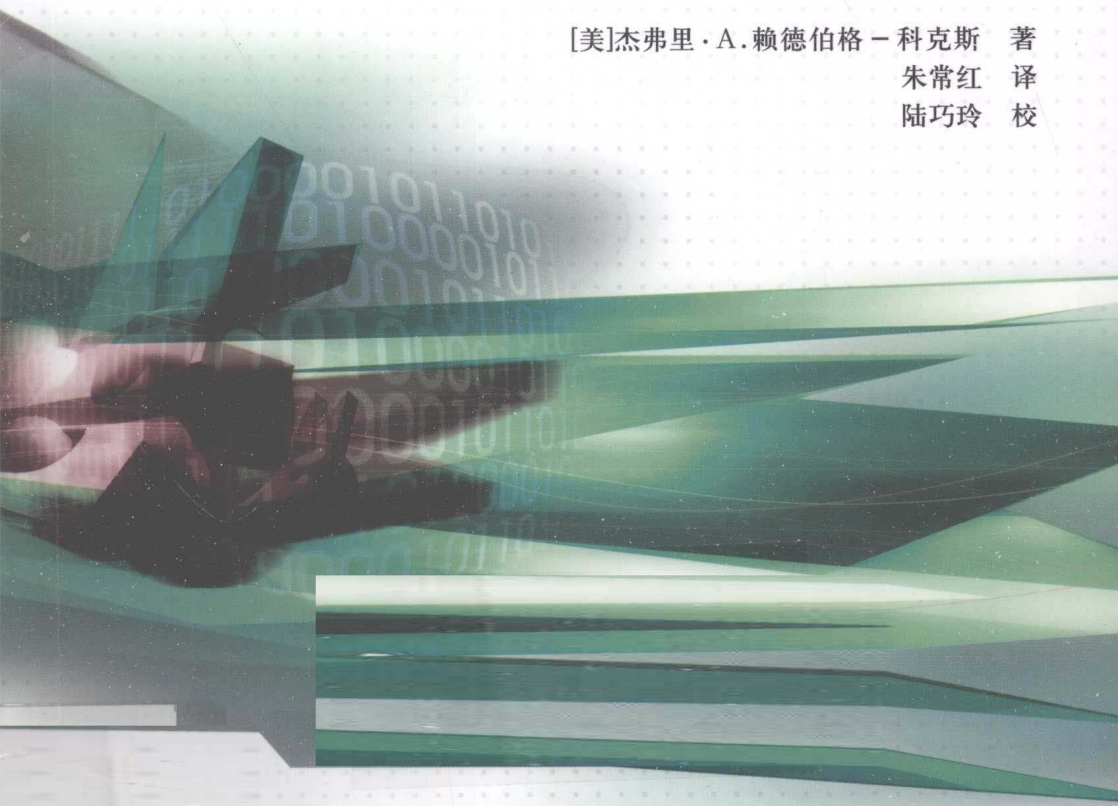
TIAOZHAN SHUZI TUSHUGUAN
HE SHUZHI RENWEN KEXUE

挑战数字图书馆和数字人文科学

[美]杰弗里·A.赖德伯格－科克斯 著
朱常红 译
陆巧玲 校

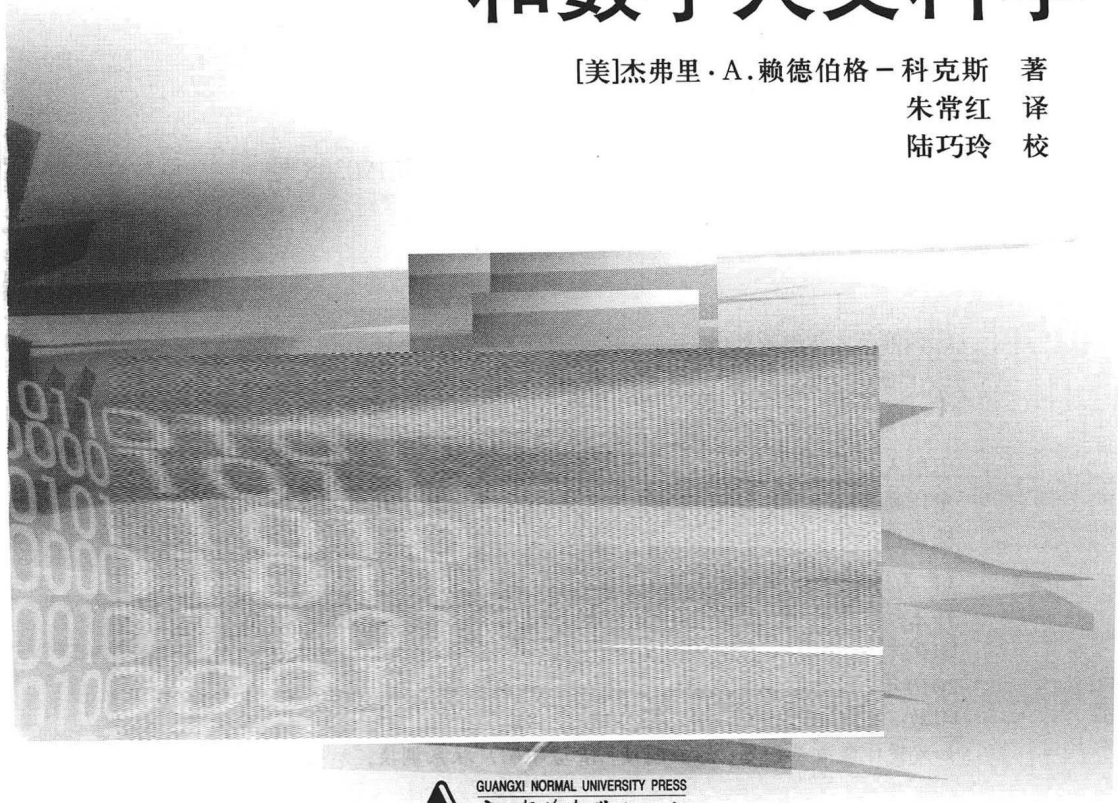


GUANGXI NORMAL UNIVERSITY PRESS
广西师范大学出版社

TIAOZHAN SHUZI TUSHUGUAN HE SHUZHI RENWEN KEXUE

挑战数字图书馆和数字人文科学

[美]杰弗里·A.赖德伯格－科克斯 著
朱常红 译
陆巧玲 校



GUANGXI NORMAL UNIVERSITY PRESS
广西师范大学出版社
·桂林·

著作权合同登记号桂图登字：20-2010-297 号

图书在版编目（CIP）数据

挑战数字图书馆和数字人文科学 /（美）赖德伯格-，科克斯著；朱常红译．—桂林：广西师范大学出版社，2010.9
ISBN 978-7-5495-0100-7

Ⅰ．挑… Ⅱ．①赖…②朱… Ⅲ．①数字图书馆—研究②数字技术—应用—人文科学—研究
Ⅳ．①G250.76②C39

中国版本图书馆 CIP 数据核字（2010）第 190308 号

广西师范大学出版社出版发行
（广西桂林市中华路 22 号　邮政编码：541001
网址：http://www.bbtpress.com）
出版人：何林夏
全国新华书店经销
桂林日报印刷厂印刷
（广西桂林市八桂路 2 号　邮政编码：541001）
开本：880 mm × 1 240 mm　1/32
印张：4.125　字数：100 千字
2010 年 9 月第 1 版　2010 年 9 月第 1 次印刷
印数：0 001~1000 册　定价：20.00 元

