白皮书

迈向国家虚拟天文台：
科学目标、技术挑战和实施计划

(草案)

2000 年 6 月 8 日

摘　　要

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。国家虚拟天文台将把空间和地面观测的数据库、多波段巡天、以及支撑利用这些资料进行相互对比和交叉关联所需要的计算资源连接在一起。这本白皮书描述了国家虚拟天文台的科学机遇和技术挑战，以及以高效益的投资实现国家虚拟天文台目标的实施方略。国家虚拟天文台依靠不同机构间的合作、进行分布式的发展和运行。它将对天文学界提出挑战，但也为几年前不可想象的新的科学发现提供机会。

执行概要

近十年来望远镜及仪器设计上的技术进步，加上计算机和通讯能力的提高，导致了天文研究的特征有了令人注目的和不可逆转的变化。空间和地面上波段从射电到 X 射线的大规模巡天已经开始，正在产生大量优质而不能替代的数据。

这些新的巡天带来的科学发现的潜力是巨大的。对这些数据集的联合使用，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。然而其规模之大和数据之复杂需要使用各种工具和恰当结构去发现蕴藏在里面的复杂现象。我们建议建立一个国家虚拟天文台，协调现有的分散努力，并集中发展现在尚不具备的功能来满足这种需要。在未来十年，国家虚拟天文台将作为一个促进与协调的实体，为实现天文数据库的全部科学潜在价值，发展各种必要的手段工具，制定协议，组织合作。全部实施后，国家虚拟天文台将成为天文发现的发动机。

国家虚拟天文台所能达到的新的科学能力,是实现已有的或即将产生的兆兆字节至千兆兆字节数据集的全部科学价值的关键。大规模星表的快速检索、统计关联的建立、数据中新的结构图案和时间变化的发现，以及与复杂的数值模拟的比对，都是通向新科学的途径，它们都可以通过国家虚拟天文台实现。此外，国家虚拟天文台及其数据档案需要与计算机科学界的合作，并提供与面临类似挑战的其它学科合作的机会，它还为教育和普及服务。以活动星系核、宇宙大尺度结构和银河系结构三个科学问题为例子就可以说明国家虚拟天文台的科学前景。国家虚拟天文台将是技术上可行的，但是以科学来推动的。

国家虚拟天文台的实施涉及到许多重大的技术挑战。这包括现有的天文数据归档努力的协调以及新能力和新结构的发展。国家虚拟天文台的主要技术组成包括数据档案、元数据标准、数据访问层、查询和计算服务，以及数据挖掘应用。这些能力的发展要求与信息技术界的密切合作。

国家虚拟天文台的实施计划分为四个阶段，始于国家虚拟天文台建立前的启动阶段，直至其完全开放运行阶段，在计划批准后要花四到五年时间。本实施计划的设计是把国家虚拟天文台一旦可交用的结果和能力尽早地交给天文界，这种立竿见影的做法对项目的成功是至关重要的。

第一部分 引言：变革之风

两百多年来，天文研究通常都是单个天文学家或者天文学家小组进行小数目天体的观测。在过去，天文学家花整个一生所获得的资料仅能勉强得出有统计意义的结论。加上威力大的设备观测时间非常有限，那些需要大量数据来解决的天体物理问题就无法进行研究。

现在情况正发生着戏剧性的迅速变化。这一变化是由过去十年中发生的前所未有的技术进展所推动的。这场天文学上革命性变化的主要领域集中在以下几个方面：望远镜的设计和制造、大尺寸探测器阵列的开发、计算能力的指数增长以及通信网络覆盖和容量的不断增长。

望远镜设计和制造的进步使得大型空基天文台成为可能，为伽玛射线、X 射线、光学和红外天文的发展开辟了新的前景。技术的进步也使得新一代的大口径地面光学和红外望远镜以及毫米波与厘米波单天线和多天线阵的建造成为可能。在光学和红外波段，这些进步已经与极灵敏的、高分辨率而尺寸不断增大的探测器阵列结合起来。拼接这些阵列的能力使得仪器的视场达到 30 角分，每幅图像具有 10^8 像元。随着这些技术进步不断成熟，功能更好、口径更大的空间和地面望远镜正在计划中，这些设备配有使用面积更大、像元更多的探测器的先进仪器。如同 Moore 定律反映计算能力随时间指数增长，在过去十年中实测天文技术的进步使得天文发展实际也要用 **Moore** 定律来描述。

更多的地面和空间大口径和复杂仪器的天文设备的出现必将产生一个关键而必然的结果：数据流极大地增加。比如目前哈勃空间望远镜(HST)每天大约产生 5 千兆字节的数据，而天文学和天体物理学发展规划委员会十年规划最近推荐建造的一项设备（大口径全景巡天望远镜，LSST）每天将产生 10 兆兆字节 的数据 !

除了数据增长速度外，进行观测的方式也有了变化。尽管新的地面和空间天文台还将继续分配给单观测者/单计划的观测模式的研究相当多的观测时间，这些时间被分成许多小时段安排给许多特定目标的研究课题，但更多的时间是用来进行大规模的巡天，经常是多波段的，涉及大量的合作者。

这些大的巡天计划将会产生大量质量均匀、标准统一的数据，通常会以兆兆字节来衡量。这种天文研究模式的变化，不仅由于新设备威力的提高允许更快速地获得这些数据，还由于计算机的硬件和软件可以快速地进行数据采集、处理及存档。

将使天文学研究的特征发生变化的一个主要的技术进步是宽带高速网络信息交换技术。

虽然大量数据通过通常的网络进行交换还慢得不能接受（传输 1 兆兆字节的数据集要花 20 多天），但是未来的网络会快得多。这种数据传输速率加上地面和空间设备的高效率的数据采集，使得在不同地点间的大量数据交换成为可能。不像以前，这一技术将使广大用户能够访问使用其中特定部分的数据，其潜在科学产出将是巨大的。

