中国科学院研究生院
博士学位论文

中国虚拟天文台系统设计

作　　者　　崔 辰 州
指导教师　　赵 永 恒
学科专业　　天 体 物 理
研究方向　　天 文 数 据 处 理 方 法
申请学位　　博　　士
培养单位　　国 家 天 文 台

二〇〇三年六月

National Astronomical Observatory
Chinese Academy of Sciences

System Design for Chinese Virtual Observatory

Chen-Zhou CUI

Advisor: Prof. Yong-Heng ZHAO

June 2003

摘　要

随着天文望远镜及终端设备的设计与制造技术不断提高，天文观测能力大大增强，天文学已从古老的光学观测变为全波段的天文学，并正在进入一个“数据雪崩”时代。计算机与互联网技术的飞速发展，网格技术、XML 技术、语义网技术等全新 IT 技术的涌现，使得海量、分布式、多波段天文数据的无缝融合和处理成为可能。在这样的背景下，旨在将世界范围内主要天文研究资源无缝透明地整合在一起的虚拟天文台（VO）设想应运而生并很快得到世界各国天文界的重视。2002 年 6 月，国际虚拟天文台联盟（IVOA）成立，VO 进入一个国际联合开发阶段。同年，以中国科学院国家天文台（NAOC）为代表的中国天文界提出中国虚拟天文台（China-VO）计划并于当年成为 IVOA 的成员。

论文系统地阐述了建设 China-VO 的意义和实施策略，给出了 China-VO 的体系结构框架；介绍了 VO 各研究领域的最新动态、相关技术，并给出了 China-VO 可能采取的解决方案；阐述了 China-VO 与 LAMOST 紧密结合的必要性，提出建设 VO-enabled LAMOST 的设想，按照 IVOA 数据互操作标准的制定思路给出了一套 VO 兼容的 LAMOST 数据模型。

China-VO 的研究与开发为中国天文界带来了许多崭新的机遇和挑战。通过 China-VO 这座通向国际虚拟天文台（IVO）的桥梁，中国天文学家可以及时的共享国际天文研究的最新技术和成果，同时将自己的资源与国际同行分享。China-VO 将与目前国内天文界唯一的大科学工程 LAMOST 项目紧密结合，以 LAMOST 数据产品作为重要的资源基础，以光谱巡天数据处理及相关分析服务为主要特色。China-VO 将把国内天文学、计算机科学、数学统计等多领域的科学家联合在一起，同时与国际合作伙伴一起推动 IVO 的研究与开发，争取早日将 VO 的美好设想变成实实在在的数据密集型在线天文研究平台。

以开放网格服务架构（OGSA）为代表的网格技术旨在消除互联网上的“信息孤岛”，实现数据资源、计算资源、存储资源等各种网络资源的全面共享。OGSA 为 China-VO 的建设提供了重要的网格平台，其架构是 China-VO 体系结构的基础。整体上看，China-VO 由构造层、资源层、汇集层和用户层构成。China-VO 采用面向服务的设计理念，整个服务体系包括应用服务提供者、数据服务提供者和 VO 注册三种角色。各种功能的实现由不同层次的 VO 服务承担。

OGSA 通过定义一系列标准的服务描述和操作为 China-VO 系统内资源与

服务的注册与发现提供了基本的实现机制。China-VO 将与 IVOA 的合作伙伴一起制定被国际同行认可的资源注册与发现、服务注册与发现等方面的一整套标准和协议，最终实现 VO 系统中资源与服务的动态发现和高效利用。

互操作性是 VO 对资源与服务的基本要求，是 VO 功能开发的基础。IVOA 成员通过制定并共同采用适应互操作性要求的 VO 数据模型、数据交换标准和元数据标准，实现 VO 对异构资源的统一访问和处理。China-VO 利用网络存储、虚拟存储等先进的存储技术实现对海量数据的存储；通过开发高效的检索算法，采用与网格兼容的数据库管理系统实现数据的高效访问。

适应性广、功能强大的应用服务是 VO 最终成功的关键。VO 对传统的数据挖掘、知识发现、可视化、高性能计算等领域提出了挑战。China-VO 将从自身实际情况出发，优先在光谱巡天数据自动分析处理领域取得突破，提供一定数量的高层服务，以中国国家网格为平台实现网格环境下的高性能计算服务。

VO 门户是用户访问 VO 资源与服务的重要途径。China-VO 将根据使用目的和需求的不同，为专业用户、非专业用户和系统管理员、开发人员等这样的特殊用户提供不同的访问环境和界面。China-VO 门户将采用模块化、标准化的开发模式，以实现不同服务提供者提供的服务在 China-VO 门户中方便的整合，让用户能简便直观的定制个性化的访问环境和使用系统服务。

作为联结国内外天文研究的桥梁，China-VO 通过采用符合国际化标准的工具和策略进行国际化和本地化开发，让国内用户方便舒适的访问国际资源，同时将国内资源，特别是珍贵的历史资料共享给国际用户。

与 LAMOST 紧密结合是 China-VO 的特色，实现 VO-enabled LAMOST 是 China-VO 的重要使命。这个目标将分两步实现。第一步，通过制定和采用符合 VO 标准的资源注册标准、数据模型、数据访问服务实现 LAMOST 数据与 VO 的兼容；第二步，将 LAMOST 望远镜 VO 化、网格化，使其成为 VO 系统中一个重要的资源节点。

VO 是一个新生事物，其赖以存在的技术基础尚未成熟。通过定义一些科学与技术范例可以帮助我们明确 VO 的科学与技术需求，为进一步的工作打下基础。在论文的最后，通过定义三个范例，对 China-VO 在数据访问、分布式计算、可视化、多波段数据的联合操作等方面的需求进行了探讨。

尽管 VO 脚下的大地尚在不停的晃动，但她以自己十足的魅力终将为古老的天文学带来一场新的革命。

关键词：天文技术—中国虚拟天文台—系统设计

Abstract

System Design for Chinese Virtual Observatory

CUI Chenzhou (Astrophysics)

Directed by Prof. ZHAO Yongheng

With the technological advances in design and manufacture of telescopes and instruments, the capability of astronomical observation is enhanced greatly. Astronomy has turned into a full electromagnetic waveband era from the traditional optical era, and is facing a data avalanche. With the breakthroughs in computer and Internet technology and emergences of brand new IT technologies, such as Grid, XML and semantic web, the concepts of seamless federation and analyzing of large distributed multi-band astronomical observation archives are no longer outlandish. Under such background, the concept of Virtual Observatory (VO) is brought forward, which rapidly draws tremendous attention of astronomers from many countries. The principal goal of VO is to integrate and utilize the worldwide major astronomical research resources transparently.

In June 2002, the International Virtual Observation Alliance (IVOA) was constituted, which indicated that VO stepped into a phase of international federal research and development (R&D). In early 2002, Chinese astronomical community, especially its leading institute, National Astronomical Observatory of China (NAOC), Chinese Academy of Sciences, announced the Virtual Observatory of China (China-VO) project. Then in October 2002, it became a member organization of the IVOA.

In the thesis, the significance and technical route of the China-VO are systematically described; the system architecture of it is designed; the latest progresses and technologies in VO-related research fields are introduced; the possible solution for the China-VO is provided; the necessary for closely cooperation between the China-VO and the LAMOST is pointed out, which brings forward the concept of VO-enabled LAMOST; And according to the IVOA standards, a set of VO-compatible data models for LAMOST data product is designed.

The development and construction of the China-VO brings both many new opportunities and challenges for Chinese astronomical community. Acting as the bridge to International Virtual Observatory (IVO), the China-VO will enable domestic astronomers to enjoy the latest progresses and discoveries of the worldwide

astronomy research, and to share our resources with international colleagues at the same time.

The China-VO will cooperate closely with the LAMOST, which is currently the only big-science project for Chinese astronomy. LAMOST sky survey data will be an important resource basis for the China-VO. An important characteristic of the China-VO is its services in spectral sky survey processing and analyzing. To improve the R&D of VO, China-VO will combine together our scientists and experts in Astronomy, Computer Science, Mathematics and Statistics, etc. Furthermore, the China-VO will cooperate with international VO partners to turn the charming VO idea into a real online data-intensive astronomical research platform.

Grid technology, with its preventative Open Grid Service Architecture (OGSA), aims to remove the "isolated islands" of information on the Internet, and to enable dynamic sharing of all kinds of network resources, including data resource, computing resource, storage resource and so on. OGSA provides an important technical platform for the China-VO. The system architecture of the China-VO is strongly based on that. The China-VO will adopt service-oriented conception during its R&D. Logically, the system is composed of four layers, i.e. fabric layer, resource layer, collective layer and user layer. The whole service system functions as three roles, i.e. application service provider, data service provider and VO registry. Different VO functions will be realized by different levels of VO/Grid services.

By defining a set of standard service descriptions and operations, OGSA provides the basic mechanisms for the resource registry and discovery of the China-VO. China-VO will cooperates with IVOA partners close to define a whole set of standards and protocols for resource registry and discovery in order to enable an astronomer to locate, get details of, and make use of, any resource located anywhere in any Virtual Observatory.

In VO environment, interoperability is the basic requirement for resources and services, and the base for developing high-level applications. By defining and adopting feasible and scalable VO data model, data exchange standard and metadata standard, heterogeneity transparency of resources access can be reached.

Chine-VO will take advantage of advanced storage technologies, including network storage and virtual storage, to store its huge datasets. By developing high efficient index arithmetic, adopting Grid compatible DBMS, the China-VO will provide high efficient data access service.

Flexible and powerful application services are the key elements for the success of VO. VO provides challenges for data mining, knowledge discovery, visualization, high-performance computing and other related research fields. China-VO will lay its development course from current condition, concentrate itself on spectral sky survey

data processing and analyzing, and provide some high-level services. Chinese National Grid will act as the test-bed for the China-VO to provide high performance computing services.

Portal is the main access to the VO services. China-VO will provide different user interfaces and access environments for different types of users with different requirements and goals. Users of the China-VO can be classified into three categories, i.e. professional users, non-professional users and special users. Special users include system administrator and VO developers. The China-VO portal will adopt modular and standard developing mode so that services from different providers can be integrated into the portal easily, and users can custom their access environments as they want.

Being a bridge linking national and international astronomy research, China-VO will follow internationalization standards and adopt compatible tools in its development. Internationalization and localization are both necessary for the China-VO to provide comfortable accessing interface for national users and to share the valuable historical Chinese observational data with international users.

A major characteristic of the China-VO is its close combination with the LAMOST. Realization of VO-enabled LAMOST is an important goal of it. That can be reached in two steps. First, to integrate LAMOST in to IVO by defining and adopting VO compatible resource registry standard, data model and data access service. Second, to integrate LAMOST telescope into VO and make it become an important resource node for the system.

VO is a brand new concept. Many technologies, on which VO is based, are still immature and in fast changing. In order to specify scientific and technical requirements for the China-VO in data access, distributed computing, visualization, interoperation of multi-waveband data, three demonstrations are designed in the last chapter of the thesis.

Although the ground is swaying constantly, VO will serve as an engine of discovery for astronomy and provide a path to the future.

Keywords: Astronomical Technology — Chinese Virtual Observatory — System Design

第一章 虚拟天文台概念的产生与研究现状

1.1 虚拟天文台概念产生的背景

四百年前伽利略首次把望远镜指向太空，使人类摆脱了仅能用肉眼直接观测太空的历史，为从哥白尼开始的天文学革命提供了大量的科学证据。历史悠久的天文学经过哥白尼、伽利略、开普勒和牛顿等人的发展，演变成了一门崭新的科学，同时也催生了现代科学技术[1]。

到一百五十年前，由于照相技术和光谱技术在天文观测中的应用，用人眼作为唯一的天文探测器的时代结束，诞生了天文学的新分支——天体物理学，并发展成为现代天文学的主流。

五十多年前，在第二次世界大战中得到蓬勃发展的无线电技术使得人类的视野跃出了可见光的波段，发展成为射电天文学。之后不久随着宇航时代的到来，空间天文学诞生，天文观测不再局限于地面。人类对宇宙的观测范围扩展到了γ射线、X射线、紫外和红外波段。天文学开始进入全波段天文学时代。

从20世纪90年代开始，天文学又经历着革命性的变化。这一变化是由前所未有的技术进步所推动的，即望远镜的设计和制造技术、大尺寸探测器阵列的设计和制造技术、高性能计算技术和互联网技术。

望远镜技术的进步使得人类可以建造大型的空间天文台，为γ射线、X射线、光学和红外天文的发展开辟了崭新的前景，同时也推动了新一代大口径地面光学望远镜和射电望远镜的建造。在光学与近红外波段，已经有了高灵敏度高分辨率而尺寸不断增大的探测器阵列。伴随着这些技术进步，天文学家们正在计划建造功能更强、口径更大的空间和地面望远镜，并将配备尺寸更大象素更多的探测器。

望远镜制造技术事实上反映了所在时代工程科学可能达到的最高设计水平[2]。由于许多崭新的技术的引入，比如主动光学技术、自适应光学技术、薄镜面拼接技术，使得人类可以制造出口径越来越大、技术越来越先进、观测效率越来越高的新型观测设备。目前30米口径的拼合镜面望远镜的方案正在计划之中，雄心勃勃的100米口径光学望远镜方案也正在形成。

大尺寸探测器技术这些年也取得了突破性的进展。从古至今，天文探测器

经历了四个时代[a]：

- 人眼（Isomerization/rhodopsin）
- 照相乳胶（Photochemics/potetial）
- 光电器件（Photoemition/workfunction）
- 半导体器件（Semiconduction/Bandgap）

并正在向第五时代过渡，也就是超导器件（Superconduction/Energygap）时代。

多年来天文探测器主要围绕四个方面进行改进：

- 更高的量子效率

通过镀波段转换介质、镀抗反膜、减薄、采用高阻半导体材料等措施提高探测器的量子效率。

- 更大的视场覆盖

7K×9K 像素、12 微米像元、86.0mm×110.6mm 大小的单片 CCD 已经由菲利浦（Philips）公司制造成功。同时，LM-Loral 公司也研制成功 9216×9216 像素、 81mm×81mm 的 CCD。

CCD的拼接技术也已经进入成熟时期，大大提高了CCD观测的视场覆盖能力。美国Sloan数字巡天计划（SDSS）的焦面接收系统就使用了 5 行 6 列由 30 块CCD组成的CCD阵列[3]。

- 更大的动态范围，CCD的动态范围可高达 10^4~10^5，远高于照相底片
- 更宽的响应波段，比如从红外、紫外到 X 射线

目前下一代天文探测器超导隧道结（Superconducting Tunnel Junctions，STJs）已经浮出水面。这种探测器在光谱响应、实时性、动态范围、覆盖视场等方面较以往的探测器都有很大的提高，可应用于：

- 暗天体成像光谱巡天
- 大动态范围两维光子计数
- 克服大气影响的观测，比如光斑成像/光谱、自适应光学中波前探测
- 短时标变化对象的观测，比如脉冲星
- 空间观测中的应用，比如空间光学干涉仪中条纹宽带光子计数系统
- 红外波段观测

等领域。

[a] 叶彬浔. CCD及其新进展. 2003

如同在计算机工业中反映计算能力随时间指数增长的摩尔定律一样，在过去十年中这些技术进步使得天文学的数据收集能力也遵循着摩尔定律[4]，光学望远镜接收面积每 25 年增长一倍，而CCD像素每 2 年增长一倍，如图 1.1 所示。

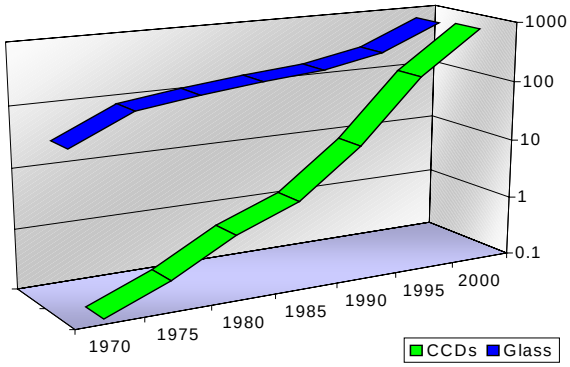


图 1.1 望远镜接收面积和CCD象素的增长曲线[b]

同时，天文学研究走向新时代离不开高性能计算和互联网技术的发展。

科学计算同理论研究和科学实验一样，已成为人类探索未知世界的重要科学手段。高性能计算在基础科学研究、国民经济发展和社会进步中具有不可替代的作用。

超级计算机已经成为 21 世纪社会经济、科技发展，国防现代化建设的最重要的基础设施，科技创新的重要工具。为了解决 21 世纪国际社会面临的巨大挑战，超级计算机今后的开发热点仍然是万亿次以上的系统，短期目标是百万亿次，长期目标是千万亿次甚至更高。

当前，超级计算机的性能持续以高于摩尔定律的速率（每两年性能提高三倍）发展。2002 年 11 月公布的全球TOP500 超级计算机[5]中前十名的LINPACK峰值性能为 82419.00Gflops，1997 年 11 月公布的全球TOP500 前十名的LINPACK峰值性能为 4102.80Gflops。五年的时间里LINPACK峰值性能提高了20.09 倍。

21 世纪高性能计算的一个重要趋势是与网络特别是网格技术结合。网格的主要特征是：资源共享，动态配置，协同工作，不存在任何集中控制；使用标准、通用、开放的协议和接口；高服务质量，包括响应时间、流量、可用性和安全性。计算网格（Computational Grid）也称元计算（Metacomputing）、远程计算（Distance Computing），是基于高性能互联网络（Internet Ⅱ）实现的高端计算技术，众多计算资源通过高性能互联网广域连接构成高端计算环境，为科

[b] 接收面积单位：平方米；CCD像素单位：百万

技人员和普通百姓提供更多的资源、功能和交互性，让人们透明地使用计算、存储等资源。

近年来，互联网技术也得到了飞速的发展。仅从我国互联网出口带宽和上网人数的增长就可以看出近十年互联网的迅猛发展。

- 国际出口带宽

根据中国互联网络信息中心（CNNIC）的统计报告[6]，截止到 2002 年 12 月 31 日，我国国际出口带宽的总容量为 9380M，是 1997 年 10 月出口带宽 25.408M的 369.2 倍。

- 上网用户人数

截止到 2002 年 12 月 31 日，我国的上网用户总人数为 5910 万人。同 1997 年 10 月第一次调查结果 62 万上网用户人数相比，5 年间我国上网用户人数增长了 95.3 倍。

高性能计算和互联网技术的快速发展，使得在不同地点间进行天文数据的交换与传输成为可能，使得世界各地的天文学家都能够访问和使用这些数据，从而具有巨大的科学产出的潜在意义。

互联网时代的天文数据有着其他学科数据所无法比拟的特点：

1. 天文数据绝大部分是开放数据。天文界的传统是所有的观测数据在一年后向公众开放，使得观测者有时间进行数据分析和发表早期结果，也使其他天文学家可以有机会使用这些数据。国际上的许多大型天文观测项目的观测数据都会及时在互联网上公布，这为数据共享提供了良好的基础，在各类学科中是独一无二的。
2. 天文数据的数据量非常大。现有的以及即将实施的天文项目每天都会产生数 GB 甚至 TB 的数据，天文数据中心的存贮容量已经达到数 TB，并开始向 PB 扩展。
3. 天文数据有比较好的归档，并提供互联网服务。当前，世界上已经有多家天文数据中心在天文数据归档方面做了大量的工作，并取得了很好的应用，甚至各类天文文献也是在线网络服务的。
4. 天文数据的格式多种多样。天文数据的内容主要有星表、星图、光谱等，数据的格式则各种各样，其内部格式都依不同的天文观测项目而变化。
5. 天文数据是全波段的数据。从 γ 射线、X 射线、紫外、光学、红外到射电波段都有观测项目在进行。这些数据是高度相关的，需要在高维参数空间内进行研究。

6. 新的观测数据总能带来全新的现象或规律的发现。原北京天文台研究生樊晓辉利用 SDSS 巡天数据发现迄今为止最高红移的类星体就是很好的例子。

随着众多先进的地面与空间天文设备的投入使用，特别是大规模CCD探测器的使用，使得观测数据量急速增长。例如目前哈勃空间望远镜（HST）每天大约产生 5GB[c]的数据；我国正在建造的大天区面积多目标光纤光谱望远镜（LAMOST）也将每天产生 3GB的数据[7]；而美国计划建造的“大口径综合巡天望远镜（LSST）”，又称为“暗物质望远镜”，每天的观测数据将达到18TB的量级[8]!

如此巨大的数据产出，在天文学历史上第一次使天文学家得到的数据多得用不了，使天文学进入一个数据富庶时代。以往那个辛辛苦苦观测许久但数据还是不够用的年代一去不复返。

除了数据量的快速增长外，天文观测的方式也发生着变化。当前天文学的观测方式主要有两种类型：定点观测和巡天观测。

定点观测就是为了完成特定的研究课题而利用观测设备对预先选定的天体进行观测，数量一般不多，很少能超过一百个。研究过程如图 1.2 所示。首先，天文学家根据课题的需要选定观测目标。然后在申请到的望远镜时间里对目标天体进行观测，得到观测数据。之后，利用自己计算机系统中的处理工具对这些数据进行分析处理。最后天文学家分析处理结果，写成文章发表。

定点观测的好处是可以对少数天体进行高精度的观测和细致的研究。比如，通过对恒星进行高分辨率光谱观测可以得到天体高精度的化学丰度、视向速度、引力常数等物理参量。缺点是效率低，无法实现需要大规模样本的研究课题。以光谱观测为例，一个晴夜最多只能完成数十颗恒星的观测。我国国家天文台兴隆观测站的 2.16 米望远镜的工作方式就是典型的定点观测。从伽利略首先将望远镜指向天空开始，这种观测方式已经延续了 400 多年。直到今天，在观测天文学上仍占有重要的位置。但是随着望远镜制造技术和探测器技术的快速发展，定点观测与现代观测设备强大的观测能力越来越不协调。传统天文学研究方法越来越跟不上现代科技的前进步伐。在这种背景下，巡天观测开始出现并正在取得越来越重要的地位。

巡天观测是对整个天区或者很大面积的天区按照预先设定的观测计划进行观测。观测流程如图 1.3 所示。首先制定巡天战略和输入星表，然后按照预定的观测计划进行观测得到原始观测数据。经过分析处理后产生巡天观测数据产

[c] GB：10^{10}字节

品，比如巡天星表、星图库、光谱库等。由于巡天观测的产出数据量一般都非常大，传统的交互式处理无法满足处理要求，必须利用自动处理工具进行处理形成数据产品；利用数据挖掘工具对巡天数据进行分析，产出研究成果。在数据挖掘的过程中往往会发现一些特殊的数据。对这样的天体进行后续定点观测又可以产生新的研究成果。这种成果往往是激动人心的。此外，巡天观测得到的星表还可能被其它巡天观测项目利用来提取输入星表进行新一轮的巡天观测。

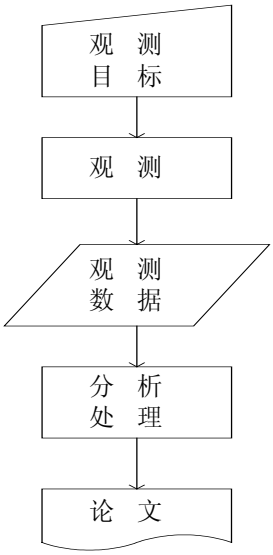


图 1.2 定点观测研究模式

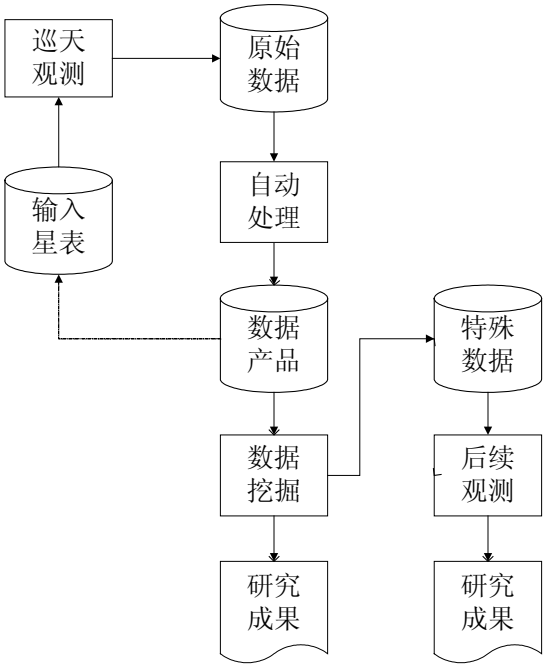


图 1.3 巡天观测研究模式

先进的地面和空间天文台进行大规模的巡天观测，产生质量均匀、标准统

一的海量数据。每个这样的巡天数据库从其本身来说都是非常有价值的，每个波段上的观测都带来了有关天体本质的重要信息。然而同样一个天体在不同波段上的表现可能是完全不同的[9]。比如，蟹状星云的光学图像显示出了电离氢的分布，射电图像显示了中性氢的分布，红外图像显示了尘埃和分子云的分布，而X射线图像显示了高温（千万度）热气体的分布和其中的所存在的中子星，如图 1.4。要研究这类天体的物理过程，就必须结合几个波段上的数据来一起进行分析。

天体辐射是全波段的辐射，从射电到红外、光学到紫外、X 射线甚至到γ射线。随着电磁波波长的变化，它们的能量不同，它们的形成机制也不同，它们所包含的宇宙奥秘也不一样。但由于技术的制约，人类只能从电磁波谱中的某个或者某些波段对这些辐射进行观测。这些观测所产生的数据只能包含天体部分的信息。从这些观测数据得到的对天体的认识往往就像“盲人摸象”一样也是片面的。随着人类技术的进步，天文观测逐渐进入多波段时代，但尚不能说是全波段，因为有些波段我们目前仍然没有能力去涉足。综合利用多个波段的观测数据，我们就能得到关于天体更全面的理解。

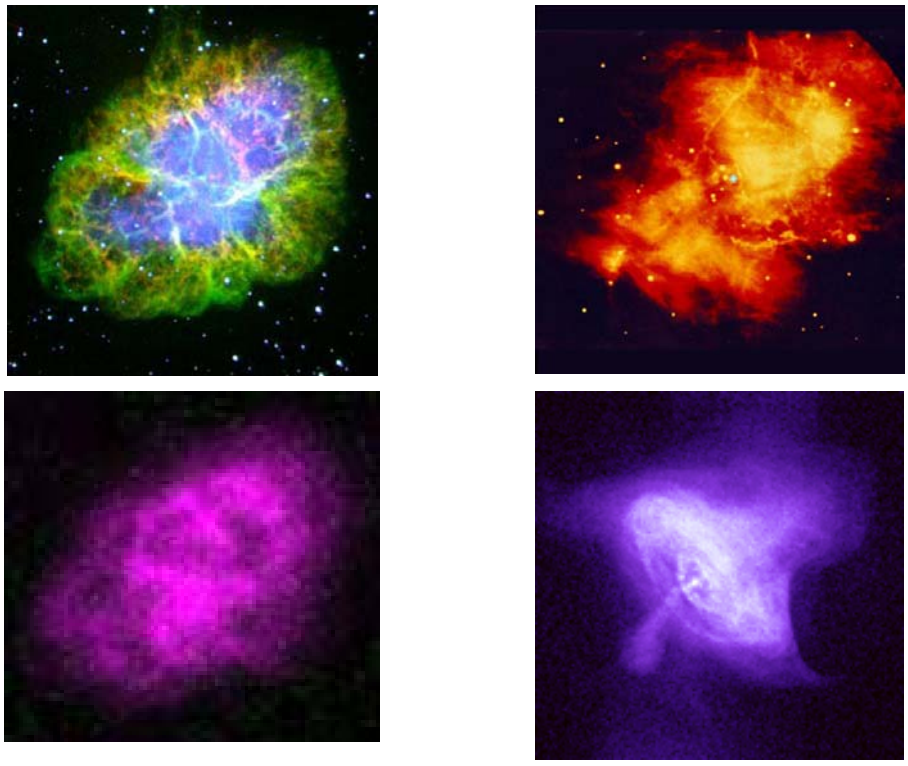


图 1.4 蟹状星云的多波段图像[d]

[d] 左上：光学、左下：射电、右上：红外、右下：X射线

1.2 虚拟天文台

如果说利用γ射线巡天、X 射线巡天、紫外巡天、光学巡天、红外巡天和射电巡天所得到的观测数据，用适合的方法对数据进行统一规范的整理、归档，便可以构成一个全波段的数字虚拟天空；而根据用户要求获得某个天区的各类数据，就仿佛是在使用一架虚拟的天文望远镜；如果再根据科学研究的要求开发出功能强大的计算工具、统计分析工具和数据挖掘工具，这就相当于拥有了虚拟的各种探测设备。这样，由虚拟的数字天空、虚拟的望远镜和虚拟的探测设备所组成的机构便是一个独一无二的虚拟天文台（Virtual Observatory，VO）。由此可见，虚拟天文台是在互联网时代里天文学发展的必然产物。

巡天观测带来了巨大的科学发现的潜力，对这些巡天数据的联合使用，将涌现出全新的、无法预见的、意义重大的科学产出，这是一种仅靠单独使用其中某一部分数据所不能产生的新科学。科学数据的获得、组织、分析和传播是持续而坚实地发展科学技术的基本要素。因此，投入一定的人力、财力、物力将所有符合特定规范的数据整合到虚拟天文台中，其科学意义是传统天文台所无法比拟也无法替代的。

虚拟天文台将使天文学取得前所未有的进展，它将成为开创“天文学发现新时代”的关键性因素[10]。虚拟天文台将是独一无二的，它将TB甚至PB[e]量级的数据库、波长遍及从γ射线到射电波段的数十亿个天体的图像库、高度复杂的数据挖掘和分析工具、具有数千PB量级容量的存储设备和每秒运算次数达到万亿次的超级计算设备、以及各主要天文数据中心之间的高速网络连成一体；它使世界各地的天文学家可以快速查询各个PB量级大小的数据库；使埋藏在庞大星表和图像数据库中的多变量模式可视化；增加发现复杂规律和稀有天体的机会；鼓励研究团体间的实时合作；允许进行大规模的统计研究，首次使数据库的内容可以和复杂精密的数值模拟结果进行对比。虚拟天文台将促进我们对许多决定宇宙演化的天体物理过程的理解。它会用更经济的投资产生新的和更好的科学。虚拟天文台将作为一个协调性的和操作性的机构来促进新型的工具、协议和合作方面的发展，以充分实现现代天文数据库的科学潜能，从而将成为“天文学发现”的推进器。

1.2.1 VO 的科学目标

目前，天文学家确定的虚拟天文台的主要科学目标是：

1）多观测参数高维空间的探索：将各个巡天数据统一到虚拟天文台中，将会有更广泛而复杂的应用。这些数据能提供全天在十多个不同波段上的信

[e] TB：10^{12}字节；PB：10^{15}字节

息，在多维空间里展示整个天空的真实面貌。可以说，多套巡天数据在虚拟天文台中的完美结合，将会得到更加完善的真实的宇宙图像（多层次的、大尺度的、系统性的等等）。

2）稀有天体与新型天体的发现：目前通过巡天来寻找稀有天体（如高红移类星体、褐矮星等）的项目正在蓬勃发展。假如某种有趣的天体或现象出现的概率是百万分之一或一亿分之一，那么就需要几百万或几亿个样本才有可能发现。这样，在海量数据中进行彻底的宇宙探索来寻找稀有的未知类型天体便具有更加诱人的前景。因此，虚拟天文台将会利用其独有的数据资源和计算资源促进新的天文发现。

3）新兴的科学领域：虚拟天文台对任何要求融合各类数据来研究天文现象的课题都具有重要的影响。虚拟天文台的出现会大大促进多波段天文学的发展，不同波段的巡天数据的联合可以从更深层次来探索宇宙；同时，虚拟天文台会推动各种各样令人兴奋的科学探索，如活动星系核和星系团的多层次研究、低表面亮度星系的形成和演化的研究、星系结构的研究等；虚拟天文台的出现还将促进统计天文学的兴起，如宇宙大尺度结构和银河系结构的定量分析、各种天体（特殊种类或特殊性质的恒星或星系、活动星系核、星系团等）完备样本的建立与研究，等等。虚拟天文台的建立可以使天文学研究在数量和质量上得到充分地提高。

4）数据挖掘技术：从海量数据中发现稀有的天体或现象, 或者发现以前未知种类的天体或新的天文现象，或者根据数据来区分不同类型的天体等，都需要充分运用在信息科学中迅速发展的数据挖掘和知识发现技术。数据挖掘技术在虚拟天文台中的应用，将使任何地方的天文学家在不依赖于大望远镜的情况下就可以做出一流的工作，而这种研究方式完全不同于传统的天文学研究。运用数据挖掘技术可以有效地解决天文学中的“数据雪崩”问题，这对天文学发展是至关重要的。

虚拟天文台的发展壮大和普及将会使得实测天文的研究模式再次发生重大变化，从巡天研究模式升级为 VO 研究模式，如图 1.5 所示。

各种天文研究资源，包括巡天观测数据、个人观测数据、天文文献、计算资源、存储资源、各种软件工具，都以某种统一的服务模式被无缝的汇集在 VO 系统中。天文学家只需登陆到 VO 系统便可以享受其提供的丰富资源和强大的服务，使自己从数据收集、数据处理这些繁琐的事务中彻底摆脱出来，而把精力集中在自己感兴趣的科学问题上。

1.2.2 VO 的技术目标

VO 的近期目标：实现世界上主要巡天观测数据的统一性访问并提供对星

表、星图等数据的基本处理服务。

VO 的长远目标：把世界上主要的天文研究资源，包括观测数据、模拟数据、多媒体数据、天文文献、数据处理工具、天文观测设备、计算/网络/存储等资源的无缝融合，使 VO 真正成为一个数据密集型的在线研究平台。

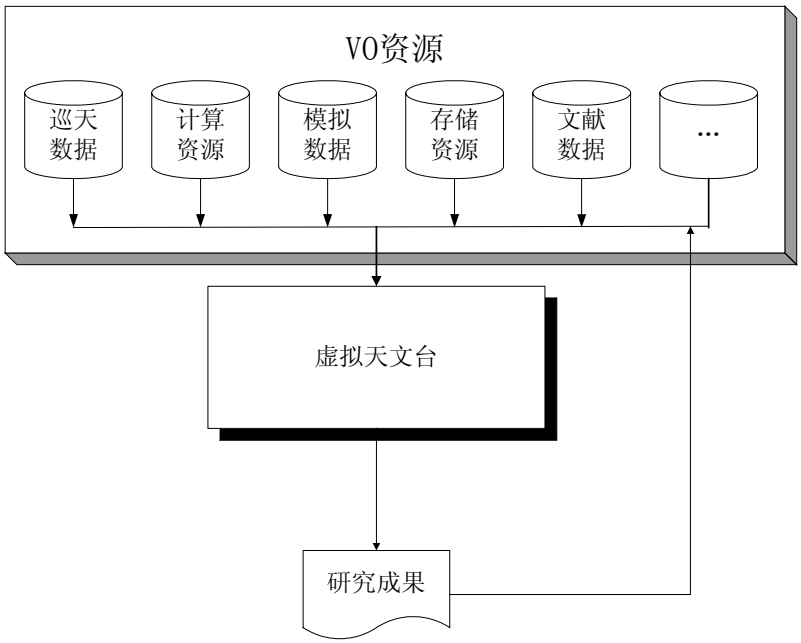


图 1.5 虚拟天文台研究模式

数据与处理工具的关系最早是“数据/天文学家”，巡天观测时代是“数据/程序”，在 VO 研究时代这种关系将转变为“数据/VO 服务”。

虚拟天文台是一个相互之间能进行互操作的数据集和软件工具的集合，它们通过互联网给天文学家提供一个可在线进行天文科学研究的科研环境。

1.2.3 VO 的收益

虚拟天文台的好处是显而易见的，主要表现在如下几个方面：

- 天文望远镜等观测设备得到的数据能被不同的用户出于不同的目的重复使用，提高了昂贵的观测设备的科学效益；
- 观测数据按照统一的方式管理起来有利于长期的保存和利用，使科学数据的价值最大化；
- 虚拟天文台可被全球各种各样的群体访问，包括那些没有经济能力建造和运行大型观测设备的群体，能大大促进发展中国家和不发达国家的天文研究；
- 虚拟天文台是非常好的公众教育设施，VO 是面向全球、全社会的。它让普通公众和天文学家一样都能接触到真实的天文资源和服务。这会

激发大众对天文学的兴趣，提高大众的天文水平，促进天文学的普及。

1.3 VO 国际研发现状

自从美国国家科学院天文学及天体物理学发展规划委员会在题为“新千年的天文学和天体物理学”的未来十年发展规划中把建立国家虚拟天文台(NVO)作为最优先推荐项目[11]，提出虚拟天文台的概念后，各国天文学界迅速响应，纷纷提出了各自的虚拟天文台计划，在全球掀起了一场虚拟天文台浪潮。当前国际上的主要虚拟天文台研究项目及其得到的资金支持如表 1.1 所示（资料截止日期 2003 年 4 月）。

正在全球掀起的虚拟天文台浪潮			
项目	地区	资金（百万美元）	资助年限
国家虚拟天文台（NVO）[12]	美国	10	5
AstroGrid（UKVO）[13]	英国	7.8	3
天体物理虚拟天文台（AVO）[14]	欧洲	3.9	3
加拿大虚拟天文台（CVO）[15]	加拿大	2.8	2
印度虚拟天文台（VO-India）[16]	印度	1.0	3
德国天体物理虚拟天文台（GAVO）[17]	德国	0.6	2.5
澳大利亚虚拟天文台（Aus-VO）[18]	澳大利亚	0.3	不详
日本虚拟天文台（JVO）[19]	日本	0.3	1
俄罗斯虚拟天文台（RVO）[20]	俄罗斯	0.01	3
韩国虚拟天文台（KVO）[21]	韩国	1.0	3

表 1.1 正在全球掀起的虚拟天文台浪潮

为了对全球范围内 VO 的研究进展有一个总体的把握，下面把各国 VO 计划的主要进展情况作一个介绍。

1.3.1 英国天文网格

在英国，2003 年和 2005 年两个新的数据爆炸点即将到来，即当UKIRT WFCAM、VISTA项目正式上马，每晚将产生上千GB的数据。英国在X射线、太阳、空间等离子体物理等方面处于领先地位。但是随着大量新观测数据的到来，英国天文界面临着许多机遇和挑战，为了保持自己的领先优势，他们提出了Astrogrid（天文网格）来应对这个挑战。AstroGrid是英国e-Science计划[22]中网格计划之一，通过英国粒子物理与天文研究委员会（PPARC）[23]得到资助。该计划开始于 2001 年 9 月 1 日，为期三年。AstroGrid项目在 1999 年到 2000 年间就已经开始酝酿，在PPARC的长期科学回顾中作为与大数据库工作相关的几

个小组之一被提到。

AstroGrid 项目分为 A、B 两个阶段。A 阶段为期一年，主要进行一些预研究工作；B 阶段为期两年，在 A 阶段的基础上进行系统的开发和实施。AstroGrid 项目 A 阶段的工作已于 2002 年 12 月结束，现在 B 阶段的工作已经开始。

项目最初制定的目标是：

- 建立用以处理英国主要天文数据的数据网格；
- 开发高效的数据库查询，数据挖掘工具；
- 提供一个标准的数据库查询和数据挖掘的软件接口；
- 实现同时检索多个数据库的能力；
- 完成一套集成化的天文数据在线分析工具；
- 完成一套在线数据库分析和挖掘工具；
- 实现用户上传算法进行数据挖掘；
- 开放式资源发现技术的研究。

在项目 A 阶段，AstroGrid 完成了许多卓有成效的工作，主要表现在：

- 科学需求制定

通过项目成员内部讨论和与外部科学家的广泛接触，完成了全面的需求列表和科学问题文档。

- 体系结构开发

从科学需求得到使用示例并由此确定 AstroGrid 的体系结构框架。体系结构开发工作仍在进行。目前已经完成了“概念模型”和“服务模型”草案。

- 国际合作

在世界范围内与各种人员和组织广泛接触，其中涉及 VO，网格，e-Science 领域等。与 NVO，AVO 共同创立了 IVOA，与 VO 国际伙伴建立了良好的合作关系；向 Globus 项目提交了 Community Authorization Service （CAS）的相关文档；与 e-Science 其他计划合作开展数据库领域的研究工作。

- 对 VO 相关技术进行评测和调研

完成了数个 AstroGrid 技术演示，比如太阳、射电研究，网格平台的建立等。

- 完成了在线协同作业工具的开发并投入使用

开发了一套功能全面的协作工具，让项目成员方便地分享经验，提出看

法，征求意见，并建立了一个健全的资料库。这些工具已被其他项目的网站采用，比如 IVOA、Aus-VO 等。

项目进行过程中，他们对项目的目标和实施方案有了进一步的认识。在 A 阶段总结报告中 AstroGrid 对项目的目标进行了修订。修订后的目标是：

科学目标：

- 改进在线天文研究的质量、效率、易用性、速度、投资回报率；
- 实现不同来源天文数据无缝透明的比较与融合；
- 消除跨学科研究工作中的数据分析障碍；
- 使数据密集型科研工作尽可能的容易开展并实现尽可能强大的功能。

主要的技术目标是：

- 与 IVOA 伙伴一同开发被国际认可的数据、元数据、数据交换和数据履历（provenance）标准；
- 为数据服务开发软件基础设施；
- 建立能为 AstroGrid 和主要数据中心共享的物理网格资源；
- 建立和维护 AstroGrid 服务与资源注册；
- 基于英国主要数据库系统和天文学家的使用要求建立真正可用的虚拟天文台；
- 为 VO 系统提供用户界面；
- 通过自主开发或者引进等方式为用户提供一套在 VO 环境下使用的科学工具；
- 为英国在 VO 领域树立领导地位。

最终目标是基于英国国内的天文资源开发出真实可用并与网格体系兼容的 VO。

1.3.2 澳大利亚虚拟天文台T

2002 年 9 月 18 日，澳大利亚虚拟天文台（Aus-VO）网站正式对外开放，当时项目的人员队伍尚未建立。2003 年 3 月，Aus-VO 全新的网站对外服务。

Aus-VO 对自己的定位是：作为国际虚拟天文台（IVO）的一部分，实现其在澳大利亚的部分。IVO 将把世界上主要天文台的数据联系起来形成一个分布式的数据库，让所有的用户都能接触到这些数据，同时提供强有力的科学工具。

项目的发起单位是澳大利亚望远镜国家设施（ATNF）[24]组织，它隶属于联邦科学与工业研究组织（CSIRO）[25]，负责澳大利亚射电天文学研究。CSIRO是澳大利亚的国家科研组织。ATNF管辖的望远镜有位于Narrabri 的射电

阵和Parkes以及Mopra的射电望远镜。 这些望远镜都是VLBI的组成成员。

Aus-VO 旨在完成 IVO 在澳大利亚的部分，主要工作包括：

- 为实现 IVO 系统内的透明访问建立相关的数据标准、格式、压缩技术、协议；
- 为望远镜的观测数据开发自动处理软件；
- 作为澳大利亚天文数据的主服务器建立全国性的分布式数据库系统并配以相应的计算能力；
- 建立连接国内数据提供者和用户以及海外数据提供者的高速网络；
- 扩展不同数据集间交叉证认和可视化的能力；
- 升级数据提供设施以保证提供高质量的数据。

Aus-VO 为 IVO 贡献的主要数据集将来自：

- ATCA ： 澳大利亚射电阵公开的观测数据；
- HIPASS ： HI Parkes 全天巡天数据 ；
- SUMSS ： 悉尼大学 Molonglo 巡天数据；
- AAT ： 英澳天文台数据。

澳大利亚的贡献将主要表现在：

- 将澳大利亚的数据共享给世界；
- 为澳大利亚的研究者提供访问海外数据和工具资源的数据网格；
- 开发建立 IVO 所必须的软件、技术、标准、格式；
- 对澳大利亚的设施进行升级以便能提供满足 IVO 需求的高质量数据。

他们希望通过 Aus-VO：

- 使得澳大利亚的天文学家能作为 VO 前沿用户进行基础科学研究；
- 促进国内交叉学科的研究发展；
- 在国际上巩固澳大利亚天文学的地位。

目前 Aus-VO 正与 AstroGrid 联合为 2003 年 7 月在悉尼举行的第二十五届 IAU 大会准备一个技术演示。

1.3.3 欧洲天体物理虚拟天文台

天体物理虚拟天文台（AVO）致力于建立欧洲天文界虚拟天文台所需技术的研究。在为期三年的 A 阶段研究中，AVO 由欧盟和六个欧洲组织联合资助。六个合作组织是：欧洲南方天文台（ESO）、欧洲太空局（ESA）、AstroGrid、CDS、法国 Louis Pasteur 大学、法国巴黎天体物理研究所 TERAPIX 天文数据中心、英国曼彻斯特大学 Jodrell Bank 天文台。

他们将历时三年来进行虚拟天文台的设计和应用的研究（A 阶段）。该提议于 2001 年 2 月份提交，已批准并得到近三百万英镑的启动资金。工作组于 2002 正式启动，主要集中在科学要求，科学互动和新的技术领域，A 阶段将为 B 阶段（预计需要六千万英镑）的完全可操作的虚拟天文台的实现打下坚实的基础。

A 阶段主要发展目标有：

1. 制定细致全套的科学要求文档，用以设计、应用、操作虚拟天文台，配套以分布的、逐步升级的网格计算设施。
2. 构造恰当的平台和界面，把来自地面和空间观测设备的天文数据库有机地统一到虚拟天文台的数据库中。
3. 推动天文学家与计算机软硬件工程师的合作。在虚拟天文台和网格技术的主要领域（例如计算机网络、数据库设计和存储管理等方面），发展与欧洲工业界的对话。

A 阶段的主要工作包括：

- 详细制定 AVO 的科学需求；
- 数据互操作问题的研究；
- 必要的网格和数据库技术的测试。

A 阶段的工作主要集中在研究、演示和试验床的搭建，目的是为 AVO 确定详细的科学、功能和操作需求，进而形成一个详细的 AVO 建设计划。B 阶段的工作将主要涉及：

- 概念上的系统设计；
- 科学操作与操作运行计划；
- 完整的科学需求和用户需求；
- 项目管理计划和经费落实；
- 项目进度安排；
- 国际协调与参与；
- 网格设施的使用；
- 欧洲研究网需求和国际连接需求。

AVO 的参与者计划与欧洲所有的天文数据中心一起推动一个综合的项目，也就是 B 阶段的 AVO，争取在 2007 年底实现一个正式运作的欧洲虚拟天文台。AVO 还要与国际 VO 伙伴如美国、加拿大、澳大利亚等加强合作与交流，最终创建全球性虚拟天文台。

1.3.4 加拿大虚拟天文台

加拿大虚拟天文台（CVO）的基本设计目标是为科学家开发一套系统。这个系统将不仅仅能让天文学家找到数据，还必须能产出科学的结果。这些结果是通过其他途径无法或很难完成的。

从技术的角度看，CVO将利用最好的技术和实践经验来实现自己的目标。CVO将不被“向前兼容”问题所困挠，不会将以前加拿大天文数据中心（CADC）[26]的服务全部囊括到新的系统中。

CVO 的基本设计目标是实现全波段的科学研究。他们的设计思路是：通过抽象将技术细节和科学分离，将技术隐藏起来从而实现对数据的统一访问；通过定义一个足够通用的数据模型来满足所有天文学家的需要。他们认识到不同波段研究工作的技术细节很不相同，可以开发一个系统对 VO 中不同类型的内容进行抽象从而去掉这些技术细节和复杂性。

目前，CADC 设计了一个通用的科学数据模型和一个访问科学数据仓库的 API。他们正利用 Java 程序设计语言开发一系列的 Jini 服务，搭建 CVO 的原型。

此外，加拿大天文数据中心决定发展从大的科学数据库中进行数据挖掘的工具，提出发展数据挖掘技术的提案。CADC 在过去十年在数据存储、加工、存档、分布等方面处于领先地位，是首家使用 WWW 界面服务器，同时又是第一个使用 CD-ROM 技术的场所，而且与国际上许多单位都有过成功的合作关系，在计算机软件方面成果斐然。在各国争相出笼虚拟天文台的形势下，加拿大天文学界为保持其在国际方面的领先优势，增加该国天文学家的工作效率，提出了在加拿大天文数据中心发展数据挖掘中心的提案，为逐步过渡到全球性虚拟天文台做准备。计划的主要目标包括：

1. 加强加拿大在发展 TERAPIX 数据管道方面的作用，包括图像处理、图像分析和数据库建立。
2. 对 TERAPIX 和所有其它 MEGACAM 数据，建立 TB 量级的数据库。
3. 提高对 TERAPIX 数据进行数据挖掘的能力，如对 5 千万个天体快速查询、提取、可视化、统计分析、与其他的数据库交叉证认，创建多波段数据库交叉证认的工具雏形。
4. 提高利用用户提供的子程序对 TERAPIX 的图像进行再现的能力。

1.3.5 印度虚拟天文台

印度虚拟天文台（VO-India） 将把天文学家和技术专家联合在一起，为IVO的建设做出贡献，在虚拟科学领域为印度树立先行者的地位。印度交通与信息技术部提供了项目的大部分资金；软件工程师主要来自Persistent Systems公司（PSPL）；基础设施、计算设施、以及其他资源主要来自（印度）大学联

合天文学与天体物理中心（IUCAA）[27]。

他们今后几年的主要工作是：

- 在数据检索和提取方面进行研究和开发；
- 开发能高效使用数据的软件；
- 让天文学家和其他感兴趣的科学家能使用虚拟天文台的数据；
- 使虚拟天文台技术也可以应用于其它涉及海量数据的领域，比如遥感、人口学、生物信息学、健康医疗等。

项目的开展将通过联合的方式进行：

- 大学和研究所中的天文学家和其他科学家的联合；
- 工业界计算机软件开发人员的联合。

项目的收益：

- 印度的天文学家可以接触到一元化的天文数据，以及用于处理这些数据的软硬件资源；
- 大学的教师和学生可以接触到 IVO 的丰富资源，接触到一个全新的研究天地；
- 促进科学界与工业界的合作；
- VO 技术可以被其他领域所借鉴；
- 为印度在虚拟天文台领域确立先行者的地位；
- 虚拟天文台是进行公众教育宣传的好工具。

当前的主要活动：

- VOTable 的 C++解析器

这是一个用来访问VOTable文件的C++函数库，整体结构与CFITSIO函数库[28]非常相似。

- VOPlot

这是一个用 JAVA 开发的可用于 VOTable 格式的天文数据可视化工具。VOPlot 可以独立运行于用户自己的计算机上也可以与 CDS 的 VizieR 整合在一起通过 WEB 进行调用。

VOPlot 是刚刚开发出来的程序。在未来的版本中将增加更多的高级画图、高维画图和高级统计功能。

- FITS 管理器

这是一个基于 WEB 的 FITS 文件生成、查看和编辑工具，同时可以把

FITS 图像转换为其它格式的图片。在未来的版本中将开发成一个独立的软件，既可安装在用户自己的计算机上也可以安装在网络服务器上。

1.3.6 意大利天体物理研究数据网格

意大利天体物理研究数据网格（IDGAR）[29]旨在证实利用特殊化的网格节点，比如观测节点、计算节点、存储节点，向科学界提供一个分布式多功能研究环境的可行性。意大利天文界将以IDGAR为桥梁参与国际上虚拟天文台建设的努力。

IDGAR 以意大利商业、工业、政府、科学、技术网格 （IG-BIGEST）为基础，由意大利国家天体物理研究所（INAF）负责天文部分的建设。主要工作内容包括三个演示：天文数据库的访问与浏览、VST 图像的提取和处理、异地望远镜观测的分布式监测以及类星体目标的实时观测。

IDGAR最初计划中的三个节点是： INAF[30]所属的三个天文台，它们是Naples、 Padova 和Trieste。三个节点分工合作完成IDGAR的技术和科学试验。

- Padova 和 Trieste 负责实现 TNG LTA 数据和 GSC-II 数据的访问服务以及与其他数据提供者之间的互操作。
- Naples 负责向天文学家提供远程访问 VST/OmegaCAM 数据处理 pipeline 的有效工具。
- Trieste 负责提供与网格兼容的观测设施观测远程监测与目标管理系统。

IDGAR 的发展计划主要体现在节点的扩展和应用的扩展上，主要目标是实现：

- 高能数据的访问；
- 射电数据的访问；
- 小规模数据集与数据库的访问研究；
- 与其他数据处理和科学应用的集成；
- 可视化和数据挖掘方面的应用。

IDGAR 将作为桥梁，沟通意大利天文界与国际天文界，共同实现 IVO 的设想。在与 IVO 合作的过程中，IDGAR 将特别的与 AVO 进行如下方面的合作：

- TNG Long-Term Archive （LTA）的访问；
- ASDC 高能数据的访问；
- 射电数据的访问；
- 可视化服务；

- 机器学习领域的研究。

1.3.7 日本虚拟天文台

日本虚拟天文台（JVO）由日本国立天文台（NAOJ）[31]的天文数据分析中心牵头，于 2002 年 4 月被确立为日本国立天文台的正式研究项目。

JVO 的主要合作伙伴有：

- 日本国立天文台
- Ochanomizu 大学
- 富士通公司

JVO 目前的成员组成包括五名天文台人员，一位 Ochanomizu 大学访问教授和数位富士通公司的系统工程师。他们希望 JVO 能在 21 世纪的天文学发展中做出重要的贡献。

经过一年多的努力，JVO 已经取得了让 IVOA 成员瞩目的成果。其中最重要的是开发了日本虚拟天文台查询语言（JVO QL），取得了虚拟天文台查询语言（VOQL）研发的领导权。

2002 年 JVO 成功了定义了一个统一的查询语言以实现对天文数据库的访问，JVOQL。作为 SQL 的一种扩展，JVOQL 可以实现对存储于数据库中的星表和图像数据的查询检索。JVO 努力通过单一的用户接口实现对多个数据服务器上星表数据和图像数据的检索提取功能。

JVO 利用网格技术建立了一个原型平台来验证 JVOQL 对多数据库查询的支持。这个平台利用 Java 实现用户接口，利用 GT2 进行相关服务开发，以便可以与 OGSA 兼容。JVO 的大多数功能都是利用自由软件实现的。JVO 计划建立自己的数据库系统，把从远程服务器上提取的数据存储到本地数据库中。

JVO 的未来一段时间的工作将主要集中在：

- OGSA 的广泛应用
- 数据分析工具的开发
- 用户访问控制

1.3.8 韩国虚拟天文台

韩国虚拟天文台（KVO）项目开始于 2003 年 2 月，挂靠于韩国天文台（KAO）[32]，这是韩国的国家天文台。KVO正在利用韩国的观测设备和天文学家收集的数据建立一个国家级的天文数据库。KVO与韩国天文数据中心（KADC）一同工作，每年得到大约一百万美元的支持，直到 2005 年。

KVO 成员来自：

- 韩国天文台
- Chungbuk 国立大学
- Kyungpook 国立大学
- Seoul 国立大学
- Yonsei 大学

他们目前的主要工作集中在：

- 组织研发队伍
- KVO 原型的开发
- 天文数据的整理

1.3.9 美国国家虚拟天文台

美国国家虚拟天文台（NVO）是完全构筑于天文数据库和网络信息技术之上的，依靠强大的计算机与网络的软硬件支持，并在开发新的数据分析技术和知识发现工具。NVO 利用最新的计算机技术、数据存储和分析技术，将来自地面和空间观测站的天文观测数据有机地统一在一起，目标是最大化来自这些观测数据的科学潜力，将数据以统一标准的格式提供给专业研究者、天文爱好者和学生。NVO 项目已通过美国国家科学基金立项，获得一千万美元资金资助，历时五年。

2001 年 8 月美国国家科学基金会（NSF）计算机与信息科学部，注意不是天文科学部，开始为“NVO 架构”项目提供为期五年的资助。项目协作的目的是 NVO 架构，项目的产出不是一个 VO，而是研究这样的工程该如何建设。

NVO 的研发阵容庞大，成员来自许多著名的天文台、大学、国家实验室和公司。主要的组织成员有约翰霍布金斯大学，加州理工学院，美国国家光学天文台，太空望远镜科学研究所，加利福尼亚大学，宾夕法尼亚大学，费米国家实验室，NASA，施密松天体物理天文台，卡耐基梅隆大学，美国海军天文台，伊利诺斯大学，匹兹堡大学，Lawrence Livermore 国家实验室，南加州大学，威斯康辛大学，芝加哥大学，微软研究院，圣地亚哥超级计算中心，等单位。

此外还包括欧洲空间局，加拿大天文数据中心，CDS，爱丁堡皇家天文台，贝尔法斯特大学，澳大利亚望远镜国家设施（ATNF），东京大学，欧洲南方天文台等国际伙伴。

自项目立项以来他们取得的主要科学和技术成就包括：

VO 标准的制定：

- 锥形检索（ConeSearch）

- 简单图像访问协议（SIAP）
- VOTable
- 统一内容描述标准（UCD）

VOTable 标准已被国际虚拟天文台界认可，并在此基础上开发出了标准的服务：锥形检索。

软件与服务的开发：

- VOTable 解析
  - JAVOT：VOTable 的 Java 语言解析器
  - Perl 语言解析器和操作程序
  - Perl 语言格式化和打印程序
  - C++ 解析器
  - SAVOT ：VOTable 简单访问程序
- VOTable 浏览
  - VOPlot：VOTable 数据画图程序
  - VOTool：VOTable 数据可视化与编辑程序
  - Treeview：天文资源浏览程序
- 服务注册
  - ConeSearch 注册
  - 简单图像访问服务注册
- Web 服务
  - NVO 正在进行基于 SOAP 协议的 WEB 服务的开发

此外，NVO 对 VO 概念的宣传、推广做了大量的工作，作为创始成员与 AVO、AstroGrid 一同创立了 IVOA，并承担了 IVOA 社区内部许多协调管理工作。

在 2003 年的美国天文学会（AAS）201 次会议上，NVO 成功的展示了自己的三个演示（demo）。

1） γ 射线暴事件追踪

利用远程服务提供的标准协议，比如 ConeSearch、SIAP、VOTable，和标准的语义编码（比如 UCD）来进行数据提取和转换。在多个波段上收集关于某个天区或者有趣的瞬间事件，比如 γ 射线暴，的图像、星表等各种相关数据。这个 demo 重点测试了 VOTable 的使用。

2）褐矮星候选体搜寻

对 SDSS EDR 光学巡天星表和 2MASS 红外巡天点源星表进行交叉证认来

搜寻褐矮星的候选体。这个 demo 重点测试了 NVO 对 ConeSearch 以及交叉证认功能的支持。

3）星系形态学分析

利用 VO 提供的：

- 通过标准接口对分布式异构星表和图像数据的访问能力，
- 建立在标准数据交换格式（FITS、VOTable）上的数据整合能力，
- 网格环境下的高性能计算能力，

对星系的表面亮度、集中指数(Concentration Index)和不对称指数(Asymmetry Index)进行了计算，并利用这些计算得到的参数与其他数据，比如测光数据、本动速度、团中位置、星系团大尺度结构，进行联合分析，来进行星系团动力学研究和大尺度结构下的星系演化研究。这个 demo 重点测试了 NVO 正在开发的简单图像访问协议。

1.3.10 俄罗斯虚拟天文台

俄罗斯虚拟天文台（RVO）是在俄罗斯科学院天文研究委员会的强烈倡导下，与俄罗斯天文数据中心和特殊天体物理天文台联合推出的。RVO 给自己的定位是 IVO 不可分割的一部分。

RVO 的主要目标及活动是：

1. 为俄罗斯天文界提供通向世界数据网格的便利接口

当前活动：

- 主要国际天文数据库镜像：ADS，Vizier，INES，VALD，BELDATA
- 文献、星表、软件列表整理
- 数据传输：DLT, DDS, ADR, CD, MO
- 星表的浏览与可视化
- 数据集的回顾与评估
- 用户需求分析

2. 整合俄罗斯和前苏联的数据提供给国际社会并将其接入 IVO

当前活动：

- 提供方便的在线资源直接访问服务
- 离线资源的分发
- 俄罗斯主要天文期刊中电子图表的访问
- 星表、照相底片、期刊的电子化
- 咨询与技术支持

- 所提供资源信息的标准化与统一化

3. 参加为建立 IVO 所需软件、技术、标准、格式的开发与研究

4. 在必要的时候使用俄罗斯观测设备提供远程观测数据

5. 利用国际天文数据加强教育和公众应用

1.4 IVOA，通向未来之路

美国首先提出建立虚拟天文台的计划后，欧洲、英国、德国、日本、加拿大、俄罗斯等国家也相继提出了类似的计划。虽然这些计划来自不同的国家，有着不同的天文和技术背景，但是他们之间有许多非常重要的共同点：每个项目都在寻求数据密集型天文研究的出路，都在力图挖掘现有以及未来海量天文数据的潜力。