前言

这本书是七年多来我在数字人文科学和数字图书馆领域研究的成果。作为一名古希腊文学研究者，我的学术培养注重结合古雅典政治和司法活动中语言的艺术来着手神话、宗教和仪式的研究。作为一名学生，我对计算机有着一种强烈的兴趣，在攻读大学和硕士学位时，我从事了各种技术和编程设计的工作。1998 年我获得博士学位，在波士顿之外，我很幸运获得一项基于塔夫斯大学(Tufts University)的珀尔修斯数字图书馆(Perseus Digital Library)博士后研究项目。这项研究工作给我提供了在一个合作研究团队的背景里，将我的文史兴趣和计算机工作经验相结合的机会。作为珀尔修斯研究团队的一员及在我后来的职业生涯中，围绕着珀尔修斯数字图书馆的资料体系的核心问题，我进行了从文学定量研究到希腊语词典编纂的各种项目研究。这本书的原始资料得自那项研究，在很多案例里，它们总结和归纳了我已经完成的、与珀尔修斯数字图书馆相关的研究。

这本书属于“文化遗产语言技术”(The Cultural Heritage Language Technologies，CHLT)研究项目成果之一，该项目获得 2001.1～2005.12 国家科学基金会和欧洲委员会国际数字图书馆方案科研基金资助。“文化遗产语言技术”研究项目将参与从有关计算机语言学、自然语言处理和信息检索技术领域里技术和技能最有效运用方式的研究，到学生和学者面对的用希腊文、拉丁文和古斯堪的那维亚文写作文本研究的挑战。在我作为美国这项项目的主要调查员时，

这项项目已经发展到拥有4个不同国家、8种不同机构的众多合作者。这个团队里的成员具有从传统人文学科、计算机科学到数字图书馆等领域里的专门技术。我得以写成此书要从两个方面感激“文化遗产语言技术”项目组。首先，这本书里的很多方面反映了我已经完成的、与珀尔修斯数字图书馆相关的研究，它也报告和归纳了我们所做这一项目的研究。其次，也许更为重要的是，这种集中时间进行两种领域交汇的研究，最后产生了这本书概念性的结构。当我反思已经完成的这项研究项目的工作时，我试图发展能汇集我们工作的众多不同方面的一般性范畴。最后，在观念上我归结到——不管我们属于何种专门技术领域——我们项目组成员在四个领域里具有共同的兴趣：(1)提供了存取珍稀、易碎的重要源资料的物质；(2)帮助读者理解用难于理解的语言写作的文本；(3)授予研究者新类型学识管理和使用的方法；(4)为未来保存数字资源。这是曾经发表在《数字图书馆》杂志上的一篇论文，在这里我再现了这篇论文的构架，这种构架提供了一种组织我们项目研究的有用的方法。

所有参与这项研究的人彼此之间都非常合作。我写成此书，得到了许多人直接地指点；或者他们作为参与这项研究的团队成员，和我一起完成了这项研究，间接地帮助我完成整本书的构思。首先，我感谢珀尔修斯数字图书馆的总编格雷戈里·克莱恩(Gregory Crane)先生，没有过去七年多来他的慷慨支持和鼓励，这项研究就不可能完成。埃尔皮达·安特汉(Elpida Anthan)、卡特·威尔逊(Cat Wilson)、凯瑟琳·弗莱彻(Catherine Fletcher)、布鲁斯·弗雷泽(Bruce Fraser)和布鲁斯·布雷德利(Bruce Bradley)等人阅读了这本书的草稿，并不断地提出修改意见。玛撒·约翰逊－奥琳(Martha Johnson－olin)是这本书定稿时不知疲倦的读者。