显然，这些技术上的驱动今后几年内会导致前所未有的天文数据流。而且这些数据集与以往不同，大部分是质量均匀的、常常在多个波段上并覆盖相当大的天区。其内容的丰富和深度是空前的，因而能为广大用户提供独一无二的机会去应用这些数据从事各种科学问题的研究。仅凭这一点，数据的系统存档就非常重要。此外，这些数据是使用昂贵的、而技术先进的设备获得的，而大大超额的对设备使用的申请，也不允许重复已经做过的观测，所有这些都要求对数据普遍存档。

面对这样包含数亿个源的多波段数据档案，天文学界显然需要访问这些档案和分析其中数据的工具。数据挖掘、先进的模式识别、大规模统计交叉关联、罕见天体以及时变的发现等机会，显然都是存在的。

此外，有了这样的数据集，将在天文学历史上第一次可以将复杂的数值模拟和统计上完备的多变量的数据进行有意义的对比。高速和广泛分布的网络的迅速进步意味着美国和其它国家的天文界都可以分享这些科学努力的成果。

过去几年这些技术进步会合在一起，它将完全改变现行的大部分观测天文学的工作方式。这些变化是不可避免和不可逆转的，它们对天文社会学本身将产生巨大的影响。无论在美国和国外，大家都越来越意识到，科学数据的获得、组织、分析以及传播对科学和技术持续而坚实的发展都是基本的要素。所有这些都表明，建立一种结构去有效地综合这些技术能力势在必行。因此现在就需要一个象国家虚拟天文台这样的实体来指导日益增多的天文数据的合理部署。

第二部分 设想的国家虚拟天文台

2.1 结构和功能：产生新科学

国家虚拟天文台是这样一个机构，它将使天文学和天体物理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。国家虚拟天文台将是独一无二的，它将兆兆字节的数据档案、波长遍及从伽玛射线到射电波段的成百万个天体的图像库、高度复杂的数据挖掘和分析工具、可访问带有千兆兆字节存储容量的每秒运算次数达到万亿次的超级计算设备、以及在各主要天文中心间的极高速的互联，连成一体。它使成千上万的研究者可以快速查询各个达兆兆字节大小的档案；使埋藏在庞大星表和图像数据库中的多变量图案可视化；增加发现复杂图案和稀有现象的机会；鼓励多个研究小组间的实时合作；允许进行大型的统计研究，这些研究将首次使数据库的内容和复杂精密的数值模拟结果进行对比。国家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、更好的科学。国家虚拟天文台将作为一个协调性的和操作性的机构促进工具、协议和合作方面的发展，去充分实现未来十年内天文数据库的科学潜能。国家虚拟天文台将成为“天文学发现”的一个发动机。

为了达到上述目标，首要的是要将国家虚拟天文台的建设作为一个科学驱动的、天文学界共同的努力；大部分投资将分在各处而由同行评议决定。建设工作将通过对项目的常规招

标来实施，既包括用于开发国家虚拟天文台基础设施的软件项目，也包括利用国家虚拟天文台来进行的科学活动的项目。更具体的说，国家虚拟天文台为发挥其作用要开展以下活动：

- 建立可以使用数据流水线、存档和读取过程的公共系统，来保证天文界大量分散的用户进行经济而高速的访问；
- 促成分布式开发一套通用的新软件工具，来进行查询、关联、可视化和上文提及的统计对比；
- 协调建立高速的数据传输网络，这个网络对连接数据档案、每秒万亿次的超级计算设备和广泛分布的用户界是必需的；
- 促进国内和国际的天文中心和主要学术机构间高产的合作，以用最少的基础设施费用产出最多的成果；
- 确保与面临类似问题的其它学科的科学家以及私立机构间的交流和可能的合作；
- 坚持一项持续性的普及和教育计划，利用国家虚拟天文台独一无二的资源,包括数据和软件，为其提供一个了解天文学和科学方法论的独特的窗口。

2.2 设计理念

国家虚拟天文台将成为一个独一无二的机构，主要是因为它的运作将是分布式的,而且是基于通信和计算机科学高速发展的技术。为了确保其持续的生命力，国家虚拟天文台必须要包含几个主旨思想：

- 国家虚拟天文台必须是进化的。自创建开始，进化的特性就要使它能快速回应不断变化的技术上和科学上的机遇以及天文界的需求。由于技术能力的不断提高，这种进化的特性将成为国家虚拟天文台自始至终不可或缺的组成部分。这种灵活应对的策略是管理其分布式开发成果和迅速开拓新机遇的管理结构所需要的。管理和监督必须是有效的、经济的、有远见的和对天文界负责的，还必须将日常开支和机构惰性减少到最低程度。

- 国家虚拟天文台本质上必须是分布式的。相当多的专家在已有的各个中心工作，从一开始就要充分利用这一优势。进一步说，迈向国家虚拟天文台目标最为经济有效的进展也许在其运行阶段就应该采取分布式的途径实现。这会要那些已经存在的中心和未来的数据中心承担国家虚拟天文台功能的关键部分，因为在这些中心做最为有效。

- 国家虚拟天文台必须是集成的。集成将会是作为分布式结构的补充而伴随国家虚拟天文台一直存在的主题。为了使其能极为有效地促进科学进步，在全波段上、在空基的和地基的设备上，信息技术功能必须是集成的。另外，与计算机科学和信息技术的集成开发将成为国家虚拟天文台的基本要素之一。

