CHINESE MEDICINE

RESEARCH Open Access

Publishing Chinese medicine knowledge as Linked Data on the Web

Jun Zhao

Abstract

Background: Chinese medicine (CM) draws growing attention from Western healthcare practitioners and patients. However, the integration of CM knowledge and Western medicine (WM) has been hindered by a barrier of languages and cultures as well as a lack of scientific evidence for CM's efficacy and safety. In addition, most of CM knowledge published with relational database technology makes the integration of databases even more challenging.

Methods: Linked Data approach was used in publishing CM knowledge. This approach was applied to publishing a CM linked dataset, namely RDF-TCM http://www.open-biomed.org.uk/rdf-tcm/ based on TCMGeneDIT, which provided association information about CM in English.

Results: The Linked Data approach made CM knowledge accessible through standards-compliant interfaces to facilitate the bridging of CM and WM. The open and programmatically-accessible RDF-TCM facilitated the creation of new data mash-up and novel federated query applications.

Conclusion: Publishing CM knowledge in Linked Data provides a point of departure for integration of CM databases.

Background

Chinese medicine (CM) is yet to become an integral part of the standard healthcare system in Western countries due to a lack of scientific evidence for its efficacy and safety as well as a language and cultural barrier. This article presents a Linked Data approach to publishing CM knowledge in hope of bridging the gap between CM and Western medicine (WM).

The World Wide Web is a scalable platform for disseminating information through documents, having transformed how knowledge is learned and shared. Similarly, the Web may also be used as the platform for disseminating data. Linked Data [1] uses the Web as the information space to publish structured data rather than documents on the Web. In Linked Data, Uniform Resource Identifiers (URIs) are used to identify resources [2] and Resource Description Framework (RDF) is used to describe resources [3]. URIs are to data as what Uniform Resource Locators (URLs) are to web pages, providing identifications to resources; and RDF is to data as what HTML is to documents, providing descriptions about a resource in a machine-processable representation format.

Linked Data promises a new and more efficient paradigm for sharing and connecting distributed data, permitting decentralization and interoperability. Since Linked Data is built upon the Web Architecture [4], it inherits its decentralization and connectivity. The Web enforces no central control points and those distributed resources on the Web are intrinsically connected to each other by two fundamental elements, namely the Hyper-Text Transfer Protocol (HTTP) [5] which permits the transportation of information resources on the Web and the URIs which provide a globally-scoped system for identifying web resources (documents or data). Furthermore, linked datasets are meant to be interoperable based upon the Semantic Web standards established by the World Wide Web Consortium (W3C). These standards comprise RDF for publishing data in a structured format with explicit semantics and the SPARQL query language and protocol [6,7] for querying and accessing RDF data through an open and HTTP-based protocol.

Correspondence: jun.zhao@zoo.ox.ac.uk
Image Bioinformatics Research Group, Department of Zoology, Oxford University, South Parks Road, Oxford, OX1 3PS, UK

A growing number of linked datasets as well as supporting tools and technologies are rapidly emerging, providing a unique opportunity for Linked Data to be applied in biomedical research and healthcare. The Linking Open Data (LOD) project [8] was founded in January 2007 and within one year the RDF published by the LOD community grew to over two billion [9]. The fast growth of Linked Data cloud cannot be achieved without the variety of open-source tools for publishing, searching, indexing and browsing linked datasets. Notably, tools such as D2R Server [10] and Triplify [11] are making relational databases accessible as RDF without transforming the source databases. Linked datasets become consumable for both humans and computers with the emergence of various Linked Data browsers such as Tabulator [12], Sig.ma [13], Linked Data query engines (e.g. SQUIN [14]) and Google-like Linked Data search engines (e.g. Sindice [15] and SWoogle [16]).

One of the earliest adopters of Linked Data for life sciences is the Bio2RDF project [17], in which various biological and bioinformatics knowledge bases have been published in the form of linked datasets using Semantic Web technologies. The knowledge bases published by Bio2RDF continue to grow, ranging from human genomics databases such as NCBI's Entrez Gene, proteiomics databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [18] and Protein Data Bank (PDB) [19] to pharmacogenomics databases such as PharmGKB [20], and cheminformatics databases such as PubChem [21]. Another active effort, similar to Bio2RDF, is the Linking Open Drug Data (LODD) project [22], founded under the umbrella of W3C Health Care and Life Science Interest Group. The goal of the LODD project is to gather requirements from the life science research community and to publish required databases in the Linked Data format. LODD has successfully published a selection of databases as Linked Data and generated their links with other Linked Data cloud [23], including the Bio2RDF datasets and the nucleus of Linked Data Cloud, namely DBpedia [24]. A missing link in the life science-oriented Linked Data cloud is a dataset about alternative medicines. Our RDF-TCM linked dataset plays a key role in connecting medical knowledge originating from different cultures and scientific disciplines. The aims of the presented article are as follows:

- Describing a CM linked dataset RDF-TCM, which is the first effort in publishing CM knowledge in a more accessible Linked Data format and is created according to our Linked Data Publication Methodology;
- Demonstrating that publishing linked CM data provides a point of departure for data integration through two efficient ways of consuming linked datasets.

Methods

TCMGeneDIT database

The RDF-TCM dataset transformed the relational TCMGeneDIT [25] as RDF. TCMGeneDIT not only provides information in English but also collects the associations among herbs, genes, diseases, CM effects and CM ingredients from public databases and literature. Existing knowledge is reused and some association information is collected through text mining techniques, such as:

- Herb names, such as *Ginkgo biloba*, were collected from the HULU TCM professional web site [26] and TCM-ID [27], a database on CM herbs and herbal ingredients;
- Ingredient data were collected from the above two resources as well as the Chinese medicine resource web [28];
- Human genes and their information were retrieved from NCBI Entrez [29];
- Disease names were extracted from the heading and entry term fields in the disease (C) section of the medical subject headings vocabulary (MeSH) [30];
- The relationship between genes and diseases were collected from PharmGKB [20];
- Many other association information between herbs and genes, diseases and effects were mined and extracted from a corpus of MEDLINE abstracts collected through PubMed.

Create RDF-TCM

The TCMGeneDIT database is available as a database dump under the Creative Commons Attribution License [31]. To publish TCMGeneDIT as Linked Data, we followed our Linked Data Publication Methodology proposed previously [32], including the following steps:

1. Choose a transformation strategy, either through RDF caching or virtualization;
2. Design an URI scheme according to the Linked Data principles and the Cool URIs style [33], providing simple and stable URIs;
3. Construct schemas or ontologies based on the source data schemas, imposing as little interpretations as possible and reusing existing ontologies where possible;
4. Construct transformation scripts and mapping files, starting with transforming a small portion of the records and a test framework, which is not only

useful for validating the sanity of the RDF dataset but also for revalidation when the transformation process is repeated;

5. Create mappings to other data sources where immediate values are foreseen, either using customized scripts or existing software tools such as Silk [34];

6. Finally, and preferably, provide metadata descriptions about the dataset, including its provenance information, and make all the scripts, configuration files, and ontologies accessible.

A skeleton of the methodology was proposed [32] and the following sections will provide details. Steps 2-5 should be applied iteratively and some design decisions must be made in accordance with fundamental principles.

