“所有大型项目如果都分解成小块，完成起来将会更容易。”

4

砌砖匠眼中的信息架构

学习如何自底向上设计架构

错，Web 上确实有太多的东西。甚至仅仅在你的网站上内容就不少！是的，内容确实需要加以组织。不过怎样组织呢？如果你像我们一样，就会感觉自己好像在面对一次大扫除，而你能做的只是呆坐在长椅上，盯着眼前的一片混乱不知如何下手。

所有大型项目如果能分解成小块，完成起来就会更容易。盖房子看起来简直是不可能完成的任务（特别是对于那些连大扫除都无法应付的人），而为内容设计一个架构也同样是困难重重。不过，与盖房子类似，你要一步一步来完成。首先要收集材料和工具。接下来打地基，再砌砖抹灰。随后挂上窗帘，摆上椅子。

4.1 获得元数据

元数据是指信息的信息。尽管听上去有些抽象，不过这对于信息架构（IA）确实是一个非常实用的工具。元数据是所有组织系统的基础，从搜索到购物网站上的分面导航系统[①]都依赖于元数据。这就像是 IA 房屋的砖瓦，它可以根据你的需要“摆放”成各种各样的检索系统。信息可以有多种不同形式，可以是一篇文章、一本电子书（e-book）、一张照片或者是一个目录。有些信息没有文字，如 Flash 影片、MP3 格式的声音或者照片。如果信息中固有的文字很少（比如照片和音乐就是如此），那么元数据将有助于这些信息的查找。

① 分面（facet）是指按某个方面来分类，如商店按规模、按价格、按品牌等分类，更多内容见后面的介绍！

元数据是一种很有效的方法，用以确保以上各种形式的内容确实都能被查找到。元数据就是关于每一项内容的所有信息。例如，对于一首歌曲，元数据可能包括：“Brown Sugar，第 2 版，花絮，作词作曲：Mick Jagger 和 Keith Richards，演唱者：The Rolling Stones（滚石乐队），唱片：Itchy Fingers，bootleg，长度：3 分 50 秒分类：摇滚乐，蓝调布鲁斯”……

如今常用的 3 类主要元数据包括以下几种。

- **固有性元数据**。与事物构成有关的元数据。这是一个 MS Word 文档、JPEG、20Kb 大小的文件，还是一个 zip 文件？
- **管理性元数据**。与事物处理方式有关的元数据。这是临时的，还是需要归档保存的？编辑是谁？已经获批发表了吗？
- **描述性元数据**。与事件本质有关的元数据。对于我们的特定用途来说，这是最重要的一类元数据，也是 Web 上最常用的元数据。这是假想的还是事实？这是一篇文章吗？主题是什么？相关主题是什么？

“酷狗”的元数据

固有性元数据：
20KB
JPEG

管理性元数据：
摄影师：Noel Franus
用途：圣诞卡

描述性元数据：狗，小狗，犬科，金毛拉布拉多猎犬，金毛拉布拉多犬，圣诞帽，圣诞老人，圣诞节，圣诞，照片，诺埃尔的狗，可爱，伤感，让人想抱的

元数据并不是总能清晰地划分到这 3 类中，请看诺埃尔的圣诞卡（印着一只酷狗）。“圣诞卡”放在这 3 类中的任何一类下面都是可以的：

- **固有性元数据**。因为这说明了这个物品是什么。
- **管理性元数据**。因为这说明了它的用途是什么。
- **描述性元数据**。因为可以这样来描述这个物品。

在这种情况下，你可能想把“圣诞卡”放在所有这 3 类下面。

如果开发过网站，也许你已经遇到过 HTML meta 标记形式的元数据。可以看看 Dean and DeLuca 网站的源代码（HTML 代码就在源代码中），它在 meta 标记中给出了以下描述性元数据：

```
<meta name="description" content="Dean and DeLuca gourmet food stores.
Offering a wide selection of California wines, custom gift baskets, cakes,
cheeses, hard to find spices, coffee, caviar, truffles, holiday and seasonal
foods." />
<meta name="keywords" content="dean; deluca; gift; gourmet; food; online;
store; caviar; cheese; steak; coffee; holiday; artisan cheeses; artisan
cheese; spices; california; napa valley; baskets; corporate sales; olive oil;
vinegar; chocolate; seafood; shellfish; wine; herbs; cooks tools; cookware;
cake; cakes; wines; cookies; pies; truffles; seasonal; bakery; salmon; shrimp;
lobster; gifts; balsamic" />
```

在 Kansas City Steaks 网站上，可以看到除了描述和关键字外还有一行：

```
<meta name="developer" content="Digital Evolution Group, LLC">
```

在 HomeBistro.com 上，可以看到以下管理性元数据：

```
<meta name="ROBOTS" content="ALL">
<meta name="revisit" content="15 days">
<meta name="robots" content="index,follow">
```

隐藏在源代码中的元数据主要用于搜索引擎。Dean and DeLuca 网站告诉扫描到其网页的搜索引擎，这是一个出售食品的网站。Home Bistro 则请搜索引擎每 15 天回来查看一次是否有新内容。不过，这些任务都相当麻烦。下面来加以分解。

美国纽约公共图书馆（New York Public Library，NYPL）收藏了自发明照相机以来拍摄的各种照片。他们存储的图像简直数不胜数，存储空间达 57TB。假设你记得在 NYPL 网站上看到过一张照片，当时你特别喜欢，现在想再看一次。元数据将帮助你在这个照片的海洋里找到你想看的照片。这张照片（如**图 4-1** 所示）目前包含固有性元数据，右键单击这张照片，查看其属性就可以看到这些元数据。你可以从属性窗口明确以下信息：

- 这是一个 JPEG 图片，是 Web 上流行的图片格式之一；
- 大小为 303.29KB，并不算大；
- 图片为 609×760 像素，大约就是一张纸的大小。

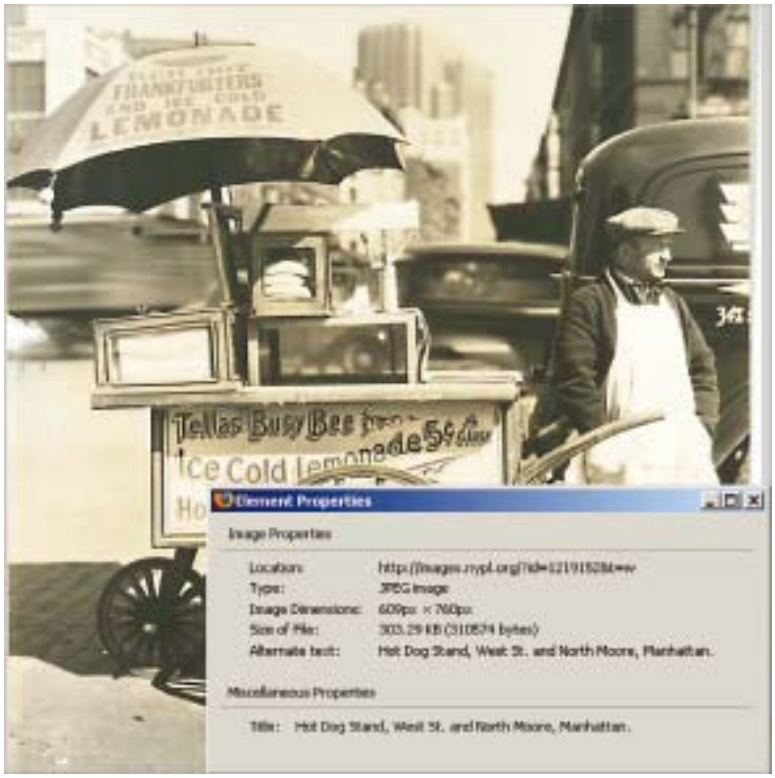


图 4-1　美国纽约公共图书馆收藏的一张照片及其元数据

不过，查看这些信息对于你再次找到这张图片确实有帮助吗？如果你原先就知道以后可能想找这张图片，就会把它的 ID 号或 URL 记下来。但是一般情况下并非如此，如果你原来并不知道将来可 能需要它，直到后来想要时才发现，又该怎么办呢？

