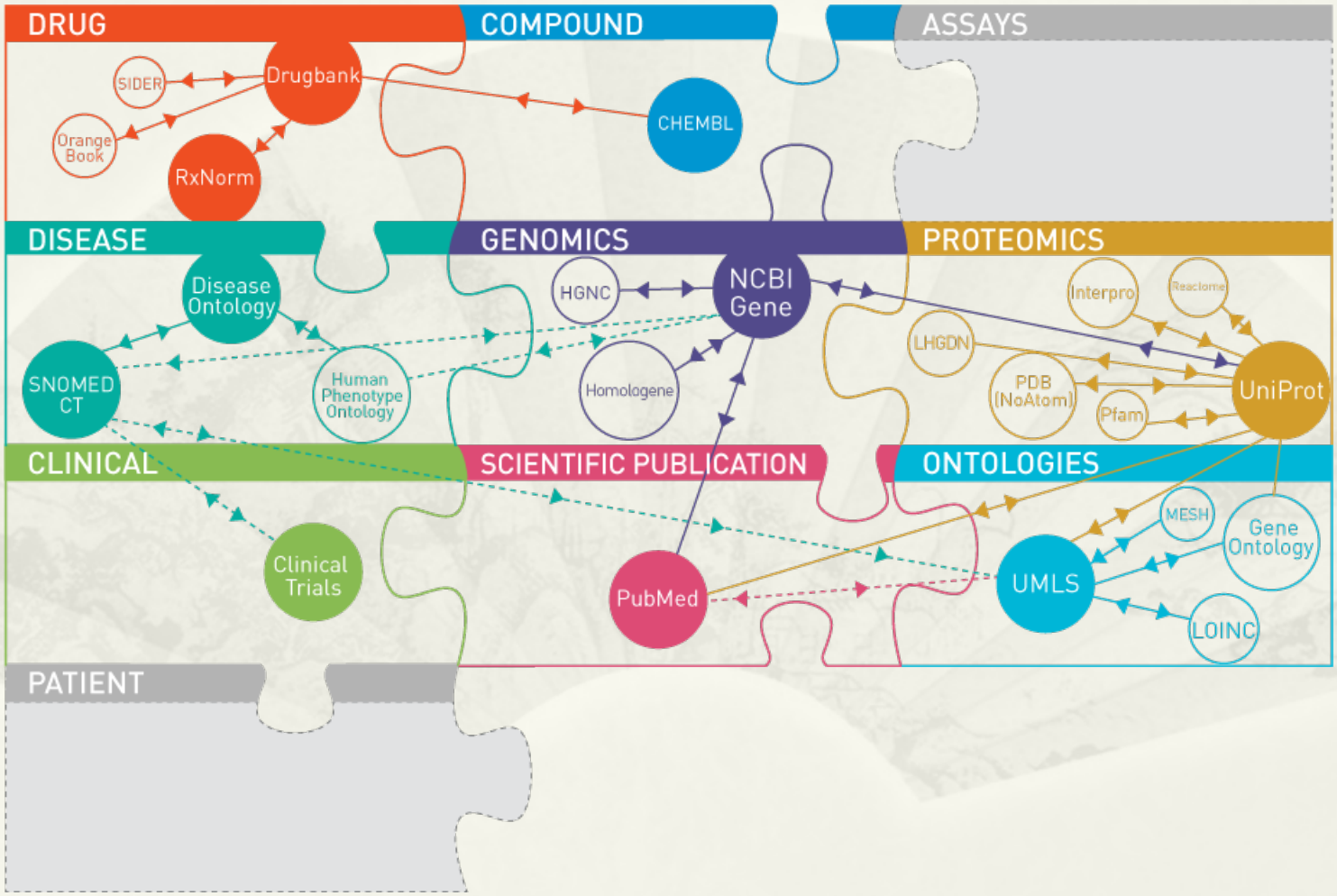
关联数据是在网络上发布和共享数据的推荐方法

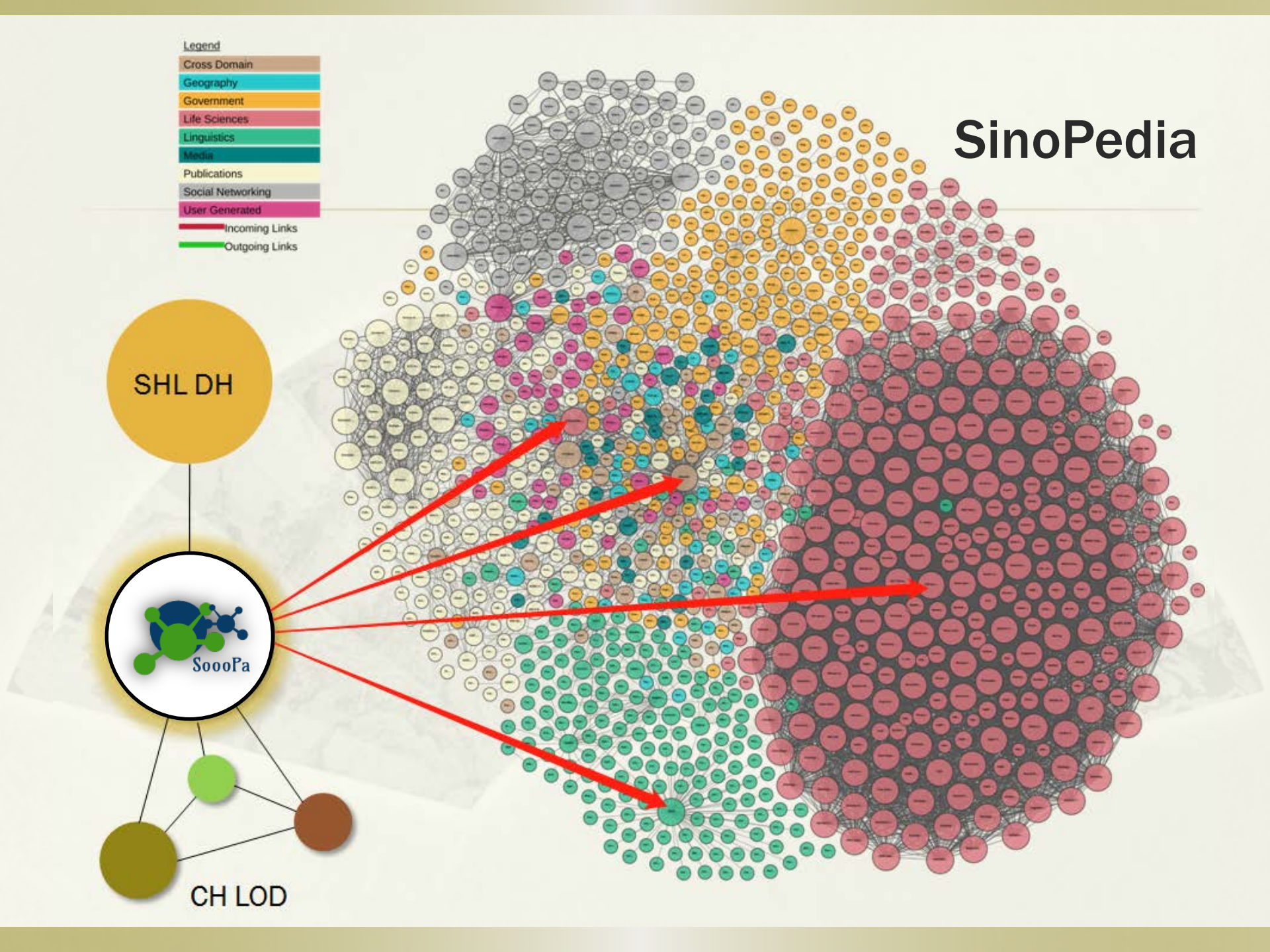


关联数据作用

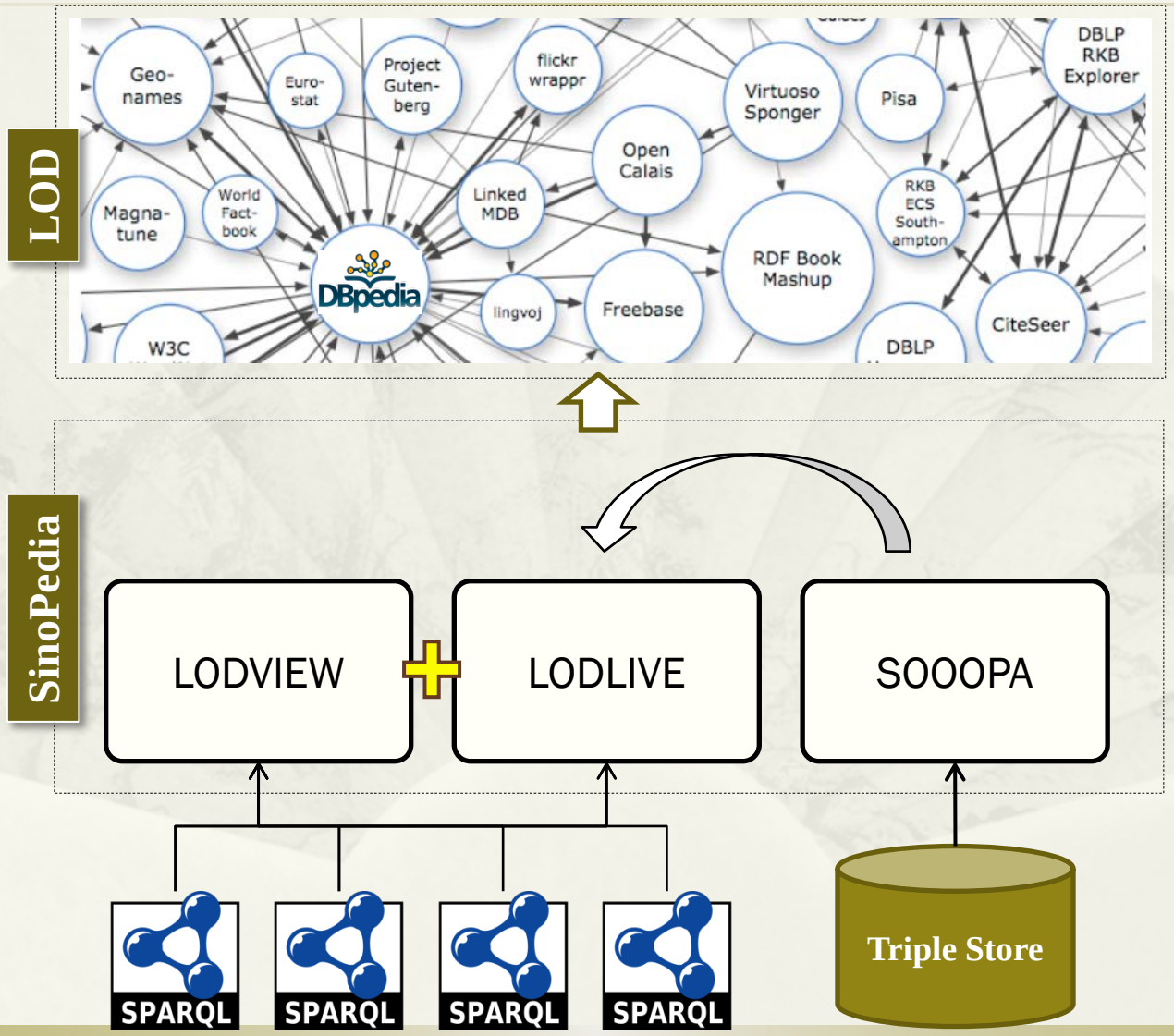


知识关联

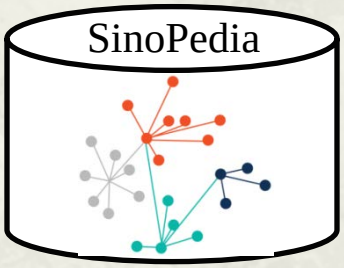

DRUG
COMPOUND
ASSAYS
SIDER
Drugbank
Orange Book
RxNorm
CHEMBL
DISEASE
GENOMICS
PROTEOMICS
Disease Ontology
HGNC
NCBI Gene
Interpro
Reactome
LHGDN
SNOMED CT
Human Phenotype Ontology
Homologene
PDB (NoAtom)
Pfam
UniProt
CLINICAL
SCIENTIFIC PUBLICATION
ONTOLOGIES
MESH
Gene Ontology
Clinical Trials
PubMed
UMLS
LOINC
PATIENT