Choose a transformation strategy

Linked datasets can be published either by creating RDF caching or through a virtualized access to the source data. RDF caching means that developers convert a snapshot of the source database into RDF and then load these cached data into an RDF store and publish it as Linked Data. The virtualization approach rewrites an HTTP-dereference request to a data URI into a query expressed in a language native to the source database (e.g. SQL) for evaluation against the data in their native form without transformation into RDF. The virtualization approach is more desirable if the source data have a high churn rate, but the performance of the current tools supporting this virtualization (such as Triplify [11]) is difficult to cope with large relational databases and complex rewriting rules. If the update rate of the source data is sufficiently low, the caching approach is more feasible. Because TCMGeneDIT is no longer updated, we chose the RDF caching approach to build RDF-TCM.

Design the URIs

URIs are required in Linked Data in order to identify entities (instances), types of entities (classes) and types of their relationships (properties). The 'Linked Data Principles' outlined by Berners-Lee [35] clarify the role of URIs in Linked Data and the set of best practices for publishing them:

> *"1. Use URIs as names for things; 2. Use HTTP URIs so that people can look up these names; 3. When someone looks up a URI, provide useful information using the standards (e.g. RDF, SPARQL); 4. Include links to other URIs, so that they can discover more things."*

In addition we recommend that new URIs should only be coined if no existing URIs can be found and that they should be persistent. Reusing existing URIs improves the connectivity of a dataset with others and help establish shared names within the community. Consortia such as SharedNames [36] and Concept Web Alliance [37] are the active ongoing efforts in creating unique, shared names for biological entities. A data publisher should have control over the namespace under which new URIs are created, not only allowing useful information about these resources to be provided but also improving the stability of these URIs. Creating links to URIs published by others is highly recommended for bridging the gap between a local namespace and the Linked Data cloud.

The URIs used for RDF-TCM followed the pattern of:

> http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/{type}/{id}

where {type} corresponds to the type of an entity (such as Gene) and {id} is an identifier derived from the source data, e.g. the gene name or the herb name, or from a sequential number assigned by the transformation program. We used PURL [38] URIs to control the persistency of these URIs and we used the namespace of the TCMGeneDIT website as part of the URI to preserve some information about the owner and origin of the dataset. For example, the URI

> http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/medicine/Ginkgo_biloba

identifies the herb *Ginkgo biloba*.

And the URI

> http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/statistics/9199

denotes a statistics entity that describes confidence in the association relationship between some entities.

Design ontologies

Ontologies can be used as a controlled vocabulary to define the type of entities in a dataset and the type of relationships between them and to achieve a consistent interpretation about different datasets. A rich body of biological ontologies has been created and accumulated over the years [39]. When designing ontologies for describing linked datasets, we should reuse existing ontologies as much as possible. When a new ontology must be created, a conservative and incremental approach is recommended. Many of the linked datasets are published by a third party, rather than by the data provider. Documentation about these datasets is not always available. Imposing personal interpretations about the semantics of the data and its schema could introduce errors and should be avoided.

As the data structure of TCMGeneDIT is very simple and there was no known TCM ontology by the time of creating the dataset, we created a simple CM ontology using OWL http://purl.org/net/tcm-onto/. The ontology contains seven classes, namely `Gene`, `Medicine`, `Disease`, `Ingredient`, `Effect`, `Association` and `Statistics`. Each entity of type `Statistics` describes statistics confidence in the associations between entities. Each entity of type `Association` represents an association between a `Medicine`, a `Gene` and a `Disease`. There are six object properties in total: five of them for relating a `Medicine` to a `Gene`, a `Disease`, its `Ingredient`, or its `Effect` and the last one, `tcm:source`, for pointing to the entities whose association relationship is described by a `Statistics` entity. There are five data properties whose domain is `Statistics` and whose value represents the statistics confidence in the association. For example, the value of `tcm:medicine_effect_association_tvalue` represents our confidence in the association between a `Medicine` and its `Effect`. A diagram capturing the structure of the ontology is shown in Figure 1. Note that the data properties associated with the `Statistics` class are not shown in the figure.

A `Statistics` entity was used to describe the statistical value of an association. Some associations relating to more than two entities such as the association relationship of medicine-gene-diseases cannot be expressed as RDF triples. To capture this n-ary relationship, we created `Statistics` entities to link together every entity involved in an association (see the example below) and to express the statistical value of the association using the data properties, e.g., `tcm:medicine_effect_association_tvalue`. The different types of data properties were created for different types of associations.

```
http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/statistics/19087 a tcm:Statistics;

  tcm:source
    http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/medicine/Acanthopanax_gracilistylus;
  tcm:source http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/disease/Retinoblastoma;
  tcm:source http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/id/gene/CDK2;
  tcm:medicine_gene_disease_association_tvalue "1.414"^^xsd:float.
```

Data transformation

Data transformation should be incremental and test-driven. When transforming a new dataset into RDF or writing the configuration files for virtualization, developers should start with a small subset and avoid transforming the complete dataset. Loading a large number of RDF triples into an RDF store or retrieving very complex RDF descriptions for data entities by query rewriting can be a very time-consuming task and block the execution of following-on tests. A test framework should be designed forefront to spot any problems with the testing data and to ensure the sanity of the datasets, such as no blank nodes, no URIs containing invalid characters (e.g. space), no wrong property cardinalities, or no missing property values. These principles were applied when the relational TCMGeneDIT database was transformed into RDF.

Data linking

Links between datasets can be expressed with RDF. These links either reflect a type of relationship

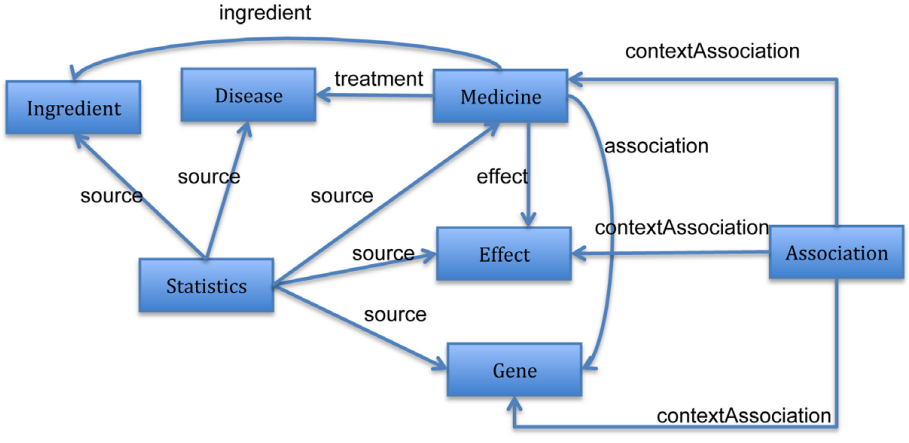


Figure 1 The diagram of the RDF-TCM ontology. The diagram illustrates the main classes (the boxes) and object properties (the directed arrows) in the RDF-TCM ontology http://purl.org/net/tcm-onto/. The data properties of the ontology are not shown.

between entities or state a reconciliation between URIs published by various authorities. An example of the relationship type of links is to associate drugs from dataset *D*1 with genes from dataset *D*2 through a property such as *ex:targets*. Properties such as *owl:sameAs* or *rdfs:seeAlso* can be used for stating identity reconciliation. These RDF links allow users and Linked Data applications to start from one dataset and then follow on these RDF data links to move through a potentially endless web of data.

