项目编号	INFO-115-C01
文档编号	TR-REC-03

中国科学院数据应用环境建设与服务

参考型数据库建设规范

（征求意见稿）

（2009.06.19 完善版）

中国科学院计算机网络信息中心科学数据中心

2009 年 6 月

目　　录

1　适用范围 1

2　术语与定义 1

2.1　参考型数据库 1
2.2　数据标准 1
2.3　元数据 1
2.4　WWW 服务 2
2.5　Web 服务 2

3　参考型数据库基本要求 2

4　总体架构 3

5　内容组织 4

5.1　参考型数据库建立 4
5.1.1　数据模型 5
5.1.2　其他规范 6
5.2　数据处理 7
5.2.1　数据的映射 7
5.2.2　数据清洗 8
5.2.3　数据转换 9
5.2.4　数据更新 9
5.3　元数据 9
5.4　数据质量 10
5.4.1　数据质量指标 10
5.4.2　数据质量控制方法 13
5.4.3　数据质量评价审核方法 14
5.4.4　质量报告 16

6　技术架构和接口规范 16

6.1　功能要求 17
6.2　应用系统与工具 18
6.3　接口规范 18
6.4　接口格式要求 20
6.4.1　通用格式定义 20
6.4.2　开放接口的安全性要求 21

7　服务 21

7.1　服务对象 22
7.2　服务方式与要求 22
7.2.1　在线发布方式 22

7.2.2 离线发布方式.....24
7.3 数据交换格式.....24
7.4 共享分级分类设置.....24
7.5 其他服务要求.....24
8 运行维护.....25
8.1 运维人员.....26
8.2 基础运行环境.....26
8.2.1 机房.....26
8.2.2 互联网接入环境.....26
8.2.3 网络服务器与存储设备.....27
8.3 运行.....27
8.3.1 运行模式.....27
8.3.2 日志管理.....27
8.4 安全保障和故障处理.....29
8.4.1 基础设施安全.....29
8.4.2 软件安全.....29
8.4.3 数据安全.....29
8.4.4 非技术防护措施.....30
8.4.5 故障处理.....30
8.5 备份和恢复.....30
8.6 运维服务的质量.....31
附录A（规范性附录）标准实施一致性测试.....32
A.1 内容组织.....32
A.1.1 数据集名称及标识符.....32
A.1.2 数据库建立.....33
A.1.3 数据处理.....33
A.1.4 数据质量.....34
A.2 技术架构与接口规范.....34
A.3 服务.....35
A.4 共享.....35
A.5 运行维护.....35

参考型数据库建设规范

1 适用范围

本规范定义了参考型数据库的总体架构，规定了参考型数据库在内容组织、质量控制和技术实现方面需要完成的工作和需要满足的要求，并提出了对参考型数据库在运行维护和服务方面的要求。

本规范适用于中国科学院数据应用环境建设与服务项目中参考型数据库的建设、运维和服务。

2 术语与定义

2.1 参考型数据库

以我院有特色、有长期积累的数据为基础，建成的符合国家或国际标准、有严格质量控制与管理、具有完整性和权威性的数据库。参考型数据库服务于广泛的用户群体，是我院多年来积累的特色资源。参考型数据库要求数据规范，数据质量有保证，学科内容权威。参考型数据库是对数据进行标准化和产品化深加工后形成的相对完整的数据集。

2.2 数据标准

在适用范围内取得共识的一组数据资源所应遵守的规则的集合。不同类型的资源可能有不同的数据标准，一般包括完整描述事实所需的数据项集合、数据项语义定义、著录规则和计算机应用时的语法规定等内容。

2.3 元数据

描述数据及其环境的数据。一般而言，元数据有两方面的用途：

- 首先，元数据提供基于用户的信息，帮助用户发现和使用数据;
- 其次，元数据支持系统对数据的管理和维护。

具体来说，在参考型数据库系统中，元数据机制主要支持以下系统管理功能：

- 资源定位，确认所在主题数据库；
- 集中管理数据集成访问的物理参数；
- 记录并客观反映数据质量相关内容。

2.4 WWW 服务

WWW 是一种交互式图形界面的 Internet 服务，基于客户机/服务器方式的信息发现技术和超文本技术的综合。WWW 服务器通过 HTML 超文本标记语言把信息组织成为图文并茂的超文本；WWW 浏览器则为用户提供基于 HTTP 超文本传输协议的用户界面。

2.5 Web 服务

应该是一个软件系统，用以支持网络间不同机器的互动操作。网络服务通常是许多应用程序界面（API）所组成的，他们通过网络，如国际互联网（internet）的远程服务器端，执行客户提交服务的请求。（W3C 概念）

3 参考型数据库基本要求

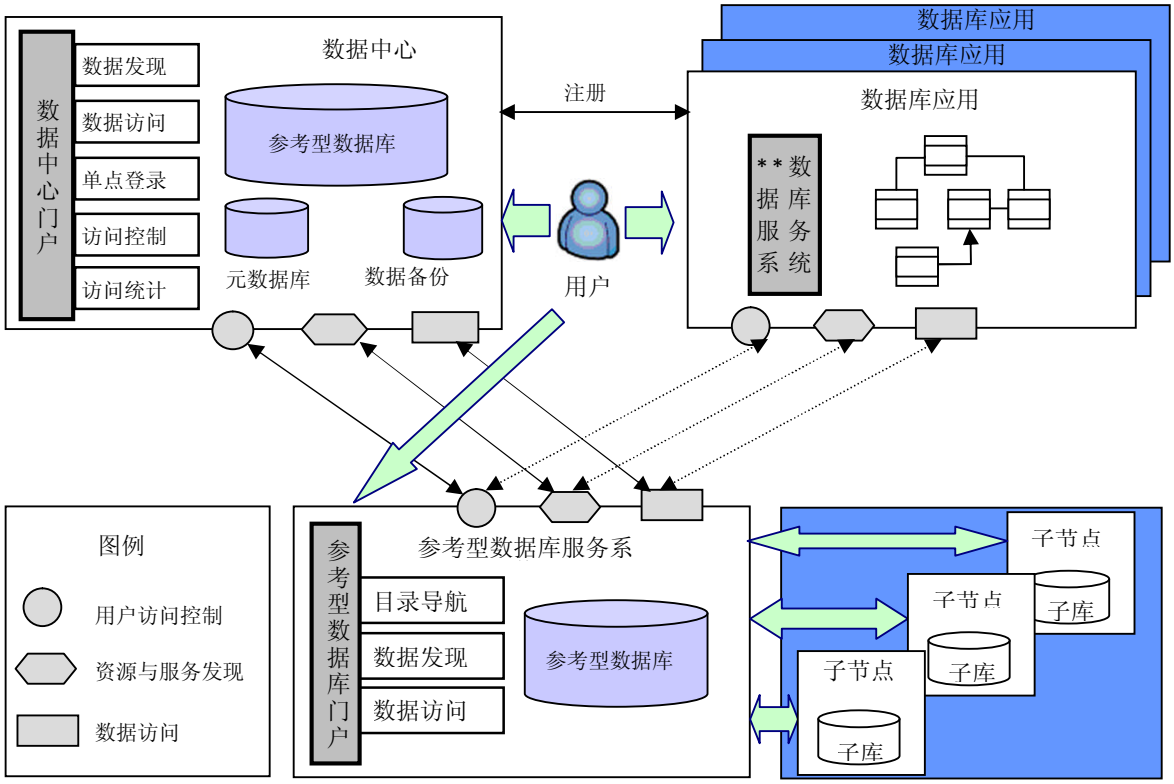
参考型数据库建设过程中，应采用“数据应用环境建设与服务”项目发布的有关标准规范，特别是《中国科学院数据应用环境建设与服务 数据库建设技术指导规范》，以及相关的国家标准、国际标准、学科领域标准规范或其应用方案，完成数据规范的构建，数据资源的整理、加工和增建；参考型数据库实行完全免费共享的服务模式，面向个人用户和数据库用户提供标准化的服务。其间应特别实现以下基本要求：