- 国家虚拟天文台必须为公众服务。通过国家虚拟天文台可以得到巨大的数据集和附带的分析工具，它们将给教育领域和公众领域提供一个前所未有的机会。一个充满活力的为公众服务的计划会发掘出国家虚拟天文台教育潜力的全部优势，这个计划将贯彻国家虚拟天文台发展的各个阶段。

- 国家虚拟天文台必须面向全球需要。国家虚拟天文台始终要保持与其它国家类似努力的国际联结。虽然国家虚拟天文台在最初不会是一个国际性合作机构，但是明显的是它必须与其它类似的研究工作保持交流，并在适当的时候进行各种层次的合作。看起来不可避免的是，国家虚拟天文台发起的工作将成为一种世界性的活动。

- 国家虚拟天文台必须提供一条通向未来的道路。这里大致描述一下国家虚拟天文台的远景：它将是一个起催化和促进功能的机构，拥有最小的结构和庞大的连接。这种机构的直接产物将使天文学的科学产出提升到一个新的水平。当然，国家虚拟天文台更大的、也许更为持久的遗产将是它在建立一个美国的和世界的天文学信息基础设施方面所起的作用。为国家虚拟天文台所力促的基础设施的增长，将提供给未来的天文学研究空前的新前景和机遇。

第三部分 国家虚拟天文台的科学应用案例

我们看到，天文界已准备好通过以下两种途径来充分利用计算速度、存储介质和探测技术上的突破性进展：（1）开展新一代的巡天，涵盖多波段，并充分利用以上进展；（2）发展软件工具以实现从这些巡天的数兆兆字节（今后的千兆兆字节）数据库中得到新发现。新一代的巡天和软件工具结合起来就为促使产出质上不同的科学提供了基础。

更为重要的是，这些数据库固有的丰富内容为研究工作提供了远远超过巡天本身目的的科学价值：比如，旨在探索柯伊伯带天体的反复成像巡天可以为发现 z>1 的超新星提供依据。确实，巡天数据库的倍增效应是巨大的，由哈勃深场（HDF）引发的世界范围的研究探索就是一个极好的例证。

现在，我们已经能在各种空间尺度上进行几乎覆盖整个电磁波谱的巡天，它们都有明确的选择判据和完全清楚的局限性。产生宇宙全色图像和某些情况下宇宙数字电影的能力，为我们提供了空前机会去发现那些能彻底改变我们对宇宙认识的新现象和新图景。过去，同一天区的光学和射电的多色图像带来了类星体的发现；应用红外数据我们发现了可见光图像中被遮掩的活动星系核以及恒星形成区；反复观测导致瞬变现象（例如超新星和最近的微引力透镜事件）的发现，以及对变化现象更深入的了解；不同尺度的大规模数字巡天联合起来将可能对参量空间进行新的探索，例如全波段低面亮度宇宙的探索。

在海量天文数据库中发现新图案和新现象所遇到的挑战同样也出现在医学、生物学和地球科学中。例如，人类基因组数据大小约为 3 千兆字节，而整个天区的数字巡天的结果约为 10 兆兆字节。处理这个量级的天文数据库的工具和技术的发展明显需要计算机新技术的支持，而当这些工具和技术发展之后，也会应用在天文之外的科学领域中。所有这些，如果没有新的工具和新的研究结构来组织那些分散的数据库和星表，提供对它们的检索，并把分析工具放在广大富有想象力的科学家手里，这些数据库的所有威力就不可能得以发挥和实现。正是这种考虑推动了国家虚拟天文台的产生。

为了实现其科学目标，国家虚拟天文台的主要功能应包括：

- 把现有的多波段的大型数据库联合起来，创造工具实现对星表类的以及图像像元类的数据查询；
- 为建构未来的大型数据库发展通用标准；
- 为新的数据库提供兼容的架构，以尽量减少新的巡天和观测的代价，同时从中得到最大的科学回报；
- 发展分析工具，以在星表数据集中进行发现和对联合数据集作统计分析；

- 发展在图像数据库中进行天体分类的工具；
- 发展实现星表和图像数据库可视化的工具；
- 为查询图像数据库，图像分析以及模式识别发展新方法；
- 综合数值模拟的结果并为观测数据与模拟结果的对比提供“工具箱”；
- 与现有和未来的数字图书馆与期刊进行链接。

以上所有功能都是可能实现的，但是由于数量、维数与复杂程度的不同将与我们现在的做法有质的区别。国家虚拟天文台将逐步实现这些功能。虽然国家虚拟天文台是由技术所实现，但决不是被技术所驱动的。相反，它的结构和发展从根本上是为科学和科学界的需求所驱动的。天文界为国家虚拟天文台提出的“科学任务指南”（“Science Reference Mission”，SRM）将是国家虚拟天文台发展一项功能和实现这些功能的步骤的主要指导决策文件。我们预想为国家虚拟天文台编制科学任务指南的步骤如下：

- 2000 年 6 月 13－16 日在 Pasadena 举办工作会议，以在天文界中开展广泛讨论。
- 在 Pasadena 会议上确定的多个工作组中进行讨论，各组要提出：
  1. 能被上述技术可能性实现的主要科学项目的细节；
  2. 阐明这些科学项目的意义，并讨论如果没有国家虚拟天文台要实现同样结果的困难；
  3. 理解从科学项目的需要到文档、档案访问和所需的软件工具的整个过程；
  4. 排出各种需要的优先级。
- 举行各工作组主席的会议，并由国家虚拟天文台临时指导委员会把工作组讨论结果综合成统一的科学任务指南，形成一个国家虚拟天文台从科学到需求的全过程计划。