Legend
Cross Domain
Geography
Government
Life Sciences
Linguistics
Media
Publications
Social Networking
User Generated
Incoming Links
Outgoing Links
SinoPedia
SHL DH
SoooPa
CH LOD

SinoPedia框架



SinoPedia数据来源



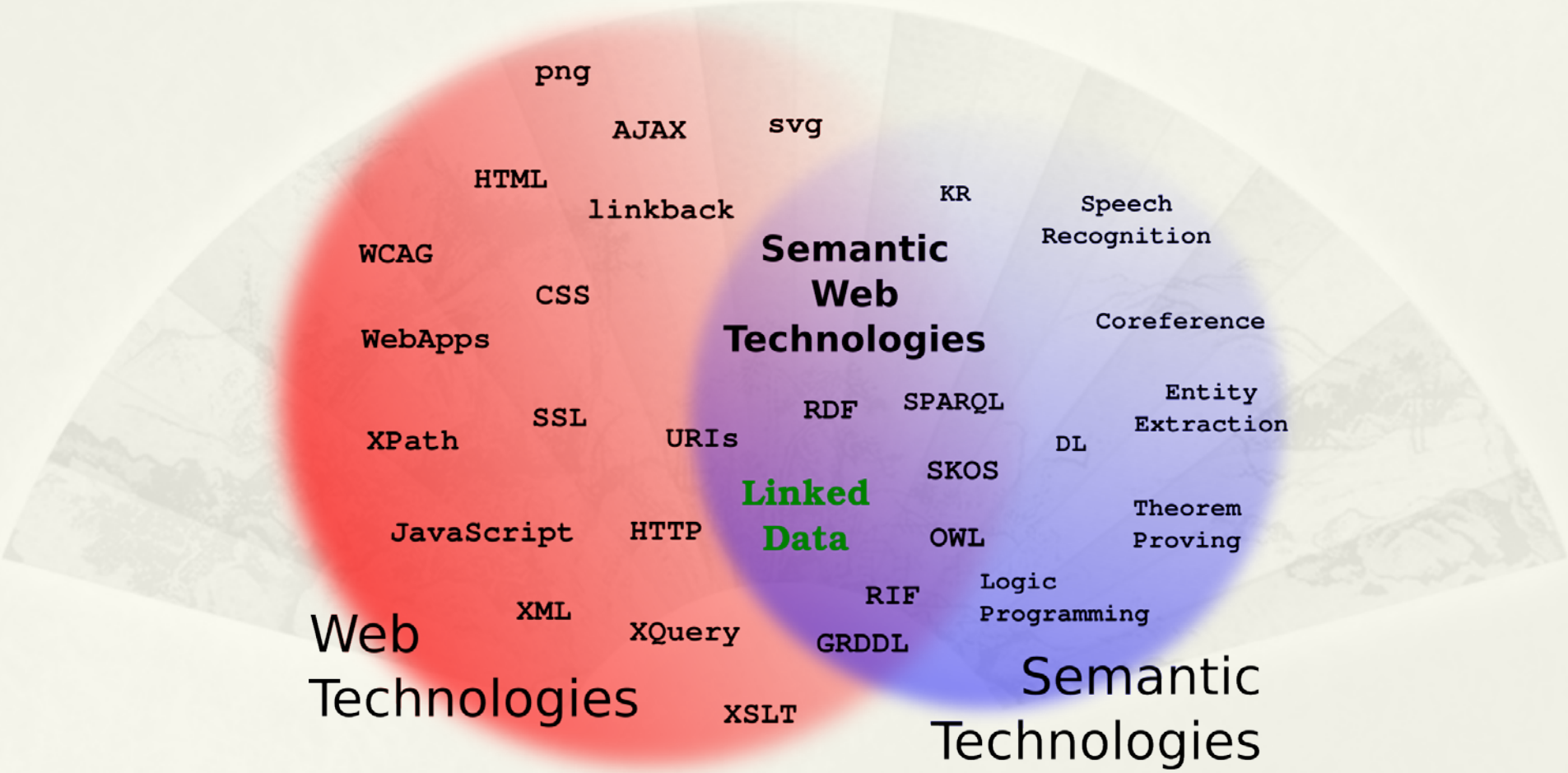
数据载入

享受关联数据发布\内容协商\知识图谱服务

开放资源

- LOD
  - 上海图书馆数字人文系统
  - DBPedia
  - BBC
  - LOC
  - VIAF - Virtual International Authority File
  - GeoNames
  - ……
- Non-LOD
  - CBDB
  - 百科资源
  - 图片
  - ……

相关技术


png
AJAX
svg
HTML
linkback
KR
Speech Recognition
WCAG
Semantic Web Technologies
CSS
Coreference
WebApps
SSL
RDF
SPARQL
Entity Extraction
XPath
URIs
DL
SKOS
Linked Data
Theorem Proving
JavaScript
HTTP
OWL
RIF
Logic Programming
XML
XQuery
GRDDL
Web Technologies
Semantic Technologies
XSLT