These data links can be created either during or after the creation of a linked dataset. Commonly, relating to another dataset (e.g., *ex:targets*) may be achieved as part of the transformation script, while mapping two URIs from different datasets may take place after a dataset is published and be executed either by their publishers or third parties.

The links may be created manually or automatically with open-source tools such as Silk [34]. However, identity reconciliation between biological entities is known to be difficult; string mapping is not always sufficient or reliable [40]. Developers should look for existing authoritative name mappings curated by data providers. Identifying the reference databases used by the source databases could help improve the precision of the mapping. For example, by understanding that the gene names used by TCMGeneDIT are from NCBI Entrez Gene for human, we can reduce the ambiguity of the mapping to the Entrez Gene dataset previously published by Neurocommons or Bio2RDF.

Extra attention should be given to any many-to-many mappings between URIs in the results. A manual cleaning of these mappings is highly recommended, requiring either the participation of domain experts or some contextual knowledge that are difficult to be expressed in computer programs.

The gene entities in the RDF-TCM dataset were linked with those from the NCBI Entrez Gene linked dataset [41] published by Neurocommons and those from the STITCH linked dataset [42] published by the Freie Universität Berlin. Gene mapping was constructed with customized Python scripts based on the label of the genes. The mapping to Entrez Gene showed that 849 out of the total 945 RDF-TCM genes had a one-to-one mapping to an Entrez gene and that 95 of them had a many-to-many mapping to an Entrez gene and one of them was not mapped. The mapping to STITCH genes showed that 539 out of 943 mapped genes had a one-to-one mapping to a STITCH gene; and that 404 of them had a many-to-many mapping and two of them were not mapped. These many-to-many mappings were manually corrected so that only one-to-one mappings were in the results. We selected some sample data to manually confirm the correctness of the automatically generated one-to-one mappings. However, these automatic gene mappings were not thoroughly evaluated and this is an limitation of the work.

To link RDF-TCM with various other linked dataset from LODD, we used Silk, as part of the LODD project [23]. The mapping results by Silk have not been formally evaluated, but the correctness and completeness of Silk's approach were evaluated with other test datasets [34].

Data documentation

To improve the visibility of a dataset to Linked Data search engines such as Sindice, we recommend data publishers to describe their datasets using vocabularies such as the Vocabulary of Interlinked Datasets (voiD) [43] or the Provenance Vocabulary [44]. voiD is an RDF vocabulary for describing linked datasets on the Web in order to facilitate the discovery of these datasets and query federation applications. The Provenance Vocabulary is the first vocabulary to describe both the data creation and data access process related to a dataset on the Web.

A voiD file was published for RDF-TCM http://www.open-biomed.org.uk/void/rdf-tcm.ttl and the provenance of each RDF-TCM entity was described with the Provenance Vocabulary, published with Pubby [45], a Linked Data publication tool extended with a provenance component. We published all our Python scripts for transforming the database dump into RDF and for linking RDF-TCM to other datasets. All the scripts can be found at http://code.google.com/p/junsbriefcase/source/browse/#svn/trunk/biordf2009_query_federation_case/tcm-data.

Results

RDF-TCM dataset

The RDF-TCM dataset contained 111,021 RDF triples, providing association information for 848 herbs, 1064 ingredients, 241 putative effects, 553 diseases and 945 genes. This dataset was linked with a variety of life science linked dataset including:

- Entrez Gene dataset, part of the HCLS knowledge base, derived from the NCBI Entrez Gene database
- DrugBank http://www4.wiwiss.fu-berlin.de/drugbank/: derived from DrugBank [46] published by the University of Alberta, containing detailed information about almost 5,000 FDA-approved small molecule and biotech drugs
- DailyMed http://www4.wiwiss.fu-berlin.de/dailymed/: derived from Dailymed [47] published by National Library of Medicine (NLM), containing high quality packaging information on 4,300 marketed drugs

Table 1 A summary of different types of links between RDF-TCM and other datasets

Dataset	Type of linked entities	Properties used for interlinking	Number of links
Entrez gene	Genes	Symbols of the genes	944
Diseasesome	Diseases	Labels of the disease names	63
	Genes	Symbols of the genes	312
SIDER	Diseases	Labels of the disease names	171
Drugbank	Genes	Symbols of the genes	384
Dailymed	Ingredients	Labels of the ingredient names	21
	Genes	Symbols of the genes	649
DBpedia	Diseases	Labels of the disease names	255
	Herbs	Labels of the herb names	438
STITCH	Genes (encoding proteins)	Names of the genes	937
PharmGKB	Genes	Names of the genes	202

• SIDER http://www4.wiwiss.fu-berlin.de/sider/: derived from SIDER database [48] published by EMBL Germany, containing side effect information on 930 marketed drugs
• Diseasome http://www4.wiwiss.fu-berlin.de/diseasome/: derived from the Diseasome dataset [49] which publishes a network of disorders and disorder genes, obtained from Online Mendelian Inheritance in Man (OMIM)
• STITCH http://www4.wiwiss.fu-berlin.de/stitch/: derived from STITCH [50] published by EMBL Germany, containing information about known or predicted interactions between proteins and chemicals
• PharmGKB http://bio2rdf.org/ published by Bio2RDF: derived from PharmGKB [51] published by Stanford University, sharing knowledge about the impact of human genetic variations on drug response and publishing data, among many others,