每个项目都制定了有自己特色的科学目标，都有自己的技术优势和兴趣。但每个 VO 项目都面临着共同的需求并寻求着相似的解决方案。从国际天文界的眼光来看，这些项目的共同目标就是建立一个国际虚拟天文台（IVO）。如果把这些项目联合起来，共同迎接虚拟天文台所面临的科学与技术挑战，这对 IVO 的建立是很有好处的。

因为 IVO 必须是一个完整的，不同部分能进行互操作的系统。为了实现不同部分之间的互操作性，就必须对许多问题在国际范围内达成一致和认可。这其中最重要的是数据和接口的标准化，此外需要协调的部分还包括软件包、源代码库、开发工具等。为了 IVO 的最终实现还有一些事情要与官方的政策、资金和安全等问题相关。

虽然各国的 VO 项目已就一些互操作性方面的标准达成一致意见，但要取得天文学界更广泛的认可，还需要完成一些更高级的科学和技术演示，同时要公布一个共同的进度安排。

共同进度计划的执行，国际上不同项目间的协调与合作，需要一个机制来实现。正是出于此目的，AstroGrid、AVO和NVO利用 2002 年 6 月在德国 Garching 举行“迈向国际虚拟天文台（Towards an International Virtual Observatory）”国际会议[33]的机会提出了成立国际虚拟天文台联盟（IVOA）[34]的倡议，希望利用IVOA把这些国际努力团结在一起。他们的倡议得到了与会代表的支持。IVOA当即表示成立，当时成员数为八个。

这次会议是由欧洲南方天文台、欧洲空间局、美国宇航局（NASA）和美国国家科学基金会联合资助的。

此后，其他一些国家包括中国在内也提出了各自的虚拟天文台计划。目前（截止到 2003 年 4 月），IVOA 的成员数已经增加到了十二个，组成情况如表 1.2 所示。此外，南非也已经表示对 VO 感兴趣。

1.4.1 使命与进度安排

联盟成立时还为 IVOA 制定自己的使命以及从 2002 年到 2005 年间的进度安排。

当时制定的 IVOA 的使命是：推进国际合作与协作，为建设一个能综合利用国际天文数据的、完整的、能协同工作的虚拟天文台，开发、配置必要的工具、系统和组织结构。

地区	项目名称
Australia	Australian Virtual Observatory (Aus-VO)
Canada	Canadian Virtual Observatory (CVO)
China	Chinese Virtual Observatory (China-VO)
Europe	Astrophysical Virtual Observatory (AVO)
German	German Astrophysical Virtual Observatory (GAVO)
India	Virtual Observatory India (VO-India)
Italy	Italian Data Grid for Astrophysical Research (IDGAR)
Japan	Japanese Virtual Observatory (JVO)
Korea	Korean Virtual Observatory (KVO)
Russia	Russian Virtual Observatory (RVO)
UK	Astrogrid
US	National Virtual Observatory (NVO)

表 1.2 IVOA成员组成

IVOA 给自己制定的 2002 年到 2005 年进度安排是：

- 2002 年 1 月
  - 关于互操作性的首次国际对话
  - VOTable 标准草案的讨论和修订
- 2002 年 4 月 15 日
  - 就 VOTable 1.0 标准达成一致意见
- 2002 年 6 月 10-14 日
  - IVOA 成立
- 2003 年 1 月
  - IVOA 成员首次联合科学演示
  - IVOA 就初步的互操作标准和工具达成一致意见
- 2003 年 5 月
  - WEB 服务发布

- 2003 年 7 月
  - 在 IAU 大会上进行带有国际数据访问和传输功能的科学演示
- 2003 年 10 月
  - 天文学查询语言 AQL 1.0 标准发布
- 2004 年 1 月
  - 带有网格计算和数据存储技术的联合科学演示
- 2004 年 5 月
  - 资源发现 1.0 标准发布
- 2004 年 7 月
  - 公布 2005 年及以后的 VO 开发计划
- 2004 年 10 月
  - 组合 WEB 服务和 Ontology 服务 1.0 标准发布
- 2005 年 1 月
  - 复杂的联合科学演示

为方便研究开发人员讨论问题、交流经验，IVOA 设立了多个技术讨论组邮件列表，如表 1.3 所示。任何对这些话题感兴趣的人员都可以申请加入相应的列表。

讨论组	主题
Hdal@ivoa.net	数据访问层
dm@ivoa.net	数据模型
grid@ivoa.net	网格
interop@ivoa.net	互操作性
net@ivoa.net	网络
radiovo@ivoa.net	射电天文学
registry@ivoa.net	资源元数据注册
semantics@ivoa.net	语义网、知识工程
stdproc@ivoa.net	标准化过程
ucd@ivoa.net	统一内容描述（UCD）
voql@ivoa.net	VO 查询语言（VOQL）
votable@ivoa.net	VOTable/XML
ws@ivoa.net	Web 服务

表 1.3 IVOA 讨论组

此外，为了方便 IVOA 各成员项目间负责人的联络，IVOA 执行委员会开设了一个电子邮件列表“ivoa@ivoa.net”。不过与上面提到的技术讨论组邮件列表不同的是只有各个 VO 项目的主要负责人才有资格加入这个列表。

IVAO 成立以来通过面对面方式或者电话方式已经组织了多次多方讨论，比如：

- 2002 年 10 月, Interoperability Meeting, Baltimore, USA
- 2003 年 1 月, IVOA Meeting, AAS, Seattle, USA
- 2003 年 1 月, Data Modeling Workshop, ESO, Garching, Germany
- 2003 年 3 月, Registry Workshop, e-Science Center, London, UK
- 2003 年 5 月, Interoperability Meeting, IoA, Cambridge, UK

正在计划中的还包括：

- 2003 年 7 月, IAU GA, IAU, Sydney, Australia
- 2003 年 10 月，Interoperability Meeting, CDS, Strasbourg, France

此外，通过邮件列表等方式，各国的 VO 人员就 IVO 涉及的许多技术问题，比如注册、UCD、数据模型、VOQL、VOTable、SIAP、互操作性等，正进行着充分的讨论。

IVOA 的成立促进了各国 VO 项目间的沟通与合作，实现了人力、物力和财力的高效利用，扩大了 VO 在国际天文界的声誉和影响，加速了国际虚拟天文台的建设步伐。这是一条将天文学研究引入新时代的未来之路。

参考文献

[1] 赵永恒. 互联网时代的天文学革命: 虚拟天文台.科学, 2002, 54(2):13

[2] 程景全. 天文望远镜原理和设计——射电、红外、光学、X 射线和射线望远镜. 第 1 版. 北京: 中国科学技术出版社, 2003. 325

[3] Alex Szalay. The Sloan Digital Sky Survey. http://tarkus.pha.jhu.edu/~szalay/powerpoint/sdssv15.ppt

[4] Paul Messina, Alex Szalay. Project Description: Building the Framework for the National Virtual Observatory. http://bill.cacr.caltech.edu/cfdocs/usvo-pubs/files/nvo-proj.pdf

[5] [TOP5000] TOP500 SUPERCOMPUTER SITES, http://www.top500.org/

[6] [CNNIC] 中国互联网络信息中心. 中国互联网络发展状况统计报告（2003. 1）. http://www.cnnic.net.cn/develst/2003-1/

[7] [LAMOST] Large Sky Area Multi-Object Fiber Spectroscopic Telescope. http://www.lamost.org

[8] [LSST] Large-aperture Synoptic Survey Telescope. http://www.lssto.org/lsst_home.html

[9] Szalay A., Gray J. The World-Wide Telescope. Science. 2001(293): 203

[10] Brunner R., Djorgovski S., Szalay A. Towards a National Virtual Observatory. In: Virtual Observatory of the Future. Michigan: 2001, p.343-372
[11] [NAS99] National Academy of Science, Astronomy and Astrophysics Survey Committee. Astronomy and Astrophysics in the New Millennium (Decadal Survey). http://www.nap.edu/books/0309070317/html/
[12] [NVO] US National Virtual Observatory. http://www.us-vo.org
[13] [AstroGrid] VO United Kingdom. http://www.astrogrid.org
[14] [AVO] Astrophysics Virtual Observatory. http://www.euro-vo.org
[15] [CVO] Canadian Virtual Observatory. http://services.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/cvo/
[16] [VO-India] Virtual Observatory India. http://vo.iucaa.ernet.in/~voi/
[17] [GAVO] German Astrophysical Virtual Observatory. http://www.g-vo.org/
[18] [Aus-VO] Australian Virtual Observatory. http://www.aus-vo.org
[19] [JVO] Japanese Virtual Observatory. http://jvo.nao.ac.jp/
[20] [RVO] Russian Virtual Observatory. http://www.inasan.rssi.ru/eng/rvo/
[21] [KVO] Korean Virtual Observatory. http://kvo.kao.re.kr/
[22] [e-Science] UK e-Science. http://umbriel.dcs.gla.ac.uk/nesc/
[23] [PPARC] Particle Physics and Astronomy Research Council. http://www.pparc.ac.uk/
[24] [ATNF] Australia Telescope National Facility. http://www.atnf.csiro.au/
[25] [CSIRO] Commonwealth Scientific & Industrial Research Organization. http://www.csiro.au/
[26] [CADC] Canadian Astronomy Data Centre. http://cadcwww.dao.nrc.ca/
[27] [IUCAA] Inter-University Centre for Astronomy and Astrophysics. http://www.iucaa.ernet.in/
[28] [CFITSIO] HEASARC CFITSIO. http://heasarc.gsfc.nasa.gov/docs/software/fitsio/fitsio.html
[29] [IDGAR] [IDGAR] Italian Data Grid for Astrophysical Research. http://wwwas.oat.ts.astro.it/idgar/IDGAR-home.htm
[30] [INAF] Istituto Nazionale di Astrofisica. http://www.inaf.it/
[31] [NAOJ] National Astronomical Observatory of Japan. http://www.nao.ac.jp/
[32] [KAO] Korean Astronomical Observatory. http://www.kao.re.kr/
[33] [Garching2002] Toward an International Virtual Observatory: Scientific Motivation, Roadmap for Development and Current Status. http://www.eso.org/gen-fac/meetings/vo2002/
[34] [IVOA] International Virtual Observatory Alliance. http://www.ivoa.net

第二章 VO 主要相关技术

虚拟天文台是“科学驱动，技术使能”的前沿科学计划。当前信息科技领域发展势头强劲的三大技术正作为VO的三大支柱，为VO宏伟科学目标的实现保驾护航。这三大技术便是：可扩展标记语言（eXtensible Markup Language, XML）[1]，WEB服务（Web Services）[2]和网格技术（Grid Technology）[3]。为了能更方便的阐述论文后面的内容，本章对这三大技术做一简单的介绍。

2.1 XML

早在 1969 年，国际商用机器公司（IBM）就开发了一种文档描述语言GML用来解决不同系统中文档格式不同的问题。GML是IBM许多文档系统的基础，包括Script和Bookmaster。这个语言在 1986 年演变成一个国际标准（ISO8879），也就是标准通用标记语言（Standard Generalized Markup Language, SGML）[4]。SGML已经成为很多大型组织，比如飞机、汽车公司和军队的文档标准。它是平台无关的、结构化的、可扩展的语言，这些特点使它在很多公司受到欢迎，被用来创建、处理和发布大量的文本信息。

1989 年，欧洲粒子物理研究中心（CERN）的研究人员开发了基于SGML的超文本版本，来给在Internet上共享的技术文件做标记。这种语言最终发展成为一种简化了的SGML应用，这就是超文本标记语言（HTML）[5]。现在，HTML已经成为互连网上信息的标准格式。HTML继承了SGML的许多重要的特点，比如结构化、平台独立性和可描述性。但是同时它也存在很多缺陷：比如它只能使用固定的有限的标记，而且只侧重对静态内容的显示。

随着Web上数据资源的增多，人们越来越不能满足于静态的图片和文本，HTML存在的缺点就变的不可被忽略。于是，万维网联盟（W3C）[6]提供了HTML的几个扩展用来解决这些问题，比如DHTML、XHTML[7]。但最后，W3C还是决定开发一个新的SGML的子集，XML。

XML 的出现是为了解决 HTML 存在的弊病。它保留了很多 SGML 标准的优点，但是更加容易操作和在 WWW 环境下实现。1998 年 2 月 10 日，XML 1.0 成为了 W3C 的推荐标准。2003 年 2 月 10 日，XML 迎来了它的 5 周岁生日。

HTML 和 XML 都是由 W3C 推荐的标准，W3C 的成员认识到随着 Web 的发展，必须有一种方法能够把数据和自身的显示分离开来，这样就导致了 XML

的诞生。但是为什么不直接使用 SGML 呢？原因是 SGML 相当复杂，它的标准超过了 500 页，而 XML 最初的标准就非常简单，只有二十几页。

几个世纪以来，人类在商务活动中已经成功地采用了交换标准化文档的方式，如采购订单、发票、货物清单、收据等。文档能够在商务活动中起作用是因为它并不要求所涉及的各方知道对方的内部处理过程。每条记录所表述的信息正是其接收者所需知道的信息，除此别无其他。

Web 起源于学术界，但推动其飞速发展的是商业。最近出现的电子商务已经引起了广泛的注意，但是 Business to Business （B2B）的商务活动向在线发展的速度却受到限制。例如，通过制造过程的商品流需要自动化。但实际上，依赖于复杂的程序到程序的交互作用系统的运行状况并不理想，其原因是它们所依赖的统一处理过程不存在。像传统商务活动中一样，交换标准化文档的方式也是在线进行商务活动的最佳方式。但是，HTML 并不是为此项任务而创建的；相反，XML 是为文档交换所设计的，而且越来越明显的是，全球化的电子商务将主要依赖于协议流，大量协议将以 XML 文档形式在 Internet 上传送。

XML 作为 SGML 的最小完备子集，继承了 SGML 的强大功能而抛弃了其繁琐的定义。作为一种自描述的数据共享机制，XML 的主要特点如下：

- 自描述性：这个特征允许差异性的存在，计算机可以在没有人为干涉的情况下理解数据的含义。
- 可扩展性：文档通过 XML Schema 或者 DTD（文档类型定义）来定义文档结构，使其他信息系统可以自动了解文档结构。
- 分层结构：保证了信息的层次性。比如一本图书可以有书名、作者、出版社、图书编号；出版社又可以有通信地址、邮编等信息。
- 丰富的链接定义：对应于 HTML 单一的单向单通道链接，XML 提供各种不同的链接方式，比如一对多、多对一和双向链接。
- 多样的样式表支持：XML 将数据内容和它们的表现形式分离，这样既可以只关心数据的逻辑关系，也可以通过样式表来格式化数据的表现，甚至还可以定义自己的个人样式来显示各种不同的 XML 数据。

由于 XML 和 HTML 有着本质的区别，XML 比 HTML 提供了更多对于内容和结构的说明和限制的机制，使得存储、查询、管理 XML 文档相对而言更容易，为基于 WEB 的应用提供了一个描述数据和交换数据的有效手段。XML 的应用从 WEB 网站的内容管理、内容表示起步，已经并正在扩散到信息管理相关的各个领域，比如内容管理发布、电子商务、数据集成、分布式系统集成、系统配置信息描述等。

2.1.1 XML 文档的基本格式

为了比较简略的了解 XML 文档的基本格式，下面以附文 2.1“cv.xml”为例来说明它的主要内容。

这个文件共分为两大部分，前两行是文件头，对文件的属性进行了说明。后面标签“<cv>”和“</cv>”之间的部分是文件的正文，描述了本人的一些基本信息。这其中又包括三部分联系方式、个人信息和发表文章，分别包括在{<contact>, </contact>}、{<personal>, </personal>}、{<publication>, </publication>}之间。标签“<!…>”中的部分是注释。这里已经体现了 XML 语言的可扩展性，文中的许多标签都是自己定义的。

2.1.2 XML 家族体系

为了满足不同群体用户的需要，经过 5 年时间的发展，XML 已经从原来二十几页的简单规范发展为庞大的协议集。之所以有这么快的发展速度，一方面是由于 XML 本身的可扩展性，另一方面是人们对于 XML 的欢迎和重视。

XML非常显著的一个特点就是其可扩展性。不同的学科、不同的应用领域都可以根据自身的特点在其基础上编写各自的标记语言，比如数学标记语言MathML[8]，天文领域也正在开发天文学标记语言AML[9]和天文仪器标记语言（AIML）[10]。

目前 XML 协议集的成员涉及以下几方面的内容：

- XML：eXtensible Markup Language，可扩展标记语言
- XBase：XML BASE
- XLINK：XML Linking Lauguage，XML 链接语言
- XSL：Extensible Stylesheet Language (XSL)，可扩展样式表语言
- XSLT：XSL Transformations，XSL 变换
- XPointer：XML Pointer Language，XML 指针语言
- XPath：XML Path Language，XML 路径语言
- DOM：Document Object Model，文档对象模型
- XML Schema：XML Schema 规范
- XQuery：XML Query Language，XML 查询语言
- XML Encryption：XML 加密规范
- XML Canonicalization：XML 规范化规范
- XML Signature：XML 签名规范

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>

<!DOCTYPE CV  SYSTEM "CV.dtd">
<cv>
        <contact>
                <! -- This is the contact information of me -->
                <name>Chenzhou CUI  </name>
                <unit>National Astronomical Observatory</unit>
                <address>
                        <street>Datun Road 20A</street>
                        <district>Chaoyang District</district>
                        <province>Beijing </province>
                        <zip>100012</zip>
                        <country>China</country>
                </address>
                <phone>86-10-64877703-1320 </phone>
                <fax>86-10-64878240</fax>
                <email>ccz@bao.ac.cn </email>
                <web>http://www.lamost.org/~cb/ </web>
        </contact>

