
元数据
METADATA
用数据的数据管理你的世界
[美]杰弗里·波梅兰茨（Jeffrey Pomerantz）◎著
李梁◎译
世界的本质是数据——数据管理时代来了！
推动建立企业元数据管理平台 深入了解无所不在的元数据
中信出版集团

METADATA
元数据

用数据管理你的数据

[美] 杰弗里·波梅兰茨（Jeffrey Pomerantz）◎著
李梁◎译

中信出版集团 · 北京

图书在版编目（CIP）数据

元数据 /（美）杰弗里·波梅兰茨著；李梁译. --
北京：中信出版社，2017.2
书名原文：Metadata
ISBN 978-7-5086-7078-2

Ⅰ.①元… Ⅱ.①杰…②李… Ⅲ.①元数据－研究
Ⅳ.①G250

中国版本图书馆CIP数据核字〔2016〕第293477号

元数据

著　　者：[美]杰弗里·波梅兰茨
译　　者：李　梁
出版发行：中信出版集团股份有限公司
（北京市朝阳区惠新东街甲4号富盛大厦2座　邮编　100029）
承 印 者：北京诚信伟业印刷有限公司

开　　本：880mm×1230mm　1/32
印　　张：7
字　　数：132千字
版　　次：2017年2月第1版
印　　次：2017年2月第1次印刷
京权图字：01-2015-8589
广告经营许可证：京朝工商广字第8087号
书　　号：ISBN 978-7-5086-7078-2
定　　价：49.00元

METADATA

目 录

METADATA

第一章
元数据概览

元数据是一张地图，是一种能用更为通俗易懂的形式表达对象复杂性的方法。

METADATA

元数据（metadata）在我们的周围无时不在，无处不在。当代社会中随处可见的电子设备，不是依靠元数据来运行，就是用于产生元数据，或者两者皆有。但当元数据真正发挥作用的时候，它却隐于幕后、默默无闻，就像根本不存在一样。一定程度上来说，这也让元数据在2013年夏天突然成了一个广受关注且具有争议性的话题。

2013年5月，美国国家安全局前外聘员工爱德华·斯诺登（Edward Snowden）飞往中国香港与英国《卫报》记者会面，向其披露了大量有关美国国家安全局在本土进行监听活动的机密文件。这些监听项目之一——“棱镜”（PRISM），涉及直接向电信公司搜集电话呼叫的数据。不用说，《卫报》对此事的报道成为轰动一时的大新闻。

美国媒体对斯诺登泄密事件反应不一，随着事态的发展，这些反应的变化耐人寻味。对于美国国家安全局暗中搜集美国

公民数据的行为，公众当时最直接的反应就是群情激愤。然而随着事件日趋明朗，让人们如释重负的是，美国国家安全局搜集的仅仅是与电话呼叫有关的元数据，而不是电话呼叫本身的内容。换句话说，美国国家安全局没有进行窃听活动。这很快缓和了公众的怒火。事态随后急转直下，媒体通过调查才发现，凭借“区区”元数据居然能推断出如此多的个人信息。在此之后，对这一事件的权威解读才终于公之于众。

MetaPhone项目是斯坦福大学法学院互联网与社会研究中心（Stanford Law School Center for Internet and Society）的研究人员于2013年年底进行的一项研究，旨在重现美国国家安全局搜集电话呼叫元数据采取的方法。他们发现，用“区区”元数据居然能推断出如此令人难以置信的信息量。MetaPhone的研究人员在报告中提到了这样一个案例：一位研究对象分别打电话联系了“家庭装修用品店、锁匠、水培植物经销商还有烟草大麻用具店”。也许，打这些电话是出于非常单纯、合理的原因，也许它们之间完全没有关系……但是这可能并不是我们大多数人会得出的结论。

许多元数据都与电话呼叫有关，尤其是手机呼叫。而在与电话呼叫有关的元数据片段中，最显而易见的就是拨打与接听双方的电话号码，其次就是电话呼叫的时间与通话时长。如果使用具备GPS（全球定位系统）功能的智能手机拨打电话，还

可搜集到拨打与接听双方的地理位置信息，至少可以精确到通话双方手机所在地区手机信号塔的信号范围。与手机呼叫关联的元数据还有很多，但是如此少量的信息也足以让倡导保护隐私的人士再三思忖。因为即使你没有在打电话，你的手机也会与本地手机信号塔之间交换数据。这样一来，移动运营商就能随时搜集你的位置信息以及一段时间内的活动轨迹——根据斯诺登披露的机密文件显示，移动运营商实际也在这样做。当然，前提是你一直带着自己的手机。

元数据这个词就这样成了一个公众话题。鉴于元数据如此广泛地存在，人们理应更好地去了解它，而公众也早应该进行这样的讨论。在当代世界中，计算活动无处不在，因此元数据像电网和高速公路网一样成了一种基础设施。这些当代基础设施的构成部分一方面发挥着不可或缺的作用，另一方面它们展现在我们面前的又只是冰山一角。比如，当你触动照明开关时，你就变成了大量技术与策略的最终用户。