在这个过程的准备之中，我们已提出了几个科学项目的例子，让人们领略一下国家虚拟天文台可能作出的发现，以及为其实施而确定的框架和基调的初步基础。请注意这些例子还不完整，还需要相当的努力以便令人信服地展示国家虚拟天文台蕴藏的威力和所需的功能。在列举于此的材料中，我们强调如何从科学到功能的过程，因为这一步对于完整理解国家虚拟天文台是必要的。

例 1 活动星系核的全波段研究

背景： 一方面因为活动星系核的高光度使他们在很大的宇宙距离上可见，另一方面因为活动星系核代表了星系演化的一个基本阶段，所以对活动星系核本质的了解非常重要。在观测上，由于活动星系核的光谱能量分布比一般黑体谱宽很多，易于从恒星中分辨出来。但是，红移、光变、遮挡（可能很大）以及内禀光谱形状与特征的不同导致从 X 射线到射电波段的“颜色”有很大的范围。

科学目标：这个课题旨在建立一个完整的活动星系核样本以便 ：

- 检验它们的观测特性是否由外在因素，如指向或遮挡所决定（所谓统一模型）；
- 比较星系环境对活动星系核类型的影响，例如其射电特征与作为星系团成员之间的关系；
- 了解活动星系核光度函数的演化，特别是区分密度演化和光度演化；
- 在不同波段上建立活动星系核光度函数，由此来了解在统计意义上活动星系核特性的演化过程。

项目概述：

1. **联合几（**N**）个巡天，它们覆盖同一（足够大）天区，并跨越很大的波长范围（从** X **射线到射电）。**
2. **包含元数据信息，以便考虑巡天选择效应及其局限性。相关的元数据将包括巡天天区、波段范围、流量极限等等。**
3. **在合成后得到的** N **维多色空间中认证可以辨别的天体“云”。**
4. **利用已有的天文知识（比如发表的星表、理论模型）以了解这些“云”的成分。**
5. **在联合的数据集中证认活动星系核候选体。注意可能需要新的观测来证实，因为这种证认也许是统计性的，给出某一天体是活动星系核的几率。**
6. **利用星表的性质或从图像数据库的进一步的测量来解决科学问题。**

对国家虚拟天文台功能的需求：

- **联合有关巡天，包括不同波段巡天中天体的交叉证认以及元数据的交换与融合；**
- **用聚类分析方法认证“云”与“链”。包括监督分析（由天文知识指导定义与分 析）与非监督分析（识别新模式）；**
- **多维数据集的可视化；**
- **在多维参量空间对给定的天区进行天体组分的统计分析与分类。**

例 2 大尺度结构的形成与演化

背景：星系团代表了已知最大的明确的质量集中现象。不同的早期宇宙演化模型预言了不同的星系团的形成和演化。这些模型可以通过各种预言与观测到的星系团质量和光度函数进行比较来加以验证。

科学目标：这个项目的目标是建立无偏的、在宇宙学意义上有相当红移跨度的星系团样本，通过与观测到的质量和光度函数随红移演化的比较，来检验各种结构形成与演化的模型。另外，这些样本也可以用来研究星系的形态—密度关系及其在特定时间尺度上的演化。

项目概述：

1. **通过以下不同方法，利用多波段像元数据生成统计星系团样本：**
   X 射线巡天：从图像数据中的热气体发射证认星系团；
   光学/红外巡天：把图像数据与设计好的核函数进行卷积来挑选星系团；
   毫米波巡天：由 Sunyaev-Zel'dovich **效应引起的宇宙微波背景温度变化来证认星系团；**
   射电巡天：根据标志星系团环境的射电源形态来证认星系团。
2. **比较不同选择方法的结果，定量给出选择效应作为星系团质量、密度、红移等的函 数。**
3. **运用各种距离标志或者红移估计来补充星系团的观测特性。**

对国家虚拟天文台功能的需求：

- **根据用户定义的算法和工具对大量图像数据进行处理；**
- **模拟巡天以了解选择效应；用模拟检验用户定义的工具；**

- 根据不同的理论预测观测样本特性，并与观测样本作比较。

例 3 数字银河系

背景：银河系在结构上由多种成分组成，即晕、薄盘、厚盘、核球和旋臂。每一部分由在年龄、质量、化学组分、轨道（位置与运动）等分布上相关的星族，以及如气体和尘埃等非恒星物质的分布来表征。它们是反映银河系形成历史的化石，从整体上说，对这些结构的全面理解还从来没有过,因为要同时定出所有变量所需的样本实在太大,使得研究非常难于进行。

科学目标：这个项目要建立一个非常大的银河系恒星样本，包括尽可能多的关于每一个天体物理特性的信息。这些资料加上银河系非恒星物质的分布图将用来：

- 产生银河系参数化的模型，包括位置和运动信息；
- 把这个模型同基于不同形成过程的银河系结构模型进行比较；
- 特别要寻找代表并合事件或潮汐碎片痕迹的共动天体群。

项目概述：

1. 联合各种光学和红外巡天来产生出彼此匹配的星表。
2. 利用位置、星等和颜色建立三维恒星分布。这需要用颜色来得到光度型并估计消光。
3. 利用远红外、 HI 和 CO 图像，定量得出尘埃分布和遮挡。
4. 反复迭代项 2 和项 3，直至自洽。
5. 用红外和射电图像来确定恒星形成区的整体流动和地点。
6. 用自行巡天（如果有可能，加上视向速度信息）推导出恒星子集的运动。
7. 用多次历元的成像寻找变星，利用它们检验距离。