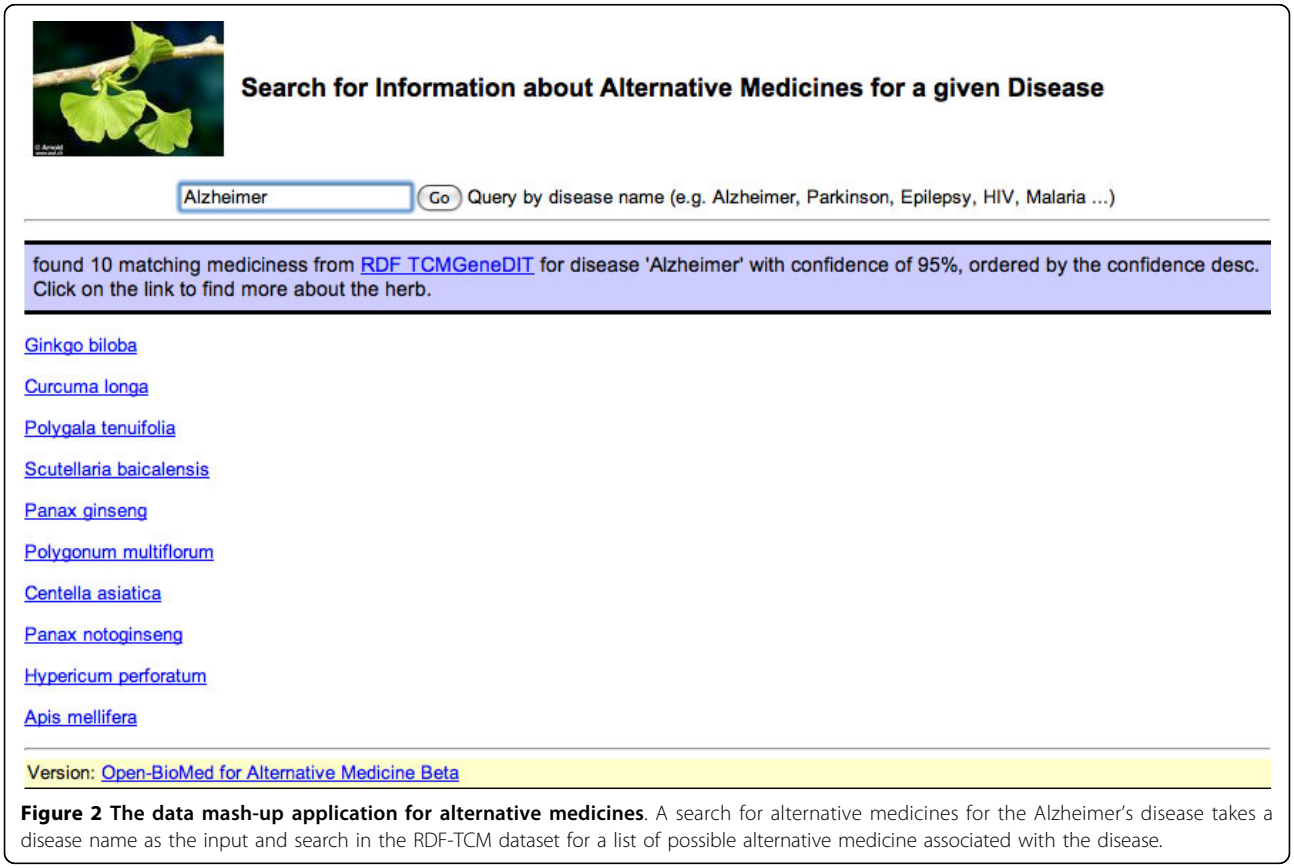
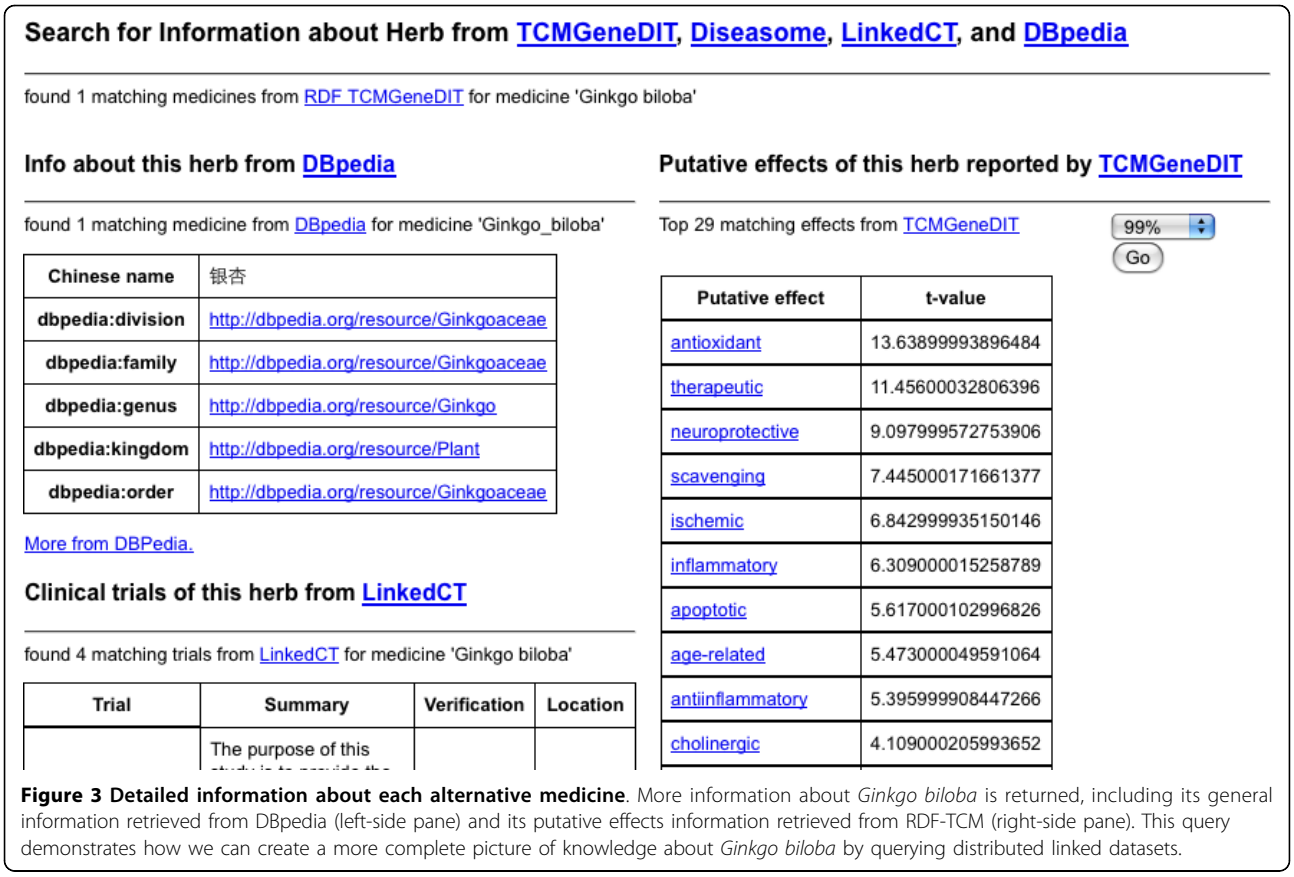


Figure 2 The data mash-up application for alternative medicines. A search for alternative medicines for the Alzheimer's disease takes a disease name as the input and search in the RDF-TCM dataset for a list of possible alternative medicine associated with the disease.



Search for Information about Herb from TCMGeneDIT, Diseasome, LinkedCT, and DBpedia

found 1 matching medicines from RDF TCMGeneDIT for medicine 'Ginkgo biloba'

Info about this herb from DBpedia

found 1 matching medicine from DBpedia for medicine 'Ginkgo_biloba'

Chinese name	银杏
dbpedia:division	http://dbpedia.org/resource/Ginkgoaceae
dbpedia:family	http://dbpedia.org/resource/Ginkgoaceae
dbpedia:genus	http://dbpedia.org/resource/Ginkgo
dbpedia:kingdom	http://dbpedia.org/resource/Plant
dbpedia:order	http://dbpedia.org/resource/Ginkgoaceae

More from DBPedia.