        <! -- There are some personal information of me -->
        <personal>
                <gender> Male</gender>
                <birthday>February 27, 1976</birthday>
                <family>Married</family>
                <nationality>China</nationality>
        </personal>
        <! -- There are some papers of me -->
        <publication>
                <paper>
                        <title>Abundance Gradients in the Galactic Disk</title>
                        <author>Chenzhou CUI</author>
                        <author>Yuqin CHEN</author>
                        <author>Gang ZHAO</author>
                        <author>Yongheng ZHAO</author>
                        <year>2003</year>
                        <magazine>Science in China </magazine>
                        <volume>(Series G),46 (1)</volume>
                        <startpage>52</startpage>
                        <endpage>61</endpage>
                </paper>
                <paper>
                        <title>Technical Progresses of International Virtual Observatories</title>
                        <author>Chenzhou CUI</author>
                        <author>Yongheng ZHAO</author>
                        <author>Yanxia ZHANG</author>
                        <year>2002</year>
                        <magazine>PABei</magazine>
                        <volume>20 (4)</volume>
                        <startpage>302</startpage>
                        <endpage>311</endpage>
                </paper>
        </publication>
</cv>
```

附文 2.1　　cv.xml

2.1.3 XML 的变革与发展

在 XML 的使用过程中人们发现不同的部分有时不能很好的协调工作。实体声明与使用的机制、格式规范（Well-formed）的文档的处理规则、在 XML 文档中嵌套另外 XML 文档的有限可能性都是问题的源泉。由于 XML 太精简了，使得它容易被采用和扩展。也是由于它太精简了，用户在使用前几乎都必须对其进行扩展。

事实也是如此，人们都在这么做。最初的 XML 开发小组已经被许多目的不同、背景不同的开发组所替代。这些开发组的加入使得 XML 更加强大。但同时，XML 也正在变得越来越复杂和混乱。当初只有 25 页的 XML 规范现在已经变为长达数百页复杂的规范集。5 年前，一个好的程序员可以在一周之内完成一个 XML 工具的开发，但现在需要一个庞大的开发团队。XML 变得越来越复杂，其可用性正在受到损害。

如何继续保持 XML 的有用性和易用性？XML 该如何进一步发展？这是 W3C 和其他的 XML 开发组织正在考虑的问题。

2.2 Web Services

Internet 的发展史同时也是 WEB 应用的发展史。基于超文本传输协议（HTTP）的 WEB 应用可以说是 Internet 上最耀眼也最多样化的服务形态。

最开始的时候仅仅使用 WEB 来共享信息。信息发布者将信息发布在 WEB 服务器上，用户登陆 WEB 站点浏览或者下载信息。随着 WEB 应用开发语言（比如 CGI、Perl、ASP、PHP、Javascript、Python）的兴起，各种各样的 WEB 应用开始出现。开发人员在 WEB 服务器上添加了编程语言的支持，同时使用数据库对动态的 WEB 数据进行管理。

随着 WEB 应用的不断发展，应用的复杂度不断提高，由 WEB 服务器和数据库服务器组成的两层模式无法满足大用户量情况下复杂应用的要求了。为了提高系统的吞吐率和应用的实现效率，架构师们在两层架构基础上增加了应用服务器，从而形成了由 WEB 服务器、应用服务器、数据库服务器构成的 WEB 应用三层架构。

但是随着 WEB 应用的不断发展，WEB 应用和传统桌面应用之间存在的连接鸿沟越来越显著，人们不得不反复的将数据从 WEB 应用迁移到桌面应用，从传统桌面应用将数据迁移到 WEB 应用。计算机的应用是要满足自动化的需求。自动化流程中的人工流，也就是人工干预，会在不同程度上降低工作效率和人们的积极性。WEB 应用和传统桌面应用之间存在的连接鸿沟成了阻碍

WEB 应用进入主流工作流的一个巨大障碍。

为了实现大量 WEB 应用之间的对接，WEB 服务技术应运而生。传统 WEB 应用技术解决的问题是如何让人来使用 WEB 应用所提供的服务，而 WEB 服务技术则要解决如何让计算机系统来使用 WEB 应用所提供的服务。

2.2.1 WEB 服务体系结构

图 2.1 简明地表述了Web Services的核心体系架构[11]。各种高级和扩展的 Web Services体系架构都是在这个三角形的基础架构上扩展而成。同时，这个 Web Services体系架构也是组织各种Web Services标准的框架。

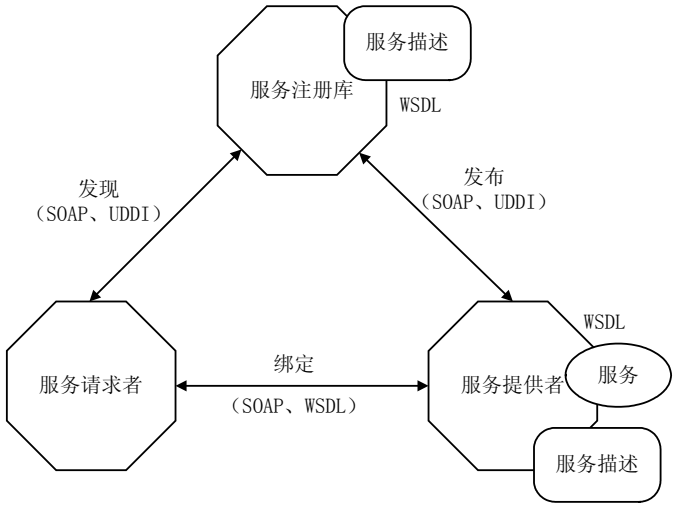


图 2.1 WEB 服务体系结构模型

Web Services 体系结构基于三种角色之间的交互；彼此的交互涉及到发布、查找和绑定三种操作。这些角色和操作一起作用于 Web Services 构件：Web Services 软件模块及其描述。

Web Services 体系结构中的角色包括：

- 服务提供者（Service Provider）：从用户的角度看，这是服务的所有者。从体系结构的角度看，这是被访问服务运行的平台。
- 服务请求者（Service Requestor）：从用户的角度看，这是要求使用某些特定功能的用户。从体系结构的角度看，这是寻找并调用服务的应用程序，或是启动与服务交互的应用程序。服务请求者角色可以由浏览器或应用程序来担当，由人或无用户界面的程序（如另外一个 Web Service）来控制。
- 服务注册库（Service Registry）：这是可检索的服务描述的注册中心，服务提供者在此发布他们的服务描述，服务请求者在此搜寻要找的服务的描述。值得注意的是，服务请求者同样可以从服务注册库以外的

其他来源得到服务描述，例如本地文件、FTP 站点、Web 站点、ADS（Advertisement and Discovery of Services）文件、DISCO（Discovery of WebServices）文件等。

实际操作中服务注册库与普通 Web Service 一样，也是一个 Web Service，因此 Web Service 体系架构中的各种操作都与调用 Web Services、描述 Web Services 相关。像发布和查找都是调用 Web Service 的操作，在具体调用服务注册库时，则要涉及与服务注册库交互的具体规范方法（UDDI），而绑定操作则涉及调用 Web Service 和 Web Service 描述。

对于利用 Web Services 的应用程序，必须发生以下三种操作或者说是行为：发布服务描述、查询或查找服务描述、根据服务描述绑定或调用服务。这些行为可以单次或反复出现。

- 发布（Publish）：为了使服务可访问，需要发布服务描述使服务请求者可以查找它。发布服务描述的位置可以根据应用程序的要求而变化。
- 查找（Find）：服务请求者可直接检索服务描述或在服务注册库中查询所要求的服务类型。对于服务请求者，可能会在两个不同的生命周期阶段中牵涉到查找操作：在设计时为了程序开发而检索服务的接口描述，而在运行时为了调用而检索服务的绑定和位置描述。
- 绑定（Bind）：在绑定操作中，服务请求者使用服务描述中的绑定细节来定位、联系和调用服务，从而在运行时调用或启动与服务的交互。

Web Services 体系架构中包含的 Web Services 构件有：

- 服务（Service）：在这里，Web Service 是一个由服务描述语言描述的接口，服务描述的实现就是该服务。服务是一个软件模块，其存在目的就是要被服务请求者调用或者与服务请求者交互。当服务的实现中利用其他 Web Service 时，它也可以作为请求者。
- 服务描述（Service Description）：服务描述包含服务的接口和实现的细节。其中包括服务的数据类型、操作、绑定信息和网络位置，还可能包括方便服务请求者发现和利用的分类及其他元数据。服务描述可以被发布给服务请求者或服务注册库。

2.2.2 Web Services 标准协议栈

为了实现Web Services体系架构的各种操作和这些操作所要达到的技术目标和商业目标，各大厂商与标准化组织一起制定了如表 2.1 所示的系列Web Services标准，这些标准组成了Web Services标准协议栈[12]。 随着WEB服务日益

广泛的应用，新的协议还在不断的制定中，WEB服务协议栈的深度在不断加深。

这个 Web Services 标准协议栈自下而上可以分为四个主要部分：

- Web Services 调用（消息传输），包括 HTTP、FTP、SMTP 等传输协议，XML、SOAP 等消息协议，以及 WS－Attachment、WS-Routing、WS-Security 等在消息协议上的扩展；
- Web Services 描述（接口、实现描述），包括 XML Schema、WSDL 和 WSCL 等；
- Web Services 注册（发布、发现），包括 UDDI、WS－Inspection 等；
- Web Services 工作流（商业流程、事务等），包含 BPEL4WS、WS－Transaction 和 WS－Coordination 等。

工具	层次
BPEL4WS	工作流
WS-Transaction WS-Coordination	事务控制
UDDI	服务发布和发现
WS-Inspection	服务发布和发现
WSCL	会话描述
WSDL	服务描述
WS-Security	安全信息传递
WS-Routing	消息路由
WS-Attachment	消息附件
SOAP	基于 XML 的消息传递
XML Schema	XML 建模
HTTP、FTP、SMTP、MQ	传输协议

表 2.1 Web Services 标准协议栈

这些标准协议都是实现 Web Services 各项功能需要的功能模块，其中，最为核心的是 XML、SOAP、WSDL 和 UDDI。

XML Schema: 描述元语言

XML Schema 是所有 Web Services 标准的基础，是描述 XML 文档格式的模式语言。Web Services 标准都基于 XML，XML Schema 就是用来对所有 Web Services 标准以及 Web Services 实例中使用的 XML 数据、文档进行建模，通过 XML Schema 能够定义所有的 Web Services 标准。

SOAP: 调用 Web Services

简单对象访问协议（Simple Object Access Protocol，简称 SOAP），为在一个松散的、分布的环境中使用 XML 对等的交换格式化和类型化的信息提供了一个简单且轻量级的机制。SOAP 目前最新的版本是 1.2，由 W3C 的 XML Protocol WG 制定。实际投入使用的 SOAP 版本是 1.1。目前 SOAP 1.1 在.NET 中的实现是内置的，其前身是 Microsoft SOAP Toolkit；在 J2EE 平台，最主要也最通用的 SOAP 实现是 Apache SOAP，当前的成熟版本是 2.3 和 3.0。

WSDL: 描述 Web Services

WSDL 是一种 XML Application。它首先对访问的操作和访问时使用的请求/响应消息进行抽象描述，然后将其绑定到具体的传输协议和消息格式上，以最终定义具体部署的服务访问点，相关的服务访问点通过组合就成为抽象的 Web Services。客户端可以通过这些服务访问点对包含面向文档信息或面向过程调用的服务进行访问。

UDDI: 注册和发现 Web Services

作为 Web Services 体系中的元服务（Meta Service），UDDI 为 Web Services 体系提供 Web Services 的注册和发现机制。UDDI 相对其他 Web Services 技术有其独特性：UDDI 是一个服务，其具体实施形式就是 Web Services。一般我们称这个提供 Web Services 注册发现服务的 Web Services 为 UDDI 注册库。

除了上面介绍的这四个核心协议外，协议栈中的其他协议主要功能如下。

WS-Security: 为 SOAP 加上安全的闸门

WS-Security 提供了一套可以帮助 Web Services 开发者保障 SOAP 消息交换安全的机制。它增强了基本的 SOAP 消息传递，通过应用消息完整性、消息机密性、单消息认证等手段提供了不同的保护级别。这些基本机制可以通过各种方式联合，以适应构建使用多种加密技术的多种安全模型。

WS-Routing: SOAP 消息的路由机制

WS-Routing 定义了路由 SOAP 消息的机制。SOAP 是一个轻量级的有限传输协议，定义了一系列传输交换机制，用来传输在应用层协议上使用的方法调用。虽然在 SOAP 规范中引用了一个虚拟的消息路径机制，但 SOAP 实际上没有定义从一点发送消息到另一点的机制。WS-Routing（以前被称做 SOAP 路由协议）是一个无状态协议，它扩展了 SOAP 协议。它通过定义一个方法来说明一个预先设计好的路由或传输路径，这个路径将从消息源，经过若干中介，最后到达消息的最终接收者。

BPEL4WS: 商业流程整合

BPEL4WS 是一种基于 XML 的工作流定义语言，它使企业能够描述由 Web Service 参与的复杂的业务流程，同时它又能将 Web Services 组合而成的工作流进一步包装成为更高级别的 Web Services 并发布出去。

WS-Coordination 和 WS-Transaction: 事务控制

通过工作流我们可以将 Web Services 单元贯穿成一个整体。但是，对于商业工作流而言，Web Services 单元的结构通常会很复杂，活动参与者之间的关系也很复杂。由于业务的延迟和用户的交互，执行这样的工作流通常要花很长时间才能够完成。

WS-Coordination 使用一组协调服务和一组协调协议定义了一个用于协调活动的可扩展框架。这个框架使参与者能够定义用于协调各种活动的协调协议，包括用于简单的短期操作协议和用于复杂的长期事务活动协议。其中的组件服务包括：激活服务、注册服务和一组特定的协调协议。

WS-Transaction 是在 WS-Coordination 框架上定义的标准，提供了原子事务（AT）和业务活动（BA）两种协调类型的定义。原子事务用于协调持续时间短并且在有限的信任域内执行的活动，具有“全做或者全不做”的特性；业务活动用于协调持续时间长的活动，并希望应用业务逻辑来处理业务异常，如资源锁定等。一个 Web Services 应用可以既包含原子事务又包含业务活动。

WS-Inspection: 分布式的发现机制

UDDI 是集中式的 Web Services 注册和发现机制。WS-Inspection 则把重点放在了分布式服务发现上，它在服务提供者的服务提供站点提供对服务描述的引用。有些 Web Services 可能尚未被列入 UDDI 注册中心，而通过 WS-Inspection 规范可以使用分布式的方式来发现这些 Web Services，从而作为 UDDI 的功能补充。

2.3 Grid 技术

清华大学李三立院士将网格与信息高速公路作了比较，他说：“将先进计算基础设施（网格）与信息高速公路相比较，可以说，信息高速公路是信息传输和获取的信息基础设施；而先进计算基础设施则是信息处理的信息基础设施。虽然，国内外都有不断把信息高速公路扩充频带宽度、改进路由器性能的计划；但是，国外科学家认为：真正的下一代信息基础设施是先进计算基础设施，它将使以计算机为主体的信息处理发生根本性的变化。”[13]

中科院计算所李国杰院士认为：“网格不同于国外正在搞的Internet 2 或下

一代Internet (NGI)，网格可以称作是第三代 Internet，其主要特点是不仅仅包括计算机和网页，而且包括各种信息资源，例如数据库、软件以及各种信息获取设备等，它们都连接成一个整体，整个网络如同一台巨大无比的计算机，向每个用户提供一体化的服务。”[14]

网格技术的产生、发展必须具备以下三个基本条件：计算资源的广域分布、网络技术（特别是互联网）以及不断增长的对资源共享的需求。在计算机技术发展的早期阶段，只有很少数量的大型计算机，它们通常被安装在相互独立的计算中心内，多个计算机用户通过使用终端来共享一台大型机的资源，但却不能同时共享多台大型机的计算资源。随着网络技术的发展，多台大型计算机可以在局域网内互连，用户通过网络便可以同时使用多台计算机的资源。而互联网的飞速发展与普及使得网格计算技术的产生成为可能。图 2.2 显示了计算资源共享的发展过程[15]。

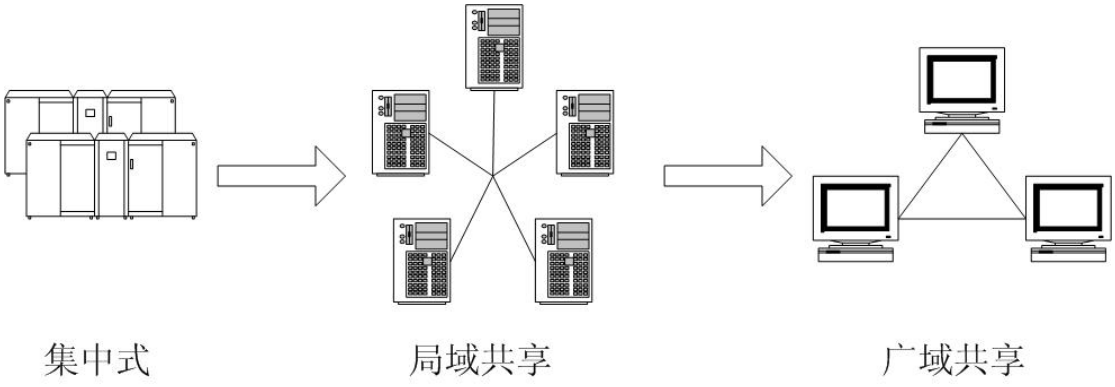


图 2.2 计算资源共享的发展过程

从上世纪 60 年代末开始研究计算机分组交换技术到今天，互联网已经走过两代历程。第一代是 20 世纪 70～80 年代，主要的成就是把分布在世界各地的计算机用 TCP/IP 协议连接起来，主要的应用是电子邮件。第二代是 20 世纪 90 年代，主要成就是把成千上万网站上的网页连接起来，主要的应用是 Web 信息浏览以及电子商务等信息服务。目前正处于从第二代互联网向第三代互联网过渡的转型期。第三代互联网也就是网格（Information Service Grid），其主要特点是不仅仅包括计算机和网页、而且包括各种信息资源，例如数据库、软件以及各种信息获取设备等，它们都连接成一个整体，整个网络如同一台巨大无比的计算机，向每个用户提供一体化的服务。简单地讲，传统互联网实现了计算机硬件的连通，Web 实现了网页的连通，而网格试图实现互联网上所有资源的全面连通。网格追求的最终目标是能够做到服务点播和一步到位的服务，把整个互联网整合成一台巨大的超级计算机，实现计算资源、存储资源、数据资源、信息资源、知识资源、专家资源的全面共享。

物质与能量原则上只能分享，一吨水、一度电你使用了我就不能使用。而信息的最大特点是可以共享，不会因使用同一信息资源的用户多而被耗尽。在以往的信息化建设中，我们往往忽视了“信息应该共享”这一最本质的应用要求，把信息当成物质与能量一样使用，这已造成极大的浪费。

网格技术要解决的信息共享不是一般的文件交换与信息浏览，而是要把所有个人与单位连接成一个虚拟组织（Virtual Organization），实现在动态变化环境中有灵活控制的协作式信息资源共享。网格与 Web 最大的区别是一体化，即用户看到的不是数不清的门类繁多的网站，而是单一的入口和单一系统映像。

现有的 Web 信息服务器就好像 Internet 世界上一个个孤立的小岛。虽然这些“小岛”之间暂时还有充足的带宽资源可用，但大量的信息还是被“锁”在各个小岛的中央数据库里，各 “孤岛”之间并不能按照用户的指令进行有意义的交流。解决这一问题的最佳途径是建立跨越 Web 的信息分布和集成应用程序逻辑——网格。

网格的兴起将改变传统的 Client/Server（C/S）和 Client/Cluster 结构，形成新的 Pervasive/Grid 体系结构。客户端是各种各样的上网设备，而连在网上的各种服务器将组成单一的逻辑上的网格。

网格的本质特征表现在应用上。网格的服务包括文件消息、计算、信息内容、事务处理和知识服务等，因此网格可大致分为数据网格、计算网格、信息网格与知识网格等。

网格系统大致可以分为五个基本层次：构造层、连接层、资源层、汇集层和应用层，如图 2.3 所示。

构造层：本地控制系统的接口，提供了共享的资源。比如计算资源、存储资源、数据、网络资源和传感器。构造层部件执行的是本地的、与资源相关的操作。

连接层：安全方便的通信，定义了网格环境下的网络交易所需的核心通讯和认证协议。通讯协议实现了构造层资源间的数据交换。认证协议建立在通讯服务基础上提供用于用户和资源识别的加密的安全机制。

资源层：共享单一资源，为个体资源的共享操作定义了安全对话、初始化、监测、控制、记帐、付帐等协议。资源层协议针对的是单个的资源，不考虑分布式资源集合广域上的状态和管理等事务。

汇集层：协调多样资源。资源层关注的是单个的资源，汇集层则关注的是资源的总体。建立在资源层和连接层构成的协议瓶颈上，汇集层实现了广泛的共享操作。汇集层的功能是屏蔽网格资源层中各种资源的分布、异构特性，向

网格应用层提供透明、一致的使用接口。汇集层也称为网格操作系统，它同时需要提供用户编程接口和相应的环境，以支持网格应用的开发。

应用层是用户需求的具体体现。在网格操作系统的支持下，网格用户可以使用其提供的工具或环境开发各种应用系统。

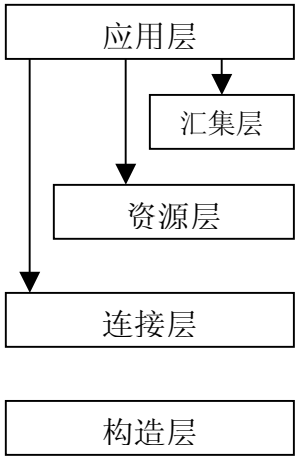


图 2.3 网格体系结构

2.4 OGSA 体系结构

2002 年 2 月，在加拿大多伦多市召开的全球网格论坛（GGF） 会议上，Globus 项目组和IBM 共同倡议了一个全新的网格标准OGSA[16]。OGSA ，开放网格服务架构（Open Grid Services Architecture），将把Globus为代表的网格技术标准与以商用为主的Web Services 的标准结合起来，网格服务统一以Services 的方式实现。OGSA 的诞生，标志着网格已经从学术界的象牙塔延伸到了商业世界中，而且从一个封闭的世界走向了开放的环境中。

OGSA 从一诞生，就得到业界的广泛支持。到目前为止，OGSA 已经广为接受，几乎所有的业界同仁都认为它就是网格的未来。

2.4.1 网格的定义

全球网格研究的领军人物、美国Argonne国家实验室的资深科学家、美国Globus 项目的领导人Ian Foster 曾在 1998 年出版的《The Grid: Blueprint for a New Computing Infrastructure》[17]一书中这样描述网格：“网格是构筑在互联网上的一组新兴技术，它将高速互联网、高性能计算机、大型数据库、传感器、远程设备等融为一体，为科技人员和普通老百姓提供更多的资源、功能和交互性。互联网主要为人们提供电子邮件、网页浏览等通信功能，而网格功能则更多更强，让人们透明地使用计算、存储等其他资源。”

2000 年，Ian Foster 在《The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration》这篇论文中把网格进一步描述为“在动态变化的多个虚拟机构间共享资源和协同解决问题。”但是，人们仍然就什么是网格而争论不休。2002 年 7 月，Ian Foster 在《What is the Grid? A Three Point Checklist》[18]一文中，限定网格必须同时满足三个条件：

1. 在非集中控制的环境中协同使用资源；
2. 使用标准的、开放的和通用的协议和接口；
3. 提供非平凡的服务。

这三个条件非常严格，将P2P[19]、SUN ONE Grid Engine[20]、Condor[21]、Entropia[22]、MultiCluster[23]等都被排除在网格之外。

Ian Foster 把他头脑中的网格概念描绘清楚了，但并不是所有人都同意他的观点。例如，有许多人赞同广义的网格概念，把它称作超级全球网格 GGG（Great Global Grid）。GGG 不仅包括计算网格、数据网格、信息网格、知识网格、商业网格，还包括一些已有的网络计算模式，例如对等计算 P2P(Peer to Peer)、寄生计算等。可以这样认为，Ian Foster 赞成狭义的“网格观”，而 GGG 是一种广义的“网格观”。

不管是狭义还是广义的网格，其目的不外乎是要利用互联网把分散在不同地理位置的电脑组织成一台“虚拟的超级计算机”， 实现计算资源、存储资源、数据资源、信息资源、软件资源、存储资源、通信资源、知识资源、专家资源等的全面共享。其中每一台参与的计算机就是一个节点，就像摆放在围棋棋盘上的棋子一样，而棋盘上纵横交错的线条对应于现实世界的网络，所以整个系统就叫做“网格”了。在网格上做计算，就像下围棋一样，不是单个棋子完成的，而是所有棋子互相配合形成合力完成的。传统互联网实现了计算机硬件的连通，Web 实现了网页的连通，而网格试图实现互联网上所有资源的全面连通。

2.4.2 面向服务的思想

网格体系结构给出了网格的基本组成和功能，描述了网格各组成部分的关系以及它们集成的方式或方法，刻画了支持网格有效运转的方式。OGSA 是在综合了 WEB 服务和网格协议两方面优势的基础上提出的，对 WEB 服务进行了与网格技术兼容性的扩展。

OGSA 是一个面向服务的体系结构，在网格中：

- 一个服务是一个基于网络的能提供某种功能的实体。
- 一个网格服务是一个遵循一套与其接口定义和行为相关的规范的由 WSDL 进行描述的服务。每个网格服务都是一个 WEB 服务，但是对其

进行了网格化扩展。

OGSA与WEB服务和网格协议的关系如图 2.4 所示[24]。基于Grid和WEB服务的思想和技术，OGSA体系定义了一个统一的对外服务语义，即Grid服务；定义了标准的瞬时Grid服务实例的创建、命名和发现机制；为服务实例提供了地域透明性和多协议绑定；并提供了与本地平台系统的集成机制；以WSDL接口和相关约束的格式定义了创建和组织高级分布式系统所需的机制，其中包括生命期管理、变动管理、通告等。

在 OGSA 中，一切都以 Grid 服务的形式体现。面向服务的模型有许多优点：环境中所有部件都进行了虚拟化，通过层层抽象以统一的方式对待这些服务。所有服务都要提供一系列核心永久接口，在此基础上，可以构造层次式的高级服务。核心网格服务与其他网格元素的关系如图 2.5 所示。虚拟化还能实现多个逻辑资源实例向同一物理资源的映射，在无需考虑底层资源的实现和管理情况下构造上层服务。Grid 服务的虚拟化还提供了共同服务语义行为向本地平台相应机制的无缝映射，从而实现平台无关性。

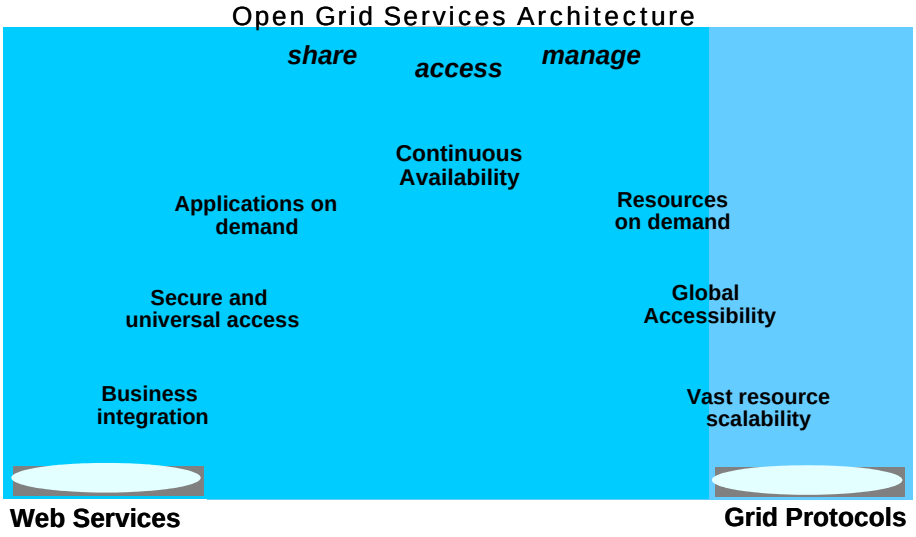


图 2.4 OGSA 与 WEB 服务和网格协议的关系

2.4.3 OGSA 平台

OGSA平台[25]旨在为各种网格系统所共同面临的基本问题定义标准的解决方案和机制。这些基本的问题包括网格服务间的通信、身份确立、授权对话、服务发现、错误通告、服务集管理等。

如图 2.6 所示，OGSA 平台包括三个基本元素：开放网格服务基础设施（OGSI）、OGSA 平台接口和 OGSA 平台模型。

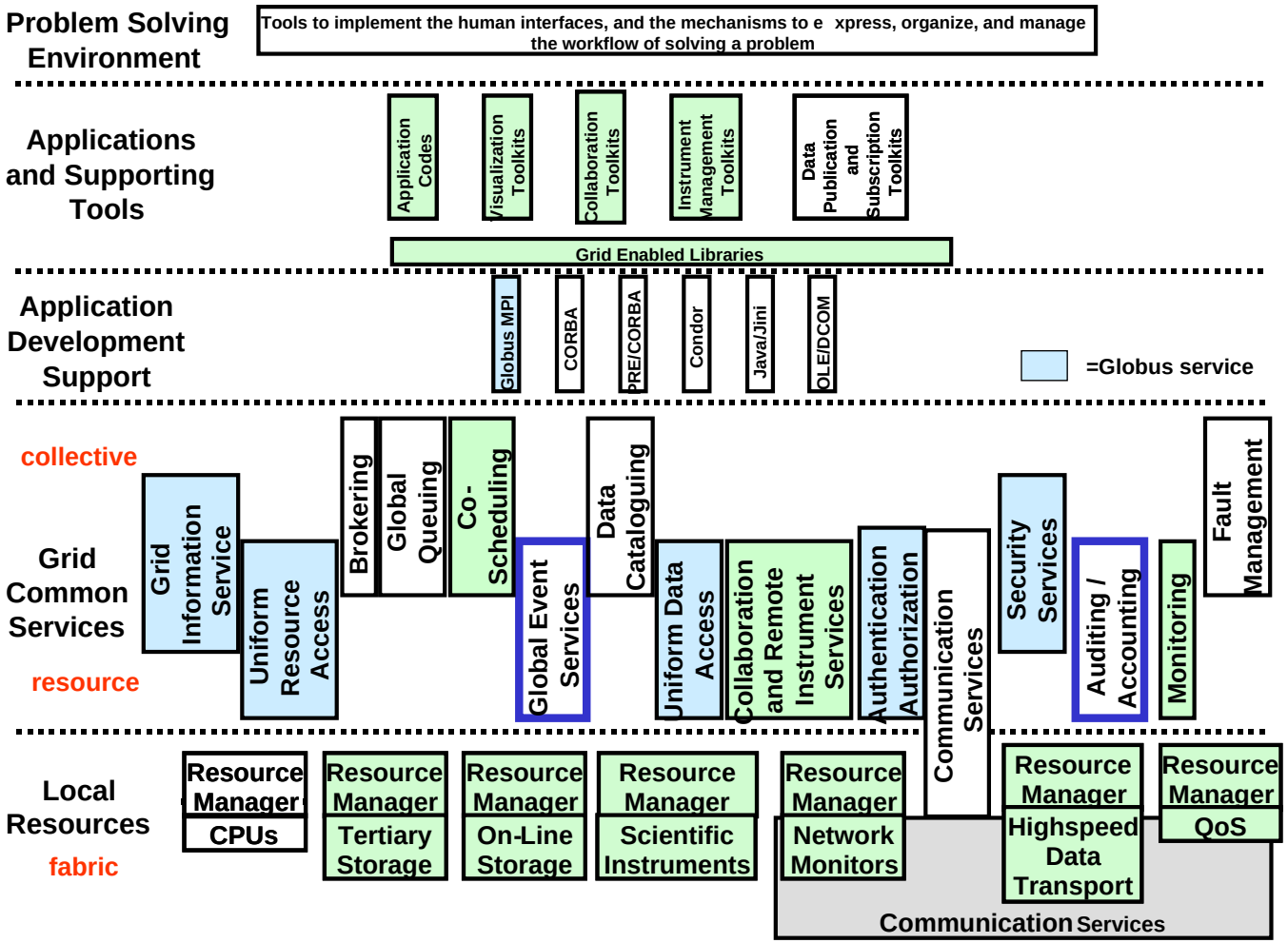


图 2.5 核心网格服务与其他网格元素的关系

- 建立在网格和 WEB 服务基础上，OGSI 定义了一套机制来实现网格服务的产生、管理以及网格服务间的信息交换。一个网格服务是一个实现了一系列规范的 WEB 服务。这些规范，包括接口和行为，定义了一个客户如何与一个网格服务进行交互。这些规范与其他一些与网格服务的产生和发现相关的 OGSI 机制一起实现了对分布式、长生命期的状态进行可控的、带错误反馈的、安全的管理。这对分布式的应用是一个基本的需求。
- OGSA 平台接口，建立在 OGSI 之上，定义了一系列 OGSI 不直接支持的各种功能的接口和相关行为，比如服务发现、数据访问、数据融合、消息、监控等。
- OGSA 平台模型，通过为公用资源和服务类型定义模型来实现上述接口规范。

在 OGSA 平台之上 OGSA 还定义了一些高层系统服务，主要包括：

- 分布式数据管理服务，实现了对分布式异构数据资源的统一透明访问；
- 工作流服务，实现多个分布式网格资源上多个应用作业的协同操作；
- 审计服务，记录资源的使用并对这些记录信息安全的保存和分析，用于发现异常和入侵检测等；
- 指导和监测服务，提供分布式环境中“传感器”的发现，检测信息的收集和分析，探测到非常状况时产生警告；
- 分布式计算的问题判定服务，实现备份、跟踪、记录等功能；
- 安全协议映射服务，实现分布式安全协议透明的映射为本地平台系统的安全服务，以满足本地安全认证和访问控制的需要。

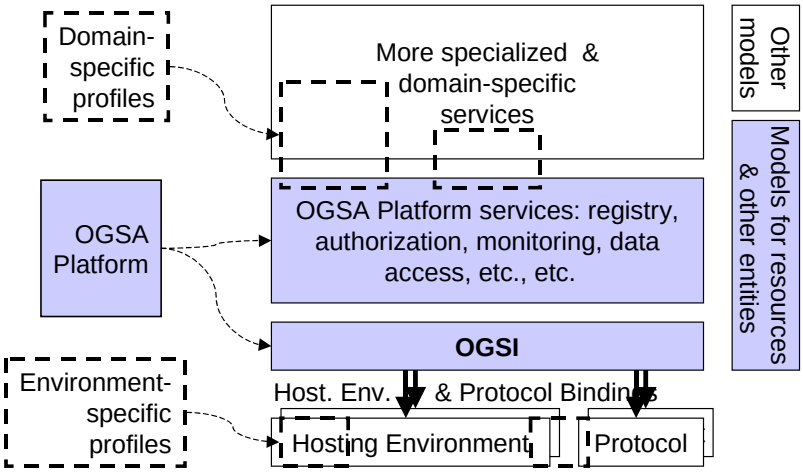


图 2.6 OGSA 平台及其相关环境

2.5 Globus Toolkit

目前，在国际上影响最大的网格开发项目是Globus[26]。这个项目的成员来自美国Argonne 国家实验室数学与计算机分部、南加州大学信息科学学院和芝加哥大学分布式系统实验室等单位，并与美国国家计算科学联盟、NASA 信息能源网格（IPG） 项目、美国国家先进计算基础设施同盟（NPACI）等建立了伙伴关系。Globus项目的主要工作是开发、解决建立网格所需要的基本技术。

目前，他们推出的工具集Globus Toolkit （GT） [27]版本是 2.4。这是一套旨在实现网格资源管理的中间件。Globus计划在 2003 年 6 月推出新的基于开放网格服务体系（简称OGSA）的版本，也就是Globus Toolkit 3.0 （GT3）[a]。目前Globus对OGSA实现的最新版本是Globus Toolkit 3.0 Beta。Globus Toolkit可以满足网格操作的主要需求，实现安全管理、资源管理、数据管理和信息管理等功能。

迄今为止，Globus 项目开发的 Globus Toolkit 已经成为事实上的网格标准。一些重要的公司，包括 IBM、Microsoft、Compaq、Cray、SGI、Sun、Fujitsu、Hitachi、NEC 等已经公开宣布支持 Globus Toolkit。大多数网格项目也都是基于 Globus Tookit 所提供的协议及服务建设的，例如美国的物理网格 GriPhyN、欧洲的数据网格 DataGrid、荷兰的集群计算机网格 DAS-2、美国能源部的科学网格和 DISCOM 网格、美国学术界的 TeraGrid 等等。

当前，Globus 的发展在很大程度上代表了网格技术的发展，其主要发展阶段包括：

1. 最初探索阶段：从 1996 年到 1999 年，在此期间推出了 Globus 1.0 标准。这个阶段主要的工作是对网格技术进行广泛的探讨，制定网格技术所必需的核心协议。
2. 数据网格阶段：从 1999 年到现在，推出了 Globus 2.0 标准。在这个标准中实现了大规模数据管理和分析功能。
3. 开放式网格服务体系：这个阶段从 2001 年开始，计划在 2003 年 6 月推出 Globus Toolkit 3.0 标准。在这个标准中将实现与 Web 服务的融合，提供更好的主机环境、资源可视化以及其他高层服务。
4. 初步可升级系统：这个阶段将从 2003 年开始，使网格技术得到初步的实现，同时把网格技术向无线技术、普遍计算、对等网络等方向进行扩展。

[a] 从目前情况来看，GT3 很可能无法按计划发布

2.5.1 GT3

作为OGSA的一种实现，GT3 的服务模块包括：核心服务、安全服务、基本服务、数据服务和其他服务。这几种服务之间的关系如图 2.7 所示[28]。

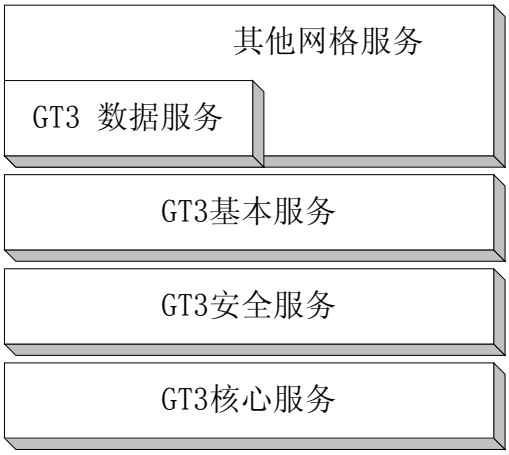


图 2.7 GT3 体系结构

GT3 核心服务包括：

- Factory：工厂服务。通过它的 CreateService 操作可以创建新的网格服务，返回网格服务句柄（Grid Service Handle，GSH），维护所需的服务数据元素。一个 GSH 是一个 URL，用来为一个网格服务实例命名。为了使用这个服务，一个 GSH 必须转换为一个网格服务参考（Grid Service Reference，GSR）。GSR 描述一个客户如何能与一个网格服务实例通信。HandleMap 接口允许一个客户把一个 GSH 映射到 GSR。GSH 仅代表名字，GSR 包括了传输协议的绑定信息和数据编码格式。
- NotificationSource：用来接受用户通告订阅
- HandleResolver：句柄解析器
- Registration：通过返回一组网格服务的 GSH 来支持服务发现
- NotificationSubscription：用户通告订阅
- NotificationSink：用来异步发送通知消息
- GridService：所有网格服务都必须支持的接口

GT3 的核心服务要实现：

- OGSI 规范
- 公用接口和 API（包括服务数据、通告、查询、状态管理等）
- 网格服务容器基础设施
- 开发与运行环境

GT3 安全服务主要包括：

- 传输层：SSL

- 消息层： WS-Security, SOAP
- 安全会话服务
- 消息签名与加密
- 信任授权、访问控制
- 基于虚拟主机环境的增强型资源安全模型

GT3 基本服务包括：

- 索引服务
- 元数据目录索引服务
- 服务数据提供者基础设施
- 可管理作业服务
- 资源管理
- GRAM 的实现（GRAM 是 Globus 资源管理体系结构中最底层的部分，通过使用一组 WSDL/OGSI 客户端接口进行作业的提交、监视和终止操作，实现作业的远程运行等活动。）
- 虚拟主机环境和路由基础设施
- 文件流式化服务
- 可靠文件传输服务（GridFTP、可重获的网格服务、进展与重新启动监视）

GT3 数据服务包括：

- 数据访问
- 数据备份
- 数据缓存
- 元数据目录和服务集的服务
- Schema 变换
- 存储
- 数据库

GT3 其他服务包括：

- 错误诊断
- 负载管理

按照 Globus 最新公布的开发计划，GT3 的推出过程如下：

- Alpha 版本：2003 年 1 月 13 日
- Beta 版本： 2003 年 6 月 6 日
- GT3 正式版：2003 年 6 月底

2.6 VO 与网格

从上面的介绍我们可以清楚的看出网格技术在很大程度上与虚拟天文台的开发目标一致，利用网格技术作为基础设施发展虚拟天文台将是必然的趋势[29]。

一方面，虚拟天文台的建立和实现需要网格技术的支持。

虚拟天文台的最终发展目标就是实现全球天文数据的高级共享，同时提供一整套的智能化工具。TB 量级甚至 PB 量级大型天文数据产出项目的不断涌现，对数据存储、数据管理、数据传输、数据检索等技术提出了更高的要求。在如此海量分布式数据的基础上进行科学研究，就必须有全新的数据共享、数据互操作、作业调度、数据可视化、数据统计分析、数据挖掘、数据安全管理等工具的支持。

虚拟天文台的这些需求正是网格技术要实现的目标。网格技术将实现把整个互联网整合成一台巨大的超级计算机，实现计算资源、存储资源、数据资源、信息资源、知识资源、专家资源的全面共享，为用户提供一步到位的服务。因此，虚拟天文台把网格技术作为自己的技术基础将是可行而明智的选择。

另一方面，虚拟天文台将为网格技术的发展提供最好的实验场。

正如前一章所述，天文数据有着其他学科数据无法比拟的特点：

- 开放性
- 海量数据
- 良好归档
- 格式多样
- 全波段数据

虚拟天文台要实现对这样数据的融合。这样的发展目标为网格技术提供了独一无二的试验场。从网格基础设施的构建，到网格操作系统的开发，最后到网格天文应用工具的实现，虚拟天文台为网格技术提供了一整套的应用需求。

虚拟天文台的实现不可能仅是天文学家的事情，必须与计算机、网络、数学等领域的专家共同努力。这无论对天文学还是对信息科学、数学都是双赢甚至多赢的合作。

参考文献

[1] [XML] W3C Extensible Markup Language. http://www.w3c.org/XML

[2] [WS] W3C Web Services Activity. http://www.w3.org/2002/ws/

[3] [GGF] Globe Grid Forum. http://www.ggf.org

[4] [SGML] Standard Generalized Markup Language. http://www.w3.org/MarkUp/SGML/

[5] [HTML] HyperText Markup Language. http://www.w3.org/MarkUp/

[6] [W3C] World Wide Web Consortium. http://www.w3c.org

[7] [XHTML] Extensible HyperText Markup Language. http://www.w3c.org/MarkUp/

[8] [MathML] Mathematical Markup Language. http://www.w3c.org/Math/

[9] [AML] Astronomical Markup Language. http://xml.coverpages.org/aml.html

[10] [AIML] Astronomical Instrument Markup Language. http://pioneer.gsfc.nasa.gov/public/aiml/

[11] 柴晓路. WEB 服务架构与开放互操作技术. 第 1 版. 北京: 清华大学出版社, 2002. p15

[12] 柴晓路. 索引 Web Services 标准. 计算机世界报, 2003 (06): B8-B11

[13] 刘鹏. 网格概念的界定. http://grid.cs.tsinghua.edu.cn/grid/paperppt/GridConcept.pdf

[14] 李国杰. 信息服务网格—第三代 Internet. 计算机世界, 2001 (40): B8-B10

[15] 李伟. 万丈高楼平地起——浅谈网格计算基础. 计算机世界, 2001 (43): B3-B5

[16] Ian Foster, Carl Kesselman, Jeffrey M. Nick, et al. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. http://www.gridforum.org/ogsi-wg/drafts/ogsa_draft2.9_2002-06-22.pdf

[17] Ian Foster, Carl Kesselman. The Grid: Blueprint for a New Computing Infrastructure. http://www.amazon.com/exec/obidos/ASIN/1558604758/o/qid=958665349/sr=2-1/103-6896860-5655839

[18] Ian Foster. WHAT IS THE GRID? A THREE POINT CHECKLIST. http://www.gridtoday.com/02/0722/100136.html

[19] [P2P] Peer-to-Peer Working Group. http://www.peer-to-peerwg.org/

[20] [SunONE] Sun ONE Grid Engine. http://wwws.sun.com/software/gridware/

[21] [Condor] Condor Project. http://www.cs.wisc.edu/condor/

[22] [Entropia] Entropia - PC Grid Computing. http://www.entropia.com/

[23] [MultiCluster] MultiCluster Project. http://www-gppd.inf.ufrgs.br/projects/mcluster/

[24] Ian Foster. The Globus Toolkit: The Open Source Solution for Grid Computing

.
http://www.globusworld.org/globusworld_web/plenary/Ian%20Foster%20Keynote.ppt
[25] I. Foster, D. Gannon. The Open Grid Services Architecture Platform. http://www.gridforum.org/Meetings/ggf7/drafts/draft-ggf-ogsa-platform-2.pdf
[26] [Globus] Globus Project. http://www.globus.org
[27] [GT] Globus Toolkit. http://www.globus.org/toolkit/
[28] The Globus Project. The Globus Toolkit v3 Developers and Administrators Tutorial. Session 1: Overview. http://www.globusworld.org/globusworld_web/gt3/Session-1-Overview1.ppt
[29] 崔辰州, 赵永恒. 虚拟天文台和网格技术. In: 孙九龄, 施慧中. 科学数据——管理与共享. 第一版. 北京: 中国科学技术出版社, 2002. 272-290

第三章 China-VO 计划

3.1 项目研发的必要性

虚拟天文台是“科学驱动，技术使能”的产物，是当天文学进入数据富庶时代后，天文界为了解决如何进行数据密集型在线科学研究这个极富挑战性的难题而给出的最富竞争力的候选答案。

截止到 2003 年 6 月，在虚拟天文台概念提出不到 4 年的时间里已经有 12 个国家和国际组织（包括中国在内）明确提出了自己的设想并开始了相关的研究工作[1]，以事实说明虚拟天文台将是 21 世纪天文学的一个重要发展方向。

在国内天文界，目前有多个重大工程项目正在建设和预研究阶段，比如国家“九五”重大科学工程项目大天区面积多目标光纤光谱望远镜（LAMOST）[2]、太阳空间望远镜（SST）[3]、500 米口径射电望远镜（FAST）[4]、国家天文台密云观测站 50 米射电望远镜、国家天文台南方天文观测基地等。这些观测设施的建成将大大增强中国天文学在世界上的竞争力。观测能力的提高、大面积CCD探测器的使用使得这些设备的观测数据产出将非常丰富，是当前老观测设备无法比拟的。如何高质量、高效率的使这些数据为国内外的天文学家乃至社会各界使用是天文学发展必须考虑的问题。这些观测设施的建成将大大增强中国天文学在国际天文学界的发言权，并将在未来的虚拟天文台中发挥重要的作用；同时，这些大型观测设备也需要虚拟天文台这样的系统，以便在最大程度上共享观测资料，取得最大的科学收益。

中国虚拟天文台（简称China-VO）[5]的建设是非常必要的：

1. 只有加入国际虚拟天文台联盟（IVOA），才能以平等的身份全方位共享 IVOA 的技术与资源。
2. 建设China-VO是实现LAMOST、BATC[6]巡天等我国自产数据与IVO数据融合的最佳途径。
3. 建设 China-VO 可以培养一批与 IVO 相适应的天文学家和技术人才，为未来中国天文学的发展提供智力支持。
4. 可以利用 IVO 丰富的资源加强教育和科学普及工作，提高我国公众的科学素质。
5. 为国内网格技术研究提供最好的实验场，推动国内 IT 技术的发展。

目前，国际虚拟天文台（简称 IVO）正处于起步阶段，中国天文界应该积

极参与，从一开始就把自身融入到这股将为天文学带来一场新的革命的浪潮中。

虚拟天文台的出现为中国天文学提供了新的发展机遇，主要表现在以下几个方面。

- **降低了进行高水平实测天文学研究的门槛**

高水平的实测天文学研究离不开大型天文观测设施，但这些设施，特别是空基观测设施的建造都需要高昂的投资。我国的经济实力虽已大大增强，但还无法满足这样的需求。虽然我们可以申请国际大型观测设施的观测时间，但能得到的观测时间毕竟是非常有限的。虚拟天文台的建成将在很大程度上改变高质量观测数据缺乏的局面。虚拟天文台借助先进的 IT 技术将世界上主要的天文数据集无缝透明的融合在一起，同时还提供功能强大的分析处理服务。国内用户只要能访问互联网就能访问这些极为丰富的天文资源，为进一步的天文研究敞开大门。

- **减小了硬件设施对天文研究的限制**

虚拟天文台对于我国这样的发展中国家来说具有更深刻的意义。国力的限制使得我们不可能全面地发展各种类型的大型天文观测设备。国际上天文观测数据的开放性和虚拟天文台的建设与发展使我国的天文学家完全可以充分利用国际上最先进的设备所获得的高质量的数据来做出一流的科学研究工作。

- **后来者居上的潜在优势**

在欧美，天文数据服务，比如 ADC、CDS，已经发展了几十年。他们积累了大量宝贵的经验和技术。但另一方面，许多提供对外服务的硬件、软件都已经老化甚至过时。为了保证原有服务的延续性，他们必须花很大的精力维持或者升级原来的系统和服务，新系统和服务的实施也必须尽量与原有的系统兼容。这给他们的 VO 建设增加了额外的难度和工作量。也就是说，欧美这些发达国家的天文数据服务提供者有许多历史包袱。与他们相反，我们可以说没有任何历史包袱，没有向前兼容性问题。China-VO 可以架构在最先进的软硬件基础设施之上。这就是我们的后发优势。

虽然有这样的机遇和优势，我们在迎接新的机遇的同时也面临着许多挑战，主要表现在：

- **人才短缺**

在国内，专业天文研究队伍包括研究生在内不足 2000 人。从事天文技术应用开发的人员不超过 200 人。国内大学，目前只有北京大学、南京大学、北

京师范大学设有天文系。而大多数的美国大学都设有天文系或者天文专业。在人才数量上，我们与欧美相差悬殊。更重要的是，在科研和技术水平上，我们与欧美更是相差悬殊。VO 的建设过程，最重要的投资是智力的投资。China-VO 的智力源泉在哪里？

- **资源短缺**

我们最大的光学望远镜口径是 2.16 米。世界上最大的光学望远镜已经是 10 米。美国的哈勃太空望远镜、钱德拉 X 射线望远镜、SIRTF 红外望远镜等一大批空间天文台已经或将要投入使用。而目前我们没有一架空间天文观测设备。没有尖端的观测设备就很难得到高质量的观测数据。没有观测数据，VO 就失去了存在的基础。在缺乏原创数据的前提下，China-VO 如何在 IVO 中定位？

- **技术短缺**

虽然如第一章所述，我国的计算机技术和互联网技术已经取得巨大的进步并正飞速发展。但我们不能不承认，IT 技术中许多关键技术，比如大规模集成电路设计制造、操作系统、数据库管理系统、超级计算机研制、下一代互联网，我们都还远没有掌握。VO 是一个“技术使能”的系统。我们现有的技术基础能否满足 VO 的需要？

- **资金短缺**

虽然近 20 多年来，我国政府在科学研究上的投入逐年增加。但我们不得不承认目前的投入力度还远不能满足科研的需要。与我国国力相近的印度在三年时间里对印度虚拟天文台项目给予了 100 万美元的资助。我们能否每年得到 200 万人民币的资金支持呢？资金短缺就很难引进人才，购买设备和软件。如何能多快好省的建设 China-VO？

China-VO 只有在 IVO 的建设中做出自己的贡献，才有可能充分利用虚拟天文台来获得科学上的发展；我们的突破口在哪里？

这些都是 China-VO 研发过程中需要考虑和解决的问题。在论文的后续部分将陆续对这些问题进行探讨。

3.2 现有的工作基础

天文数据的急剧增长和计算机网络技术的突破性进展是国际虚拟天文台计划产生和实现的两大直接动力。经过近几年的努力，我们在这两方面已经积累了一定的基础，为 China-VO 的建设准备了条件。

经过数十年的奋斗，我国现代天文学取得了长足的发展，一批大型天文观测设备已经投入使用并取得了丰富的观测资料。更重要的是，像LAMOST、SST、FAST等一批正在研制或预研中的大观测设备在投入使用后将极大丰富我国的天文数据资源，这为我们建设China-VO奠定了数据资源基础。此外，除了这些自产的天文数据，经过多年的努力我们积累了大量的国际天文数据和相关软件，主要资源如表 3.1 所示[7]。此外，SDSS的第一批巡天数据DR1，大约3TB刚刚释放。China-VO正与LAMOST一起在和SDSS协商，以期能在今年下半年获得这批数据。

数据集	在线情况	数据量(GB)
NASA/ADC 在线星表	√	12
USNO-A1.0 星表		5
USNO-A2.0 星表	√	6
Hipparcos/Tycho 星表		6
RealSky（第一期巡天图像）	√	5
DSS（第一期巡天图像）	√	60
DSS-II（第二期巡天图像）		>360
Einstein X 射线卫星数据		5
ROSAT X 射线卫星数据		27
2MASS 红外星表		>10
ADS 文献数据服务	√	350
SDSS EDR	√	60
常用天文软件		60
（合计）		**>970**

表 3.1 国家天文台积累的主要国际数据资源

经过对国际虚拟天文台技术持续的跟踪，我们已经对 WEB 服务、网格服务等技术以及 IVOA 正在探讨中的各种标准有了一定的掌握和积累。最近，我们已经成功开发了自己的第一个网格服务、第一个 VOTable 编码器，实现了基于 GT3 的锥形检索（ConeSearch）服务。

同时，国内的计算机和网络技术也正在飞速发展，取得了许多世界先进水平的研究成果。第二代互联网和网格技术的研究正在与世界其他国家同步进行。为网格时代开发的曙光 4000 服务器将达到 4 万亿次的运算速度。国家天文台已经能自行搭建近十 TB 数据容量的数据服务器，网络速度也将很快从目前的 100Mbps 升级为 1000Mbps。

China-VO 与国内 IT 业界已经建立了良好的合作关系。在 863 计划专项资助的中国国家网格项目中，China-VO 与中科院网络信息中心共同承担了“科学

数据网格”子课题的研究工作。此外，China-VO 与中科院计算所、清华大学计算机系等单位也建立了良好的合作关系。

3.3 主要工作内容

虽然近年来我国的天文研究事业和 IT 技术产业得到了飞速的发展，取得很大的成绩，但不得不承认我们目前的研究和技术水平与世界先进水平还存在相当的差距。从国内现有的基础出发，遵循“一切从实际出发，有所为有所不为”的指导思想，China-VO 的主要工作内容将主要涵盖如下几个方面：

- 配合 IVOA 的国际伙伴制定虚拟天文台相关标准；
- 熟悉、掌握、应用、开发虚拟天文台相关标准和技术；
- 开发中国虚拟天文台门户；
- 实现国内主要天文观测数据和观测项目与 VO 的兼容。

上面这四个方面的工作基本上是按照在 China-VO 项目研发过程中开展的先后顺序为序给出的，但它们之间又是并行的。当前我们的工作主要集中在前两个方面。当虚拟天文台的相关标准制定工作基本完毕，相关技术逐渐成熟后，工作的重心便会转到后两方面的任务上。

- **配合 IVOA 的国际伙伴制定虚拟天文台相关标准**

除了直接采用和借鉴 IT 界的许多标准，为了实现 VO 既定的奋斗目标，IVOA 社会还必须结合天文学自身的特点和 VO 的具体工作要求制定一系列标准和协议。当前主要是解决资源注册、数据访问和互操作性问题。根据目前的情况，China-VO 很难在这些标准的制定过程中发挥领导作用，但我们需要尽自身可能积极的参与进去。这其中必须要由 China-VO 解决的一个问题是 IVO 对中文的支持。

由于我国是一个非英语国家，按照语言习惯，中国虚拟天文台的用户可分为英文用户和中文用户两大群体。一方面，为了融入 IVO 大家庭，为国际用户提供服务，China-VO 必须把现有的中文文献资料、观测数据转化为英文并提供相应的英文用户界面，即“国际化”。另一方面，为了防止因为语言障碍而把国内广大用户挡在门外的现象出现，China-VO 必须承担起把丰富的英文资源转化为中文的使命并同时提供相应的中文用户界面，即“本地化”。

- **熟悉、掌握、应用、开发虚拟天文台相关标准、技术**

虚拟天文台是一个全新的概念，无论是其从 IT 业界采用的技术还是自己定义的标准都需要学习、掌握的过程。把这些技术成功的应用和实现也需要很多的努力。

为了实现以“VO-enabled LAMOST”为特色的 China-VO，除了 IVO 共同面临的挑战，我们还必须在下面一些方面进行研发：

- 海量多光纤光谱观测数据的自动处理；
- 光纤光谱谱线的自动提取；
- 光谱自动分类；
- 光谱红移的自动测量；
- 光纤光谱数据与其它类型天文数据的融合；
- 光谱数据的可视化。

- **开发中国虚拟天文台门户**

为国际 IVOA 伙伴提供的 VO 服务开发适合中国用户习惯的客户端程序，让国内用户能方便的访问 IVO 的各种资源。

给 China-VO 定好位非常重要。能否能让国内用户方便的使用国际虚拟天文台伙伴提供的各种资源与服务将在很大程度上决定 China-VO 的成败。根据国内用户的需求和使用习惯做好 China-VO 与 IVO 的接口是 China-VO 工作的一个非常重要的方面。

- **实现国内主要天文观测数据和观测项目与 VO 的兼容**

将国内的天文数据资源与观测、处理资源与国际同行共享。这方面的工作中特别重要的是要实现 LAMOST 与 VO 的兼容，建设 VO-enabled LAMOST。

虽然我们无法领导 IVO 的发展方向，但我们要尽可能多的为国际社会做出贡献。重要的就是为国内现有的天文观测数据提供数据访问接口和相应的分析处理服务，丰富 IVO 的数据资源。其中 LAMOST 的观测数据将非常重要，无论是其数据量还是观测数据的珍贵性都决定了它的价值。实现 LAMOST 项目与 VO 的兼容将是 China-VO 为 IVO 做出的最重要贡献之一。

3.4 项目特色

China-VO 的历史使命是：***作为 IVO 不可或缺的一部分，完成 IVO 的中国部分，引领中国天文学进入数据密集型在线科学研究新时代。***

China-VO 是在现代天文观测模式与最新 IT 技术的共同推动下，力图实现国内、国际天文数据资源的高级共享而提出的一项前瞻性研究计划。无论对于国内天文研究还是 IT 技术，China-VO 都提供了极好的实验场。

相对于欧美这些科技强国而言，China-VO 无论从科技上还是人财物资源上都存在一定差距。China-VO 必须按照“一切从实际出发，有所为有所不为”的指导思想，挖掘自身潜力，建设出自己的特色。

科学上，China-VO 将采用与 LAMOST 项目紧密结合的方式，充分发挥 LAMOST 光学光谱数据中心的作用，建设“VO-enabled LAMOST”，使光谱数据及其相关处理技术成为 China-VO 的核心与特色。

技术上，China-VO 必须最大程度的集中国内各天文研究机构、IT 研究机构和数学等相关领域的人才资源，共同努力实现目标。

此外，非常重要的一点是 China-VO 采用开放的运作方式，与国际上各虚拟天文台计划开展充分的合作，在 IVOA 中发挥积极作用。

3.5 总体实施方案

China-VO 将利用最新的以网格为代表的 IT 技术，实现国内天文数据的无缝透明融合并向 IVO 提供一定数量的 VO 服务，实现与国际 VO 资源的互联共享，同时配合国家重大科学工程 LAMOST 项目和天文创新工程的需求，逐步把自己建成连接国内外天文研究资源的“网关”，为我国的天文学科发展提供重要的基础科研环境。

China-VO 将利用中国国家网格（CNGrid）这一良好桥梁，与国内 IT 领域的兄弟单位合作，抓住“网格”这一新的互联网发展机遇，提高我国的网格研发和应用水平。

China-VO 将基于真实的天文观测数据、科学事实、科学文献和科学思想而将其加工为大众所能理解的形式，让大众享受真实的天文学体验，从而成为宣传科学思想和方法、反对封建迷信的坚强阵地。

虚拟天文台是天文学未来一个长远的发展方向。作为国内在虚拟天文台领域的首次尝试，China-VO 第一阶段计划周期为五年。其中，2003－2004 年的主要工作是凝聚技术力量、确定实施方案并检验技术方案的可行性，建立 China-VO 的基本框架。2005－2007 年将在此基础上丰富天文数据资源，充实和完善各种数据访问服务，开发数据挖掘和知识发现等高级服务。

China-VO 整体进度将实行三步走的战略。首先，在国家天文台研发成功 China-VO 试验系统，验证技术方案的可行性。其次，在中国国家网格平台上实现京区范围的 China-VO。最后，把 China-VO 推向全国。

在项目整个建设过程中，China-VO 将始终保持与 IVOA 的密切合作，参与各种协议标准的制定和推广，实现 China-VO 与 IVOA 合作伙伴间的互联。

图 3.1 是计划中的 China-VO 拓扑结构图。国家天文台（NAOC）、国家天文台云南天文台（YNAO）、中科大天体物理中心（USTC）、北京大学天文系（PKU）、清华天体物理中心等天文单位将构成 China-VO 的应用主体。中国

国家网格（CNGrid）拥有强大的计算资源和 IT 技术优势，比如上海超级计算中心（SSC）、中科院计算机网络信息中心超算中心（SCCAS），中科院计算机网络信息中心科学数据网格（CNIC），清华大学高性能计算研究所（THU）等，将成为 China-VO 的大规模计算平台和技术试验床。作为 IVO 不可或缺的一部分，China-VO 将与全球的 VO 项目一道为实现 IVO 的宏伟目标共同努力。China-VO 是面向全社会的，它将利用 Internet 便利的途径与全社会成员共享 IVO 丰富的资源和激动人心的天文发现与研究成果。

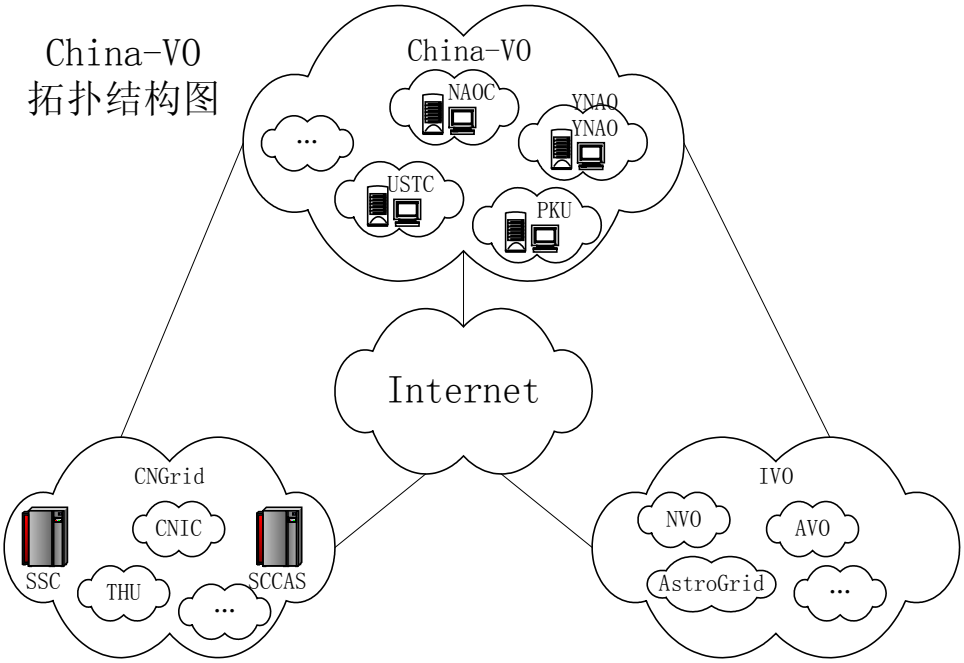


图 3.1 China-VO 拓扑结构

3.6 VO-enabled LAMOST

3.6.1 LAMOST 项目简介及科学目标

LAMOST 项目由中国科学院院士王绶琯、苏定强为首的研究集体倡议，并于 1996 年列为国家重大科学工程项目。项目总投资 2.35 亿元。根据 2002 年底的进展情况望远镜的硬件建设将在 2005 年建成，届时将安装在国家天文台兴隆观测站。

LAMOST 是一架横卧南北方向的中星仪式反射施密特望远镜，应用主动光学技术控制反射改正板，使它兼俱大口径和大视场的特点。

综合了科学目标的要求、望远镜的总体技术指标和观测台址的条件，望远镜建成后应达到的性能参数如表 3.2 所示。

LAMOST 每个晴夜可观测上万个天文目标，每年估计可获得二三百万个天体的光谱，建成后将使我国在大规模天文光学光谱观测研究中占据国际领先地位，为我国在天文学和天体物理学许多研究领域中取得重大科研成果奠定基础。

有效通光孔径	4 米
焦距	20 米
视场角直径	5 度（焦面线直径 1.75 米）
光学质量	80%光能量集中在 2.0 角秒直径的圆内
光纤数	4000 根
光谱覆盖范围	370－900 纳米
观测天区	赤纬从－10 度到＋90 度的 24000 平方度
光谱分辨率	1－0.25 纳米
光谱观测能力	以 1 纳米的光谱分辨率，在 1.5 小时曝光时间内，极限星等达到 20.5 等
实时和离线数据处理能力	

表 3.2 LAMOST 望远镜主要性能参数

天文观测设备的科学价值在很大程度上体现在其能实现的科学目标上。LAMOST 最优先的科学目标是：星系的红移巡天、恒星及银河系结构、多波段天体目标的证认。

- 星系红移巡天

利用星系红移巡天数据开展宇宙模型、宇宙大尺度结构形成与演化、星系的形成与演化、活动星系核、星系团等方面的研究。其中宇宙大尺度结构的研究又是 LAMOST 核心课题中占首位的课题。LAMOST 红移巡天将为我们提供关于宇宙物质的空间分布及演化的一个较为完备的统计样本，它将对宇宙学模型及基本构成很强的限制，并为探索新的理论提供一定的线索，这对宇宙学的发展将是意义深远的。

- 恒星和银河系结构

利用 LAMOST 强大的光谱观测能力对银河系内恒星进行中低分辨率的光谱观测，进而开展银河系结构、动力学、化学丰度、视向速度等方面的大样本统计研究。

- 多波段天体目标的证认

利用 LAMOST 大样本光谱巡天数据与其他地基或空基巡天观测数据进行交叉证认可以发现一大批新的天体，极大地推动人类对宇宙的了解。

3.6.2 LAMOST 输入星表

LAMOST 是一台能同时观测大量天体光谱的高效率望远镜，它使天文观测从每个夜晚仅观测几个到几十个天体，一下跃迁到观测数万个天体。这给天文学家提出了新的、严峻的任务。为了使 LAMOST 建成后能顺利有效地开展研究工作，在望远镜建造的同时必须进行输入星表的准备和巡天战略的研究工作，这是 LAMOST 工程建设中一个不可缺少的组成部分。

LAMOST 所需的输入星表主要包含三个部分：

1. 目标星表（观测对象）
   LAMOST 的观测对象将选自国际上已有的或正在进行中的大视场成像巡天资料。根据不同的观测课题，选取目标，构成目标星表，供 LAMOST 使用。
2. 导星星表
   为了使 4000 根光纤都能准确地对准观测目标，必须要有一批位置精度足够高的恒星，来确定望远镜的指向、跟踪和光纤定位。
3. 光谱流量定标星表
   收集和整理供处理光谱资料时确定流量定标的光谱标准星或多色测光标准星资料。

观测对象选取的原则：

1. 在观测天区内选择标准的一致性，样品的均匀性，而且对所选样品的选择效应需有定量估计，这对大样本统计工作是至关重要的；
2. 同时要得到样品的精确位置，以便光纤定位时使用；按照光纤定位技术要求定位误差小于 0.5 角秒，输入星表中样品的位置误差是其主要因素。
3. 选取样品的参数，必须具有明确的天体物理意义。

输入星表的来源

LAMOST 工程于 1999 年完成了初步设计。在初步设计时，通过对当时国际已有的和正在进行中的大视场成像巡天资料的分析，认为可供 LAMOST 项目使用的输入星表的主要来源有二种，SDSS 巡天和 Palomar 巡天。

1. 利用 SDSS 巡天资料选取样品

SDSS 成像巡天覆盖北天区 π 立体角的天区，对于研究河外天体物理学而言，若取高银纬 b≥30 度，则北天区也只能观测大约 π 立体角的天区。因此除了 Apach Point 天文台与兴隆在纬度上的差异可能引起一些微小差别外，就研究河外天体而言，LAMOST 观测天区必然与 SDSS 的天区完全重合。

2. 利用 POSS 巡天资料选取样品

在 SDSS 巡天的一万平方度以外，LAMOST 尚有一万余平方度的天区可供观测。在 SDSS 天区以外利用 POSS（Palomar Observatory Sky Survey）巡天的资料是目前唯一可供选择的方案。POSS-Ⅱ的优点是它的天区覆盖范围远远大于 SDSS，特别是与南天 UK Schmidt 配合在一起，构成了覆盖全天球的巡天资料，这是当时唯一可供利用的资料。

初步设计时，VO 的概念还没有提出。现在 LAMOST 在输入星表的制定上又多了一个重要的选择。VO 首要的任务就是实现世界上主要巡天数据无缝透明的访问并提供相应的检索、查询、提取、融合等工具以及每个数据集的元数据。这对 LAMOST 输入星表的选取提供了极好的选择和参考。

3.6.3 LAMOST 数据产品

LAMOST顶层数据流图如图 3.2 所示[8]。可以看出与数据产品直接相关的有两个数据处理系统：图像处理系统（图中编号 3）和光谱分析系统（图中编号 4）。

图像处理系统主要有两个任务，一是对当天的观测数据进行质量分析，以确定其是否达到科学要求，形成质量报告文件提交给观测战略软件。另一是对原始观测数据依照实测天体物理学的方法一步步地进行处理，得到每个待观测天体的一维光谱，并存储到光谱数据库中。

光谱分析系统将对 LAMOST 的观测数据进行处理与分析，得到被观测天体的物理信息，这些可提供给天文学家进行科学研究。LAMOST 的观测数据是在每次观测后由三十个 CCD 上采集下来的原始数据，数据量每天近 10GB，每年达数 TB。对这些数据按照一定的步骤和程序进行处理后得到每个被观测天体的一维光谱图（即流量与波长的关系），再对这些光谱进行分析和测量，得到被观测天体的各种物理信息，如类型、红移、视向速度、光谱型、线强比、等值宽度等，从而可直接提供给天文学家进行大样本天文学的科学研究。

从数据流图我们可以看出 LAMOST 的数据产品将包括：

- 二维原始观测数据。这是 LAMOST 望远镜光谱观测的直接成果。按照目前的设计，CCD 的图像读出后将以 FITS 文件的格式保存。其中每个 FITS 文件中包含了 250 条二维光纤光谱，文件大小为 16MB（每像素 16 比特深度）。
- 巡天一维光谱库。红蓝两段二维原始光谱合并，经过 pipeline 处理得到目标天体的一维光谱。这些光谱可能以 FITS 格式保存，也可能以 VOTable 格式或者 ASCII 格式保存。

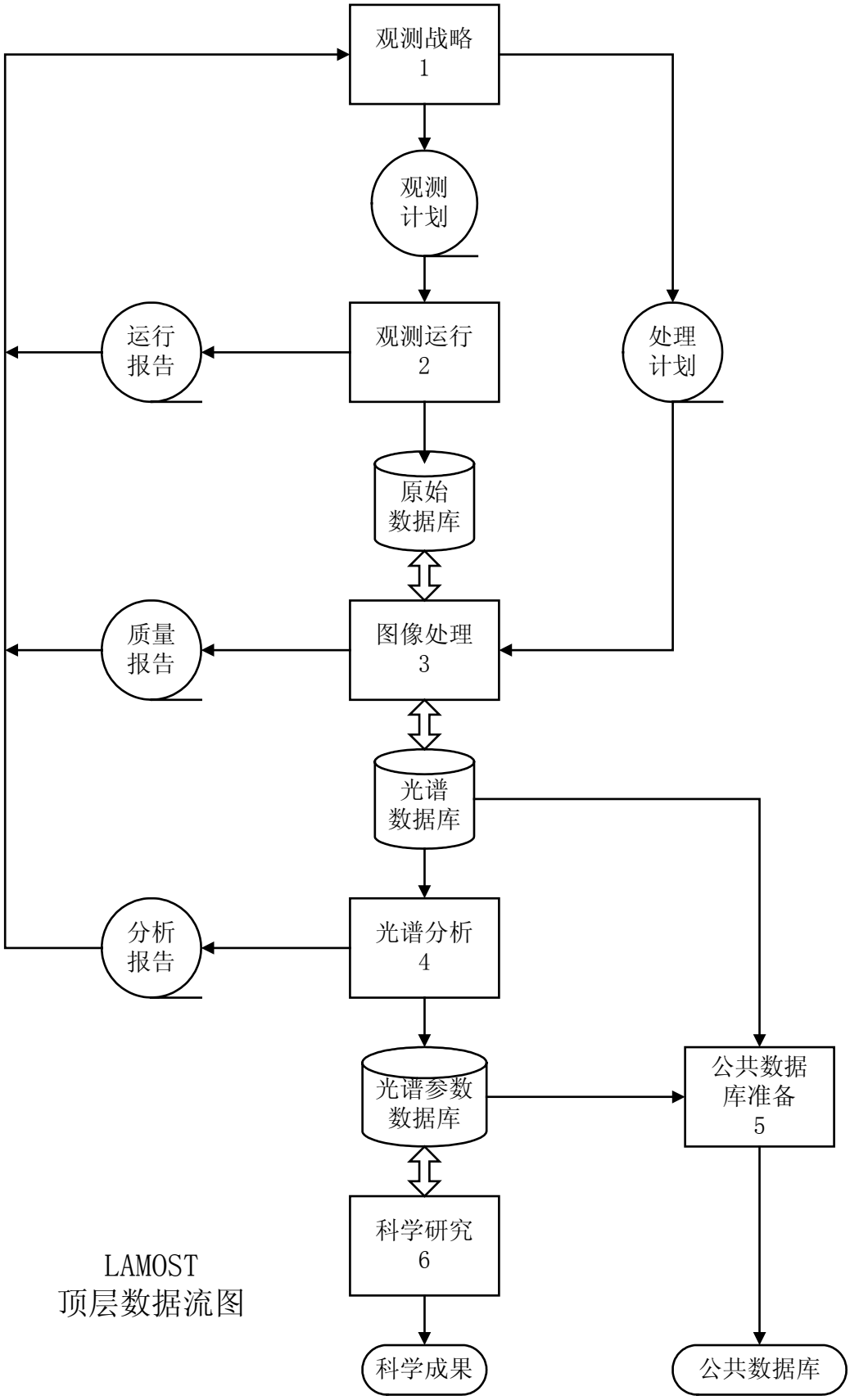


图 3.2 LAMOST 顶层数据流图

- 巡天星表。经过光谱自动处理和分析后得到的包含目标天体各种测量参数的星表，比如天体类型、光谱型、光度级、红移、视向速度等。

此外，经过项目精心准备的三种工作星表，输入星表、导星星表、光谱流量定标星表，也可以为其他天文学家和观测项目所利用。

3.6.4 VO-enabled LAMOST 思想

China-VO 将与大科学工程 LAMOST 紧密配合，把 LAMOST 建成为**“**VO-enabled LAMOST**”**。

“VO-enabled LAMOST”包括两层含义：

第一层含义是指“***VO-enabled LAMOST Data***”。指的是 LAMOST 的工作星表在 VO 技术和资源的支持下产生，同时项目的数据产品通过 VO 进行共享，如图 3.3 所示。

第二层含义是“***VO-enabled LAMOST Telescope***”。实现 LAMOST 望远镜的 VO 化，让其成为 VO 资源中的一个结点。

目前，在 China-VO 的第一阶段“VO-enabled LAMOST”主要指的是第一层含义。

大样本光谱巡天是 LAMOST 工程的本质特点，把 LAMOST 建成为***“VO-enabled LAMOST”***不但是 LAMOST 本身的需要，也将是中国对世界天文的重大贡献。

首先，LAMOST 的观测目标从虚拟天文台中选取。LAMOST 是对选定目标进行光谱巡天的望远镜，输入星表的选取决定了 LAMOST 的科学产出。虚拟天文台是最富有的天文数据拥有者，LAMOST 从中提取输入星表是极好的途径。

其次，LAMOST 的观测结果，将成为虚拟天文台的一个重要组成部分。目前正在进行中的大型巡天观测，大部分是成像观测，为了深入研究天体物理过程，大量光谱资料的获得将成为天体物理学发展的瓶颈。可同时获得 4000 个天体光谱的 LAMOST 望远镜建成后将成为世界上威力最大的光谱巡天望远镜，LAMOST 数据中心也将成为世界光学光谱数据中心。

再次，LAMOST 望远镜观测获得的海量光谱数据不可能依靠传统的数据处理方法进行处理，必须开发一套自动处理和分析工具，比如光谱处理、谱线提取、天体自动分类、红移测量、统计分析等工具。这些工具对于实现虚拟天文台的数据挖掘和知识发现功能将起到重要的推动作用。

另外，LAMOST 与虚拟天文台一样都面临 TB、PB 量级数据的归档和管理

问题。LAMOST 可以借鉴虚拟天文台的数据管理经验，同样 LAMOST 的自身经验也可为虚拟天文台所借鉴。

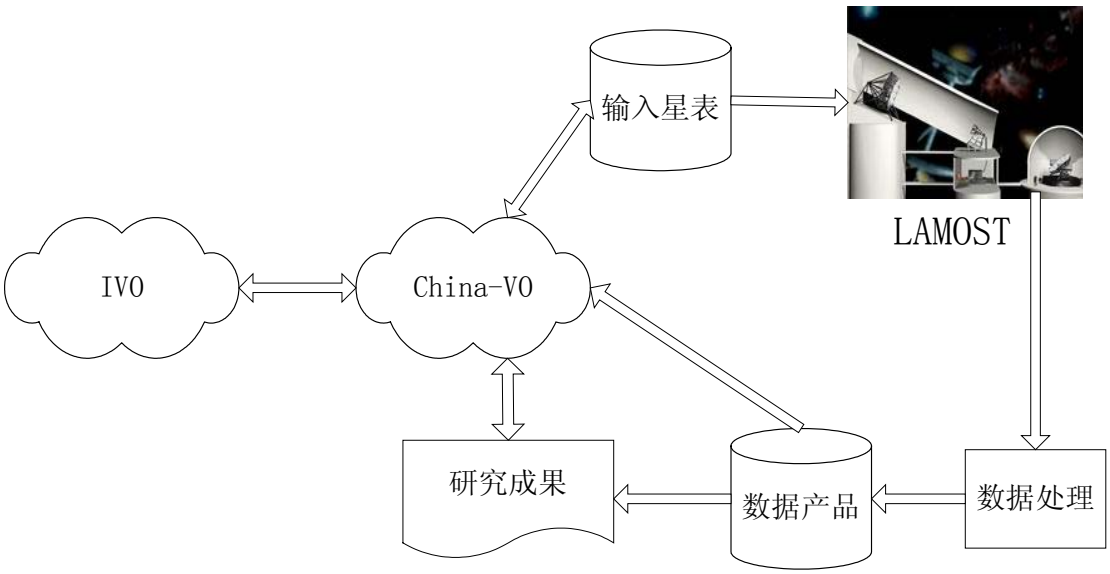


图 3.3 VO-enabled LAMOST

LAMOST 工程是目前我国天文学界唯一的大科学工程，它的观测数据将成为 China-VO 的重要基础，为国际虚拟天文台和天文学的发展做出贡献。

参考文献

[1] [IVOA] International Virtual Observatory Alliance Member Organizations. http://www.ivoa.net/pub/members/

[2] [LAMOST] Large Sky Area Multi-Object Fiber Spectroscopic Telescope. http://www.lamost.org

[3] [SST] Space Solar Telescope. http://www.bao.ac.cn/bao/SST/index.html

[4] [FAST] Five hundred meter Aperture Spherical Telescope. http://www.bao.ac.cn/bao/LT/

[5] [China-VO] Virtual Observatory of China. http://www.china-vo.org

[6] [BATC] Beijing - Arizona - Taiwan - Connecticut 大视场 CCD 多色巡天. http://vega.bac.pku.edu.cn/batc/index.htm

