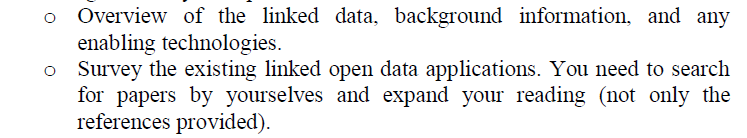
A Semantic Web Primer

The general vision of semantic web can be summarized in a single phrase: to make the web more accessible to computers.



Motivation for the semantic web (linked open data)

Where are we now?

There is rapidly maturing technology to support all phases of deployment of Semantic Web technology, and the number of substantial adoptions, both in commercial and public organizations, is growing rapidly. However, major challenges remain, such as dealing with the ever-increasing scale, lowering the barrier of adoption, and of course fighting that omnipresent bane of information systems: semantic heterogeneity.

What is the background before LOD?

A key aspect of the traditional web is the fact that its content is distributed, both in location and in ownership: web pages that link to each other often live on different web servers, and these servers are in different physical locations and owned by different parties.



In recent years, the concept of Linked Data, and the so-called Web of Linked Data, has attracted tremendous attention from both the academic world and real application world.

From Bibliographic Records to Data

Within this context, therefore, the term “Linked Data” refers to the method by which data can be displayed, published, linked and exchanged using Uniform Resource Identifier (URIs) and Resource Description Framework (RDF). Linked Data (LD) technologies make it possible to connect and enrich metadata, so that different representations of the same content can be searched and linked, thus relating resources that come from different sources and domains. Background

The Open Data initiative is closely related and linked to the conception of open government, with its “philosophy of open access to certain data without copyright restrictions” (Ferrer ; Peset; Benavent, 2011, p. 162).

It is necessary to clarify that the term Open Data does not refer to the mere availability of data on the network, i.e., Internet publication in such a way that data can be read and down­loaded. Hernández and García (2013, p. 260) claim that “to be truly open the data must be available online, preferably for download, but they must also have some sort of legal license so that they can be used, reused and redistributed, even mixed with other data, at the minimum sub­ject to “attribution” (recognition of authorship, the person who created it), or “share alike “(the norm that any exploitation made of such data, including derivative works, should maintain the same license to be disclosed).”

The *Guía breve de Linked Data* (2014) specifies that “the Semantic Web is not just about publishing data on the Web, but about doing it in such a way that they can be linked to others, so that people and machines can explore the web of data, and be able to reach related information referred from other initial data. “

The term Linked Data (LD) thus refers to a set of best practices for publishing and interlinking structured data on the Web. In the case of LD, data should be published “in accordance with the principles designed to fa­cilitate linkages among datasets, element sets, and value vocabularies” (Berners-Lee, 2011). These best practices were introduced in 2006 by Tim Berners-Lee in his Web architecture note Linked Data and they have become known as the Linked Data principles. These principles are the following:

1. Use URIs as names for things.

2. Use HTTP URIs, so that people can look up those names.

3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL).

4. Include links to other URIs, so that people can discover more things.

That is, this set of best practices or recom­mendations included under this concept will allow sets of data, information, and knowledge to be exhibited, shared, and connected on the Semantic Web using URIs and RDF.”

LOD从定义上看，是将互联网打碎成无数个数据，再根据数据之间的关系相连，这样可以实现各种不同的、很有意义的应用。但是这和搜索引擎的区别在哪里呢？搜索引擎不也是通过关键词查找，能够快速从整个互联网获取资源吗？不，这并不一样，试想一下，如果我现在这个应用程序需要从互联网上获取迪士尼公司拍过的电影的名单，我该怎么办？直接Google “Movies that produced by Disneyland company” ? 显然如果机器和人一样聪明，可以从一大串字符里面，准确的提取出名单来，并且格式正确地提供给应用程序，这样当然是可行的。但是计算机没有这么聪明，互联网上也没有为每一种数据，每一串数据提供了获取数据的应用程序接口，所以我们该怎么办？答案就是LOD和语义网，我的理解是，语义网是给机器来理解的，HTML是给人来理解的，这样就相当于把网络分成两种版本，一种给机器，一种给人类，机器那种遵循着LOD的规范，这样我的应用程序就不需要通过搜索明文了，直接生成一个RDF格式的搜索命令，发到互联网上面，比如我想知道迪士尼公司拍过那些电影，我就可以先寻找Disneyland company这个数据串，再根据RDF和我想知道的数据之间的关系，进行查询，这整个过程就很直接，我通过Disneyland词条Link到Movie那里，或者可以通过Movie词条link到Disneyland那里，这就是所谓的，Linked Data概念。