对国家虚拟天文台功能的需求：

- 联合多波段和多历元的星表数据；
- 以用户定义的工具对大量图像数据进行处理；
- 大量多维数据集的可视化工具；
- 成分分析及寻找相关天体群的统计工具。

我们再次强调这些项目在于演示这类过去难以实现的科学任务，在应用国家虚拟天文台的功能后是能够完成的。显然，随着国家虚拟天文台的逐步形成，天文界会通过更多的定义更周密、范围更广泛的科学项目来丰富它的内容。

第四部分 技术问题

4.1 概述

从北美天文学的现状评估中，可以看到用来支持新兴的国家虚拟天文台的如下资源已经到位：

- 数据中心和超级计算机中心：已有来自于各种空间计划、公用望远镜和巡天观测的数十兆兆字节的数据产品（星表、图像和光谱）；到本十年代末，数据将扩充至千兆兆字节

量级。NASA的几个主要中心（STScI、IPAC、HEASARC与CXC）和加拿大的CADC已经具备了数据归档和数据分析的能力；此外还有很多规模小一些的或更专门化的档案。象SDSC和NCSA这样的超级计算机中心可以用于解决涉及大规模计算的问题。而一个高性能的国家网络基础设施业已就位。

- 天文信息服务：已有信息服务如ADS、NED与SIMBAD可以提供有关河内和河外天体的命名办法和相互参照，并且可以提供文献目录信息、已发表的和预印本的文献以及档案数据中心之间的日益复杂的联结。

- 数据分析软件：已有各种软件包如AIPS、AIPS++、IRAF、IDL、FTOOLS、SkyView等可以用来进行天文数据的一般分析。虽然针对大规模数据挖掘的复杂软件的开发还处于萌芽阶段，但是诸如NPACI资助的“数字化天空”和IPAC的“红外科学档案”的一些开创性的工作以及对数据档案互联与挖掘技术方面的研究已经展现出了一定的潜力。

尽管这些资源很重要，但是如果有人打算对分布于各地的种类繁多的档案数据和各种星表进行多波段数据分析或者进行大规模统计研究，就会知道目前的状况与国家虚拟天文台的目标还相差多少。对于地面观测得到的光学红外和射电波段的数据，需要象现在对空间数据做的那样进行流水线式的处理和归档。需要为分布广泛的档案之间的数据交换和互操作开发相应的标准和协议。天文数据分析软件有待发展成为如同访问本地数据集那样方便地对分布在各处的不同波段的档案中数据进行访问。需要开发新的算法、应用程序与工具包来对数兆兆字节的数据档案进行挖掘。同样需要把超级计算机一级的计算系统发展成具备对海量的多波段数据档案进行大规模统计研究的能力。需要在可获得的最大网络带宽上将数据、软件与计算资源实施互联。

4.1.1 数据档案

无论是通过国家虚拟天文台来进行的科学研究，还是涉及国家虚拟天文台实施的有关技术问题，都必须从数据开始考虑。尽管在过去十多年里，NASA的多数空间装置所获得的数据都依惯例进行存档；但是地面望远镜观测得来的数据，除了几个主要的巡天观测外，目前还很少能够在线获取。随着现代大视场及多谱仪器在地面望远镜中的应用，将会产生更大量的数据；而基于地面的巡天观测将变得几乎同常规观测一样普通，从而迫切需要对地面仪器和巡天得到的高质量数据集进行存档并发表。如果国家虚拟天文台无法成功地产生真实的全色的图像和星表，并对地面和空间档案中的数据进行完美的整合，以用来对几乎整个电磁波谱中的天文现象进行探索，国家虚拟天文台所承诺的科学目标将无法实现。

过去十年的经验告诉我们，天文档案是复杂多样的，并且还在不断地增加，最好由那些接触并了解这些数据的人来维护。事实上，这意味着大多数在线数据，或者通过各个大的巡天项目，如2微米全天巡天（2MASS）或斯隆数字化巡天（SDSS），或者通过服务于某个特定团体的学科档案中心来提供。为满足将地面天文观测数据实现大规模存档的需要，有必要在国家中心（NOAO、NRAO、NSO、NAIC）建立存档设施，并且有必要与主要的私人及大学运行的设施建立伙伴关系。面向地面和空间的主要国家数据中心将构成国家虚拟天文台分布式数据系统的主干结点。

4.1.2 技术挑战

设想从天文学各个分支所获得的具有存档质量的数据分布在10到20个主要档案中心和许多辅助数据集里，而需要服务于分布各地的数以千计的科学用户界，就可以确定究竟需要什么样的新型软件和服务才能使国家虚拟天文台正常运转。由于来自天文学不同分支的数据集具有不同特性，并且由于为适应日益精密的现代天文仪器而采用了日益复杂的数据结构（在 FITS 数据格式标准的一般框架下），数据的分析将变得非常复杂。

该问题的规模就已令人吃惊：星表的大小会接近兆兆字节量级，而数据的总量将达到千兆兆字节量级。然而，更严峻的挑战来自于这些数据集的复杂性，因为面对的是上千万甚至更多的天体，每个天体又有几十或几百个属性。这对于数据挖掘而言是一个至关重要的新问题，对如此大的星表进行多变量相互关联将是一个极大规模计算的问题。如果还要对候选天体进行像元层次上的分析，计算的问题将更加突出。重要的是，要认识到目前蛮干的分析技术已不能继续推广去应付这种规模的问题！还需要对元数据的表示与处理、大规模统计分析与相互关联、以及分布式并行计算技术等进行多学科研究，来解决国家虚拟天文台所面临的前所未有的数据访问和计算问题。