你可能记得这张照片是一个名叫 Beatrice Abbott 的人拍摄的，或者你记得它拍摄于 1936 年。这就是管理性元数据，它不仅包含信息的作者或创造者，还包括创建日期、发布日期等，有关于事物 / 信息管理方式的所有信息都属于管理性元数据。但是（参见**图 4-2**）你到底能从这张照片 k 记住多少信息？这是一个人在纽约大街卖热狗吗？ 这就是第三种元数据：描述性元数据。这可能是搜索和浏览时最重要的信息，因为我们是人，不是机器，我们往往会记住作为人所感兴趣的东西，比如故事和影像。

图 4-2 “卖热狗的人”（West St. North Moore, Manhattan）有关属性

1. 讲讲“可查找性”的故事

在历史课上，老师总要求我们记住日期和地点，但是这些往往都是我们记不住的，我们能记住的通常是所听到的故事。对于大多数人来说，我们记不住拿破仑是出生在 1769 年 8 月 15 日（出生日期）的法国统治者，也记不住他是一个身高 159cm 的皇帝。我们所能记住的是，他是一个胳膊上搭着上衣，头戴斜边帽，在征服欧洲大片地盘的同时还在不断向约瑟芬写情书的那个人。那些铁的事实对我们来说黯淡无光，倒是这些充满罗曼蒂克的细节被我们牢牢记

住。可以利用人类的这一弱点帮助我们改善“可查找性”。可查找性（Findability）是由 Peter Morville[①]创造并普及的一个术语。这是指一个对象能够通过搜索或浏览而被找到的能力。我们很欣赏这个术语，因为由此可以清楚地看出：要把责任交给要被找到的对象，而不是把负担压在试图建立有效搜索查询的用户身上。在 Web 上，用户都希望减少麻烦。最近对 Yahoo! 和 Google 上最热门搜索的调查表明，80% 的搜索都是单字或双字查询[②]。在某种程度上，这样的一个词或两个词必须足以找出用户查找的对象，而有效的元数据就是一种很好的方法，可以将这个词延伸得更……远。

创建一组描述性元数据时，要提取出人们关于这个对象所讲述的故事。这些才是人们能记住的细节。在第 3 章中完成卡片分拣时，你会听到人们这样谈论他们分拣的东西：“Sarah 姨妈的苹果脆总是酥脆之极，真不知道她是怎么做的。”现在该用这些故事来选择有效的元数据了。

要找到“纽约大街上卖热狗的人”那张照片，用户可以搜索 1219152，这是这幅图片的 ID 号。但是用户一般不会这么做。在当今这个充斥着各种数字的世界里，我们要记住电话号码、ATM 码和密码，而这又是一个数字需要我们去记。我们的大脑可能没有足够的空间来存放更多随机数据。用户可能会搜索“Bernice Abbott”（摄影师）或者搜索“Changing New York”（这是这个系列的名字）。这些作为搜索项的可能性更大，而且名字比数字更容易记。不过，用户可能会得到很多页的结果，如**图 4-3** 所示。想想看要翻阅 29 页的结果！用户更有可能搜索“Hot Dog Vendor”（热狗小贩）。这是描述性信息，它取自于照片所讲的故事，也是照片故事所固有的主题。

这些照片创建时并没有故事。摄影师可能会给出一个标题或题目，但是这对于照片的实质内容只是一个不充分的线索。不过，既然图书馆的任务是选择、收集、保存并允许人们访问“这个世界所累积的精神财富，而不考虑物质财富、信仰、国籍或其他人文条件的差别”，所以图书馆要确保任何人只要加入早先在属性中看到的某个文本就能找到这张照片。主题包括“food vendors”（食品小贩）和“lower west side”（曼哈顿西区底层社会），另外还有一个说明“Vendor stands next to his Tellas Busy Bee cart, advertising ‘Red Hot Frankfurters and Ice Cold Lemonade’ traffic a blur in the background.”（小贩站在兜售车旁，背景上模糊印有广告‘Red Hot Frankfurters and Ice Cold Lemonade’）。这些元素构成了一个完整的故事，将与用户搜索时讲的故事匹配（参见**图 4-4**）。

---

① *Information Architecture for the World Wide Web*（O’Reilly, 1998）的作者，他还是一家信息架构咨询公司 Semantic Studios 的 CEO。

② Google Zeitgeist（http://www.google.com/press/zeitgeist.html）和 Yahoo!’s Buzz（http://buzz.yahoo.com）都很有意思，可以通过这两个网站了解在某个给定时刻人们在搜索什么。

图 4-3 搜索 Beatrice Abbott，会得到图书馆收藏的 Beatrice Abbott 的所有照片

纽约公共图书馆只是要履行使命，而盈利性网站可不能止步于此。更多情况下文本搜索可能无法成功地找到所要的结果，所以这些盈利性网站不能冒这个险。如果图书馆未能返回搜索结果，我们可能只是对图书馆感到失望，但是如果一个盈利性网站有这种欠佳表现，这个网站的末日也就到了。

图 4-4　成功的搜索会找到标题以及说明域中的文本

2. 利用手工元数据尽享查找的快乐

iStockphoto 网站拥有成百上千张珍藏照片，它充分利用了手工元数据（handcrafted metadata）。他们的业务模型依赖于查找照片的用户有足够强烈的愿望付费购买。如果用户没能找到照片，这就意味着他们的公司无法盈利。

在每一张照片的右下角，iStockphoto 都会显示一个关键字列表，分别链接到所有标有相同关键字的照片（参见图 4-5）。如果想查找以下内容，你可以按如下方式搜索：

- 另一张关于家庭的图像（描述性）；
- 这个摄影师拍摄的另外一张照片（管理性）；
- 另一张具有相同配色的照片（固有性）。

图 4-5 幸福家庭，由 Jpmediainc 拍摄

每一张照片都会由某个人来查看，他会仔细考虑这个图像并选择关键字（元数据），这些元数据可能会由设计人员用于搜索过程。看这张照片的信息架构师会这样考虑：“一个家庭，坐在草坪上，他们很幸福，这可能正是吸引人的地方”，如此等等。信息架构师自己会想出十几个关于这个图片的小故事，选出其中最震撼、最有可能被搜索的故事作为搜索项来提高照片的可查找性。然后信息架构师可以使用这些搜索项创建更有效的搜索，不仅如此，还可以创建更便于浏览的结构。如果一位有创意的导演想找一张优秀的图片完成一个设计，但对这张照片不满意，也可以非常容易地按照其他关键字查找（其中可能涵盖他要查找的东西）。只需轻轻单击就可以轻松获得更多照片（参见**图 4-6**、**图 4-7** 和**图 4-8**）。