Clinical trials of this herb from LinkedCT

found 4 matching trials from LinkedCT for medicine 'Ginkgo biloba'

Trial	Summary	Verification	Location
	The purpose of this		

Putative effects of this herb reported by TCMGeneDIT

Top 29 matching effects from TCMGeneDIT 99% Go

Putative effect	t-value
antioxidant	13.63899993896484
therapeutic	11.45600032806396
neuroprotective	9.097999572753906
scavenging	7.445000171661377
ischemic	6.842999935150146
inflammatory	6.309000015258789
apoptotic	5.617000102996826
age-related	5.473000049591064
antiinflammatory	5.395999908447266
cholinergic	4.109000205993652

Figure 3 Detailed information about each alternative medicine. More information about *Ginkgo biloba* is returned, including its general information retrieved from DBpedia (left-side pane) and its putative effects information retrieved from RDF-TCM (right-side pane). This query demonstrates how we can create a more complete picture of knowledge about *Ginkgo biloba* by querying distributed linked datasets.

about the associations between drugs, genes and diseases curated by domain experts

Table 1 summarizes the type of entities that link RDF-TCM with each of the above dataset and the number of each type of links. All these link datasets can be downloaded as RDF dumps http://purl.org/net/tcmdata/ or accessed through the public SPARQL endpoint http://www.open-biomed.org.uk/sparql/. In the following section, we will demonstrate how this RDF dataset and these RDF links data are used to assist the exploitation of CM and WM.

Search for potential alternative medicines by mash-ups

Here we present an application [52] of the RDF-TCM dataset as an example. As shown in Figure 2, the data mash-up application allows users to first search for alternative medicines for a diseases using the disease and herb association information from RDF-TCM. The result was ranked by the statistical value from the TCMGeneDIT database that states the confidence in the association between diseases and herbs, i.e. Ginkgo biloba has the highest score for its association with the Alzheimer's Disease. Users may then retrieve detailed information about each alternative medicine (Figure 3, 4 and 5). The scientific classification information was retrieved from DBPedia and putative effects of herbs were retrieved from RDF-TCM (Figure 3). Related clinical trial information were retrieved from the LinkedCT dataset (Figure 4) hosted by the EU LarKC project [53] with string matching SPARQL queries. Figure 5 shows how this application may also help confirm the association relationship between a herb, its possible disease targets and the genes affected by these diseases by combining the WM knowledge from Diseasome and RDF-TCM. The application is an Ajax application implemented with Javascript. Each widget in the application executed a SPARQL query to one or multiple SPARQL endpoints and presented the query result in the web browser in a user-friendly way. The application requires that a data source must be accessible through a SPARQL endpoint. This data mash-up application bridged the knowledge connection between CM and WM. Instead of making users browse various possible data sources to gather information about herbs, the mash-up provides a central point for searching for knowledge about CM gathered from various sources published by these two scientific communities.

Clinical trials of this herb from LinkedCT

found 4 matching trials from LinkedCT for medicine 'Ginkgo biloba'

Trial	Summary	Verification	Location
Ginkgo Biloba: Antidepressant-Induced Sexual Dysfunction	The purpose of this study is to provide the first empirical examination of the effects of Ginkgo biloba (GBE), sex therapy, and a combination of the two on subjective and physiological measures of sexual function in women who are experiencing sexual disorders secondary to antidepressants.	July 2006	Austin, United States
Ginkgo Biloba Extract and the Insulin Resistance Syndrome	The purpose of this study is to examine whether the ingestion of the herbal dietary supplement Ginkgo biloba extract has any effect on the efficacy of three classes of diabetic medications - (Glucotrol, Glucophage and Actose). Additionally, the study will examine the effect of Ginkgo biloba extract on pancreatic insulin production in non-diabetic subjects between the ages of 20 and 75 years old.	July 2006	San Antonio, United States
Ginkgo Biloba Prevention Trial in Older Individuals	This study will determine the effect of 240mg/day Ginkgo biloba in decreasing the incidence of dementia and specifically Alzheimer's disease (AD), slowing cognitive decline and functional disability, reducing incidence of cardiovascular disease, and decreasing total mortality.	September 2007	Winston-Salem, United States
Ginkgo Biloba to Improve Short-Term Memory Losses Associated With Electroconvulsive Therapy (ECT)	Electroconvulsive therapy (ECT) is an effective treatment for severe or medication-resistant depression and other psychiatric disorders. A common side effect of ECT is problems with short-term memory during treatment. This study will test whether taking ginkgo biloba (GB) prior to and during the course of ECT will lessen the effects of ECT on short-term memory.	August 2008	Newark, United States

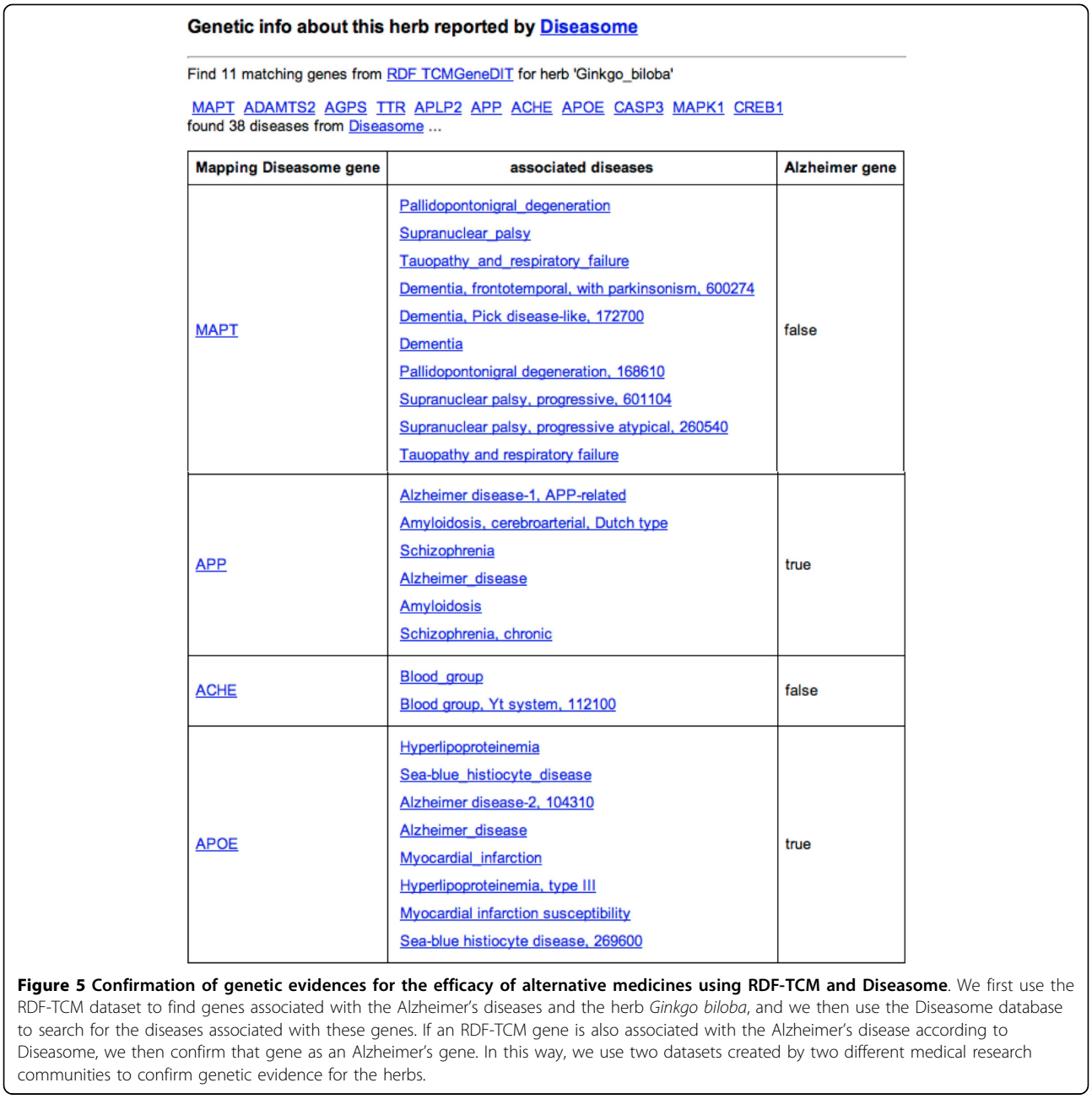
Figure 4 Clinical trials related to Ginkgo biloba. Clinical trials related to *Ginkgo biloba* are found from the LinkedCT dataset. These results are also linked to LinkedCT where more information about these trials can be found.