Challenges and Opportunities

Due to the massive growth of available data, conventional methods of data integration are bound to fail while the complexity of processes within organizations ask for more agile options to link and mash-up data in a qualified way. Availability and matching of diverse data sources become more crucial and therefor the need for standards-based tools is growing. 传统整合数据的方式会失效，当数据复杂度急剧加快的时候

In most cases public sector information is not published in a machine-processable format that would allow data re-users from the public and private sector to automate combining public data with other public or proprietary data sources (Analysis Report of Public Sector Data and Knowledge Sources^ 2012). 这段我觉得阐述的是从方法论的层面解释Linked Data存在的必要，因为现阶段的框架并不支持语义网

The vast majority of Public Data sources do not provide the datasets in a standard format which would support true semantic enrichment and interlinking of data (such as Resource Description Framework, RDF). 承接上一段

During the last years, significant research activities have appeared that focus on industrial relevant scenarios, such as the LATC (LATC Project 2016) and the LOD2 (LOD2 Project 2016) projects that aim to contribute high-quality interlinked versions of public semantic web datasets and promoting their use in new cross-domain applications by developers across the globe (Verborgh et al. 2014). In the context of these efforts and emerging tools, while there is considerable support for linked data in other issues, such as storage (Virtuoso, Sesame), linkage (Silk Framework), discovery and publishing (SPARQL standard) and even visualization of RDF graphs (LodLive, CubeViz), there are very limited options for renovating existing data into Linked Data. Currently available solutions either support specific structured data formats, such as spreadsheets (XLWrap) and relational databases (D2R, Triplify) or provide RDF representations of data for specific sources (DBpedia). Lastly, most existing work related to exploring and visualizing RDF is limited on concrete domains and concrete datatypes and is mainly focused towards academic researchers that are familiar with the semantic web technologies.

Despite the envisioned benefits of Linked Data, a number of significant challenges and issues prevent the massive adoption of the Linked Data tools presented in the previous section by public data providers. 局限性

Linked Data-The Story So Far

Technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets.

With significant volumes of Linked Data being published on the Web, numerous efforts are under­way to research and build applications that exploit this Web of Data. At present these efforts can be broadly classified into three categories: Linked Data browsers, Linked Data search engines, and domain-specific Linked Data applications. 分类

Structure:

1. Background/Intro.
2. Reason
3. How
4. Importance
5. Technologies

Architecture of Linked Data Applications

1. The Crawling Pattern

Applications that implement this pattern crawl the Web of Data in advance by traversing RDF links. Afterwards, they integrate and cleanse the discovered data and provide the higher layers of the applications with an integrated view on the original data. Separating the tasks of building up the cache and using this cache later in the application context enables applications to execute complex queries with reasonable performance over large amounts of data. The disadvantage of the crawling pattern is that data is replicated and that applications may work with stale data, as the crawler might only manage to re-crawl data sources at certain intervals.

Case study:

These search engines integrate data from thousands of data sources and thus nicely demonstrate the advantages of the open, standards-based Linked Data architecture, compared to Web 2.0 mashups which rely on a fixed set of data sources exposing proprietary interfaces. 这就像是，传统Web使用传统数据结构比如链表、集合之类的容器存储数据，LOD直接采用各种树来存储，算法复杂度直接下降一个等级

The user is presented with a search box into which they can enter keywords related to the item or topic in which they are interested, and the application returns a list of results that may be relevant to the query. However, rather than simply providing links from search results through to the source documents in which the queried keywords are mentioned, Linked Data search engines provide richer interaction capabilities to the user which exploit the underlying structure of the data. For instance, Falcons enables the user to filter search results by class and therefore limit the results to show, for example, only persons or entities belonging to a specific subclass of person, such as athlete or politician. Sig.ma, Falcons and SWSE provide summary views of the entity the user selects from the results list, alongside additional structured data crawled from the Web and links to related entities. 搜索结果更加丰富，这些可能相对于主流的搜索引擎而言并不算什么，用户应该也没有很惊喜，可能认为是理所当然的，但是如果用传统的关键词查找匹配，是根本找不到这么全的，网状的信息结构的，而且效率也没这么高

The Sig.ma search engine applies vocabulary mappings to integrate Web data as well as specific display templates to properly render data for human consumption. 搜索引擎会综合多个数据源，体现了Linked的思想

Queries are formulated by the user in an exploratory fashion and can be far more expressive than queries that Google and Yahoo can currently answer. For instance, VisiNav answers the query *"give me the URLs of all blogs that are written by people that Tim Berners-Lee knows!"* with a list of 54 correct URLs. Google and Yahoo just return links to arbitrary web pages describing Tim Berners-Lee himself. 这才是我理解的基于LOD的搜索引擎，可以通过解析搜索字串，得到数据的关系，再精确查找，智能化。