[7] [WDC-Astronomy] World Data Center for Astronomy. http://badc.lamost.org

[8] LAMOST 初步设计文档. http://www.lamost.org/design.htm

第四章 China-VO 体系结构

体系结构也称为体系架构，为整个系统提供了一个结构、行为和属性的高级抽象，由对构成系统的元素的描述、元素间的相互作用、元素集成的模式以及这些模式的约束组成。体系结构不仅指定了系统的组织结构和拓扑结构，并且显示了系统需求和构成系统的元素之间的相互关系，提供了一些设计决策的基本原理。

目前国际上的 VO 项目中英国的 AstroGrid 和美国的 NVO 对体系结构进行了较多的讨论。本章将首先对这两个项目在这方面的研究状况做一介绍，然后参考 OGSA 的体系结构给出 China-VO 体系结构模型。

4.1 AstroGrid 体系结构设计

体系结构的设计对于每个工程项目都是一项重要的工作内容。各国VO项目都需要结合各自的资源特色和技术特色设计体系结构。英国AstroGrid 2002 年底公布的A阶段的总结报告（红皮书）[1]对他们的体系结构设计工作进行了描述。

AstroGrid 在体系结构设计上采用了统一过程方法（Unified Process methodology, UP），并对这种方法进行了扩充。

他们从项目内外的天文学家手中征集了许多便于显示 VO 特点的科学问题作为范例。为了说明 VO 如何完成范例中描述的科学任务，AstroGrid 完成了一个“概念模型”。这个模型将整个 AstroGrid 系统或称为“AstroGrid 域”分成了许多“实体”。模型阐述了各实体的功用及其相互关系。利用这个“概念模型”可以为一些典型的科学范例生成工作流程。这对了解 VO 的工作原理是十分重要的。最后，通过这个模型可以了解 AstroGrid 要建立的是怎样的一个 VO 以及如何细化下一步的设计。

图 4.1 是AstroGrid红皮书中给出的概念模型。这个概念模型是在项目的早期用Together[2]建模软件绘制的。这个模型将AstroGrid域中的主要实体组合在一起。其中的许多实体将进一步发展成为功能健全的类或者服务。

整个系统围绕四个主要实体展开。这四个主要实体包括：

- DataObject （数据对象）：一个非常广义的数据对象，涵盖从一个单一的 FITS 文件到整个数据集。

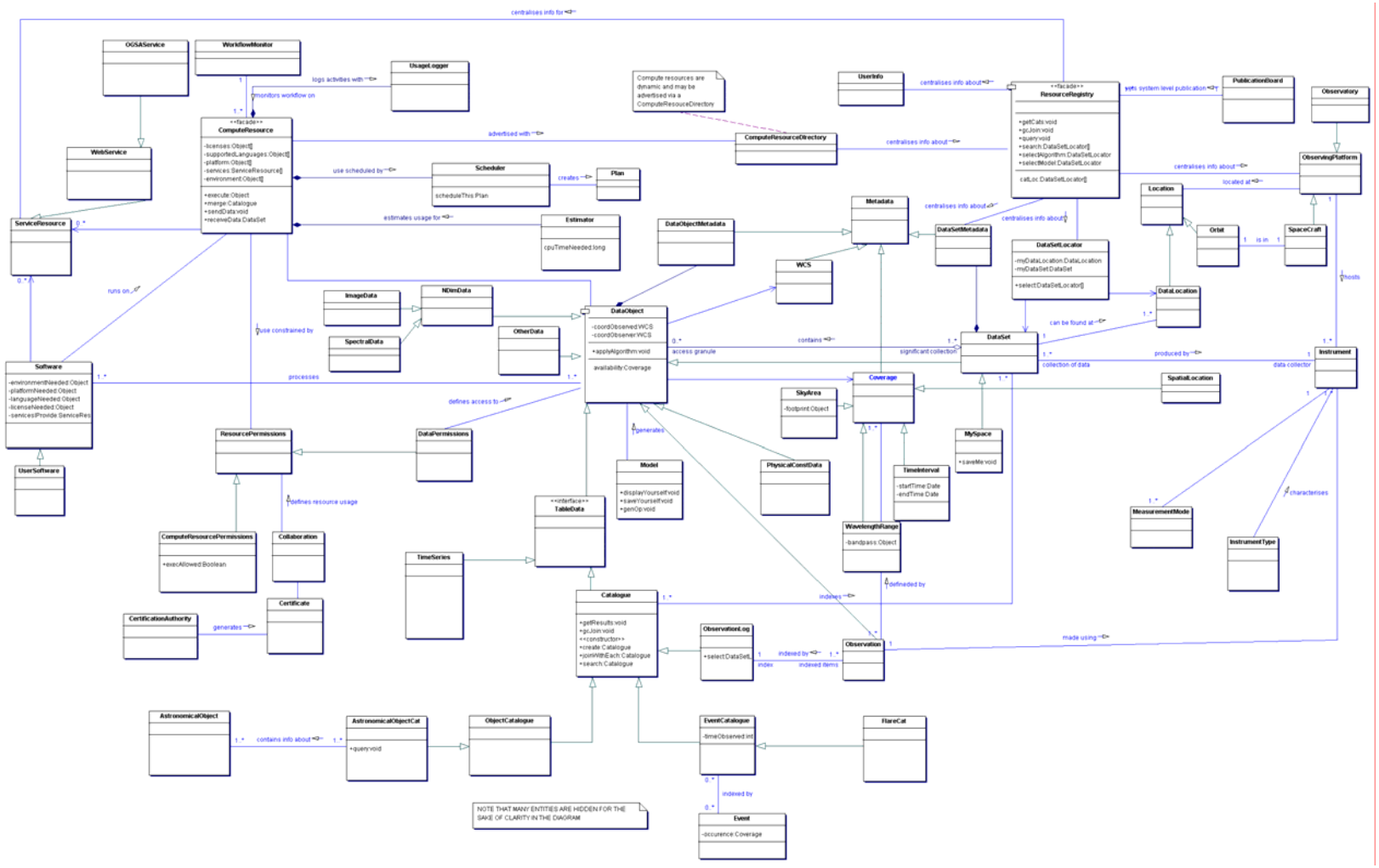


图 4.1 AstroGrid 概念模型

- DataSet （数据集）：DataObject 或者 DataObject 的集合，作为一个资源进行注册。
- ResourceRegistry （资源注册）：一种注册，可以直接或间接的定位需要的资源。这些资源有可能是一个特定的数据集或者是一个特定的软件服务。
- ComputeResource（计算资源）：一种封装形式，提供对所有形式的 CPU、软件、网络和存储资源的访问。每个对象的任何动作都会与某个计算资源相关。大多数的处理请求必须首先得到可用的计算资源，然后才能访问感兴趣的数据对象。

经过考虑，AstroGrid 决定采用 Web 服务的观点，利用 GT3 架构。 但同时他们强调要保证让 AstroGrid 的组件可以独立于 OGSA 或者其他的网格环境而正常工作。为了满足 OGSA 面向服务的体系架构的要求，他们完成了 AstroGrid 组件定义及其组件之间的消息传递方案草案，即“服务模型”（Services Model）。

概念模型是一个整体模型或者说逻辑模型，它将 VO 看成是一个单一的系统。系统设计人员在项目初级阶段用它来分析系统的动态行为，而不用关心组件的具体实现。服务模型体现了 OGSA 面向服务的设计理念，将整个系统看成是一系列服务的集合。每个服务实现一定的功能，以 WEB 服务或者网格服务的形式体现出来。

红皮书中给出的 AstroGrid 服务模型将 AstroGrid 域分成了数十个功能模块。除了门户“Potal”和客户程序不以网格服务的形式体现外，其他各部分都将以网格服务的形式体现。这个服务模型中列出的服务绝大多数都是高级的复合服务或者服务集，它们将建立在更多的低层服务和原子服务的基础上实现这些高级的功能。

在 OGSA 的框架下，每个服务都利用经标准化描述的接口和行为与其他服务交换信息。每个服务都是相对独立的，可以利用模块化的思想进行开发。

AstroGrid服务模型的设计工作才刚刚开始，目前他们已经确定的基本服务有[3]：

- 活动日志（Activity Log）
- 分析工具（Analysis Tools）
- 应用资源（Application Resource ）
- 天文查询语言翻译器（AQL Translator）
- CAS 服务器（Cas Server）
- 计算资源（Compute Resource）
- 数据挖掘（Data Mining）
- 数据库输出（Database Export）

- 数据集访问（Dataset Access）
- 数据路由（Data Router）
- 作业控制（Job Control）
- 作业估计（Job Estimator）
- 作业调度（Job Scheduler）
- 用户空间（MySpace）
- 查询估计与优化（Query Estimator/Optimizer）
- 备份管理（Replica Builder）
- 资源注册（Resource Registry）
- 用户通告（User Notification）
- 用户偏好设置（User Preferences）
- 工作流（Workflow）

图 4.2 简要的描绘了这些服务之间的关系。

除了这些服务，AstroGrid 计划开发一个基于 WEB 的门户“Portal”和一个基于 PC 的客户程序“Client”，以便让用户能发现 VO 的资源和向 VO 提交作业。此外，他们还计划开发一个基于 WEB 的日志分析器“Log Analyzer”进行资源分析。

4.2 NVO 体系结构研究

到目前为止，NVO 还没有给出一个明确的体系结构。这其中一个重要的原因应该是 NVO 的机构组成过于复杂，涉及数十个大学、国家实验室和公司。不同的机构都有自己的运作机制和技术优势，有些还处于相互竞争的状态，很难形成一个统一的体系，是一个真正“异构”的系统。

今年初，在NVO刚刚完成的三个demo中就采用了不同的实现途径[4]。仅就前端界面来说，“伽马射线暴事件追踪”采用的是PERL编写的CGI程序；“褐矮星候选体搜寻”用到了CGI和Java Servlet；“星系形态学分析”用到了ASPX。JHU与微软合作，在.Net平台上开发WEB服务。SDSC的数据网格则架构在SRB技术上。

NVO 项目内 WEB 服务、数据分析、数据网格等方面的研究工作分别由不同的人员和组织负责。他们之间的协调程度将直接影响 NVO 的体系结构。

不过，NVO在其 2002 年第四季度的报告[5]中已经明确表示“NVO的系统设计将在很大程度上集中于正在浮现的网格服务”。在其 2003 年第一季度的报告[6]中，NVO建议的体系组成如下：

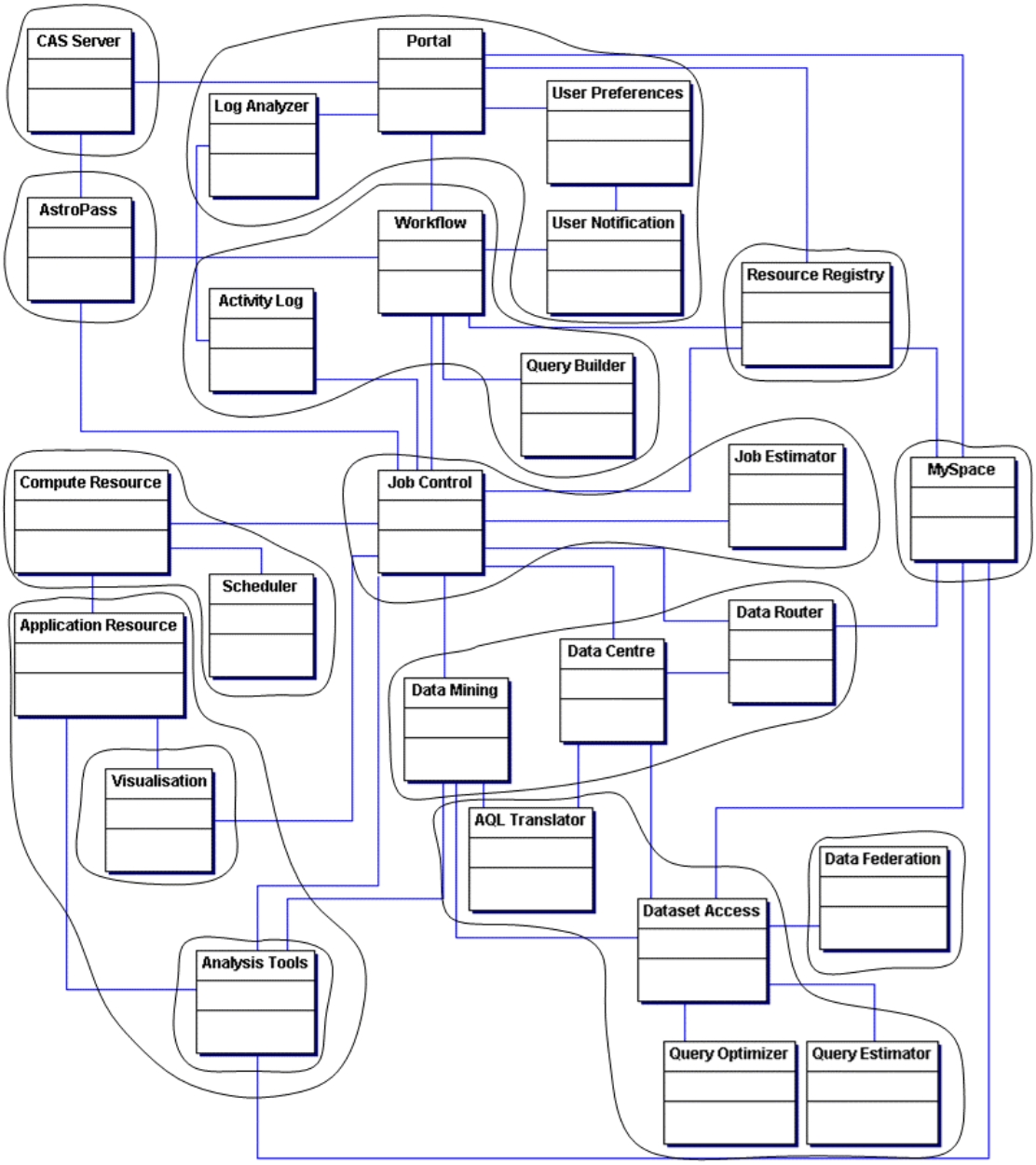


图 4.2 AstroGrid 服务模型

- 门户 – 用户作业分析的 WEB 服务接口 (OASIS)
- 作业管理系统– 能产生衍生数据产品的数据批处理程序 (Chimera, Montage)
- Web 服务 – 跨系统的星表和图像统一访问服务 (ConeSearch, VOTable catalog query, simple image access)
- 数据访问层 – 数据编码格式管理以实现基于物理单位的数据访问 (UCDs)
- 数据网格 – 分布式数据集管理，逻辑名字空间管理(SRB)
- 计算网格 – 分布式计算资源 (Globus toolkit)
- 持久数据 – 技术发展管理 (SRB)

- 星表和图像数据 (SDSS, 2MASS, DPOSS, USNO-B, MACHO)
- 持久磁盘系统– 对巡天图像数据的交互式访问 (Grid Bricks)
- 高性能磁盘缓存– 用于大数据分析的高速访问 (SAN)
- 计算平台 – NSF Teragrid

可以看出上面的组成方案仅是个粗略的考虑，并没有一个系统的划分标准和整体结构。

在系统设计中他们目前有几个关心的问题。

NVO 认为在基于 OGSA 进行体系结构设计时一个重要的挑战是决定“在网格服务体系中知识管理服务在哪里实现”？对虚拟天文台数据实体操作所需的知识或者说元数据可以封装到每个文件中，或者在数据访问层，或者在信息目录中，或者在注册中心，或者在一个单独设计的知识管理层。这个问题的实质是元数据服务和资源的注册与发现功能定位在何处？信息与知识服务是否需要分级实现，如何分级？

NVO 对网格技术主要的担心主要是因为网格体系尚不成熟，处于快速变化之中。在过去的三年中网格技术进行了三次体系结构的迁移，从基于库的环境到基于联合的客户端/服务器环境，再到网格服务环境。网格服务将基于OGSA，但这是个很不成熟的标准。OGSA 是 SOAP/WSDL 的一个扩展，提供了实例化临时服务和认证与访问控制等功能。关于这些扩展的规范尚没有制定完成。NVO 希望 GGF 能尽快提供一个健壮、稳定的版本。

按照目前的计划，OGSA第一个正式标准将在 2003 年 6 月 24 至 27 日在美国西雅图举行的全球网格论坛（GGF）第八次会议[7]上发布。Globus也计划在今年 6 月发布GT3 的正式版本。

我们期待今年下半年，在 OGSA 标准和 GT3 正式推出后能对 VO 的体系结构设计有较大的推动作用。

此外，NVO 还担心 OGSA 能否提供 NVO 需要的安全和访问控制功能。目前 NVO 的 WEB 服务开发仍是基于 SOAP 和 WSDL。

4.3 VO 工作流程

实施虚拟天文台所面临的技术挑战是如何来解决许多相互对立的需求。一方面数据是广泛分布的，另一方面虚拟天文台的大型科学前沿研究需要巨大的计算资源和快速的数据访问；一方面要采用复杂的元数据标准和访问协议将分布的数据集和网络服务连为一体，另一方面还要使小型数据集接入虚拟天文台简便可行，以鼓励人们发布新收集的数据；一方面数据集和计算服务是广泛分

布的，另一方面还需要标准的系统接口以使数据和服务的定位和存储资源发现尽可能地透明。

体系结构的设计就是要充分认识到各种不同使用模式的重要性，使用最简单的结构来实现所需要的功能。在体系结构设计过程中要：

- 明确实现数据融合和大规模数据计算对网格技术的要求；
- 明确要实现跨数据集的发现对注册与元数据服务的要求；
- 明确 VO 的一些核心、基本服务。

为了比较清晰的了解 VO 的工作原理，我们以 NVO 最早实现的三个原型之一，褐矮星候选体的搜寻，作为范例阐述 VO 的任务实现过程。

近年来，深度巡天计划的开展，比如 2MASS、SDSS，使得褐矮星候选体的搜寻研究工作取得了突破性的进展。其中一个关键性的因素就是多波段巡天数据的联合应用。当前，巨型星表之间的交叉证认工作进行的特别费时费力，是天体物理学研究中经常面临的典型问题。

范例，褐矮星候选体的搜寻，利用 SDSS EDR 光学巡天数据和 2MASS 红外巡天点源星表作为数据源，通过基于 WEB 的 VO 服务实现两个大规模星表之间的交叉证认，寻找 SDSS EDR 星表中天体的红外对应体并从中寻找褐矮星候选体。

通过 VO 来实现这样的证认工作有两个突出的优点。第一，可以免去用户下载大量数据的烦恼。VO 服务程序可以直接对分布式的数据进行操作。第二，VO 中的交叉证认服务在算法和软硬件配置等方面进行了优化，可以在数分钟内将结果返回。如果用户在自己的桌面系统中进行同样的操作可能要耗费很长的时间，几个小时甚至几天。

NVO 的褐矮星候选体搜寻演示已经取得了可喜的科学成果。在 2003 年 1 月美国西雅图举行的 AAS 第 201 次会议上 NVO 公布的 5 个最有可能的候选体中已经有一个经凯克望远镜光谱观测证实是褐矮星。通过对 2MASS IDR2 和 SDSS EDR 交叉证认，这个 Demo 得到了 326, 020 个对应体。经过进一步的分析，最后公布了 5 个最有可能的候选体，分别是：

- 2MASSI J0016084-004301
- 2MASSI J0104075-005328
- 2MASSI J0229279-005328
- 2MASSI J1326298-003831
- 2MASSI J1346464-003150

经凯克望远镜观测证实的是第二颗。其中后面两颗是早已经被确认的褐矮

星。

范例在 VO 中的工作过程，大致可以描述如下。

1. 用户登陆 VO 门户，进入自己的工作室（MyVO）。这里需要进行身份证认。根据用户名确定用户类型及其相应的访问权限。当然 VO 也会允许公开访问，但访问权限与 VO 正常用户相比会有所限制。
2. 用户在线编写、制定工作计划并提交。修改任务和工作计划并在工作室保存。当用户设定的任务提交时刻到来时，工作室负责向 VO 门户提交工作任务。
3. VO 门户分析作业，查找并选取用以完成任务所需的数据资源和服务。
4. 如果任务可行，则接受用户作业。如果不可行则告知用户无权执行此项作业或者找不到合适的资源完成作业并把信息反馈给用户。这里，需要利用 VO 信息和知识管理系统对所提交的问题进行合理性判断；利用数据/资源注册机制获得可用资源信息；核实用户对资源的访问权限；根据可用资源情况编排工作计划，估计任务执行的工作量和其他相关信息。
5. VO 门户针对所选取的不同数据服务提供者（Data Service Provider，DSP）的数据描述生成 VO 查询语句（VOQL）并分别提交到相应的数据服务提供者。
6. DSP 接受 VO 门户提交的 VOQL，根据当前资源的使用情况进行作业调度，给出作业调度报告。
7. VOQL 解析。DSP 利用数据集元数据信息将 VOQL 语句转换为与其（虚拟）主机环境对应的查询语句。VOQL 解析过程需要用到 VOQL 解析器。
8. DSP 数据访问服务执行查询操作，向 DSP 返回查询结果
9. DSP 将查询结果转换成 VO 制定的统一的数据交换格式，比如 VOTable、FITS 等，返还给 VO 门户。
10. VO 门户远程查询结果返回的过程中从 VO 资源管理系统接受系统检测数据，跟踪作业的执行情况并及时通告用户。
11. 当所需远程结果返回后，进行质量评估。
12. 对来自不同 DSP 的查询结果利用 Ontology/UCD 进行必要的格式转换。然后调用相应的数据处理工具（数据挖掘、统计分析）。这里需要本体（Ontology）和格式转换服务来完成转换功能；数据处理过程还涉及数据缓存。
13. 最后对处理结果进行必要的整理，比如可视化，然后返回给用户。
14. 任务执行过程中，VO 系统持续监听用户的请求和系统各部分的情况并

做出相应的响应。如果在任务执行期间用户主动提出暂停或者终止的请求，VO 门户要向相关的 DSP 或者数据处理等服务发出消息，终止作业，释放所占用的资源。如果系统的某个服务或者资源出现变化，比如服务访问被终止或者出现硬件故障，VO 也将及时通告用户或者将任务转移到其他的服务上。这需要有保存中间结果的功能，暂停服务的重新激活功能，检查点设置等。

上面的流程显示出 VO 的工作过程大致可以分为四个阶段：

1. 作业提交：这其中包括用户登录，身份验证，工作计划编制，提交等操作。
2. 作业调度：其中可能涉及资源和服务的发现，访问权限认证，任务估计，系统监控等。
3. 作业处理：其中可能涉及数据集访问，计算服务，数据挖掘等。
4. 结果返回：其中可能涉及数据编码，格式转换，可视化等。

4.4 体系结构

从基本的功能模块角度来分析，China-VO 系统结构组成将如图 4.3 所示。

整个体系结构分为四层，从下到上依次是构造层、资源层、汇集层和用户层。

构造层是整个虚拟天文台系统的资源基础，其中包括各种数据资源，计算资源，网络资源，存储资源等。各种数据资源在虚拟天文台这样一个数据密集型在线研究平台中占有非常关键的作用，是 VO 成功运作的基础和前提。它主要包括星表、星图、光谱、时序数据、计数测量数据、模拟数据、多媒体数据、天文文献等。

资源层将以开放网格服务架构（简称 OGSA）为基础，配合其他网格系统服务工具，利用标准的数据模型和服务模型，通过抽象化实现统一的数据访问和统一的计算访问以及网格系统管理等功能。前面提到的数据访问层的功能将在这部分实现。这里，系统管理主要涉及作业管理、安全管理、资源状态管理、数据管理等。

汇集层包括最能体现天文特色的各种 VO 服务，比如数据处理、数据挖掘、统计分析、可视化等应用服务。当 OGSA 体系架构及其实现工具成熟以后，这些服务的开发和发布将是 VO 建设的重点。

用户层，包括 VO 客户端服务和 VO 门户，是整个体系的最高层，直接与虚拟天文台用户接触。用户层的基本职能是用户任务提交和处理结果返回，主

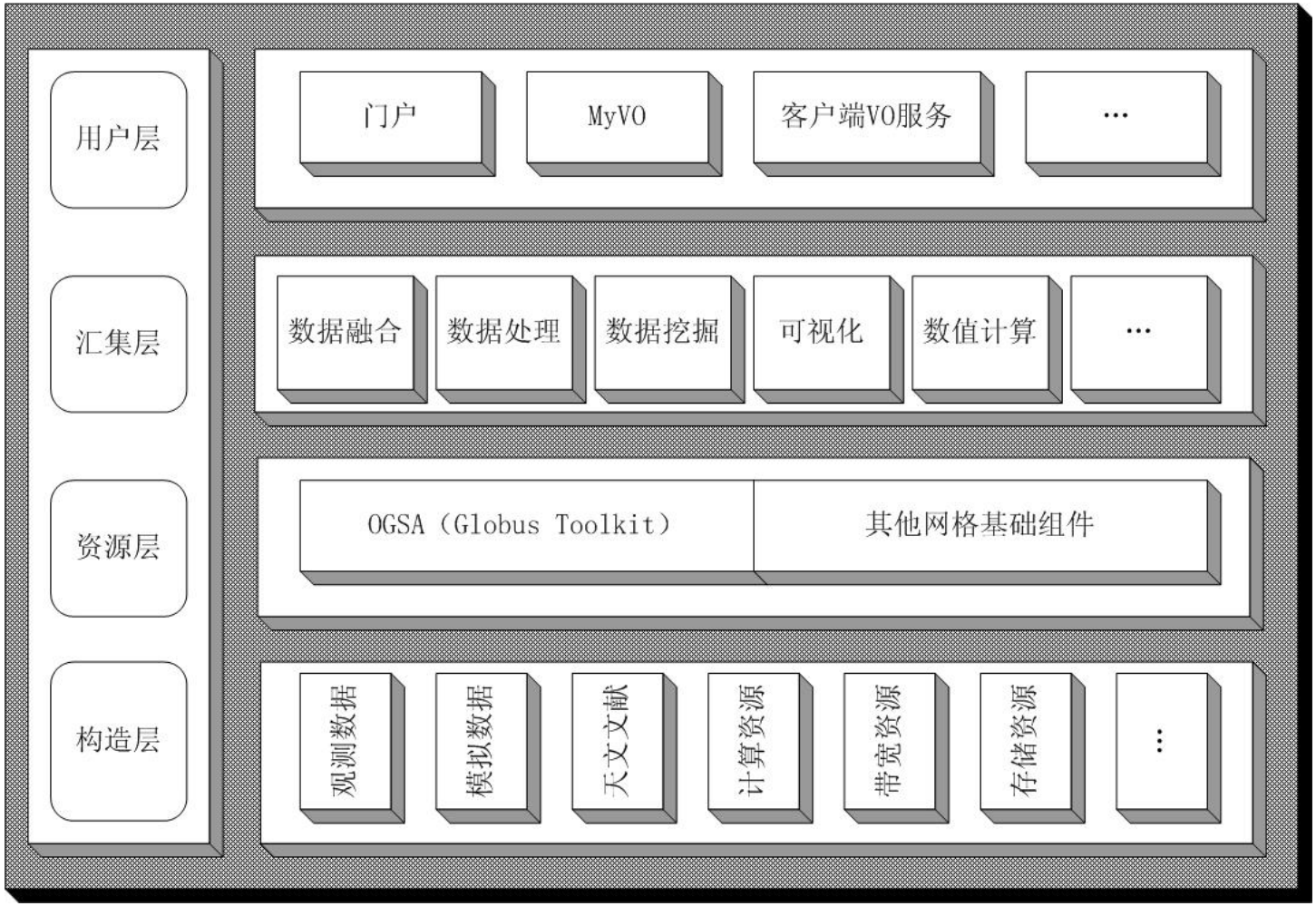


图 4.3 中国虚拟天文台系统结构

要功能包括用户登录、身份认证、VO 资源浏览、任务编制和提交、结果显示、数据下载、偏好设置等。

China-VO 的体系结构建立在 OGSA 的基础之上。物理上，整个系统是分布式的，在网络环境下实现的；逻辑上，通过网格操作系统的管理，它是一个统一的整体。

4.5 服务模型

体系结构可以从多个方面来设计。现在，我们换一个角度来审视整个 VO 系统。其实在上节中我们已经提到了许多 VO 需要实现的功能，比如数据访问、数据处理、数据互操作、资源发现等。为了进一步明确 VO 的核心服务以及对网格平台的要求，下面我们按照 OGSA 面向服务的设计理念来重新分析 VO 系统。

OGSA 是一个面向服务的体系。在整个网格环境中所有组件都是以服务的形式来体现的。服务可以是不依赖于其他服务而独立存在的原子服务，也可以是建立在其他服务之上的复合服务。不管是原子服务还是复合服务，它们最基本的共同点就是在网格环境中可以提供某种功能。

根据上面两节内容的阐述，我们将 VO 系统中需要用到的服务或功能综合如下。

数据访问

- 数据库的简单访问：SQL92 直接支持的访问，比如简单的数据库插入、更新、检索；
- 数据库的高级访问：比如统计分析型访问、计算型访问、交叉证认等；
- 文件访问：单个文件的访问，文件系统的访问；
- VOQL 的生成、解析和优化；
- 数据整合：不同数据集查询结果的整合；
- 数据迁移：不同数据类型（数据、图像、流媒体）利用不同的迁移机制（GridFTP、可靠文件传输服务 RFT）在不同地点或不同服务间的传输；

数据挖掘

- 分类：目的是提出一个分类函数或分类模型，利用该模型把数据库中的数据项映射到给定类别中的某一个；
- 聚类：根据数据的不同特征，将其划分为不同的数据类，使得属于同

一类别的个体之间的距离尽可能的小，而不同类别的个体间的距离尽可能的大；

- 相关分析：目的是发现特征之间或数据之间的相互依赖关系；
- 偏差分析：发现观测结果与参照量之间有意义的差别。通过发现离群数据可以发现一些不同寻常的或奇异的天体。

计算服务

- pipeline 处理：海量数据的自动处理是 VO 应该提供的服务
- 光谱处理：包括光谱的自动分类和自动测量
- 动力学计算：比如星系演化、日月食、天体轨道的计算等
- 数值模拟计算：N 体模拟，宇宙大爆炸过程模拟等

可视化服务

- 二维和三维空间的散点图、直方图、曲线图、轮廓图，馅饼图
- 一维和二维光谱显示
- 常用天文图像的显示
- 动画模拟
- 流媒体播放

数据转换

- 数据编码、数据解码
- 物理单位换算
- 坐标换算
- 历元换算
- 图像格式转换
- 数据压缩与解压缩

注册与发现

- 数据集的注册与发现
- 应用服务的注册与发现
- 物理资源（计算资源、存储资源、网络资源等）的注册与发现
- 数据路由
- 服务路由

元数据服务

- 天文学本体
- 元数据目录

- 国际化多语种支持

安全服务

- 身份认证
- 访问授权
- 单点登陆
- 访问代理
- 安全策略

资源管理

- 状态管理
- 资源动态上下线管理
- 数据备份
- 数据缓存

作业调度

- 作业估计
- 负载均衡
- 工作流管理
- 断点管理
- 服务的重新获取

系统监测

- 系统日志
- 性能监控
- 日志分析
- 系统通告
- 用户通告

MyVO（我的 VO），VO 的个性化服务

- 活动日志
- 任务编制
- 中间结果和常用数据存储
- 及时通告
- 偏好设置

这其中有些功能可以由符合 OGSA 标准的 Globus Toolkit 为代表的网格操作系统来提供，比如资源管理、作业管理、系统检测、安全服务等，其余则需

要由 China-VO 提供。不过由 China-VO 提供的服务还可以分为两种情况：某些功能服务是天文学研究所特有的，此类服务必须由 VO 开发者独立提供，比如元数据服务、数据交换格式编码、pipeline 计算等；而还有一些功能服务是许多科学研究领域甚至非科研领域所共同需要的，比如可视化、数据挖掘、统计分析等。对于这类服务，China-VO 将借鉴其他领域和行业的现有工具和实现方案，经过改造加工后融入到 China-VO 中。

China-VO 是一个数据密集型的在线研究平台。为了实现数据密集型在线研究这个目标，它必须实现三个方面的基本功能：数据访问、数据处理、数据互操作。这三个基本任务在 VO 系统中可由三个角色来承担：数据服务提供者（Data Service Provider, DSP），应用服务提供者（Application Service Provider, ASP）和注册（Registry）。把上面列出的 VO 系统的基本功能服务按照所处角色的不同将其分配到 DSP、ASP 和 Registry 中，便得到 VO 的服务模型。图 4.4 给出了 China-VO 的服务模型。

从模型中可以看出，资源管理、系统监控、安全服务、任务调度对所有角色都是必须的。但这部分对所有网格应用都是必须的，将由符合 OGSA 标准的网格操作系统来提供。这部分内容将不在论文中进行详细讨论。

虽然元数据服务对所有角色也是必须的，但由于这部分涉及许多天文学的学科特点和内容，所以 China-VO 将参考 OGSA 体系提供的元数据服务机制结合天文学的需求给出 China-VO 的元数据服务。资源注册与发现、数据访问、计算服务、数据挖掘、可视化、MyVO 等服务则将由 VO 社区独立开发或者借鉴其他学科领域的实践经验来实现。

除了上面这些 VO 服务，还有两个比较特殊的部件，即 VO 门户和客户端程序。它们可能不以网格服务的形式存在，而利用现有的 WWW 技术，比如 Java Servlet、Java Applet、PHP、ASP 等为基础开发。

在论文以后的部分中我们将重点讨论由 China-VO 独立（包括借鉴）提供的服务和功能。

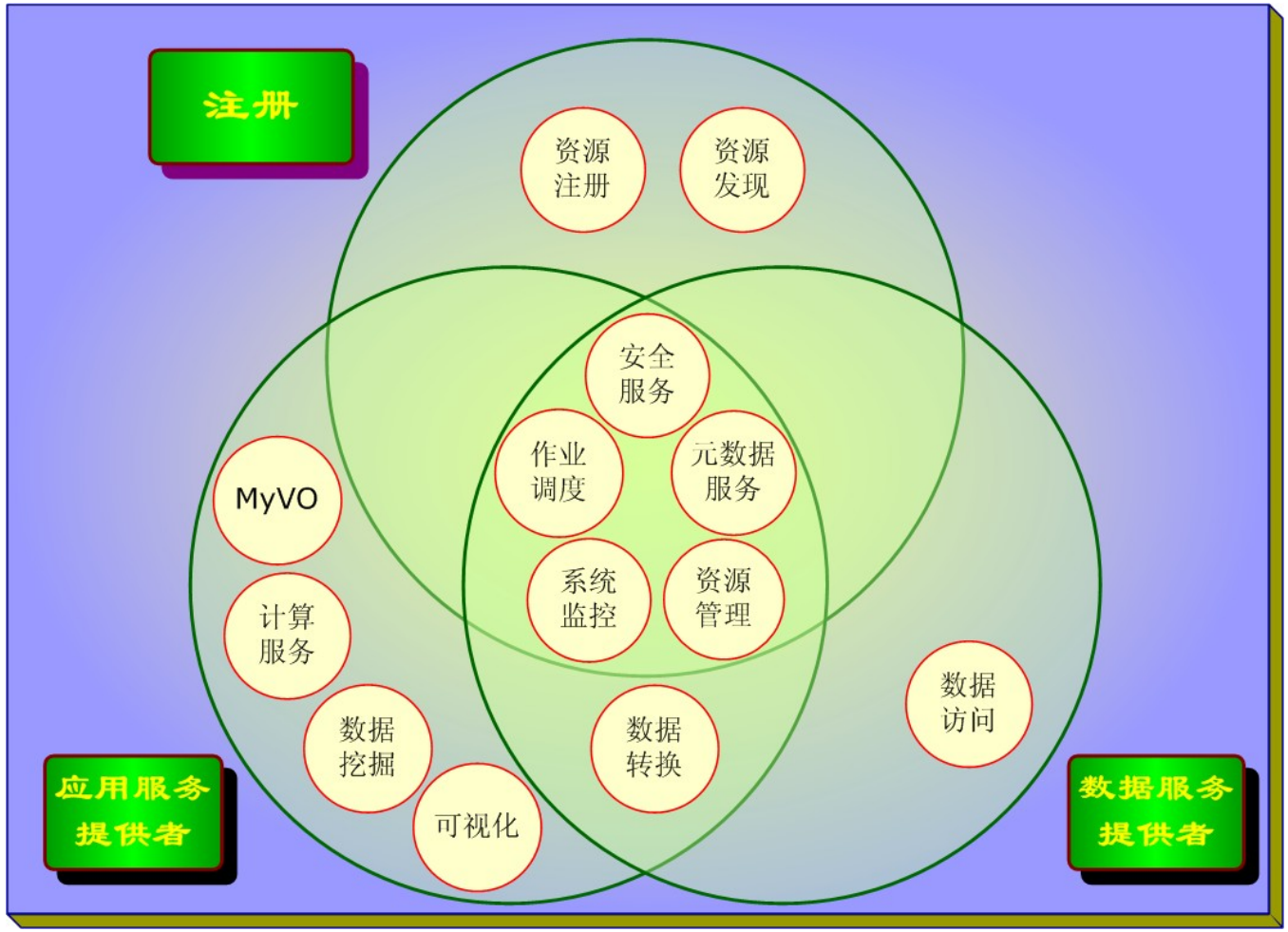


图 4.4 中国虚拟天文台服务模型

参考文献

[1] [RedBook] The AstroGrid Phase A Report. http://wiki.astrogrid.org/pub/Astrogrid/PhaseAReport/redbook.pdf

[2] [Together] Borland Together ControlCenter. http://www.togethersoft.com/

[3] AstroGrid Phase-B Structure. http://wiki.astrogrid.org/bin/view/Astrogrid/PhaseBStructure

[4] NVO Prototypes. http://www.us-vo.org/prototypes.html

[5] Robert Hanisch. Building the Framework for the National Virtual Observatory, NSF Cooperative Agreement AST0122449, Quarterly Report, Oct-Dec 2002.

[6] Robert Hanisch. Building the Framework for the National Virtual Observatory, NSF Cooperative Agreement AST0122449, Quarterly Report, Jan-Mar 2003.

[7] [GGF8] Global Grid Forum 8. http://www.ggf.org/Meetings/ggf8/

第五章 注册与发现

资源的注册与发现是网格技术的精华，同样也是虚拟天文台的精华。在任何时间，一旦某个资源可为虚拟天文台所用就应该被及时地发现。应用需要发现机制来发现可用的服务和资源并确定这些服务和资源的性质和功能，以便对自身进行配置，组织对可用服务的请求操作。China-VO 应用和用户必须能发现和确定可用服务的属性、特点和调用要求并能够创建瞬时服务。

5.1 OGSA 服务发现机制的基本思想

在OGSA中服务发现机制是通过一系列标准化过程来实现的[1]，主要体现在：

- 标准的服务数据（Service Data）描述。服务数据是 Grid 服务实例的描述信息，在 OGSA 中由一系列被命名的、有类型的 XML 元素，即服务数据元素（Service Data Element）组成。
- 标准的操作（FindServiceData），来提取 Grid 服务实例的服务数据。
- 标准的接口，来注册（Registry）和解析（HandleMap）服务信息。

在 OGSA 体系中，网格服务的 Factory、Registry、GridService 和 HandleMap 接口实现瞬时服务实例的创建和发现。

同时，OGSA将定义完善的注册服务来实现服务的注册和发现。注册服务是一个服务实例，提供了用于注册操作的Registry接口和用于FindServiceData操作的GridService接口。具体到Globus Toolkit[2]，它主要利用网格资源分配和管理（Grid Resource Allocation and Management，GRAM） 服务实现资源的调度和管理。

在 OGSA 体系中所有的组件都以服务的形式体现。但为了表述的方便和更符合天文学家的使用习惯，将 VO 中的服务注册分为两大类：资源注册和服务注册。这里的资源包括星表、星图、光谱、数据集、时序数据、模拟数据、文档等各种观测、模拟数据，衍生数据和文档，以及存储资源、网络资源、观测设备等物理资源与设施。服务则指的是数据库访问、数据处理、计算服务、数据挖掘、可视化等 VO 应用服务。

OGSA 提供了服务注册和发现的基本机制。但真正要实现 VO 环境中资源与服务的注册和发现还需要天文学家进行许多方面的努力。这主要体现在三个

方面：

- 明确 VO 资源与服务注册的内容和功能需求；
- 制定 VO 资源与服务注册数据模型和相应的元数据标准；
- VO 资源与服务注册在 OGSA 体系中的实现。

目前在IT业界以及电子商务领域两项技术对VO资源与服务注册与发现有很好的借鉴价值。其一是W3C领导下的语义网技术；另一个是WEB服务架构下电子商务领域中已经开始应用的UDDI（统一描述、发现和集成，Universal Description, Discovery and Integration）技术[3]。

5.1.1 语义网

语义网主要是利用XML和资源描述基础架构（简称RDF）[4]技术，借助本体（ontology）[5]的作用，使网络中信息的含义得到良好的定义，让计算机和人都能对网页的内容有深入的了解。语义网最大的优点是可以让机器获得对网络空间储存的数据进行智能评估的能力，推动人类知识的发展。

语义网并不是一个单独的网络，而是对现有互联网的一种扩展。在语义网中，本体起到了非常重要的作用。本体对绝大多数天文学家来说肯定是一个模糊的概念，有必要介绍一下儿什么是本体。

ontologies 是某个科研领域的研究者为了实现信息共享而定义的一个公用的词汇表，它包括让机器可以理解的这个学科领域内基本概念的定义以及概念之间的相互关系。

一个本体由三部分组成：

- 类（classes）或者概念(concepts)
- 槽（ slots） 或者角色(roles）或者属性（properties）：类各方面的属性和特征
- 面（facets）或者角色约束(role restrictions) ：也就是与其他类的关系

可见，这里的“ontologies”并不是哲学上“本体论”的意思，而是一个规范地定义术语之间关系的文档。这个文档按照分类学原理和事先制订的推理法则和规范对学科术语进行管理。网页中的术语或者 XML 代码的含义通过网页内链接到“ontologies”的指针来定义。“ontologies”的使用能大大增强网络的功能，促进知识的共享。

5.1.2 UDDI

UDDI最早应用于电子商务，是由商业界和IT业界为了加速Web 服务的推广、加强Web 服务的互操作能力而推出的。UDDI系统中使用商业注册中心来存储全球商家的信息。商家通过服务注册发布其可用的服务，服务需求者通过

注册中心查找服务。这个注册中心被看成是全球商家的白页、绿页和黄页，它和语义网中的ontologies、SRB（存储资源中介）[6]中的元数据目录有着许多相似之处。

UDDI 注册表中的数据由“白页”、“黄页”、“绿页”构成，有四种主要数据结构：商业实体信息(businessEntity)、服务信息(businessService)、绑定信息(bindingTemplate)、技术规范信息(tModel)。其中“白页”包含了关于商业体名称、地址、电话号码等信息；“黄页”包含基于某些商业类型的商业体的列表，或者说是 UDDI 按照商业类型或者其所在行业的类型提供的入口；“绿页”用于显示每个商业体提供的服务，包括与之有关的或者使用这种服务的所有诸如参数、终点值(endpoint)等技术信息。

UDDI 本身就是一个 WEB 服务，它的调用接口包含查询 API 和发布 API。查询 API 用来快速的定位候选的商业实体、WEB 服务，及其调用规范和相关的信息细节。

UDDI 最新的 V3 规范改变了 V2 规范平行的体系结构，实现了层次式的结构，为全球的 UDDI 运营商的统一管理和服务提供了坚实的基础。

在 V2 规范中，所有的 UDDI Registry 都是同级的关系，他们之间形成一个环，而它们之间的数据都要通过一个安全通道进行复制，最终要使得所有 UDDI Registry 的数据完全相同，以达到在服务请求者发现服务的时候无论通过哪个 UDDI Registry 查询得到的数据都是相同的。

V3 规范融入了多注册中心的拓扑结构（multi-registry topology）、增强的安全特征、改进了的 WSDL 支持，以及订阅 API 和核心信息模型的先进性，使得在多 WEB 服务集结构的情况下，UDDI 可以提供给客户或使用者更复杂、更完善的描述和发现功能。

5.2 服务发现在天文学上的尝试

虽然资源与服务的注册与发现机制是随着网格技术的发展才日益受到人们重视的，但在此之前，天文学上已经在这些领域有所尝试和涉足。也许没有明确提出注册和发现的概念，但在某些功能上已经向这方面做出了努力。这其中代表性的工作来自于法国斯特拉斯堡天文数据中心（CDS）[7]。他们开发了统一内容描述（UCD）[8]和统一链接生成器（GLU）[9]。其中UCD正越来越受到IVOA各成员的重视，专门成立有讨论组研究UCD在IVO注册中的应用。

5.2.1 统一内容描述

UCD是由ESO-CDS数据挖掘项目组开发的，最早的介绍来自ADASS[10] 第

八次会议[11]。它是一种参数分类法，旨在实现交叉相关和数据挖掘操作中天文参数的自动转换。

UCD 对天文学上常用参数，特别是星表中列及其相关属性进行了归纳和分类。目前 UCD 拥有一个 4 层的树状结构，包括有大约 1500 个叶子节点。

UCD 是设计用来描述数据表中列的内容，从而实现天文表列数据的互操作。但 UCD 不是列名，你可能在同一表中找到数个相同的 UCD 条目。同时 UCD 不是准确的定义、物理维度表达和单位描述。UCD 条目和物理单位是不同的参数，但是又有很强的相关性。

目前，UCD 已经被用于对星表中数列内容的描述、星表和数据库检索、数据过滤、可视化等方面。UCD 最主要的应用是在 CDS 的 VizieR 星表服务系统中。此外，VOTable 标准、ConeSearch 服务、CDS 的 Aladin 服务、简单图像访问协议（SIAP）、IDHA Aladin 服务器输出都提供了对 UCD 的支持。

UCD 与本体

从目前的 UCD 来看，它与本体有一些相似之处。UCD 的树状结构代表了某种程度的 class。但是在 UCD 中没有明确定义类、属性和实例之间的差别，每一个都可以被一个单独的 UCD 描述。所以整个 UCD 体系说成是基于知识的更合适，而不是一个现成的本体。同时，UCD 没有定义分级结构间不同部分的相互关系，也就是说没有与本体中面（facets）对应的部分。

在某个学科领域创建一个本体是一个不断反复的过程，需要开发者之间、开发者与用户之间、用户之间对知识体系结构达成广泛的一致。如果我们想利用 UCD 作为天文本体的标准部件，那么我们需要在其内容表达和分类标准上达成一致意见。

本体非常适合于描述概念之间一般或者说通常的关系，但它在如何描述概念之间的数学关系方面还没有进行尝试，比如如何描述波长、频率和光速之间的关系。

UCD 与数据模型

UCD 是被设计用来描述数据表中列的内容的，从而实现天文表列数据的互操作。UCD 通常不包括全局性的元数据，比如关于天文台的说明。

将数据模型中的元素附加上 UCD 属性就可以利用这个词汇表使得数据模型中的元素有一个统一的描述。论文第九章，LAMOST 的数据模型设计就为模型中的每个元素定义了相应的 UCD 条目。美国约翰·霍布金斯大学的 A. Szalay 博士也曾试图为 SDSS 数据库结构赋予 UCD 条目，他还设计了少量新的

UCD 条目。

虽然 UCD 在 VizieR 系统中得到了成功的应用，但还不能完全满足 VO 数据模型的要求。为了将 UCD 用于某个特定的数据模型，比如 LAMOST 光谱数据模型，需要对 UCD 进行扩展：

- 定义数据模型中用到但现在 UCD 中没有的新的概念，比如赤纬位置误差；
- 创立相应的 UCD 并将其放在元数据结构中，比如 POS_EQ_RA_ERROR；

目前，UCD 的结构灵活度还不够，在 UCD 树中进行条目检索比较困难。在 UCD 应用到 VO 注册服务之前，我们必须对其进行必要的改造和扩展。比如：

- 重新整理根节点，形成一个天文界公认的分类标准；
- 增加新的 UCD 条目，删除无用的条目；
- 定义标准的前缀、后缀，使得条目更容易理解和检索；
- 打破现有的树状结构，进行条目原子化；
- 对 UCD 结构进行原子化调整；
- 将 UCD 参数化，给每个条目都加上参数或者属性，就像一个个函数一样，但这可能会使 UCD 变得异常复杂；

这些建议是不成熟和不系统的，需要在 VO 实践中不断的摸索。

5.2.2 统一链接生成器

WWW 通过提供单一的用户界面、单一的网络协议、URL 超文本连接机制，在推动远程数据库访问的统一化方面起到了非常重要的作用。然而，虽然 URL 的生成很容易，但是这些 URL 的维护是数据管理者们面临的一个挑战。我们在浏览网页时经常会遇到“Error 404，网页无法访问”这样的错误，说明无效的 URL 在网络上实在太多了。

为了增进 CDS 不同服务之间的互操作性，CDS 开发了 GLU（统一连接生成器）。GLU 允许数据库管理者在 WEB 页面中使用符号名而不是直接的 URL 代码。GLU 系统包括三个部分：

- GLU 词典（GLU dictionary）：将符号名与相应的 URL 和其他信息对应起来；
- GLU 过滤器（GLU Filter）：动态的将符号名替换为真实的 URL；
- GLU 监守程序（GLU Daemon）：利用 GLU 词典实现 GLU 标签到 URL 的替换。

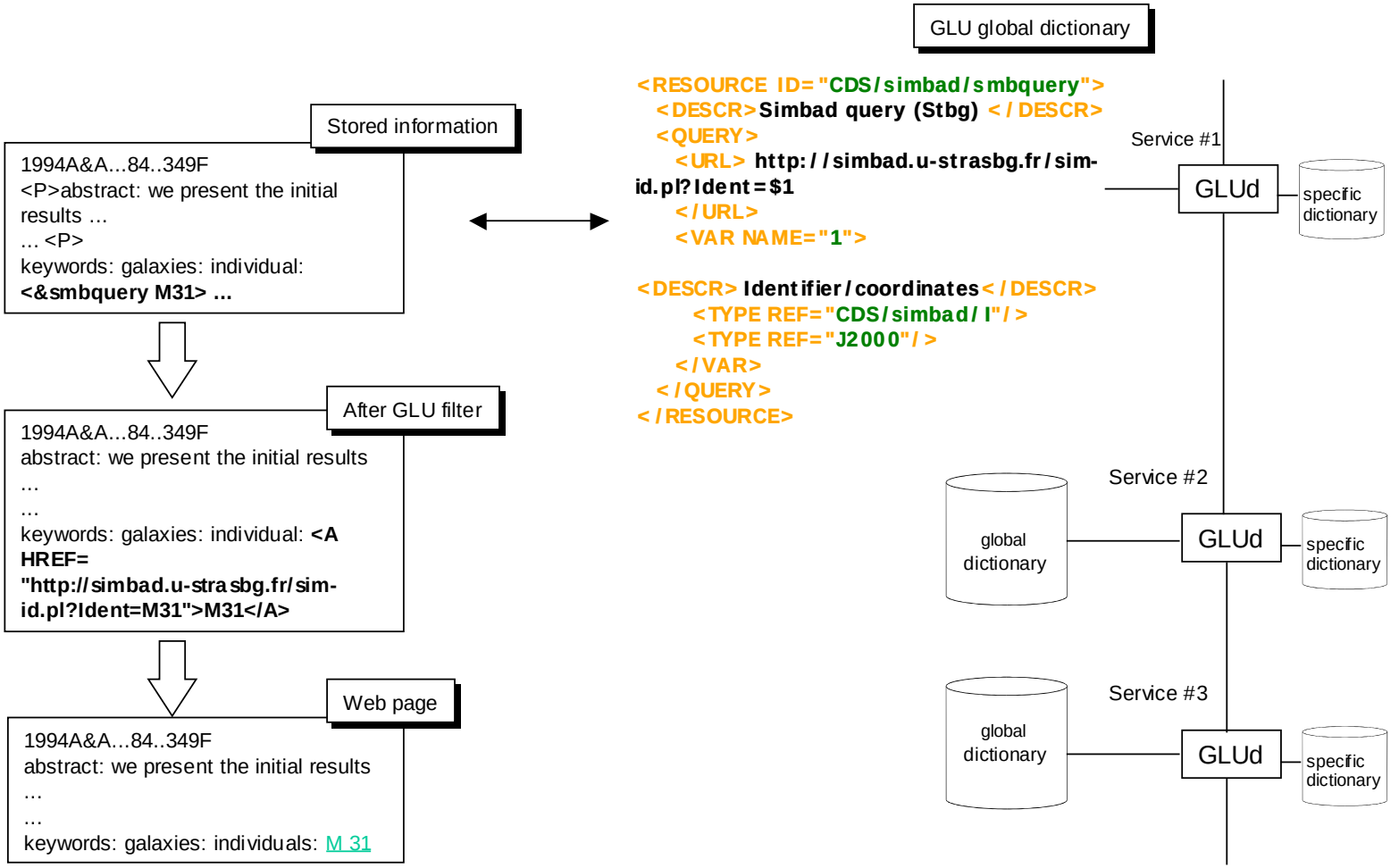


图 5.1 GLU 工作过程示意图

在一个 WEB 文档中使用符号名，即 GLU 标签，代替 URL。每次这个 WEB 文档被访问时，GLU 动态的将 GLU 标签替换为 URL。借助 GLU 词典对 URL 进行管理，降低了管理的复杂性，减小了无效“URL”发生的机会。GLU 的工作过程如图 5.1 所示。

如果在 IVO 中采用 GLU 方案，那么结构应该是：这个系统要由多个域组成，许多注册中心同时工作。每个数据中心负责本地注册的维护，但每个注册都作为整个系统的一个入口。有一个管理员管理不同域中的服务器。每一个服务器有一个本地词典，包含本地注册中的条目，以及其他注册的副本。当一个服务器的本地词典进行了更新，通过同步机制对其它服务器进行相应的更新。

除了天文学领域在资源注册和发现方面的努力，其他领域也进行了许多探索。比如NVO就对数字图书馆领域的开放式文档动议（OAI）[12]进行了评估，研究在VO注册领域使用的可能性[13]。

5.2.3 开放式文档动议

开放式文档动议（OAI）是数字图书馆领域开发和倡议的一项互操作标准，旨在完成电子内容的发布。动议的核心是元数据获取标准（PMH）。利用 PMH，一个数据仓库可以发布自己数据的元数据。

OAI PMH 协议是一个轻量级的服务提供者从数据提供者获取元数据记录的机制。它允许服务提供者向数据提供者索取各种资源的元数据。OAI 基于 HTTP 和 XML，简单快捷、部署方便。虽然 OAI 不是一个搜索协议，但可以与基于 Z39.50 、SRW 、SOAP 等协议上的搜索服务协同工作。OAI 不但可以提供元数据还可以通过 URL 等方式提供数据内容。虽然都柏林核心元数据格式是 OAI PMH 的默认记录格式，但 OAI PMH 与 XML 有很好的兼容。OAI 还提供了访问控制功能，可以对元数据和数据内容有不同的访问权限。此外，在 HTTP 协议的基础上，OAI 提供了访问控制、压缩和加密等机制。PMH 特别适用于利用自动代理工具将元数据收集到中心站点，并支持跨集合的搜索。

在PMH模型中，元数据以记录的形式存放，每条记录对应一个数据条目。一个数据条目可能是一本书或者是一个大的数据集。每一个记录有一个唯一的标识，这个标识符合IETF的URI标准(RFC2396, Berners-Lee 1998)[14]。

PMH 接口接受 URL 编码的 HTTP GET 查询，返回 XML 格式的结果。3 个核心操作分别是：

- ListIdentifiers：标识 ID 列表
- GetRecord：从 ID 得到记录
- ListRecords：记录列表

OAI 站点维护着一个资源的注册。这个注册周期性的检查注册记录的接

口，删除失败的接口记录。对 OAI PMH 这样的工具进行评估可以为 VO 带来一些好处，比如：

- 借用一个现成的、经过实际应用检验的框架，可以节省我们设计和调试的时间；
- 可以利用已经存在的软件和工具包，即软件复用；
- 可以将 VO 领域外的符合 OAI 标准的资源纳入到 VO 系统中；
- 可以让其他领域的人更了解天文数据。

PMH 是一个很灵活的模型，它在数字图书馆领域的活跃使用证明了其可行性。与 IVO 注册需求相比，它不能满足全部，但提供了许多可以借鉴的功能。

- 对于注册内容方面，PMH 可以很好的满足 VO 的需求。注册信息由注册人负责，可以对这些信息进行更新。
- PMH 不能提供用户如何查询注册方面的功能，PMH 协议不支持复杂的查询，不能满足资源和服务的发现的要求。
- VO 资源是一个层次式的，但 PMH 在这方面的支持不够。

5.3 IVOA 注册工作组

IVOA Registry[15]属于互操作性的范畴，最重要目标是让天文学家能定位、提取详细信息和使用IVO空间内任意地点的任何资源。

Registry 领域在 IVOA 有最高的优先权。注册标准旨在为如何满足天文学家在 IVO 或任何 VO 环境中对任意地点任意资源的发现和使用的需求提供解决方案。IVOA 力争能在 2003 年底推出 1.0 版的 VO 注册标准，以便提交 2004 年 1 月 IVOA 会议讨论通过。

VO 注册标准初步计划将包括如下规范：

- 注册元数据（ RegistryMetadata ）
- 资源元数据（ ResourceMetadata ）
- 注册查询（RegistryQuery ）
- 注册响应（RegistryResponse ）
- 资源标识（ResourceIdentifier ）
- 资源体系（ResourceHierarchy ）
- 注册定位（RegistryLocation ）
- 注册同步（RegistrySynchronisation ）
- 注册扩展（RegistryExtension ）

上述规范集仅是一个设想，很可能无法全部完成。但 IVOA 将力争完成其中最基本和最关键的部分。

为了协调各个 VO 计划在资源注册方面的努力，同时推动全球天文界认可的注册标准，IVOA 成立了注册工作组，并按照职责的不同细分为 5 个工作小组。工作组以邮件列表“registry@ivoa.net”为主要的信息交换途径。

IVOA 注册工作组划分的 5 个小组及其职责如表 5.1 所示。

5.3.1 注册标准制定

旨在实现互操作性的资源和服务的注册机制虽然有 UDDI 等作为实践先行者，但还是相当的不成熟，与 VO 的理想目标还有太大的距离。VO 注册标准的制定是一个复杂的过程，需要多方的协调和反复迭代过程。这里有一种可行的标准制定思路。首先挑选或制定一些科学范例，对这些范例定义相应的使用范例和测试范例。在这些范例的基础上获取 VO 注册的功能需求，进而制定适应需求的注册规范。

为了定义注册需求，IVOA 从项目成员的 Demo 中抽取了一些与 VO 注册相关的科学范例。主要包括：

AstroGrid 计划的：

- Brown Dwarf Selection
- Deep Field Surveys
- Galaxy Clustering
- Hi-Z Quasars
- Low Surface Brightness Galaxy Discovery
- Magnetic Storm Onset
- Solar Coronal Waves
- Solar Stellar Flare Comparison
- STP Solar Event Coincidence
- Supernova Galaxy Environment

NVO 计划的：

- Find Super Novae Pre-Burst Observations
- Select Dwarf Galaxies by Colour for Observational Follow-up
- Construct Highest Signal-to-Noise V-band image of sky region
- Find all object catalogs derived from a mission
- Find queriable image-cutout services with sky coverage and galactic coordinate criteria.
- Gamma Ray Burst Demo

- 褐矮星候选体搜寻
- 星系形态学分析

工作小组	工作范围	工作描述
Rwp01	管理、政策、文档标准	● 产生文档模板 ● 与 IVOA 执委会联络 ● 协调工作组计划和发布联合的文档 ● 制定相关策略
Rwp02	需求、科学范例、使用范例、测试范例	● 确定一系列关键的科学范例以用来定义注册的需求。这些科学范例要能例证发给注册的请求所涉及内容的范围。 ● 描述一系列使用范例，这些要能代表注册的典型使用情况。这其中要包括实现科学范例所需的使用情况。 ● 定义一系列具体的查询以用来对注册系统进行测试。 ● 制定一套功能需求，要能满足使用范例的需要并有助于注册的开发。
Rwp03	元数据规范	● 明确在注册中描述和定位一个资源时所需的相关信息 ● 提供一个基本的、可扩展的数据模型来在注册中描述资源 ● 开发一个 XML Schema 来利用资源数据模型中的词条描述资源
Rwp04	注册复制和接口	设计一个机制来实现注册信息以鲁棒、灵活的方式进行管理和分布，同时让注册之间实现对话和协调。4 个主要目标： ● 设计分布式机制 ● 设计注册查询方案（Schema） ● 定义实现以上两方面内容的接口 ● 定义注册内容管理接口
Rwp05	应用协调	● 协调 VO 项目间注册标准的实现 ● 生成软件库 ● 开发相关组件 ● 争取实现一个相对简单的参考原型

表 5.1 IVOA 注册工作组职责分工

AstroVirtel 项目的：

- 近邻星系中星团的光度函数研究

此外，我们 China-VO 计划提出的两个 Demo 也有助于用来定义 VO 注册需求：

- 银河系丰度梯度分析
- 利用多波段观测数据检验 SVM 算法在天体自动分类中的应用

对这些范例进行精选，IVOA 注册工作组初步确定了下面范例作为关键科学范例。

AstroGrid: Brown Dwarf Selection
NVO: Select Dwarf Galaxies by Colour for Observational Follow-up
功能需求：基于参数的星表搜索

AstroGrid: Deep Field Surveys
功能需求：基于天区或时间覆盖的数据搜索

NVO: Gamma Ray Burst
NVO: Find Super Novae Pre-Burst Observation
功能需求：特定天区的相关资源显示

AstroVirtel: 近邻星系中星团的光度函数研究
功能需求：注册功能的详细需求设计

NVO的Ray Plante在IVOA关于注册问题的讨论中对注册的需求给出了一个初步的设想[16]，他指出VO的注册：

注册类型分为：资源注册和服务注册两类

注册内容包括：

- 资源描述
- 服务描述（调用方法、响应数据格式、兼容性说明等）
- 元数据结构需要标准化

注册查询：

- 资源和服务可以进行基于特征的发现
- 查询结果是一种机器能理解的形式，比如可以考虑使用 VOTable 格式
- 能通过 ID 或者少量的元数据就可以定位唯一的资源

注册和发现过程：

- 注册过程应该尽可能简单
- 资源和服务的元数据更新尽可能简单或者能自动完成
- 取消注册的过程尽可能简单
- 注册应该允许注册的服务临时下线
- 要保证注册资源和服务的有效性，对永久无效的注册服务要有适当的管理措施
- 需要明确资源、服务间的依存关系

结合 IVOA 的讨论，China-VO 注册服务应该满足下面的需求：

- 提供足够的元数据信息，包括数据履历信息、数据内容信息和相应的访问服务信息
- 标识资源

- 可用资源列表
- 通过给定天区发现资源
- 通过给定时段发现资源
- 通过给定目标发现资源
- 通过给定波段发现资源
- 对表列数据的支持
- 对光谱数据的支持
- 对时变事件的支持
- 对理论模拟数据的支持
- 对太阳系领域观测资源的支持（比如太阳物理、行星物理）

注册中常用查询服务比如：

- 星表浏览服务
- 关于特定天体类型的服务（星系、恒星、星际介质）
- 关于特定数据类型的服务（星表、星图、光谱）
- 提供特定输出参数的服务（NGC 名、恒星光谱型、红移、视向速度）
- 受特定参量约束的服务（距离、星等、自行）

5.3.2 Astrogrid 注册的初步设计

AstroGrid目前给出的注册设计[17]中，每个服务包括三种类型的元数据：基本（basic）、履历（ curation）和元数据格式（metadata format）。

服务种类按照提供的功能不同分为：

- 数据存档（Data Archive）: 数据集和星表资源
- 数据存储（Data Storage）: VO 用户可以利用的存储设施
- 数据转换（Data Transformation）: 模型、图像处理、批处理程序、单个程序、等
- 注册（Registry）: 注册本身以及其他注册

他们还给出了一些基本的操作动作：

- 列出所有服务: 返回所有服务的基本元数据
- 查询服务: 返回满足查询条件的服务的基本元数据
- 列出服务元数据: 返回满足查询条件的服务的履历元数据
- 列出服务元数据格式: 返回满足查询条件的服务的基本元数据和元数据格式
- 列出注册元数据: 返回满足查询条件的注册的履历元数据

下一步功能发展目标：

- 服务查询
- 组织的元数据查询
- 数据集查找
- 数据集详细信息查找
- 地理位置查找

5.4 IVOA 建议的 VO 资源元数据

IVOA在借鉴Dublin Core Metadata[18]的基础上经过广泛的讨论给出了一个初步的VO资源元数据（简称RSM）的体系结构[19]。

5.4.1 体系结构

在这个体系结构的顶层，只需要少量的信息说明资源的存在和其出处。在下面的层次中，元数据越来越丰富和复杂，对数据内容、查询语法、访问协议、使用规则进行说明。这样的体系结构可以让天文信息服务能方便的加入到VO中。

在RSM的模型中：

- 资源（Resource）在 VO 中是一个非常宽泛的词汇，几乎所有东西都可称为资源。一个资源是一个或多个服务或其它资源的集合。资源有一些共同的元数据特征，比如发布者、创立者、标识、类型等。资源和服务都由元数据来描述。
- 组织（Organization）是一个特定类型的资源，由人员组成。
- 服务（Service）是 VO 中能被用户调用来执行某些操作的实体。
- 查询服务（Query Service）支持查询/响应操作。用户向服务提出查询，服务向用户反馈一些信息。
- 注册（Registry）是一个查询服务，其响应是一个关于其它服务的结构化的描述。

资源元数据构成了天文信息的“黄页”，这类似于 UDDI 服务，同时也类似于 CDS GLU 中的高层描述。它由三部分组成：标识元数据（Identity Metadata）、履历元数据（Curation Metadata）、内容元数据（Content Metadata）。

服务元数据对服务进行描述，包括接口元数据（Interface metadata）和功能元数据（Capability metadata）两部分。接口元数据描述了如何访问服务，输入输出格式。VO 服务需要提供一些标准的接口，比如基于 WEB 浏览器的接口（HTML 表格）、Grid/WEB Service 接口（WSDL 文档）、传统的 HTTP GET 接口等。功能元数据描述了服务能做什么、有哪些限制以及其他行为特征等。

5.4.2 RSM 主要内容

IVOA 给出的资源服务元数据标准（0.7 版）主要的内容如下：

资源元数据

标识元数据

标题（Title）：
数据类型：string
定义：资源的名称

标签（Ticker）
数据类型：string
定义： 资源名称的一个短小缩写

标识（Identifier）
数据类型：URI
定义： 在给定环境中对资源的一个明确指向

履历元数据

发布者（Publisher）
数据类型：string
定义：发布资源的实体，它要对资源的可用性负责

发布者标识（PublisherID）
数据类型：URI
定义：资源发布者的标识，可以是一个网络链接

创建者（Creator）
数据类型：string
定义： 资源主要内容的创建实体，可以是个人或者组织

主题（Subject）
数据类型：string, list
定义：对资源的主题、天体类型或其他属性进行描述的一系列关键词

描述（Description）
数据类型：string, free text
定义：对资源内容的一个概括

贡献者（Contributor）
数据类型：string
定义：对资源内容有贡献的实体

日期（Date）
数据类型：string
定义：在资源的生命期中一个重要的日期，比如创建或者发布日期

版本（Version）
数据类型：string

定义：与资源创建或者发布相关的一个标号

资源参考（ReferenceURL）
数据类型：URL
定义： 一个指向资源额外信息的 URL 地址

联系方式（Contact）
数据类型：string, e-mail address
定义：资源负责人，包括联系人的姓名和电子邮件地址

内容元数据

类型（Type）
数据类型：string, list
定义：资源内容的种类或类型，可以是 Archive、Bibliography、Catalog、Journal、Library、Simulation、Survey、Education、Outreach、EPOResource、Animation、Artwork、Background、BasicData、Historical、Photographic、Press 等其中的一个或几个。

范畴（Coverage）
数据类型：string
定义：资源内容的涵盖范畴，包括空间覆盖（Coverage.Spatial）、相关区域覆盖（Coverage.RegionOfRegard）、光谱覆盖（Coverage.Spectral）、时间覆盖（Coverage.Temporal）、深度覆盖（Coverage.Depth）、目标密度（Coverage.ObjectDensity）、目标计数（Coverage.ObjectCount）等子项。

内容等级（ContentLevel）
数据类型：string, list
定义：内容等级描述，指出资源的潜在用户，可以是公众（General）、小学教育（Elementary Education）、初级中学教育（Middle School Education）、高级中学教育（Secondary Education）、社区学院（Community College）、大学（University）、业余研究（Amateur）、信息教育（Informal Education）中的一项或几项。

机构（Facility）
数据类型：string
定义：获取数据的天文台和研究机构

设备（Instrument）
数据类型：string
定义：收集数据所用设备

格式（Format）
数据类型：string, list
定义：资源所提供信息的物理或者数字显示方式，比如文件格式（FITS，ASCII，VOTable，GIF，…）、存储方式（CDROM、DVD、在线、录像带、出版物，…）等

权限（Rights）
数据类型：string
定义：资源的所属及访问权限

服务元数据

接口元数据

服务接口 URL（ServiceInterfaceURL）
数据类型： URL
定义：一个指向服务接口说明文档的 URL

服务基准 URL（ServiceBaseURL）
数据类型： URL
定义：用户激活服务的 URL 的基本部分

HTTP 服务接口返回结果（ServiceHTTPResults）
数据类型： (MIME type)
定义：服务返回结果的 MIME 类型

功能元数据

服务标准 URI（ServiceStandardURI）
数据类型：URI
定义：标识一个服务标准的 URI

服务标准 URL（ServiceStandardURL）
数据类型： URL
定义：一个指向服务标准描述的 URL

服务最大搜索范围（ServiceMSR）
数据类型：float, decimal degrees
定义：服务提供者为服务附加的最大搜索范围，以度的形式规定最大的搜索半径

UIUC的Ray Plante博士基于上述的元数据模型给出了一个不同类型的VO资源描述Schema，如图 5.2 所示[20]。

5.5 VO 注册的几点考虑

5.5.1 集中注册还是分布式注册

对于整个 VO 系统，我们需要的是一个单一的集中式的注册还是一个分布式的相互连接起来的注册网络？如果是分布式拓扑结构，每个注册都需要包含到其他注册的连接。当需要搜索资源时，客户需要遍历这些连接。这些连接的拓扑结构如何，树状、网状还是环状？

虽然天文学相对于电子商务或者物理、化学等其他学科是一个相对简单的领域，但是我们的数据和服务还是有相当的复杂性，一个集中式的注册机制可能是不实际的，最好采用分布式的注册网络。每个对外服务系统或者说每个 VO 项目拥有自己的本地注册系统来描述自身资源和服务，各自负责本地信息

的管理和更新。互联网域名服务系统（DNS）的拓扑结构和同步、缓冲机制都是很值得借鉴的。

5.5.2 注册元数据粒度问题

注册元数据的粒度，也就是注册内容的详细程度分级，是VO注册需要解决的问题[21]。以天文学星表注册为例，注册内容可以有不同的内容等级。比如在注册中：

- 仅存储服务器的名称和连接
- 存储所有注册服务器中所有星表的名称和连接
- 存储所有注册服务器中所有星表的名称、连接和标准元数据
- 存储所有注册服务器中所有星表的名称、连接和标准元数据，此外还保存每个星表中列和参数的名称
- 存储所有注册服务器中所有星表的名称、连接和标准元数据，此外还保存每个星表中列和参数的全部信息（名称、数据类型、单位、UCD，等）

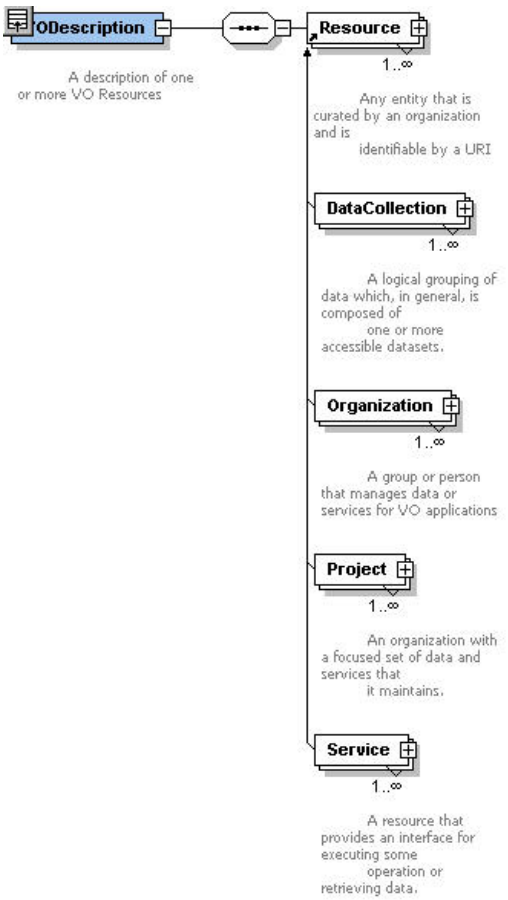


图 5.2 VO 注册元数据模型

不同的作业对元数据的要求是不同的。比如，“浏览全部北天光学巡天星表资源”这样的操作，对星表的天区覆盖和波段覆盖提出了要求，并不需要知道星表的具体表列结构信息；“找出 LAMOST 巡天星表中与 SDSS 巡天的对应体”，这样的操作就不但需要星表的天区覆盖信息还必须提供星表各列的详细信息。

为了能成功的表达对某个星表的查询，客户程序可能需要得到关于星表的所有上面列出的信息。然而这些信息不必要都直接从注册中心取得，客户可以直接从星表服务器上得到。这就涉及到各级元数据在 VO 系统的哪部分提供的问题。如果将尽可能详细的元数据都进行注册，对像上面后一种查询这样的请求比较合适，但对前段第一种请求就会造成性能的下降和资源的浪费。如何平衡这两者的利弊？

5.5.3 ID 与元数据

资源和服务 ID 是注册内容的一部分，但它有其特殊性。虽然 ID 和注册是相互依存的，对用户来说可能直接使用 ID 更方便。在需要的时候通过 ID 取得关于资源的更多的细节。

这里有个一致性问题，是不是通过一个 ID 应该可以唯一确定一个资源？不同格式但相同内容的数据，比如同一天体 VOTable 格式和 FIT 格式的光谱，LAMOST 巡天星表在不同地点的镜像，是否应该拥有同一个 ID？

ID应具有的特点[22]：

- 适应性广，VO 应用所涉及到的资源、项目、数据集、服务等都可以用
- 和相对 URL 和绝对 URL 的概念相似，可以定义局部 ID 和全局 ID
- 通过 ID 取得资源的唯一描述
- 有效性和持久性
- 给数据和服务提供者提供最大的方便

5.5.4 其他观点和问题

- 数据集本身应该自带一些 VO 服务，至少是数据访问服务；
- VO 中的注册将提供 VO 资源的元数据，简化资源搜索过程；
- IVOA 的注册将以标准的方式组织起来，不应该需要集中授权；
- 每一个注册都要有一个命名空间和自己唯一的 ID；
- 每一个资源都应有一个唯一 ID；
- 资源的拥有者可以方便的对自己的注册进行更新和修改；
- 资源之间的依赖关系（比如'is derived from', 'was copied from', 'was extracted from', 'provides access to', 等）需要注明；
- IVOA 互操作的协议数量要最少化；

- 超过 IVOA 最精粒度的元数据要直接从所有者处取得；
- 注册需要同时提供 Grid 服务和 GUI 界面以方便注册和更新；
- 注册也需要生命期管理；
- 虽然 OGSA 中的所有组件都是服务，但 VO 注册是否需要分类（服务、文档、数据）。

5.6 China-VO 的资源注册与发现

资源的注册与发现是网格技术的精华，同样也是虚拟天文台的精华。不过，这方面的研究还处于刚刚起步阶段，无论是网格技术领域、电子商务领域还是 VO 领域，都没有给出一套切实可行的方案。GIIS、SRB、UDDI 这些应用工具都只能提供少量简单的功能，与“资源的动态发现”这个目标还相差甚远。

VO 注册元数据标准的制定需要 IVOA 各成员紧密合作，在注册内容、注册格式、查询接口、语义表达等多方面达成共识。China-VO 将与 IVOA 的合作伙伴一起制定被国际同行认可的资源注册、服务注册、服务发现方面的一整套标准和协议，最终实现 VO 系统中资源与服务的动态发现和高效利用。

参考文献

[1] Ian Foster, Carl Kesselman, Jeffrey M. Nick, et al. The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration. http://www.gridforum.org/ogsi-wg/drafts/ogsa_draft2.9_2002-06-22.pdf

[2] [GT] Globus Toolkit. http://www.globus.org/toolkit/

[3] [UDDI] Universal Description, Discovery and Integration. http://www.uddi.org/

[4] [RDF] Resource Description Framework. http://www.w3.org/RDF

[5] Natalya F. Noy, Deborah L. McGuinness. Ontology Development 101: A Guide to Creating Your First Ontology. http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html

[6] [SRB] SDSC Storage Resource Broker. http://www.npaci.edu/DICE/SRB/

[7] [CDS] Centre de Données astronomiques de Strasbourg. http://cdsweb.u-strasbg.fr/

[8] [UCD] Unified Content Descriptors. http://cdsweb.u-strasbg.fr/UCD/

[9] [GLU] Générateur de Liens Uniformes. http://simbad.u-strasbg.fr/glu/glu.htx

[10] [ADASS] Astronomical Data Analysis Software and Systems. http://www.adass.org

[11] Patrcio F. Ortiz, François Ochsenbein. ESO/CDS Data-mining Tool

Development Project. In: D.M. Mehringer, R.L. Plante, & D.A. Roberts, eds. ASP Conf. Ser., Vol. 172. ADASS VIII, 1998: 379

[12] [OAI] The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html

[13] Ray Plante. An Evaluation of the Open Archives Initiative for VO Registries. http://bill.cacr.caltech.edu/cfdocs/usvo-pubs/files/evaloai.html

[14] [RFC2396] T. Berners-Lee. Uniform Resource Identifiers (URI): Generic Syntax. http://www.ietf.org/rfc/rfc2396.txt

[15] [Registry] IVOA Resource Registry. http://www.ivoa.net/twiki/bin/view/IVOA/IvoaResReg

[16] Ray Plante. Registry Requirements. http://www.ivoa.net/forum/registry/att-0009/01-registryreq.txt

[17] AstroGrid Registry. http://wiki.astrogrid.org/bin/view/Astrogrid/AgCd06Registry

[18] [DublinCore] Dublin Core Metadata Initiative Metadata Terms. http://dublincore.org/documents/dcmi-terms/

[19] [RSM] Robert Hanisch. Resource and Service Metadata for the Virtual Observatory. http://www.ivoa.net/internal/IVOA/IvoaResReg/ResourceServiceMetadataV7.pdf

[20] Ray Plante. VOResource. http://rai.ncsa.uiuc.edu/~rplante/VO/schemas/VOResource.xsd

[21] Clive Davenhall. Preliminary Thoughts on the Contents of a VO Registry. http://wiki.astrogrid.org/bin/view/Astrogrid/RegistryContents

[22] VO Resource Identifiers. http://www.ivoa.net/forum/registry/

第六章 数据存储、访问与互操作

虚拟天文台需要面对“数据雪崩”所带来的各种挑战。如何对 TB、PB 量级的数据进行存储？如何实现对全球各地海量异构数据资源的统一高效访问？海量数据的存储、访问与互操作是 China-VO 应用服务实现的基础，是当前阶段工作的重心。本章将就以上问题进行探讨。

6.1 互操作实现的基本思想

网格技术最重要的特点和目标就是消除“信息孤岛”。同样，VO 的基本功能和目标也是要实现全球范围主要天文研究资源，特别是数据资源的统一访问。这其中最关键和最重要的就是不同数据集、不同数据服务间互操作性的实现。

不同数据中心的数据存储形式各式各样，提供的访问服务功能和形式也很不一致。为了能实现异构数据和服务的统一访问，一条重要的途径就是抽象化。通过定义能满足所有 VO 数据提供者和用户需求的数据模型以及相应的资源、服务注册机制，来屏蔽数据源在数据格式、存储格式、主机环境、访问形式等诸多方面的异构性、复杂性，实现对这些分布的、异构的数据源的统一、透明访问。

在 VO 系统中，为了实现这个目的，需要在许多方面进行努力，主要涉及的工作包括：VO 数据模型的建立、数据集与服务注册机制的实现、VO 查询语言定义与实现、VO 标准数据交换格式的制定和实现。

互操作性是 IVOA 的一个主要课题。目前在 IVOA 内最活跃的讨论也集中在互操作性方面。IVOA 成立有数据访问层（DAL）工作组，旨在定义和规范 VO 远程数据访问标准。在 VO 的架构内，客户端数据分析软件利用这些服务访问数据；数据提供者在向 VO 发布数据时要实现这些服务。

DAL 工作组要制定 DAL 标准。这些标准将指导数据中心和巡天项目如何开发与 VO 兼容的接口，评估实施和维护过程中的资源分配。

数据访问层功能的实现在OGSA中体现为一系列的Grid服务。这些服务存在于数据源和上层应用之间，为上层应用提供统一的数据访问服务。其原理就像沙漏，如图 6.1 所示[1]。DAL底部是各种异构的数据资源，上面是各种数据应用。但数据与应用的接口则是统一的，当然具体的接口形式还须研究和探讨。

6.2 数据模型

数据模型（DM）描述了不同数据集之间有什么不同和相同之处，但不考虑具体的数据格式，是天文数据的信息内容模型而不是格式的模型。与保存数据的特定的数据格式和交换数据的元数据关键词不同，数据模型是对数据抽象内容的描述，是对数据结构的描述。数据模型可以通过一套元数据或者 Schema 的形式体现。数据提供者和用户通过元数据来理解数据的内容。

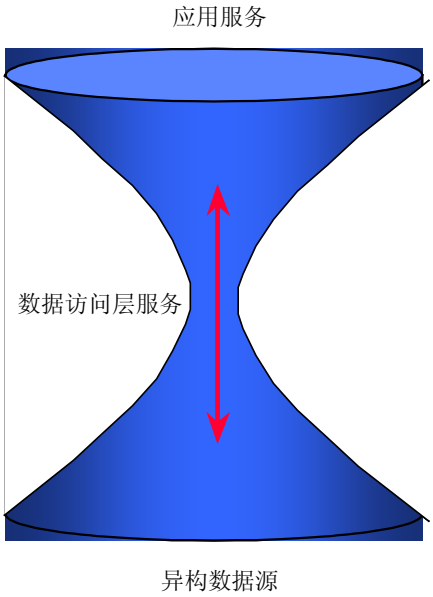


图 6.1 数据访问层沙漏模型

数据模型和元数据的界限需要明确。VO 的元数据力图提供一致数据资源语义描述，而数据模型旨在定义数据结构和接口以及新结构的加入机制。元数据标准描述的是通过数据模型定义的数据内容如何进行格式化、标准化。

数据格式、元数据和数据模型三者之间的关系可表述如下[2]：

- 格式：数据描述的语法
- 元数据：How，描述数据的语义
- 数据模型：What，描述数据的内容

虽然数据模型不是元数据，但它与元数据关系密切。VO 的设计者们将利用数据模型设计元数据。

从数据集的角度来看，DM 有多方面的用途。可以用它来对数据的逻辑结构和内容进行标准的描述。数据提供者可以用 DM 来向 VO 描述自己的数据。

一个统一的数据模型对于 VO 互操作性的实现起着非常关键的作用。这个模型将为 VO 中的各种数据资源给出一个统一的逻辑视图，从而屏蔽数据格式、数据内容的异构性。数据访问服务通过标准的数据访问接口实现对数据的

访问。

数据资源的体系层次划分是数据模型建立的基础，同时也涉及到元数据标准和资源与服务注册标准的制定。J. McDowell 等人为 VO 的数据资源给出了一个如图 6.2 所示的可能的分级方案。在这个方案中图与表是明确的分开的。但是在许多情况下，图像可以用来存储表列数据，表也可以内嵌图像数据。比如一维光谱，既可以用 FITS 存储也可以用简单的数据表存储。在他们给出的另一个数据体系结构中将图和表整合在了一起，如图 6.3 所示。

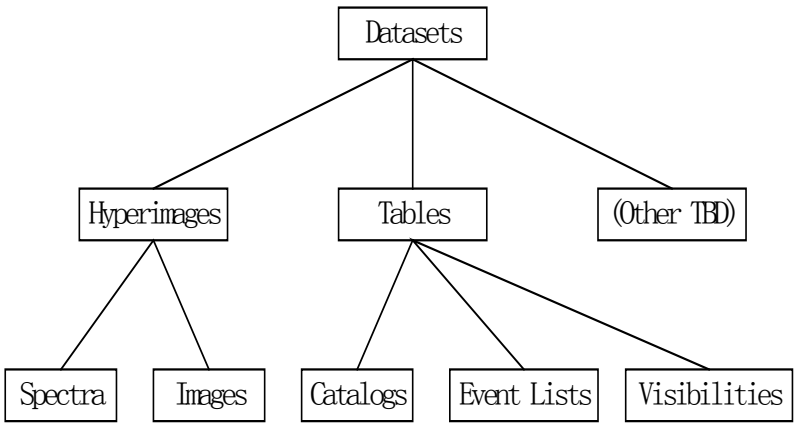


图 6.2 一种可能的数据体系划分

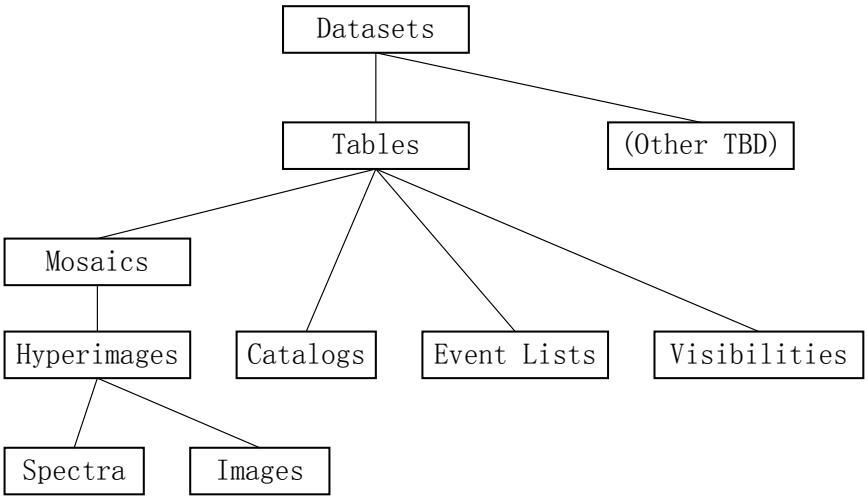


图 6.3 另一种可能的数据体系划分

