十、以图模式组织信息

关联数据

1. 一组在Web上发布和关联结构化数据的最佳实践
2. 首先，对事物进行结构化和有意义的描述；然后将数据相互联系起来，以支持知识的整合和融合

关联数据原则

1. 采用URI作为事物的名称进行标识
2. 采用HTTP URIs，使得人们可以查找这些名称（html），在web上可以打开
3. 当有人查找URI时，提供有用的RDF信息
4. 将该事物与其他相关事物相关联
    a. 用URI代替独立的对象
    b. 将前一个RDF三元组中被标识的宾语作为主语，然后创建新的RDF三元组来描述它

注意

1. 传统Web是面向人阅读的；传统Web上的结构化数据只能通过Web API来访问
2. 传统Web上的超链接无法显示网页和网页之间的语义关系
3. 关联数据在Web浏览器上可浏览，计算机可读可理解；不是采用超链接进行连接

DC元数据规范

1. 简单性、通用性和可扩展性的基本原则

MARC

1. 机器可读目录的缩写
2. 这是一种计算机化的记录编目所需信息的方法，包含描述性编目、主题标题和其他访问点、分类号和其他索书号信息

DC元数据元素

1. 一个用于跨领域信息资源描述的标准
2. 所有的DC元素都是可选的，可重复的，可以任何顺序出现

DC 元素	含义	值
title	资源的题名，中英文皆可	使用纯文本
description	资源的描述，摘要	纯文本
identifier	标识	纯文本（ISBN, URL）或URI标识符
rights	资源的版权信息	纯文本的描述/URL网址中的描述
creator	作者	人名地名或者机构名，来自人名地名规范文档；文本/URI标识的实体（说明已经被关联数据表示了）——可以将本资源与其他资源进行关联
contributor	合作者	
publisher	出版社	
date	日期	W3CDTF格式：2022-06-10
coverage	日期范围	同上
	地理空间范围	地名规范文档、地名词表、URI
relation	另一个相关联的资源	该资源的文本题名/URI
source	一个资源，是被描述资源的来源	同上
language	ISO 639-2/639-3	英文/中文前三个字母
type	DCMI提供的一组资源list	Resource Type List中的受控词
format	IMT，Internet媒体类型	Media Type List中的值
subject	资源的分类号：DDC、UDC、CLC；主题词：主题标题词表一个或多个词；汉语主题词表	受控词

type：`http://purl.org/dc/terms/DCMIType` – `http://purl.org/dc/dcmitype/Collection`

format：`http://purl.org/dc/terms/IMT`

subject：`http://purl.org/dc/terms/DDC`

例子 XML 复制代码

```xml
<?xml version="1.0"?>
<metadata
  xmlns="http://example.org/myapp/"
  xmlns:dc=http://purl.org/dc/elements/1.1/
  xmlns:dcterms="http://purl.org/dc/terms/">
  <dc:title>UKOLN Homepage</dc:title>
  <dc:creator>UKOLN, University of Bath</dc:creator>
  <dc:subject xsi:type="dcterms:LCSH">network information support</dc:subject>
  <dc:subject xsi:type="dcterms:DDC">062</dc:subject>
  <dc:description>UKOLN is a national focus of expertise …</dc:description>
  <dc:description xml:lang="fr">UKOLN est un centre national ...</dc:description>
  <dc:language xsi:type="dcterms:ISO639-2">eng</dc:language>
  <dc:relation xsi:type="dcterms:URI">http://www.bath.ac.uk/</dc:relation>
  <dc:identifier xsi:type="dcterms:URI">http://www.ukoln.ac.uk/</dc:identifier>
  <dc:date xsi:type="dcterms:W3CDTF">2001-07-18</dc:date>
  <dc:format xsi:type="dcterms:IMT">text/html</dc:format>
  <dc:type xsi:type="dcterms:DCMIType">http://purl.org/dc/dcmitype/Text</dc:type>
</metadata>
```

注意

1. MARC和DC都是对信息资源进行描述，都可采用受控词表控制标引词汇，MARC数据可无损地转换为DC数据
2. 采用关联数据发布书目数据可以为图书馆书目记录提供更丰富的相关信息，可以实现对书目信息的语义检索，实现书目数据的开放和共享

知识图谱

1. 知识图谱是一种用图模型来描述知识和建模世界万物之间的关联关系的技术方法
2. 知识图谱是由一些相互连接的实体和它们的属性构成的
3. 知识图谱由节点和边组成，节点可以是实体，或是抽象概念，边是节点之间的关系

4. 知识图谱可以看作是Semantic Web的一种简化后的商业实现
5. 知识图谱是相互关联的实体描述的集合
    a. 描述是以形式化结构进行的, 使人和计算机都能以有效和明确的方式处理它们
    b. 实体描述相互促进，形成一个网络，每个实体是其相关实体的描述的一部分
6. 知识图谱可以被理解为
    a. 数据库：能采用结构化查询查询数据
    b. 图：可以作为网络数据结构来被分析
    c. 知识库：数据具有形式语义，可被用来解释数据并推断出新的事实
7. 不是所有的rdf图都是知识图谱
    a. 是连接和图，而不是用来表示数据的语言，构成了知识图谱
    b. 数据的图表示通常是有用的，但可能没有必要了解数据的语义知识（？）
    c. RDF会用来表示统计数据，而这些数据间并没有语义关联
8. 不是所有的知识库都是知识图谱
    a. 知识图谱的一个关键特点是实体描述应该相互关联
    b. 没有形式结构和语义的知识库不是知识图谱

知识图谱与关联数据

1. 联系
    a. 两者在本质上相同，都采用图模型，都采用“实体–关系–实体”三元组结构表示知识
    b. 关联数据其实就是一个采用RDF数据模型表示的全球性知识图谱
    c. 语义网和关联数据是知识图谱的前身，而知识图谱则是工程化的语义技术，具有更大的实际应用价值
2. 区别
    a. 两者的目标并不相同：语义网（关联数据）的目标是构建一个开放共享的全球知识库，知识图谱则更倾向于成为一个机构私有的封闭知识库
    b. 不是所有的知识图谱都是关联数据
    c. 知识图谱并不强调底层数据一定要采用RDF数据库存储，可以采用任何其他的与RDF兼容的数据库
    d. 知识图谱不一定要链接到外部的知识图谱，更强调有一个本体层来定义实体的类型和实体之间的关系

知识图谱的构成

1. 实体：指的是具有可区别性且独立存在的某种事物 `某一个人、某一个城市`
2. 语义类（概念）：具有同种特性的实体构成的集合 `国家、民族、书籍`
3. 内容：通常作为实体和语义类的名字、描述、解释等
4. 属性：从一个实体指向它的属性值 `面积`
5. 属性值主要指对象指定属性的值 `960万平方公里`
6. 关系：连接两个实体

知识图谱的价值

1. 辅助搜索
2. 辅助自然语言理解与处理
3. 辅助自动问答
4. 辅助大数据分析

注意

1. 知识图谱帮助对检索目标进行消歧，为检索到的事物提供额外的信息
2. ~~无法从检索到的信息中抽取出结构化信息；无法从HTML网页中抽取出嵌入的语义信息~~
3. 维基百科的语义网版DBPedia、从OCLC中下载的RDF格式的书目数据、采用语义网络描述事物及其相互间关系都是知识图谱
4. 关联数据是开放的、全局的；知识图谱是封闭的、局部的
5. 关联数据是知识图谱的前身