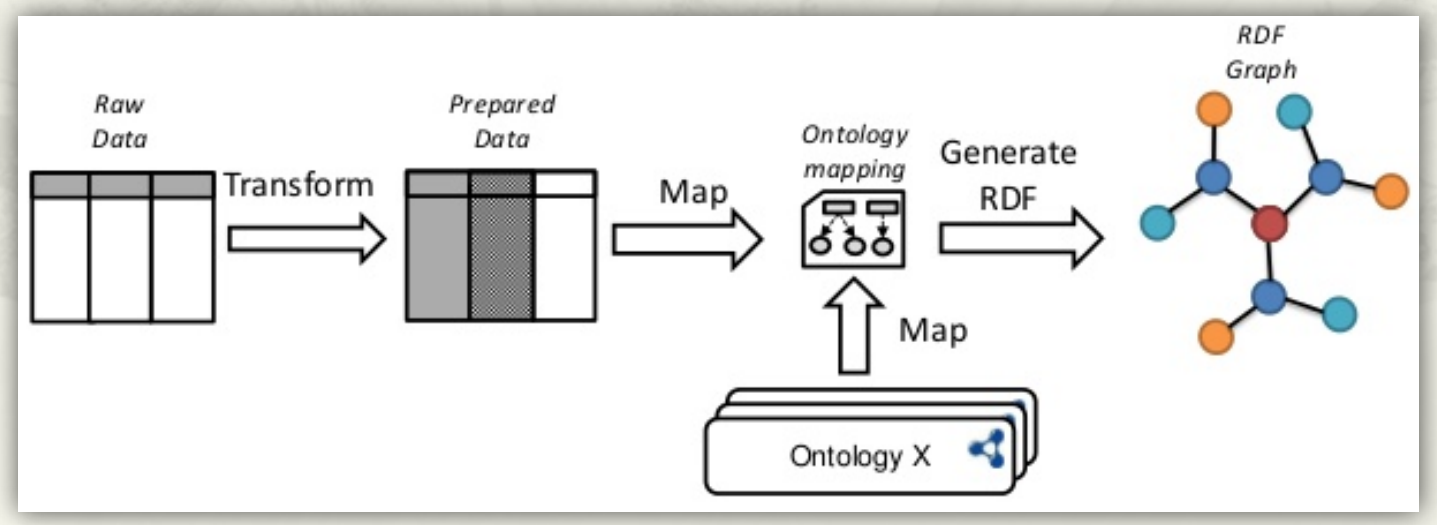
RDF数据生成

① 准备数据

② 转换/清洗数据

③ 准备映射文件（数据字段到本体的映射）

④ 生成RDF数据



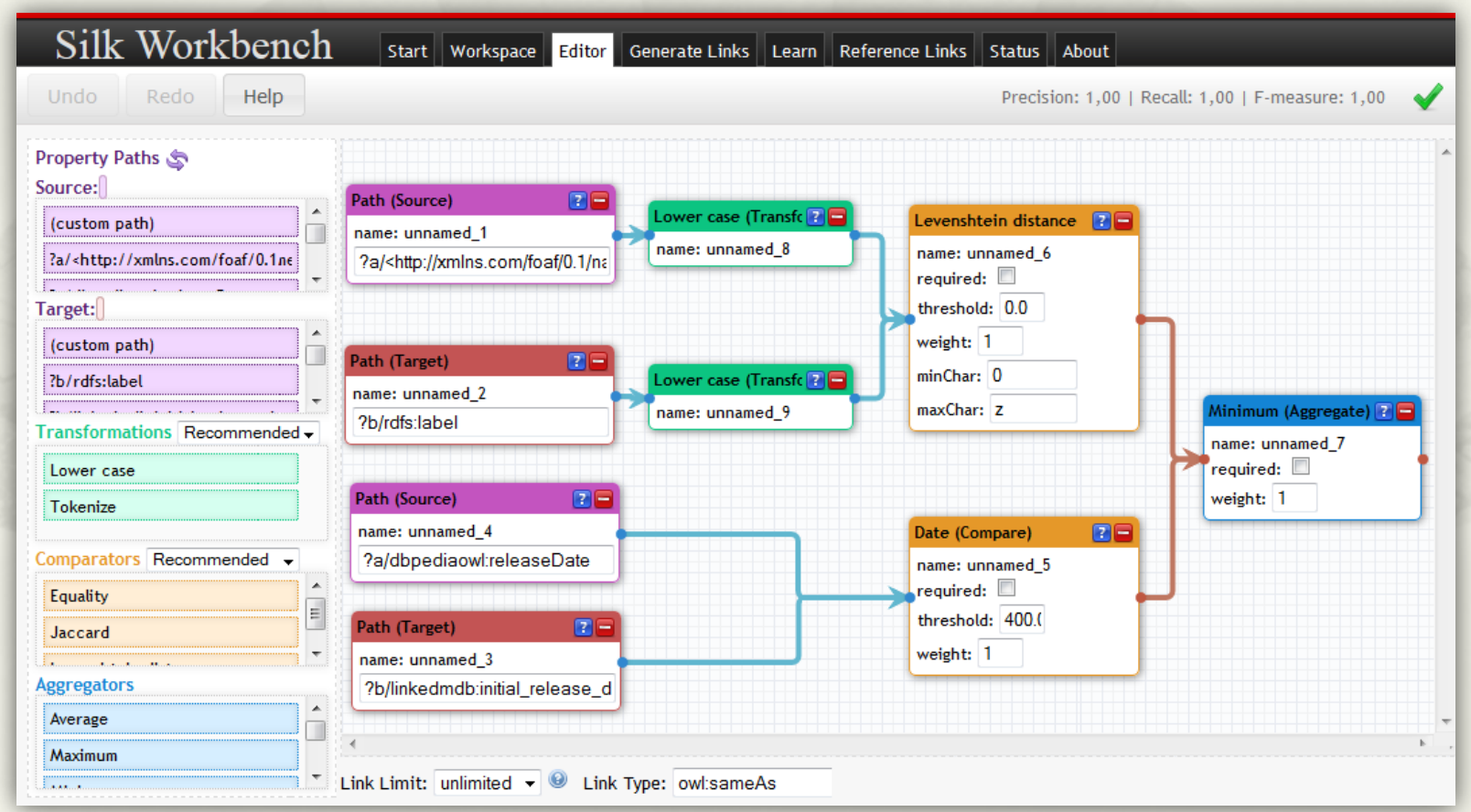
Link to other Data Sets

- **Linking Approaches**
  - Common Key Matching
    - Matching based on common keys
      - e.g. ISBN, DOI, Wikipedia Article ID, Musicbrainz IDs
    - Matching locations based on geographic coordinates
  - Label Matching
    - Comparing labels using string similarity measures
      - e.g. object/page with title/label "The Shining (film)" on DBpedia/Wikipedia is the same as movie object with the "The Shining" on LinkedMDB
    - Comparing labels using semantic similarity measures
      - e.g. "UofT" is the same "University of Toronto", or a drug named "Tylenol" is the same another dury "Acetaminophen" (scientific name of brand name Tylenol)
  - Graph/Ontology Matching
    - Compare labels, schema elements (e.g., types), and related objects (e.g., matching papers if they have the same set of authors)

Link to other Data Sets

* **Linking Frameworks**

  Silk | Limes | LinQuer



关联属性

常用的链接属性（Predicates）

- owl:sameAs

- foaf:homepage
- foaf:topic
- foaf:based_near
- foaf:maker/foaf:made
- foaf:depiction

- foaf:page
- foaf:primaryTopic
- rdfs:seeAlso

SPARQL Queries

HOW TO DO?