我也非常感谢我的堪萨斯州同僚在许多不同项目上帮助我；非常感谢密苏里大学堪萨斯分校(UMKC)的埃尔皮达·安特汉、卡特·威

尔逊、凯瑟琳·弗莱彻、琳达·沃格特斯(Linda Voigts)、琼·迪安(Joan Dean)、乔治·威廉斯(George Williams)、汤姆·斯特罗克(Tom Stroik)、劳里·艾林豪森(Laurie Ellinghausen)和拉腊·维特(Lara Vetter);非常感谢林达·霍尔图书馆的布鲁斯·布雷德利、比尔·阿什沃思(Bill Ashworth)、赖安·费根(Ryan Fagan)和辛迪·罗杰斯(Cyndi Rogers);非常感谢国家医学图书馆的迈克尔·诺斯(Michael North);非常感谢密苏里大学堪萨斯分校图书馆的罗伯特·雷(Robert Ray)、布鲁斯·舍伍德(Bruce Sherwood)、特丽萨·吉普森(Theresa Gipson)、玛里琳·卡博内尔(Marilyn Carbonell)、布伦达·丁格莱(Brenda Dingley)和格温·威廉斯(Gwen Williams)。我也深深地感谢我的珀尔修斯同僚:大卫·史密斯(David Smith)、安妮·马奥尼(Anne Mahoney)、利萨·塞拉托(Lisa Cerrato)、罗伯特·夏维兹(Robert Chavez)、艾米·史密斯(Amy Smith)、玛丽娅·丹尼尔斯(Maria Daniels)和大卫·米诺(David Mimno)。作为"文化遗产语言技术"项目组部分获得资助,我在珀尔修斯主要研究项目之一是与剑桥大学写作组一起合编一部新中级水平希腊语—英语词典(有关这方面的描写在第4章)。我的这个项目伙伴,安妮·科尔曼(Anne Coleman)和布鲁斯·弗雷泽的帮助使我有关研究项目和数字图书馆相互作用的思考更完善。"文化遗产语言技术"项目组的其他同僚,也给我的许多思考带来影响,他们是伦敦皇家大学的多洛雷斯·艾奥利佐(Dolores Iorizzo)、斯蒂芬·拉格(Stefan Ruger)和丹尼尔·希斯克(Daniel Heesch),肯塔基大学的罗斯·斯凯夫(Ross Scaife),加利福尼亚大学洛杉矶分校(UCLA)的提姆·坦赫利尼(Tim Tangherlini)和克利兹托夫·厄本(Kryztof Urban),哥本哈根大学的马特·德里斯科尔(Matt Driscoll),以及在意大利比萨的意大利国家科学研究院的计算机语言学研究所的安德烈亚·博悉(Andrea Bozzi)、帮洛·鲁弗罗(Paolo Ruffolo)和马可·帕萨罗蒂

(Marco Passarotti)。在肯塔基大学计算机科学中心介绍了第一章初稿后，我得到了极其有益的评价。

这项工作也得到了许多组织的慷慨资助，其中包括国家科学基金会国际数字图书馆项目、国家基金会人文科学保存和存取部、密苏里大学研究部和国家医学图书馆医学史部等提供的资金资助。

注释

1. 赖德伯格－科克斯（Rydberg－Cox）（2005a）著作中将这一概念性体系结构应用到“文化遗产语言技术”研究项目。“文化遗产语言技术”主页，http://www.chlt.org. 赖德伯格－科克斯（2003a）著作中作出了“文化遗产语言技术”研究项目最初的描述。

作者介绍

杰弗雷·A. 赖德伯格－科克斯（Jeffrey A. Rydberg－Cox）博士是密苏里大学堪萨斯分校英语系主任、副教授，古典研究项目主管，古典研究、宗教研究项目成员以及计算机工程学院相关学科的大学教学人员。1998年，芝加哥大学古地中海社会学术委员会授予他博士学位之后，他以程序员和计算机词典编纂者的身份在塔夫斯大学的珀尔修斯数字图书馆工作了两年之久，直到2000年他才开始从事现在的工作。

他的研究主要集中在这两方面：使用计算方法学研究希腊和拉丁语言，以及研究希腊宗教和神话中的修辞学。他的计算机语言研究服务于一所网络中心，该网络中心拥有众多单独但相关的项目，产生了包括新中级水平希腊语－英语词典编纂研究、英语－拉丁文本数字化语料库群研究，特别关注数字图书馆知识管理和信息检索的研究。在这些领域里，他已经发表了30多篇文章，还写完另一部有关利西阿斯（Lysias）演说方面的书。