Search for potential alternative medicines by the Linked Data approach

RDF-TCM together with LODD forms a web of medical data, accessible through Linked Data query engines as a single dataspace. SQUIN [14] is one such Linked Data query engine that traverses the whole Web of Data to retrieve all relevant data sources for a query by taking the URIs in the query or in the intermediate results and following links of these URIs to other data sources. In this second application [54], to search for an alternative medicine to a Western medicine (Figure 6) we used SQUIN to take the example SPARQL query in Listing 1 to traverse 7 distributed Linked Datasets including Drugbank, Diseasome, SIDER, LinkedCT, Dailymed and RDF-TCM.

Listing 1: The SPARQL query for finding alternative medicines to Simvastatin.

```
PREFIX tcm: http://purl.org/net/tcm/tcm.lifescience.ntu.edu.tw/
PREFIX drugbank: http://www4.wiwiss.fu-berlin.de/drugbank/resource/drugs/
PREFIX rdfs: http://www.w3.org/2000/01/rdf-schema#
PREFIX owl: http://www.w3.org/2002/07/owl#
PREFIX rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
SELECT DISTINCT ? diseaseLabel ? altMedicineLabel
```



Genetic info about this herb reported by Diseasome

Find 11 matching genes from RDF TCMGeneDIT for herb 'Ginkgo_biloba'

MAPT ADAMTS2 AGPS TTR APLP2 APP ACHE APOE CASP3 MAPK1 CREB1
found 38 diseases from Diseasome ...

Mapping Diseasome gene	associated diseases	Alzheimer gene
MAPT	Pallidopontonigral_degeneration Supranuclear_palsy Tauopathy_and_respiratory_failure Dementia, frontotemporal, with parkinsonism, 600274 Dementia, Pick disease-like, 172700 Dementia Pallidopontonigral degeneration, 168610 Supranuclear palsy, progressive, 601104 Supranuclear palsy, progressive atypical, 260540 Tauopathy and respiratory failure	false
APP	Alzheimer disease-1, APP-related Amyloidosis, cerebroarterial, Dutch type Schizophrenia Alzheimer_disease Amyloidosis Schizophrenia, chronic	true
ACHE	Blood_group Blood group, Yt system, 112100	false
APOE	Hyperlipoproteinemia Sea-blue_histiocyte_disease Alzheimer disease-2, 104310 Alzheimer_disease Myocardial_infarction Hyperlipoproteinemia, type III Myocardial infarction susceptibility Sea-blue histiocyte disease, 269600	true

Figure 5 Confirmation of genetic evidences for the efficacy of alternative medicines using RDF-TCM and Diseasome. We first use the RDF-TCM dataset to find genes associated with the Alzheimer's diseases and the herb *Ginkgo biloba*, and we then use the Diseasome database to search for the diseases associated with these genes. If an RDF-TCM gene is also associated with the Alzheimer's disease according to Diseasome, we then confirm that gene as an Alzheimer's gene. In this way, we use two datasets created by two different medical research communities to confirm genetic evidence for the herbs.

```
WHERE {

  http://www4.wiwiss.fu-berlin.de/drug-
  bank/resource/drugs/DB01273
     drugbank: possibleDiseaseTarget ?
     disease.
  ? disease owl: sameAs ? sameDisease.
  ?  altMedicine  tcm:  treatment  ?
  sameDisease.
  ? altMedicine rdf: type tcm: Medicine.
  ?  sameDisease  rdfs:  label  ?
  diseaseLabel.
  ?  altMedicine  rdfs:  label  ?
  altMedicineLabel.

}
```

Discussion

The data mashups and the SQUIN-powered application demonstrate how Linked Data may serve as the point of departure for data integration. It allows developers to access machine-processable datasets either using the exible SPARQL query language or using Linked Data query engines (e.g. SQUIN) to access distributed