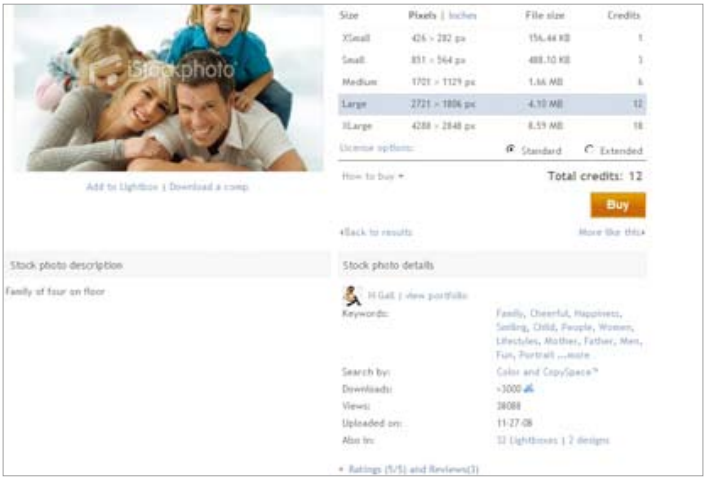


图 4-6 另一张关于家庭的图片

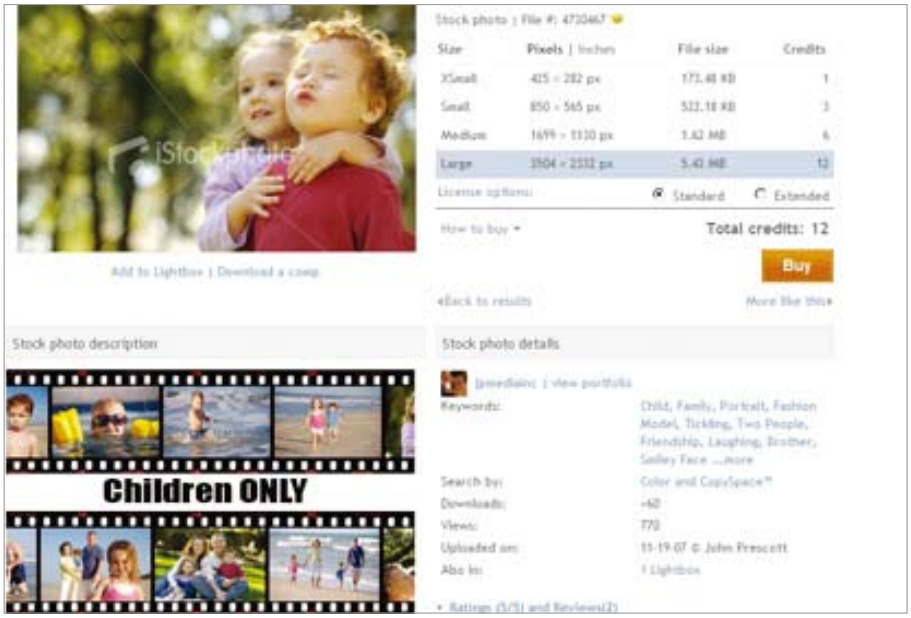


图 4-7　Jpmediainc 拍摄的另一张照片

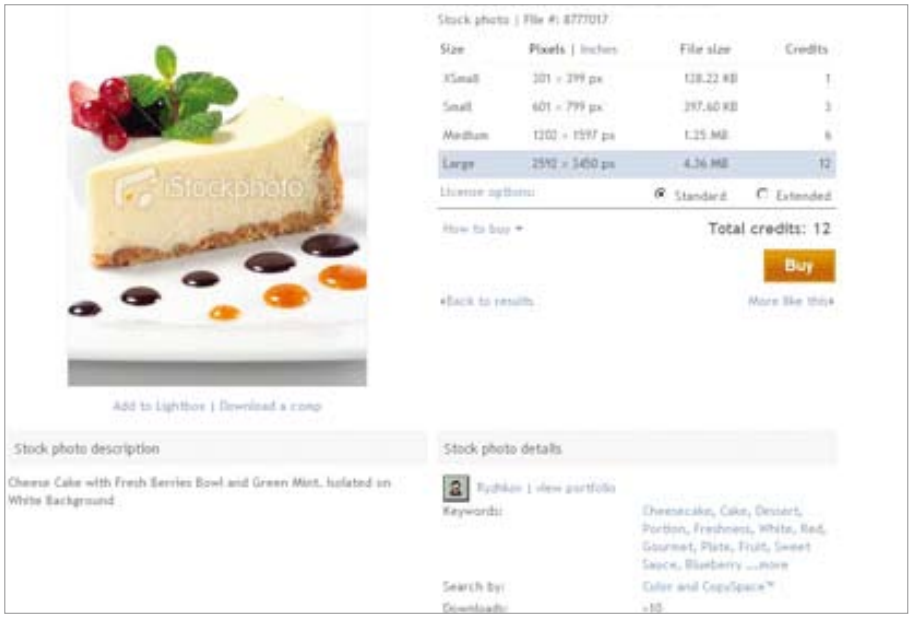


图 4-8　另一张有相同配色（调色板）的照片

如果手工地增加了元数据，下表（**表 4-1**）中的各项将更有可能被查找者找到。

表 4-1 元数据类型示例

对 象	描 述	可能的关键字	描述性元数据	固有性元数据	管理性元数据
 照片：Santorini 教堂	Santorini（位于 Fira 城）希腊岛上的一座教堂	教堂、小教堂、小路、希腊、希腊的、Santorini、悬崖、大海、海洋、蓝色、白色、地中海	文件类型：JPEG 图像文件 比例：1600×1200 px 文件大小：1814KB	创建日期：4/20/2002 修改日期：4/21/2002	摄影师：Christina Wodtke 用途：Greek picture book 2002
原创歌曲：“Ain’t nobody here but us chickens”	随兴而作的即兴歌曲。闲散风格的“Ain’t nobody here but us chickens”	爵士乐、闲散风格、男性歌手、Mark Murphy、顽皮的、即兴而作、休闲音乐、幽默, 糟糕、娱乐、聚会音乐	文件类型：MP3 文件大小：2234KB	节奏：慢板	演唱者：Mark Murphy 创作者：（A.Kramer / J.Whitney） 制作：David Bram 与 Mark Murphy 授权：32 Records 录制速度：96Kbps 用途：The Best of Jazz Juice
 产品：防滑条	就像为脚下安上了汽车轮胎的防滑链，这种便于安装的防滑条让穿着者在冰雪中也能大步前进	鞋配件、皮靴配件、防寒装备、防雪装备、防滑、滑、冰、雪，防雪鞋，防雪靴、轮胎防滑链、金属、橡胶、交织	制造商：Hammacher Schlemmer	库存：有货	出品：2002

（续）

对　象	描　述	可能的关键字	描述性元数据	固有性元数据	管理性元数据
Adobe InDesign 教程	使一个 Adobe Photoshop 文件的透明度不变，为对象和文本应用阴影，混合向量图片和位图图片之间的颜色来达到有趣的效果	Adobe、InDesign、透明、阴影、文本、向量处理、教程、学习、电子教程、在线学习、flash 影片	Flash 教程，布局中使用透明效果	Flash 影片 Windows: 2.6MB Flash 影片 Macintosh: 3.6MB	支持 : InDesign 2.0 停止使用 : 2003 年秋季
网站 : COMMON GROUND: 一种面向人机界面设计的模式语言	研究报告，解释如何使用模式来设计交互式系统	设计、网站设计、交互设计、交互式、模式语言、Christopher、Alexander、内容表示、导航设计、HTML	283KB	www.mit.edu/~jtidwell/common_ground.html	作者 : Jenifer Tidwell 最后修改日期 : May 17, 1999 版权 : 1999

很多东西都需要有意识地增加手工元数据，如动画和电影、目录中的产品、新闻栏目以及杂志文章和研究论文。“不应该包括文章”，你可能会这样说，“它们本身就是文本，难道不能直接搜索文本吗？”

下面来看一个小说专栏作家的简短专栏报道 :“Bonds 今天又打出了另一个本垒打。巨人队的球迷们简直疯狂了！大家都为他的破记录之年而欢欣鼓舞。Barry 这个赛季打出了 0.863，看来球迷们势必要为这一届世界职业棒球锦标赛之旅做准备了。”

假设你想搜索出 San Francisco 队最近有没有打出过本垒打，可能找不出这篇文章。真糟糕，你可能甚至无法看出这是否有关于棒球！但是如果增加一些描述性元数据：“San Francisco 巨人队、本垒打、世界职业棒球锦标赛、Barry Bonds 和棒球”。现在搜索时就很可能会在结果中给出这个专栏报道。

3. 都使用同一种语言

还有一种方法可以让搜索更有效，即创建一个受控词汇表。几个月以前，Christina 在做餐厅服务员。一天她的经理通知所有服务员从现在起都要把顾客称为“客人”。还要把上菜顺序称为“第一道菜”（fi rst course）和“第二道菜”（second course）。他们的厨师是一位法国人，他发现美国用“entrée”一词表示主菜，这让他很不舒服。这是 Christina 第一次接触到受控词汇表。

英语是一种复杂、灵活而且功能强大的语言。Steve Martin 曾说过，“天哪，看看那些法国人，每一样东西都分别对应一个不同的词！”但实际上，英语中更是布满了陷阱。正餐之前你可能会有一些开胃菜，而开胃菜有各种各样不同的说法。