1. 参考型数据库是本学科领域的基础性数据，以我院有特色、有长期积累的、成熟的数据库作为基础，并建立包括多种来源的持续更新机制；
2. 参考型数据库按照既定的数据标准建立和提供服务，具有完整性和权威性，是对数据进行标准化和产品化深加工后形成的数据库；
3. 参考型数据库建设过程中执行全面的质量措施，建立有严格的面向数据资源全生命周期的质量控制和管理、审核与发布、服务跟踪等制度；
4. 参考型数据库内容完全公开共享，构建统一的数据服务平台，服务于广泛的数据库用户和个人用户；
5. 对数据中心门户系统开放符合规范的接口，以支持通过数据中心门户系统实现对数据资源的访问。

上述参考型数据库建设基本要求，是本规范关于内容组织、资源建设、接口规范、运维与服务等具体内容的概括，所有基本要求的满足情况可以通过本规范附录 A“标准实施一致性测试”确认。

4 总体架构

参考型数据库面向给定学科或应用领域中的一组常用或核心内容，按照严格的数据规范建立数据存储和应用环境，以院内多年积累的数据资源为基础并辅以多种来源的数据更新，在严格的质量控制要求下，通过参考型数据库站点，web 服务或其他服务形式，为用户提供查询或数据库引用服务，支持研究人员的工作和相关数据资源的建设。参考型数据库的服务还应集成到数据中心门户系统，支持通过数据中心门户系统统一发现和访问。



参考型数据库总体架构

参考型数据库前期建设阶段的主要工作包括内容组织和技术实现两个方面，后期阶段的主要工作是运维和服务。

参考型数据库按照统一的数据规范组织数据内容，数据规范包括参考型数据库数据模型和其他相关规范，将院内多年积累的原始数据按照参考型数据库规范要求进行映射、清洗和转换后进入参考型数据库，同时参考型数据库还需建立持续更新机制以确保其具有长期有效的参考价值。参考型数据库还应建立相应的元数据信息。

参考型数据库建设在技术上涉及到异构数据资源整合，数据管理与服务，元数据的生成和管理、用户认证与授权、服务监控以及参考型数据库、数据中心二者之间的信息交互与通信等方面。

运维和服务是保障建成的参考型数据库发挥出其价值的重要工作。通过机制、环境、人员等在安全保障和故障处理、备份和恢复、数据更新等方面的高水平运维以保证参考型数据

库正常稳定的对外服务而服务过程中应紧密结合资源特点配置支撑队伍，特别应面向领域内关键项目的需求提供定向的数据服务支持。

5 内容组织

参考型数据库面向给定学科或应用领域中各类研究所共同关注的基础性数据内容，根据领域科研人员的共识选择或建立数据规范，以我院有特色、长期积累的数据库为主要来源，经过数据规范化加工处理，审核和补充，从而建成权威的可供学科领域内数据建设和科研活动共同参考引用的参考型数据集合。

参考型数据库不追求对一个学科领域数据内容的完整覆盖，而是致力于对其中最为核心的或常用或共性的一组基础内容提供高质量和规范的表达。建成后的参考型数据库可以作为权威的参考资料面向研究者提供查询，同时也可对其他相关数据库的建设提供帮助——专业数据库通过引用参考型数据库，可以在这些共性内容上取得较高的数据质量、部分避免重复建设、并且易于建立数据库之间的关联。

参考型数据库的建立不仅是聚集原始数据，还包括对原始数据的规范化加工。为实现前述价值，参考型数据库在内容组织方面的工作主要应包括：参考型数据库建立、数据处理以及元数据建立三个方面：

- 参考型数据库建立：参考型数据库的建设必须建立或参照已有的权威数据规范，保证参考型数据库内容具有正确、统一的结构，数据元素具有明确的含义和规范的表达；数据规范本身必须符合业内通行的规则和习惯，并具有科学性和合理性；最重要的数据规范是参考型数据库数据模型，参考型数据库仅包括学科领域中一组最为核心的基础内容，其结构良好易于参考引用。
- 数据处理：参考型数据库以我院长期积累的特色数据库为基础，原始数据未经统一规划，质量、格式互有不同，被抽取后必须经过清洗、转换、规约等加工操作，方可形成符合参考型数据库规范的数据。此外，为确保参考型数据库的完整、正确和及时，还需搜集其他资料来源进行数据补充并建立数据持续更新机制。
- 元数据：参考型数据库应建立相应的元数据信息，为用户提供便捷的资源发现、理解、评价和应用，同时承建单位需将元数据向数据中心进行注册，以实现数据应用环境门户系统对参考型数据库的集成。

5.1 参考型数据库建立

参考型数据库的建设必须按照数据规范进行。数据规范的主要任务是针对参考型数据库的建设目标和论域，制订一套结构化的标准，这组标准确定数据的定义、描述、表示以及生成规则，从而实现参考型数据库的规范性和可共享性。

数据规范的首要内容是参考型数据库数据模型，它确保数据库内容具有统一的结构，并且使数据元素具有明确的含义和规范的表达，数据模型可以直接使用业内权威性的数据规范，也可在保持良好兼容性的前提下自行制订。

为凸显内容设计的科学性，除统一的数据模型以外，参考型数据库在设计与建设过程中还应注重遵循学科领域中已约定俗成或业已形成国际或国家标准的内容。

参考型数据库基本要求

- 参考型数据库必须建立参考型数据库数据模型，模型应遵循领域内相关的国际或国家标准，国内外权威数据整合项目发布的数据格式，或至少对其提供良好的兼容性，承建单位提交相关文档中应对数据的规范化情况有所分析；
- 参考型数据库一般仅包含学科领域中一组核心内容而不追求对领域知识的完整覆盖，通常保持简洁的数据结构，该结构应正确、简洁，利于为专业数据所引用；
- 参考型数据库具有统一的数据编码标准，统一的数据录入、存储、输出标准；
- 承建单位对参考型数据库所遵循的标准规范必须在其元数据中和数据服务系统页面上明确说明；
- 参考型数据库应向数据中心门户系统开放符合 6.3 节“接口规范”要求的接口；

5.1.1 数据模型

参考型数据库数据模型负责保证参考型数据库内容具有正确、统一的结构，数据元素具有明确的含义和规范的表达。

参考型数据库数据模型基于对学科领域数据实体和现有原始数据两方面的分析而建立。

通常，参考型数据库数据模型应优先考虑直接采用学科领域内现有的且已被公认的数据规范，在现有规范不能完全满足参考型数据库要求时承建单位在充分尊重和借鉴现有通行数据规范的前提下应自行建立参考型数据库数据模型。