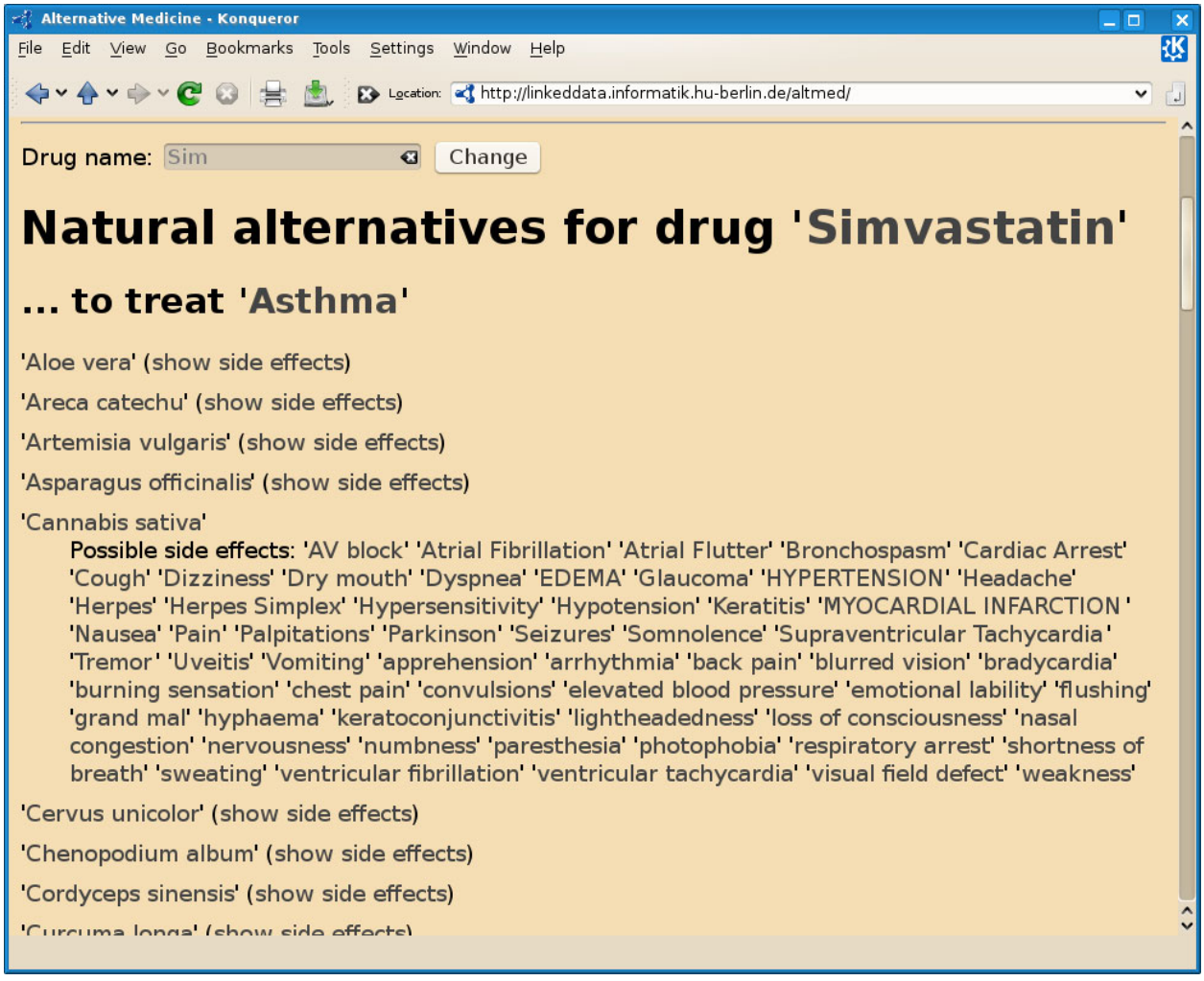


Figure 6 Finding alternative medicines as well as their side effects powered by SQUIN. To find alternative medicines to *Simvastatin* as well as their side effects powered by SQUIN, we use a Linked Data query engine, which allows one SPARQL query to access 6 distributed linked datasets published at different sources, including Drugbank, Diseasome, SIDER, LinkedCT, Dailymed and RDF-TCM.

information as one Web of Data. These two different approaches are complementary: the SQUIN-powered application may be included as one of the widgets in the mash-up application, and the mash-up approach may be used to support applications that need to perform schema and semantic mappings between datasets, which cannot be achieved with SQUIN.

Publishing RDF-TCM as Linked Data enables us to address some disadvantages of data integration approaches based on the relational database technologies [55], which are not necessarily unique to CM data resources. Firstly, Linked Data helps us address the identity linking and management. Most relational life science databases tend to use a local identifier for their data resources, even though overlapping information or existing identifiers have been provided elsewhere. Integrating these databases must first overcome the identity mapping problem. Linked Data promotes the use of uniform resource identifiers, i.e the URIs. Although uniform identifiers are yet to be established, there are ongoing active efforts in drawing together the community. Moreover, Linked Data allows the interlinking between URIs to be expressed in structured and explicit statements, such as RDF statements. Such RDF data links may be published by anyone and kept independent of the datasets. The other issue related to relational database integration is that often no programmatic access is provided for these databases and only a data dump is available. Linked Data on the other hand enables descriptions about an entity to be expressed in structured format (i.e. RDF) and retrievable by its URI. Linked Data also allows datasets to be accessible through the standard SPARQL query language and protocol. Our example applications have demonstrated how

these two ways of consuming RDF-TCM provide the flexibility of integrating biomedical knowledge available in Linked Data format.

In contrast to the existing ontology-based approach [56,57], our RDF-TCM dataset is described with a very lightweight schema to publish a large number of instances. Associating lightweight semantics reduces the cost in publishing data and such datasets can satisfy most initial user requirements; while the heavier semantic approach would require more efforts in ontology engineering that makes data publication much more expensive. Linked data is most useful to data integration tasks at a syntactic level, such as the two example applications presented here; an ontology-based approach would be more useful for addressing requirements and issues requiring a controlled vocabulary to link together information at the semantic level. Investigating whether the latter approach would be needed for a Linked Data approach, such as one providing the integration of medical datasets by the disease names (and their classifications), is part of our future work.

Conclusion

The Linked Data approach provides a set of best practices encouraging data providers to publish their data in an openly-accessible and programmatically-accessible manner. The benefit of such approach is demonstrated by the two examples in this study, consuming linked datasets to build useful applications. As improved tools and technologies of Linked Data are being made available, the CM and WM linked datasets will increase in number and volume through stepwise changes in multilingual publication and query practices among the CM community and become openly accessible to a larger community. Our Linked Data publication methodology reduces the efforts and errors in publishing linked datasets by systematizing and explicating the design decisions. Our further work is the evaluation of the correctness and completeness of the mapping between different datasets.

Abbreviations

CM: Chinese Medicine; WM: Western medicine; URIS: Uniform Resource Identifiers; RDF: Resource Description Framework; URLS: Uniform Resource Locators; HTTP: Hyper-Text Transfer Protocol; W3C: World Wide Web Consortium; LOD: Linking Open Data; KEGG: Kyoto Encyclopedia of Genes and Genomes; PDB: Protein Data Bank; LODD: Linking Open Drug Data; MESH: Medical Subject Headings Vocabulary; VOID: Vocabulary of Interlinked Datasets; NLM: National Library of Medicine; OMIM: Online Mendelian Inheritance in Man.

Acknowledgements

The work of JZ is funded by EPSRC grant EP/G049327/1. JZ would also like to thank Anja Jentzsch and the Linked Open Drug Data project members for helping create some of the data links between RDF-TCM and LODD and Olaf Hartig for his contribution in the SQUIN-powered application.

Authors' contributions

The author conducted the research and wrote this article.

Competing interests

The author declares that they have no competing interests.