分开来看，这些技术或策略也许微不足道、无关紧要……但是聚合在一起，就能带来深远的文化与经济影响。元数据亦是如此。就像电网和高速公路网一样，元数据不知不觉地融入日常生活的背景之中，理所当然地成为当代社会得以稳步前进的动力之一。

作为生活在现代世界的公民，我们熟悉电网、高速公路网

以及其他现代的基础设施，也对其有着合理（尽管可能并不完整）的了解。但是除非你是一位信息技术科学家，或者是为美国国家安全局工作的情报分析师，否则可能无法对元数据形成这样的认识。

这就是我写作本书的目的——向你介绍元数据，以及元数据涉及的诸多主题与问题。我将探讨什么是元数据及其存在的原因、适用于不同用户与用例的各类元数据以及使现代元数据成为可能的一些技术，还会预测元数据的未来路在何方。读完本书，你无论身在何处都会看到元数据。

这是一个元数据的世界，而你就身处其中。

隐形的元数据

当你走进书店、从书架上拿起这本书的时候，你就已经用到了元数据。什么吸引你来选择或拿起这本书？是书名、出版社还是封面设计？无论怎样，毫无疑问不会是本书的内容。当然，现在你正在读这本书，所以对其中的内容已经有了一些了解，但是在你拿起这本书之前没有这样的认识。这样一来，你就不得不依赖有关这本书的其他提示或信息片段才能做出这样的选择。而这些所谓的“其他信息片段”就是元数据，也就是“有关这本书的数据”。

元数据真正发挥作用时，会隐于幕后、默默无闻，就像根

本不存在一样。你对书名、出版社和封面设计等要素已经习以为常，甚至不会注意这本书是否有这些部分。但是如果这本书没有书名、出版社或封面设计，你反而会意识到这些部分的缺失。我们对有关书籍的元数据已经如此习惯，以至认为这是购书环境的一部分，不会对此多加思考。同样，我们对许多事物的元数据也已经习以为常，把它们作为日常环境的一部分，因此也不会去多加思考。为什么会这样呢？

元数据简史

英语中的元数据一词最早出现于 1968 年，但是其概念可以追溯到世界上第一座图书馆。这是根据亚里士多德的著作集《形而上学》（*Metaphysics*）特别创造的一个词。尽管亚里士多德从未用“形而上学”这个词来命名这些著作，但是在历史上这些著作都被收录在这本作品集之中，以示它们是《物理学》（*Physics*）的延续或讨论超脱于这一主题的内容。元数据一词与此类似，它是指超脱于数据的事物，即有关于数据的一条或多条陈述。从语言学角度来看，这个词虽是对希腊语前缀“meta-”的粗略翻译，却能与“meta”的日常用法保持一致，用于表明更高抽象层次的事物。

尽管元数据一词只有几十年的历史，然而几千年来图书馆管理员们一直在工作中使用着元数据，只不过我们现在所谓的

“元数据”在历史上被称为“图书馆目录信息”（information in the library catalog）。

图书目录中的信息解决了一个十分具体的问题：如何帮助用户在图书馆的馆藏书籍中找到具体的资料。历史学家们认为，卡利马科斯（Callimachus）在公元前 245 年前后为亚历山大图书馆制作的“卷录”（Pinakes）是世界上第一套图书目录。虽然接下来的几千年中只有部分“卷录”得以保存，但是人们仍然可以从中了解到以下几点：按体裁、书名以及作者姓名排列著作，并且对每位作者的生平进行一定的介绍。卷录除了收录著作摘要，还列出了每本著作共有多少行文字。回到 2 000 多年后的今天，图书目录中仍然采用了许多相同的信息片段：作者、主题、简介和篇幅等。然而公平地说，与卡利马科斯的卷录相比，如今的图书目录采用了更多信息片段。每本著作都有唯一的编目号码——根据某种编码方案（例如，杜威十进制分类法）设计的纯数字或字母数字混编字符串——来帮助图书馆用户在书架上找到著作。对于藏书量巨大的图书馆来说，图书编目号码尤为关键，因为读者必须首先找到相应的较大藏书区，然后才能进一步寻找单本书籍。很难想象卡利马科斯如何在没有发明图书编目号码的情况下构思出了卷录——据说亚历山大图书馆藏书有 50 万册之多，用当代的标准来看，这个级别的藏书量也相当之大。