- starter。
- first course。
- appetizer。

或者是从其他语言借用的词汇：

- Hors d’ oeuvres。
- Anamuse-gueule[①]。

另外，一家西餐馆可能把第一道菜称作“grazing”，运动主题酒吧则会把这称为“warmups”。可以看到，这就会带来混乱。在餐厅里，如果 Christina 询问客人是否需要开胃

① 在法国，这就是“正餐前少得可怜的一点点东西”。比如说四分之一汤匙的鱼子酱再加上放在一个小白盘里的两片油煎面包块。

菜（first course），他们可能会认为她很滑稽并反问一句“嗯？”，这样一来，她可能会接着说，“appetizer? Hors d’oeuvres? 还是 nibble?”但是在 Web 上，即使你大声叫也没有人能听到你的解释，所以我们认识到需要创建一个受控词汇表。

4. 受控词汇表

顾名思义，受控词汇表（controlled vocabulary）就是一种控制所用词汇含义并跟踪相关词的方法。在 Christina 的餐厅，他们选用的优选术语是“first course”，而客人们可能用到的词包括“starter、fi rst course、hors d’oeuvres 和 appetizer”，可谓多种多样。所以，如果一位客人想要一份熏鲑鱼作开胃菜，服务员就会在账单上写上“first course: 熏鲑鱼”。餐厅还可能跟踪相关的概念：“夫人，想来点开胃酒（aperitif）吗？”，或者更随意的，“您看菜单时介意给您先上点饮料吗？”

就像那位法国厨师一样，计算机往往很不灵活。假设你在考虑为早午餐做一熏份鲑鱼（cured salmon）。如果你搜索“鲑鱼（salmon）”，计算机就会给出提到鲑鱼一词的所有结果，也许你能从中发现所要找的东西。但是如果你输入“鱼（fish）”或者“鲑鱼（gravlax）”，等你找到结果并为客人上菜时，客人们可能早已经饥肠辘辘了，除非搜索设计者创建了某种类型的受控词汇表。受控词汇表有多种不同类型，可能是由等价关系构成的简单词汇表，“对，gravlax 和 cured salmon 是一样的”，也可能是一个复杂的辞典，“gravlax 是一种鲑鱼，等同于 cured salmon，这是百吉圈和熏鲑鱼的一种配料”。接下来让我们更深入地讨论这个问题。

5. 等价关系

最简单的受控词汇表是一组等价关系：cured salmon 和 gravlax 对于搜索来说含义相同。**表 4-2** 给出了一个例子。这种关系可以很简单，如表示同一个事物的两个词：cat 和 kittycat（都表示猫）。它们是同义词，也可以是表示同一个事物的不同拼法或缩写。Lion 与 lyon 是一样的（狮子）；SPCA 代表 Society for Prevention of Cruelty to Animals（动物保护协会），这些都认为是变体。单词可能稍有差别，但是对于搜索来说，可以选择 cat 和 kitten 中任何一个，结果都是一样的。你可能有一个贺年卡网站，有人想要一张印有小猫（kitten）图片的贺年卡，但是你只有一张带 cat 画面的卡片。最好是把这张有 cat 的贺年卡提供给用户，而不要简单地显示“未找到任何结果”。

这非常类似于书最后的索引。你在一本关于太阳系的书中查找“moon”，它指出“参见 satellites”。对这本书而言，satellite 和 moon 都是一样的（表示卫星）。另外一本书（可能更厚）可能会对二者加以区分。关键是要考虑哪些人搜索，他们会用哪些词搜索，然后再为他们提供你的内容。

表 4-2 等价关系示例

优选术语	变 体
Smoked salmon（熏鲑鱼）	Fish, gravlax, lox, cured salmon, smoked fish, preserved fish, nova

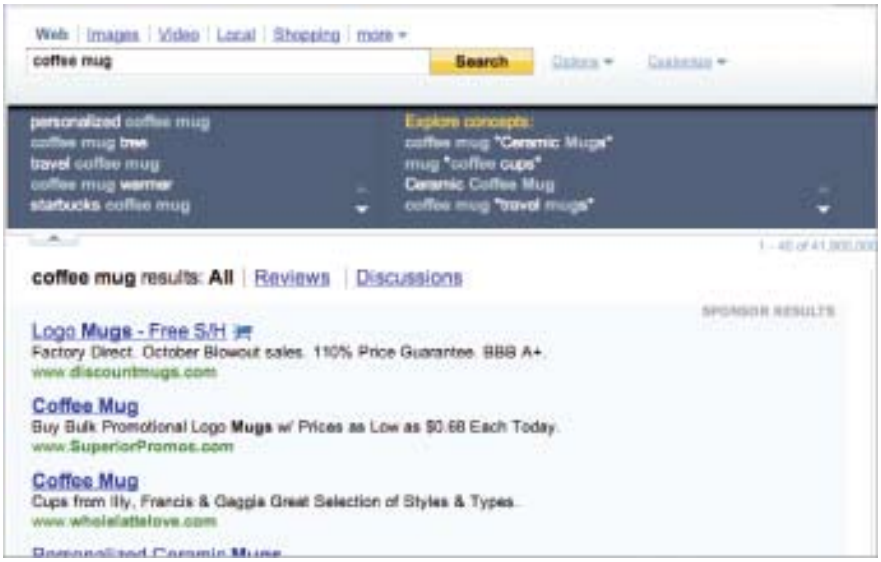
6. 层次关系

分类系统是一种更复杂的受控词汇表。除了等价关系外，它还显示了层次关系。这不仅对搜索很有用，还有助于建立有效的浏览层次体系结构以及将二者结合起来。**表 4-3** 给出了一个例子。

表 4-3 层次关系示例

优选术语	变 体	父词（广义词）	子词（狭义词）
Smoked salmon	Gravlox, lox, cured salmon	Fish, smoked fish, cured meats, preserved fish	Smoked salmon flatbread with crème fraise, linguini with smoked salmon and asparagus

Yahoo! 中可以看到一个实际的分类系统。如果搜索“coffee mug”（马克杯），将得到相当多的结果（参见图 4-9）。



www.yahoo.com

图 4-9 Yahoo! 上搜索“coffee mug”的结果

可以更仔细地查看这些结果。每个结果不仅包括标题、描述和链接，还提供了对箸名的 Yahoo! 层次体系的一个链接。如果搜索者要查找 tchotchke（小摆设品）来印上公司的 logo，可以单击 Promotional Items > Mugs 并找到提供这项服务的公司，或者找到一个马克杯收藏家欣赏其他收藏品。这种分类还可以为搜索者提供上下文。例如，如果马克杯收藏家注意到第二个结果归入 Punk and Hardcore Artists 一类，也许就不会再单击这个结果了。

7. 关联关系

受控词汇表中的泰姬陵[①]就是分类辞典。你可能还记得上小学时使用分类辞典的日子。使用分类辞典能让你看起来更明智。并非简单地写作“她说（said）”，利用分类辞典，你可以写为“她大声叫嚷（yelled），她讲（spoke），她低声耳语（whispered），她旁敲侧击地暗示（insinuated），她明白地说（articulated），她说出（uttered），她坚持说（insisted）”，等等[②]。

借助 Web 的力量，分类辞典又回到了我们的日常生活中。它不仅仅是一个能提供更多、更好词汇的工具，分类辞典还可以用来建立相互关联的词汇网络，帮助人们找到他们原先没有的东西。分类辞典不只能显示层次关系，还展示了关联关系。

如**表 4-4** 所示，将元数据组织到一个受控词汇表中时存在一定的主观性。在不同的网站上，Jewish cuisine 可能作为父词，而 preserved fish 作为关联词。这取决于网站的类型以及网站的访问者。

表 4-4　分类辞典雏形

优选术语	变　体	相关词	父　词	兄弟词	子　词	关联词
Smoked salmon	Gravlax, lox, Cured salmon	Preserved fish	Smoke trout, bacalao, salt-cured sardines, pickled anchovies	Smokes salmon flatbread with crème fraise, linguini with smoked salmon and asparagus	Jewish cuisine, kosher foods	Crème fraise, bagels, capers, dill, crackers, fish knife, caviar

关联词就是同属一个范畴但却并不相同的一些词，而且这些词也不是更广义或更狭义的词。它们只是同在一起而已。例如，如果表 4-4 是一个食谱网站的分类辞典，给出主要词的同时还