好在不单单是天文学面临这个问题，诸如高能物理、计算基因学、全球气候研究和海洋学等其它科学分支也同国家虚拟天文台一样面临类似的技术挑战。信息系统技术的研究与开发已在一些领域中展开，如大型档案的统计分析与数据挖掘、分布式计算网格、数据集中网格计算（数据网格）、以及结构化数字信息管理（数字图书馆），这样的研究许多都和国家虚拟天文台所面临的问题有关。贯穿于这些科学分支中的信息技术和数据管理将会推动国家虚拟天文台，并为国家虚拟天文台所推动。

大规模的数据集和用户与资源的地理分布同样对互联性提出巨大挑战。能提供每秒100兆字节的洲际传输带宽的新一代网络业已投入使用、但尚未被充分利用，这种情况将很快地得到改变。对国家虚拟天文台来说非常重要的一点是，其主要数据中心必须用极高速的网络来建立互联，并使用智能型服务器端的软件代理，以便与最终用户进行交互时能够最有效地利用网络资源。

4.2 体系结构

实施国家虚拟天文台所面临的技术挑战是如何来解决相互对立的需求。一方面数据是广泛分布的，另一方面国家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本地访问；一方面要采用复杂的元数据标准和访问协议将分布的档案和网络服务连为一体，另一方面还要使小型档案接入国家虚拟天文台简便可行，以鼓励人们发表新收集的数据；一方面数据收集和计算服务是广泛分布的，另一方面还需要直截了当的系统接口以使数据和服务的定位和存储表示尽可能地透明。

为满足如此广泛的要求，国家虚拟天文台需要一个分布式的体系结构来提供统一且有效的数据访问和服务，而不管所在地点和具体实现。数据库假设已经存在，但在实施过程和访问政策方面会有相当的改变。元数据标准要提供明确办法来描述档案、数据集合和各种服务。数据访问层提供面向所有数据和服务的单一标准接口，既能在国家虚拟天文台的体系结构中连接数据库和服务，又允许用户应用程序访问国家虚拟天文台的资源；查询和计算服务提供信息发现、大规模相互关联和不同数据集分析的工具。数据挖掘应用程序运行于研究机

构中的用户工作站上，如 Web 浏览器中的 Java 程序那样；或者运行在国家虚拟天文台的主要数据中心，提供主要的用户接口以使国家虚拟天文台在科学研究中发挥作用。

4.3 组成部分

4.3.1 数据档案

数据档案里存储着数据集（如星表、图像和光谱），组织成逻辑相关的数据集合，也存储着描述数据档案及其数据保存的元数据。用户可以通过各种方式进行访问，如结构式 Web 界面、象 FTP 这样基于文件的标准界面，或其它因数据档案而异的访问协议。

国家虚拟天文台对数据档案不作要求，但要求能通过数据访问层让国家虚拟天文台可以访问，数据访问层是国家虚拟天文台访问数据库的入口。在最简单的情况下，要将一个档案接入国家虚拟天文台，只要安装数据访问层软件并修改几个配置文件来反映本地档案的数据保存与访问权限就可以了，就象安装一个 Web 服务器那样简单。更复杂的安装可提供元数据访问和服务器端功能的扩展支持，这些将在下面数据访问层一节中讨论。

4.3.2 元数据标准

元数据（字面意思就是“关于数据的数据”）是描述国家虚拟天文台的一些要素的结构化信息。元数据用来描述档案、档案提供的服务、其中的数据集合、每个数据集合的结构和语义（即含义）、以及数据集合中每个数据集的结构和语义。典型的天文数据集是星表、图像或光谱等数据对象。例如，一幅天文图像的语义元数据就是这幅图像 FITS 头的逻辑内容。

描述天文数据的元数据对数据发现和数据互操作是十分必要的，国家虚拟天文台各组成部分的自动化交互操作必需要有描述档案和服务的元数据。为使这些问题更容易解决，需要一套元数据标准。实际上将数据集特定元数据进行标准化是有一些限制的，不过在数字图书馆研究中发展起来的调解技术就有办法把不同学术界为相似类型的数据而开发的各种元数据“方言”融合起来。目前象 Astrobrowse（天文浏览）和 ISAIA（数据库信息访问交互系统）这样的项目表明天文界已经开始着手建立元数据标准了。

4.3.3 数据访问层

数据访问层（DAL）将提供统一的界面来访问国家虚拟天文台中所有的数据、元数据和计算服务。数据访问层的最底层是一个标准协议，定义了国家虚拟天文台软件的各部分之间如何进行通话。它也提供实现此协议的基准级软件，既可以直接应用，也可以作为天文界进一步开发的基础。这个软件包括将数据档案和计算服务接入国家虚拟天文台的服务器端软件，以及用来编写“国家虚拟天文台式”的分布式数据挖掘程序的客户端应用程序接口（API）。由于数据访问层基本上是一个协议，因此可以实现多个应用程序接口，用于支持诸如传统软件或多语言环境等。

数据访问层的关键在于它提供了国家虚拟天文台中所有数据和服务的统一接口。 用户应用程序通过数据访问层来访问国家虚拟天文台的数据和服务，而国家虚拟天文台内的档案和计算服务通过数据访问层来对其它档案的数据和服务进行内部访问，潜在地进行级联式的连接。因此，国家虚拟天文台是多层次分布式的系统，不过国家虚拟天文台的结构是简单的，

因为所有的组成部分共享同样的接口。除了这种位置透明性外，数据访问层还提供存储透明性，它隐藏了数据如何存储在数据档案中的细节。最后，数据访问层协议还对图像与光谱这类天文数据对象定义了标准的数据模型（在协议层上）。数据档案维护者将提供服务器端模块，当数据对象被访问时进行数据模型转换以允许应用程序处理远程的数据，而不管数据来源或数据在特定的档案中是如何存储的。