对数据实体的分析应确保应用无关，旨在构建该领域中核心数据内容的原子粒度的数据模型。对数据实体的分析需要结合以下内容：

- 相关的学科背景知识
- 数据内容已有的数据标准规范
- 相似的权威性数据库的内容分析

参考型数据库数据模型是面向系统实现和关系型数据库数据存储的，内容包括三个部分：数据结构、数据操作、数据约束。

- 数据结构：主要描述数据的类型、内容、性质以及数据间的联系等。数据结构是数据模型的基础，数据操作和约束都建立在数据结构上。不同的数据结构具有不同的操作和约束。
- 数据操作：主要描述在相应的数据结构上的操作类型和操作方式。
- 数据约束：主要描述数据结构内数据间的语法、词义联系、他们之间的制约和依存

关系，以及数据动态变化的规则，以保证数据的正确、有效和相容。

参考型数据库数据模具体应包括以下内容：

- 实体
- 属性
- 关系
- 约束条件

参考型数据库数据模型使用实体-关系模型描述。

参考型数据库数据模型应具备如下特征

- 参考型数据库应专注于对领域内核心或常用的基础内容提供高质量和规范的表达，模型应为学科领域内的共识性内容，应在符合研究者习惯的前提下使用较小的粒度表达并保持简洁，利于被各种专业数据库所引用；
- 数据模型不应针对具体应用而设计；
- 数据模型应具有良好的结构，一般而言应满足第三范式的要求，数据元素的命名应遵循统一的风格；
- 参考型数据库中每条记录必须分配具有在整个参考型数据库中唯一的标示符（UID），承建单位负责唯一标示符命名规则的建立与实施。

公共数据模型的表达

分析工作完成的成果应体现为实体关系图和数据字典，数据字典如下：

数据字典

表名：		**ID：**		
字段名	**ID**	**定义**	**数据类型**	**约束条件**

5.1.2 其他规范

参考型数据库除必须遵循统一的数据模型规范外，还应尽可能参考已经存在的学科领域内其他与数据内容相关的标准规范，如：

- 命名规范
- 表达方法
- 分类或代码
- 数据交换格式等

承建单位应充分了解研究领域的表达习惯、搜集相关规范并将其实现于参考型数据库当中，一般而言这些规范可以落实为：

- 数据模型；
- 约束条件；
- 质量要求等。

参考型数据库所遵循的相关规范也应在参考型数据库元数据和服务页面上加以说明。

参考型数据库的数据模型和数据规范完全确定后，即可按内容规范进行技术实现，技术实现的相关细节规定参见第 6 章。

5.2 数据处理

参考型数据库的原始数据一般具有各自不同的结构，并且由于质量问题或结构差异，往往不能直接进入参考型数据库，需加以规整。数据内容的整理包括清洗、转换、映射等。

5.2.1 数据的映射

参考型数据库的数据模型确定以后，原始数据应建立与参考型数据库数据模型之间的映射规则。

映射规则的建立包括两部分内容：映射关系的建立和转换规则的建立。

5.2.1.1 映射关系

映射规则不是简单的对应关系，根据实际迁移的源和目的结构，还可能包含字段的拆分、合并等。对于用实体关系模型表达的公共数据模型而言，映射体现在实体层面和属性层面。

- 实体映射：按公共数据模型中实体的属性来源和拆分的情况，源表和目的表在数量上可分为一对一映射，一对多映射，多对一映射，多对多映射。
- 字段映射：关系可以分为 3 种，即直接映射、主键映射和外键映射。其中：主键映射是为了保证专业库中主外键约束在参考型数据库中被保留，外键映射是保证参考型数据库实体表中外键字段能够从专业库表中对应字段正确迁移。直接映射就是专业库表中的字段直接映射到数据模型表中的字段上，不发生拆分、合并等运算。
- 类型映射：数据类型在不同的实现方法当中存在不同的具体表达，类型映射的两端可能是该实现方法下的一个数据类型，也可能是一个数据类型加格式约束。

5.2.1.2 映射转换的表达

每个映射关系都可以表达为一个来源、对象和生产式的三元组。

- 对象：被映射端对象的集合，对象可以是基础层级（字段层面的，变量）的也可以是基础层级对象组成的集合，如果存在非基础层级的对象，其映射规则最终应落实到基础层级；
- 来源：每个对象在源端对应的数据来源；
- 生产式：源端的数据来源均可按照生产式进行加工，并形成符合目标端数据模型要求的对象内容。

映射关系可按如下的表格建立说明：

映射 ID	对象标识	来源	生产式

5.2.2 数据清洗

原始数据中有可能存在着大量的脏数据，需要利用有关技术如数理统计、数据挖掘或预定义的数据清洗规则将脏数据筛选、转化成满足数据质量要求的数据。不符合要求的数据主要是有不完整的数据、错误的数据和重复的数据三大类。

- 不完整的数据，其特征是是一些应该有的信息缺失。需要将这一类数据过滤出来，列出其缺失的内容，要求在规定的时间内补全。补全后才写入数据库。
- 错误的数据，产生原因可能是在接收输入后没有进行判断直接写入后台数据库造成的，比如数值数据输成全角数字字符、字符串数据后面有一个回车、日期格式不正确、日期越界等。这一类错误需要用 SQL 的方式挑出来，要求限期修正。
- 重复的数据，将重复的数据的记录所有字段导出来，然后进行确认并清除。

经过清洗的数据满足以下要求：

- 单一字段中不存在多种信息；
- 相同对象的名称表达一致；
- 缩写词、惯用语的表达一致；
- 值与字段名含义匹配；
- 同类数据的计量单位统一；
- 同一字段内的数据格式统一。

5.2.3 数据转换

原始数据往往不能直接对应到参考型数据库数据模型当中，经过映射、清洗后尚可能需要一系列的变换和运算。凡此类变换应对其详细的变换规则加以明确。

5.2.4 数据更新

为维持参考型数据库的权威性和及时性，参考型数据库必须建立持续更新机制。

数据更新应满足以下要求：

- 根据描述对象资源发展变化的规律及学科领域内研究进展的特点，或者是错误修正、质量提升或时间延续等的因素，制订明确的数据更新计划，并向数据中心提交备案，作为对参考型数据库资源更新情况进行检查的参考和依据。数据更新可分为周更新、月更新、季度更新、半年度更新、年度更新和不定期更新等，不定期更新的时间间隔一般不应超过一年。
- 数据更新的来源包括资源数据库更新的同步、国内外相关研究进展的报道、权威的文献杂志、学报、专著，以及整合新的数据资源等，承建单位有责任确保更新数据的质量。
- 更新结果应实时反映于参考型数据库线上服务，如有基于参考型数据库构建的产品，更新内容至少应每年一次反映到该产品当中。
- 参考型数据库承担建设单位应安排专门的人员负责参考型数据库内容的更新。

5.3 元数据

元数据在数据的组织管理、发现、理解、评价、利用等方面起着重要的作用。同时为了实现数据应用环境门户系统对参考型数据库的集成，承建单位应将参考型数据库及其元数据注册到数据中心的资源与服务注册系统中，并向数据应用环境门户系统开放符合 6.3 小节规定的接口。

参考型数据库的元数据包含以下内容：

1) 核心元数据：主要描述参考型数据的基本内容特征、外部特征和结构特征，符合核心元数据规范的规定。核心元数据包含以下元数据元素：
    - 唯一标识符
    - 标题
    - 关键词
    - 摘要

- 类型：默认为参考型数据库
- 所属分类
- 创建者
- URL
- 模式信息
- 质量报告：具体要求参见 5.4.4“质量报告”