① 印度知名度最高的古迹之一，被誉为“完美建筑”，有极高的艺术价值。——译者注

② 当然，后来我们又学到谈话时要惜字如金，只需简单的“说”（said）就可以了。这样一来，分类辞典只能沦为压书的镇纸。

列出配料通常会很有用（crème fraise、bagels、capers、dill 和 cream cheese）。在一个美食家食品商店网站上，列出顾客可能想买的其他物品可能很有帮助（crackers、fish knife 和 caviar）。这些都是与 smoked salmon 相关的词，但是没有人会把它们混淆成同一个事物。所有这些受控词汇表都是为了让人们得到他们寻找的东西（不论在搜索框里的输入是多么千奇百怪）。下面来具体看一看。

8. 每个人有不同的拼法

没错，我们当中一些人肯定有不同的拼法。

图 4-10 和图 4-11 显示了最近尝试查找不同类型 chedder 的结果。根据这个搜索可以看出，Zabar’s 网站没有 chedder。除非搜索者把拼法改为正确的 cheddar，并要求你使用这种拼法，才能显示出信息。不过 Yahoo! 则不同，它认识到人们可能会创造各种各样的拼法，尽管“chedder”也能查出些结果，但他们还建议你试试“cheddar”。

图 4-10 在 Zabar’s 网站上搜索 chedder

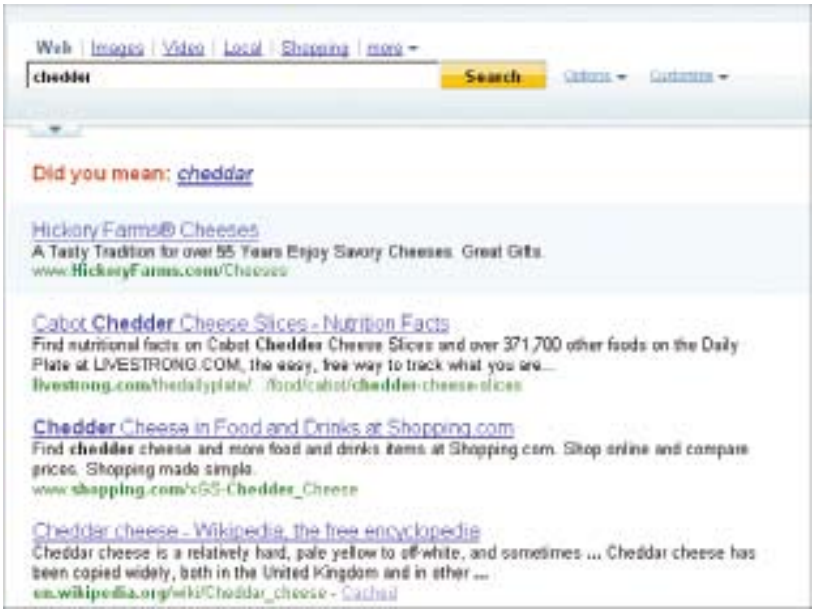


图 4-11　在 Yahoo! 网站上搜索 chedder

下面尝试对 Zabar’s 网站完成逆向工程[①]。网站不是我们构建的，而且我们也不认识建网站的人，但是通过“摆弄”这个网站，可以准确地猜测出它是怎样工作的。所以，如果我们认为 Zabar’s 会出售英国干酪，可以试着搜索“cheese”。这会返回相当多的奶酪产品，其中就包括英国干酪（cheddar）。Zabar’s 的受控词汇表包括一些层次信息，指出 cheddar 是多个父词的子集：“Semi-firm cheese”、“English Cheese”，甚至有自己的一个特殊“cheddar”子集放在“All cheeses A-Z”下面（参见**图 4-12**）。

如果网站像这样有多个父词，这称为分面分类（faceted classification）。分面可以包括多个项共有的任何性质，包括价格、重量和颜色。在这里，分面可以是品牌、产地和硬度。我们很奇怪他们居然没有包括口味的轻重，不过我们并不是在为 Zabar’s 工作（所以不必为他们考虑这种问题）。用户想缩小商品选择范围来找到最满意的商品时，分面就很有用。商业网站就经常使用分面分类[②]。

① 技术人员或者五岁小孩子想搞清楚一件东西的原理时，他们就会把东西拆开来。如果一个五岁小孩子这样做，会被认为是搞破坏。而技术人员这样做时，往往会学着重新装配一遍，然后再造出自己的产品，可能这个产品比原来的更好，这称为逆向工程（reverse engineering）。这是一种了解事物原理的好方法：先研究产品，把它拆开，重新组装，再尝试建造自己的产品。

② 在 Christina 的杂志 Boxes and Arrows 上可以了解更多有关分面分类的知识：
http://www.boxesandarrows.com/view/ranganathan_for_ias。
http://www.boxesandarrows.com/view/all_about_facets_controlled_vocabularies。

图 4-12 绝妙的奶酪世界

如果继续查看 Keen’s Farmhouse Cheddar 页面，可以看到这里使用了分类辞典来引诱购买者买下更多商品（参见**图 4-13**）。检查 You May Also Like 部分，你可能会猜出 Keen’s Farmhouse 的父词——English cheese，尽管这里没有列出这个元数据。English cheese 在相关项选择中有丰富的含义，包括：

- 有相同父品牌的兄弟词（Colston Bassett Stilton），这也由 Neal’s Yard 出品；
- 有相同父产地的兄弟词（Shopshire Blue），这是一种非干酪型英国奶酪，有点臭但很好吃；
- 一个关联项 Pumpernickel（裸麦粉粗面包），这并不是奶酪，但是可以让食物吃起来更美味。

如果访问者要找一种好的英国干酪，不仅要能让他们找到想要的东西，最好还能让他们不由自主地被引诱买下原来不知道自己想要的一些东西。分类辞典认为他们可能会这样做（如**表 4-5** 所示），这样一来，Zabar’s 就要赚大钱了。只要能理解这个小小的拼写问题，他们的网站就会相当完美。

图 4-13　相关和关联产品

表 4-5　cheddar 的辞条

优选术语	变　体	相关词	关联词	父词（可能多个）	子　词	兄弟词
Cheddar	chedar, Chedder, cheder	English Cheese, Semi-Firm Cheese	Keen’s Farmhouse Cheddar	Colson Basset	Stilton, Shopshire Blue, Cabbott’s Extra Sharp Vintage Cheddar	Pumpernickle

9. 建立一个受控词汇表

产品大卖了！答案就在于受控词汇表。嗯，那么受控词汇表从哪里来呢？与很多有意义的东西一样，建立一个受控词汇表很费时间。在这里我们可以给出一些基本步骤，不过也许你希望再买一本更厚的书来深入地学习。

(1) 收集内容

你的第一个问题应当是“我想要组织的到底是什么？”我们发现，对此最有效的方法就是建立一个内容目录。内容目录（content inventory）是对网站上现存的所有东西以及你希望网站能够增加的所有东西的一个记录。

假设你在创建一个 MP3 网站，希望统计目前网站上的所有 MP3，以及所有音乐评论、艺术家访谈和可以从网站下载的其他支持材料（如 MP3 播放器信息）。你可能还想知道接下来会做什么。也许这个网站还计划增加音乐视频的 MPEG 文件，以便用户观看。不过这可能安排到第二年才具体实施，没有人明确知道这到底是怎样的。你想指出将会有音乐视频，但是无法组织你并不清楚的东西。增加新内容时很有可能会出现重复工作。

如果你有时间，而且想利用这个机会挑选一些内容，可能还希望完成一个内容审计，这样一来不仅要统计每一个内容，还必须根据某些准则对各个内容做出评价，如冗余度、时效性和有效性。完成内容审计后，你就能全面地了解目前有些什么，将会有什么以及哪些内容真正具有价值。

(2) 从尽可能多的来源收集词汇

现在要得出元数据了。如果内容中包含有词语，可以先从内容本身入手，挑出当前主题独有的术语。还可以查看现有的分类辞典。当今世界已经有相当多这样的分类辞典。这些辞典并不总能“原封不动”地直接使用，因为每个分类辞典都是为某个特定的用途而设计的，不过它们往往有助于更好地理解你要描述的领域，并帮助你找出相关术语。