Received: 29 January 2010 Accepted: 27 July 2010
Published: 27 July 2010

References

1. Bizer C, Heath T, Berners-Lee T: **Linked data - the story so far.** *Int J Semant Web Inf Syst, Special Issue on Linked Data* 2009, **53(3)**:1-22.
2. Berners-Lee T, Fielding R, Masinter L: **Uniform Resource Identifiers (URI): Generic Syntax.** [http://www.ietf.org/rfc/rfc2396.txt].
3. Klyne G, Carroll JJ, McBride B: **Resource Description Framework (RDF): Concepts and Abstract Syntax.** [http://www.w3.org/TR/rdf-concepts/].
4. Fielding RT, Taylor RN: **Principled design of the modern Web architecture.** *ACM Transactions on Internet Technology* 2002, **2(2)**:115-150.
5. Fielding R, Gettys J, Mogul J, Frystyk H, Masinter L, Leach P, Berners-Lee T: **Hypertext Transfer Protocol-HTTP/1.1.** [http://www.w3.org/Protocols/rfc2616/rfc2616.html].
6. Prud'hommeaux E, Seaborne A: **SPARQL query language for RDF.** 2008 [http://www.w3.org/TR/rdf-sparql-query/].
7. Clark KG, Feigenbaum L, Torres E: **SPARQL protocol for RDF.** 2008 [http://www.w3.org/TR/rdf-sparql-protocol/].
8. **The Linking Open Data project.** [http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData].
9. Bizer C, Heath T, Idehen K, Berners-Lee T: **Linked data on the web (LDOW2008).** *Proceeding of the Seventeenth international conference on World Wide Web, Beijing, China* 2008, 1265-1266.
10. Bizer C, Cyganiak R: **D2R server-publishing relational databases on the Semantic Web.** *Poster at the Fifth International Semantic Web Conference, Athens, GA, USA* 2006.
11. Auer S, Dietzold S, Lehmann J, Hellmann S, Aumueller D: **Triplify: light-weight Linked Data publication from relational databases.** *Proceedings of the Eighteenth International Conference on World Wide Web, Madrid, Spain* 2009, 621-630.
12. Berners-Lee T, Chen Y, Chilton L, Connolly D, Dhanaraj R, Hollenbach J, Lerer A, Sheets D: **Tabulator: exploring and analyzing Linked Data on the Semantic Web.** *Proceedings of the Third International Semantic Web User Interaction Workshop, Athens, Georgia, USA* 2006.
13. **sig.ma: live Web views on the Web of Data.** [http://sig.ma/].
14. Hartig O, Bizer C, Freytag JC: **Executing SPARQL queries over the Web of Linked Data.** *Proceedings of the Eighth International Semantic Web Conference 2009, Washington D.C., USA* 2009, 293-309.
15. Oren E, Delbru R, Catasta M, Cyganiak R, Stenzhorn H, Tummarello G: **Sindice. com: a document-oriented lookup index for open Linked Data.** *International Journal of Metadata, Semantics and Ontologies* 2008, **3**:37-52.
16. Ding L, Finin T, Joshi A, Pan R, Cost RS, Peng Y, Reddivari P, Doshi V, Sachs J: **Swoogle: a Search and metadata engine for the Semantic Web.** *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management, Washington D.C., USA* 2004, 652-659.
17. Belleau F, Nolin MA, Tourigny N, Rigault P, Morissette J: **Bio2RDF: towards a mashup to build bioinformatics knowledge systems.** *J Biomed Inform* 2008, **41**:706-716.
18. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28**:27-30.
19. Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C: **The Protein Data Bank.** *Acta Crystallographica Section D: Biological Crystallography* 2002, **58(6)**:899-907.
20. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE: **PharmGKB: the pharmacogenetics knowledge base.** *Nucleic Acids Res* 2002, **30**:163-165.
21. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH: **PubChem: a Public information system for analyzing bioactivities of small molecules.** *Nucleic Acids Res* 2009, , **37 Web Server:** W623-W633.
22. **Linking Open Drug Data.** [http://esw.w3.org/topic/HCLSIG/LODD].

23. Jentzsch A, Zhao J, Hassanzadeh O, Cheung KH, Samwald M, Andersson B: **Linking Open Drug Data.** *Proceedings of the Second Triplification Challenge 2009, Graz, Austria* 2009.
24. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z: **Dbpedia: a nucleus for a Web of open data.** *Proceedings of the Sixth International Semantic Web Conference, Busan, Korea* 2007, 722-735.
25. Fang YC, Huang HC, Chen HH, Juan HF: **TCMGeneDIT: a database for associated Traditional Chinese Medicine, gene and disease information using text mining.** *BMC Complement Altern Med* 2008, **8**:58.
26. **HULU TCM professional web site.** [http://www.hulu.com.tw/].
27. Chen X, Zhou H, Liu YB, Wang JF, Li H, Ung CY, Han LY, Cao ZW, Chen YZ: **Database of Traditional Chinese Medicine and its application to studies of mechanism and to prescription validation.** *Br J Pharmacol* 2006, **149(8)**:1092-1103.
28. **Chinese medicine resource web.** [http://web.archive.org/web/20080612040654/http://www.spec-g.com.tw/newherb/], [W3C Interest Group Note 03 December 2008].
34. Volz J, Bizer C, Gaedke M, Kobilarov G: **Discovering and maintaining links on the Web of data.** *Proceedings of the Eighth International Semantic Web Conference, Washington D.C., USA* 2009, 650-665.
35. Berners-Lee T: **Linked Data.** [http://www.w3.org/DesignIssues/LinkedData.html].
36. **Shared Names.** [http://sharedname.org/].
37. **Concept Web Alliance.** [http://conceptweblog.wordpress.com/].
38. **PURL.** [http://purl.org/].
39. Bodenreider O, Stevens R: **Bio-ontologies: current trends and future directions.** *Brief Bioinform* 2006, **7(3)**:256-274.
40. Pearson H: **Biology's name game.** *Nature* 2001, **411(6838)**:631-632.
41. **W3C HCLS knowledge base.** [http://hcls.deri.org/].
42. **STITCH Linked Data.** [http://www4.wiwiss.fu-berlin.de/stitch/].
43. Alexander K, Cyganiak R, Hausenblas M, Zhao J: **Describing linked datasets-on the design and usage of voiD, the 'Vocabulary of Interlinked Datasets'.** *Proceedings of the Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09), Madrid, Spain* 2009 [http://vocab.deri.ie/void].
44. Hartig O, Zhao J: **Using Web data provenance for quality assessment.** *Proceedings of the International Workshop on Semantic Web and Provenance Management, Washington D.C., USA* 2009.
45. **Pubby: a Linked Data frontend for SPARQL endpoints.** [http://www4.wiwiss.fu-berlin.de/pubby/].
46. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M: **DrugBank: a knowledgebase for drugs, drug actions and drug targets.** *Nucleic Acids Res* 2007, , **Database:** D901-6.
47. **DailyMed.** [http://dailymed.nlm.nih.gov/dailymed/].
48. Campillos M, Kuhn M, Gavin AC, Jensen LJ, Bork P: **Drug target identification using side-effect similarity.** *Science* 2008, **321(5886)**:263-6.
49. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL: **The human disease network.** *PNAS* 2007, **104(21)**:8685-8690.
50. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P: **STITCH: interaction networks of chemicals and proteins.** *Nucleic Acids Res* 2008, , **36 Database:** D684-8.
51. Klein TE, Chang JT, Cho MK, Easton KL, Fergerson R, Hewett M, Lin Z, Liu Y, Liu S, Oliver DE, Rubin DL, Shafa F, Stuart JM, Altman RB: **Integrating genotype and phenotype information: an overview of the PharmGKB project.** *Pharmacogenomics* 2001, **1**:167-170.
52. **Search for potential alternative medicines by mash-ups.** [http://www.open-biomed.org.uk/admed/admedapps/searchTCMByDiseaseName/].
53. **Linked Life Data, part of the EU LarKC project.** [http://linkedlifedata.com/sparql].
54. Hartig O, Zhao J: **Find Traditional Chinese Medicine as an alternative to western drugs.** Washington D.C., USA 2009, [In First Linked Data-a-thon, in conjunction with International Semantic Web Conference 2009].
55. Goble C, Stevens R: **State of the nation in data integration for bioinformatics.** *J Biomed Inform* 2008, **41(5)**:687-693.
56. Cheung KH, Yip KY, Smith A, Deknikker R, Masiar A, Gerstein M: **YeastHub: a Semantic Web use case for integrating data in the life sciences domain.** *Bioinformatics* 2005, **21**:85-96.
57. Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, Doherty D, Forsberg K, Gao Y, Kashyap V, Kinoshita J, Luciano J, Marshall MS, Ogbuji C, Rees J, Stephens S, Wong GT, Wu E, Zaccagnini D, Hongsermeier T, Neumann E, Herman I, Cheung KH: **Advancing translational research with The Semantic Web.** *Brief Bioinform* 2007, **8(Suppl 3)**:S2.

doi:10.1186/1749-8546-5-27
Cite this article as: Zhao: **Publishing Chinese medicine knowledge as Linked Data on the Web.** *Chinese Medicine* 2010 **5**:27.