2) 系统元数据：主要描述参考型数据库的有关连接信息，元数据元素包括：

- 数据库连接主机 IP
- 端口号
- 数据库名
- 用户名
- 密码
- Web Service 服务地址

5.4 数据质量

数据质量指参考型数据库数据内容的质量状况，主要包括数据质量指标、数据质量控制和数据质量评价三个方面。

5.4.1 数据质量指标

数据质量指标是进行质量活动中客体的具体质量反映，如正确性、准确性等，是控制和评价数据质量的主要内容。

科学数据质量指标与学科领域关系密切，每个的参考型数据库质量指标应根据具体情况由承建单位组织领域专家研究和制定。

评价指标选取的原则

评价指标的选择应与科学数据的主要质量特征基本一致，特别应注重科学数据的真实性、可达性和实用性方面的指标，确定的主要原则包括：

- 指标选取要有系统性，以保证综合评价的全面性和可信度；
- 指标应意思明确，含义明确，不产生歧义；
- 选取的指标要有可测性，能被客观测量，易于掌握，而且能把数据质量在时间上进行比较，且其测量方法应长期保持有效；；
- 指标之间应尽可能避免明显的包含关系和相互冲突，对隐含的相关关系和相互冲突的指标，在模型中加以适当的消除和取舍；
- 指标的选择要保持同趋势化，以保证可比性；
- 指标设置要有重点，抓住主要因素。

常用数据质量指标

具体的质量指标应基于以上原则建立，指标的筛选、权重都需由领域专家根据评价对象具体决定。以下仅列出常用的数据质量指标供领域专家参考。

- **数据库层面上常用的质量指标**

指标	评价依据
A 可获取性	1 数据能够很方便的为用户获取 2 在数据库中需要很费劲才能找到所需数据 3 所需数据应能很快检索到 4 系统中还有很多所需数据不能自动、快捷查到，查全率不高 5 数据检索（或查找）流程简洁、清晰
B 准确性	1 提供的数据准确无误 2 数据的表述（或值）很好地反映源数据的真实状态 3 数据的表述不会引起歧义 4 经过加工整理后的数据表述不够准确，与原始数据有较大误差 5 数据的表述（或值）与实际误差在可接受的范围内
C 正确性	1 提供的数据符合数据质量控制标准或规范 2 采集、传递、加工和整理后的数据偏离标准误差大 3 有专门机构（或专业人员）审核检查数据的正确性 4 有必要的程序或反馈流程来监测、修改数据的正确性 5 对目前提供的数据的正确性不太满意
D 一致性	1 数据一致性对数据共享很关键 2 经过加工整理前、后的数据经常出现不匹配、不一致 3 提供的数据元数据（如名称描述、时间、空间、来源、载体、表达方式、数据量）之间不应存在逻辑冲突 4 数据集合内各个个体数据之间经常有冲突（例如某一指标有多个数值，多个版本；编码相同但数据实体不同等） 5 普遍存在的相同数据实体使用不同的表达符号或不同的描述名称的情况
E 相关性	1 查找到的数据与主题不完全一致，但却是其中的某一方面的阐述 2 查找到的数据集合多数在用户需要的检索主题内 3 提供的数据主题与用户检索主题意思匹配 4 查找到的数据多数和用户需要数据无关 5 数据必须要和用户需求（目的）有相关性
F 有用性	1 数据有用性对共享数据很重要 2 数据能过帮助解决问题 3 一般经过加工、整理过的数据可用性较好 4 数据具有增值性 5 数据是对传统文献数据的有用补充
G 完整性	1 数据（数据要素）尽可能完整对共享数据很重要 2 科学数据记录格式、条目不完整 3 数据库内普遍存在数据要素残缺、不完整的情况 4 检索到的数据内容完整性应能够满足所需查找要求 5 目前可检索到的有价值科学数据量仍然不够

H 可信性	1 数据可信度不高 2 对专业网站的数据基本可靠 3 数据要素齐全的数据可信度高 4 对数据来源背景描述清楚的数据认为可信 5 经过专家审核或专业人员编辑过的数据可信度高
I 可理解性	1 科学数据（内容、格式等）应清晰易懂 2 提供的数据非常容易判断出是否符合需要 3 科学数据描述有太多专业术语难以明白，影响数据使用 4 对数据描述、分类及编码等的不规范性易造成对数据难以理解 5 用户对共享数据的技术规范、质量控制标准的了解、熟悉有助于数据理解
J 客观性	1 科学数据应符合所述事实 2 提供的数据应经得起再验证 3 存在虚假的数据 4 多数网络科学数据的表述基于事实，不带有个人主观看法 5 数据提交前数据内容没经过专家或专业人员的质量审核，造成与事实偏差
K 适量性	1 检索到的数据有多余的、与要求不符的数据 2 重复数据很多 3 数据过量容易对数据吸收造成负荷（时间、精力、消耗） 4 查询到的数据越多越好 5 目前查找的数据量能足够满足要求
L 及时性	1 服务是否稳定，响应是否及时 2 用户反馈的问题能否及时得到解决 3 存在过时的数据 4 是否经常更新，更新是否及时 5 查询所花的时间和精力是否好过预期
M 有效性	1 数据的有效性对其共享使用非常重要 2 能查询到最新的数据 3 查询到的科学数据满足当前的工作任务 4 查到的结果比要求的还要好很多 5 更新的数据与原数据没有区别标识
N 可靠性	1 数据的可靠性对数据用户很重要 2 数据来源标注齐全且真实可靠 3 数据来源标注普遍不齐全 4 对来源清楚的数据可以放心采用 5 加工编辑过的数据基本可靠
O 元数据	1 了解数据的背景资料对数据使用很有必要 2 提供的科学数据有必要的背景资料说明 3 所查询的数据库有完备的元数据说明 4 所查询的数据库有清洗的数据加工、整理的数据质量说明

- **记录层面常用的质量指标**

指标	评价依据
专家审核意见	领域专家对该记录的评审意见

用户反馈	使用者对记录提出质疑的情况
数据质量历史	专家对该记录评审的历史记录，反复争议的内容往往有较多质量问题
文献证明	可以佐证该条记录的文献数量，以及这些文献的影响因子
被引用频次	每条记录被参考引用的频次，该指标应结合参考型数据库的服务由系统统计得出
上下文	相关数据元素的质量状况，如子元素具有高质量并且完备，则一般父元素质量较高
访问量	用户访问该记录的次数较多则一般质量较高

质量指标权重的确定

质量指标确定后还应为各项指标确定权重。参考型数据库质量指标的权重可由制定指标的领域专家打分决定。

指标与数据质量可以呈正相关性或负相关性，并根据强弱划分若干档次。确定权重使用的表格可参见以下范例。

指标																		
	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9
A 可获得性																			
B 准确性																			
C 正确性																			
D 一致性																			
E 相关性																			
F 有用性																			
G 完整性																			
H 可信性																			
I 可理解性																			
J 客观性																			
K 适量性																			
L 及时性																			
M 有效性																			
N 可靠性																			
O 元数据																			
……																			

5.4.2 数据质量控制方法

数据质量的控制是贯穿于数据从生产到服务的整个生命流程当中的所有质量方法和质量行为的总和。参考型数据库建设过程本身往往已经包含一些质量控制方法，但这些举措未经合理和统一的规划，也未形成制度，其执行力度是无法确保的。