一些可以借用的分类辞典！

Getty Art and Architecture Thesaurus（http://www.getty.edu/research/tools/vocabulary/aat/index.html）——与艺术、建筑和物质文化有关的概念的一个可搜索数据库和研究工具。包括风格和时代、媒介、建筑工作、材料和技术，等等。

The Astronomy Thesaurus（http://www.mso.anu.edu.au/research/library/thesaurus/）——英语、法语、德语、西班牙语和意大利语的天文学术语，所有词汇都可以交叉引用。

Legislative Indexing Vocabulary (LIV)（http://www.loc.gov/lexico/servlet/lexico?usr=pub&op=sessioncheck&db=LIV）——专为法律和国家政策相关主题而开发。

Thesaurus for Graphic Materials I: Subject Terms (TGM I)（http://www.loc.gov/rr/print/tgm1/）——包括索引可视化材料的有关术语和交叉引用。

Maths Thesaurus（http://thesaurus.maths.org/）——常用数学术语的定义，以及广义、狭义和相关概念。

NASA Thesaurus（http://www.sti.nasa.gov/thesfrm1.htm）——美国国家航空航天局（National Aeronautics and Space Administration）术语表，广泛涵盖了相关的科学术语。

National Monuments Record Thesauri（http://thesaurus.english-heritage.org.uk/）——English Heritage 提供的文化遗产数据，提供了大量分类辞典的层次和字母列表。

你可以与主题问题专家进行交谈，也可以利用卡片分拣得出搜索者考虑术语的方式。如果你希望得出关键概念，并把有关术语分组在一起——对此，单独的卡片分拣就是一个很有效的方法。这些关键概念或“主题词”应当包括从内容目录中收集的所有重要概念的同义词与缩写、首字母缩写以及可选的拼写方法。**表 4-6** 给出了一个示例。

表 4-6　主题词示例

优选术语	同 义词	缩　　写	首字母缩写	候选拼法
Rock music	Rock and Roll	Rock	R&R	Rawk

(3) 定义优选术语

优选术语（preferred terms）是一种在内部控制词汇表并保证所有人都能达成共识的工具，同时它也是一种了解标记过程的方法。在**表 4-4** 中，所有选术语都可以作为优选术语。选择其中之一时，首先应当考虑用户。如果 MP3 网站专注于 20 世纪 50 年代的音乐，使用全称“Rock and Roll”就很好。如果是 Jake’s Rawking Out 网站，“Rawk”可能最合适。“Rock music”和“Rock”之间的区别可以忽略不计。如果要从中做出选择，最安全的做法就是选择最不容易含混的选术语，或者是最适合放在导航条中的选术语。

做这些决策时，一定要注意根据你的选择而制定的规则。分类系统是一个不断发展变化的事物，随着新内容的增加，你往往希望保证它与其余内容有一致的组织结构。

(4) 链接同义词和近义词

好好梳理你的术语。将之前没有连接起来的所有关系链接起来。找出常见的拼法错误（对此的研究报告会很有帮助）。还可能要做一些艰难的抉择。（World Music 和 Global Beats 真的是两个不同的概念吗？）把优选术语归入核心概念集合中。

(5) 按主题对优选术语分组

卡片分拣！你首先会看到信息架构师比拉斯维加斯赌城的发牌人洗牌次数还要多。现在要抽出那些优选术语，组织到同类的组中。Rock、Hip-Hop、Rap 和 Techno 都是一类，而 Jazz、Bebop 和 Fusion 则同属另一类（至少从某种意义上是这样）。

这里也很适合让网站可能的最终用户完成一个卡片分拣，了解他们如何考虑各种分类。或者，如果你很有自信，认为根据用户先前的一次卡片分拣你已经很了解用户的心理，就可以继续用一个更完善的层次体系进行测试。测试往往事关时机、时间和金钱。这是肯定要发生的，问题在于它在这个过程中的哪个阶段对你的帮助最大。

不论怎样，都需要把这些优选术语放入相关的堆中。

(6) 找出广义术语和狭义术语

现在需要确定每个术语最适合放在层次结构的哪个位置。请看看你分出的各个堆，可能你认为 Hip-Hop、Rap 和 Techno 都是 Rock 的子集，也可能认为 Hip-Hop 是 Rap 的子集，再看 Hip-Hop 和 Techno，也许你认为它们都属于 Club Music。现在可以开始建立一个层次体系，甚

至你可能会发现存在多个可能的层次体系。正是这一点让人心生恐惧，重要的是需要后退一步，纵观全局。如果有一个分面分类系统会对你的用户很有好处，用户可以浏览多种不同类型的层次体系，如摇滚乐（Rock）、舞曲（Dance Music）、娱乐音乐（Upbeat Music）、按艺术家组织音乐（Music by Artist），等等。不过企业可能无法承受花费这么多时间，或者基础设施（运行网站的基础技术）可能无法处理如此繁多的信息。如果网站是 Virgin Records World Wide，采用分面也许是明智之举。但如果是 Jake’s Rawking Out 网站，这可能就不是合适的选择了。当然，在这两者之间还有很多其他层次。

按照艺术家分类并不是一个好方法，因为他们总是随着艺术灵感的灵光乍现从一个流派变到另一个流派。也许上一周还在做乡村音乐，这一周可能就成了摇滚明星！从**图 4-14** 可以看到 iTunes 是怎么处理的。它提供了一组不同类型的摇滚乐，并把艺术家放在合适的地方。所以 Allman Brothers 会同时出现在 Southern Rock 和 Best of the’70s 中，另外他们的名字还特别出现在 Legends 中。用户对于你的内容会有某种理解，要在用户的心理与内容的实质之间取得平衡，考虑到建立辞典所要花费的时间，并尽量选择最佳的折衷方案。最后你可能会建立一个基于音乐类型的网站，但是仍会链接艺术家的名字，使得 Allman Brothers 的歌迷总能找到他们的所有歌曲。然后你可能决定换换心情，采用一种基于事件的组织。或者你可以采用所有这些做法。

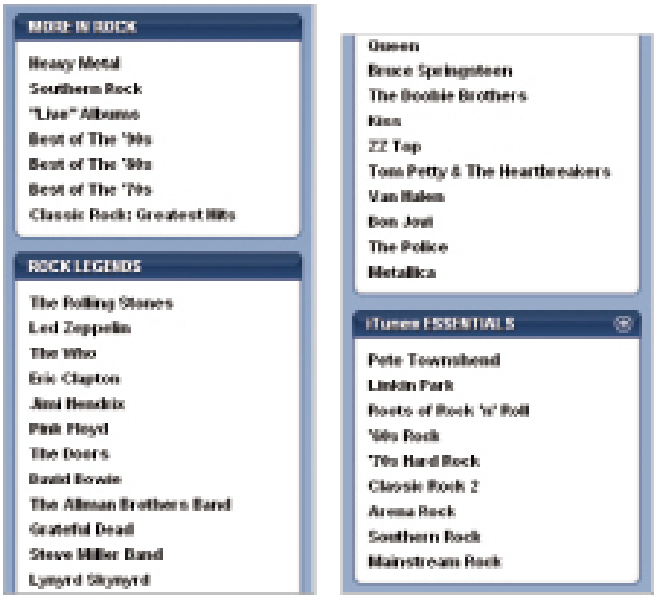


图 4-14　iTunes 的摇滚子音乐类型

(7) 完成关联链接

现在该洒蛋糕最上面的那层糖霜了。或者说，该在收款台为你买的糖结账了。对于各个优选术语，问问自己，“用户下一步可能想去哪里？”

不过，要有所克制——只选择最明显和最重要的关系：

- 奶酪可以链接饼干；
- Beck CD 可以链接音乐会门票；
- 锤子可以链接钉子；
- 驱动程序下载可以链接支持文档。

要充分利用你对于内容的关系、企业驱动力以及用户需求和任务行为的理解。仔细地设计用户下一步去的地方。

(8) 对选择及相应原因建立文档

这是最令人厌烦的一步，也是每个人都想跳过的一步。要当心，千万不要忽略这一步，请为你的后来者（或者是忘记了前车之鉴的你自己）考虑，应当以某种方式写下你已经做了什么，以便过后别人可以借助你的这些经验。