1. Issue follow-up queries to different endpoints
2. Querying a central collection of datasets
3. Build store with copies of relevant datasets
4. Use query federation system

Follow-up Queries

Idea: issue follow-up queries over other datasets based on results from previous queries

Substituting placeholders in query templates

✓ Advantage

- Queried data is up-to-date

✗ Drawbacks

- Requires the existence of a SPARQL endpoint for each dataset
- Requires program logic
- Very inefficient

Query a Collection

Idea: Use an existing SPARQL endpoint that provides access to a set of copies of relevant datasets

Example:

SPARQL endpoint over a majority of datasets from the LOD cloud

✓ Advantage

No need for specific program logic

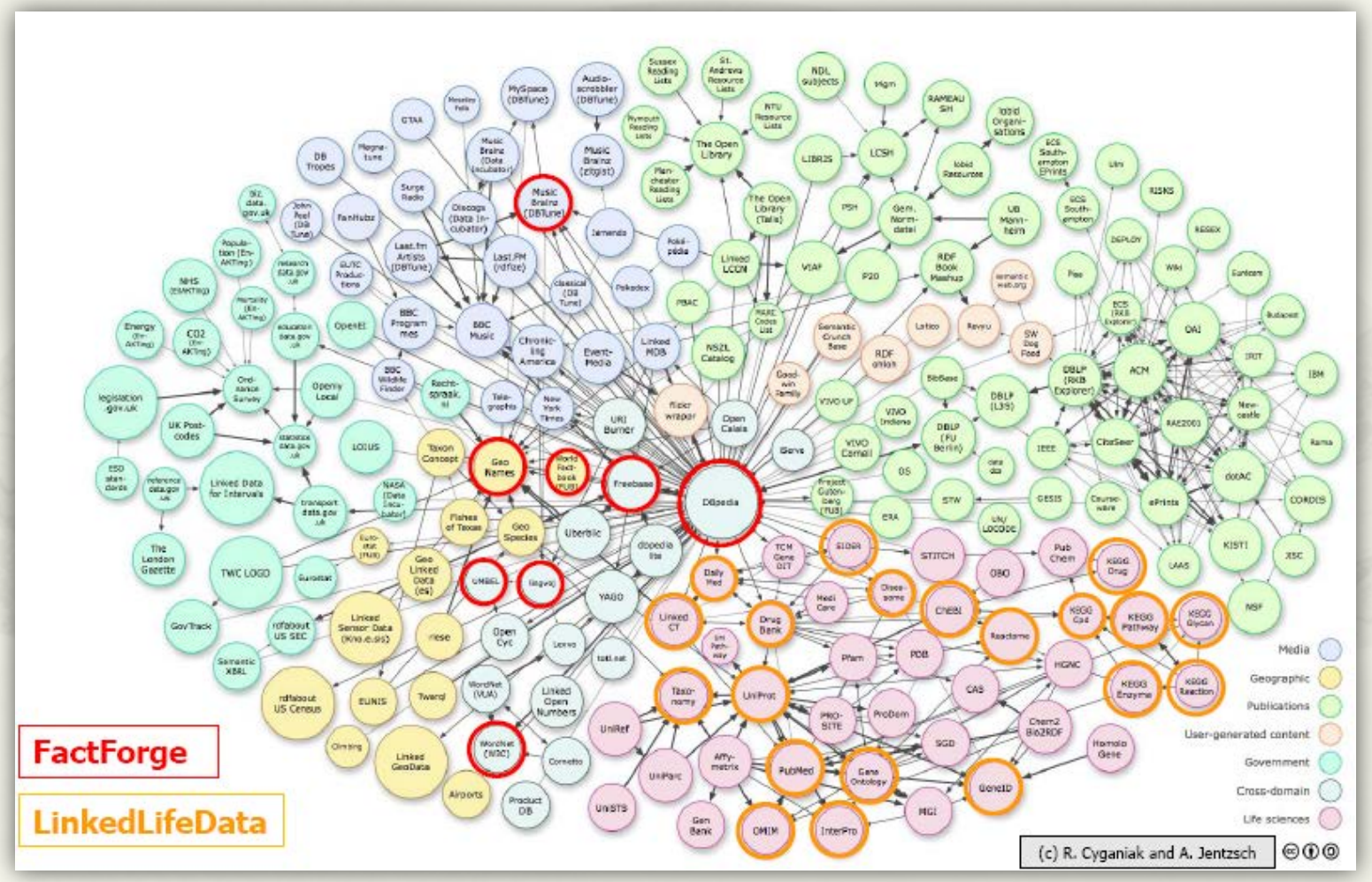
✗ Drawbacks

Queried data might be out of date

Not all relevant datasets in the collection

Query a Collection

http://www.linkedlifedata.com



Build Store

Idea: Build your own store with copies of relevant datasets and query it

✓ Advantages

- No need for specific program logic
- Can include all datasets
- Independent of the existence, availability, and efficiency of SPARQL endpoints

✗ Drawbacks

- Requires effort to set up and to operate the store
- Ideally, data sources provide RDF dumps; if not?
- How to keep the copies in sync with the originals?
- Queried data might be out of date

Federate Query

Idea: Querying a mediator which distributes sub-queries to relevant sources and integrates the results