承建单位应组织领域数据专家根据参考型数据库数据内容的特点以及生产流程，制订参

考型数据库质量控制规范，并在生产过程中严格遵守执行。

质量控制可以是技术方法，也可以是组织管理等各个方面。不同数据对象数据质量控制方法的方面和细节均有不同。以下方法是质量控制流程的范例：

- 多路输入互相校验
- 不同来源同类数据互相校验
- 统计方法
- 基于规则库的自动检测
- 用户反馈触发的内容检查

参考型数据库的质量控制应满足以下要求：

承建单位对所承建参考型数据库的质量负责，因数据质量问题给用户造成损失的，由承担建设单位与用户协商后做出处理。

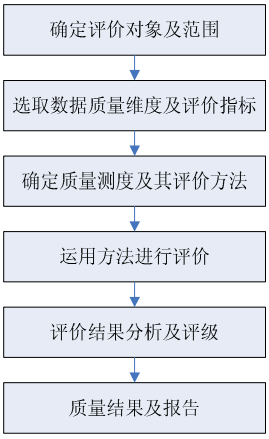
数据用户有权就数据质量问题和数据错误提出修改意见，接到用户相关意见的参考型数据库承担建设单位应在 10 天内就用户所提问题给予答复，对确实存在质量问题和数据错误的数据，应及时修正。无法及时修正的，应暂时撤出共享范围，待修正后再行列入。

参考型数据库承担建设单位负责组织专家建立数据质量控制规范并严格遵照执行。规范文档应上报数据中心备案，同时，对规范的执行情况应反映于数据质量报告当中，两者作为将来对参考型数据库数据质量进行评价的指标之一。

5.4.3 数据质量评价审核方法

数据中心质量评价流程

参考型数据库的数据质量评价根据已提出的数据质量指标和维度展开，一般而言参考型数据库的质量审核参照如下流程进行：



- 确定评价对象及范围：评价对象主要是参考型数据库及其内容。
- 选取数据质量维度及评价指标：参考型数据库的评价指标由承建单位组织领域专家提出，具体参照 5.4.1“数据质量指标”执行。
- 确定质量测度及其评价方法：数据中心将主要利用组织专家对参考型数据库中数据进行抽查、利用数据完整性检测与评价工具相结合的方式，来对参考型数据库的数

据质量进行评测。

- 运用方法进行评价：基于选定的质量指标、测度及其评价方法，借助于辅助软件等实现对数据的客观评价。
- 评价结果分析及评级：对评价结果进行分析，并参照相应的等级划分策略确定质量级别。
- 质量结果及报告：发布完整的数据质量报告，包括质量评价的整个流程和最终结论。

参考型数据库数据质量评价应注重以下原则：

- 科学性原则
- 客观性原则
- 系统性原则
- 可操作性原则
- 针对性原则
- 引导性原则

参考型数据库的数据质量至少应达到如下要求：

- 准确率
  - 数据准确率不低于 99%
- 完整性
  - 数据库的核心属性（学科领域专家讨论确定）不能空，非关系型文件的元数据库的核心元素不能为空；
  - 关系数据库的填充率不能低于 70%；
- 数据中心将通过数据完整性检测工具软件对上述完整性指标进行检测。重复率
  - 数据重复率不超过 1%
  -

专家评审制度

除接受中心数据质量检查以外，一般而言参考型数据库应建立长期持续的专家评审制度。

- 承建单位负责组织领域专家，不定期对参考型数据库内容进行逐条评审。
- 参考型数据库中的每条记录设定四种状态：未审核、审核有效、有争议和审核无效，由资源数据库整合进入参考型数据库的数据应标记为未审核；
- 领域专家审核后方可修改记录状态，未审核、审核有效和有争议数据可以对用户提供服务，但必须注明数据审核状态，判定审核无效的数据将被暂时拿出参考型数据库， 但必须由其他专家附议方可彻底删除，专家可以对其他人已审核过的数据再做审核；
- 存在争议的数据应由承建单位负责组织专家讨论解决，状态被多名专家反复修改的记录也被认为是存在争议的数据；
- 评审专家对评审后的数据负责，参考型数据库提供的服务当中应列出评审专家的姓名，同时列举的资料应包括以下内容：

- 专家姓名
- 评审时间
- 评审依据（所参考的文献资料等）
- 评审历史

- 研究对象频繁发生变化的领域，专家评审应持续进行。

5.4.4 质量报告

参考型数据库的数据质量状况应定期形成数据质量报告。数据质量报告更新的频率根据数据内容更新的频率决定，至少每年应更新一次。

数据质量报告的内容是对参考型数据库数据质量状况的综合概括，至少应包含以下内容：

- 数据质量指标
- 数据质量控制方法及执行情况
- 数据质量评审方法极结果

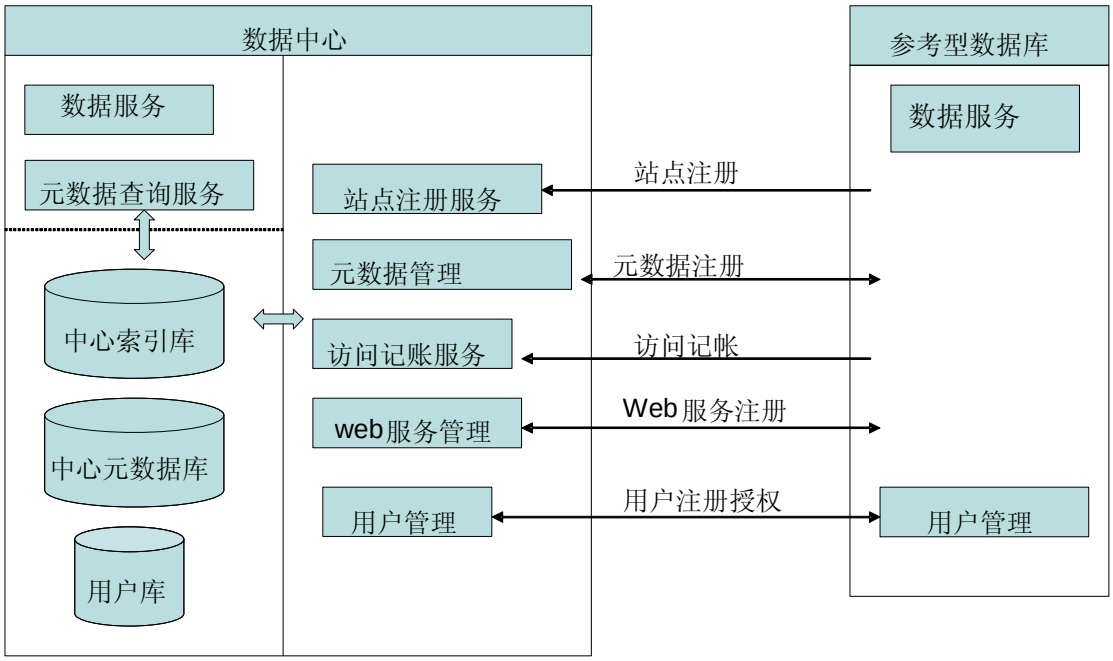
参考型数据库质量报告除提交数据中心以外，也是参考型数据库元数据的一个组成部分。

关于参考型数据库质量的详细规定，见数据应用环境建设和服务项目发布的《数据质量评测方法与指标体系》、《共享服务评价指标体系》。

6 技术架构和接口规范

参考型数据库建设在技术上涉及到异构数据资源整合，数据管理与服务，元数据的生成和管理、用户认证与授权、服务监控以及参考型数据库、数据中心二者之间的信息交互与通信等方面。