通常，调用数据访问层的客户程序不需要整个数据集，而只要其中一部分。服务器端函数将允许对单个数据集来修改设置、进行过滤和数据模型转换。在某些情况下，可以下载用户定义函数来进行数据处理并将计算结果返回到远程客户端，这对减少网络负载和分配计算是至关重要的。

由于数据访问层可以用于从远程档案中读取元数据和实际的数据集，所以就可以进行数据集复制来维持一个本地数据缓存，这对于优化整个国家虚拟天文台的数据访问是至关重要的，而对许多大规模的统计研究和相互关联则是必须的。数据集复制还可以实现全部数据集合的复制以及将数据档案及时地迁移。通过复制和采用元数据可以使中心站点能够自动地将所有的远程数据档案编入索引。

4.3.4 查询和计算服务

当数据访问层和元数据标准允许国家虚拟天文台连接档案和访问数据时，就需要查询和计算服务来支持信息发现和提供数据挖掘所需要的统计相关和图像分析的能力。

即使大多数的档案提供了其数据集合的基本查询服务，但只有将多重星表融合（相互关联）起来并能搜索符合一定统计特征的天体时，大规模的数据挖掘才可能实现。较大的国家虚拟天文台数据中心将提供这种大规模相互关联所需要的数据和计算资源。当一个查询或关联导致对远程数据档案进行子查询时，将广泛应用数据集的复制和缓存，以优化对经常访问的星表或档案的查询。为了融合各种星表的结果，需要完善的的元数据调解技术。

在某些情形下，用户可能需要下载一个算法函数来对原始数据进行像元水平的分析，计算出新的天体参数，从而提炼出一个参数搜索（事实上这种操作能动态地在已有的星表中加入新列，它是一项极有威力的技术）。由于国家虚拟天文台的候选天体表中可能包含几亿个天体，这将导致庞大的并行计算问题，可能需要万亿次超级计算机来解决。在大规模统计研究中，即使分布式计算技术和快速网络使得用户可以在他们自己的研究所里工作，但仍需要某种形式的同行评议来保证用户能分配到必需的计算和存储资源。对于某些更大的研究项目，用户可能要亲自前往某个国家虚拟天文台数据中心，以便有效地与各类人员进行联系，并有效地访问数据、软件和计算资源。

4.3.5 数据挖掘应用程序

数据挖掘领域，包括大型的多变量的数据集的可视化和统计分析，目前尚仍处于幼年期，但在未来许多年中将是非常活跃的研究领域。需要将目前大多数的天文数据分析软件升级为“国家虚拟天文台式”的软件，以便能很好地使用本地数据和远程数据。作为正在进行的数据挖掘技术研究的一部分，要开发新的应用程序。国家虚拟天文台应该提供支持这种开发所需要的接口和工具包，以及国家虚拟天文台主要中心的初步数据挖掘应用程序。但该问题的开放性质表明，一旦国家虚拟天文台开始运作，还将需要某种资助项目来支持多学科的

数据挖掘研究。

4.3.6 信息系统研究

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，国家虚拟天文台将突破现有技术的限制。这就要求学术界和工业界携起手来共同研究和开发国家虚拟天文台所需要的信息系统技术 ;要求与其它科学分支和与国家超级计算机中心进行合作，来开发元数据处理、数据处理和分布式计算的标准。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。

新一代的高速国家研究因特网已经出现，但目前还用得不多，还缺少有水平的学术应用问题来使用这个高性能的网络。国家虚拟天文台将是利用广域的高性能网络来进行学术研究的一个富有创造性的例子。

4.3.7 教育和普及

由于大量真正的科学数据能够免费地从因特网上获得，而且公众对天文学有着浓厚的兴趣，国家虚拟天文台将特别适合教育和科学普及。国家虚拟天文台具有基于因特网的内禀性质，因此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方法。

我们预计职业教育和普及人士（教育家、天文馆和科学展览馆的工作人员、科普作家和科学记者等）将广泛地使用这些资源创建科普网站，编写课程教材（从小学到研究生院）和制作精美的演示等等。只要有适度科学教育资源的学校就可以从网上找到各种现成的演示。公众可以利用商用网页浏览器中的 Java 程序来访问国家虚拟天文台的数据并可视化，进行虚拟观测并对所得数据进行分析和解释。我们期望国家虚拟天文台作为中心和催化剂从而建立起一系列科学普及合作伙伴关系。

尤为有趣的是，作为科技教育的焦点，国家虚拟天文台在一门物理科学（天文学）和计算机应用科学之间架起了一座桥梁。它用到了一系列事关二十一世纪整个经济与社会许多方面的技术和技能。运用这些方法论的活生生的事例，使得原本可能非常枯燥难懂的东西变得很容易接受。例如，我们注意到 SETI@home 计划的广泛成功，可以预期会有更多更好的实例，在其中许多人会用数据挖掘技术来研究国家虚拟天文台中提出的引人入胜的问题。

第五部分 实施计划

前面已经描述了导致“ 新天文学 ”的技术进步，以及国家虚拟天文台的特征，它利用和依靠这些技术进步，以更经济有效的投资来实现本来不可能做到的新的科学。

国家虚拟天文台管理的核心基础，是要认识到它由科学驱动的特性，并尽量使科学界更多的参与。它的活动和资金支持将分三个层面，并组织起来保证提供好用而又有说明文件的基础设施，保证软件项目是由科学驱动的，并保证大头的资金分配给经过同行评议的重点明确的科学项目。