IVOA 需要对 VO 数据资源的体系结构达成共识，从而在此基础上为每个组成部分构建数据模型。在论文的第九章，本人采用 VOTable 的格式给出了一种 LAMOST 数据产品的数据模型方案。IVOA 成员也在积极的制定各种 VO 资源的数据模型。

J. McDowell和S. Lowe给出了一个光谱数据模型方案[3]。这个数据模型主要

包括以下几个部分：

- 顶层属性（SPECTRUM）：对光谱的总体描述。
- 坐标轴对象（SPECTRUM _AXIS）：光谱各坐标轴参量描述。
- 物理量对象（SPECTRUM _VALUE）：光谱数据中物理量的定义，比如波长、流量单位。
- 履历元数据（SPECTRUM_CURATION）：观测者、发布者、软硬件环境等的说明。
- 观测对象（SPECTRUM _OBS）：与观测相关的数据，比如天区、观测波长、起止时间。
- 质量对象（SPECTRUM _QUALITY）：光谱质量的总体说明，比如是否经过定标、是否存在饱和问题等。
- 光谱提取对象（SPECTRUM _EXTRACT）：光谱提取过程相关描述，比如狭缝大小、光谱级次。

F. Valdes定义了一个称为“天球归并光子观测（Celestial Sphere Binned Photon Observations，4DBIN）”的天文数据超类[4]，把各类观测数据看作是一定时间内特定天区位置光子能量的集合。这个类只有 4 个参量：位置（2个）、能量和时间。VO数据与观测数据的主要区别是对能量进行了标准化处理，或者说是“Calibrated”。J. McDowell和S. Lowe模型中的图像和光谱都只是 4DBIN.CALIBRATED类的投影或者说子类。

英国的Starlink开发了HDX数据模型[5]，用一个Java数据访问库来实现，并实现了与VOTable的联合操作。Starlink在灵活性和可用性之间进行了充分的权衡，使得HDX成为一个灵活的可扩展的天文数据模型，适用于图像、表列和其他元数据。这个模型是基于XML的，可以适用于FITS和其他一些数据存贮格式。对VO数据模型的建立有很好的借鉴价值。

IDHA也制定了一套天文图像数据的数据模型[6]。这个数据模型主要对图像数据集查询过程中涉及的元数据进行了描述。从一个对数据（图像、光谱）分析和理解感兴趣的天文学家的视角出发，定义了用于定量分析相关的元数据、观测过程相关的定性信息以及数据质量信息等。其“data”包中子类的情况如图 6.4。

6.3 数据格式

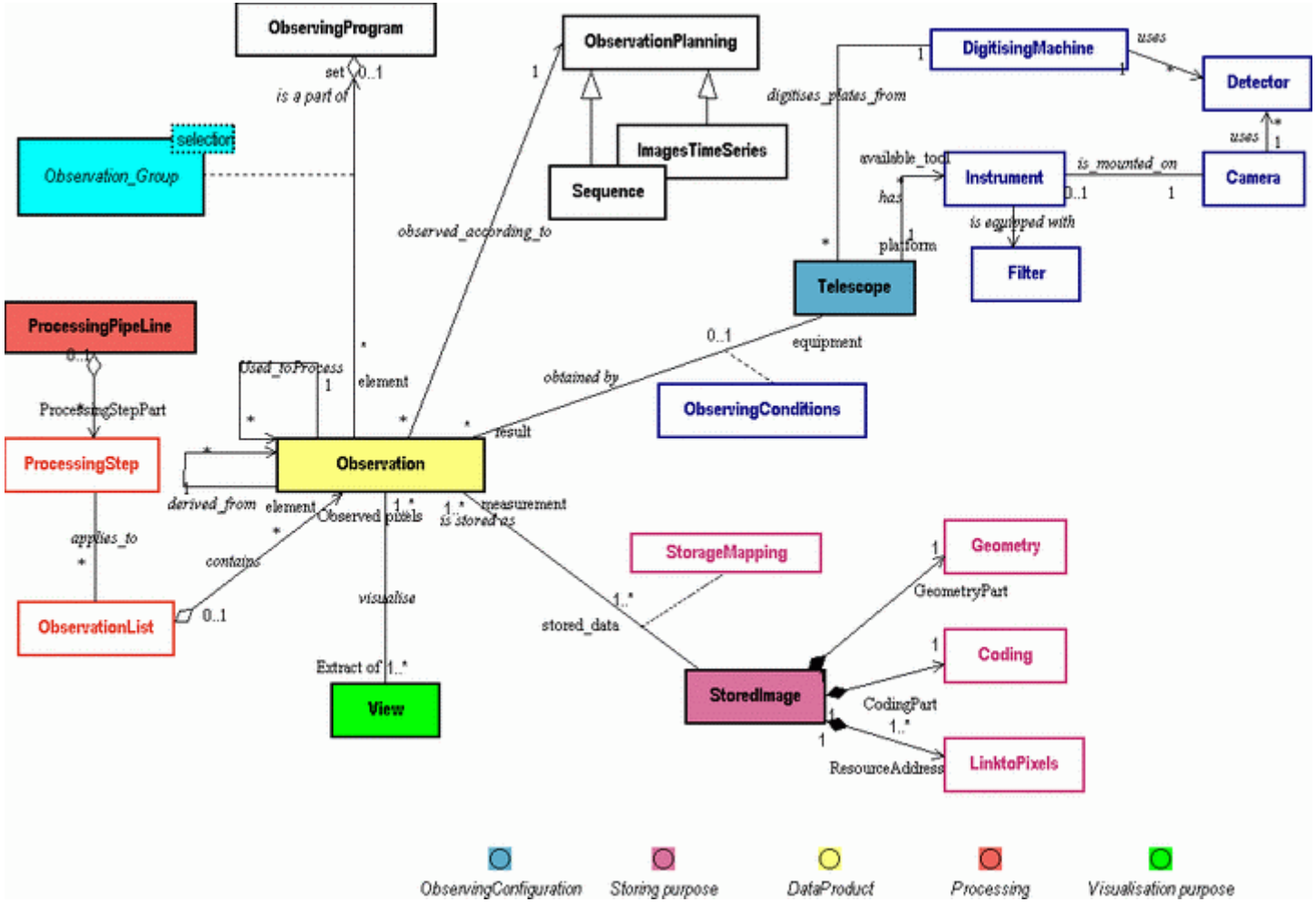


图 6.4 IDHA 数据模型

在 VO 系统中，数据格式按用途不同可分为数据的存储格式和数据的交换格式，其中数据的交换格式与互操作性的实现关系甚为密切。

数据的存储格式广义上可以分为文本格式和二进制格式。文本格式又可以分为普通的无结构文本和标记语言文本。我们常用的HTML、XML、LaTex[7]都属于标记语言文本。二进制形式具体的类型非常多，在天文学中常用的有FITS格式，JPEG、GIF、TIFF等图像格式，PS、PDF、DOC等文档格式，以及二进制的数据库文件等。

上面提到的这些存储格式在 VO 系统中都将用到。VO 是以现有技术为基础，一个逐步进化的系统。现在流行的数据、文档的存储格式和使用习惯仍将在 VO 中得到支持。

数据的交换格式是指数据在用户间、程序间进行信息交换时所采用的编码格式。由于数据交换过程是一个双方或多方的过程，交换格式的选用影响到交换过程各方。标准的、灵活的、适应性广的交换格式将减少各方为交换过程所付出的代价，提高效率，增强互操作能力。

从目前情况来看，两种数据格式将作为VO的主要数据交换格式。它们就是已在天文学上使用多年的FITS[8]和 2002 年刚刚推出 1.0 标准的VOTable[9]。

6.3.1 FITS

FITS 是“灵活图像传输系统，Flexible Image Transport System”的首字母缩写，是目前天文学上标准的数据格式。1982 年，国际天文联合会 (IAU) 确定 FITS 为世界各天文台之间用于数据传输、交换的统一标准格式。FITS 标准的维护当前主要由 NASA GSFC 的 FITS 支持办公室（http://fits.gsfc.nasa.gov/）负责。

FITS 描述了数据定义和数据编码的一般方法，被用于多维矩阵数据（比如一维光谱、二维图像、三维数据立方体）、表列数据的传输、分析和存档。它是与机器无关的，提供了图像的单值转换，精度包括符号在内可以达到 32 位。对一维、二维、三维、甚至多维的数据类型都提供了合适的转换。

FITS 格式多年的使用证明它对天文学家是非常有用的。现在天文学家可以在不同天文台之间或不同的图像处理系统进行数据交换，只要说明使用的是 FITS 格式就可。目前，不论是像兴隆 216 望远镜这样的传统观测，还是像 SDSS、2dF 这样的现代巡天观测都使用 FITS 格式来存储交换数据。

天文学上大量的数据都是以FITS文件形式存储的。但直到今天都没有一个互联网标准可以让WEB服务器通知WEB客户端如何处理FITS文件，而只是简单的下载到本地。Lick天文台和NRAO天文台的天文学家正努力，将FITS注册为

MIME的一种类型[10]。如果他们的努力成功，那么FITS文件类型将得到Apache、Microsoft IIS等这样HTTP服务器的支持，用现在流行的Netscape、Mozilla、IE等网络浏览器就可以显示甚至处理，将大大推动FITS格式的普及。

6.3.2 VOTable

VOTable标准是为了实现VO环境中在线数据的互操作和可升级性要求而开发制定的。VOTable使用XML文档格式，是一个灵活的数据存储和交换标准。VOTable格式来源于Astrores[11]格式，而它又基于FITS的表格式，设计时尽量靠近FITS的二级制表格式。

VOTable通过使用XML标准，增强了互操作性。物理量不但标明单位还通过UCD[12]来对物理量的属性进行描述。元数据分为对表自身的描述和对域（列）的描述。VOTable实现了元数据和数据的分离存储，适合于海量数据和网格计算。

VOTable 的数据部分可以用三种格式表示：“TABLEDATA”、“FITS”和“BINARY”。TABLEDATA 是纯粹的 XML 格式，适合小型表的表示和处理。FITS 二进制表是天文学家所熟悉的，VOTable 可以对这样的表或相关元数据进行封装。BINARY 格式适用于程序处理，不需要 FITS 库，同时支持数据流。因为 FITS 的数据块大小有特殊的规定，FITS 数据流式化比较困难。

VOTable 的数据模型如表 6.1 所示。

VOTable = hierarchy of Metadata + Tables
Metadata = Parameters + Infos + Descriptions + Links + Fields
Table = list of Fields + Data
Data = stream of Rows
Row = list of Cells
Cell = Primitive or variable-length list of Primitives or multidimensional array of Primitives
Primitive = integer, character, float, floatComplex, 等

表 6.1 VOTable 数据模型

VOTable 支持的原子数据类型如表 6.2 所示。

文档结构

VOTable 文档的根元素称为“VOTAble”，它可能包含一个“DESCRIPTION”元素、一个“DEFINITIONS”元素、若干个“INFO”元素，包含至少一个“RESOURCE”元素。“RESOURCE”元素可以嵌套子“RESOURCE”元素，也可以包含表（Table）和参数（PARAM）。VOTable

的文档结构如图 6.5 所示。

数据类型	含义	FITS对应类型	长度（Bytes ）
"boolean"	Logical	"L"	1
"bit"	Bit	"X"	*
"unsignedByte"	Byte (0 to 255)	"B"	1
"short"	Short Integer	"I"	2
"int"	Integer	"J"	4
"long"	Long integer	"K"	8
"char"	ASCII Character	"A"	1
"unicodeChar"	Unicode Character		2
"float"	Floating point	"E"	4
"double"	Double	"D"	8
"floatComplex"	Float Complex	"C"	8
"doubleComplex"	Double Complex	"M"	16

表 6.2 VOTable 支持的原子数据类型

在 1.0 标准中 VOTable 能处理的最复杂的数据结构是多维矩阵。在未来标准中，将努力实现 VOTable 与 FITS 的更好兼容，让 VOTable 能处理更大规模的数据流。

6.3.3 FITS 与 VOTable 的功能差别

FITS 的“子串矩阵”是一个复杂的语义规范，可以用一个单一字符串表示一个子串的集合。目前，VOTable 不支持这个功能。VOTable 支持定长和变长的字符串以及由定长字符串构成的变长矩阵。

VOTable 支持数据与元数据的分离，支持表数据流以及适应现代分布式计算的一些功能。它在两种结构化数据表示方法（XML、FITS）之间架起了一座桥梁，利用 UCD 来描述参数和列的属性。它继承了 XML 的体系结构特点和灵活性，支持 Unicode 编码。

一个 FITS 文件可以无损转换为一个 VOTable 文件，但 VOTable 文件向 FITS 文件的转换可能就意味着某些信息的丢失。

6.4 数据存储

虚拟天文台是数据密集型在线天文研究平台，丰富的数据资源和强大的数

据处理能力是其内在的要求。China-VO 对数据存储的要求主要包含以下几个方面：

- TB 甚至 PB 量级的存储能力
- 良好的可扩展性
- 良好的 I/O 性能
- 良好的安全管理
- 良好的数据管理

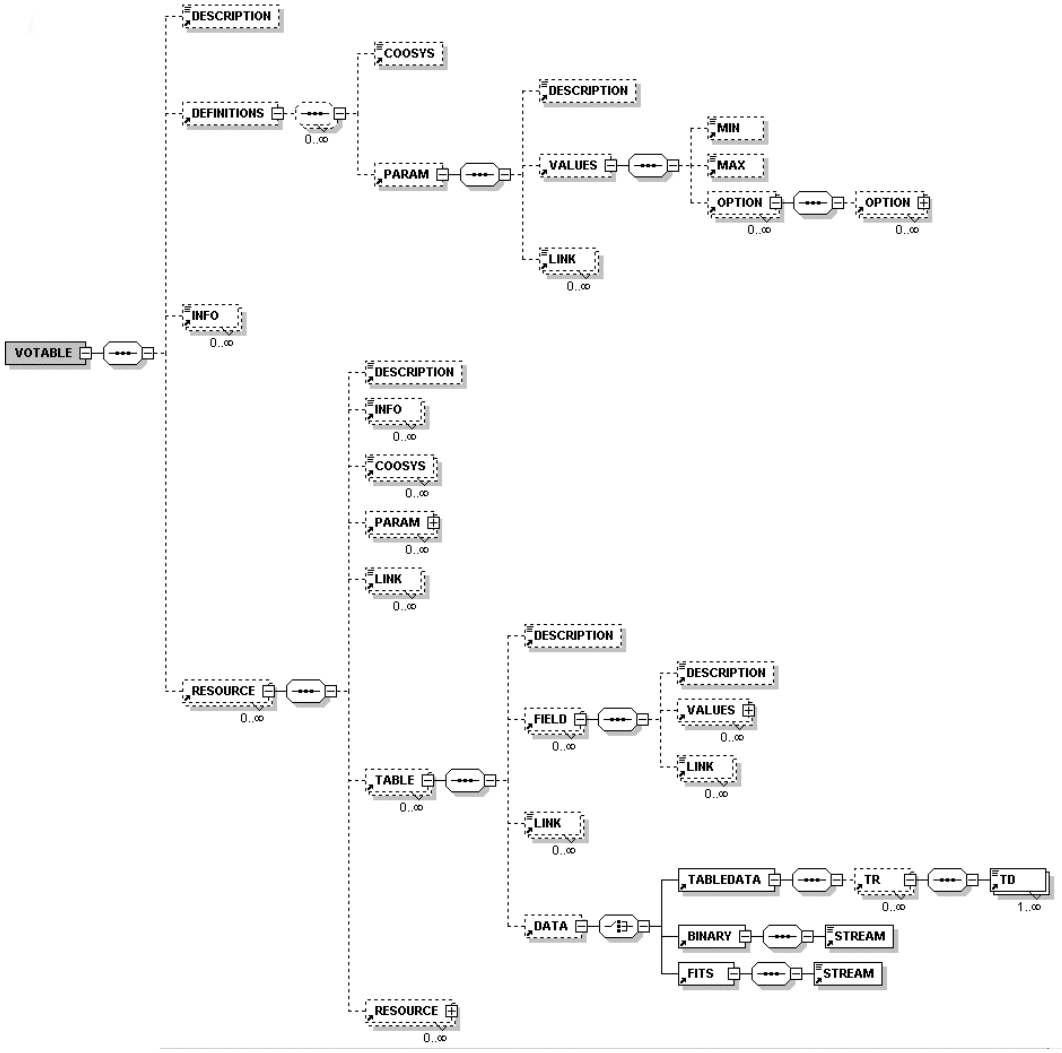


图 6.5 VOTable 文档结构

九十年代以前，存储产品大多作为服务器的组成部分之一，这种形式的存储被称为SAS（Server Attached Storage，服务器附属存储）或DAS（Direct Attached Storage，直接附属存储）。随着技术发展，进入九十年代以后，人们逐渐意识到数据集中和共享成为一个亟待解决的问题。于是，网络化存储的概念被提出并得到了迅速发展[13]。从架构上来分，今天的网络化存储系统主要包

括SAN（Storage Area Network，存储区域网）、NAS（Network Attached Storage，网络附加存储）和iSCSI三大类。

直接存储不利用网络协议，直接利用本地磁盘空间资源向应用提供存储服务。网络存储则将存储器从应用服务器中分离出来，利用网络协议组成一个存储网络。使用存储网络的好处在于：

- 统一性，在逻辑上是完全一体的；
- 实现数据集中管理；
- 容易扩充，即收缩性很强；
- 具有容错功能，整个网络无单点故障。

6.4.1 DAS

直连的磁盘阵列是 20 世纪 80~90 年代比较流行的计算机存储设备，DAS 被定义为直接连接在各种服务器或客户端扩展接口下的数据存储设备系统。它完全以服务器为中心，寄生在相应服务器或客户端上，其本身是硬件的堆叠，不带有任何存储操作系统。

现在 BADC 的 4TB 数据服务器使用的是 DAS 技术。我们选用的是通用的 LINUX 操作系统，没有针对磁盘阵列进行存储方面的优化或者说专门化。4 个独立的文件系统服务器通过 NFS 机制挂在一台门户服务器上对外服务。文件服务器间使用的是普通 5 类双绞线连接。

6.4.2 NAS

网络附加存储设备（NAS）是一种专业的网络文件存储及文件备份设备，或称为网络直联存储设备、网络磁盘阵列。NAS的典型组成是使用TCP/IP协议的以太网文件服务器，每个NAS拥有单独的IP地址，可以直接挂接在主干网的交换机或其他局域网的Hub上。一个NAS里面包括核心处理器，文件服务管理工具，一个或者多个的硬盘驱动器用于数据的存储。 NAS 可以应用在任何的网络环境当中。主服务器和客户端可以非常方便地在NAS上存取任意格式的文件，包括SMB格式（Windows）、NFS格式(Unix, Linux)和CIFS格式等等。NAS网络存储的拓扑结构如图 6.6 所示[14]。

为了明晰NAS与DAS的主要差异，表 6.3[15]将其二者做了一个简单的比较。

6.4.3 SAN

SAN（Storage Area Network，存储区域网）可以定义为是以数据存储为中心，采用可伸缩的网络拓扑结构，通过具有高传输速率的光通道的直接连接方式，提供 SAN 内部任意节点之间的多路可选择的数据交换，并且将数据存储管

理集中在相对独立的区域网内的一种存储网络。SAN 是一种通过 SCSI、SSA、ESCON 和光纤通道等外围通道协议连接存储设备的后端网，是一个高速的专用子网。通常 SAN 由 RAID 阵列连接光纤通道组成，SAN 和服务器和客户机的数据通信通过 SCSI 命令而非 TCP/IP，数据访问是“块级”（block level）。SAN 拓扑结构如图 6.7 所示。 采用 SAN 可以实现局域网中的任何服务器、任何磁盘阵列子系统、任何磁带系统之间的互连，实现数据的共享和集中的管理。

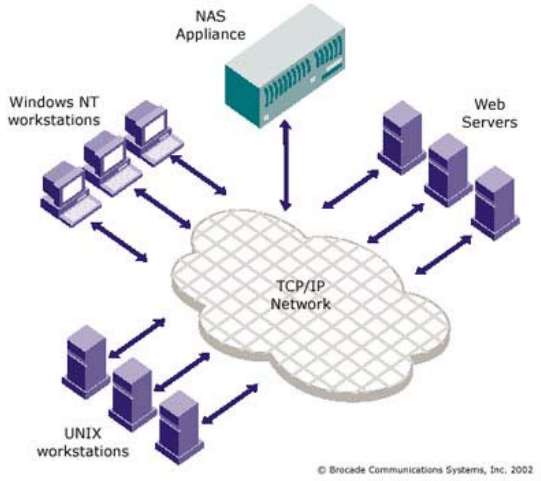


图 6.6 NAS 网络存储拓扑结构

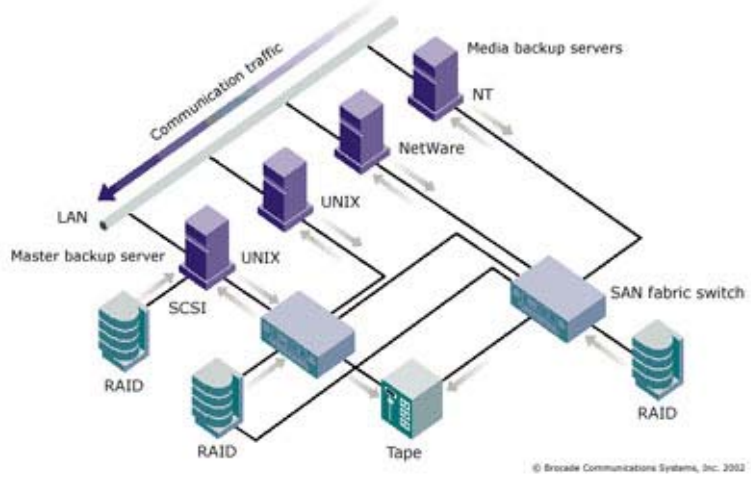


图 6.7 SAN 网络存储拓扑结构

6.4.4 NAS 和 SAN

NAS 是基于 IP 的文件级访问存储技术，SAN 是基于通道的块级访问存储技术。NAS 和 SAN 的区别如表 6.4 所示。

NAS 是在 RAID 的基础上增加了存储操作系统，通常是一个服务器群：应用服务器、邮件服务器等等，存储设备附加在这个系统上。 由于 NAS 设备在网络中占用一个 IP 地址，本身就相当于一台高性能的文件服务器，操作系统、文件管理系统与物理设备专用化的软硬一体化的整体设计。

比较项目	NAS	DAS
安装	安装简便快捷，即插即用。只需要10 分钟便可顺利独立安装成功。	系统软件安装较为烦琐，初始化RAID 及调试第三方软件一般需要两天时间。
异构网络环境下文件共享	完全跨平台文件共享，支持Windows、NT、UNIX（Linux）等操作系统。	不能提供跨平台文件共享功能，各系统平台下文件需分别存储。
操作系统	独立的优化存储操作系统，完全不受服务器干预，有效释放带宽，可提高网络整体性能。	无独立的存储操作系统，需相应服务器或客户端支持，容易造成网络瘫痪。
存储数据结构	集中式数据存储模式，将不同系统平台下文件存储在一台 NAS 设备中，方便网络管理员集中管理大量的数据，降低维护成本。	分散式数据存储模式。网络管理员需要耗费大量时间奔波于不同服务器下分别管理各自的数据，维护成本增加。
数据管理	管理简单，基于 Web 的 GUI 管理界面使 NAS 设备的管理一目了然。	管理较复杂。需要第三方软件支持。由于各系统平台文件系统不同，增容时需对各自系统分别增加数据存储设备及管理软件。
软件功能	自带支持多种协议的管理软件，功能多样，支持日志文件系统，并一般集成本地备份软件。	没有自身管理软件，需要针对现有系统情况另行购买。
扩充性	在线增加设备，无需停顿网络，而且与已建立起的网络完全融合，充分保护用户原有投资。良好的扩充性完全满足 24X7 不间断服务。	增加硬盘后重新做 RAID 须宕机，会影响网络服务。
总拥有成本（TCO）	单台设备的价格高，但选择 NAS 后，以后的投入会很少，降低用户的后续成本，从而使总拥有成本降低。	前期单台设备的价格较便宜，但后续成本会增加，总拥有成本升高。
数据备份与灾难恢复	集成本地备份软件，可实现无服务器备份。日志文件系统和检查点设计，以求全面保护数据，恢复数据准确及时。双引擎设计理念，即使服务器发生故障，用户仍可进行数据存取。	异地备份，备份过程麻烦。依靠双服务器和相关软件实现双机容错功能，但两服务器同时发生故障，用户就不能进行数据存储。

表 6.3 NAS 与 DAS 的主要差异

SAN 是独立出一个数据存储网络，网络内部的数据传输率很快，但操作系统仍停留在服务器端，用户不是在直接访问 SAN 的网络，因此这就造成 SAN 在异构环境下不能实现文件共享。

SAN 是只能独享的数据存储池，NAS 是共享与独享兼顾的数据存储池。

NAS 与 SAN 的关系也可以表述为：NAS 是 Network-attached，而 SAN 是 Channel-attached。

		网络类型	
		IP	通道
访问方式	文件	NAS	MFS（Multi-path File Serving）
	数据块	Block Storage Over IP	SAN

表 6.4 网络存储分类

NAS 是网络技术在存储领域的延伸和发展，数据以文件的形式按照网络协议在客户机与存储设备之间流动，它可以利用 NFS 实现异构平台的客户机对数据的共享，集成在存储设备内的专用文件服务器提高了文件传输的 I/O 速度。

但是，当数据存储发展到一定规模，NAS 的缺陷就显现出来，如数据服务和数据管理形成了网络的双重负担；磁盘阵列必须配置专用文件服务器，后期扩容成本高；一般文件服务器没有高可用配置，有单点故障；通过网络协议的访问方式，对存储系统的数据安全构成威胁等。

由于 SAN 采用了网络结构，服务器可以访问存储网络上的任何一个存储设备，因此用户可以自由增加磁盘阵列、带库和服务器等设备，使得整个系统的存储空间和处理能力得以按客户需求不断扩大。

在数据量不大，网络资源充裕的中、小型系统中，NAS 具有良好的性能价格比。因此，在 China-VO 的初期，我们可以考虑使用 NAS，而将 SAN 作为后期的扩充。

6.4.5 iSCSI

2003 年 2 月 11 日，IETF（Internet Engineering Task Force，互联网工程任务组）通过了由 IBM 和 Cisco 共同发起的 iSCSI（Internet SCSI）标准。

iSCSI 是一种在 Internet 协议网络上，特别是以太网上进行数据块（block）传输的标准；是一个供硬件设备使用的可以在 IP 协议上层运行的 SCSI 指令集。它可以实现在 IP 网络上运行 SCSI 协议，使其能够在以太网上进行路由选择。 利用 iSCSI，用户能够在诸如高速千兆以太网上进行路由选择，从而通过现有的 TCP/IP 网络来构建存储局域网（SAN）。通过相对比较熟悉的 IP 网络技术，用户得以更容易地管理 SAN 存储。这就是 IP 存储，即通过 IP 协议进行数据交换的存储技术。IP 存储让用户得到了比光纤通道存储局域网解决方案低得多的整体拥有成本。iSCSI 利用现有的 TCP/IP 基础设施，解决了 SCSI 的访问和距离问题。iSCSI 主机适配器把服务器的 SCSI 命令和数据转换成网络

包后，通过 IP 网络进行传送。

SAN 是独立出来的一个数据存储网络，网络内部的数据传输率很快。最关键的是，现在构建 SAN 结构的存储环境投资较大，对于天文科研领域是一个昂贵的投资。 iSCSI 正逐渐成为用于 SAN 的光纤通道以外的另一种选择，我们可以利用现有的以太网线缆部署 SAN。

6.4.6 RAID

虽然单个磁盘的容量不断增加，目前 320GB 的磁盘已经出现，但与 TB 甚至 PB 量级的存储需求来比还相差甚远。磁盘跨越技术可将多个磁盘作为一个逻辑磁盘来管理，从而实现大容量存储。这样的技术主要包括 LVM（逻辑卷管理）和 RAID。其中 RAID 适用范围广，适用性好，在 VO 的存储系统中必将大量的使用。

RAID，为“Redundant Arrays of Independent Disks”的头字母缩写[16]，中文为独立冗余磁盘阵列，通常简称“磁盘阵列”。如何增加磁盘的访问速度，如何实现超大容量的存储空间，如何防止数据因磁盘故障而丢失，一直是IT专业人员和用户的困扰。磁盘阵列技术的产生一举解决了这些问题。

磁盘阵列的两个基本技术：

- 磁盘跨越（Disk Spanning）： 磁盘阵列控制器将多个磁盘视为单一的磁盘，把小容量的磁盘整合为大容量的单一磁盘。用户不必规划数据在各磁盘的分布，提高了磁盘空间的使用率；使磁盘容量可作几乎无限的扩展；多个磁盘共同完成读写动作，提高了 I/O 性能。
- 磁盘分割（Disk Striping）：磁盘阵列将同一阵列中的多个磁盘视为单一的虚拟磁盘，数据以分块的方式顺序存放在磁盘阵列中。数据按需要分块，从第一个磁盘开始放，放到最后一个磁盘再回到第一个磁盘，直到数据分布完毕。数据以分块方式存放在不同的磁盘上，整个阵列的多个磁盘可同时读写，所以这大大提高了整个阵列的 I／O 性能。

磁盘阵列分为软阵列 （Software Raid）和硬阵列 （Hardware Raid） 两种。软阵列即通过软件程序并由计算机的 CPU 提供运行能力。 由于软件程序不是一个完整系统故只能提供最基本的 RAID 容错功能，其他如热备份硬盘的设置，远程管理等功能均不能实现。硬阵列由独立操作的硬件提供整个磁盘阵列的控制和计算功能，不依靠系统的 CPU 资源。由于硬阵列是一个完整的系统，所有需要的功能均可以做进去，所以硬阵列所提供的功能和性能均比软阵列好，缺点是投资较大。

作为高性能的存储系统，RAID得到了越来越广泛的应用。RAID的级别从RAID概念的提出到现在，已经发展了六个基本级别，分别是 0、1、2、3、4、5 等。最常用的是 0、1、3、5 这四个级别[17]。

RAID 0：将多个较小的磁盘合并成一个大的磁盘，不具有冗余，并行I/O，速度最快。它将多个磁盘并列起来，成为一个大磁盘。在存放数据时，将数据按磁盘的个数来进行分段，然后同时将这些数据写进这些磁盘中。在所有的级别中，RAID 0 的写速度是最快的。但是 RAID 0 没有冗余功能，如果一个磁盘（物理）损坏，则所有的数据都无法使用。

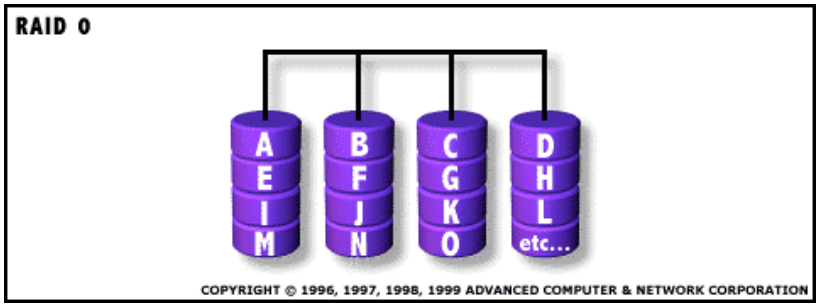


图 6.8 RAID 0 存储示意图

RAID 1：两组相同的磁盘系统互作镜像，速度没有提高，但是允许单个磁盘出错，可靠性最高。其原理为在主磁盘上存放数据的同时也在镜像磁盘上写一样的数据。当主磁盘（物理）损坏时，镜像磁盘则代替主磁盘的工作。因为有镜像磁盘做数据备份，所以 RAID 1 的数据安全性在所有的 RAID 级别上来说是最好的。但是其磁盘的利用率却只有 50%，是所有 RAID 级别中空间利用率最低的。

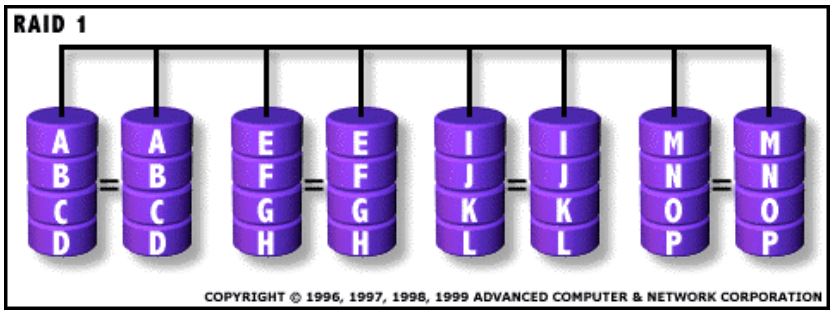


图 6.9 RAID 1 存储示意图

RAID 3 存放数据的原理和 RAID0、RAID1 不同。RAID 3 是用一个硬盘来存放数据的奇偶校验位，数据则分段存储于其余硬盘中。它象 RAID 0 一样以并行的方式来存放数据，但速度没有 RAID 0 快。如果数据盘（物理）损坏，只要将坏硬盘换掉，RAID 控制系统则会根据校验盘的数据校验位在新盘中重建坏盘上的数据。不过，如果校验盘（物理）损坏的话，则全部数据都无法使用。利用单独的校验盘来保护数据虽然没有镜像的安全性高，但是硬盘利用率

得到了很大的提高，为单硬盘容量的（n-1）倍。

RAID 5：向阵列中的磁盘写数据，奇偶校验数据存放在阵列中的各个盘上，允许单个磁盘出错。RAID 5 也是以数据的校验位来保证数据的安全，但它不是以单独硬盘来存放数据的校验位，而是将数据段的校验位交互存放于各个硬盘上。这样，任何一个硬盘损坏，都可以根据其它硬盘上的校验位来重建损坏的数据。硬盘的利用率为(n-1) 倍的单硬盘容量。RAID 是使用最多的阵列级别。

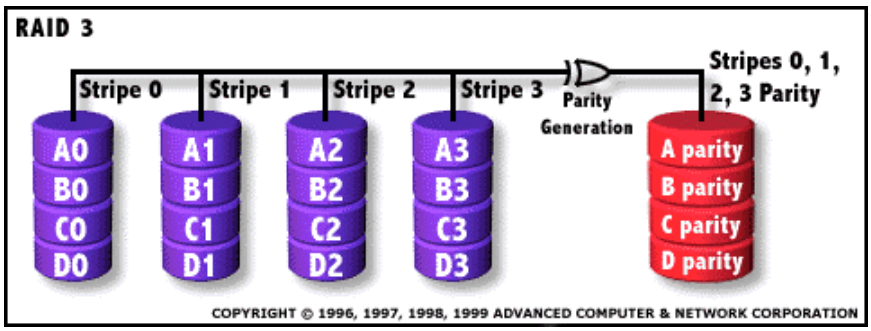


图 6.10　　RAID 3 存储示意图

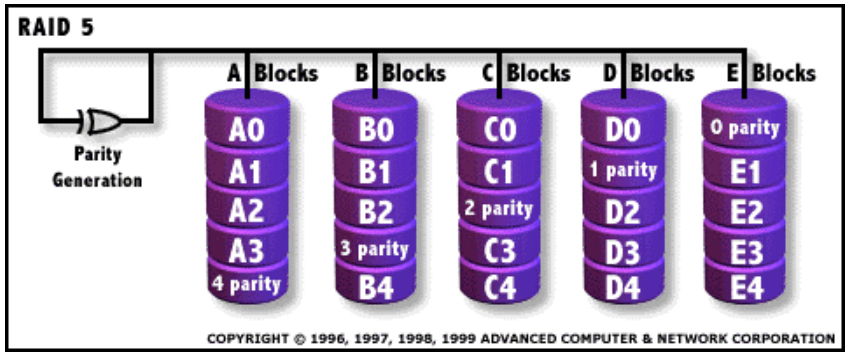


图 6.11　　RAID 5 存储示意图

此外，为了满足一些特殊的需求，还出现了复合式的 RAID 级别，比如 RAID 1+0、RAID 0+1、RAID 5+3 等。综合利用不同级别的 RAID 阵列可以实现如下的需求：

- 增加存取速度
- 容错(fault tolerance)，即安全性
- 有效的利用磁盘空间
- 尽量的平衡 CPU，内存及磁盘的性能差异，提高电脑的整体工作性能

在 VO 系统中，对读取性能的要求是第一位的，对数据安全和可靠性要求相对较低。所以，RAID 5 比较合适。同时，在一些关键的数据应用部分，也将考虑使用 RAID 1 或者 RAID 1＋0，提高系统的可靠性。

6.4.7　China-VO 的存储方案

上面介绍的这几种存储技术各有利弊。总体而言，SAN 性能最好，但投资最高；NAS 适用性广，性能逊于 SAN；DAS 价格最低但性能差；iSCSI 标准刚刚出炉，尚有待检验和完善。

China-VO 的数据存储将采用逐步升级的方案，根据当时数据量以及访问需求采用相应的存储方案。采用自行搭建 DAS 或者 NAS 和购买商业的数据服务器相结合的途径。目前，我们的 4TB 数据存储采用的是 DAS 的方式，“操作系统＋RAID 磁盘阵列”。随着数据量和访问量的增加，这种传统的方式将很难满足 VO 高性能数据访问的需求。在项目的初期，China-VO 国际数据和 LAMOST 巡天数据不会超过 20TB，使用数个 NAS 服务器将可满足要求。

存储技术正在快速发展，NAS、SAN、iSCSI 之间的共性会越来越多，互操作性会越来越好。同时，网格存储、虚拟存储技术会不断成熟，China-VO 将会在需求、投资、主流技术这几个方面平衡利益做出选择。

6.5 数据库访问

6.5.1 天文查询的特点

在VO环境中，数据库查询可以分为两种类型：位置查询和非位置查询[18]。位置查询是指通过天体名称或者天区位置进行查询，比如：

- Are there any X-ray sources near HD123456?
- I would like an infra-red image centred on PSR1234-456

这种查询方式是目前查询操作的最主要形式 。位置查询以外的其他查询方式都属于非位置查询，比如：

- 不同星表的交叉证认
- 统计分析
- 回归分析以发现相关性
- 稀有天体的发现
- 聚类算法
- 时序分析，寻找周期性、爆发等事件
- 相似性或者差异分析，寻找移动或者亮度变化天体
- 大尺度结构测量，比如傅里叶分析

非位置查询一般涉及的计算量和数据量比较大。上面列出的这些操作，其中一部分很适合当前的 DBMS 实现，但另一部分在现在的数据库系统中就很难实现，需要研究新的算法和数据结构。

6.5.2 VO 查询语言

现代 DBMS 通常都是使用“查询语言”对存储的数据表进行操作。这些“查询语言”基本上都以 SQL 为基础。通常，这种查询语言是与数据库中的数据表进行操作的唯一机制，在 DBMS 中起着核心作用。

现在 DBMS 所使用的标准查询语言有许多功能上的缺陷，为在天文学的应用带来困难。标准的 SQL 连求幂运算都没有明确的规范，不同的 DBMS 有各自的实现形式，如表 6.5 所示。

	SQUARE(x)	POWER(x,2)	POW(x,2)	x^2
DB2		√		
MySQL		√	√	
Oracle		√		
Postgres			√	√
Microsoft SQL Server	√	√		

表 6.5 不同数据库系统中的平方运算

SQL 也没有提供对六十进制支持问题。在天文学上大部分的位置数据使用度分秒表示的，但在所有的 DBMS 中都没有提供对度分秒直接运算的支持。同时，SQL 提供的查询功能也过于简单，比如无法实现天文学上最常用的锥形检索。

开发VO Query Language（VOQL）[19]的目的就是屏蔽不同数据库对SQL实现程度的不一致以及语法规则的不同，为不同类型的数据库访问提供统一的接口，同时针对SQL先天性的缺陷，开发提供能满足天文学查询需求和使用习惯的功能。

VOQL 的基本要求是满足天文学家对表列数据处理的需求。在此基础上还尽量提供图像操作、数据处理等方面的功能。

1996 年CDS推出的Astronomical Server URL（ASU）[20]在HTTP协议上提供了一种标准的天文查询方式。ASU提供了锥形检索在内的一些天文常用查询功能，但它还不能处理像交叉证认、数据挖掘和可视化这样的高级功能。此外，CDS还开发了GLU[21]，它保存了每一个服务的特性，可以将一个查询请求转换为适当的参数名和物理单位。当然，GLU只是暂时的解决方法，缺乏标准化。

VOQL 查询操作的基本过程：

1. 用户构造查询条件；
2. 用户将查询请求提交工作流引擎；
3. 工作流引擎生成 VOQL 语句；
4. VOQL 解析器在注册服务和元数据服务的协助下对 VOQL 语句进行解

析；

5. 生成作业计划，根据涉及的数据源和查询工作的性质可能会分解为多个子 VOQL 查询；
6. 在元数据服务的协助下，VOQL 解析器将子 VOQL 查询翻译为特定的数据库语言；
7. 执行查询作业；
8. 结果返回。

VOQL 对标准 SQL 的功能扩充将主要表现在：

- 输入输出格式过滤器，提供 XML、VOTable、FITS 格式的输入输出；提供 PS、PDF、FITS 、GIF、JPEG、TIFF 等图像格式的输出；提供 gzip、bzip2 等压缩格式输入输出。
- 天区形状的定义及相关操作，比如矩形（BOX）、圆形（CIRCLE）、椭圆（OVAL）、三角形（TRIANGLE）、多边形（POLY）等；对区域的操作包括在区域内部、在区域外部、区域的并与非操作等。
- 高级子集查询，比如锥形检索（Cone），给定位置和搜索半径的查询；代数表达式查询（ARB）；间隔查询（FREQ），每隔 n 行抽出一行；子集查询（LITBIG），提取最大或最小的 n 行；多边形范围查询（POLY）等。
- 交叉相关操作，提供基于位置的交叉证认和非位置的其他交叉相关操作。
- 基本的绘图和统计分析功能，提供散点图、曲线图、柱状图等基本图形的输出。
- 一维光谱显示支持，能以图像的形式显示一维光谱数据。

目前，IVOA 内的 VOQL 的开发工作是由 JVO 领导的，他们已经开发出了一个原型系统 JVOQL，在前面的 JVO 介绍中已经介绍过了。

6.6 DBMS 选取

6.6.1 DBMS 在 VO 中的应用

DBMS 在 VO 中的应用主要体现在数据存储和数据挖掘两个方面：

数据存储，这是 DBMS 的基本用途。VO 对 DBMS 在存储方面的要求主要有以下几个方面：

- 元数据处理能力
- 大数据量支持，比如大于 2GB 文件的支持

- 复杂数据结构支持
- XML 支持，这样才能满足网格服务、WEB 服务或者说 SOAP 调用的需求。
- 64 位操作系统支持，随着数据量增加、计算精度要求提高，64 位计算将显得越来越重要。
- 分布式数据库

数据挖掘，这对不同的用户有不同的含义，但从底层来说是对现有数据集的一系列操作，这些操作大部分是通过网络分布式完成的。这需要 DBMS 提供：

- 远程访问支持
- 并行处理能力
- 大数据集顺序扫描
- 空间索引支持
- 统计分析功能
- 高可靠性和高可用性
- 支持事务处理和回滚操作

目前，比较流行的数据库系统主要有 Oracle、DB2、Sybase、Microsoft SQL Server、MySQL、PostgreSQL。其中前四种是商业产品，后两种是自由软件产品。

6.6.2 AstroGrid 的 DBMS 测试

天文DBMS应用的基本功能是存储大型表列数据集，其次是对大型图像、光谱数据的存储。2002 年 8 月AstroGrid对PostgreSQL 、 MySQL、Oracle、Microsoft SQL Server这四个DBMS进行了测试[22]。AstroGrid测试的主要目的是：对不同的DBMS进行测试以确定它们是否适合天文星表和数据集的存储要求以及VO访问的需求。

测试标准

由于受到客观环境的限制，AstroGrid 的测试大部分是性能方面的测试，少部分是功能方面的测试。他们的测试内容主要包括下面这些：

功能特性：

- 查询语言（应该通过准代数表达式支持任意的投影、选择和联合操作）
- 数据表的输入输出（是否支持 HTML、XML、FITS 等格式）
- 数据类型（基本数据类型；特殊数据类型，比如图像、流媒体等）

- 空值处理
- 适应性（行数、列数的限制；表大小和数目限制；64 位操作系统）
- 索引功能（基本索引功能、2 维索引，能否支持 HTM 或者 Healpix）
- 元数据（元数据存储机制）
- 接口和 API（JDBC、ODBC 支持；API 功能）

互操作性：

- 远程访问（通过 WEB 或 Grid 服务进行远程调用）
- 分布式数据库
- 并行机制

DBMS 管理：

- 安全管理（访问授权与控制）
- 系统安装（是否容易）
- 系统管理（对数据库管理员的需求程度）
- 系统资源（安装大小、数据库文件的压缩比、磁盘空间需求）
- 到主机文件系统的映射（数据表、数据库、索引等向主机文件系统的映射；备份与移植功能）

市场位置：

- 产品提供者状态
- 价格
- 运行平台

测试结论

他们测试的结果如下：

- 对于数据存储，大部分的 DBMS 都可以胜任；
- MySQL 总体上比 Postgres 要快，同时使用简便；
- 在数据挖掘方面，所有系统都有很多的限制；
- 没有面向过程的语法，SQL 语言是面向结果的语言；
- 都支持多行命令，但调试困难；
- 标准兼容性差，虽然有国际标准 SQL92，但没有一个 DBMS 能真正全部实现；
- 除了列名及其数据类型不支持其他的元数据；
- 浮点数的显示很不规范；
- 所有测试的 DBMS 都支持空值（Null），但在文本文件的输入和输出中没有统一的表示标准；

- 都不支持二进制输入输出，VO 很需要能直接对 FITS 进行 I/O 操作的 DBMS；
- 都支持用户自定义函数，但标准不统一；
- 资源使用，对于大部分系统很难估计，只有 MySQL 是个例外，其资源占用情况比较低；
- 网络访问，两种免费的 DBMS 都没有网络访问数据表信息的能力，也不能真正支持分布式数据库；
- 都没有提供图形输出功能。

在参加测试的 4 款产品中：

SQL Server 具有非常好的易用性，大部分基本的数据库操作都可以通过 GUI 完成。SQL Server 在 SDSS SkyServer 的成功应用说明其对天文星表是适用的，同时 HTM 天区像素化方案也可在 SQL Server 上得到很好的支持。当然，SQL Server 一个很显著的缺点就是只能在 Microsoft 的 Windows 平台上运行。这对习惯使用 UNIX/Linux 的天文学家是一个重要缺陷。

Oracle 是一个庞大的、功能强大的、适应性广的 DBMS，在商业 DBMS 领域处于领先地位。Oracle 被许多天文数据中心所采用。这个 DBMS 已经支持 XML 和 WEB 服务。

总体上，MySQL 比 Postgres 更好用，也更快一些。MySQL 是使用最广泛的免费数据库系统。但是 Postgres 提供了比 MySQL 更强的功能，比如对 R 树空间索引的支持，是功能最强大的免费数据库系统。

AstroGrid 的数据库测试是从 VO 使用的角度出发的，与 IT 产品评测部门给出的评测清单以及各个数据库提供商给出的功能与性能是有着许多不同的。同时，我们也要清楚，AstroGrid 的测试并不是权威的测试。这个测试是由非 DBMS 专业人员在软硬件条件有限、时间有限、测试工具有限的前提下完成的，可以供我们参考。

6.6.3 CERN 的数据库测试

在数据库功能方面，2000 年下半年CERN完成了一个对Oracle、MySQL、PostgreSQL三个DBMS系统的测试[23]。测试结果可供我们参考。

他们从以下几个方面对这三个 DBMS 进行测试：

- 基本功能（基本数据类型、SQL 语言特征、完整性约束、编程能力、ID 的自动生成、国际语言支持等）
- 事务处理与多用户访问
- 数据库编程

- 数据库管理（访问控制、备份机制、数据移植）
- 灵活性与可扩展性
- 超大数据库性能（查询优化、作业分析、磁盘空间分配、数据大小限制等）
- 分布式数据库（多数据库访问、异构系统支持）
- 特殊数据类型（大对象、非关系型扩展、特殊数据类型）
- 应用开发与接口（内嵌 SQL、标准接口、与 WEB 技术的互操作、XML 支持等）
- 可靠性（错误恢复）
- 商业考虑（技术支持、市场份额）

测试的结果如表 6.6 所示。

在 DBMS 的选取上，商业软件提供了明显高于自由软件产品的功能。但从实际使用情况分析，其中的许多功能我们很少甚至从未使用的。根据天文学多年来对 DBMS 的使用经验，China-VO 倾向于使用 Oracle 系统。在免费的产品中，我们倾向于使用 MySQL。虽然 Postgres 提供了比 MySQL 更好的功能，但它们的功能差别不是本质上的。为了利用这些功能，我们可能需要在其他方面浮出更大的代价。

我们最终将从 Oracle 和 MySQL 中挑选一个作为 China-VO 的首选 DBMS。但在实际应用中可能会同时使用。

6.7 Grid 环境下的数据访问

虽然网络传输速度不断增加，但仍然是数据迁移过程的瓶颈所在。TB 甚至 PB 量级的海量数据的迁移对目前的网络而言实现起来还有相当的困难。在 VO 环境中数据迁移的原则是：***能迁移结果则不迁移数据，能迁移服务则不迁移数据***。

VO环境中数据的访问主要表现为两种形式：文件访问和数据库访问。VO是网格的一种高级应用，其环境中的数据访问也是在网格体系基础上实现的。网格标准的制定者GGF对于网格环境下的数据访问设有专门的研究领域，即“Data Area”[24]。这个部分下设两个工作组和四个研究组。两个工作组是数据库访问与集成服务（Database Access and Intergration Services，DAIS）和GridFTP工作组；四个研究组是：数据复制、数据传输、网格高性能网络、永久文档。其中两个工作组与VO数据访问关系密切。

种类	问题	评价		
		MySQL	Oracle8	PostgreSQL
基本特性	Basic data types	B	C	A
	SQL	C	B	B
	Declarative constraints	C	A	A
	Programming abstractions	D	A	C
	Generation of ids	C	A	A
	National chars	B	A	B
事务处理	Transactions	D	A	A
	Locks	D	A	A
	Multi-user access	C	A	C
数据库内编程	Stored procedures and triggers	D	A	A
管理	Access control	A	A	B
	Backup	C	A	C
	Data migration	A	B	A
灵活性与适应性	Portability	B	A	B
	Scalability	B	A	C
性能与超大数据支持	Query optimization	B	A	B
	Structures supporting optimization	D	A	B
	Support for OLAP	D	A	D
	Allocation of the disk space	C	A	C
	Size limits	B	A	C
	VLDB implementations	D	A	B
分布式数据库	Access to multiple databases	C	A	C
	Heterogeneous systems support	D	B	D
特殊数据类型	Large objects	B	A	C
	Post-relational extensions	D	A	B
	Support for special data types	D	A	C
应用开发与接口	Embedded SQL	D	A	B
	Standard interfaces	B	A	B
	Additional interfaces	A	A	A
	Web technology	B	A	B
	XML	D	A	D
	CASE	D	A	D
可靠性	Recovery	C	A	C
商业考虑	Prices	A	D	A
	Technical support	C	B	D
	Position on the market	D	A	D

表 6.6 CERN数据库测试结果[a]

GGF正在研究制定OGSA体系中的数据访问标准。GGF计划通过一层虚拟化服务实现网格应用对复杂数据的透明访问[25]。他们通过定义一层“网格数据

[a] A: 很好；B: 好；C: 一般；D: 差

虚拟化服务，GDVS”来提供数据访问和数据处理的透明化。这些服务将支持分布式数据的联合访问、基于内容的数据源动态发现、动态数据整合、数据并行处理、联合等功能。

网格上的数据应用按照使用目的的不同可以分为：

- 协作性（Collaboration）：数据来自许多站点，需要以统一的方式访问和处理；
- 扩展性（Scalability）：大量数据来源于少量几个站点，各站点进行高度的并行处理，站点间通讯不多。

在 VO 系统中，这两方面的情况都是存在的。不过，在目前阶段互操作性，即协作性，的需求是第一位的。在 GGF 中，他们将这种统一访问和处理称为虚拟化或透明性。这涉及数据访问透明性和数据处理透明性两方面的内容。

数据访问透明性包括：

- 异构透明性：使用统一的数据模型，从数据存储的真实实现中独立出来；
- 名称透明性：通过定义基于属性的逻辑域或者名字空间实现数据地理位置和副本的透明性；
- 属主和成本透明性： 这是长期发展时需要考虑的问题。

数据处理透明性包括：

- 并行透明性：处理应用能自动利用网格环境的并行处理能力
- 分布透明性：处理应用能自动利用网格环境的分布式处理能力

虚拟化几乎总是伴随着性能的损失。所以，网格除了支持虚拟化访问还必须支持对数据源的直接访问方式以满足性能的要求。

“GDVS”是 OGSA 的长期目标之一，这层服务位于应用与数据资源之间，屏蔽掉数据源的分布、异构和孤立的性质，它属于 OGSA 中的数据管理功能。

图 6.12 给出了 GGF 数据服务的体系结构，其中灰色部分是 OGSA 其他部分提供的服务。GGF 在网格应用和数据源之间定义了一层虚拟化服务。这些服务将网格的各方面进行了虚拟化，从终端用户角度看网格是一个单一的实体。

体系结构中包含一些核心数据虚拟化服务和附属服务。其中核心服务包括发现、联合访问、一致性管理、协作、工作流协调；附属服务包括授权、复制与缓存、Schema 管理。这些服务与 OGSA 中的其他一些服务紧密合作完成数据虚拟化的任务。OGSA 中涉及的其他服务主要包括注册、认证、记帐、通告

等。

6.7.1 GridFTP

在VO的文件访问中GridFTP将作为基本的数据传输方式。GridFTP协议[26]由GGF所属的GridFTP工作组开发，旨在实现文件在互联网上的高效、灵活、可靠的传输，这是网格应用必需的功能。目前 1.0 版协议草案已经完成，并将成为GGF的正式文档。

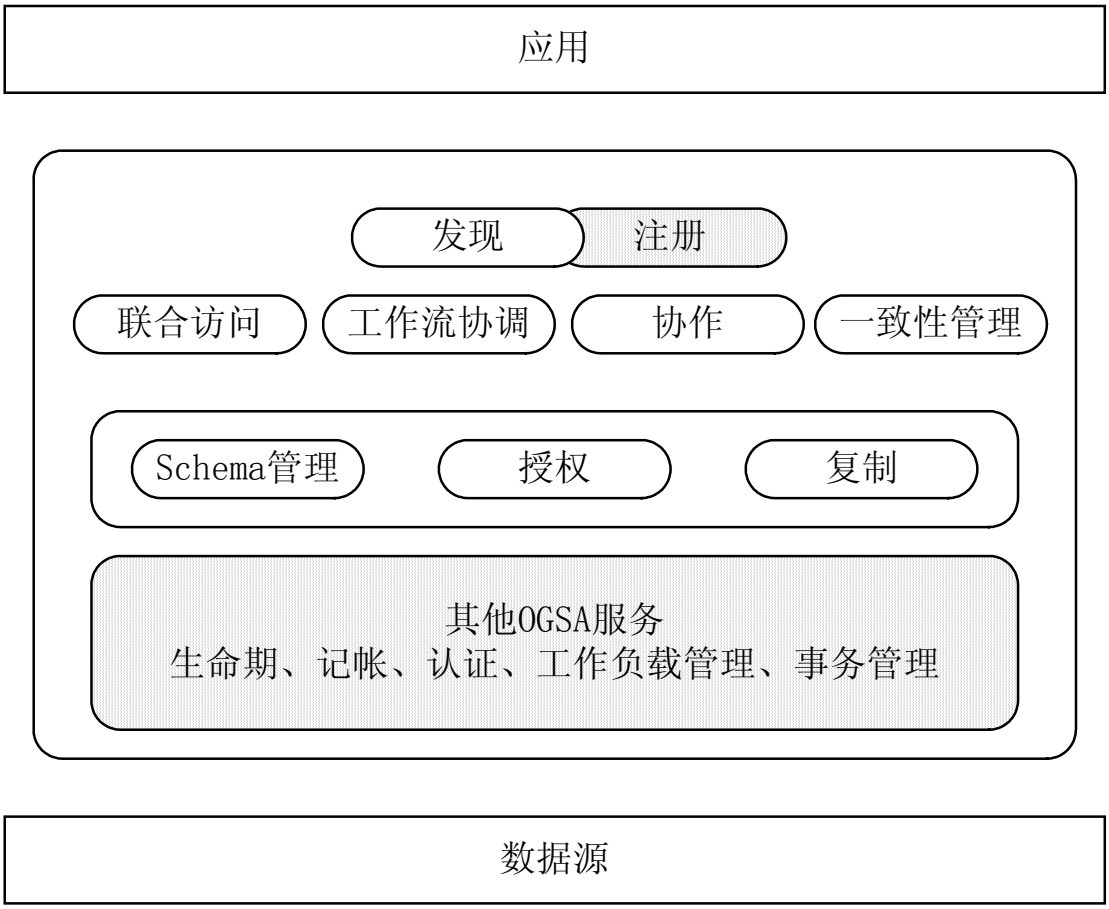


图 6.12 OGSA 数据服务的体系结构

这是一个通用的数据传输和访问协议，提供了网格环境下安全高效的数据迁移机制。这个协议对标准的 FTP 协议进行了可扩展，吸收了现有网格存储系统的一些功能。GridFTP 1.0 中规定的主要功能包括：

- GSI（网格安全基础架构）和 Kerberos 支持
- 第三方控制下的数据传输
- 并行数据传输
- 多点传输
- 部分文件传输
- TCP 缓冲和窗口大小的自动调整

- 可靠传输断点续传
- 数据完整性控制

GridFTP 协议已经有了多个实现工具，服务端和客户端工具都已经包含在了 Globus Toolkit 套件中。

6.7.2 SRB

除了按照OGSA的思想，利用GT2 和GT3 来搭建数据网格，China-VO还会考虑借鉴圣地亚哥超算中心（SDSC）开发的存储资源中介（SRB）[27]来实现分布式数据的管理。SRB虽然是“数据网格”的一种实现途径，但它不遵循OGSA标准。NVO的技术报告中指出SRB为分布式的数据资源管理提供了很好的实践经验[28]。

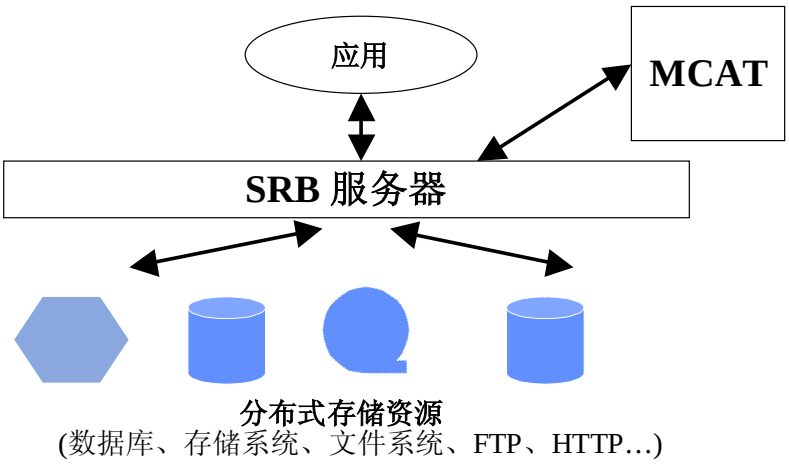


图 6.13 SRB 数据网格

SRB 是一个中间件，向分布式的客户端提供对异构计算环境中各种存储资源的统一访问。

图 6.13 中显示了 SRB 的一个简化体系结构。模型包括三部分：元数据目录（MCAT）服务、SRB 服务器、SRB 客户。不同部分通过网络连接在一起。

MCAT 存储有 SRB 管理下的数据集、用户和资源的元数据。MCAT 负责处理来自 SRB 服务器的请求。这些请求既有信息查询请求也有元数据建立与更新操作指令。

客户端应用的体现形式是一套 API。通过这些 API 向 SRB 发送请求和接受响应。SRB 服务器负责执行作业以满足客户端提交的请求。这些作业包括与 MCAT 的交互、I/O 操作。客户端通过共同的 API 访问 SRB 管理的所有存储系统。与各种存储系统以及各种操作系统、硬件体系之间的复杂交互都由 SRB 服务器处理。

6.7.3 数据库访问与集成

网格环境下数据库的访问与传统的数据库访问相比有其自身的特点，比如：高度动态的联合，在网格动态环境下的协同作业；极端性能要求，这是对（数据/计算）网格高性能的要求；数据源可替换，根据网格资源的使用情况动态的选取数据源；语义网格探索，需要借助资源注册发现机制实现作业；使用网格标准和服务，是一种平台无关的透明访问和处理。

网格数据库访问与集成的功能应该包括：

- 发布和发现，使用数据库服务描述发布数据库服务；
- 状态：通过一个标准的数据库状态接口，提供状态操作。每一个数据库操作都经过三个阶段：准备和检验、应用、结果返回。
- 结构化数据传输
- 高层次网格数据传输服务：
  - 通过一系列通道从一个数据源向多个目的地进行迁移
  - 提供系统的加密和压缩方法
  - 提供一致性的迁移成功与否通告和进程检测机制
  - 在不同的通道上可使用不同的协议
- 数据翻译和转换
- 事务处理
- 授权、访问控制、记帐
- 元数据：定义数据服务操作、数据内容、数据库功能描述标准
- 管理：实现异构数据库的操作管理和性能管理
- 数据复制
- 会话和连接
- 集成

China-VO 将在 OGSA 提供的网格数据库访问与集成机制的基础上，采用与网格系统兼容性好的 DBMS 实现数据资源的存储、管理与访问。

参考文献

[1] The Globus Project. Globus Toolkit Developer Tutorial - Basics. http://www.globus.org/about/events/US_tutorial/slides/Dev-02-Basics1.ppt

[2] J. McDowell, M. Cresitello-Dittmar, et al.. Data Model for the VO: Version 0.03, Part I: Overview. http://bill.cacr.caltech.edu/cfdocs/usvo-pubs/files/vodm003.ps

[3] J. McDowell, S. Lowe. Spectral Data Models, Draft: May 7, 2003. http://hea-www.harvard.edu/~jcm/vo/vospec.ps

[4] Francisco Valdes. A Virtual Observatory Data Model. http://iraf.noao.edu/projects/vo/dal/datamodel.html
[5] [HDX] http://www.astro.gla.ac.uk/users/norman/star/java/hdx/
[6] [IDHADM] IDHA Data Model for Astronomical Images. http://alinda.u-strasbg.fr/IDHA/lastmodel/
[7] [LaTeX] LaTeX project home. http://www.latex-project.org/
[8] Wells D.C., Greisen E.W. `FITS: a Flexible Image Transport System' in Image Processing in Astronomy. In: G.Sedmax, M.Capacciol. R.J.Allen (eds). Trieste. 1979：445
[9] [VOTable] http://vizier.u-strasbg.fr/doc/VOTable/
[10] S. Allen, D. Wells. Proposal to register MIME media types for FITS. http://www.ucolick.org/~sla/fits/mime/
[11] [Astrores] VizieR. Describing Astronomical Catalogues and Query Results with XML. http://vizier.u-strasbg.fr/doc/astrores.htx
[12] [UCD] Unified Content Descriptors. http://vizier.u-strasbg.fr//UCD/
[13] 姜志, 边歆. NAS 成为存储新宠. http://industry.ccidnet.com/pub/disp/Article?columnID=48&articleID=2493&pageNO=1
[14] Todd Volz. SAN and NAS defined. http://techupdate.zdnet.com/techupdate/stories/main/0,14179,2853222,00.html
[15] NAS 与 DAS 的比较. http://industry.ccidnet.com/pub/disp/Article?columnID=48&articleID=847&pageNO=1
[16] [RAID] KATHLEEN OHLSON. Redundant Arrays of Independent Disks. http://www.computerworld.com/hardwaretopics/hardware/story/0,10801,45211,00.html
[17] AC&NC. Get To Know RAID. http://www.acnc.com/04_01_00.html
[18] Clive Page. Database Technology for AstroGrid: a Discussion Document. http://wiki.astrogrid.org/bin/view/Astrogrid/DatabaseDiscussion
[19] [VOQL] VO Query Language. http://www.ivoa.net/twiki/bin/view/IVOA/IvoaVOQL
[20] [ASU] Astronomical Server URL. http://vizier.u-strasbg.fr/doc/asu.html
[21] [GLU] Générateur de Liens Uniformes. http://simbad.u-strasbg.fr/glu/glu.htx
[22] Clive Page, et al. DBMS Evaluations. http://wiki.astrogrid.org/bin/view/Astrogrid/DbmsEvaluations
[23] Konrad Bohuszewicz, Maciej Czyzowicz, Michal Janik, et al. Comparison of Oracle, MySQL and PostgreSQL DBMS. http://hep-proj-database.web.cern.ch/hep-proj-database/db_compar.htm
[24] GGF Data Area. http://www.ggf.org/6_DATA/data.htm

[25] Vijayshanker Raman, Inderpal Narang, Chris Crone, et al. Services for Data Access and Processing on Grids. http://www.cs.man.ac.uk/grid-db/papers/DAIS_DataServices.pdf

[26] [GridFTP] W. Allcock, J. Bester, J. Bresnahan, et al. GridFTP: Protocol Extensions to FTP for the Grid. http://www-fp.mcs.anl.gov/dsl/GridFTP-Protocol-RFC-Draft.pdf

[27] [SRB] SDSC Storage Resource Broker. http://www.npaci.edu/DICE/SRB/

[28] Robert Hanisch. Building the Framework for the National Virtual Observatory, NSF Cooperative Agreement AST0122449, Quarterly Report, Oct-Dec 2002.

第七章　　应用服务

数据挖掘、可视化和高性能计算服务等高层应用服务是最能体现虚拟天文台天文特色的服务，也是 VO 能否最终让天文学家接受和使用的关键。由于在体系结构上的分布式、网格化等特点，VO 对这些领域提出了挑战。但同时我们看到这些挑战也是其他应用领域以及 IT 业界所共同面临的。China-VO 将本着“有所为有所不为”的方针，主要采用移植现成技术的方法将其他领域的成功案例应用到 VO 中，服务于天文学。China-VO 将突出自身的特色，把研究与开发的重点放在光谱巡天数据的自动处理与分析领域。同时，China-VO 将与中国国家网格紧密合作，利用其强大的计算和网络资源为用户提供高性能计算服务。

7.1　从数据到知识

数据分析产生信息，信息产生知识，最终形成“理解”。“理解”产生新的问题，推动人类收集更多的数据。这是一个不断循环，不断深化的过程，也是科学研究方法的本质。

由于各种技术（如计算机技术、互联网技术、空间观测技术等）的飞速发展，各个领域正面临着一场“数据爆炸”，数据量呈指数增长。在未来数年里，将产生比过去所有数据总和还要多的数据。人们已逐步意识到蕴含在这些数据中的威力。但是我们对宇宙的理解并没有与观测数据的增长同步，而是慢得多。正如Ian H. Witten和E. Frank在他们的著作“Data Mining”中所指出的“数据与我们对其理解之间的鸿沟正变得越来越大，这是不争的事实。”[1]

造成这样的结果主要是两方面的原因：

- 人脑本身的限制

数据越来越复杂，对人脑的理解力提出了挑战。我们最熟悉的空间维度是二维、三维，也就是平面或者立体。虽然我们能在更高维的空间里进行数学上的理解，但对理解力的要求越来越高。虽然我们自己的神经网络是强大的模式识别工具，但维度的增加使我们理解的难度呈指数增长。我们需要更好的分析工具和可视化工具来帮助我们对这些数据进行理解。

- 方法上的限制

我们分析和处理数据的方法和技术远远滞后于数据的增长，获取知识的方法并没有与数据增长同步，而是大大滞后。

数据复杂性增加的一个主要表现是维度的提高。高维数据带来的问题主要集中在两个方面：计算量的急剧增加和可视化问题。虽然存在一些降低维数的方法，但这远远不够。为了发现新类型的天体和现象，VO 需要借助数据挖掘和可视化工具，需要 TB 和 PB 量级数据情况下高效的数据挖掘和可视化工具，需要更好的聚类算法和统计探索工具。虚拟天文台将在存储技术、信息管理、数据处理、分布和并行计算、高速网络、数据可视化和数据挖掘等各个领域突破现有技术的限制。这需要天文学家与 CS/IT 专家、统计学家的更好合作。

如果一个天文台只有望远镜而没有处理观测数据所需的计算机、软件包等软硬件配套设施，与其说这是一个天文台倒不如称之为“天文观测基地”。一个真正的天文台不但要能取得观测数据还要对这些数据进行处理，让天文学家能从中获得信息和知识，最终加深我们对宇宙的理解。

虚拟天文台也是如此。在这个系统中，各种数据资源构成了数字星空，建立在标准数据模型基础上的数据访问服务实现了望远镜的功能。为了让虚拟天文台成为一个健全的“天文台”，还必须有各种配套研究设施，这就是本章所要讨论的数据挖掘服务、可视化服务和计算服务等应用服务。

7.2 数据挖掘

面对海量数据，我们将面临许多非常现实的挑战，例如怎样记录、加工原始数据；怎样通过现代计算机硬件和网络系统存储、合并、获取数据；怎样快速有效地探索及分析数据并将这些数据可视化。在这种形势下，数据挖掘是虚拟天文台成功应用的重要因素。虚拟天文台要实现全球范围内分布式数据的透明访问，让天文学家以一种友好的方式从大量异构的数据集中发现、访问、分析和整合天文数据。VO 需要数据挖掘：

- VO 的主要功能及其相关技术涉及到数据挖掘的诸多方面；
- VO 是一个对异构数据与信息服务进行分布式数据挖掘的基础设施。

随着计算机技术、数据库技术、统计学、数学、机器学习等方面在近几十年的长足进步，数据挖掘和知识发现从中分流并发展成为一门新型学科。知识发现就是对数据进行抽取和提炼，从而取得新知识的过程，是数据库研究中的一个很有价值的新领域。它融合了数据库技术、人工智能、机器学习、神经网

络、统计学、模式识别、知识库系统、知识获取、信息检索、高性能计算、数据可视化等多个领域的理论和技术。相对而言，数据挖掘主要流行于统计领域、数据分析、数据库和管理信息领域；而知识发现则主要流行于人工智能和机器学习领域。

具体到天文学中，数据挖掘是指从天文数据中提取信息和发现知识，更具体地说，就是从海量数据中发现稀有的天体或现象, 或者发现以前未知种类的天体或新天文现象。近年来，这方面的研究已成为天文数据研究领域的热点。

数据挖掘利用复杂的技术建立模型，从数据中发现模型和相关性。模型分为两类：描述性模型和预测性模型。描述性模型，即描述数据中的模式，并用以创建有意义的群或者子群；预测性模型，即利用从已有的数据中推出的模型来预测未知事件。

数据挖掘分为事件性数据挖掘和相关性数据挖掘。事件性数据挖掘进一步分为四类：

- 已知事件/已知算法：用已有的物理模型去确定数据中存在的人们感兴趣的已知现象；
- 已知事件/未知算法：用模式识别或数据的聚类特性来发现已知现象中存在的新的观测相关性；
- 未知事件/已知算法：以天文现象的观测参数中存在的预期的相关性来预测数据中存在的以前未知的事件；
- 未知事件/未知算法：用临界值确定瞬时事件或独特事件，从而发现新现象。

相关性数据挖掘则分为三类：

- 空间相关：在天空中的同一位置证认天体；
- 时间相关：证认发生在相同时间或相关时间的事件或现象；
- 一致相关：用聚类方法证认存在于同一多维参数空间的现象。

7.2.1 数据挖掘的功能

数据挖掘的功能主要包括数据的汇总、分类、聚类、关联规则发现或序列模式发现等。

分类在数据挖掘中占据着重要地位。分类的目的是提出一个分类函数或分类模型（也常常称作分类器），该模型能把数据库中的数据项映射到给定类别中的某一个。分类的方法有统计方法、机器学习方法、神经网络方法等等。统计方法包括贝叶斯法和非参数法（近邻学习或基于范例的学习）；机器学习包括决策树法和规则归纳法；神经网络方法主要是前向神经网络的反向传播算法

（Backpropagation Algorithm， BP 算法）及 Kohonen 学习矢量量化方法 (Learning Vector Quantization, LVQ)；另外，最近又兴起了一种新的方法，粗集 (Rough Set)。

聚类是根据数据的不同特征，将其划分为不同的数据类。其原理是使得属于同一类别的个体之间的距离尽可能的小，而不同类别的个体间的距离尽可能的大。聚类方法包括统计方法、机器学习方法、神经网络方法和面向数据库的方法等。

相关性分析的目的是发现特征之间或数据之间的相互依赖关系。经常用的技术有回归分析、关联规则等。

偏差分析的基本思想是发现观测结果与参照量之间的有意义的差别，包括分类中的反常实例、例外模式、观测结果与期望值的偏离以及量值随时间的变化等。通过发现离群数据 outliers，可以发现一些不同寻常的或奇异的天体，如褐矮星和高红移类星体的发现。

7.2.2 VO 数据挖掘的特点

- 海量数据

目前单个天文数据集已经达到数TB的量级，比如SDSS DR1[2]总数据量近3TB。未来几年，单个数据集将突破 10TB，而总的天文数据将突破PB量级。

- 非线性

天文数据属性之间的非线性关系是天文系统复杂性的重要标志，其中蕴含着系统内部作用的复杂机制，因而被作为天文数据知识发现的主要任务之一。

- 高维

多波段性是指天文数据在不同观测波段上所遵循的规律以及体现出的特征不尽相同。这是天文数据复杂性的又一表现形式。天文数据的属性增加极为迅速，例如由于空间技术的飞速发展，覆盖的波段的数目也由几个增加到几十个甚至上百个。如何从几十甚至几百维空间中提取信息、发现知识则成为研究中的又一重要任务。

- 缺值

缺值现象起源于某种不可抗拒的外力（如仪器的灵敏度低、天气恶化等，一些天体在一个或多个波段探测不到）而使数据无法获得或丢失，如何对丢失数据进行恢复并估计数据的固有分布参量，成为解决数据复杂性的难点之一。

天文数据所表现出的上述复杂性特征为相应的数据挖掘和知识发现研究提出了更高的要求，并成为推动其发展的强大动力。

7.2.3 VO 数据挖掘的主要任务

- 天体的交叉证认：以源的位置为参量，将存在于不同数据库中的源联系起来，用来加深对证认源的天文理解，例如寻找 γ 暴对应体。
- 天体的交叉相关：用假定分析方法处理数据中的所有参数，例如在 DPOSS 和 SDSS 巡天中，通过在双色图中远离正常恒星区的特性发现高红移类星体。
- 最近邻规则证认：在多维空间中运用聚类算法证认天体或天文现象，如在长蛇座中通过天体具有相似的运动学特征、X 射线发射特征、Hα 线特征和 Li 丰度，发现了人们最熟悉的年轻恒星族。
- 系统的数据探索： 在数据库中广泛地应用事件性和相关性数据挖掘技术可以偶然发现一种新天体或新类型天体，例如在 MACHO（Massive Compact Halo Object）数据中发现的“bumpers”。

7.2.4 主要数据挖掘技术

在天文中会遇到各种各样的问题，面对这些问题如何处理，这是摆在天文学家面前的重要课题。表 7.1 中列出了一些天文中经常遇到的问题及其处理的方法[3]。

神经网络

神经网络是模仿人脑神经网络的结构和某些工作机制而建立的一种计算模型。其特点是利用大量的简单计算单元连成网络，来实现大规模并行计算。由于神经网络非常适用于处理天文数据的非线性复杂关系，并且在处理复杂问题时不需要了解网络内部所发生的结构变化，因而被广泛地应用于天文数据挖掘和知识发现的研究中，并以不同的网络结构实现了空间聚类、分类、关联、回归、模式识别等多种算法。

神经网络在天文中有广泛的应用，如星表的提取、恒星与星系的分类、星系形态的分类、恒星光谱的分类等。

统计方法

统计方法是从事物外在数量上的表现去推断该事物可能的规律性。常用的统计方法有回归分析（多元回归和自回归等）、判别分析（贝叶斯判别和非参数判别等）、聚类分析（系统聚类和动态聚类等）以及探索性分析（主分量分析法和相关分析法等）等。

高维数据的挖掘算法

问 题	例 子	常 用 方 法
天体分类	恒星/星系分类 星系形态分类 恒星/星系/类星体	学习矢量量化(LVQ) 支持矢量机(SVM) 主分量分析(PCA) 自组织映射(SOM) 模糊集理论 神经网络(NN) 小波变换 决策树
图像分类	数字巡天中的恒星/星系区别	学习矢量量化(LVQ) 自组织映射(SOM) 模糊集理论 神经网络(NN) 最近邻规则 聚类分析 决策树
数据压缩 与分类	光谱压缩和分类	主分量分析（或KL变换） 独立分量分析(ICA) 信息瓶颈(IB) Fisher 矩阵
重建方法	大尺度巡天中的图像重建	均方差估计（UMV） 小波分析 维纳滤波 shapelet 公式 傅立叶拟合 变像素线性重建 最大熵方法（MEM） Massive Inference Pixon 方法
大尺度结构 分析	有关大尺度结构和微波背景辐射的大尺度巡天	独立分量分析（ICA） 最大熵方法(MEM) 贝叶斯分析 小波分析 错误发现率（FDR） N点相关函数 FastICA

表 7.1 天文中常遇到的问题及其处理方法

在近期的研究中，对高维数据进行挖掘的思路一般有两条，一是将高维数据通过线性变换投影到低维空间，然后再实施其他挖掘算法；另一种就是采用适合处理高维数据的算法直接对其进行信息提取。

7.2.5 VO 数据挖掘所面临的主要任务

- 扩充数据挖掘算法：随着观测记录和观测次数的增长，观测参数的增长，分析观测数据的预测模型数增长，同时对交互式反应和真实反应时间减少的要求加强，需要多种算法组合或发展新的算法。
- 数据挖掘方法应用到新的数据类型：如时间序列的数据、无结构的数据（文本数据）、半结构性的数据（HTML 和 XML 文件）、多媒体关联数据、多层次多度量单位的关联数据、集合数据。
- 发展分布式的数据挖掘算法：由于数据的分布特性和计算环境越来越分散，故必须发展与之匹配的数据挖掘系统和算法。
- 提高数据挖掘方法的易用性：这包括数据挖掘自动化程度的提高，优化用户界面以便支持随机用户的浏览，提高大型分布数据的可视化程度，发展用以管理数据挖掘的元数据技术和系统，发展恰当的语言和协议支持数据随机提取，改善数据挖掘和知识发现的环境，这涵盖从数据收集到数据加工，到数据挖掘，再到可视化以及结果的评估和解释的整个过程。

数据挖掘领域，包括大型、多变量数据集的可视化和统计分析，目前尚处于幼年期，但在未来许多年中将是非常活跃的研究领域。VO 数据挖掘是一个需要天文学家、计算机科学家、数学家和软件专家一起合作才能解决的多学科问题。

同时，在使用 VO 的数据挖掘和知识发现服务过程中，VO 用户的作用是至关重要的。可以说 VO 用户是联系 VO 数据与数据挖掘工具的枢纽。如何利用数据挖掘工具将数据转化为知识，这对用户提出了更高的要求，他不仅要懂天文而且要会使用数据挖掘工具，需要既具备扎实的天文功底，还要了解数学、统计学、计算机和模式识别等方面的知识。

7.3 可视化

当前，天文仪器产出大量的观测数据越来越容易，但天文学家从这些数据中提取信息的难度并没有降低。虽然计算机的运算速度越来越快、价格越来越便宜，可以承担大量的数据处理，但是天文学家和计算机之间的界面却一直阻碍着天文学家把这些数据以有效的方式聚集在头脑中。数据中的有用信息不能

得到有效提取，从而导致许多信息丢失。

可视化是数据理解的关键。在模式识别、形态学分析方面，人眼是一种有效的工具。视觉信息是产生知识和形成理解的重要桥梁，俗话说得好“一幅画胜过千言万语”，许多情况是“只能意会不易言传”的。可视化工具对 VO 是必不可少的。

数据可视化的好处：

- 有利于对数据直观的理解；
- 有利于突出其他方式不容易察觉的数据的特征，对数据有更全面的理解；
- 有利于得到数据定量的分析结果；
- 有利于将结果与他人进行定性或者定量的交流。

可视化服务是 VO 系统必要的组成部分，它可以引导查询过程的执行和显示查询结果。用户要利用可视化服务对不同目标参数的组合进行绘图。可视化工具可以作为理论天文学家和实测天文学家之间的接口，实现理论模型的预测结果和基于统计相关分析的经验结果之间的比对。

7.3.1 VO 可视化的特点与要求

在 VO 体系中，由于所有的服务都基于标准的数据模型和服务模型，系统中的数据有比较统一的交换格式，所以在进行 VO 服务的开发时 I/O 格式、类型的适用性、兼容性要求较低，更注重的是服务的功能和与其他服务的互操作性。

对数据挖掘和可视化工具的使用要求：易学易用、与其他工具的兼容性好、功能强大、良好的用户开发环境、与网格环境和分布式环境良好融合。

对 VO 应用服务的一个重要要求是易用性好。如果用户在不必参阅任何手册的前提下就能完成所需的操作，这样的 VO 服务必是非常成功的。

7.3.2 可视化功能范畴

为了熟悉可视化可以实现那些功能，下面以天文学上使用的比较普遍的 IDL 软件包为例说明。

IDL（Interactive Data Language）[4]是美国RSI公司(Research System Inc.) 开发的交互式数据语言，是进行二维及多维数据可视化表现和分析及应用开发的软件工具。作为面向矩阵、语法简单的第四代可视化语言，IDL致力于科学数据的可视化和分析，适用于跨平台的应用开发。它集可视化、交互分析、大型商业开发为一体，为用户提供了完善、灵活、有效的开发环境。

IDL 被广泛应用于地球科学、航空航天、医学影像、图像处理、软件开发、大学教学、实验室研究、测试技术、信号处理、防御工程、数学分析、统计等诸多领域。在天文研究中也得到了广泛的应用，许多大的科研项目都基于 IDL 进行了应用软件的开发，比如 HST、SDSS。

下面我们就以 IDL 目前最近推出的 5.6 版本为例来了解传统数据可视化工具所能提供的功能，并在此基础上讨论 VO 系统中提供的可视化服务所应具备的功能。

IDL 的主要特性包括：高级图象处理能力、交互式二维和三维图形技术、面向对象的编程方式、OpenGL 图形加速、量化可视化表现、集成数学与统计学算法、灵活的数据输入输出方式、跨平台图形用户界面工具包、连接 ODBC 兼容数据库及多种程序连接工具等。

图 7.1 是 IDL 的功能结构图，包括数据访问、数据分析、数据可视化、GUI 工具、开发环境、外部开发环境、发布七大功能模块。

从图中可见，数据访问、数据分析和数据可视化处于整套软件的核心地位，其许多功能对 VO 应用服务有借鉴价值。与这三部分相关的功能特色主要包括：

1. 数据访问

提供对不同数据源（本地、网络）、文件格式、数据库、数据格式、数据类型、数据/文件大小的支持。

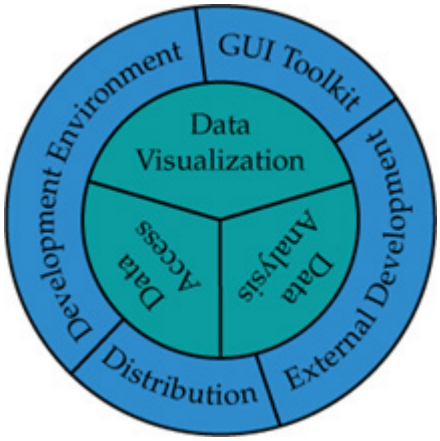