DARQ (Distributed ARQ) http://darq.sourceforge.net/

Semantic Web Integrator and Query Engine(SemWIQ) http://semwiq.sourceforge.net/

✓ Advantages

- No need for specific program logic
- Queried data is up to date

✗ Drawbacks

- Requires the existence of a SPARQL endpoint for each dataset

SinoPedia入口:嗖啪

扫一扫，直接在手机上打开

嗖啪检索平台

RDF/XML

JSON-LD

NTriples

内容协商

```
- <rdf:RDF>
  - <rdf:Description rdf:about="http://sinopedia.library.sh.cn/entity
    /person/dd411250a2c24fa8a2bdecc1bc819ab0">
      <owl:sameAs rdf:resource="http://dbpedia.org/resource
      /Lu_Xun"/>
      <foaf:depiction rdf:resource="http://commons.wikimedia.org
      /wiki/Special:FilePath/LuXun1930.jpg"/>
      <owl:sameAs rdf:resource="http://data.library.sh.cn/entity
      /person/x5xiy94zid7u351t"/>
    - <rdfs:comment xml:lang="zh">
        原名周树人，字豫才。浙江绍兴人。现代文学家、思想家。1898年就
        读于江南水师学堂。1902年赴日本留学，原学医，后转志于文学。
        1905年参加革命派与改良派的大论战，力主反清革命。1907年加入光
        复会。1909年回国后，曾在杭州、绍兴等地任教。辛亥革命后任职于
        中华民国政府教育部社会教育司，曾参与筹建京师图书馆和历史博物
        馆，同时在北京大学和北京女子师范大学等校兼课。后投身于五四新文
        化运动。1918年1月参加《新青年》杂志编委会。同年5月发表中国现
        代文学史上第一篇白话文小说《狂人日记》，奠定了新文学运动的基
        石。此后，又陆续创作发表了《呐喊》、《坟》、《热风》、《彷徨》
        等小说专集，并开始接触马克思主义。1925年发起并领导语丝社、未
        名社，主编《莽原》杂志。1926年因支持北京爱国学生运动而遭军阀
        政府通缉，同年8月南下，执教于厦门大学。1927年1月转赴广州，任
        中山大学文学系主任兼教务主任，并与中国共产党组织建立联系。四一
        五反革命政变发生后愤而辞职。同年10月迁居上海，同时彻底放弃进
        化论思想，接受马克思主义观点。1928年创办《奔流》杂志，并编辑
```

RDF/XML

```
<http://sinopedia.library.sh.cn/entity/person
/dd411250a2c24fa8a2bdecc1bc819ab0> <http://www.w3.org/2002/07
/owl#sameAs> <http://dbpedia.org/resource/Lu_Xun> .
<http://sinopedia.library.sh.cn/entity/person
/dd411250a2c24fa8a2bdecc1bc819ab0> <http://xmlns.com/foaf/0.1
/depiction> <http://commons.wikimedia.org/wiki/Special:FilePath
/LuXun1930.jpg> .
<http://sinopedia.library.sh.cn/entity/person
/dd411250a2c24fa8a2bdecc1bc819ab0> <http://www.w3.org/2002/07
/owl#sameAs> <http://data.library.sh.cn/entity/person
/x5xiy94zid7u351t> .
<http://sinopedia.library.sh.cn/entity/person
/dd411250a2c24fa8a2bdecc1bc819ab0> <http://www.w3.org/2000/01/rdf-
schema#comment> "原名周树人，字豫才。浙江绍兴人。现代文学家、思想
家。1898年就读于江南水师学堂。1902年赴日本留学，原学医，后转志于文
学。1905年参加革命派与改良派的大论战，力主反清革命。1907年加入光复
会。1909年回国后，曾在杭州、绍兴等地任教。辛亥革命后任职于中华民国政
府教育部社会教育司，曾参与筹建京师图书馆和历史博物馆，同时在北京大学
和北京女子师范大学等校兼课。后投身于五四新文化运动。1918年1月参加
《新青年》杂志编委会。同年5月发表中国现代文学史上第一篇白话文小说
《狂人日记》，奠定了新文学运动的基石。此后，又陆续创作发表了《呐
喊》、《坟》、《热风》、《彷徨》等小说专集，并开始接触马克思主义。
1925年发起并领导语丝社、未名社，主编《莽原》杂志。1926年因支持北京爱
国学生运动而遭军阀政府通缉，同年8月南下，执教于厦门大学。1927年1月转
赴广州，任中山大学文学系主任兼教务主任，并与中国共产党组织建立联系。
四一五反革命政变发生后愤而辞职。同年10月迁居上海，同时彻底放弃进化论
思想，接受马克思主义观点。1928年创办《奔流》杂志，并编辑出版《马克思
主义文艺论丛》。1930年起，先后参加和参与发起中国自由运动大同盟、中国
左翼作家联盟、中国民权保障同盟等，曾任“左联”常务委员，民权保障同盟
上海分会执行委员等，同时继续进行小说、杂文及其他文学创作，并与国民党
```