从国家人文基金会、国家科学基金会、国家医学图书馆医学史部和密苏里大学研究部，他已经得到这方面的研究项目。他也参与了两项国际项目：一项由国家科学基金会和欧洲委员会联合资助，另一项则由国家科学基金会和德国科学研究协会联合资助。

作者联系方式：

地址：美国堪萨斯州，密苏里，密苏里大学堪萨斯分校
英语系科克事务大厅 106

邮编：（美国）64110

插图目录

1.1 佛罗伦萨科学史研究所和博物馆的伽利略军用圆规目录词条 …… 5
1.2 珀尔修斯数字图书馆的希罗多德《历史》网页展示自动生成超链接 …… 9
1.3 屏幕展示珀尔修斯数字图书馆的荷马《伊利亚特》文本 μῆνις 一词的形态学分析 …… 10
1.4 特拉华大学图书馆科学史学科导航 …… 24
2.1 电子文本语料库里,除了《吉尔伽美什》、《伊奇丢》和《地狱》译本之外的苏美尔文学的 HTML 代码 …… 30
2.2 1880 年出版的爱德华·卡洛斯译著《西达瑞斯·奴奇斯》书名页的 XML 编码 …… 35
2.3 为编码书名页,简单假设 DTD …… 36
2.4 根据文本编码倡议方针,1880 年出版的卡洛斯译著《西达瑞斯·奴奇斯》书名页的 XML 编码 …… 38
2.5 1880 年出版的原作者为伽利略的《西达瑞斯·奴奇斯》译著正文的第一页 …… 40
2.6 符合 TEI 标准的 XML 代表——《中级希腊语－英语词典》geras 词条 …… 42
2.7 1472 年出版的普林尼《自然史》中某些缩写词、单词区别和单词违例的典型问题的举例说明 …… 45
2.8 具有相应唯一标识符的未知象形文字样表 …… 47
3.1 美国股票市场的 Smartmoney.com 自组织地图 …… 70
3.2 格洛克地图展示搜索《西达瑞斯·奴奇斯》的结果的文献集 …… 71

CONTENTS

目录

前言/ i

作者介绍/ v

插图目录/ vii

1 **引言**/1

案例研究/2

挑战数字图书馆/12

什么是数字图书馆/13

数字人文科学的前数字类比/19

关于这本书的内容/24

注释/26

2 **提供存取文本**/29

什么是贴标签/30

文本编码倡议/37

怎样显示 XML 文本/42

挑战编码/43

它是怎样实行帮助的/47

注释/52

3 帮助读者理解学识/56
解析器和解析了的语料库群/56
关键词分析/59
信息检索/63
高级信息检索：查询扩展、可视化和多语种检索/65
链接相似段落/73
结论/75
注释/76

4 有用的新学识/79
文学定量研究/79
词汇定量研究/84
词典编纂/88
结论/91
注释/92

5 新学识：数字图书馆和特色学术机构库/94
注释/99

6 结论/100

参考书目/102
索引/113

1. 引言

数字技术已经对许多学者的人文科学管理和分享他们的研究产生了一种意义深远的影响。一旦文本数字化，即使最普通的搜索设备也允许用户交互工作和以完全崭新的方式研究文本。电子媒介打开了传播的新模式和思考文本的新方法；学者们可以使用交互式乐谱、动态生成地图，或者使用其他多媒介元素，以不同于纸质印刷文章的方式来传达信息。同时，这些电子资源可以从根本上改变读者使用资料的方式和完成人文科学学者承担的工作。