图 7.1 IDL 功能结构图

2. 数据分析

丰富的类库，其中包含：数学、统计学、图像处理、信号处理等领域，以

帮助用户分析数据；对多线程计算的支持。

数学与统计学

- 数学操作符和基本函数
- 线性代数包，包含特征向量和特征值的计算、同步线性等式解决的各种方法
- 非线性计算
- 稀疏型线性系统和稀疏型矩阵处理
- 从基本的统计计算到误差分析、假设检验和随机数据生产的统计学方法
- 时序分析
- 多元数据分析

图像处理

从基本的处理手段，例如平滑、锐化和对比度拉伸，到高级应用，例如边缘检测、生态学应用和离散小波变换等等；完备的小波分析工具。

信号处理

- 1-、2- & 3-D 卷积
- 自适应快速傅立叶变换
- 小波变换工具包
- Bi-level、假彩色、真彩色阈值处理
- 块卷积
- 真彩色转换为假彩色
- 彩色系统：RGB、HLS、HSV，彩色索引表
- 卷积和频率域块卷积
- 傅立叶变换：1 到 8 维，任意点数，多线程
- 频率域滤波和分析
- 普通图像算法
- 几何变换：放大、缩小、转动、多项式
- 规则或不规则网格旋转
- 高、低通滤波
- 自适应直方图均衡和处理
- 图像注记、统计功能
- 交互式拉伸增强
- Lomb 周期表
- 中值滤波

- 形态学算子：腐蚀（erode）、膨胀(dilate)、开(open)、闭(close)等
- ROI 选择及 ROI 对象生成
- 区域增长算法
- Roberts、Sobel 边缘增强
- 信号编辑
- 波谱分析
- 时间序列分析
- 波形产生
- 使用 Daubechies 系数的小波变换
- Zoom 窗口及放大
- 估算图形的 N 维欧几里得距离
- 对图像局部进行基于像元的统计
- 一系列形态函数及分水岭分割算法
- 矢量区域路径跟踪
- Hough 及 Radon 变换功能
- Savitsky-Golay 滤波
- 支持 Date/Time 数据绘图
- watershed 图像分割
- 2D、3D 的网格与插值
- 曲线与曲面的线性拟合、独立变量拟合以及非线性拟合

3. 数据可视化

IDL 对复杂的海量数据的可视化进行了许多优化设计，适用于 2D 绘图、图像显示、交互式 3D 图形设计以及体数据的显示、处理和分析。IDL 提供两种图形显示系统：直接图形法和对象图形法。

- 直接图形法就是将图形直接画到当前设备中去，而无论该设备是显示器、打印机或者其他。该方式通常是静态的，实现速度快是它的优势。
- 对象图形系统则可以利用 OpenGL 的图形加速技术进行显示，而没有当前设备的概念。该图形法支持真三维对象的显示与分析，使用户更易完成三维显示之后的高级分析处理工作。

IDL的可视化功能主要包括：

2D 绘图：

- 简单的 2D 绘图常常是数据通信与理解的关键。线图、散点图、对数坐

标绘图、极坐标绘图、直方图、时序图等；误差棒、多坐标轴、多数据集绘图；图形显示方式的定制，比如线性、字体、符号、颜色等等。

基本的图像处理：

- 比如平滑、增强、边缘检测、形态学操作、离散小波变换

绘制等值线：

- 在离散数据上生成网格和等值线，用封闭曲线创建填充等值区域和添加标签，在图像、三维地形上叠加等值线。

矢量应用：创建矢量流表格和三维显示

地图绘制

曲面、多边形网格和 3D 绘图：

- 浏览和交互式处理曲面绘图、3 维线条、离散数据以及复杂对象
- 在曲面和 3 维视景显示时控制光源与阴影
- 纹理叠加，快速生成三维景观
- 使用辅助数据来实现 3D 曲面或者对象腔体的色彩匹配
- 离散网格显示或者三角网格显示
- 任何地理数据都可以用多边形网格的形式显示
- 输入显示 DXF 对象和属性

体数据分析和四面体网格：

- 实体数据交互式的任意切割和显示
- 实体数据的分析
- 显示质量和速度的控制
- 定义光源模型和运用色彩以及透明显示
- 将离散 3D 数据生成网格或者通过四面体网格显示不规则几何形状
- 从体数据和四面体网格数据提取等值表面和内腔体
- 在体数据中生成流线

动画显示：

- 通过改变数据或者图形属性进行动态可视化显示
- 可以生成有虚拟地形和高分辨率纹理叠加的飞行视景
- 可以生成 MPEG 文件

7.3.3 IDL 在天文学上的应用

由于IDL杰出的数据分析和可视化功能，使得它在天文学上得到了广泛的使用[5]。天文学家利用IDL提供的编程环境开发了各式各样的程序库。应用范畴涉及许多方面：

常用天文工具：

- 格式转换（度分秒，弧度，实数，十进制、六十进制）
- 坐标转换（赤道、黄道、银道坐标；地平、赤道）
- 基本绘图
- 文件读取
- 常用计算（进动、历法变换、从流量到星等、空间速度、球面天文学计算、天测计算）

测光处理

图像处理：

- 图像与点扩展函数的卷积处理
- 图像相关处理
- 图像匹配
- 宇宙线剔除
- 天光线剔除
- 图像伸缩
- 逆卷积运算
- 中值滤波

数据库处理

磁盘输入输出

FITS 数据处理

数据和统计分析

绘图

在线数据访问与检索

7.3.4 China-VO 可视化工作

当前，VO 还处于最初阶段，主要目标是实现数据访问和互操作。数据挖掘、可视化功能的实现都要以此为基础。

VO 可视化工具的实施过程可以分为两个阶段：

- 一个相对简单的浏览器，提供一些基本的功能

- 一个相对高级的浏览器，提供更高级的功能

VO 提供的基本可视化功能应该实现：

- 基本的二维绘图，包括直线曲线图、散点图、对数坐标绘图、极坐标绘图、直方图等
- 光谱图像的显示
- 时序数据的显示，对时间坐标的支持
- 在坐标元数据的配合下，巡天图像的拼接与匹配
- 图像的伸缩与转动
- 多数据集的图像、数据迭加
- 图像标记功能

VO 的高级可视化可能实现：

- 等值线绘制
- 实时数据的动态显示
- 曲线与曲面的拟合
- 对磁场观测矢量数据的可视化
- 图像的傅里叶变换
- 图像的增强与目标探测
- 图像滤波

从目前看来，VO 最重要的可视化工具是一个好的浏览器，能方便有效的实现图像、光谱和星表数据的综合显示。现有的浏览器可以对本地数据实现上述功能。我们有必要对它进行扩充以便能处理网络上的有关联的数据（利用元数据决定数据之间是否有关联），将大量的数据以某种有意义的图像形式显示出来。同时，将其他数据的连接和操作模块也包含在浏览器中，让浏览器成为探索数据资源的向导。

IVOA成员已经开始了在可视化方面的尝试，比如实时数据可视化和分析工具PartiView[6]，VOTable数据显示工具VOPlot[7]。

目前，一些商业的和免费的软件产品，比如上面介绍的 IDL，提供了一些很有价值的功能。我们需要根据 VO 的情况对这些产品进行评估，借鉴它们的功能和实现方法。这些软件都没有实现网络化，VO 的数据挖掘和可视化服务是在分布式的网格环境下实现的。如何将这些高级功能在崭新的环境中实现，这是 VO 以及 IT 业界共同面临的一个挑战。

在目前阶段，应用服务的开发与提供不是 China-VO 的重点，但我们要从现在就开始考虑需要提供哪些高层服务、如何在 VO 环境中实现这些服务。正

如前面章节所述，China-VO 的高层服务将突出与 LAMOST 相结合的特色，在光谱自动处理与分析、红移自动测量、光谱自动分类等方面进行重点开发。

对于计算应用，China-VO 将与中国国家网格（CNGrid）紧密合作，充分利用 CNGrid 强大的网格计算资源，以其为试验床为用户提供一定数量的高级计算服务。

参考文献

[1] Ian H. Witten, Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.

[2] [DR1] SDSS Data Release 1. http://www.sdss.org/dr1/

[3] 张彦霞, 赵永恒, 崔辰州. 天文学中的数据挖掘和知识发现. 天文学进展, 2003 (4): 312-323

[4] [IDL] Interactive Data Language. http://www.rsinc.com/idl/

[5] [IDL-Astro] IDL Astronomy User's Library. http://idlastro.gsfc.nasa.gov/homepage.html

[6] [PartiView] Partiview Visualization Software. http://www.haydenplanetarium.org/hp/vo/partiview/

[7] [VOPlot] VO-India VOPlot. http://vo.iucaa.ernet.in/~voi/voplot.htm

第八章 门户与用户

在 China-VO 体系结构的最高层，即用户层，要实现用户对 VO 资源与服务的访问以及处理结果的返回。China-VO 门户是用户访问系统资源的重要途径。为了迎合不同类型用户的需要，China-VO 需要提供不同类型的用户界面和访问环境。虽然 China-VO 的服务主体是专业用户，但 China-VO 也将重视非专业用户的需求。China-VO 将充分利用 IVO 丰富的资源在科普教育方面有所作为，为国内的非专业用户带来真实的天文学体验。

8.1 门户

VO 门户提供了虚拟天文台与用户的直接接口，是访问 VO 服务的主要甚至唯一的途径。用户通过 VO 门户访问 VO 的各种资源和服务。门户的基本职能是用户任务提交和处理结果返回。门户的最终目标是为虚拟天文台的用户提供一站式服务。让用户从任何一个虚拟天文台门户都可以使用到所有的可用资源和服务，包括天文数据、计算、应用工具等，同时用户也可以从任何一个门户发布自己的数据和服务。

简单的说，VO 门户类似一个网站。它提供了 VO 资源浏览、任务提交和结果返回功能。现在天文学界已经存在许多很受用户欢迎的门户网站，它们都在不断的优化自己的服务和用户界面。但是，它们大多是各自为政，相互之间缺乏协调。将来的虚拟天文台并不会抛弃这些现有的网站，而是要对它们在互操作性、可扩展性、适应性等方面的功能进行扩充。在此基础上，再开发新的面向网格的超级天文门户。

8.1.1 VO 原型调研

为了明确VO门户应该提供哪些功能，AstroGrid对当前一些有代表性的天文门户和工具进行了调研[1]。几乎所有的这些软件和网站都是基于数据网格和WEB服务之前的技术。这些已经存在的设施为我们提供了大量的用户需求信息，从中可以了解VO应该提供怎样的服务。

AstroGrid 调研的门户和工具包括：

- Astrobrowse[2]
- Browse/W3Browse[3]
- CURSA[4]

- ISAIA[5]
- MAST[6]
- NED[7]
- Querator[8]
- Simbad[9], VizieR[10] 和 Aladin[11]
- Skyview [12]和Skymorph[13]
- Skycat[14], GAIA[15] 和JSkyCat[16]
- Starcast[17]
- Starview[18]

有些门户提供的查询界面与我们希望 VO 门户提供的界面比较相似，比如 Vizier、NED 和 Astrobrowse。并且这些站点已经可以提供对多个网站的同时访问。

一些站点，比如 Vizier、Aladin 和 Skyview，它们可以让用户以统一的方式访问大量的数据集。但前提条件是这些数据集必须在本地以一种特定的格式进行管理。VO 一个基本的目标就是实现数据集的统一访问，不过 VO 的数据集是分布式的，不可能存在统一的管理中心。

现有的站点和工具通常把结果以网页的形式展现出来，还不能以某种语义形式提供以便进行更深层的处理。VO 的要求是查询，然后对结果进行处理。这需要 VO 提供一套完善的语义机制以实现机器对数据的理解。这方面必须要与资源和服务的注册、元数据紧密结合。

在AstroGrid于 2002 年早些时候完成对以上系统的调研后，Clive Page等在 2002 年 11 月又对新近的两个VO相关原型站点进行了测试[19]。这两个原型分别是：Sky Server[20]和Virtual Sky[21]。

Sky Server （skyserver.sdss.org）提供了SDSS EDR数据产品（大约 80GB、1400 万天体）的公共访问服务。这个站点大量使用了Microsoft的产品，比如数据库采用的是Microsoft SQL Server ，Web服务器采用Terraserver 以及IIS，同时网页中大量采用Active Server Pages。这是 JHU和Microsoft密切合作的体现。通过采用JHU开发的Hierarchical Triangular Mesh（HTM）空间索引方法，Sky Server成功实现了SDSS EDR和另外两大巡天数据 2MASS和VLA FIRST的联合操作。在Jim Gray博士的帮助下， Sky Server已经在China-VO实现了镜像[22]。

Virtual Sky（virtualsky.org）项目是 CalTech、Microsoft Research、SDSS、JHU 联合开发的。项目的目的是：提供高质量的无缝的夜空图像。Virtual Sky 通过易用直观的界面提供了对整套 DPOSS（Digital Palomar Observatory Sky Survey）巡天数据的访问。从站点的风格和服务内容来看，它主要的服务对象

是业余天文学家和普通公众。

上面调研的系统中有很多很好的特性，许多都值得我们效仿。但是我们还是可以明显的看到许多功能都没有实现或者做得很不够。例如，大部分站点都支持交互式的查询，但其中很少能提供批处理功能，比如以一系列坐标为输入进行查询。VO 门户需要提供对工作流和批处理的支持，让工作流、资源浏览器等组件能在门户中运行。

VO 门户必须提供丰富的调用 VO 后端资源和服务的客户端服务。这些服务相对于 VO 用户来说是服务端，接收用户提交的指令和作业请求，并向用户返回结果。但它们对于 VO 的后端服务来说则是客户端，它们需要调用后端的 VO 服务来完成用户的作业。

8.1.2 AstroGrid 门户

AstroGrid 最近对其项目门户（http://www.astrogrid.org）进行了重新设计。虽然他们称其为 AstroGrid 门户，但与我们讨论的 VO 门户有许多明显的差别。目前他们的这个门户更准确的说是一个项目网站，而不是 VO 服务门户。

关于这个新门户，他们目前的基本思路是：

- 利用Cocoon[23]重新开发简化的门户；
- 为工作流、MySpace 浏览器、数据浏览器开发门户；
- 与 AstroPass 协作开发简单的组件。

在重新设计的门户上，静态网页是用Cocoon生成的，包括标题、脚注、导航条和网页内容。Cocoon的输出是纯正的XHTML。布局和颜色由单独的CSS（层叠式样式表）控制。除了静态网页，门户还提供了一些协同工具。AstroGrid从工具软件的功能特征和开发队伍的情况（是否容易介入、可否融洽的共同开发），从开放源码项目中选取的这些工具。比如，新闻站点使用的是GeekLog；论坛目前使用的是OpenBB，计划选用phpBB[24]；最有特色的是他们选用的综合性的协同套件Wiki[25]，这已经影响了包括IVOA和Aus-VO在内的一些VO项目。

AstroGrid 门户希望能在将来提供下面的一些功能：

- 门户模板
- 单一注册和登陆
- 文档上传和版本控制
- 源码上传和版本控制
- 项目外部网：日历、任务列表等
- 学习工具：手册、文档等

- 在线会议系统：聊天室方式、BBS 方式、WEBCAM 方式

AstroGrid 成立了四个工作小组负责门户的开发，分别是：门户结构设计、登陆、用户注册和用户环境。他们从开放源码软件中选择门户开发平台，比如他们选取的：

- **Coccon** Apache Cocoon 是一个XML发布框架，使用XML和XSLT技术。
- **JetSpeed**[26] 一套利用Java和XML的开发源码企业信息门户套件。
- **Velocity**[27] 基于Java的一个模板工具。
- **Turbine**[28] 基于servlet的一个开发环境，让有经验的Java开发者快速搭建安全网络应用

上面这 4 套软件均来自 Apache 社区。是 Apache 为实现 Java 在互连网上的应用提供的开放源码软件。

8.1.3 MySpace

作为与门户紧密相关的一个网格服务，AstroGrid 提出了“我的空间（MySpace）”的概念，以 Grid 服务的形式为 VO 用户提供个性化的 VO 存储空间。

MySpace 是个很有趣的服务。这个空间就好像是用户在 VO 环境中的家。这个空间也许分布于不同的计算机和硬盘上，但对于用户是完全透明的。就像在自己的电脑上利用资源管理器管理自己的文件一样，用户可以在 VO 环境中管理自己 VO 中的资源。

MySpace 为天文学家提供了在 AstroGrid 系统中的存储功能。天文学家可以在 MySpace 中保存在 AstroGrid 系统中工作时会经常用到的数据、查询结果、工作流、个人数据等。用户可以像对待本机上的文件一样对待 MySpace 中的资料，可以方便的在本机和 MySpace 之间转移资料。

AstroGrid 设计的 MySpace 的第一个对外版本将由这些组件构成：MySpace 服务器、MySpace 注册、数据迁移器、MySpace 浏览器。他们的 MySpace 开发队伍分工如表 8.1 所示。

MySpace 将用于 VO 用户保存临时结果和数据，是协同工作的关键部件。为了定义 MySpace 的功能，AstroGrid 制定了一些使用范例，并从这些使用范例中定义了 Phase B 阶段迭代 II 中将要实现的功能。MySpace 基本操作及其流程如下：

1. 存储空间分配
    a. 数据设置代理向服务器请求缓存空间

b. 服务器分配缓存空间
c. 服务器将新分配的缓存空间信息更新到 MySpace 注册
d. 服务器将空间的 URL 返回数据设置代理（DataSetAgent）

2. 保存数据

a. 数据设置代理应用户的要求请求缓存保存数据（DataItem）
b. 服务器保存数据（缓存可能是分布式的，数据有可能保存到其他的服务器上）
c. 服务器更新 MySpace 注册，将缓存的所有权交给用户

3. 删除数据

a. 操作者（Actor）明确要删除的数据
b. 操作者向服务器提出请求删除指定的数据
c. 服务器给要删除的数据加上删除标签
d. 服务器更新注册
e. 服务器删除数据
f. 服务器向操作者返回操作成功信号

4. 查找数据

a. 操作者向注册发送数据查找请求
b. 注册定位数据
c. 注册向操作者返回数据信息

5. 复制数据

a. 操作者明确要复制的数据
b. 操作者向目标服务器发送复制请求
c. 目标服务器对源服务器执行获取数据（GetDataItem）操作
d. 服务器更新注册

6. 迁移数据

a. 操作者明确要迁移的数据
b. 操作者向目标服务器发出迁移请求
c. 目标服务器执行复制操作
d. 源服务器执行删除操作

7. 延长租期

a. 操作者明确要延期的数据
b. 操作者向注册发送延期请求（ExtendLeaseRequest）
c. 如果数据在公用服务器上（CommunityServer），注册延长租期
d. 如果数据在缓存服务器上，注册将数据转移到公用服务器上
e. 注册对自身进行更新

MySpace 服务 (MySpaceService) 和 MySpace 目录服务 (MySpaceDirectoryService) 将作为两个主要的 Grid 服务，实现 MySpace 的所有接口。

8.1.4 MyVO

AstroGrid提出MySpace的设想，不知是巧合，还是他们借鉴了商业网络服

务提供商（ISP）的经验。国际上著名的ISP，比如Yahoo、AOL、Netscape、MicroSoft都提供了类似的个性化服务，分别称为MyYahoo![29]、MyAOL[30]、MyNetscape[31]、MyMSN[32]。

上世纪 90 年代中后期在IT界悄然兴起的Internet个性化服务技术[33]，比如用户建模技术、个性化推荐技术、网站自适应技术、用户隐私保护技术将会对VO的个性化服务提供重要借鉴。

代码	名称	当前状况	预期的发布日期
AgCd07-001	基本 MySpace	进行中	30 Jun 2003
AgCd07-001.1	MySpace 注册	进行中	30 Jun 2003
AgCd07-001.2	MySpace 服务器	进行中	30 Jun 2003
AgCd07-001.3	MySpace 浏览器	进行中	30 Jun 2003
AgCd07-001.4	数据迁移器	进行中	30 Jun 2003
AgCd07-002	带上载功能的多 MySpace 站点	尚未启动	30 Sep 2003
AgCd07-003	多样化的数据存储支持，比如数据库、文件等	尚未启动	31 Dec 2003

表 8.1 AstroGrid 的 MySpace 开发组及其分工

为了明确 VO 门户个性化服务可能需要提供的功能，本人对 MyYahoo!、MyNetscape 和 MyMSN 三个代表性的专业 ISP 服务网站进行了调研。总体印象是 MyYahoo！提供了更丰富的功能和更好的浏览器兼容性，MyNetscape 提供了更好的技术性能，MyMSN 则无论在内容上还是技术上都显得稍逊一筹。下面以 MyYahoo!为例了解这些商业 ISP 提供的个性化服务。

My Yahoo! 是完全免费的，个性化版的 Yahoo!。My Yahoo! 让用户将 Yahoo! 中所有喜欢的部分收集到一个地方。 用户可以选择要看的内容，如新闻、天气、股票价格、体育战况、电视和电影节目表等。 My Yahoo 是完全移动的， 不需要下载任何东西，可以在办公室、家里、甚至世界任何一个可以上网的地方查看同一个 My Yahoo!。

MyYahoo！的个人空间提供了电子邮件、地址簿、日历、记事本等功能。这其中许多功能都是 VO 门户可以借鉴的，特别是记事本和日历，它们在某种程度上与 AstroGrid 的 MySpace 功能相似，可以进行文件的操作和工作进程的安排。

每个享受个性化服务的用户都需要申请一个 Yahoo! ID。Yahoo! 提供的许多功能都可以进行个性化设置。例如：Yahoo! 财经的股票投资组合、Yahoo! 聊

天室、MyYahoo！的个性化新闻和 Yahoo! 电邮均可由用户按自己的需要进行配置。通过 Yahoo! ID 系统可以保存用户的设置、兴趣和首选项；通过相同的 Yahoo! ID，用户可以使用网络上所有这些可自定义的、个性化的服务。仅需注册一次便可以使用该 ID 享受 Yahoo! 中的各项服务，这与 VO 中的单点登录功能相似。

My Yahoo! 的每个方面几乎都可以个性化，包括：

- 要在网页上看到什么样的内容
- 要以何顺序查看这些内容
- My Yahoo! 使用何颜色，并采用何种问候语
- 内容有多少页

MyNetscape 提供了与 My Yahoo 大致相似的功能。两者都实现了不同板块的单一登录功能。但后者提供了更好的界面，比如国际语言的支持。但前者有个很好的功能就是能在浏览器中用鼠标拖放和布置内容模块。这个功能 MyYahoo 和 MyMSN 都没有提供，它们的内容定制只能通过网页底部的菜单实现。My Netscape 不能在微软的 IE 上正常运行，浏览器的兼容性不如 Yahoo 和 MSN。微软 My MSN 也提供了类似的模块定制功能，但定制的内容比较简单，功能也不如上面两个灵活。

此外，Yahoo还提供了一个很方便的协作工具Yahoo! Groups[34]。它通过网站和电子邮件组的形式将家庭、朋友、协会成员等联系在一起。为有相同兴趣和观点的人提供了一种方便的交际渠道。这要比IVOA现在使用的邮件列表更高级、更安全。

Yahoo! Groups 功能特性包括：

- 消息（Messages ）：发送和接收讨论组消息。
- 文件（Files ）：文件上传、下载；组织文件和文件夹；自动分发文件。
- 聊天（Chat ）：同组内的朋友聊天。
- 照片（Photos ）：增加、编辑和删除照片；组织相册；访问控制和共享。
- 书签（Bookmarks ）：为其他网页建立快捷通道。
- 表决（Polls ）：投票和更改自己的投票；创建、修改和总结表决。
- 日历（Calendar ）：日程安排；浏览日历；日程提醒。
- 数据库（Database ）：创建表格、组织信息。
- 成员（Members ）：查看成员信息。
- 讨论组管理（Management）：成员管理、消息管理、环境设置。

上面列举的这几个商业 ISP 个性化服务，其共同的特点就是网站内容的模块化设计。VO 门户也应该采用结构化的设计思路，以便实现个性化，更重要的是可以让第三方组件能方便的融入到门户中。

VO 门户的结构化设计可以带来许多好处。通过定义标准化的 VO 门户模块接口，可以实现全球 VO 门户资源的共享，方便客户服务的开发。将不同的 VO 服务集成为多个模块，用户可以根据自己的兴趣选择服务模块，并以自己喜欢的方式进行布置。这就像我们对自己的新家进行装修一样，室内用怎样的风格装修，买什么家具，怎样布置这些家具，都是自己的选择。

按照同样的游戏规则，我们可以将这样的 VO 门户称为：MyVO。

8.2 用户分类

VO 将拥有大量的用户。他们出于各种目的使用 VO 的服务。对这些用户进行分类不仅是有用的也是必要的。比如，用户分类可以帮助我们更有效的进行资源配置，哪部分服务应该加强，哪部分服务需要减弱或取消。

VO 门户是 VO 系统与用户的直接接口，VO 需要根据不同用户的使用需求和使用习惯来为不同的用户提供不同的用户界面。

根据用户身份和对 VO 使用的目的，VO 用户可以分为三大类：专业用户、非专业用户、特殊用户。

专业用户主要包括天文学家、天文和天体物理专业的研究生、物理学领域与天体物理关系紧密的专业的科研人员等。

非专业用户指除了专业用户和特殊用户以外的其他用户，主要包括大中小学学生和教师、业余天文学家和天文爱好者、科学观/太空馆/天文馆工作人员、新闻记者和普通公众。

除了上面两类用户，VO 还有少量特殊的使用群体，我把他们称为 VO 的特殊用户，主要包括 VO 系统管理员、资源服务提供者和开发人员。

VO 的科学目标是建设数据密集型在线天体物理研究环境。VO 的主要服务对象是专业研究人员，所以专业用户是 VO 用户的主体。前面各章所讨论的各种功能需求都是从专业用户的角度出发的。

非专业用户虽然不是 VO 系统服务的主体，但在数量上很可能会超过专业用户。同时，VO 的社会效益主要体现在非专业用户的使用上。所以，VO 系统必须重视非专业用户的需求，为他们提供良好的服务。

非专业用户成分复杂，需求各异。在 VO 为这些用户开发服务的时候必须

与他们紧密合作或者请他们参与到 VO 服务的建设中来。

对于管理用户，VO 系统需要提供友好的管理界面，最好是图形化的管理界面，让管理员能方便快捷的管理和配置 VO 系统中的资源与服务，完成资源管理、服务管理、用户管理、系统检测、统计汇总等管理工作。

对于服务提供者和开发人员，VO 系统需要提供标准的、良好的开发接口和齐全的开发文档。让开发人员能方便的进行 VO 服务的开发和发布，让资源服务提供者能以标准、方便的方式共享自己的资源。VO 资源的访问和使用与向 VO 系统发布资源和服务有很大的差异，需要通过不同的机制和途径。

门户用户和 VO 服务访问控制部分所涉及的用户和授权是两种不同的概念。VO 资源和服务的安全访问用户通常情况下是 VO 服务和程序。访问指的是客户端 VO 服务对服务器端 VO 服务的调用。门户用户是直接通过 VO 门户浏览、使用 VO 服务的用户，是人，他们接触的主要是 VO 服务中的客户端服务。

8.3 教育与普及

由于大量真正的科学数据能够无偿地从互联网上获得，而且公众对天文学有着浓厚的兴趣，因此虚拟天文台将特别适合教育和科学普及，是一个理想的科学教育平台。虚拟天文台与互联网的紧密结合使得它能够在前所未有的社会和地理范围内提供各式各样的高质量的科普和教育服务。

VO 在教育与普及方面的目标是把 VO 建成天文学公众和教育门户。针对不同层次的用户开发定制合适的界面和功能。既要做到好用，又不让界面阻碍高级用户能力的发挥，给用户充分的自由空间。

在 VO 的教育服务开发建设过程中要采用广泛联合的方式，吸收社会各界的力量和优势，这也是满足不同用户需求的需要。VO 的教育和普及工作不仅仅需要提供一些服务，更需要培养与各个行业用户的合作。这其中业余天文学家加盟 VO 的潜力是巨大的。他们对天文充满热情，同时也是 VO 非专业用户中的最忠实用户。在某些领域，比如彗星、变星，他们有自己的优势，是专业天文科研人员无法替代的。

VO 需要针对非专业用户提供一系列的服务基础设施。按照开发过程中优先权和功能层次的不同，这些服务可以分为基础服务、优先服务和其他服务。基本服务是其他服务开发的基础。

基础服务主要包括以下这些方面[35]：

- 教育普及资源相关元数据，比如描述天文数据访问方面的元数据、单

个天体相关的元数据、教育资源相关的元数据等。

- 非专业查询协议，让各种相关服务以统一的方式查询 VO。VO 的非专业入口可以多种多样，比如可以通过桌面天文馆软件访问。
- 最受欢迎天体索引，让非专业用户更容易的找到感兴趣的目标。广大非专业用户感兴趣的天体只是天文数据中的很小一块。为了不至于让他们在信息的海洋中失去兴趣，VO 需要这样做以保护他们的好奇心。像 HubbleSite 这样的网站都保存有用户的点击信息。这些点击信息对统计网络用户的兴趣点是非常有用的。
- 最受欢迎天文数据索引，这可以加快在 VO 系统中的数据检索速度。通常非专业用户并不了解怎样的专业数据是他们最感兴趣的。
- 教育资源目录，方便定位与 VO 相关的教育资源。

优先提供的服务主要包括：

- 面向公众的天文图像的坐标元数据，方便多波段图像的比对处理和将面向公众的图像整合到巡天图像中。多波段数据的整合是 VO 的科学核心。面向公众的图像的坐标信息将使得图像到 VO 数据资源的融合变得相对容易。
- 业余天体摄影数据库，允许业余天文学界将自己的作品融入到 VO 中。当今的天文爱好者创作了许多高质量的图像，这些大都很符合大众的口味。这样做会大大激发业余天文界的热情。
- 面向公众的天文图像拼接服务，生成大视场的图像以满足数字天文馆的需要。
- 实时天文数据的访问，让学生能访问到最新的天文数据，甚至可以让学生通过 VO 遥控望远镜的观测，激发他们的热情。
- 互联网天空数据库，业余天文学家非常希望能接触到真实的天文数据。现有的桌面天文馆软件已经提供了关于许多天体的丰富的信息。VO 的这个功能将大大增强公众对天体的探索能力。
- 天文及时消息服务，将最近观测到的天文事件通知给感兴趣的组织，比如新闻部门、天文馆、业余天文学家等。
- 太空科学艺术资源目录，让用户方便的找到与天文学和太空科学相关的艺术资源，比如太空画等。

8.4 本地化和国际化

VO 实现了全球主要天文资源和服务无缝透明的融合，为天文研究和教育普及提供了极大的方便。但我们必须认识到这其中还有一个阻碍中国用户充分使用 VO 资源的一个因素，这就是语言。

语言是通过互联网实现全球性沟通的最大障碍。对于整个世界来说，英语是网络上的常用语言。能够用母语交流是我们感到最惬意的事，非英语用户很少会把浏览英文的网站当成一件乐事，除非是在进行外语学习。

并非只有英文水平较低的用户才喜欢使用中文界面。绝大多数人还是使用自己的母语进行交流最为自如。即使外语水平很高的用户，使用外语进行工作的效率也往往比用母语低。同时，用自己的母语工作、学习和娱乐也是用户的一项权利。

由于我国是一个非英语国家，按照语言习惯，中国虚拟天文台的用户可分为英文用户和中文用户两大群体。一方面，为了融入 IVO 大家庭，为国际用户提供服务，China-VO 必须把现有的中文文献资料、观测数据转化为英文并提供相应的英文用户界面，即“国际化”。另一方面，为了防止因为语言障碍而把国内广大用户挡在门外的现象出现，China-VO 必须承担起把丰富的英文资源，特别是教育和普及资源，转化为中文的使命并同时提供相应的中文用户界面，即“本地化”。

本地化（Localization 或简写为 L10N）：又称为本土化，是指将某一事物转换成符合本地特定要求的过程。

资源本地化的最佳效果是既能适应本地要求，又尽可能地保持资源原有的特定情境含义。人们常说的“汉化”其实所指的是“中国本地化”，即从其他语言文化系统到汉语文化系统的转换过程；而“翻译”则是专指“文法层面的本地化”，即只从字符、格式等文法层面上对资源进行的转换改制。翻译和本地化是不同的概念。翻译是本地化的子集，主要指把原文字从一种语言转换到另一种语言。然而，当文字被翻译后，必然要对资源和服务相应地进行许多其它更改。这些更改包括技术上和文化上的更改。本地化包括但不限于以下方面：翻译、文化本地化、图形处理、编译、测试，等等。

国际化（Internationalization 或 I18N）：也常称为全球化（Globalization 或 G11N），指将一个事物转变成超越所有本地局部特征的过程。与本地化相对，国际化在本质上是一种消除或隐去个性要素的过程。

VO 是国际化的系统。VO 的开发过程应遵循国际化的标准。遵循国际化标准，可以更高效地开发和调试系统，降低系统的开发费用，使用户更方便地使用。在 VO 中，可以通过采用国际化的开发语言、数据编码方式、数据库平台等途径实现国际化。

8.4.1 技术方案

目前国际化的主要途径是采用符合国际化标准的开发工具，提供对国际语

言的支持。这里所说的国际化标准是国际化标准组织或一些相关组织制定的一些标准。国际化标准涉及到字符集、编码、字体处理、打印、文本绘制、用户界面、语言输入方法、数据交换、文化习俗等方方面面。几个有代表性的标准化组织有：

- ANSI（美国国家标准化组织）[36]
- ISO（国际标准化组织）[37]
- Unicode组织[38]
- IEEE（美国电气电子工程师协会）[39]

其中，ANSI/ISO 制定了使用 C 编程语言编写国际化软件的通用接口。ISO 制定了字符集标准和其它影响 locale 名字的标准。 IEEE 提供了一些国际化的通用库函数和设置管理不同 locale 的用户命令。Unicode 提供了国际化统一字符编码标准。

Locale 是 ANSI C 语言中最基本的支持国际化的标志。Locale 是软件在运行时的语言环境，它包括语言(Language)，地域 (Territory)和字符集(Codeset)。其格式为：语言[_地域[.字符集]]。 如对中文 GB18030 字符集，locale 的格式是：zh_CN.GB18030。

目前所使用的 Unicode 是一种 16 位字宽的字符编码, 它由非赢利的计算机组织 Unicode 协会维护和改进。Unicode 协会是一个非盈利的组织，是为发展，扩展和推广使用 Unicode 标准而建立的。

Unicode 给每个字符提供了一个唯一的数字，不论是什么平台，不论是什么程序，不论什么语言。Unicode 包含了当今计算机领域中广泛使用的所有字符，如世界上大部分的书面语言、印刷字符、数字和技术符号、地理图形和标点符号。由于 Unicode 的一致性，它在大多数情况下都能简化软件的国际化过程。Unicode 标准的出现和其支持工具的迅速增多，是近来全球软件技术重要的发展趋势。

在 VO 各种服务组件的开发过程中都采用国际化的工具便可以保证 VO 服务对国际化的支持。可喜的是，越来越多的软件和工具都已经符合国际化标准，支持 Unicode。比如我们在开发中很可能用到的开发环境、字体、数据库等。

支持国际化的高层库：

- OSF/Motif
- Qt/kdelib
- gtk+/gnome-lib

- Perl
- Java
- TCL
- GAWK
- MicroSoft Visual Studio

字体：

- FreeType
- TrueType and Open
- UniMath （TeX, LaTeX）

数据库产品：

- IBM DB2
- Microsoft SQL Server
- Oracle
- Sybase
- PostgreSQL (PgSQL)
- MySQL

特别需要指出的是 VO 的主要开发语言 Java 和 VO 数据主要编码格式 XML 都提供了对国际化标准和 Unicode 的良好支持。这为 VO 国际化的实现提供了极好的前提条件。

8.4.2 主要工作内容

作为 IVO 在中国的部分，China-VO 有义务将 IVO 丰富的资源与服务提供给国内用户，不管是专业用户还是非专业用户。国内的专业用户主要来自中国科学院的几个天文台，国内几个大学的天文系，以及物理、地球等学科的科研人员。非专业用户则人数众多，背景各异。虽然 China-VO 的主要服务对象是专业用户，但也必须向非专业用户提供适当的服务，特别是基础的服务架构。

对于专业用户，他们一般都拥有很高的英文水平，可以直接利用 IVO 的国际资源。China-VO 的本地化工作主要是为了满足国内非专业用户的需要。China-VO 将与非专业用户群体开展广泛的合作，让非专业用户，特别是有较高文化水平的业余天文学家，参加到项目的本地化工作中。本地化工作范畴主要包括：

- 资料翻译：包括文法翻译、字符集转换、格式编码调整等。翻译是本地化工作的基础，作用在于使资源既能保持原有情境含义，又能够很好地支持本地语言；
- 界面重置：对信息资源中图形图像、菜单、对话框等界面要素加以必

要的改动，以符合本地习惯；

- 语义转换：针对各种本地语义敏感因素（如文化、语言、宗教、地域、政治、历史、语境等）进行资源改制，使其符合本地特性；
- 开发本地功能：根据本地情境的实际需求，对资源进行本地化应用，即开展本地化二次开发；
- 支持本地技术要求：支持或兼容本地资源环境的软、硬件等技术现状，以保证信息化资源的正常技术运行；
- 本地测试：根据本地具体要求，参照对上述资源本地化工作的效果进行测试修订。

在 China-VO 的本地化开发中，需要开发人员有良好的外语，特别是英语水平和良好的汉语水平；同时，需要良好的专业背景知识，对天文学有兴趣，热爱自己的工作。

此外，在 China-VO 建设的过程中，我们需要将本国的天文资源共享给国际社会，特别是我们非常珍贵的历史观测资料，比如古天象记录，古星图等。这就涉及到国际化的问题。其中一项重要的工作便是英文化，把古天象记录、古星图的内容翻译为英文。中国科学院北京天文台曾根据几千种地方志和其他史籍，整理出太阳黑子、极光、陨石、日月食、超新星、彗星、流星及有关天文学的人物、著作、学说、机构、仪器等的记录一共几百万字，编成《中国古代天象记录总集》和《中国天文史料汇编》。中国古代天文文献是世界天文学的一笔重要财富。这些珍贵的资料需要与国际同行共享，以得到更好的保存和利用。

参考文献

[1] [RedBook] The AstroGrid Phase A Report. http://wiki.astrogrid.org/pub/Astrogrid/PhaseAReport/redbook.pdf
[2] [Astrobrowse] HEASARC Astrobrowse service. http://heasarc.gsfc.nasa.gov/ab/
[3] [Browse] HEASARC Browse. http://heasarc.gsfc.nasa.gov/W3Browse/
[4] [CURSA] StarLink CURSA. http://www.roe.ac.uk/acdwww/cursa/home.html
[5] [ISAIA] Interoperable Systems for Archival Information Access. http://heasarc.gsfc.nasa.gov/isaia/
[6] [MAST] Multimission Archive at STScI. http://archive.stsci.edu/mast.html
[7] [NED] NASA/IPAC Extragalactic Database. http://nedwww.ipac.caltech.edu/
[8] [Querator] Querator Query Builder . http://archive.eso.org/querator/
[9] [Simbad] CDS Simbad. http://simbad.u-strasbg.fr/sim-fid.pl
[10] [VizieR] CDS VizieR. http://vizier.u-strasbg.fr/viz-bin/VizieR

[11] [Aladin] CDS Aladin. http://aladin.u-strasbg.fr/aladin.gml
[12] [SkyWiew] SkyView virtual observatory. http://skyview.gsfc.nasa.gov/
[13] [Skymorph] SkyMorph GSFC. http://skyview.gsfc.nasa.gov/skymorph/skymorph.html
[14] [SkyCat] ESO SkyCat. http://cadcwww.dao.nrc.ca/skycat/skycat.html
[15] [GAIA] Graphical Astronomy and Image Analysis Tool. http://star-www.dur.ac.uk/~pdraper/gaia/gaia.html
[16] [JSky] Java Components for Astronomy. http://archive.eso.org/JSky/
[17] [Starcast] STScI Starcast. http://archive.stsci.edu/starcast/
[18] [StarView] STScI StarView. http://starview.stsci.edu/
[19] Clive Page. Virtual Observatory Prototypes. http://wiki.astrogrid.org/bin/view/Astrogrid/RbProtoVOSurveyReport
[20] [SkyServer] Sloan Digital Sky Survey SkyServer. http://skyserver.sdss.org/en/
[21] [VirtualSky] Virtual Sky. http://virtualsky.org/
[22] SkyServer mirror site at China-VO. http://skyserver.china-vo.org
[23] [Cocoon] Apache Cocoon. http://cocoon.apache.org/
[24] [phpBB] phpBB bulletin board package. http://www.phpbb.com/
[25] [TWiKi] http://twiki.org
[26] [JetSpeed] Apache Jetspeed. http://jakarta.apache.org/jetspeed/
[27] [Velocity] Apache Velocity. http://jakarta.apache.org/velocity/
[28] [Turbine] Apache Turbine. http://jakarta.apache.org/turbine/
[29] [MyYahoo] My Yahoo!, http://my.yahoo.com/
[30] [MyAOL]My AOL, http://my.aol.com/
[31] [MyNetscape]My Netscape, http://my.netscape.com/
[32] [MyMSN]My MSN, http://my.msn.com/
[33] 应晓敏, 窦文华. 技术架构—Internet 个性化服务的关键技术. 计算机世界, 2003 (22): B10~B11
[34] Yahoo! Groups. http://groups.yahoo.com
[35] Mark Voit, et al. ENABLING Outreach With NVO. http://bill.cacr.caltech.edu/cfdocs/usvo-pubs/files/NVO_EPO_Recs.pdf
[36] [ANSI] American National Standards Institute. http://www.ansi.org/
[37] [ISO] International Organization for Standardization. http://www.iso.ch/
[38] [Unicode] Unicode consortium. http://www.unicode.org/
[39] [IEEE] Institute of Electrical and Electronics Engineers. http://www.ieee.org/

第九章 VO-enabled LAMOST

VO-enable LAMOST 是 China-VO 的主要特色，也是项目的重要使命。VO-enabled LAMOST 的实现分两步完成：VO-enabled LAMOST Data 和 VO-enabled LAMOST Telescope。在 China-VO 的开始阶段主要是要实现第一阶段的目标，将 LAMOST 的数据产品和工作星表整合到 VO 中。

9.1 VO 使能的必要条件

在论文的前面部分已经讨论过，VO-enabled LAMOST 包括两层含义：“VO-enabled LAMOST Dataset”和“VO-enabled LAMOST Telescope”。其中第二层含义将把整个观测系统纳入到 VO 环境中，这涉及到观测仪器的 VO 化问题。目前这方面的技术和标准都很不成熟，将作为 VO 的长远发展目标。本章中仅对第一层含义进行讨论，也就是如何将 LAMOST 的数据融入到 VO 环境中。

LAMOST的数据集按照取得方式不同分为两种：工作数据和数据产品[1]。工作数据是项目建设者为了巡天观测的需要从已有的数据中提取出来的，包括目标星表和导星星表。数据产品是项目的观测成果，包括二维光谱数据、一维光谱数据和巡天星表。其中二维光谱数据不等于望远镜的原始观测数据，而是利用图像处理软件进行质量分析、宇宙线去除、CCD改正等处理后得到的结果。一维光谱则是图像处理软件在进行了光谱抽取、波长定标、减天光、流量定标、红蓝两段光谱合并等操作后给出的关于每个观测天体完整的一维光谱。

要实现VO-enabled LAMOST，需要在两方面与VO接轨。一方面，LAMOST能够接受利用VO系统筛选的工作星表，包括导星星表和目标星表。VO环境中数据的处理结果输出格式包括多种，比如普通的文本文件、HTML格式、FITS格式，但默认的格式很可能是VOTable格式[2]的XML数据。这要求LAMOST工作星表制定软件系统能够对VOTable格式的数据进行解析和处理，直接利用或者转换为自己的工作数据格式。另一方面，LAMOST的数据产品要能共享到VO环境中，成为VO资源的一部分。要使LAMOST作为一个数据服务提供者，将自己的数据产品共享到VO环境中，最关键的工作有两项。其一，按照VO注册的标准对数据产品和相应的访问服务进行注册，以便让VO用户能利用VO资源与服务发现机制发现这些资源进而访问、利用这些资源。其二，按照VO的数据标准对数据集进行封装并按照数据访问标准提供访问服务。

目前，VO的注册标准、数据标准、数据访问标准都不成熟。相关标准草案都处于讨论和不断修改之中。本章第二节给出的注册元数据模型主要参考了Robert Hanisch给出的“VO资源和服务元数据” 0.7 版[3]以及Ray Plante对它做的修正[4]、“都柏林核心元数据”1.1 版[5]，并结合LAMOST数据的实际情况进行了修改。

本章第三节给出的LAMOST数据模型采用了VOTable的封装格式。对VOTable标准进行了扩展，比如对“INFO”元素增加了“ucd”属性，使得可以在VOTable文件的“INFO”部分提供一些基本的元数据信息。在数据模型的定义过程中，还对UCD[6]进行了扩展。UCD是目前很有希望在VO注册和互操作性方面发挥重要作用的工具和标准。但UCD还不能完全满足VO的需要，其分级结构需要进行一定的调整，条目设置需要进行一定的增减。由于LAMOST数据有其特殊性，一些条目在UCD中没有对应者，所以在设计数据模型时按照UCD的命名规则进行了适当的扩充。

9.2 资源注册元数据模型

每个注册记录由四部分组成，分别是标识元数据、履历元数据、内容元数据和服务元数据。

标识元数据对资源进行命名并给出一个标识符，这部分包括：标题（Title）、标记（Ticker）和标识（Identifier）。

履历元数据描述了资源维护者相关的信息，主要包括：发布者（Publisher）、创建者（Creator）、贡献者（Contributor）、参考链接（ReferencURL）、联系信息（Contact）等内容。

内容元数据对资源的内容、数据属性进行描述，主要包括：日期（Date）、版本（Version）、数据类型（Type）、数据格式（Format）、天区覆盖（Sky Coverage）、波段覆盖（Spectral Coverage）、时间覆盖（Temporal Coverage）、内容等级（Content Level）、访问权限（Right）等。

服务元数据对数据访问服务进行描述，包括接口元数据和功能元数据两方面。

注册元数据模型如图 9.1 所示，图 9.2 是根据这个元数据模型给出的一个模拟 LAMOST 巡天星表的注册元数据。

9.3 LAMOST 数据模型

按照数据的内容和存储方式不同，数据模型分为两类：

- 以“TABLEDATA”形式提供的表列数据。这包括巡天星表、导星星表和目标星表。
- 以“BINARY”或者“FITS”格式存储的流数据。这包括二维光谱和一维光谱。

对于一维光谱，将以二进制流的形式存储。虽然按照 VOTable 标准，每个 VOTable 文件可以存储多个“TABLE”对象，也就是说可以将大量信息都存储在同一文件中，但为了使用方便，每个文件只存储一条一维光谱。每个 VOTable 文件包括两个“TABLE”元素。其中第一个“TABLE”元素以“TABLEDATA”的形式存储光谱的相关背景信息，第二个“TABLE”元素存储光谱的本体，数据结构如图 9.3 所示。

LAMOST 的光谱观测在光谱仪处分为红蓝两端。考虑到相同天体的数据保存在同一文件中将有助于数据的处理和检索，在进行二维光谱图像数据模型设计时将同一光谱仪红蓝两端的观测数据用一个文件保存。一个 VOTable 文件包括三个“TABLE”元素，其中第一个以“TABLEDATA”格式描述二维光谱背景信息，第二和第三个“TABLE”元素分别存储红蓝两端光谱图像。数据模型如图 9.4 所示。

巡天星表、导星星表和目标星表都使用“TABLEDATA”格式存储。三者主要的区别是星表的列不同，比如导星星表和目标星表中坐标信息是最重要的信息，而巡天星表中则需要包括更多的光谱处理得出的物理参量。三个星表的结构分别如图 9.5、9.6、9.7 所示。

XML Schema 格式的注册元数据模型文件和 VOTable 格式的 LAMOST 数据模型文件可从下面网址下载：

http://www.china-vo.org/lamost/

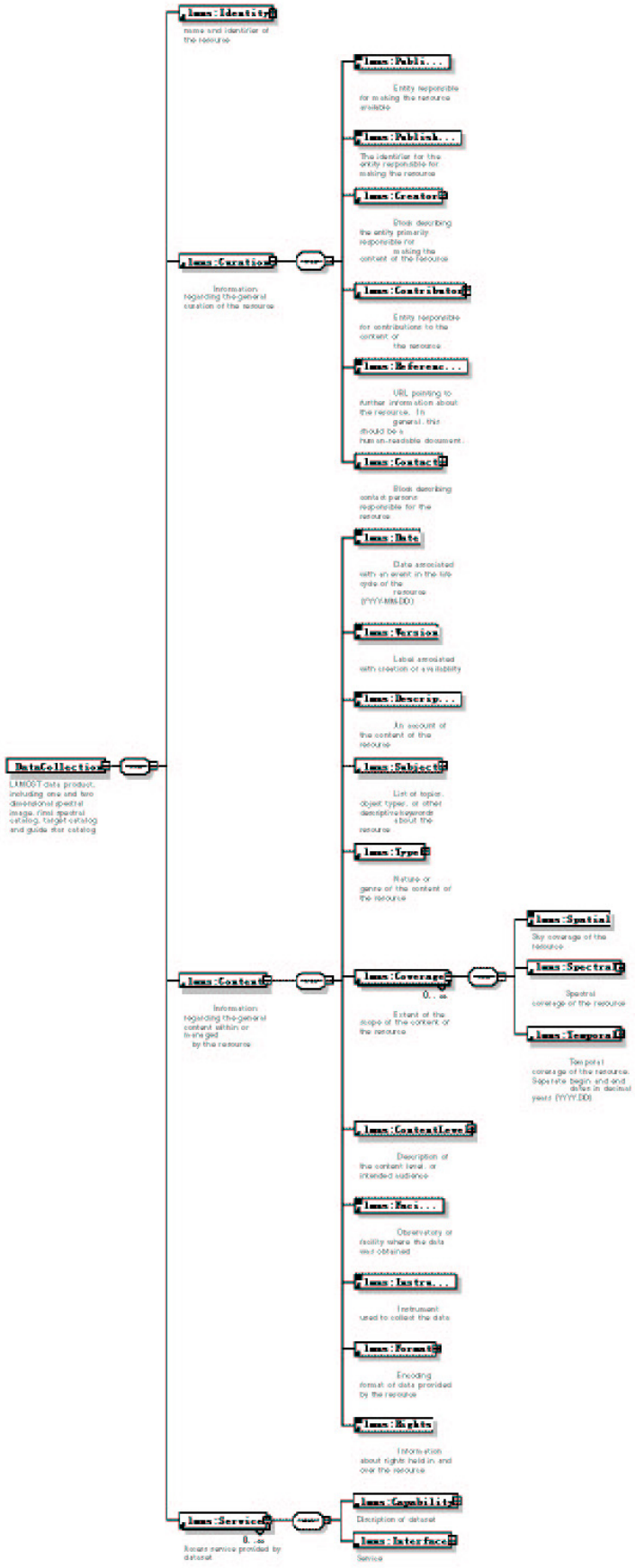
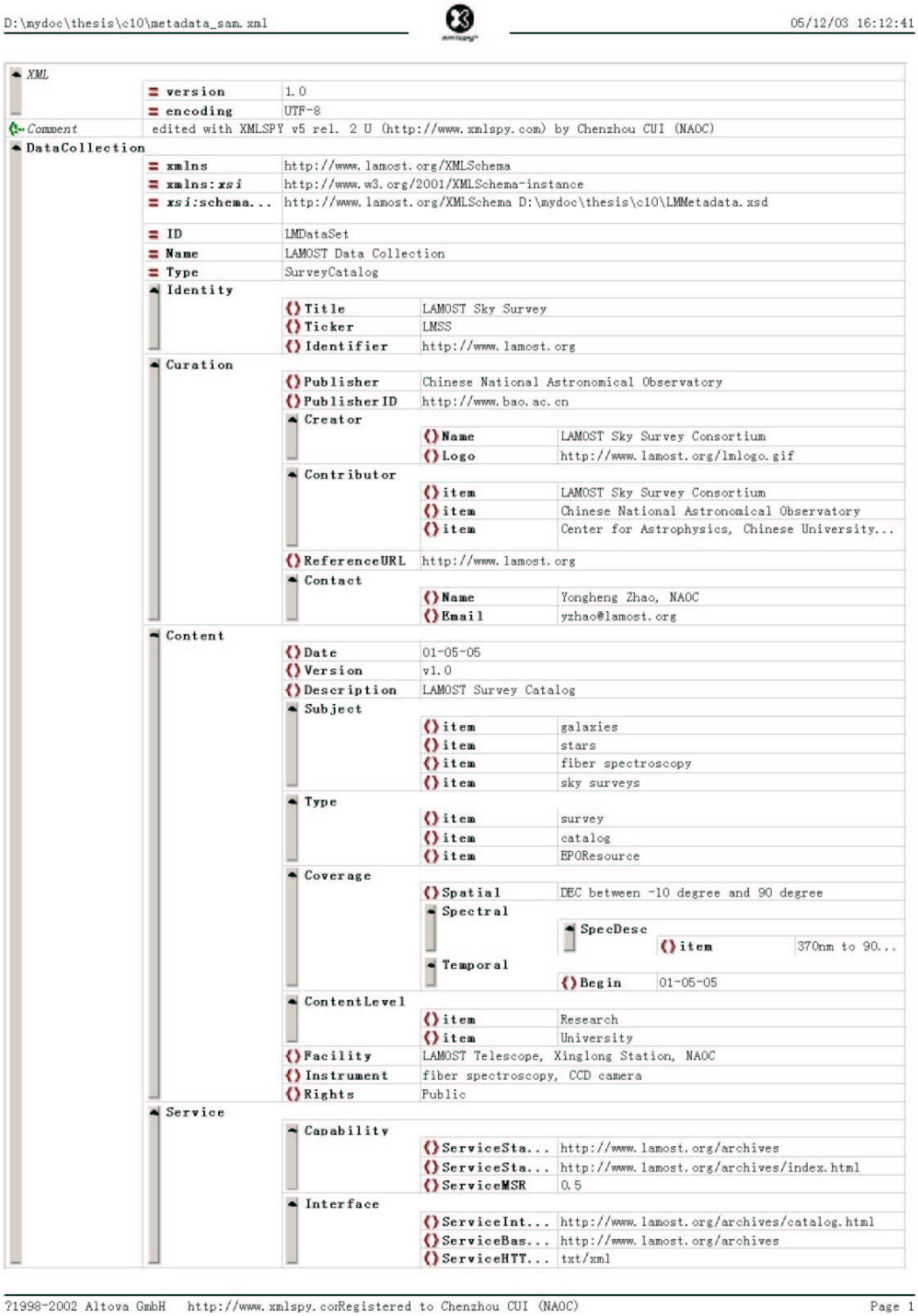


图 9.1 LAMOST 资源注册元数据模型



D:\mydoc\thesis\c10\metadata_sam.xml 05/12/03 16:12:41

XML		
	version	1.0
	encoding	UTF-8
Comment	edited with XMLSPY v5 rel. 2 U (http://www.xmlspy.com) by Chenzhou CUI (NAOC)	
DataCollection		
	xmlns	http://www.lamost.org/XMLSchema
	xmlns:xsi	http://www.w3.org/2001/XMLSchema-instance
	xsi:schema...	http://www.lamost.org/XMLSchema D:\mydoc\thesis\c10\LMMetadata.xsd
	ID	LMDataSet
	Name	LAMOST Data Collection
	Type	SurveyCatalog
	Identity	
	Title	LAMOST Sky Survey
	Ticker	LMSS
	Identifier	http://www.lamost.org
	Curation	
	Publisher	Chinese National Astronomical Observatory
	PublisherID	http://www.bao.ac.cn
	Creator	
	Name	LAMOST Sky Survey Consortium
	Logo	http://www.lamost.org/lmlogo.gif
	Contributor	
	item	LAMOST Sky Survey Consortium
	item	Chinese National Astronomical Observatory
	item	Center for Astrophysics, Chinese University...
	ReferenceURL	http://www.lamost.org
	Contact	
	Name	Yongheng Zhao, NAOC
	Email	yzhao@lamost.org
	Content	
	Date	01-05-05
	Version	v1.0
	Description	LAMOST Survey Catalog
	Subject	
	item	galaxies
	item	stars
	item	fiber spectroscopy
	item	sky surveys
	Type	
	item	survey
	item	catalog
	item	EPOResource
	Coverage	
	Spatial	DEC between -10 degree and 90 degree
	Spectral	
	SpecDesc	
	item	370nm to 90...
	Temporal	
	Begin	01-05-05
	ContentLevel	
	item	Research
	item	University
	Facility	LAMOST Telescope, Xinglong Station, NAOC
	Instrument	fiber spectroscopy, CCD camera
	Rights	Public
	Service	
	Capability	
	ServiceSta...	http://www.lamost.org/archives
	ServiceSta...	http://www.lamost.org/archives/index.html
	ServiceMSR	0.5
	Interface	
	ServiceInt...	http://www.lamost.org/archives/catalog.html
	ServiceBas...	http://www.lamost.org/archives
	ServiceHTT...	txt/xml

?1998-2002 Altova GmbH http://www.xmlspy.comRegistered to Chenzhou CUI (NAOC) Page 1

图 9.2 注册元数据示例

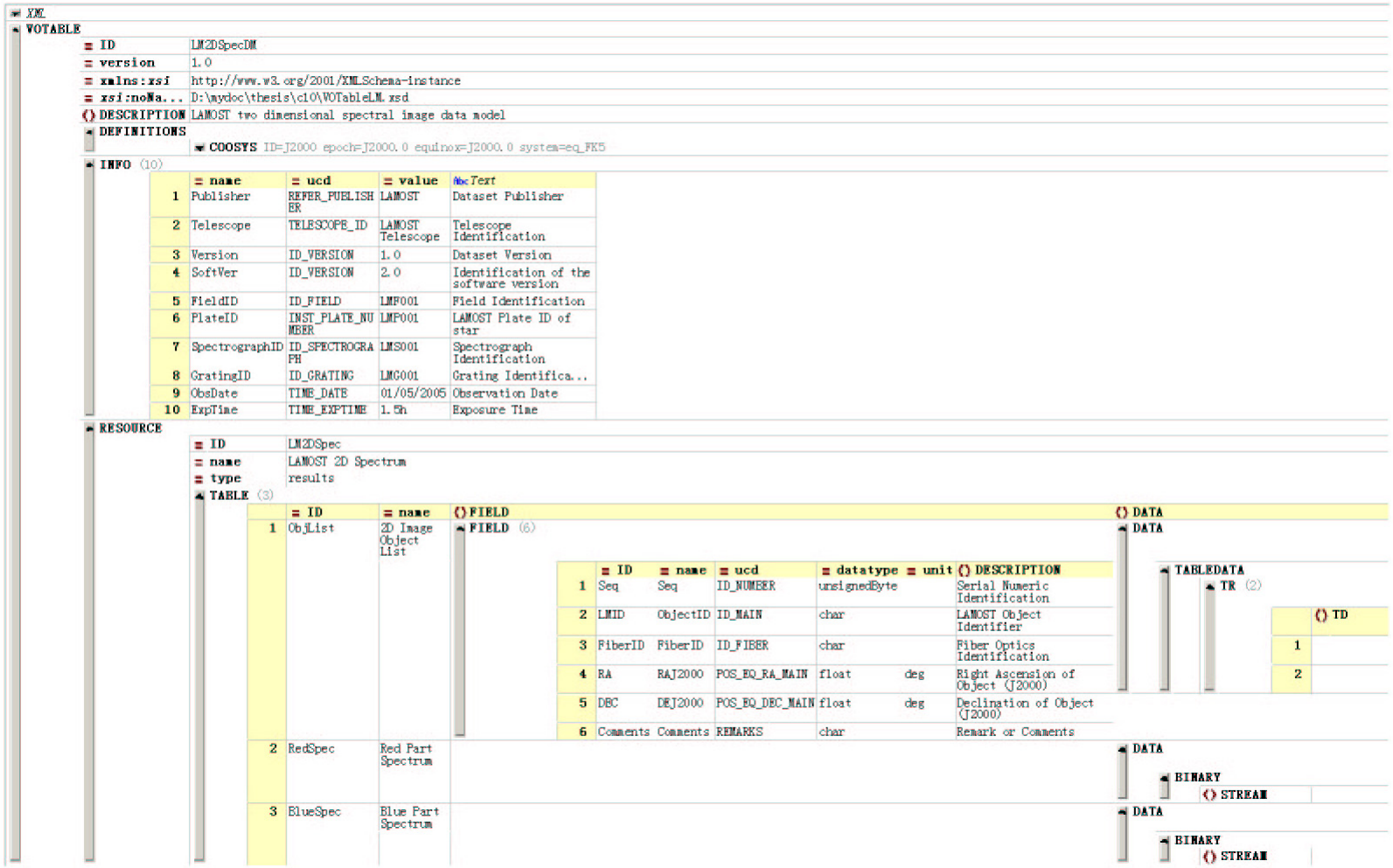


图 9.3 LAMOST 二维光谱数据模型

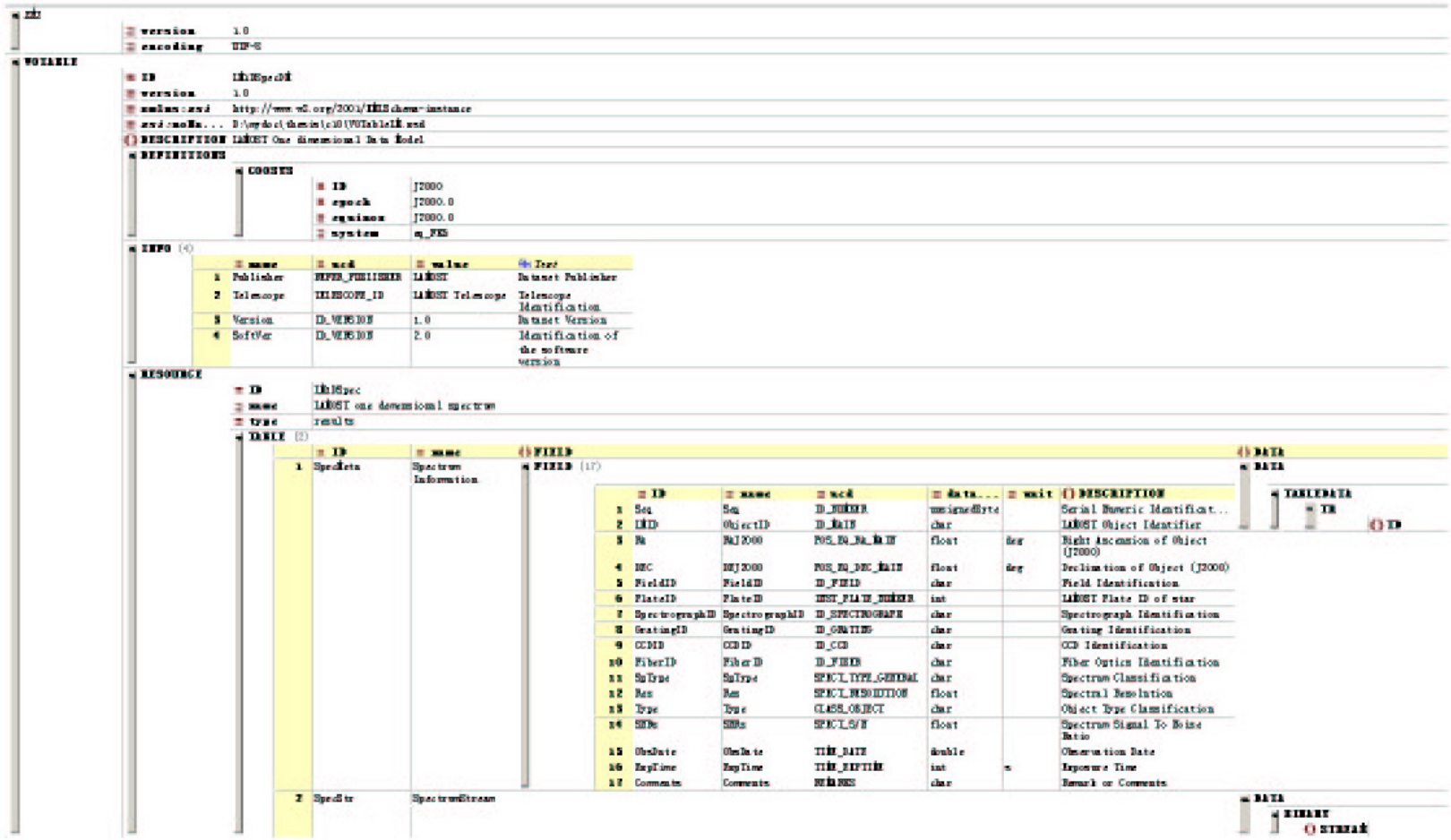
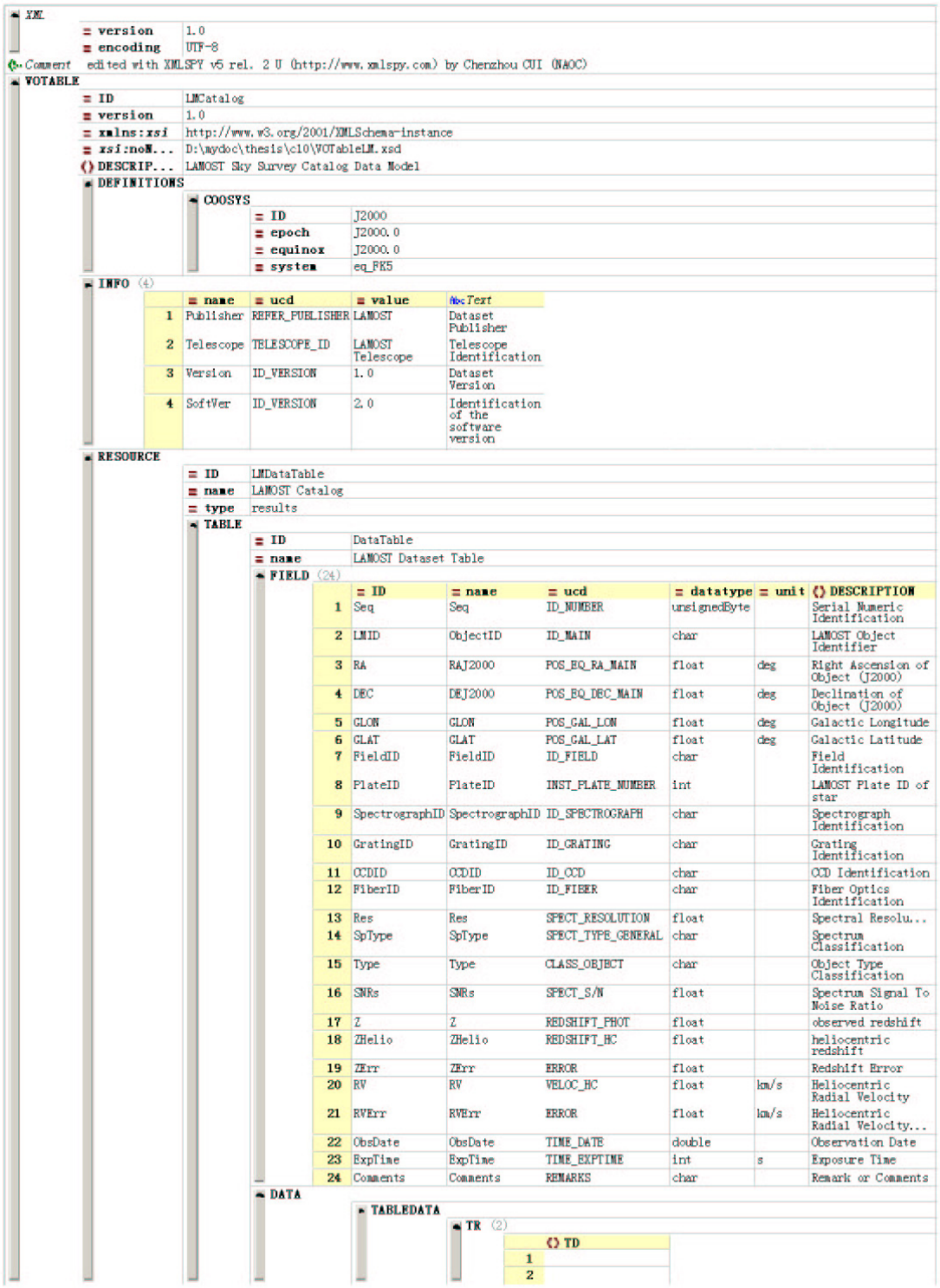


图 9.4 LAMOST 一维光谱数据模型



XML
- version: 1.0
- encoding: UTF-8

Comment: edited with XMLSPY v5 rel. 2 U (http://www.xmlspy.com) by Chenzhou CUI (NAOC)

VOTABLE
- ID: LMCatalog
- version: 1.0
- xmlns:xsi: http://www.w3.org/2001/XMLSchema-instance
- xsi:noN...: D:\mydoc\thesis\c10\VOTableLM.xsd
- DESCRIP...: LAMOST Sky Survey Catalog Data Model

DEFINITIONS
- COOSYS
  - ID: J2000
  - epoch: J2000.0
  - equinox: J2000.0
  - system: eq_FK5

INFO (4)

	name	ucd	value	Text
1	Publisher	REFER_PUBLISHER	LAMOST	Dataset Publisher
2	Telescope	TELESCOPE_ID	LAMOST Telescope	Telescope Identification
3	Version	ID_VERSION	1.0	Dataset Version
4	SoftVer	ID_VERSION	2.0	Identification of the software version

RESOURCE
- ID: LMDataTable
- name: LAMOST Catalog
- type: results

TABLE
- ID: DataTable
- name: LAMOST Dataset Table

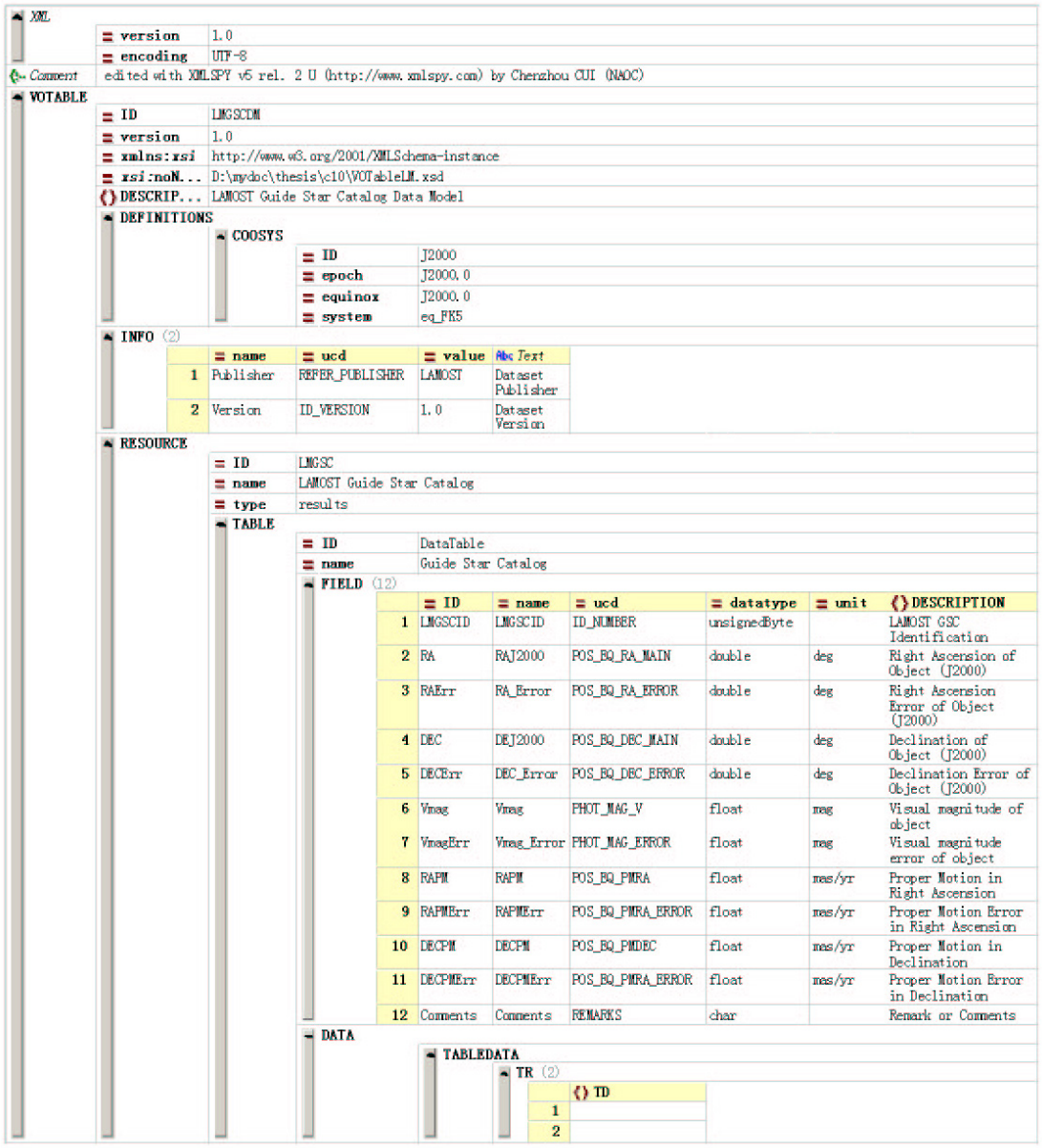
FIELD (24)

	ID	name	ucd	datatype	unit	DESCRIPTION
1	Seq	Seq	ID_NUMBER	unsignedByte		Serial Numeric Identification
2	LMID	ObjectID	ID_MAIN	char		LAMOST Object Identifier
3	RA	RAJ2000	POS_EQ_RA_MAIN	float	deg	Right Ascension of Object (J2000)
4	DEC	DEJ2000	POS_EQ_DEC_MAIN	float	deg	Declination of Object (J2000)
5	GLON	GLON	POS_GAL_LON	float	deg	Galactic Longitude
6	GLAT	GLAT	POS_GAL_LAT	float	deg	Galactic Latitude
7	FieldID	FieldID	ID_FIELD	char		Field Identification
8	PlateID	PlateID	INST_PLATE_NUMBER	int		LAMOST Plate ID of star
9	SpectrographID	SpectrographID	ID_SPECTROGRAPH	char		Spectrograph Identification
10	GratingID	GratingID	ID_GRATING	char		Grating Identification
11	CCDID	CCDID	ID_CCD	char		CCD Identification
12	FiberID	FiberID	ID_FIBER	char		Fiber Optics Identification
13	Res	Res	SPECT_RESOLUTION	float		Spectral Resolu...
14	SpType	SpType	SPECT_TYPE_GENERAL	char		Spectrum Classification
15	Type	Type	CLASS_OBJECT	char		Object Type Classification
16	SNRs	SNRs	SPECT_S/N	float		Spectrum Signal To Noise Ratio
17	Z	Z	REDSHIFT_PHOT	float		observed redshift
18	ZHelio	ZHelio	REDSHIFT_HC	float		heliocentric redshift
19	ZErr	ZErr	ERROR	float		Redshift Error
20	RV	RV	VELOC_HC	float	km/s	Heliocentric Radial Velocity
21	RVErr	RVErr	ERROR	float	km/s	Heliocentric Radial Velocity...
22	ObsDate	ObsDate	TIME_DATE	double		Observation Date
23	ExpTime	ExpTime	TIME_EXPTIME	int	s	Exposure Time
24	Comments	Comments	REMARKS	char		Remark or Comments

DATA
- TABLEDATA
  - TR (2)

	TD
1	
2	

图 9.5 LAMOST 巡天星表数据模型



XML
- version: 1.0
- encoding: UTF-8

Comment: edited with XMLSPY v5 rel. 2 U (http://www.xmlspy.com) by Chenzhou CUI (NAOC)

VOTABLE
- ID: LMGSCDM
- version: 1.0
- xmlns:xsi: http://www.w3.org/2001/XMLSchema-instance
- xsi:noN...: D:\mydoc\thesis\c10\VOTableLM.xsd
- DESCRIP...: LAMOST Guide Star Catalog Data Model

DEFINITIONS
- COOSYS
  - ID: J2000
  - epoch: J2000.0
  - equinox: J2000.0
  - system: eq_FK5

INFO (2)

	name	ucd	value	Text
1	Publisher	REFER_PUBLISHER	LAMOST	Dataset Publisher
2	Version	ID_VERSION	1.0	Dataset Version

RESOURCE
- ID: LMGSC
- name: LAMOST Guide Star Catalog
- type: results
- TABLE
  - ID: DataTable
  - name: Guide Star Catalog
  - FIELD (12)

	ID	name	ucd	datatype	unit	DESCRIPTION
1	LMGSCID	LMGSCID	ID_NUMBER	unsignedByte		LAMOST GSC Identification
2	RA	RAJ2000	POS_EQ_RA_MAIN	double	deg	Right Ascension of Object (J2000)
3	RAErr	RA_Error	POS_EQ_RA_ERROR	double	deg	Right Ascension Error of Object (J2000)
4	DEC	DEJ2000	POS_EQ_DEC_MAIN	double	deg	Declination of Object (J2000)
5	DECErr	DEC_Error	POS_EQ_DEC_ERROR	double	deg	Declination Error of Object (J2000)
6	Vmag	Vmag	PHOT_MAG_V	float	mag	Visual magnitude of object
7	VmagErr	Vmag_Error	PHOT_MAG_ERROR	float	mag	Visual magnitude error of object
8	RAPM	RAPM	POS_EQ_PMRA	float	mas/yr	Proper Motion in Right Ascension
9	RAPMErr	RAPMErr	POS_EQ_PMRA_ERROR	float	mas/yr	Proper Motion Error in Right Ascension
10	DECPM	DECPM	POS_EQ_PMDEC	float	mas/yr	Proper Motion in Declination
11	DECPMErr	DECPMErr	POS_EQ_PMRA_ERROR	float	mas/yr	Proper Motion Error in Declination
12	Comments	Comments	REMARKS	char		Remark or Comments

  - DATA
    - TABLEDATA
      - TR (2)

	TD
1	
2	

图 9.6 LAMOST 导星星表数据模型

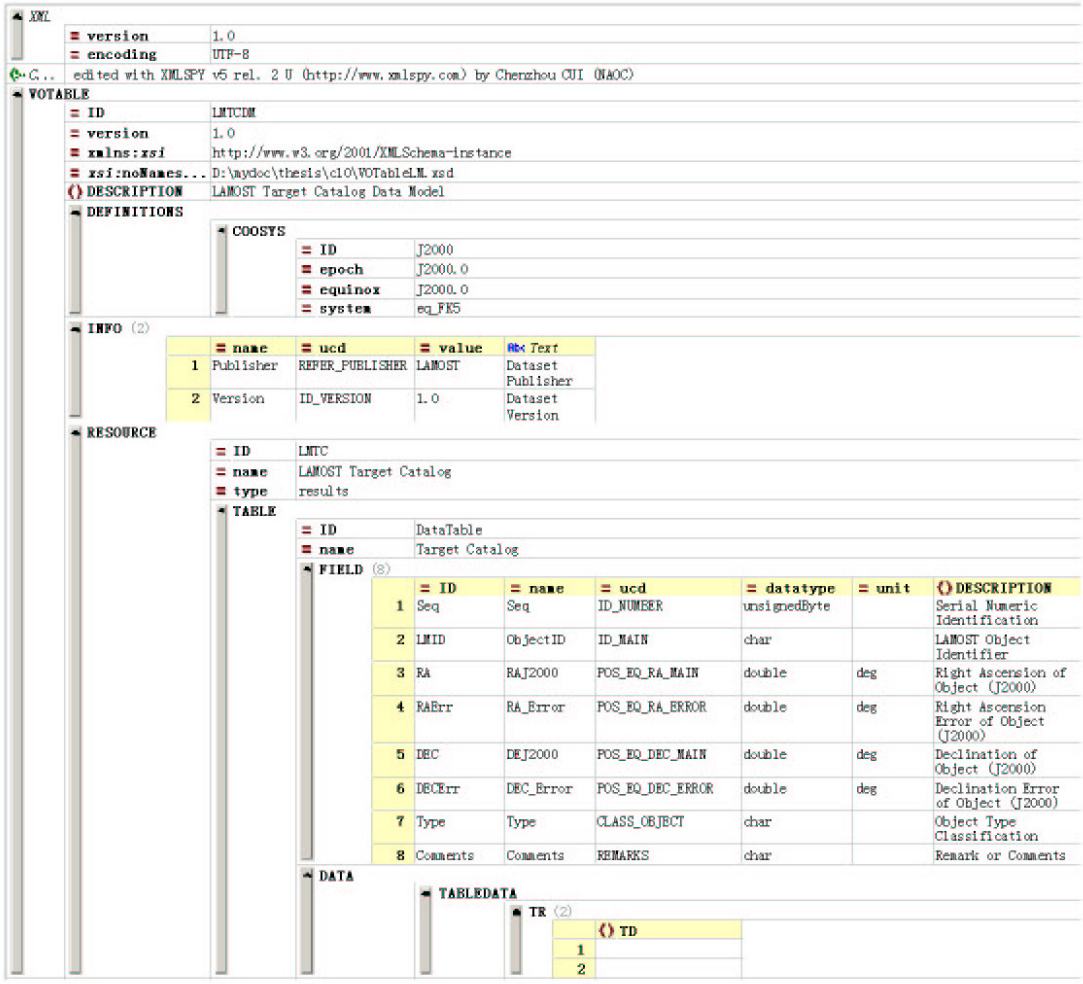


图 9.7 LAMOST 目标星表数据模型

参考文献

[1] LAMOST 初步设计文档. http://www.lamost.org/design.htm

[2] [VOTable] http://cdsweb.u-strasbg.fr/doc/VOTable/

[3] [RSM] Robert Hanisch. Resource and Service Metadata for the Virtual Observatory. http://www.ivoa.net/internal/IVOA/IvoaResReg/ResourceServiceMetadataV7.pdf

[4] Ray Plante. VOResource. http://rai.ncsa.uiuc.edu/~rplante/VO/schemas/VOResource.xsd

[5] [DublinCore] Dublin Core Metadata Initiative Metadata Terms. http://dublincore.org/documents/dcmi-terms/

[6] [UCD] Unified Content Descriptors. http://cdsweb.u-strasbg.fr/UCD/

第十章 系统范例

VO 是一个新生事物，其赖以存在的技术基础尚不成熟。天文学家对 VO 的功能需求也不十分明确。通过定义一些科学与技术范例可以帮助我们明确这些需求，为进一步的工作打下基础。论文的最后将通过三个系统范例：锥形检索、银盘金属丰度梯度统计研究、利用多波段数据检测 SVM 算法在天体自动分类中的应用，从数据访问、计算服务、数据互操作等方面对 China-VO 的功能需求进行探讨。

10.1 锥形检索

天文学上最常用的数据检索是基于天体位置的检索。然而，天体位置通常都存在测量误差，在数据检索时应该考虑进去。这就出现了给定位置（RA, DEC）和误差范围（R）的检索，也就是锥形检索（Cone Search）。

给定天球上一位置坐标（α，δ）和检索半径（r），找出与（α，δ）的角距离小于 r 的所有天体，这个过程称为锥形检索。给定圆心和半径，这不是圆形检索吗？怎么是锥形检索？我们知道，天体的坐标并不等于天体的真实位置。这里没有考虑天体的距离，是天体在天球上的投影。与给定位置（α，δ）的角距离小于 r 的天体在宇宙中的真实分布是一个无底的圆锥体。这个锥体的顶点是赤道坐标系的原点，即太阳。太阳到（α，δ）天体的连线为中心线，锥顶角为 r。如图 10.1 所示。

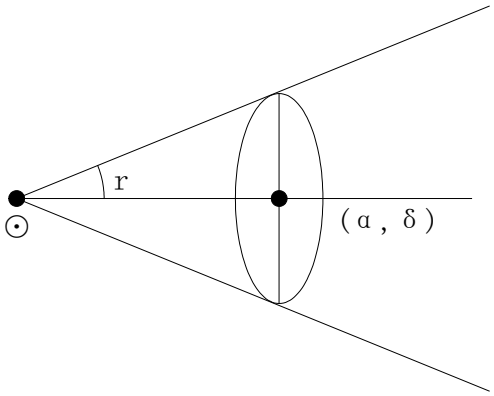


图 10.1 锥形检索

锥形检索是天文学最常用的数据检索方式，是联合查询（Join）、模糊查询（Fuzzy Join）、交叉证认等高级数据库查询的基础。高效、准确的锥形检索作为 VO 的基本服务，对 VO 其他高级服务的实现有重要的意义。

10.1.1 锥形检索的特点

锥形检索不是一般意义上的数据库联合查询。它有两个显著特点：

- 双参数查询

星表中天体的位置基本上都是以天球坐标系的方式存储的。每个天体有两个位置参数，比如赤经、赤纬。天文上常用的坐标系统有赤道坐标系、银道坐标系和黄道坐标系，其中赤道坐标系使用的是最普遍的。下面的讨论中将以赤道坐标系为例。两个天体之间的角距离必须通过赤经、赤纬的运算产生，然后才能与查询条件进行比较。

天文赤道坐标系是一种球面极坐标系，而不是直角坐标系。这里还存在一个球面环绕问题。赤经的取值范围是[0，360]，但是 0 度和 360 度是重合的。这与地球上的经度坐标类似，西经 180 度和东经 180 度是一条线。为了避免坐标环绕带来的复杂性，许多星表利用赤纬 DEC 进行索引。

- 非直接查询

从赤经、赤纬到角距离的计算过程不是简单的线性运算，而是球面坐标系中球面上两点间的大圆距离公式，如下所示：

$$r=\cos^{-1}[\sin(\delta)*\sin(\delta_0)+\cos(\delta)*\cos(\delta_0)*\cos(\alpha-\alpha_0)]$$

上面公式中存在数次三角函数运算。这对一般的数据库系统来说将使其检索性能大为下降。

10.1.2 传统的 SQL 语言实现

正如前面章节所述，大多数 DBMS 对幂运算和三角函数的实现格式不统一。下面给出的是 SQL 标准中锥形检索的实现语句。

情形如下：从星表 CAT 中查找到赤经、赤纬（α，δ）＝（123.45, -45.67）角距离（r0）小于 0.01 度的所有天体。其中 CAT 星表中赤经、赤纬的列名分别是（RA, DEC），单位是度，以实数形式表示。

则 SQL 语句如下：