1． 最优先要做的就是建设归档数据库的基础设施，和访问数据所需的充分说明的协议。数据访问的标准化要使得天文界可以靠它建造更高一级的工具。它们随技术进步而发展，

但应向后兼容。基础设施由基本预算支持，但由国家虚拟天文台主要的分布式站点来发展和维护。

2．定期“招标”在基础设施上开发软件工具。它们要遵循国家虚拟天文台所定义的标准，并交付给国家虚拟天文台让天文界更广泛地使用。重要的是，这些工具的开发机会要考虑各种可能性，吸引一大部分天文界人士参与。开发每一个软件工具都要针对一个重要科学项目，但又是整个天文界用户可以用来做研究的通用工具。

3．定期“招标”使用国家虚拟天文台，这是一些有更明确科学目标的特定项目，也可能包括软件开发（这很象现在 NASA 的 ADP 计划）。作为资助项目来说较为宽松，结果也许是一篇杂志论文。

在国家虚拟天文台的早期阶段，重点将放在前两个方面，但是随着国家虚拟天文台基础设施的发展，这三个方面之间的资金投入比例也要有相应的变化。

国家虚拟天文台的实施从最初的准备到全面运转要分几个阶段。计划实施的一个主要目标是通过现有的工具和服务设施尽快开始提供一定程度的功能。

第零阶段：前期准备到国家虚拟天文台启动

目标：国家虚拟天文台的概念设计；在一些中心开始提供实施国家虚拟天文台所必需的功能。

- 准备有关的宣传介绍文章、相关文件和确认国家虚拟天文台关键科学目标的“科学任务指南”；
- 在参与机构内启动，以确保数据可访问性和档案建设；
- 发展关键性技术，如信息交换协议和元数据标准；
- 在所有主要站点对有选择的数据子集建立星表搜索和图像数据抽取功能；
- 通过会议和工作讨论会发动天文界参与；
- 与国际天文界就一般国家虚拟天文台创建问题进行交流。

第一阶段：1—18 个月

目标：建立集成的数据发现、数据传送和数据对比服务功能。

- 扩充并定型数据发现和数据传送系统，包括建立元数据标准、传输协议和显示服务；
- 继续和所有站点合作，改进对在线服务的访问；
- 制订最后的网络连接和计算需求的计划；
- 部署小规模的交叉关联功能和可视化工具；
- 发展大规模交叉关联的原型工具；
- 通过会议和工作讨论会继续吸引天文界参与，组织用户委员会与访问委员会；
- 建立核心的技术和管理小组，并 建立报告和责任追究程序；
- 划定可由现有的数据中心和其他实体最有效发展的国家虚拟天文台功能子集；
- 设立向公众开放的计划；
- 设计并 建立国际天文信息基础设施；

- 通过与其他领域交流和合作促进信息技术发展。

第二阶段：18—36 个月

目标：建立初始的大规模交叉关联的功能；开始全面运作；

- 网络和配套的计算设备开始到位；
- 开发并配置复杂数据集的可视化工具；
- 开发并启动数据访问层；
- 确保数据发现和对比的工具成熟并投入常规运转；
- 建立国际机构之间合作关系以保证美国和非美国的设备和服务的互用性；
- 管理机构和咨询委员会进入常规运转。

第三阶段：36—60 个月

目标：建设全面运转的基本的国家虚拟天文台；实现全规模的、由计算和网络恰当组合的系统来支持的交叉关联能力。

- 扩展数据服务的范围，包括国际合作者的设备；
- 配置好数据访问层并投入常规运转；
- 支持用户定义的便携式处理器；
- 支持更高层的数据产品，例如预先准备的交叉认证。

这仅是大致的实施时间表，必然随问题更加明确以及对国家虚拟天文台支持程度的进一步明朗而作调整。但这是一份“优化”的时间表，它反映的是对国家虚拟天文台功能理想的实施途径的估计。鉴于技术上的可行性，估计服务很快会得到落实；当然，低于优化水平的资金投入会减慢其进程。

（本文件更详细的版本将包括有关可能的管理结构和预算等的讨论）

附：本文缩写词

AASC：Astronomy and Astrophysics Survey Committee

NVO：The National Virtual Observatory

LSST：The Large-aperture Synoptic Survey Telescope, a 6.5-m-class optical telescope

STScI：Space Telescope Science Institute
IPAC：Infrared Processing and Analysis Center, at JPL/CIT
HEASARC：High Energy Astrophysics Science Archive Research Center, at NASA/GSFC
CXC：Chandra X-ray Observatory Center, at SAO
CADC：Canadian Astronomy Data Center

SDSC：San Diego Supercomputing Center, at University of California, San Diago
NCSA：National Center for Supercomputing Applications, at University of Illinois, Urbana-Champaign

ADS：NASA Astrophysics Data System
NED：NASA/IPAC Extragalactic Database
SIMBAD：The SIMBAD astronomical database, at CDS, Strasbourg

AIPS： Astronomical Image Processing System, at NRAO
IRAF： Image Reduction and Analysis Facility, at NOAO
IDL： Interactive Data Language
FTOOLS：a general package of software to manipulate FITS files, at NASA/GSFC/HEASARC
SkyView： at NASA/GSFC/HEASARC
FITS：the Flexible Image Transport System

NPACI：National Partnership for Advanced Computational Infrastructure
NASA/IPAC Infrared Science Archive

2MASS：Two Micron All Sky Survey
SDSS：Sloan Digital Sky Survey

NOAO：National Optical Astronomy Observatories
NRAO：National Radio Astronomy Observatory
NSO：National Solar Observatory
NAIC：National Astronomy and Ionosphere Center

HST：Hubble Space Telescope
HDF：Hubble Deep Fields

NASA ADP：Astrophysics Data Program for data analysis support

SETI：Search for extraterrestrial intelligence

SRM：Science Reference Mission
DAL：Data Access Layer
API：applications programming interface