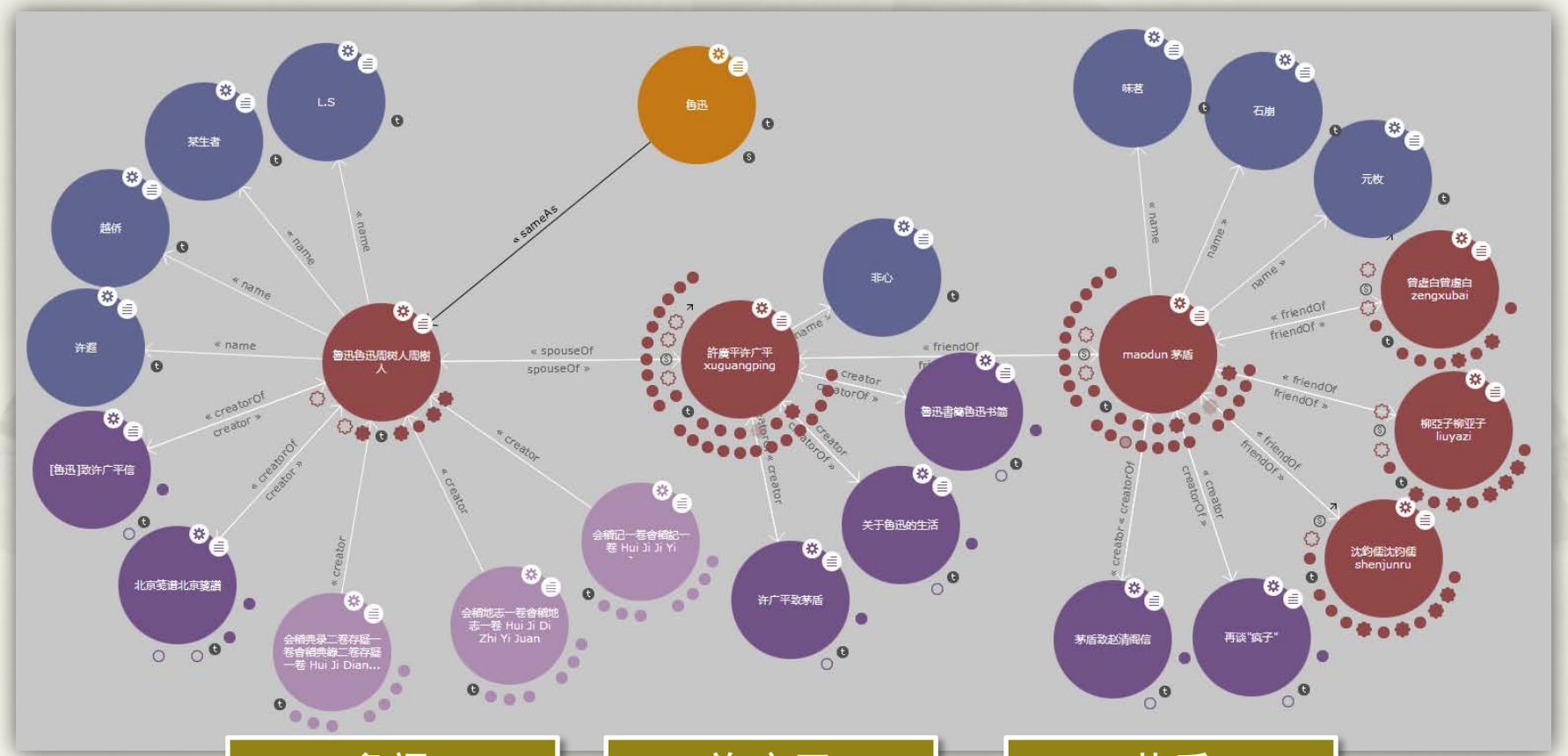
NTriples

链接数据

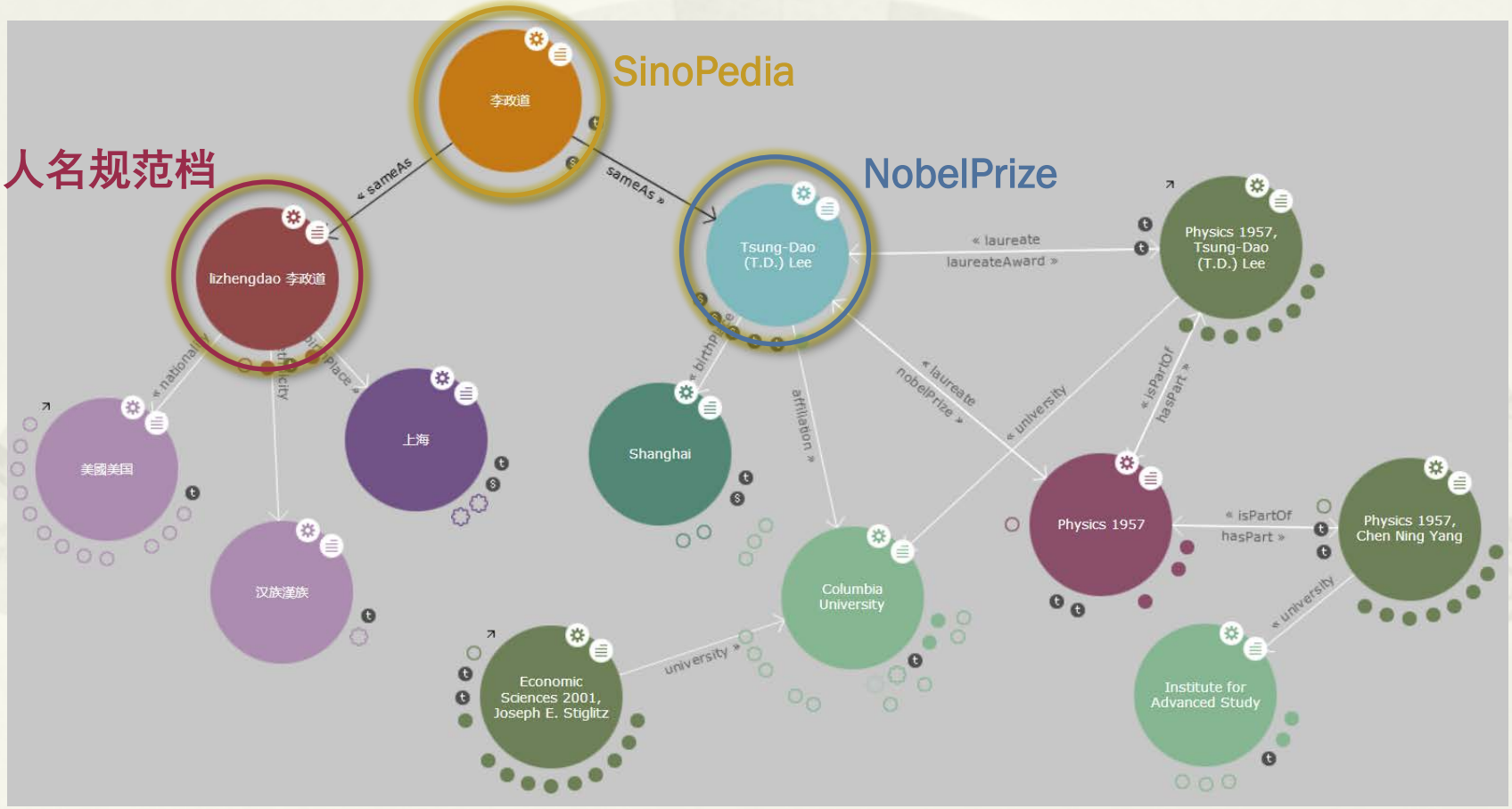
rdfs:**label**	鲁迅
foaf:**gender**	男 @zh
dbpedia-owl:**name**	鲁迅
dbpedia-owl:**birthDate**	1881-09-25
dbpedia-owl:**deathDate**	1936-10-19
rdf:**type**	foaf:Person
owl:**sameAs**	<http://data.library.sh.cn/entity/person/x5xiy94zid7u351t> 上图人名规范档
	dbpedia:Lu_Xun DBpedia
foaf:**depiction**	<http://commons.wikimedia.org/wiki/Special:FilePath/LuXun1930.jpg>