```
SELECT * FROM CAT WHERE DEGREES(ACOS(SIN(RADIANS(DEC))
* SIN(RADIANS(-45.67)) + COS(RADIANS(DEC)) * COS(RADIANS(-
45.67)) * COS(RADIANS(RA-123.45)))) < 0.01;
```

这个语句已经足够复杂，现有 DBMS 的查询优化器很难对付，只能以顺序扫描的方式执行查询语句。

先提取数据的子集，再进行球面距离计算能在一定程度上提高性能。比如在数据库对 DEC 进行索引的前提下，首先检索出满足 DEC 条件的子集然后再

进一步查询。如下面语句所示：

```
SELECT * FROM CAT WHERE (DEC BETWEEN (-45.67-0.01) AND (-45.67+0.01)) AND (DEGREES(ACOS(SIN(RADIANS(DEC)) * SIN(RADIANS(-45.67)) + COS(RADIANS(DEC)) * COS(RADIANS(-45.67)) * COS(RADIANS(RA-123.45)))) < 0.01);
```

10.1.3 快速锥形检索的实现

索引可以大大提高 DBMS 的检索性能。如果不使用索引，大多数 DBMS 的顺序扫描性能都不如 C 或者 FORTRAN 程序对文件的直接操作。

由于锥形检索是双参数查询，如果仅对赤经（RA）或者赤纬（DEC）进行索引，语句执行时仍然需要进行数据的顺序扫描，不能从根本上改善检索的性能。要从根本上提高空间检索的性能必须实现空间索引。目前在 DBMS 中主要有两种实现方式：一种是对二维或多维参数的索引，即真正的空间索引；另一种是利用某种映射关系将二维的空间参量映射到一维参数空间，把（RA、DEC）合二为一，降低检索维度。这可以称为伪空间索引。

第一种途径必须要采用支持二维索引的 DBMS。目前，大多数商业的 DBMS 产品，比如 Oracle、Sybase、DB2 都以某种方式提供了对空间索引的支持。此外，开放源码产品 PostgreSQL 也提供了对空间索引的支持。MS SQL Server 和 MySQL 目前还不支持空间索引。但是这些索引功能的性能都不尽人意，用户使用也非常少。

Clive Page对一些数据库系统进行了调研[1]。DB2 提供了“Spatial Data Extender”的附加软件包。Informix 提供了“Spatial Datablade Module”，实现了基于R-tree的空间索引。Oracle提供了一个用于空间数据的附加软件包，实现基于HHCODE的空间索引，同时提供R-tree索引。Sybase也提供了空间索引功能支持。PostgreSQL也提供了对R-tree空间索引的支持。

根据 Clive Page 等人的测试，数据库空间索引的数据导入和索引建立的过程非常缓慢。同时，空间索引所占用的磁盘空间相当可观。比如 PostgreSQL 的 R-tree 索引，索引后的数据表大小是原始数据大小的近 12 倍。

地理信息系统（GIS）中坐标系与天文坐标系有一定的相似性。但遗憾的是 GIS 并没有提供太多可让天文学家借鉴的方案。

平衡二叉树（B-tree）一维索引与 DBMS 中的查询优化器实现了完美的结合。几乎所有的 DBMS 都使用 B-tree 作为默认的索引方式。大量的实践证明一维 B-tree 索引是非常成功的。相比之下，空间索引还是一个不成熟的领域，没有一种方法可以提供像一维 B-tree 索引那样快捷的数据插入、删除、查找，平衡结构的保持，高空间利用率等性能。

第二种途径，降维索引，目前有两种比较好的候选方案，即多级三角划分法（Hierarchical Triangular Mesh，简称HTM）[2]和多级等面积同纬度划分法（Hierarchical Equal Area isoLatitude Pixelisation，简称HEALPix）[3]。

它们的基本思路大致相同，都采用天区分割方式，对整个天区进行迭代式的层层划分。然后按照一定的编码规则给每个子天区单元编号，我们称其为PCODE。这个 PCODE 便与（RA，DEC）产生了一定的对应关系，从而实现了二维空间到一维空间的映射。数据库系统便可以采用成熟的索引技术对 PCODE 进行索引和操作，进而实现快速检索的目的。

10.1.4 HEALPix

HEALPix 最初是由 Krzysztof M. Górski 在 1997 年初设计的。当时他工作于哥本哈根的理论天体物理中心，现在工作于美国喷气推进实验室（JPL）。

HEALPix 方案，首先将整个天区分为 12 个面积相等的球面四边形。然后每个子天区再细分为 4 个更小的四边形天区单元。依此类推，如图 10.2 所示。

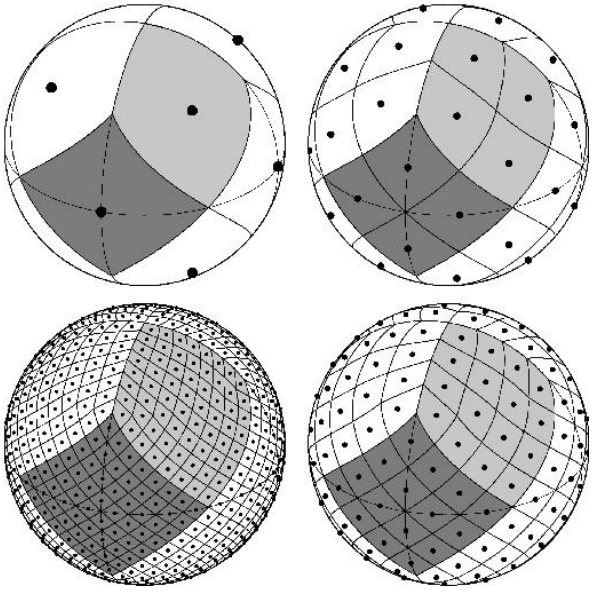


图 10.2　HEALPix 天区划分方案

HEALPix 天区单元编号方式有两种方式：环形（ring）和迭代方式（nested）。其中迭代方式更符合我们的要求：相邻单元的编号只在低位字节发生变化，这非常有助于索引和计算。

HEALPix 最初应用于宇宙微波背景卫星（CMB）数据。目前 COBE 卫星数据、IRAS 数据和 Planck 数据都采用或计划采用 HEALPix 编码。

10.1.5 HTM

HTM 是由约翰·霍布金斯大学的 Alex Szalay、Peter Kunszt、Ani Thakar 设计的一种天区划分方案，首先应用于 SDSS 巡天数据。

起始状态将整个天区分为 8 等份，上下各四个球面直角三角形。然后以每个球面三角形各边中点为新的顶点对原三角形四等分，依此类推。0 到 5 级的划分情况如图 10.3 所示。

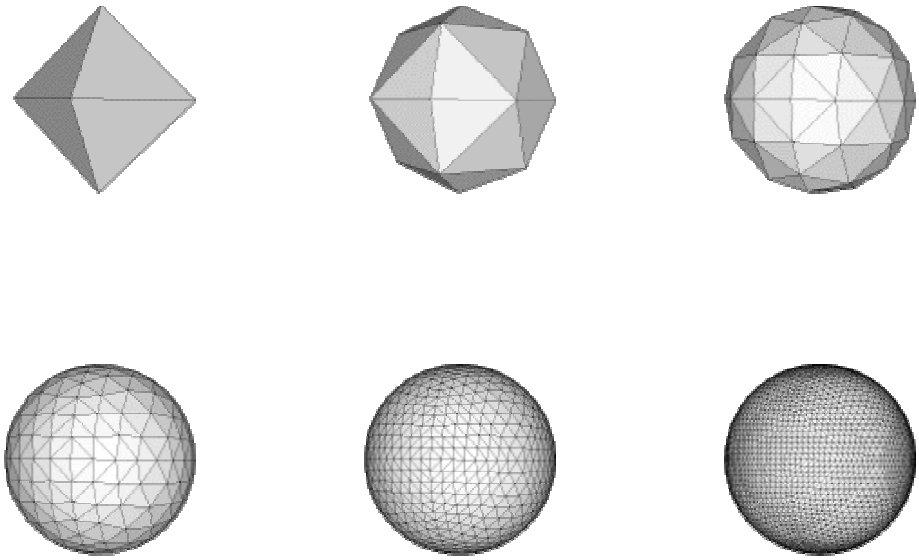


图 10.3 HTM 天区划分

HTM 的编码方式如图 10.4 所示。最顶层 8 块按顺时针方向依次命名为 N0、N1、N2、N3、S0、S1、S2、S3。由于都是四等份，父单元与子单元编码长度相差两个比特。

JHU 开发的软件包在数据库层面上提供了许多非常有利于锥形检索、模糊联合和交叉证认等工作的功能，比如：

- HTM ID 的生成
- 判定某点是否在给定的区域内
- 给出指定的 HTM 三角形的中心和面积
- 多种天区形状表达方式，比如三角形、圆形、矩形、多边形等
- 根据对天区的描述返回一个域
- 自动调整 HTM 级次，以适应对空间分辨率的需要
- 返回给定天区的特定级次的所有 HTM 单元
- 不同天区的联合和交叉操作
- 返回给定天区内的所有天体
- 搜索与给定天区有交叉关系的单元
- 判定一个单元是否与其他单元相交

HTM 已经开始被空间望远镜科学研究所（STScI）用来处理 GSC-II 星表。欧洲空间局（ESA）的 GAIA 项目也计划采用 HTM 进行数据处理。

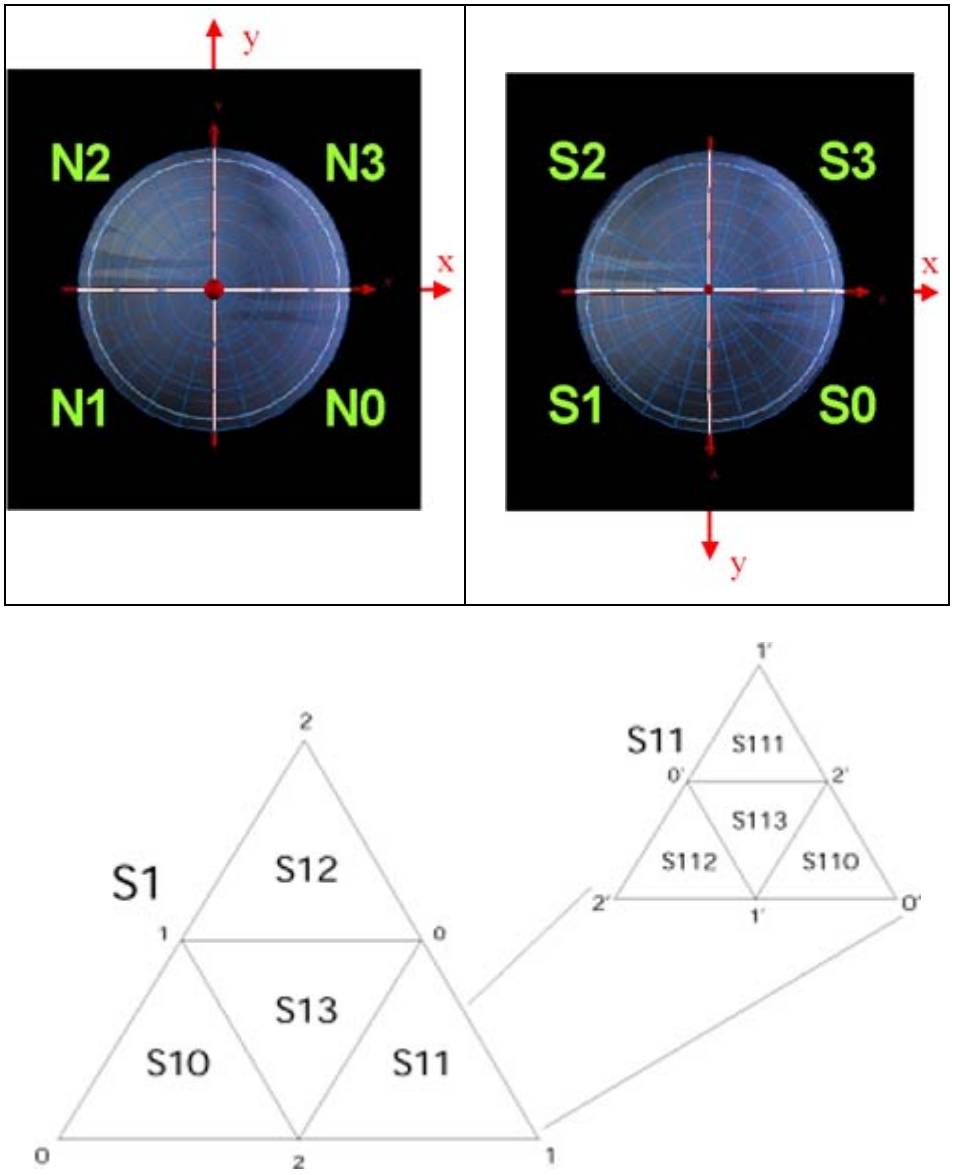


图 10.4　HTM 编码方案

10.1.6 PCODE 使用过程中需要注意的问题

划分级次的选择

如何选择合适的划分级次，以便每个单元中源的个数不会太多，同时单元面积要大于源的位置误差范围。

表 10.1 给出了 HTM 方案中不同划分级次对应的单元数和单元大小。

数据库结构调整

为了使用 PCODE，不得不对现有的数据表进行结构上的调整。至少要在

每个表中增加一列以存储相应的 PCODE。这需要一定的工作量。

级次	面积 (arcsec^2)	单元数
10	1.77E1	8,388,608
11	4.43E0	33,554,432
12	1.11E0	134,217,728
13	2.77E-1	536,870,912
14	6.92E-2	2,147,483,648
15	1.73E-2	8,589,934,592
20	1.69E-5	8,796,093,022,208
25	1.65E-8	9,007,199,254,740,922

表 10.1　HTM 不同划分级次对应的单元数和单元大小

编码长度问题

现在 32 位计算机是主流设备，32 比特长度编码的 PCODE 对应的天区单元大小在 20 到 30 角秒，这符合现在大多数天文观测的巡天深度。不过，随着巡天深度的加大，这个尺度大小内源的数据将很快增加。幸运的是，64 位计算已经出现，并在今后几年内不断普及。64 位编码长度的 PCODE 将是未来的必然选择。

当前，HEALPix 软件包提供了 F90 和 IDL 版本，HTM 则提供了 C、C++ 和 Java 版本。相比而言，HTM 更适合 VO 开发者的使用。

利用 PCODE 索引的好处是对 DBMS 没有特殊的要求，不需要数据库系统提供空间索引的功能。

10.1.7 锥形检索在 China-VO 系统的实现

目前，China-VO以Globus Toolkit 3.0[4] Beta为OGSA网格平台已经成功地实现了传统索引方式基础上的锥形检索，如图 10.5 所示。

考虑到真正空间索引实现的复杂性和对数据库管理系统的要求，China-VO 计划采用 PCODE 的降维索引方案来实现快速锥形检索。因为 SDSS 和 LAMOST 项目存在许多相似之处。HTM 在 SDSS 的数据发布中得到成功的应用，同时 HTM 提供 Java 版本的实现程序，所以 China-VO 计划对 HTM 方法进行性能测试。如果取得满意的结果，那么在 China-VO 系统中 HTM 将作为标准的空间索引方式。

10.2　银盘金属丰度梯度统计研究

10.2.1 研究背景

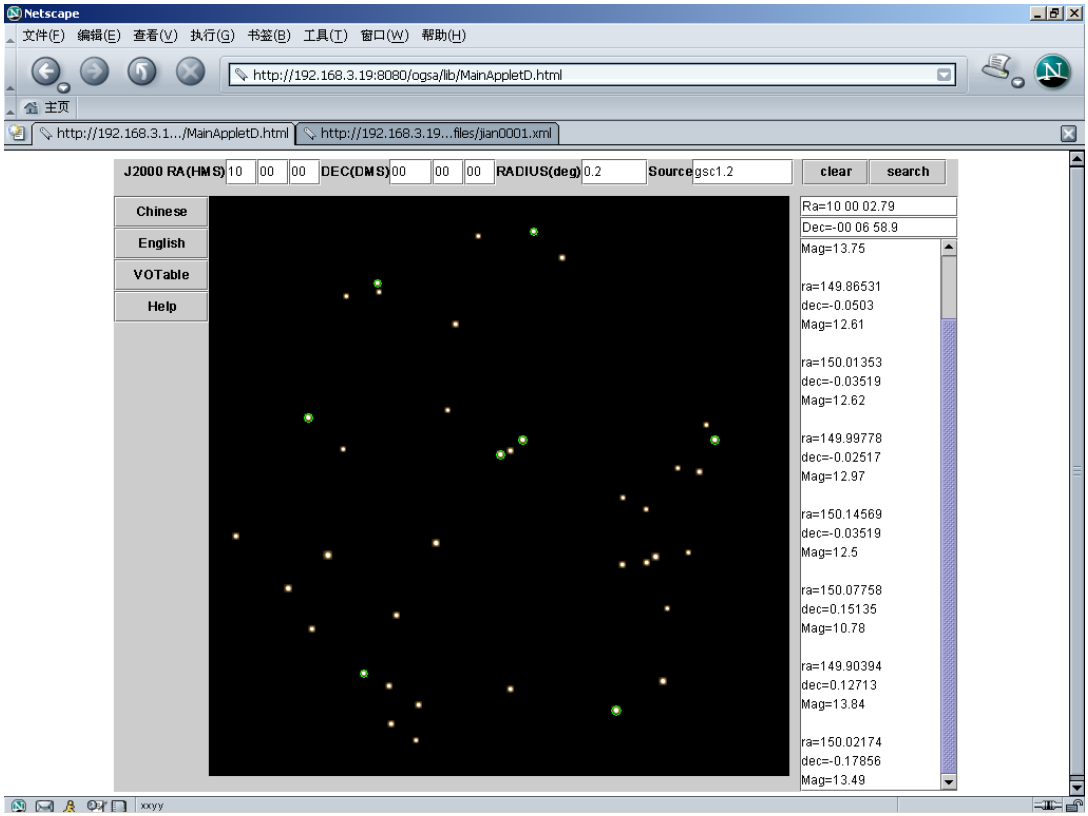
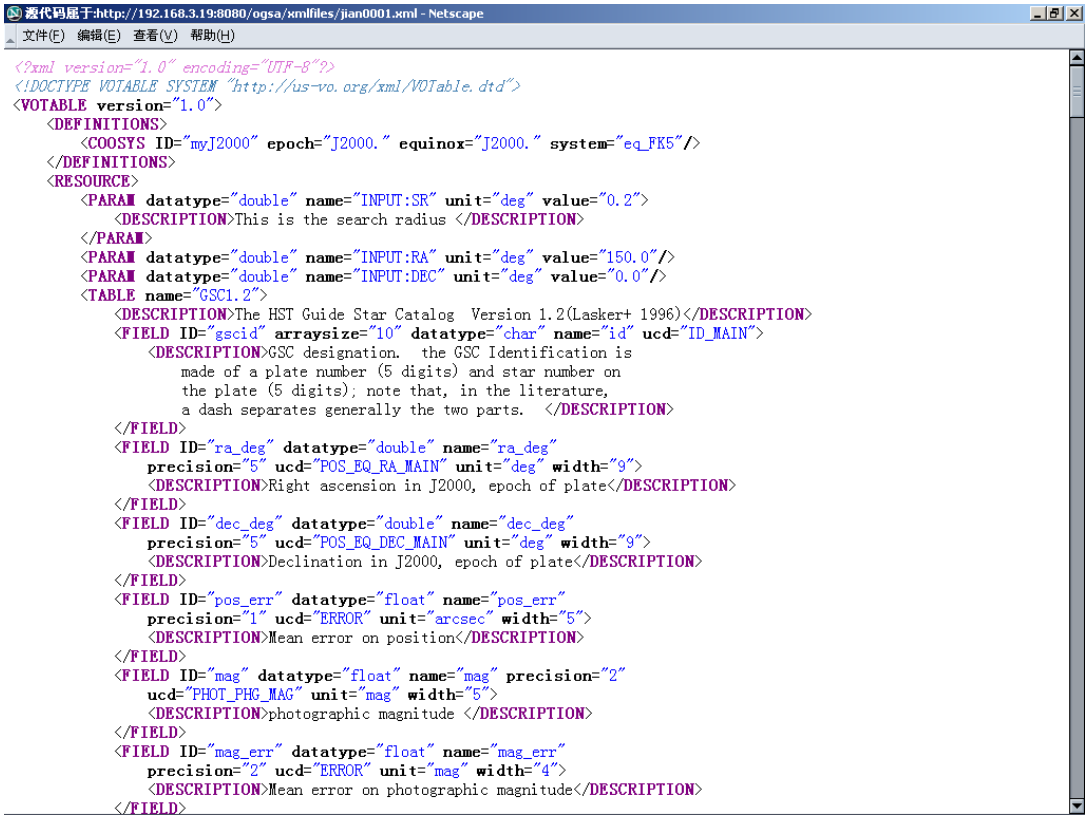


图 10.5 (a)　　China-VO 锥形检索图形界面