卷录是一套卷轴。如果你曾经在犹太教会堂里读过《妥

拉》（*Torah*），就会知道卷轴的“用户界面”并不是那么友好——在不同章节之间切换简直就是一种挑战。实际上，犹太历（Simchat Torah）的节假日都是为了庆祝《妥拉》诵经即将结束、一切要重新来过。如果你从来没有读过《妥拉》，也可以比较其他类似卷轴的技术，比如录音磁带或VHS（家用录像系统）磁带，过去，磁带上面往往有即时贴来提醒我们“听后倒带、举手之劳”。简而言之，从实用性的角度来看，编写卷录绝非易事。

从很多方面来看，我们这些现代人直接称之为“书”的手抄本采用的“用户界面”要比卷轴先进得多。因此，手抄本自然而然在发明后就被用作图书馆目录。以图书形式呈现的图书馆目录往往被名副其实地称为“排架表”（shelf list）。物如其名，它是排架上的书籍列表，列表上的书目往往按图书的采购顺序先后排列。这种排序方法的优点在于方便添加新条目——只需在列表末尾增加即可，但是当人们想要从列表上查找某条书目的时候仍然不太方便。

法国人在大革命前后发明了卡片式目录（card catalog）后，图书馆目录的发展向前迈了一大步。这种创新的方法分解了排架表，让添加、删除条目以及查找单独条目变得更加方便。卷轴或手抄本式的目录在完成后不便编辑，但是向卡片式目录中增加条目时，你只需把新的卡片插到正确的位置即可。

卡片式目录分解了图书馆目录，让每条记录（即为某本书建立的每个条目）成为可以独立操控的对象。每条记录中的数据片段（书名、作者等等）早在卷录时代就已经被分离了出来。即使目录卡没有采用书名、作者等方式来标记单独数据条目，每个数据片段所表达的类别仍然一目了然。因此，我们可以将目录卡分离为两个维度：单独项目的记录以及所有项目共用的数据类别。

如果我们按照这两个维度分解目录卡，那么就能形成多个数据库以及现代化的元数据管理法（见图 1–1）。

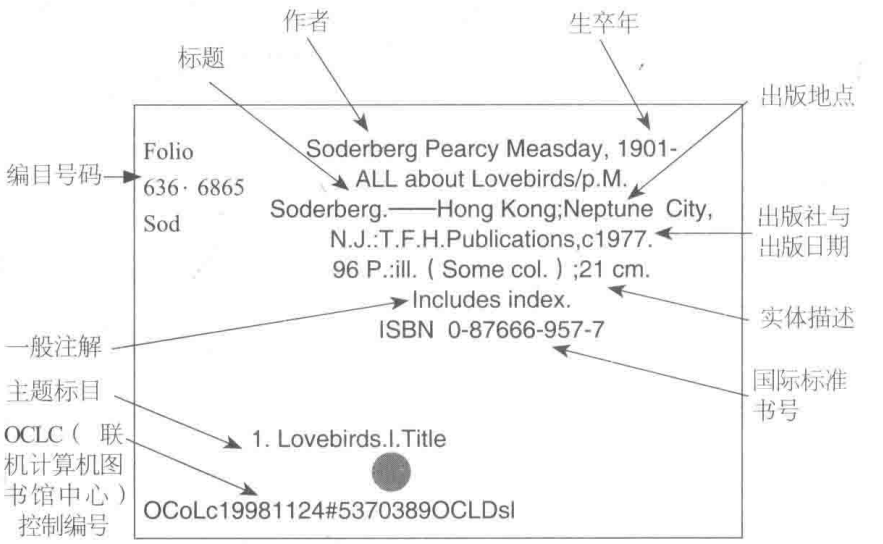

图 1–1

如果把一个数据集分解为记录，那么每条记录可以表达一个单独数据项，记录所包含的数据类别，其中多个数据项共用

一个类别，这实际上就是在创建电子表格。

想一想电子表格的布局：每行是单个对象的一条记录，而每列是这些对象的单独特性。现在，假设你要绘制一张电子表格，其中包含关于书籍的数据，各列的表头标题应该写什么？书名、作者、出版社、出版数据、主题、编目号码、页数、格式、维度等任何你可以想到的元素。接下来，每行则是单独一本书的记录，包含有关这本书的所有数据片段。这样的表格就可以作为图书馆目录（见表 1–1）。

表 1–1

书名	作者	出版日期	主题	编目号码	页数
《知识产权策略》	约翰·帕尔弗里	2012 年	知识产权——管理	HD53.P35 2012	172 页
《开放存取》	彼得·苏贝尔	2012 年	开放存取出版	Z286.O63 S83 2012	242 页
《数字文化中的模因》	利默·西弗曼	2014 年	社会进化、模因、文化传播、互联网——社会层面、模因学	HM626.S55 2014	200 页

既然你已经拥有了对象本身，为什么还要保存有关对象的数据呢？

科学家、哲学家阿尔弗雷德·科日布斯基（Alfred Korzybski）最为脍炙人口的名言也许就是“地图非疆域”（The map is not the territory），但人们往往认为这是马歇尔·麦克卢