参考型数据库采用集中的方式来存放数据。



6.1 功能要求

1. 数据访问

提供数据的简单、高级等检索方式和数据浏览功能：

（1）检索结果超过 30 项时，必须提供翻页功能；

（2）每个检索结果显示该索引记录的字段信息。

2. 参考型数据库的生成与管理

根据 5.1 参考性数据库建立与 5.2 数据处理，实现参考型数据库的建立。参考型数据库的建立应满足《数据库建设技术指导规范》提出的要求。

3. 开放服务状态监控接口

参考型数据库向数据中心提供服务状态监控接口，使数据中心对参考型数据库的站点连接状态进行监控。

4. 用户管理

参考型数据库可以管理和维护自己的注册用户信息，并开发用户注册及登录功能。如欲进行单点登录，与数据中心使用同一用户库，则参考型数据库的用户注册和管理功能由数据中心提供。

6.2 应用系统与工具

1. 参考型数据库检索系统

提供参考型数据的简单，高级检索方式和数据浏览功能。

2. 访问记账信息采集

通过调用数据中心的访问记账接口，基于 web 提供用户针对本参考型数据库的访问记账信息。

数据中心门户系统提供统一的数据记帐接口，该接口以 javascript 提供。调用方式如：

```
http://log.csdb.cn/log.js?dataset=<datasetID>&action=<action>
```

参数含义如下：

参数	类型	说明
datasetID	in	数据集的标识
action	in	操作，包括： view：查看当前数据集 query：查询当前数据集 download：下载数据

参考型数据库在页面中包含该 javascript 文件即可，包含形式如下：

```
<script language="javascript" src="http://log.csdb.cn/log.js?datasetID= cn.csdb.plant.a
&action=view"></script>
```

6.3 接口规范

1. 数据访问接口--一条记录

根据数据的唯一标识（uri）获得该数据具体内容。

名称：getRecord

参数

参数	类型	说明
uri	in	数据的唯一标识
body	out	包含该数据具体信息的 XML

Xml 格式定义：

<record>

<metadata>

<fields>

```
        <field>
        <name>name</name>
        <type>string</type>
        <title>名称</title>
    </field>
    <field>
        <name>character</name>
        <type>string</type>
        <title>形态特征</title>
    </field>
    ...
    <field>
        <name>periods</name>
        <type>collection</type>
        <title>生长期</title>
    </field>
    </fields>
</metadata>
<data>
    <fields>
      <field name="name">油松</field>
        <field name="character">乔木，高达 25m，胸径约 1m 余</field>
        <field name="periods">
        <records>
            <metadata>...</metadata>
            <data>记录 1...</data>
            <data>记录 2...</data>
            <data>记录 3...</data>
            <data>记录 4...</data>
            </records>
        </field>
    </fields>
    <link>http://xxx.csdb.cn/pines?id=chinesepine</link>
</data>
</record>
```

2. 数据访问接口—多条记录

根据数据的查询关键字（keyword）获得多条数据。

名称：getRecords

参数

参数	类型	**说明**
Keyword	in	查询关键字
from	in	起始条数
size	in	结果条数

body	out	包含该数据具体信息的 XML

Xml 格式定义：

```
<records>
    <range>
        <from>1000</from>
        <size>20</size>
        <total>12345</total>
        <next>1021</next>
    </range>
    <record>记录 1</record>
    <record>记录 2</record>
    <record>记录 3</record>
    <record>…</record>
</records>
```

3. 服务状态监控接口

获取参考型数据库当前服务状态。

名称：getStates

参数

参数	类型	**说明**
body	out	包含该参考型数据库状态信息的 XML

6.4 接口格式要求

6.4.1 通用格式定义

本部分所讨论的接口主要包括 Web Service 接口、HTTP GET 访问接口。

本部分中所有 Web Services 接口返回的信息皆采用 XML 格式来封装，XML 格式如下：

```
<response>
    <head>
        <code>状态码</code>
        <info>对状态码的注解，如：指定的用户不存在！</info>
    </head>
    <body>
        返回信息
    </body>
</response>
```

如果成功访问，所有接口返回的信息封装在<body></body>中。

通用状态码如下表所示：

状态码	说明
200	服务调用成功
300-399	待定
400-499	数据中心门户系统服务错误信息
400	服务调用失败（不可控制）
401	用户信息验证失败（CheckUser）
402	加密方式未知（CheckUser）
403	用户未知
500-599	主题库服务错误信息
600-699	专业库服务错误信息
700-799	参考型数据库错误信息
800-899	专题数据库错误信息

6.4.2 开放接口的安全性要求

出于安全性考虑， 所开放的服务都需要进行访问控制，增加如下参数用于安全性校验：参数：

参数	类型	说明
site	in	访问来源标识，分别对应于参考型数据库和数据中心的 uri

服务接口的提供端应根据该 site 标识以及来源 IP 地址信息进行身份验证。

数据库访问接口应只对数据中心和授权访问数据的应用开放，并建议分别对其提供独立账号。

7 服务

构建丰富便捷的应用服务是参考型数据库的建设目标。承担建设单位应充分利用计算机和互联网技术条件，以参考型数据库内容为基础，为用户提供丰富的应用。参考型数据库原则上面向个人用户和数据库应用提供完全免费的公开共享。

- 提供用户多种直接进入数据库的方式，至少应包括通过元数据访问、搜索、分类体系等方式；
- 关系型数据库数据查询、下载服务；
- 非关系型数据对象基于元数据查询、下载服务；

- 符合数据类型特点的数据展示；
- 基于元数据、分类体系等的数据导航；
- 为下载数据提供数据符合公开声明的数据交换格式；
- 专业数据应用工具的下载。

参考型数据库数据服务应遵循以下基本原则进行：

- 规范化原则：数据建设、共享和服务优先采用国家、行业标准，积极采用国际标准，并结合实际应用制订相关标准，确保数据在最大范围内实现有效的交换共享；
- 共享最大化原则：参考型数据库提供完全免费的公开共享。承建单位应确保不设置无谓的壁垒，使共享在最大范围内进行。（举例如下载操作需注册用户权限，则登陆操作应设置在下载这一环节之前，而此前的查询操作不应强制用户登陆系统。）
  - 确保所有元数据不需注册即可查询、浏览；
  - 确保不低于总量 20%的数据在互联网上提供无需注册即可获得的查询、浏览；
  - 确保全部数据可实现用户下载。
- 网络化原则：任何发布数据，在可以使用互联网发布的情况下，必须使用互联网实现共享与发布。

7.1 服务对象

参考型数据库的服务对象主要包括个人用户和数据库应用两类：

- 个人用户：面向个人用户提供的查询、浏览和下载等服务；
- 数据库应用：面向内容相关的数据库应用提供的服务接口。

7.2 服务方式与要求

参考型数据库的服务方式主要包括在线服务和离线服务两类：

- 在线服务：以基于互联网的方法提供服务的形式，主要包含 www 服务或 web 服务等；
- 离线服务：在线服务以外的其他离线方式提供服务的形式，可包含光盘寄送等。

7.2.1 在线发布方式

参考型数据库承建单位必须建立数据服务系统，应以完全免费的方式至少提供以下服务：

- 建立参考型数据库服务网站，提供数据或非关系型数据库元数据的查询和浏览（对于非关系型数据库，还应包括下载）；
- 面向数据库应用提供基于 web 服务封装的服务。