```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE VOTABLE SYSTEM "http://us-vo.org/xml/VOTable.dtd">
<VOTABLE version="1.0">
    <DEFINITIONS>
        <COOSYS ID="myJ2000" epoch="J2000." equinox="J2000." system="eq_FK5"/>
    </DEFINITIONS>
    <RESOURCE>
        <PARAM datatype="double" name="INPUT:SR" unit="deg" value="0.2">
            <DESCRIPTION>This is the search radius </DESCRIPTION>
        </PARAM>
        <PARAM datatype="double" name="INPUT:RA" unit="deg" value="150.0"/>
        <PARAM datatype="double" name="INPUT:DEC" unit="deg" value="0.0"/>
        <TABLE name="GSC1.2">
            <DESCRIPTION>The HST Guide Star Catalog  Version 1.2(Lasker+ 1996)</DESCRIPTION>
            <FIELD ID="gscid" arraysize="10" datatype="char" name="id" ucd="ID_MAIN">
                <DESCRIPTION>GSC designation.  the GSC Identification is
                    made of a plate number (5 digits) and star number on
                    the plate (5 digits); note that, in the literature,
                    a dash separates generally the two parts.  </DESCRIPTION>
            </FIELD>
            <FIELD ID="ra_deg" datatype="double" name="ra_deg"
                precision="5" ucd="POS_EQ_RA_MAIN" unit="deg" width="9">
                <DESCRIPTION>Right ascension in J2000, epoch of plate</DESCRIPTION>
            </FIELD>
            <FIELD ID="dec_deg" datatype="double" name="dec_deg"
                precision="5" ucd="POS_EQ_DEC_MAIN" unit="deg" width="9">
                <DESCRIPTION>Declination in J2000, epoch of plate</DESCRIPTION>
            </FIELD>
            <FIELD ID="pos_err" datatype="float" name="pos_err"
                precision="1" ucd="ERROR" unit="arcsec" width="5">
                <DESCRIPTION>Mean error on position</DESCRIPTION>
            </FIELD>
            <FIELD ID="mag" datatype="float" name="mag" precision="2"
                ucd="PHOT_PHG_MAG" unit="mag" width="5">
                <DESCRIPTION>photographic magnitude </DESCRIPTION>
            </FIELD>
            <FIELD ID="mag_err" datatype="float" name="mag_err"
                precision="2" ucd="ERROR" unit="mag" width="4">
                <DESCRIPTION>Mean error on photographic magnitude</DESCRIPTION>
            </FIELD>
```

图 10.5 (b)　　China-VO 锥形检索返回的 VOTable 格式数据

银河系的形成和演化多年来一直是天文学家们研究的热点。关于银河系的形成，目前主要有两种观点：ELS模型[5]和SZ模型[6]。ELS模型认为银河系是由一个单一的原始星系云塌缩而成。在塌缩过程中银河系从一个弥散的、球体形的、缓慢旋转的、贫金属的晕平稳而迅速地演化成一个高密度的、扁平的、快速旋转的、富金属盘。在塌缩过程中，不断爆发的超新星持续增加着星际介质的金属丰度。因此，可以期望从银晕到银盘将存在明显的金属丰度从低到高的梯度。SZ模型认为原始的银河系不是仅由单一的原始星云演化而来，而是由多个具有不同演化历史的原始星云团块相互碰撞或吸积而成。在SZ模型中多种成分的混合使得年龄、金属丰度、动力学之间没有明显的相关性。

银盘上化学元素的丰度梯度对研究银河系特别是银盘的形成与演化过程有着重要的意义。丰度梯度的时空变化趋势是银河系星际介质增丰历史以及内落、外流等过程的反映，是建立银河系化学演化模型的重要约束。按照银河系是从内到外形成的观点，银盘内部区域的密度比外部区域高，内部区域的恒星形成和化学演化速度都要比外部区域快。因此，内部区域的化学丰度要比外部区域高——即表现出丰度梯度。然而，银河系演化过程中有些效应（如银河系的动力学演化、恒星本身的运动、恒星与巨分子云的相互作用）会影响化学元素的分布，从而将这种丰度梯度抹平。总而言之，银河系或银盘上是否存在化学元素的丰度梯度，以及梯度随空间、时间的变化不仅为银河系化学演化模型提供重要限制，而且与银河系形成机制、银河系发生的动力学过程直接相关。

陈玉琴等[7]（简称Ch2000）利用中国科学院国家天文台 2.16m望远镜对 90 颗F、G型星样本进行了高分辨率、高信噪比的光谱观测，给出了这 90 颗样本的Fe、O、Na等 13 种元素的金属丰度。如果能得到这些样本的轨道参数，那么把陈玉琴等给出的丰度数据和相应样本的轨道参数进行相关分析便可以用来研究银盘上金属丰度梯度在空域、时域的变化，进而对银河系化学演化模型的建立给出更多的约束。幸运的是，Hipparcos星表提供了大批恒星的动力学数据，利用Allen和Santillán [8]1991 年提出的银河系质量分布模型作为势函数进行数值积分便可以得到样本的轨道参数。Hipparcos卫星是ESA1989 年 8 月发射的一颗高精度天体测量卫星。经过 3 年多的观测，它以空前的精度获得了 118218 颗恒星的位置、自行、三角视差等数据，为天文学家提供了一大批精确的恒星基本数据。

为了利用 Ch2000 和 Hipparcos 星表完成丰度梯度统计分析工作，必须要完成以下工作：

1. 从 Hipparcos 星表中找出与 Ch2000 相对应样本的位置数据（赤经、赤

纬）和动力学数据；

2. 利用上面的位置数据和动力学参数进行数值积分得到样本的轨道参数；
3. 对样本的金属丰度和轨道参数进行相关分析；
4. 以图表的形式给出分析结果。

10.2.2 传统的研究模式

下面就以本人 2003 年发表在《中国科学》上的统计分析工作[9]为例说明按照传统的研究方式上述工作是如何完成的。

1. 从互联网上搜寻并下载文本格式的 Hipparcos 星表；
2. 利用星表附带的“Readme”文件对星表进行必要的格式转换，然后输入到 MySQL 数据库；
3. 对 Ch2000 丰度数据文件进行格式整理，输入 MySQL 数据库；
4. 以两个星表共有的“HD”号作为关键字对两个数据表进行联合查询，返回结果包括 Ch2000 中的金属丰度数据和 Hipparcos 星表中的位置与动力学数据；
5. 利用 Allen 和 Santillán 1991 年提出的银河系质量分布模型作为势函数编写轨道计算数值积分程序；
6. 对 Ch2000 中的样本进行轨道计算；
7. 把计算得到的轨道参数与丰度数据合并，以 CSV 格式保存；
8. 利用 Windows 平台上的 Origin 数据分析软件对样本的丰度和轨道参数进行相关分析，通过线性拟合，得到拟合参数（如表 10.2）和丰度、轨道参数关系图（如图 10.6）。

10.2.3 VO 环境下的实现

1. 将 Ch2000 丰度表进行必要的格式转换，输入到 MySQL 或者其他数据库；
2. 对丰度表进行数据集注册，发布相应的元数据；
3. 通过 VO Registry，在 IVO 系统内发现可用的 Hipparcos 星表服务，可能不止一个，如果多于一个便可利用数据路由服务从中选取最优服务；
4. 通过 VOQL 调用数据查询服务，对 Ch2000 和发现的 Hipparcos 两个数据集进行联合查询，把结果保存在 VO 用户空间 MyVO；
5. 调用银河系轨道计算服务，得到样本的轨道参数；
6. 调用统计分析服务对丰度和轨道参数进行线性拟合得到拟合参数；
7. 调用 VO 可视化服务，给出丰度、轨道参数关系图。

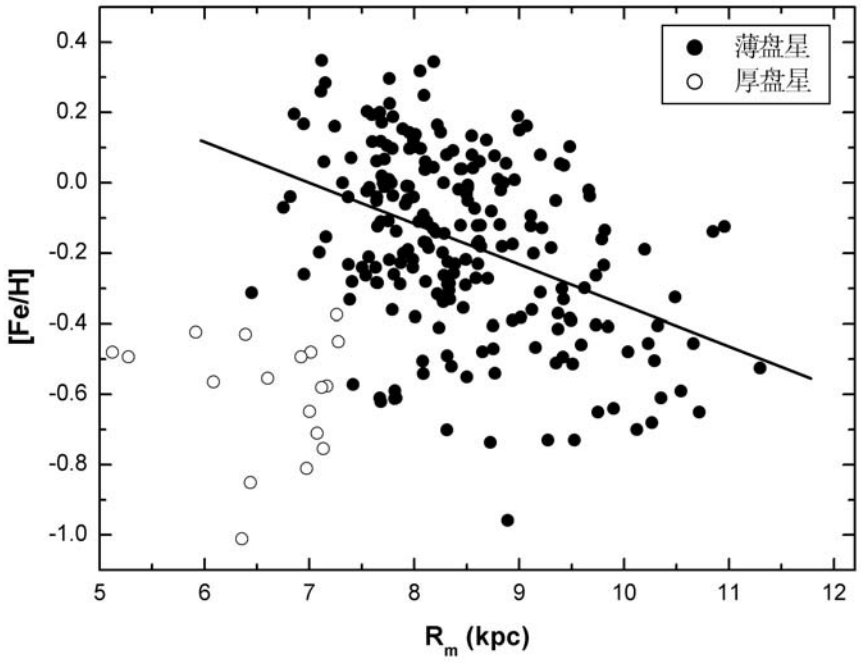


图 10.6 [Fe/H]与平均轨道半径R_m的关系

元素	分组依据	A	B	R	SD	N
Fe	不分组	0.811±0.150	-0.116±0.018	0.410	0.233	217
	Age≤4Gyr	0.666±0.205	-0.088±0.024	0.378	0.169	79
	4Gyr<Age≤6Gyr	0.660±0.304	-0.096±0.036	0.336	0.219	56
	6Gyr<Age≤8Gyr	0.263±0.399	-0.061±0.046	0.233	0.264	33
	Age>8Gyr	0.938±0.301	-0.145±0.035	0.521	0.248	49
O	不分组	0.502±0.132	-0.065±0.016	0.336	0.165	140
Na	不分组	0.822±0.157	-0.116±0.018	0.403	0.242	208
Mg	不分组	0.888±0.139	-0.114±0.016	0.445	0.207	200
Al	不分组	0.740±0.170	-0.101±0.020	0.337	0.248	197
Si	不分组	0.761±0.138	-0.104±0.016	0.400	0.214	217
Ca	不分组	0.734±0.127	-0.102±0.015	0.422	0.197	217
Ti	不分组	0.612±0.140	-0.090±0.016	0.355	0.213	206
Ni	不分组	0.886±0.165	-0.123±0.019	0.401	0.251	209
Ba	不分组	0.711±0.180	-0.101±0.021	0.346	0.245	169

表 10.2 金属丰度与平均轨道半径R_m的统计结果[a]

[a] 拟合公式为Y=A+B·X；A、B为线性拟合的截距和斜率；R为相关系数；SD为剩余标准差；N为参加统计的样本数

10.2.4 VO 工作模式的优越性

- 提高工作效率

在上面传统的工作模式中只有 4 和 8 两步用到了现成的软件工具，其余各步均需要天文学家手工或者半手工完成。VO 工作模式中，只有 1、2 两步以天文学家为主完成，其余服务均有 VO 系统提供，只需要直接调用即可。

- 提高工作质量

VO 提供的服务都经过严格的开发和测试过程并经不同用户使用检验过，有很好的服务保证。相反，研究者个人操作和编写程序出错的几率则要大得多。

- 复用性好

个人研究中所完成的工作，比如编写的程序和完成的数据处理操作，很难能被其他的天文学家重复使用。相反，VO 中提供的服务有良好的互操作性，可以被其他用户或服务调用。不但如此，1、2 步骤完成的丰度表服务的发布将使得 Ch2000 成为 VO 系统中的一个数据集服务，充实了 VO 的资源，也可被其他服务发现使用。

10.2.5 对 China-VO 的主要功能要求

为了在 VO 架构中完成上面范例的银河系丰度梯度研究，China-VO 必须提供以下几方面的功能：

- 数据集注册、元数据发布
- 资源发现
- 数据集访问
- 计算服务
- 可视化服务
- 统计分析服务
- 用户空间 MyVO

10.3 利用多波段数据检测 SVM 算法在天体分类中的应用

进入虚拟天文台时代的天文学面临的是海量数据问题。在虚拟天文台的系统中效率高、鲁棒性好的数据挖掘工具是必不可少的。VO 不是为某个特定研究课题而开发的，为了满足不同用户的需求，需要我们开发和测试各种数据挖掘算法。

支持向量机（SVM）算法是一种自动分类算法，最初是由Vapnik在 1995

年提出的[10]。因其许多诱人的特点和卓越的性能受到越来越多的欢迎，被广泛用于预测和分类。目前SVM算法已经有许多成功的实用案例，比如文字分类、人脸识别、手写签名识别等。一些观点认为SVM算法在许多情况下的表现比神经网络更好。

Wozniak et al. (2001) [11]和Humphreys et al. (2001) [12]率先将SVM算法应用到了天文学领域。Wozniak等人利用SVM进行变星的自动分类，结果证明SVM是一种高效、高准确度的分类算法。Humphreys等人利用SVM算法进行星系形态分类，同样证明了SVM是一种非常有效的分类算法。

10.3.1 基本思想

利用多波段数据来检验SVM算法对于天体自动分类的效力。我们利用一批已知类型的样本从不同波段的巡天观测数据中提取其多个特征参量。然后以这批参量数据同时作为SVM分类算法的训练样本及其检验样本。将SVM的最终分类结果与已知结果比较便可以评价出其分类效果[13]。

10.3.2 工作过程

这个示例中使用的样本数据取自四个不同的数据源：Veron 2000 AGN星表[14]、USNO-A2.0 光学巡天星表[15]、ROSAT X射线巡天星表[16]、2MASS 红外巡天星表[17]。

- Veron 2000 AGN 星表：Veron-Cetty 和 Veron 2000 对他们 1998 年公布的 AGN 星表进行了修正，给出了 13214 颗类星体、462 颗 BL Lac 天体和 4428 个 AGN 的位置等基本数据。示例中我们利用这个星表与其他三个星表作交叉证认，找出相应对应体的多维参数。
- USNO-A2.0 光学巡天星表：USNO-A2.0 是美国海军天文台对精确测量仪（Precision Measuring Machine，PMM）的扫描数据重新处理后得到的。整个星表包括 526,280,881 颗恒星。此示例中我们将使用其提供的（B, R）两个参量的数据。
- ROSAT X 射线巡天星表：ROSAT 巡天星表是 ROSAT 卫星全天巡天观测的成果，其亮源表和暗源表共给出了 124730 个 X 射线源的位置、X 射线流量、光谱型等数据。本示例中将使用该星表中的（ct, HR1, HR2, ext, ext1）五个参量的数据。
- 2MASS 红外巡天星表：2 微米红外巡天星表给出了目标天体 J、H、K 三个红外波段的测光结果，示例中将使用其提供的（J, H, K）这三个参数的数据。

在这个示例中，我们利用 Veron 星表依次与 USNO、ROSAT、2MASS 三个不同波段的巡天星表作交叉证认，从中提取出 Veron 天体在这三个巡天星表

中（B, R, ct, HR1, HR2, ext, ext1, J, H, K）共 10 个参量的数据作为初始数据。

1. 将 Veron 星表与 USNO 进行证认，从 USNO 星表中提取对应体的（B, R）数据，产生第一层证认数据（L1）；
2. 将 L1 与 ROSAT 星表进行证认，从 ROSAT 中提取对应体的（ct, HR1, HR2, ext, ext1）五个参量的数据，产生第二层证认数据（L2）。此时每个样本已经拥有 7 个参量的观测结果。
3. 将 L2 与 2MASS 进行证认，从中提取（J, H, K）三个参量数据，产生第三层证认数据（L3）。此时每个样本对应的观测参量增加到 10 个。

最后，为了更好的体现类星体、BL Lac 天体、AGN 在不同参数空间的差别和便于分类，需要对其中的少量数据项进行简单的计算处理，得到“B-R”，“B+2.5lg(ct)”，“J+2.5lg(ct)”，“J-H”，“H-K”。最后作为 SVM 算法输入数据（L4）的参数为{B-R, B+2.5lg(ct), ct, HR1, HR2, ext, ext1, J-H, H-K, J+2.5lg(ct)}。

接下来便是将得到的 10 维样本数据作为训练样本对 SVM 分类器进行训练。训练结束后再利用同样的数据作为检验样本，让 SVM 分类器进行自动分类。

最后，将 SVM 的分类结果与样本的真实类型进行比较，评估算法的分类能力。数据采集流程如图 10.7 所示。

10.3.3 检测工作在 VO 环境下的实现

1. 用户登陆 VO 门户；确定对数据访问服务、计算服务、数据挖掘服务的需求；向 VO 系统提交任务请求。
2. VO 系统发现所需要的数据访问服务（包括 Veron、USNO、ROSAT 和 2MASS 星表访问服务和交叉证认服务）、计算服务（本例中只需要基本数学运算服务）和数据挖掘服务（SVM 分类服务）。
3. 调用证认服务，对 Veron 和 USNO 进行交叉证认，保存证认结果 L1。
4. 调用证认服务，对 L1 和 ROSAT 进行交叉证认，保存证认结果 L2。
5. 调用证认服务，对 L2 和 2MASS 进行交叉证认，保存证认结果 L3。
6. 调用数学计算服务对 L3 中的部分参量进行计算，得到输入数据 L4。
7. 以 L4 为训练样本对 SVM 分类服务进行训练。
8. 以 L4 为检测数据，调用经过训练的 SVM 分类服务，返回分类结果。

10.3.4 对 China-VO 的主要功能要求

为了利用多维数据检测 SVM 算法在天体自动分类中的应用效果， China-VO 必须提供以下几方面的功能：

- 资源发现
- 数据集访问
- 多波段数据的互操作
- 交叉证认
- 数学计算
- 数据挖掘
- 中间结果保存
- 用户空间 MyVO

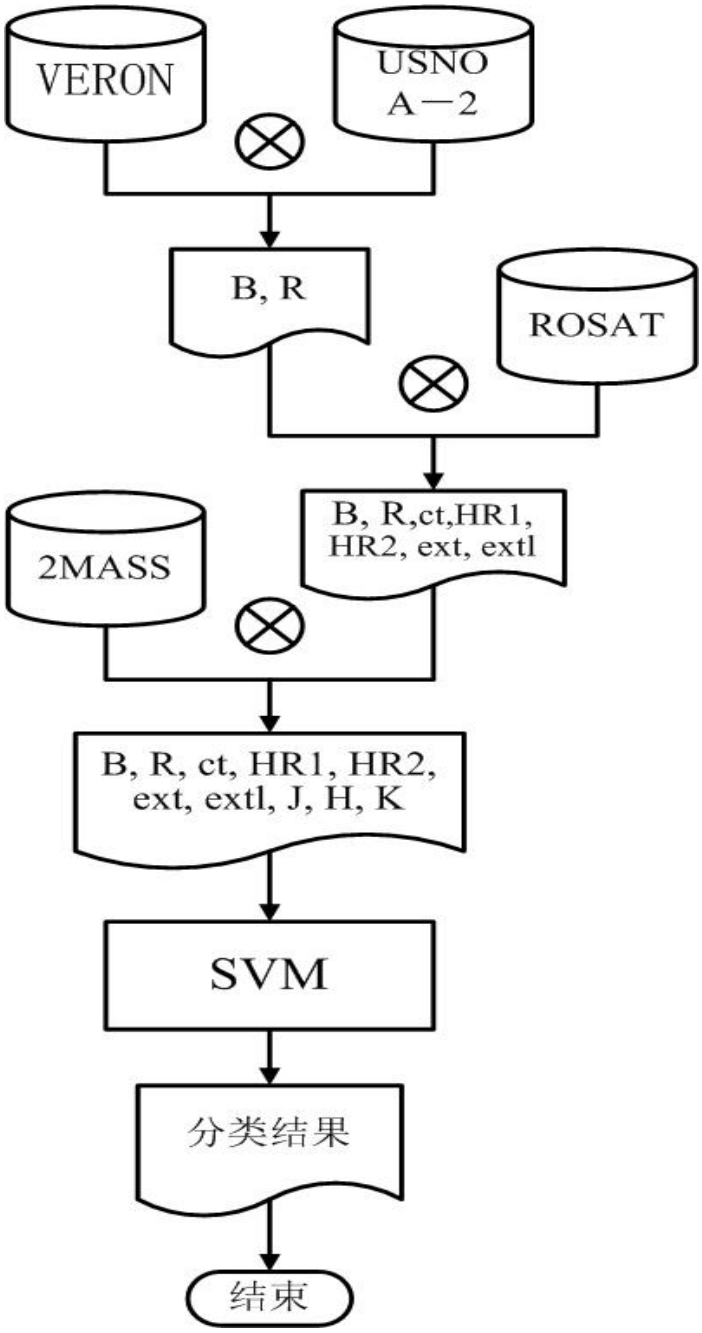


图 10.7　SVM 分类数据采集流程

参考文献

[1] Clive Page. Indexing the Sky. http://wiki.astrogrid.org/bin/view/Astrogrid/SkyIndexing

[2] [HTM] Hierarchical Triangular Mesh. http://taltos.pha.jhu.edu/htm

[3] [HEALPix]Hierarchical Equal Area isoLatitude Pixelisation. http://www.eso.org/science/healpix/

[4] [GT3] Globus Toolkit 3. http://www.globus.org/ogsa/

[5] Eggen O J, Lynden-Bell D, Sandage A. Evidence from the motions of old stars that the Galaxy collapsed. ApJ, 1962 (136): 748~766

[6] Searle L, Zinn R. Compositions of Halo Clusters and the formation of the Galactic Halo. ApJ, 1978 (225): 357~379

[7] Chen Y Q, Nissen P E, Zhao G, et al. Chemical composition of 90 F and G disk dwarfs. A&AS, 2000 (141): 491~506

[8] Allen C, Santillán A. An improved model of the galactic mass distribution for orbit computations. RMAA, 1991 (22): 255~263

[9] Chenzhou CUI, Yuqin CHEN, Gang ZHAO, Yongheng ZHAO. Abundance Gradients in the Galactic Disk. Science in China (Series G), 2003 (1): 52-61

[10] V. Vapnik. The Nature of Statistical Learning Theory. New York: Springer-Verlag, 1995.

[11] P. R. Wozniak, C. Akerlof, S. Amrose, et al. Classification of ROTSE Variable Stars using Machine Learning. AAS, 2001 (199): 130.04

[12] R. M. Humphreys, G. Karypis, M. Hasan, et al. Experiments in Automating the Morphological Classification of Galaxies. AAS, 2001 (199): 10.15

[13] Yanxia ZHANG, Chenzhou CUI, Yongheng ZHAO. Classification of AGNs from Stars and Normal Galaxies by Support Vector Machines. In: Jean-Luc Starck, Fionn D. Murtagh. Astronomical Data Analysis II. Proc. of SPIE, 2002. 371-178

[14] M. P. Véron-Cetty, P. Véron. ESO Scientific Report. 2000 (19)

[15] [USNO-A2] http://ftp.nofs.navy.mil/projects/pmm/catalogs.html

[16] [ROSAT] ROSAT All Sky Survey Catalogs. http://wave.xray.mpe.mpg.de/rosat/catalogue

[17] [2MASS] Two Micron All Sky Survey. http://pegasus.phast.umass.edu/

结　论

虚拟天文台（VO）是一个诞生还不到五年的新概念。作为中国虚拟天文台计划（China-VO）的系统设计，本论文主要包括以下内容：

- 第一、二章分别介绍了 VO 的产生背景、发展现状以及与其相关的主要信息技术；
- 第三章阐述了 China-VO 建设的必要性、现有条件、实施路线，提出并阐述了 VO-enabled LAMOST 的观点；
- 第四章以开放网格服务架构（OGSA）为基础给出了 China-VO 系统体系结构的概念模型和服务模型，并指明 China-VO 的研发重点是天文科学相关的 VO 服务；
- 第五到第八章分别从资源注册与发现、数据访问与互操作、应用服务、系统门户四个方面介绍了国际上相关研究的最新进展，阐述了 China-VO 的观点和可能采取的相应策略；
- 第九章以国际虚拟天文台联盟（IVOA）正在讨论的资源服务元数据模型（RSM）为基础，结合 LAMOST 的具体特点给出了 LAMOST 巡天数据产品与工作星表的资源元数据模型和数据模型；
- 论文的最后为 China-VO 设计了三个系统范例，从数据访问、计算服务、数据互操作等方面明确了 China-VO 的功能需求。

虚拟天文台是“科学驱动，技术使能”的产物。天文观测数据的爆炸性增长使得天文学家对它们进行系统的处理、分析和理解越来越困难。天文学家希望借助网格、XML、语义网等新兴的 IT 技术打破目前世界上不同天文研究机构，特别是数据中心之间研究资源相互孤立的状态，实现全球主要天文研究资源，特别是海量巡天观测数据的协同工作和高级共享。这样的系统就是虚拟天文台。

VO 将是一个数据密集型在线天文研究平台。其中一个关键的科学目标就是挖掘多波段巡天数据综合分析处理所带来的科学潜能。巡天观测带来了巨大的科学发现的潜力，对这些巡天数据的联合利用，将涌现出全新的、无法预见的、意义重大的科学产出。VO 将以其独一无二的资源和技术优势使天文学取得前所未有的成果，让普通公众得到真实的天文体验，成为开创“天文学发现新时代”的关键性因素。

在国际上建设虚拟天文台的浪潮一浪高过一浪之时，中国天文界也提出了

建设中国虚拟天文台的计划。作为国际虚拟天文台的一部分和联结国内外天文研究的桥梁，China-VO 将为国内天文学家和普通公众带来国际虚拟天文台（IVO）丰富的资源与技术，同时实现国内主要天文资源与国际同行的共享。China-VO 将与国内天文界目前唯一的大科学工程 LAMOST 紧密合作，以 LAMOST 光纤光谱巡天数据为重要的数据资源，以光谱自动处理与分析服务为主要特色，实现 VO-enabled LAMOST。

被称为第三代互联网的网格技术旨在把整个互联网整合成一台巨大的超级计算机，实现计算资源、存储资源、数据资源、信息资源、知识资源、专家资源的全面共享。网格技术领域最具代表性的标准及其实现分别是 OGSA 和 Globus Toolkit V3（GT3）。VO 的许多科学目标都可以看成是它们的高级应用。因此， China-VO 将以 OGSA 架构为体系结构的基础，以 GT3 为主要的运行平台。

逻辑上，China-VO 是层次式的体系结构，从底层到上层依次是构造层、资源层、汇集层和用户层。构造层是整个系统的资源基础；资源层利用标准的资源模型实现统一的资源访问。汇集层整合了最能体现天文特色的各种 VO 服务，比如数据处理、数据挖掘、统计分析、可视化、网格计算等。用户层是系统与用户直接的接口，将实现用户登录、资源浏览、任务编制与提交、结果显示、数据保存、偏好设定等功能。

开发上，China-VO 将采用面向服务的设计思路。以 Java 为主要的程序设计语言，开发与网格服务兼容的 VO 服务。作为网格技术的应用系统，China-VO 将把研究与开发的重点放在与天文科学相关的领域，比如数据与服务的互操作、资源的注册与发现、元数据服务、数据模型等。这也是本论文重点探讨的内容。资源管理、安全服务、作业调度、系统监测等方面的功能尽量直接采用 GT3 或者其他符合 OGSA 标准的网格系统提供的服务。

资源的注册与发现是网格技术的精华，同样也是虚拟天文台的精华。在任何时间，一旦某个资源可为虚拟天文台所用就应该被及时地发现。OGSA 通过定义一系列标准化的服务描述和操作提供了服务注册和发现的基本机制。但要真正实现 VO 环境中资源的注册和发现还需要 VO 开发者在资源注册标准、数据模型、元数据标准等方面做出大量的努力。要实现 IVO 范围内资源的统一注册与发现，需要各国 VO 计划的相互配合。China-VO 作为国际虚拟天文台联盟的一员，将积极参与 IVOA 关于资源注册的研究与开发工作。

面对 TB 甚至 PB 如此海量的数据，China-VO 将根据实际需求从 NAS、SAN、iSCSI 等先进的网络存储和虚拟存储方案中择优而用，同时选择与网格环境兼容的数据库管理系统，做到既满足 VO 对数据存储与访问的高性能需求又

节省和保护投资。为了实现海量数据的快速检索，特别是天文学上最常用的锥形检索和交叉证认，China-VO 将开发和采用高效的伪空间索引，将两维位置参量投影到一维空间。同时，针对标准查询语言（SQL）在许多情况下无法满足天文查询和数据挖掘的需要，China-VO 将与 IVOA 伙伴一道在 XML 的体系下开发新型的虚拟天文台查询语言（VOQL）。

为了能实现 VO 系统中异构数据和服务的统一访问，抽象化是一条重要的途径。通过定义适用性广、扩展性好的数据模型以及相应的资源注册与发现标准，来屏蔽数据资源在数据格式、存储格式、主机环境、访问形式等诸多方面的异构性、复杂性，实现统一、透明的访问。China-VO 根据目前国际上相关领域的研究状况，采用 XML Schema，按照与 IVOA 数据模型兼容的方式对 LAMOST 数据产品等系统资源进行数据模型定义。

数据挖掘、可视化和高性能计算服务等高层应用服务是最能体现虚拟天文台天文特色的服务，也是 VO 能否最终让天文学家接受和使用的关键。由于在体系结构上的分布式、网格化等特点，VO 对这些领域提出了挑战。但同时我们看到这些挑战也是其他应用领域以及 IT 业界所共同面临的。China-VO 将本着“有所为有所不为”的方针，主要采用移植现成技术的方法将其他领域的成功案例应用到 VO 中，服务于天文学。China-VO 将突出自身的特色，把研究与开发的重点放在光谱巡天数据的自动处理与分析领域。同时，China-VO 将与中国国家网格（CNGrid）紧密合作，利用其强大的计算和网络资源为用户提供高性能计算服务。

China-VO 的用户分为专业用户、非专业用户和特殊用户三类。其中专业用户是 China-VO 服务的主体，非专业用户是 China-VO 社会价值的主要体现，特殊用户是 China-VO 的管理者和开发者。China-VO 将根据不同用户的特点和需求为其提供不同的资源、服务以及访问界面。

国内大部分的非专业用户都希望能以自己的母语使用 China-VO。为了方便这些用户的使用， China-VO 需要对部分 IVO 优秀的资源与服务进行本地化。同时，为了将一些珍贵的中文资料与 IVO 共享，China-VO 也需要在国际化方面开展一些工作。

在 VO 门户的开发与设计过程中要采用结构化、模块化的方式，使得不同 VO 服务提供者的服务模块可以方便的融合到 China-VO 的门户中；为用户提供个性化服务，让用户可以方便的定值自己的访问环境，让每个用户都感觉 China-VO 是属于自己的。

目标是美好的，但任务是艰巨的。网格技术目前还没有成熟的标准和产品。网格环境下的海量数据挖掘、可视化、计算、资源注册与发现等技术的研

究还都处于起步阶段。虚拟天文台建立在如此不稳定的基础上，要实现美好的目标需要付出许多的努力。但我们相信，VO定会将天文学、计算机科学、数理统计等多个领域科学家的智慧凝聚在一起，推动相关技术取得突破性的发展，让深邃的宇宙奥秘就在你我的指尖上揭开[*]。

China-VO 为中国天文界带来了挑战，但更多的是机遇。知识创新逐步深化，大型观测项目不断启动，IT 技术飞速发展，在这历史性的机遇面前，中国科学家们应该也一定能够借助 China-VO 这座桥梁为 IVO 的建设与发展做出自己应有的贡献。

[*] Geoff Brumfiel. The Heavens at Your Fingertips. Nature, 2002 (420): 262-264

缩略语表

缩写	英文	中文
ADASS	Astronomical Data Analysis Software and Systems	天文数据分析软件与系统
ADC	Astronomical Data Center	（美国宇航局）天文数据中心
AIML	Astronomical Instrument Markup Language	天文仪器标记语言
ANSI	American National Standards Institute	美国国家标准化组织
API	Application Programming Interface	应用编程接口
ASP	Active Server Pages	动态服务器网页
ASP	Application Service Provider	应用服务提供者
ATNF	Australia Telescope National Facility	澳大利亚望远镜国家设施
Aus-VO	Australian Virtual Observatory	澳大利亚虚拟天文台
AVO	Astrophysical Virtual Observatory	（欧洲）天体物理虚拟天文台
B2B	Business to Business	企业对企业
BATC	Beijing - Arizona - Taiwan - Connecticut	BATC 大视场 CCD 多色巡天
BBS	Bulletin Board System	电子布告栏系统
BP 算法	Backpropagation Algorithm	反向传播算法
CADC	Canadian Astronomy Data Centre	加拿大天文数据中心
CAS	Community Authorization Service	公共授权服务
CCD	Charge-Coupled Device	电荷耦合器件
CDS	Centre de Données astronomiques de Strasbourg	法国斯特拉斯堡天文数据中心
CERN	European Organization for Nuclear Research	欧洲粒子物理研究中心
CGI	Common Gateway Interface	公共网关接口
China-VO	Chinese Virtual Observatory	中国虚拟天文台

缩写（续）	英文（续）	中文（续）
CIFS	Common Internet File System	公共网络文件系统
CNNIC	China Internet Network Information Center	中国互联网络信息中心
C/S	Client/Server	客户端/服务器模式
CSIRO	Commonwealth Scientific & Industrial Research Organization	（澳大利亚）联邦科学与工业研究组织
CSS	Cascading Style Sheets	层叠式样式表
CVO	Canadian Virtual Observatory	加拿大虚拟天文台
DAIS	Database Access and Intergration Services	数据库访问与集成服务
DAL	Data Access Layer	数据访问层
DAS	Direct Attached Storage	直接附属存储
DBMS	Database Management system	数据库管理系统
DEC	Declination	赤纬
DNS	Domain Name System	域名系统
DOM	Document Object Model	文档对象模型
DSP	Data Service Provider	数据服务提供者
DTD	Document Type Definition	文档类型定义
ESA	European Space Agency	欧洲太空局
ESO	European Southern Observatory	欧洲南方天文台
FAST	Five hundred meter Aperture Spherical Telescope	500 米口径射电望远镜
FITS	Flexible Image Transport System	灵活图像传输系统
FTP	File Transfer Protocol	文件传输协议
GAVO	German Astrophysical Virtual Observatory	德国天体物理虚拟天文台
GDVS	Grid Data Virtualization Services	网格数据虚拟化服务
GGF	Global Grid Forum	全球网格论坛
GGG	Great Global Grid	超级全球网格

缩写（续）	英文（续）	中文（续）
GIS	Geographic Information System	地理信息系统
GLU	Générateur de Liens Uniformes	统一链接生成器
GRAM	Grid Resource Allocation and Management	网格资源分配和管理服务
GSH	Grid Service Handle	网格服务句柄
GSR	Grid Service Reference	网格服务参考
GSC	Guide Star Catalog	导星星表
GT3	Globus Toolkit 3.0	Globus 网格工具集 3.0 版
HEALPix	Hierarchical Equal Area isoLatitude Pixelisation	多级等面积同纬度（天区）划分法
HTM	Hierarchical Triangular Mesh	多级三角（天区）划分法
HTML	Hypertext Markup Language	超文本标记语言
HTTP	Hypertext Transfer Protocol	超文本传输协议
IDGAR	Italian Data Grid for Astrophysical Research	意大利天体物理研究数据网格
IDHA	Images Distribuées Hétérogènes pour l'Astronomie	天文分布式异构图像
IDL	Interactive Data Language	交互式数据语言
IEEE	Institute of Electrical and Electronics Engineers	美国电气电子工程师协会
INAF	Istituto Nazionale di Astrofisica	（意大利）国家天体物理研究所
IP	Internet Protocol	网际协议
iSCSI	Internet SCSI	网络 SCSI
ISO	International Organization for Standardization	国际标准化组织
ISP	Internet Service Provider	网络服务提供商
IUCAA	Inter-University Centre for Astronomy and Astrophysics	（印度）大学联合天文学与天体物理中心
IVO	International Virtual Observatory	国际虚拟天文台
IVOA	International Virtual Observatory Alliance	国际虚拟天文台联盟
JVO	Japanese Virtual Observatory	日本虚拟天文台

缩写（续）	英文（续）	中文（续）
KADC	Korean Astronomical Data Center	韩国天文数据中心
KAO	Korea Astronomy Observatory	韩国天文台
KVO	Korean Virtual Observatory	韩国虚拟天文台
LAMOST	Large Sky Area Multi-Object Fiber Spectroscopic Telescope	大天区面积多目标光纤光谱望远镜
LSST	Large-aperture Synoptic Survey Telescope	大口径综合巡天望远镜
LVQ	Learning Vector Quantization	学习矢量量化方法
MathML	Mathematical Markup Language	数学标记语言
MIME	Multipurpose Internet Mail Extensions	多用途网络邮件扩充协议
NAOC	Chinese National Astronomical Observatory	中国国家天文台
NAOJ	National Astronomical Observatory of Japan	日本国立天文台
NAS	Network Attached Storage	网络附加存储
NASA	National Aeronautics and Space Administration	美国宇航局
NFS	Network File System	网络文件系统
NVO	National Virtual Observatory	（美国）国家虚拟天文台
OAI	Open Archives Initiative	开放式文档动议
OGSI	Open Grid Services Infrastructure	开放网格服务基础设施
OGSA	Open Grid Services Architecture	开放网格服务架构
P2P	Peer to Peer	对等计算
PHP	Hypertext Preprocessor	超文本预处理器
PMH	Protocol for Metadata Harvesting	元数据获取标准
POSS	Palomar Observatory Sky Survey	帕洛马巡天计划
PPARC	Particle Physics and Astronomy Research Council	（英国）粒子物理与天文研究委员会
RA	Right ascension	赤经
RAID	Redundant Arrays of Independent Disks	独立冗余磁盘阵列

缩写（续）	英文（续）	中文（续）
RDF	Resource Description Framework	资源描述基础架构
RFT	Reliable File Transfer Service	可靠文件传输服务
RVO	Russian Virtual Observatory	俄罗斯虚拟天文台
SAN	Storage Area Network	存储区域网
SAS	Server Attached Storage	服务器附属存储
SCSI	Small Computer Standard Interface	小型计算机标准接口
SDSC	San Diego Supercomputer Center	圣地亚哥超级计算机中心
SDSS	Sloan Digital Sky Survey	斯隆数字巡天
SGML	Standard Generalized Markup Language	标准通用标记语言
SIAP	Simple Image Access Protocol	简单图像访问协议
SMB	Server Message Block	服务器消息块
SMTP	Simple Mail Transfer Protocol	简单邮件传输协议
SOAP	Simple Object Access Protocol	简单对象访问协议
SQL	Standard Query Language	标准查询语言
SRB	Storage Resource Broker	存储资源中介
SST	Solar Space Telescope	太阳空间望远镜
SVM	Support Vector Machines	支持向量机
TCP	Transfer Control Protocol	传输控制协议
UCD	Unified Content Descriptor	统一内容描述
UDDI	Universal Description, Discovery and Integration	统一描述、发现和集成标准
UP	Unified Process methodology	统一过程方法
URI	Unified Resource Identifier	统一资源描述符
URL	Uniform Resource Locator	统一资源定位器
VLBI	Very Long Baseline Interferometer	甚长基干涉仪

缩写（续）	英文（续）	中文（续）
VO	Virtual Observatory	虚拟天文台
VO	Virtual Organization	虚拟组织
VO-India	Virtual Observatory India	印度虚拟天文台
VOQL	Virtual Observatory Query Language	虚拟天文台查询语言
W3C	World Wide Web Consortium	万维网联盟
WSCL	Web Services Conversation Language	Web 服务会话语言
WSDL	Web Services Description Language	Web 服务描述语言
XHTML	Extensible Hypertext Markup Language	可扩展超文本标记语言
XML	Extensible Markup Language	可扩展标记语言
XPath	XML Path Language	XML 路径语言
XQuery	XML Query Language	XML 查询语言
XSL	Extensible Stylesheet Language	可扩展样式表语言
XSLT	XSL Transformations	XSL 转换语言

发表文章目录

专业文章

1. Chenzhou CUI, Yuqin CHEN, Gang ZHAO, Yongheng ZHAO. Abundance Gradients in the Galactic Disk. Science in China (Series G), 2003 (1): 52-61
2. 崔辰州, 赵永恒, 赵刚, 张彦霞. 虚拟天文台的技术进展. 天文学进展, 2002 (4): 302-311
3. 崔辰州, 赵永恒. 虚拟天文台和网格技术. In: 孙九龄, 施慧中. *科学数据——管理与共享*. 第一版. 北京: 中国科学技术出版社, 2002. 272-290
4. 张彦霞, 赵永恒, 崔辰州. 天文学中的数据挖掘和知识发现. 天文学进展, 2002 (4): 312- 323
5. Yanxia ZHANG, Chenzhou CUI, Yongheng ZHAO. Classification of AGNs from Stars and Normal Galaxies by Support Vector Machines. In: Jean-Luc Starck, Fionn D. Murtagh. *Astronomical Data Analysis II*. Proc. of SPIE, 2002. 371-178
6. Chenzhou CUI, Yongheng ZHAO. Grid Based Chinese Virtual Observatory System Design. IAU 25th GA. Accepted.
7. Chenzhou CUI, Yongheng ZHAO. The Most Popular Web Server for Astronomy Education in China. IAU 25th GA. Accepted.

科普文章

8. 崔辰州, 苏丽颖. 登陆火星. 第一版. 石家庄: 河北少儿出版社, 2003
9. 崔辰州, 苏丽颖. 走进太空——航天知识百问. 石家庄: 河北少儿出版社, 2003
10. 崔辰州. 天文软件概述. 天文馆研究, 2001 (71&72): 16-17
11. 崔辰州. 虚拟天文台——天文学的新革命. 天文爱好者, 2001 (5): 24-26
12. 崔辰州. 互联网上的港台天文. 天文爱好者, 2001 (2): 12-14
13. 崔辰州. 国内天文网站纵览. 天文爱好者, 2001 (1): 12-14

翻译文章

14. 崔辰州. 虚拟天文台. 世界科学, 2003 (2): 16-18
15. 王二超, 崔辰州. 不断走向精确的宇宙学. 世界科学, 2003 (5): 2-5

其他工作

16. 中国虚拟天文台. http://www.china-vo.org
17. 世界数据中心天文中心. http://badc.lamost.org
18. 全国主要城市标准国旗升降时间. http://sunriseset.lamost.org
19. 业余天文网络服务器. http://amateur.lamost.org
20. 中国天文网络与软件. http://www.lamost.org/amateur/

致　　谢

1997 年当我刚来到北京天文台的时候一些老师问我是哪年出生的，我说是 76 年。当时他们都感叹道“好年轻！”。“多年的媳妇熬成婆”，一晃 6 年的时间过去了，我已经成了国家天文台资格最老的学生之一。每年的春季和秋季都会送走几个老友迎来许多新朋。当年与我一同步入天文台大门的四位学友，如今，一位求学于太平洋彼岸，一位忙碌于欧亚大陆的另一端，一位融于首善之区忙忙碌碌的上班人流，还有一位仍在国家天文台与我并肩战斗。

在此，我首先要感谢的是自己的导师赵永恒研究员。赵老师知识渊博，待人热诚、宽厚。他博采众长，善于纳新，能很好把握天文学的发展动向。在天文台六年的学习期间，我一直得益于赵老师的谆谆教诲。特别是 2000 年硕士毕业后，我的学习、工作更是在赵老师的亲手指导下进行。在国内他率先洞察到虚拟天文台将为天文学带来一场重大变革，于是引导我走上了虚拟天文台的研究之路。几年来，赵老师不仅直接指导我的学习和研究，还为我提供了许多宝贵的锻炼机会，使自己得到全面的训练和发展。这篇论文的完成凝聚了赵老师很多的心血。在自己学生生涯即将结束之际，我要衷心的向赵老师道一声“辛苦了”。

感谢国家天文台赵刚研究员。他是我的硕士生导师，是他把我领入了科学研究的神圣殿堂。自己六年的求学生涯离不开赵刚老师一贯的关心和指点。

感谢国家天文台研究生管理办公室杜红荣老师。她的无私奉献为天文台的研究生创造了优越的生活和学习环境。几年来，杜红荣老师一直对我的生活、学习、工作给予热情的关心和支持。我今天的收获离不开杜老师的亲切关怀。

感谢国家天文台陈玉琴博士。论文的顺利完成离不开与她真诚、愉快的合作。

感谢中国科学技术大学天体物理中心褚耀泉教授。作为参加第一次国际虚拟天文台会议（Caltech, 2000）的唯一中国代表和 LAMOST 科学部主任，他一直关注和支持着我在虚拟天文台方面的研究工作。

感谢与我一同进行中国虚拟天文台研发工作的桑健、邵惠娟同学。与他们经常性的讨论和切磋使自己的工作得以不断进步和提高。

感谢中科院软件所徐志伟研究员、刘利民博士，中科院计算机网络信息中心南凯、罗泽博士，清华大学计算机系刘鹏博士，国防科技大学肖侬教授，江

南计算所谢向辉研究员。他们在网格技术方面给予我许多指导和帮助。China-VO 的成功离不开这些重要的网格节点。

感谢美国微软旧金山湾研究院院长 Jim Gray 博士。他为我的工作提出了非常有益的建议。与他的交往让我懂得了许多做科学的道理。

感谢国家天文台胡景耀、南仁东、彭勃、武向平、邓李才、魏建彦、汪景琇、叶彬浔、周旭、薛随建、李启斌、韩金林研究员。他们对我的研究工作给予了许多帮助，提出了很好的建议。与他们的讨论使我受益匪浅。

感谢 LAMOST 大科学工程项目组中曾经和正在与我一同生活、学习、工作的同学们。有他们的陪伴，自己仿佛置身于一个和睦的大家庭中，总有笑声和欢乐伴随着。他们是罗阿里、陈建军、孙浩峰、张彦霞、覃冬梅、邱波、程林鹏、王伟、汴维豪、王辉、朱光华、吴潮、张昊彤、李博、贾磊、赵瑞珍、许馨、刘中田。

感谢国家天文台 LAMOST 总部的老师和同事。他们是苏洪均总经理、王刚研究员、陈英老师、袁辉老师、孙盛慈老师、石火明博士、李有刚老师、李祈老师、冯磊师傅、门力老师。

感谢国家天文台许多帮助过我的老师、同学和同事。他们是郑宪忠、苏彦、梁艳春、施建荣、王俊杰、王菲鹿、李冀、朱镇熹、姜晓军、陆烨、杨克时、孙超、赵景芝、朱爱萍。

感谢北京师范大学一直关心着我的李宗伟教授、刘学富教授、姜碧沩教授。感谢北师大热心帮助过我的陈黎教授。

感谢北京大学吴学兵教授、张华伟博士，北京天文馆馆长朱进研究员。

感谢国家天文台多年来一直对我在团委和学生会的工作给予大力支持的蒋协助书记、王春秋老师、耿丽老师、董素珍老师、马燕霞老师、王强老师、王志华师傅、方明俊。感谢中科院团委、京区各单位团委，特别是北郊片各兄弟单位团委的青年朋友们。“年轻的朋友在一起，比什么都快乐！”

感谢悉尼大学张承民博士。他是带我走进职业天文研究领域的引路人，多年来一直关心着我的发展。

感谢远在大洋彼岸的刘宏霞、陈锐夫妇。他们多年来一直关心和支持着我。

感谢我的爱人，苏丽颖。无需千言万语，一切尽在“爱”中。

我要向自己的父母表达最真挚的感激之情和无尽的谢意。二十多年来，父母在自己的身上注入了无穷的关爱，付出了太多的心血。没有他们时时的牵挂

就没有我今天的成就。

在此，我要将本论文献给最疼爱自己的祖父。十二年朝夕相伴，祖父给了我幸福快乐的童年。多年来，祖父无时不挂念着在外地读书的我，哪怕是在最后病重之时。一年半前，他离我远去。

正当我全力以赴准备这篇毕业论文的时候，首都北京正经历着一个非常特殊的春天，进行着一场没有硝烟的战争。这里，我要衷心的感谢和祝愿战斗在抗击“SARS”第一线的白衣天使们，感谢在这样的“非典”时期仍在岗位上坚持工作的同胞们。没有你们忘我的工作，我将不能安心和顺利地完成自己的论文。