当这种围绕着印刷的文化实践，已经有助于适合在特殊地理位置的支持者研究怎样创建图书馆时，数字图书馆在互联网上的使用，可以实现远离大学图书馆，甚至远离学校、公共图书馆、工场和私家居所的读者要求。没有地理位置限制的广泛检索得到加强或者成为学术机构成员的需要，这允许人文科学学识在学生、专业学者和普通公众等人的生活中，起着一种新的不同的作用。更为重要的是，由利用电子文本进行人文科学研究的学者设计的工具和技术，能允许答问读者有关他们使用传统纸质印刷资料绝对不可能提出的问题。充分利用由人文科学学者承担的计算机研究来设计数字图书馆，戏剧性地改变了广泛分散的公众怎么样和为什么阅读、研究，以及与文学、历史和档案资料交互作用的潜能。数字图书馆实践者面临的主要挑战将给广大读者建构一种能带来工具的系统。

案例研究

许多文化批评家已经评论计算机、互联网发展对社会的潜在影响，以及相关技术能改变我们所知的知识和所思的方式。其中，最著名的例子就是范内瓦·布什(Vannevar Bush)的文章《遥想未来》(*As we may think*)(1945)。在这篇文章中，布什设想一种他称之为"麦麦克斯存储器"(memex)的设备存取人类知识的假定系统。这种系统预见了万维网结构和数字图书馆的许多方面。一旦这种系统制造出来，布什推测：

> 全新形式的百科全书将出现，它准备制成一种具有横贯其中的联合线索的网状书，（外来信息）随时可以添加到麦麦克斯存储器并不断沿着联合线索扩展。譬如，律师轻轻一碰这种网状书，就具有自己以及对方律师和法官以往经历所作出的意见和决定。拥有这种书，专利代理人就拥有数百万种已发行的专利，熟悉有关他客户利益的每一点线索。被病人反应困扰的医生，可以在这种百科全书中找到以往类似病例研究的线索，通过从侧面参考一流的解剖学和组织学的相关案例，迅速地参看处理类似病例的历史从而诊断。为综合有机化合物而奋斗的化学家，在他的实验室里利用这种百科全书，可以找到他之前所有化学家的化学文献，既有列属相似化合物的正面参考资料，也有这些化合物的物理和化学反应的侧面参考资料。
>
> 利用这种百科全书，历史学家可以有着大量的依时序而编的人类史料，他们可以跳过整个跟踪的线索，只停留在那些显著的条目上，并按照当下任何一个时代的线索，来跟踪这个特定时代的文明史。由于这种百科全书开发者是一批具有新专业知识的开拓者，他们通过大量共同的记录，可以在确定有用线索的任务中找到乐趣。通过这种工具，大师的遗产不仅仅增添了世界文明的记录，而且因为这种百科全书具有联

合线索，从而能为他的弟子竖立起整个研究系统的架构。（布什，1945：180）

在这一段的结论中，布什巧妙地轻描淡写道，“所有类型的技术困难已经被忽视”。确实，60年之后，当计算机技术已经传遍整个社会和学术界时，这种想象力明显不是狂妄幻想，但它也不完全是事实。在学术情境中，那些为布什清晰描述的幻境所激动的学者们，一直在研究创造这些新作品和克服这些“技术困难”。通过使用计算机技能和技术，这些学者已经完成了有趣的、重要的，甚至是惊人的项目。四个研究案例阐释了数字文本和工具转变人文科学学识的潜能。[1]

幽谷影响域项目

幽谷影响域项目（Valley of the Shadow Project）：有关美国南北战争中的两个乡镇的项目[2]，它集中了美国南北战争期间两个美国乡镇——南部弗吉尼亚州的奥古斯塔镇和北部宾夕法尼亚州的弗兰克林镇的一种广泛、多样性的档案证物。这个项目集合了诸如人口普查记录、日记、书信、报纸和许多其他重要原始资料。从最简单的层面上说，这种档案从根本上改变了观众和读者利用资料的潜力。没有这种档案，希望使用这些原始资料的任何人将不得不需要在许多不同的图书馆里走动和花费时间来查找有关档案。此外，一旦在图书馆里，在诸如手记、旧报纸和政府记录等许多不同形式、不同类型的原始资料中，这个研究者将面临繁重的劳动任务，只有在大量劳作之后，才可能查到有关任何特别人、地点，或者特定事件的信息。随着幽谷影响域项目档案归档管理和可在互联网上使用，现在世界上任何研究者都可以从他们自己家或者办公室里存取这些资料。再者，如果他们对一个特殊月份或者年份里发生的事件感兴趣，那么他们可以无休止地观察得自当地报纸的信息，使用它们来寻找私人信件、传真记录、人口普查