基于LOD的搜索引擎的应用主要体现在了为人类操作便捷来考虑，user-friendly

1. **The On-The-Fly Dereferencing Pattern** is implemented by Linked Data browsers discussed in Section [6.1.1.1](http://linkeddatabook.com/editions/1.0/#sec:browsers). Within this pattern, URIs are dereferenced and links are followed the moment the application requires the data. The advantage of this pattern is that applications never process stale data. The disadvantage is that more complex operations are very slow as they might involve dereferencing thousands of URIs in the background. [[57]](http://linkeddatabook.com/editions/1.0/#DBLP:conf-semweb-HartigBF09) propose an architecture for answering complex queries over the Web of Data that relies on the on-the-fly dereferencing pattern. As the results of this work show, data currency and a very high degree of completeness are achieved at the price of very slow query execution. 解引用意思是，获取引用的那个地址的数据，若p是引用（指针），\*p就是解引用，获取p地址的数据

Case:

Linked Data Browser: Just as traditional Web browsers allow users to navigate between HTML pages by following hypertext links, Linked Data browsers allow users to navigate between data sources by following RDF links. For example, a user may view DBpedia’s RDF description of the city of Bristol (UK), follow a *hometown* link to the description of the band Portishead (which originated in the city), and from there onward into RDF data from Freebase describing songs and albums by that band. The result is that a user may begin navigation in one data source and progressively traverse the Web by following RDF rather than HTML links.

1. **The Query Federation Pattern**

不同模式之间的对比，总结

Hartig and Langegger present a deeper comparison of the advantages and disadvantages of the architectural patterns in [[58]](http://linkeddatabook.com/editions/1.0/#hartigQueryArchitectures). The appropriate pattern (or mixture of these patterns) will always depend on the specific needs of a Linked Data application. The factors that determine the decision for a specific pattern are:

1. the number of data sources that an application intends to use,
2. the degree of data freshness that is required by the application,
3. the required response time for queries and user interactions,
4. the extent to which the applications aims to discover new data sources at runtime.

Olaf Hartig and Andreas Langegger. A database perspective on consuming linked data on the web. Datenbank-Spektrum, 10:57–66, 2010. 10.1007/s13222-010-0021-7. http://dx.doi.org/10.1007/s13222-010-0021-7DOI: 10.1007/s13222-010-0021-7

**1. Accessing the Web of Data.**

The basic means to access Linked Data on the Web is to dereference HTTP URIs into RDF descriptions and to discover additional data sources by traversing RDF links. In addition, relevant data can also be discovered via Linked Data search engines and might be accessed via SPARQL endpoints or in the form of RDF data dumps. 获取原始数据

**2. Vocabulary Mapping.**

Different Linked Data sources may use different RDF vocabularies to represent the same type of information. In order to understand as much Web data as possible, Linked Data applications translate terms from different vocabularies into a single target schema. This translation may rely on vocabulary links that are published on the Web by vocabulary maintainers, data providers or third parties. Linked Data applications which discover data that is represented using terms that are unknown to the application may therefore search the Web for mappings and apply the discovered mappings to translate data to their local schemata. 将原始数据与应用层的Schema结合

**3. Identity Resolution.**

Different Linked Data sources use different URIs to identify the same entity, for instance, a person or a place. Data sources may provide owl:sameAs links pointing at data about the same real-world entity provided by other data sources. In cases where data sources do not provide such links, Linked Data applications may apply identity resolution heuristics in order to discover additional links. 解析

**4. Provenance Tracking.**

Linked Data applications rely on data from open sets of data sources. In order to process data more efficiently, they often cache data locally. For cached data, it is important to keep track of data provenance in order to be able to assess the quality of the data and to go back to the original source if required. 获取缓存

**5. Data Quality Assessment.**

Due to the open nature of the Web, any Web data needs to be treated with suspicion, and Linked Data applications should thus consider Web data as claims by different sources rather than as facts. Data quality issues might not be too relevant if an application integrates data from a relatively small set of known sources. However, in cases where applications integrate data from the open Web, applications should employ data quality assessment methods in order to determine which claims to accept and which to reject as untrustworthy. 评估

**6. Using the Data in the Application Context.**

After completing tasks 1 to 5, the application has integrated and cleansed Web data to an extent that is required for more sophisticated processing. Such processing may in the most simple case involve displaying data to the user in various forms (tables, diagrams, other interactive visualizations). More complex applications may aggregate and/or mine the data, and they may employ logical reasoning in order to make implicit relationships explicit.应用