设计一个受控词汇表时，最好进度慢一些，考虑周全一些（很多事情都是如此）。先建立一个草稿，可以在此基础上改进，而不是草率上马（往往速度极快）。要尽量考虑到下一个版本。受控词汇表（就像一个网站，而不像一本书）是在不断变化的。

现在对于分类辞典你已经有了很好的基础。不过要记住，分类辞典堪称受控词汇表中的泰姬陵，有时你所需要的只是一个简单的候车棚而已。不一定非得完成上述的每一个步骤。可能有一个简单的网站，只需要为之定义同义词或者一个简单的分类系统。随着网站的扩展，受控词汇表也要随之扩展，总有一天你会发现需要一个全面的分类辞典。

10. 社交分类

信息架构主要关注于分类（你可能已经注意到了这一点），出现社交分类时存在很大争议。很多信息架构师甚至认为这种新方法（尽管麻烦但可扩展）的出现可能会让他们丢掉工作。不过，最后可以看到，如果没有做一些准备，从用户那里根本无法得到一个有用的分类体系，就好像把一堆石头交给一群村民让他们开工，希冀能得到一个教堂一样。需要有一种方法鼓励有意义的参与，然后还需要一些规则来约束如何参与以及如何创建一个可查找的系统。

11. 标签

标签就是公开的关键字。长期以来，图书馆管理员、科学家、技术人员和其他对顺序和检索感兴趣的人不仅会把东西分类存放，还会为之关联关键字，以便通过搜索可以找到。Delicious 是第一个允许人们为某个对象增加关键字的网站，而不论他们是否是这个对象的官方所有者。这些关键字被称为标签（tag）而不是普通意义上的关键字（key），这是因为，类似于纽约的装配工，你在对别人的东西加标志。Delicious 还做了另外一件别人没有做过的事情，它允许公开地选择对网站建书签。

默认为公开共享书签本身很简单，为什么这种简单的事情如此有用呢？ Delicious 从所有用户那里积累了最流行的网站和标签，这让他们的首页成为有关各种主题最新最酷内容的一个指南。如果你不关心当天首页上的主题，可以使用标签把范围缩小到你个人感兴趣的主题。

例如，如果你关心健康（health）和生产力（productivity），可以看到 280 个人认为这个 lifehacker 网站很有用。如果你更关心分析学（analytics），Delicious 会通过 short urls 推荐 cli.gs 文章。

标签还有另一个很有意义的功能。在这个世界中，用户生成的内容越来越普遍，标签已经成为创建可扩展分类系统的唯一途径。YouTube、Flickr、Slideshare 和其他网站都完全由用户提供的内容构成，它们别无选择，只能依赖于由用户提供的分类系统。

最后一点，标签从另一个方面也对最终用户很有用，即个性化组织。标签“toread”是 Delicious 上一个流行的选择。Flickr 上超过 4 000 张照片都有“toprint”标签。标签不仅有助于检索，还代表了缺乏的“保存”功能。其灵活的本质可以向网站指出用户真正想做什么。

一方面是灵活的可扩展分类系统，另一方面是聚集、保存甚至推荐新项目的标记功能，二者的结合使标签成为所有 Web 2.0 网站的一个选择，甚至一些早期的 Web 网站也尽力跟踪这种技术。综合传统的分类方法，标签会更加有用。

例如，Etsy 是一家供人们出售其个人手工制品的网站，它同时使用了标签和分类系统，从而能很容易地找到最棒的礼物。通过仔细查看，可以看出它没有你想像中那么随意。尽管描述性元数据放在“标签”（Tags）系列中，网站的设计者决定再为固有元数据增加另一个系列，称之为“材料”（materials）。这也是在加标签，因为这些关键字由网站的用户增加（而不是网站的管理者）。但是哪些标签能放在这个系列中？这要由系列名来约束。在**图 4-15** 中，可以看到幸运星的制作离不开纸（paper）、折纸方法（origami）和爱。

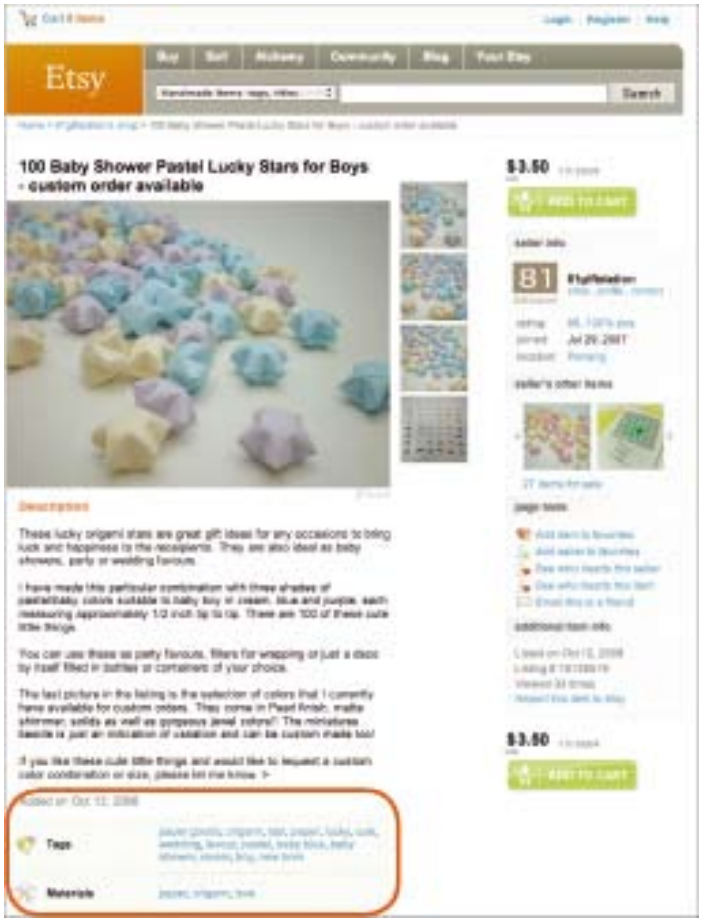


图 4-15 origami 彩色幸运星的标签给出了如何使用这些幸运星的建议

Buzzillions 对它们的 Review Snapshot 也做了同样的处理（参见**图 4-16**），这个网站创建了不同种类的标签，包括优点（pros），缺点（cons）和最佳用途（best uses）。这些可以限制用户请求的反馈，并帮助用户考虑一组更丰富的标签。

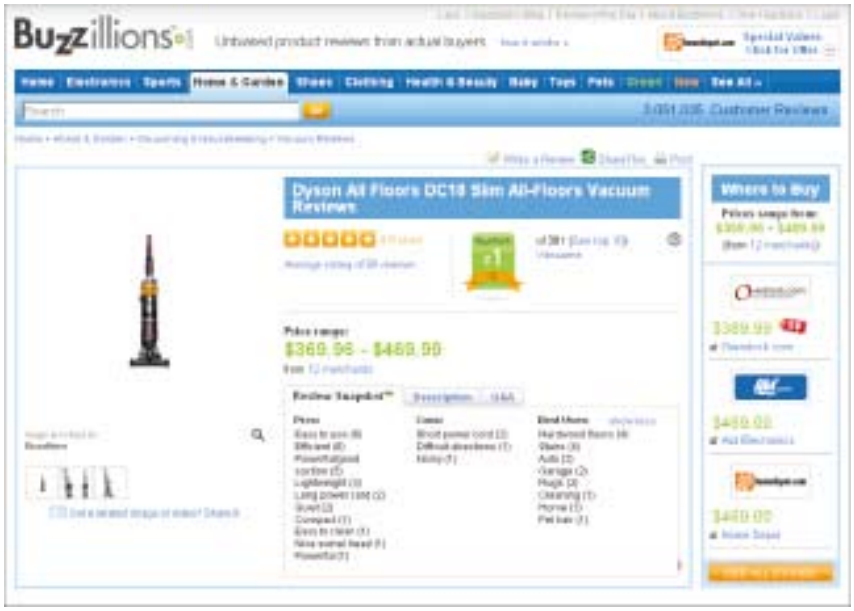


图 4-16　Buzzillions 显示一个 Dyson 的评论，标有“easy to use”和“nice swivel head”