7.2.1.1服务网站规范

参考型数据库网站建设应满足以下要求：

7.2.1.1.1 页面风格

- 页面尺寸至少支持分辨率 1024×768（页面实际尺寸 1000×618 px）或 800×600 （页面实际尺寸 780×600 px）；
- 页面风格整体一致，与科学数据库主页相近，符合对科学数据库网站页面设计的有关要求：
  - 页面布局及色调应简洁、大气、明快；
  - 文字字体及大小以清晰、协调、突出重点为原则；
  - 各页面页头统一采用中国科学院科学数据库的 LOGO 且须突出显示（建议优选左上角），LOGO 须与页面融为一体，搭配协调，LOGO 上有回科学数据中心网站的连接；
  - 明确呈现参与本主题库建设的所有单位，尊重各方的权益；
  - 页脚应包含参考型数据库的版权声明，内容如下：

    版权所有：中国科学院科学数据库

    备案序号：京 ICP 备 05036949 号

    资源建设维护单位：XXXXXXXX　　电话：XXXXXXXXXX

    地址 ：XXXXXX　邮编：XXXXXXXXXX
  - 数据中心发布页面风格模版后，可套用相关模版。

7.2.1.1.2 内容组织

- 为用户提供检索查询功能，并保证用户对查询结果的访问；
- 特别应包含：
  - 返回数据中心门户系统的链接；
  - 全科学数据库范围内元数据搜索；
  - 建设单位（含承建单位和参建单位）列表；
  - 合作或得到支持的声明，等。

7.2.1.2WEB 服务

参考型数据库及其各组成部分应提供相应的 web 服务，应完全遵照《元数据访问服务接口规范》、《数据跨域互操作技术规范》、《跨域用户认证接口规范》定义服务或接口方法，并实现部署，以支持应用的调用，并保证能够提供 7*24 小时可靠服务。

7.2.2 离线发布方式

承建单位应配备相应的设备、软件和介质，为用户提供数据光盘复制服务和数据定制等离线方式服务。

7.3 数据交换格式

参考型数据库应尽可能遵照国际标准、国家标准、行业标准或科学数据库标准规定的格式提供数据服务。

7.4 共享分级分类设置

- 原则上参考型数据库应以完全免费服务为主；
- 承建单位可根据用户及其对数据使用方法的性质设定一定的用户权限分级，但保证各级用户能够获取与其身份一致的服务。

7.5 其他服务要求

- 参考型数据库承建单位必须建立数据服务系统提供服务，为用户数据下载、数据浏览、数据查询检索、元数据查询检索等服务。
- 可按照一定程序提供离线数据服务，如提供数据光盘等。
- 配备至少一名专门的服务人员，并在数据中心门户系统中注册为咨询员，为用户提供与参考型数据库有关的咨询服务及更深层次的服务。咨询员每周累计至少 1.5 天提供实时咨询服务，对用户非实时咨询的响应时间不超过两天。
- 参考型数据库数据服务系统可以将数据中心门户系统上的参考咨询系统链接为其数据服务系统的一个栏目。
- 为了保证数据中心门户系统实现对参考型数据库服务的集成，参考型数据库承建单位应将参考型数据库注册到数据中心开发的资源与服务注册系统中，并向数据中心门户系统开放符合 6.2.3 小节规定的接口。

- 因数据原始质量问题或非人为故意的数据错误造成用户相关损失的，科学数据库及参考型数据库承建单位不承担赔偿责任。对人为故意造成数据错误进而导致用户损失的，科学数据库及参考型数据库承建单位应向用户道歉，承担道义责任，并有关规定对相关责任人做出处理。如存在人为故意错误的数据属于收取数据费用（不包括服务费）的数据，参考型数据库承建单位应退还所收取的数据费用，并向用户做出与收取的数据费用相等的赔偿，但不承担由于数据错误导致用户的其它损失的赔偿。

7.6 服务案例

- 参考型数据库建设单位应按照数据中心提供的服务案例模板积累和整理服务案例，并在数据中心门户系统的服务案例管理子系统中填写和发布。公开发布的服务案例应得到客户认同意。每年公开发布案例的个数应不少于 2 项。服务案例模板大纲如下表：

案例名称	
服务项目/课题/用户描述	
服务需求	
服务类别	
提供服务单位	
主要利用的数据库/科研应用服务系统	
主要利用的软件工具	
服务响应情况	
服务成效	
服务时间	
需求联系人	
可否公开	

8 运行维护

参考型数据库的运行维护通常涉及机房管理，服务器、网络设备、存储设备及其他必要硬件设施的管理，相关操作系统、数据库系统及应用系统等软件系统的管理，系统用户管理、网络安全管理、数据库备份管理、磁盘监控与整理等系统安全管理，数据更新等工作。以下

若干要求都是从数据应用环境建设和服务项目管理角度提出的对参考型数据库运维的要求，目的是使参考型数据库的运维作为参考型数据库建设工作的重要组成部分能够被考核，促使参考型数据库的运维达到较高的水准，并保障数据应用环境建设和服务项目在总体上的运维和服务水平。为了保障参考型数据库运行维护工作的顺利开展，参考型数据库承担建设单位可制定应用于本参考型数据库的运维制度和规范，在满足从数据应用环境建设和服务项目管理角度提出的对参考型数据库运维的下述各项要求的基础上，实现对参考型数据库运维工作的规范化管理。

8.1 运维人员

参考型数据库承担建设单位应组织成立健全的运维人员队伍（至少配备一名以上工作人员），定人、定责和定规，承担参考型数据库各项运维工作。参考型数据库的运维人员队伍是整个数据应用环境运维队伍的一个有机组成部分。运维人员队伍中应安排一名总负责人，并作为数据中心与参考型数据库在运维工作方面的联络人注册到在数据中心部署的“科学数据库站点服务状态监控系统”中。

8.2 基础运行环境

8.2.1 机房

参考型数据库信息基础设备应具有良好的工作环境，电源要有良好的接地，并具有防尘、防磁、防静电保护，抑制和防止电磁泄漏，以及良好的电源设备环境、防火、防水等备灾设施和条件，机房工作场所应符合下述国家标准所要求的建设标准：

- GB9361-1988　计算站场地安全要求
- GB6650-1986　计算机机房用活动地板技术条件
- GB50016-2006　建筑设计防火规范
- GB2887-2000　电子计算机场地通用规范
- GB50174-1993　电子计算机机房设计规范
- SJ/T30003-1993　电子计算机机房施工及验收规范
- GB7450-1987　通信机房静电防护通则
- GB17859-1999　计算机信息系统安全保护等级划分准则

8.2.2 互联网接入环境

具备 20M/bps 以上的至 INTERNET 出口带宽，并实施必要的网络安全保障措施。

8.2.3 网络服务器与存储设备

- 参考型数据库在线服务的服务器等主要硬件设备应放置于机房中，且机房具备必要的防火、防水等备灾设施和条件；
- 具有必要的网络相关设备、数据库服务器、Web 服务器、数据存储设备和其它必要的硬件设施；
- 服务器和网络能力应至少满足 40 个并发用户访问的需要。

8.3 运行

8.3.1 运行模式

参考型数据库服务网站在线服务应保证 7×24 小时开机运行，全年因故中断运行时间不得大于 5%。[1]

8.3.2 日志管理

参考型数据库应利用日志文件或其它方式对用户访问情况进行记录，并保障数据中心及时获得用户访问情况记录。向数据中心提供用户访问情况记录的方式包括以下四种，参考型数据库承担建设单位应根据参考型数据库的实际情况采用其中的几种或全部。[2]