知识图谱：鲁迅

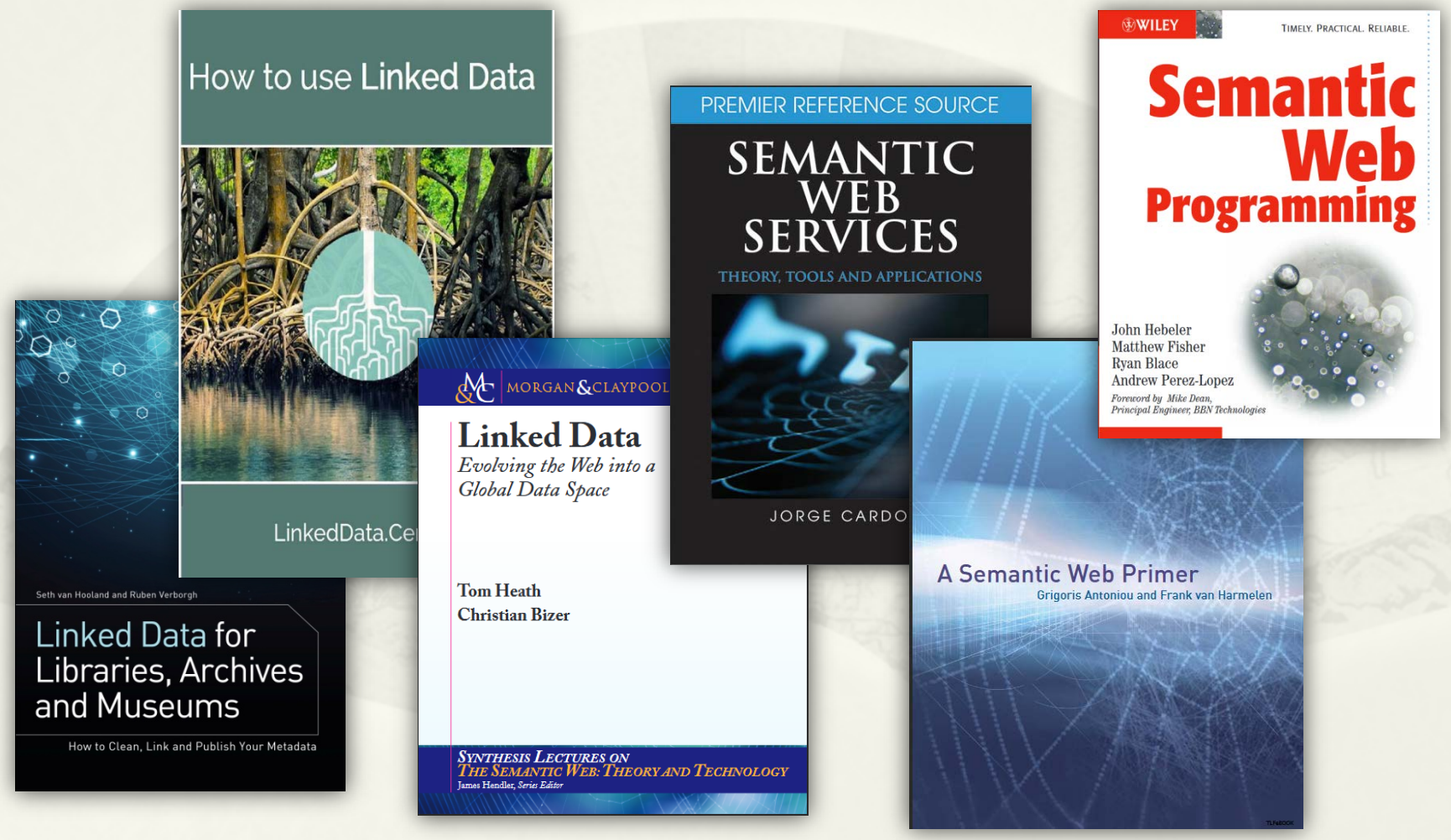
人名规范档　手稿　古籍　SinoPedia



知识图谱：李政道


SinoPedia
人名规范档
NobelPrize
李政道
« sameAs
sameAs »
lizhengdao 李政道
Tsung-Dao (T.D.) Lee
« laureate
laureateAward »
Physics 1957, Tsung-Dao (T.D.) Lee
« nationality
« birthPlace
« birthPlace
« laureate
nobelPrize »
« university
« isPartOf
hasPart »
美國美国
上海
Shanghai
affiliation »
Physics 1957
« isPartOf
hasPart »
Physics 1957, Chen Ning Yang
汉族漢族
Columbia University
« university
Economic Sciences 2001, Joseph E. Stiglitz
university »
Institute for Advanced Study

好书推荐


How to use Linked Data
LinkedData.Ce
Seth van Hooland and Ruben Verborgh
Linked Data for Libraries, Archives and Museums
How to Clean, Link and Publish Your Metadata
PREMIER REFERENCE SOURCE
SEMANTIC WEB SERVICES
THEORY, TOOLS AND APPLICATIONS
JORGE CARDO
MORGAN & CLAYPOOL
Linked Data
Evolving the Web into a Global Data Space
Tom Heath
Christian Bizer
SYNTHESIS LECTURES ON THE SEMANTIC WEB: THEORY AND TECHNOLOGY
James Hendler, Series Editor
WILEY
TIMELY. PRACTICAL. RELIABLE.
Semantic Web Programming
John Hebeler
Matthew Fisher
Ryan Blace
Andrew Perez-Lopez
Foreword by Mike Dean, Principal Engineer, BBN Technologies
A Semantic Web Primer
Grigoris Antoniou and Frank van Harmelen

嗖啪 http://sinopedia.library.sh.cn/sooopa

SinoPedia

Share it now

共建 共享
共链 共赢