可以使用这些标签来缩小搜索时的选择范围。由于标签在标准化时最为有效（更少同义词，更少拼法错误），Buzzillions 通过建议之前使用的标签来鼓励标准化（参见**图 4-17**），这并不是强制性的。Buzzillions 后期没有聘请信息架构师来创建一个词汇表。他们只是希望把用户轻轻导向一个正确的方向，并利用人的懒惰本性（我们总喜欢简单的单击而不是直接键入），所得到的分类系统会比随意的标签更为有用。

图 4-17　Buzzillions 在之前使用的标签旁提供了一些可以轻松单击的方框，以便很容易地增加标签

12. 标签类型

类似于元数据，存在多种不同类型的标签。如果能让你的系统识别这些标签，就可以开发一个基本的分类系统，使检索更容易。

图 4-18 显示了用户标记数据库中的项时可能使用的标签。不同类型可能适用于不同的企业，不过所有这些对于用户群体增长以及更好的自我表达都很有用。

标签类型	示　例
描述性	Massachusetts, city, Cambridge, architecture, U, building, night, ma, skyline, sky
资源	Book, video, podcast, photo, illustration
所有权 / 来源	NYTimes, austingovella (author), boxesandarrows
观点	Lame, tooshort, dontwasteyourhardearnedmoney
自我称谓	Me, mine, sawlive, ownit
任务组织	Todo, toread, toprint
演出和表演	Squaredcircle, akavogonpoetry, defectivebydesign

图 4-18 Gene Smith 的标签类型表，选自他所著的一本绝妙的书 Tagging，这里的一些新示例由本书作者提供

别老想着对用户打标签加以控制，从社区的特性（参见第 9 章）出发，更好的做法往往是鼓励好的行为而不是严令禁止不好的行为。比如，AkaVogonPoetry 这个标签出自《银河系漫游指南》这部著名科幻小说，Vogon Poetry 口碑很不好，会导致身体不适、痉挛和恶心[①]。对于不招人喜欢的艺术作品，这个标签比那些难以启齿的脏字要好得多，所以 Amazon 保留了这个标签，还提供了 Defectivebydesign，这是所有憎恶 DRM（数字权限管理）的人选择的标签。不过，对某个对象显示标签供参考时，更多地应当从描述性类选择，而少从观点类选择。

①“当然，Vogon 的诗还只是这个世界第三差劲的。
第二差的是 Kria 的 Azagoths。一次著名诗人 Grunthos Flatulent
朗诵了 Azagoths 的诗作‘Ode To A Small Lump of Green Putty I Found In My Armpit One Midsummer Morning’，其间 4 位听众脑溢血死亡，Mid-Galactic Arts Nobbling Council 的院长几乎把自己一条腿咬掉才保住性命。报告称 Grunthos 对这首诗的收听效果“很失望”，并且着手开始读他的 12 部长诗“My Favourite Bathtime Gurgles”，而为此他自己也终于不堪忍受这种极度煎熬。
最差的诗由其缔造者英国艾塞克斯郡 Greenbridge 的 Paula Nancy Millstone Jennings 带来，它简直能毁灭这个星球。”Douglas Adams 所著的《银河系漫游指南》，这是所有信息架构师必读的一本重要的书。

13. 标签系统的难题

“哇鸣！”你可能会说。干脆一切都用标签好了！嗯，先别那么快下结论。从 Vogon Poetry 的例子中可以看到，那些疯狂的用户可能并不总站在你那一边。下面来讨论关于标签的一些难题。

1. 冷启动问题

Amazon 第一次在网站中引入标签时，几乎没有人用它。确实用到标签的少数人也是从字面上理解“标签”的含义，并采用前面提到的纽约装配工的方式把他们的名字加到商品上作为标签。标签还没有得到广泛接受，如果提供一个空的表单域，前面显示有“标签”一词，并不是所有人都能知道该做什么。另外，如果你的系统中没有给出标签的例子，人们考虑如何描述一个给定对象时就会有困难。

对此有下面几种解决方法。

- 尽量使用对人们更有意义的标签，如“材料”（“materials”）、“主题”（“topics”）、“关键字”（“keywords”），或者是任何适合网站内容的标签。
- 包含一个简要的说明，指出如何对标签域旁边的对象加标签（见图 4-19）。
- 通过找出描述中与众不同的词来创建初始标签。技术人员应当能通过一个程序找出这种与众不同的词，并显示为标签。这会让人们对于什么是标签以及标签有什么作用有所认识。
- 让公司的同事也来加标签，创建初始的示例和活动。如果无法向他们解释怎样做，向用户讲解时同样会遇到麻烦。

2. 明显标签问题

建议工具可能会对“白纸”问题有所帮助，但也可能创建一个不走运的反馈循环，相同的标签可能会反复使用。大量标签能够使事物更可查找。那么如何让人们增加更多的标签呢？

- 允许很容易地增加大量标签。这正是很多标签系统使用一个包含逗号分隔标签的表单域的主要原因，你可以把所有标签全部放在一起。
- 建议大量标签，包括不太流行的标签。
- 回顾建议工具提供的流行标签和过于通用的黑名单。

图 4-19 Amazon 提供了一个可供使用的标签视图

3. 重复标签问题

来看图 4-20。在各个时期 Flickr 最流行的标签中，我们看到的是 photo、photos 和 photography。如果这些都算在一起，这可能是系统中最流行的一个标签。目前，如果搜索“photo”，你不会得到标签为 photos 或 photography 的项。也许这些项确实有足够大的差别，有必要使用单独的标签，但一般情况下都并非如此。在你的系统中，你可能还希望人们查找“boots”时能找到旅行靴（hiking boots），而不论使用哪个标签。你可以自己担负这个重任来设计一个同义词链（参见第 5 章），或者也可以鼓励用户创建他们自己的同义词。

- 建议相关词。
- 允许用户建议相关词。
- 允许用户轻松关联标签的创建工具。

图 4-20　Flickr 最流行的标签

4. 竞争标签问题

在 Cory Doctorow 有趣的评论“Metacrap”[①]中，他指出：

“元数据存在于一个竞争的世界。供应者竞争着要出售他们的商品，古怪的人竞争着要把他们想入非非的怪理论（请原谅我的措辞）告诉别人，艺术家竞争着向观众展示。注意力是发散的，钱包可能并不发散，但是它们有一定的相似性。

这就是为什么：

- 在一个搜索引擎（如 Altavista）上搜索任何经常引用的词时通常前 10 个结果中至少有一个黄色链接。
- 你的邮箱充斥着主题类似“Re: The information you requested.”（回复：你请求的信息）的垃圾邮件。
- 出版商的发行机构发出广告，鼓吹“你可能已经是赢家！”
- 出版发行机构有规模庞大的目录，满是空话和漂亮口号。

① 如果你是一个信息架构师而且在考虑标记，这是必读的（http://www.well.com/~doctorow/metacrap.htm）。

让你的生活中充斥着废话的人是不是应该下地狱？这个话题可以留到其他论坛中讨论。不过人们会在你的系统中为所欲为，而不是做你真正希望他们做的事情。

这个问题解决起来很困难。可以限制一个人为某一项所关联的标签个数，但是这会降低系统的整体有效性。可以向任何人开放标签，而不只是内容的提供者，希望所有标签的累积会比“专家”的想法更妙，但是这也会为不合适的标签大开方便之门。你可以进行监视，但是由于支持成本的增加，这会削弱让用户使用你的分类而得到的一些好处。你可以要求用户采用 wiki 方式来监督，但是当提供者相互竞争时这很难做到（例如，在一个贸易环境中）。如果鞋子的重要制造商从对方的产品中去掉“shoe”标签该怎么办？

为了解决所有这些问题，必须了解你的用户、你的群体、你的内容，以及你的员工。上下文是关键，你选择的解决方案必须反映你要达到的结果。