记录和其他文献的文本。私人信件附带有进一步促进他们使用的额外特征。个人手稿可能很难阅读，因此这种档案提供手稿的抄本。19世纪50年代和60年代的拼写习惯不同于现代的拼写习惯，因此在原稿旁边，这种档案提供现代版。最后这种档案提供许多信件的图像，因为这种信件本身可能具有令人感兴趣的元素，诸如附带有图画和图形之类。

当这些功能使得使用这些档案资料较便捷时，人们也许可能会这么说，倒不如介绍任何一种新事物，因为这些档案工具研究者，他们仅仅促进查找资料传统任务的执行。实际并非如此，幽谷影响域项目也提供允许人们以全新的方式，探索档案信息的机制。例如，这个项目提供活生生的地图来阐明南北战争中两个乡镇的活动，连同有关每个单元里他们主要阻击敌人的实际情况。这些地图是根据不同时间发生的不同事件来综合的，因此，用户可以把每个单元里的生活活动与这些活动发生的时间联系起来。这种地图也提供当下参考点，诸如现代城市和道路之类的，允许用户把每个单元里的活动与他们可能熟悉的地标联系起来。具有这种功能性，用户查阅不熟悉的人和地点，不再因为面临大量的、包含广泛多样资源的信息而不知所措。相反，他们可以使用清晰、可视的传达来探索每一个军队的行动。

佛罗伦萨科学史博物馆

幽谷影响域项目和许多诸如此类的项目，清晰地阐明了数字技术在历史和档案研究工作中的运用，诸如在时间和空间上确定个体和定位过去重要事件方面。在帮助读者理解不熟悉的历史文物方面，这些技术可能是相当有用的。我们发现最好的例子之一就是由佛罗伦萨科学史研究所和博物馆(The Institute and Museum of the History of Science in Florence)所创造的网上资料(http://galileo.imss.firenze.it/)。在20世纪早期，创建这个研究所主要是用来收集、保存和修复

与科学史相关的各种工具。它最初的馆藏由许多仪器组成，曾经由梅第奇家族拥有，之后由佛罗伦萨大学所持有。这所博物馆积极地使用了新信息技术作为他们展览的一部分，并向广大观众解释他们的收藏。与幽谷影响域项目一样，某些这种电子工具推动了探索博物馆书目主题的传统工作、查找背景材料和找到博物馆收藏主题之间的联系。组织这种网上目录，以便人们能够以各种方式开始他们的调查：博物馆的交互式地图；发明者、制造者的字母索引，或者人们引用的博物馆原始资料；一种虑及主题－中心论题索引的知识集合的探索；以及由仪器之类组成的博物馆馆藏一览表。

例如，一名喜爱博物馆里伽利略(Galileo)专用展厅的访问者可能转为喜爱网上博物馆，在那间展厅中查找陈列对象的目录，然后点击其中选中的一条，查看目录词条和图片。当用户查看目录词条时，边注栏提供与其他对象特殊背景的链接，视频为这种条目提供一种历史和文化的背景，提供更多目录的详细信息，链接其他能进一步提供信息的网站。插图 1.1，我们看到伽利略军用圆规(Galileo's Compass)目录词条，其中包括传统目录信息、实物图像、相关人和物及词条右边的背景化信息的链接。[3]

插图 1.1　佛罗伦萨科学史研究所和博物馆的伽利略军用圆规目录词条