- 记录和上传参考型数据库服务网站 Web 日志
  - 部署数据中心提供的日志上传工具，在日志上传工具中配置网站 IP 地址、日志保存位置及有关参数，并将 Web 日志上传频率设置为“每日”，将参考型数据库服务网站的 web 日志每日自动上传到数据中心。
  - 为了便于访问统计工具对日志中信息的统计分析，参考型数据库服务网站 Web 日志应遵循一定的格式规范。具体要求是：
    - IIS 服务器设置的 Web 日志格式中必须包含下列日志字段（Web 日志中包含的字段可以多于但不能少于下述字段）：

[1] 数据中心将通过科学数据库“站点服务状态监控系统”对各专业数据库服务网站的运行情况进行 7×24 小时监控，并将监控记录归档，作为考核各数据库正常运行率的依据。

[2] 数据中心数据中心将每个月做一次关于整个科学数据库系统及其包含的各个主题数据库、专题数据库、参考型数据库和专业库的用户访问情况的统计分析报告，并将报告报送院信息办。各参考型数据库的用户访问统计信息也将作为评价其服务效果的一部分重要指标。每个参考型数据库的用户访问统计信息将是对参考型数据库服务网站、包含的关系数据库、包含的 FTP 服务系统（如果有）、离线服务（如果有）等各部分访问统计信息的综合。

日期 date
时间 time
客户IP地址 c-ip
用户名 cs-username
方法 cs-method
URI资源 cs-uri-stem
协议状态 sc-status
发送字节数 sc-bytes
协议版本 cs-version
用户代理 cs(User-Agent)
参照 cs(Referer)

- RESIN 和 APACHE 或者 TOMCAT 的服务器，使用默认日志格式。

■ 记录和上传参考型数据库有关的 FTP 日志

◆ 若参考型数据库中的数据通过 FTP 提供服务，那么既应上传参考型数据库服务网站的 Web 日志，也应上传有关 FTP 的日志。

◆ FTP 日志的上传也是通过数据中心提供的日志上传工具，在日志上传工具中配置 FTP 地址、日志保存位置及有关参数，并将 FTP 日志上传频率设置为“每日”。相关 FTP 的日志也将每日自动上传到数据中心。

◆ FTP 日志也应遵循一定的格式规范。对于 FTP 日志的统计，目前仅限于支持 wu-ftpd，vsftpd 的 FTP 日志，且 FTP 日志必须采用 xferlog 的格式，至少包含如下字段的信息，一般只需在 FTP 服务器配置文件中将日志格式设置成 xferlog 格式即可，默认字段内容，无需定义其各字段信息。

时间	TIME
IP来源	IP
字节数	BYTES
文件名	FILE
上传下载模式	DIRECTION
用户名	USER
完成状态	COMPLETE_STATUS

■ 针对关系数据库的访问情况，按照 6.2 小节中的规定调用数据中心门户系统提供的记账接口。

■ 上报数据离线服务情况

◆ 对于通过非网络访问形式向用户提供的数据，应于每月第一周将上个月的数据提供情况上报数据中心。

◆ 数据离线服务情况按照模板整理和填报。

用户联系信息	
数据提供方式	
数据提供量	
数据提供时间	

8.4 安全保障和故障处理

参考型数据库的安全管理工作依据国家有关法规及《计算机信息网络国际联网安全保护管理办法》进行。参考型数据库承担建设单位必须采取有效措施保障其WEB服务器、数据库服务器、应用服务器的安全，建立必要的防火墙系统，加强对黑客攻击的防护，建立及时更新的防病毒系统，保护系统和数据库的安全。

8.4.1 基础设施安全

参考型数据库及其数据服务系统应具备性能较为完善的网络信息安全设施，包括：网络防火墙、入侵检测、病毒防范、用户识别等信息安全软硬件系统，并设专人进行日常管理监控与更新；

8.4.2 软件安全

- 系统软件（包括操作系统、数据库系统）和应用软件应定期进行完全备份，系统软件的配置修改和应用软件的改动都要及时备份，并做好相应的记录文档。
- 及时了解系统软件和应用软件厂家公布的软件漏洞，并立即进行更新修正；安装入侵检测系统，对网络攻击和非法扫描实时检测、及时报警；
- 应用软件的开发要有完整的技术文档，源代码要有详尽的注释。

8.4.3 数据安全

- 所有科学数据资料分类妥善保存；
- 所有入库的科学数据资料都要按照预定备份策略进行备份，包括异地备份，确保在任何情况下数据都不丢失。
- 对外提供科学数据资料要依据国家有关保密和知识产权法律法规。

8.4.4 非技术防护措施

- 参考型数据库承建单位可在项目规范和国家标准的指导下制定并遵循运维制度开展日常运行工作，具体可包括：
  - 机房管理制度；
  - 值班制度；
  - 系统维护制度；
  - 运行操作规程；
  - 技术档案管理制度；
- 参考型数据库承担建设单位应制订应急工作预案，对故障恢复相关事宜做出应急处置规定。参考型数据库运维应急工作预案应写成文档，并提交数据中心备案。

8.4.5 故障处理

当发现参考型数据库出现故障而不能正常服务时，参考型数据库运维人员应对故障进行及时处理，使参考型数据库尽快恢复正常运行，有关时间要求如下：

- 对于影响很小的一般故障，若在工作日的 8-17 时出现，那么应在 8 个小时内使参考型数据库恢复服务；若在工作日的 0-8 时、17-24 时出现故障，那么应在 1 天内使参考型数据库恢复服务；若在双休日、节假日出现故障，那么应在 2 天内使参考型数据库恢复服务。
- 由于服务器软硬件损坏、黑客攻击、病毒感染等原因导致参考型数据库数据服务系统出现故障致使停止服务时，应尽可能在 2 个工作日内恢复参考型数据库数据服务系统运作。
- 如果参考型数据库出现的故障是不能短时间恢复的，运维负责人应在规定时间（自出现故障起 2 个工作日内）提请数据中心启动紧急预案，由数据中心运维人员根据应急工作预案启动在数据中心的备份系统，代替原系统提供服务直至原系统恢复正常。

出现故障的起始时间以科学数据库站点服务状态监控系统监测到参考型数据库无法正常访问并向参考型数据库运维负责人发出服务异常通知的时间为准。

8.5 备份和恢复

参考型数据库数据服务系统应具备一定的容灾能力，除可在本地备份外，还必须备份到数据中心，以保障参考型数据库服务的持续稳定开展。

- 参考型数据库承担建设单位应制定明确的数据备份计划，并上报数据中心备份。参

考型数据库除在本地对数据库进行备份外，还应定期（至少每半年一次）或不定期（每有数据更新时）将数据备份到数据中心保存，以便在需要时能够由数据中心的备份系统代替原系统提供服务；

- 参考型数据库中的数据及应用都需要在数据中心备份，可利用数据中心提供的备份工具实现这些资源的备份；
- 参考型数据库应用系统应能够在数据中心提供镜像服务。

8.6 运维服务的质量

运维服务质量主要指参考型数据库为用户提供数据服务系统正常服务率、数据服务系统可用性、人工干预型服务的响应速度和用户满意度、用户访问情况和服务案例数量，共同构成参考型数据库的服务质量。

对于参考型数据库而言，其成功服务于相关数据库建设、数据库改造或数据集成的评价权重应相对大于面向个人用户的服务权重。

参考型数据库的运行维护应满足以下要求：

- 数据服务系统正常服务率：要求不低于 95%，将通过站点服务状态监控系统中的监控统计数据进行评价。
- 数据服务系统可用性：利用关于数据服务系统可用性的在线调查问卷得到评价数据；
- 人工干预型服务的响应速度和用户满意度：利用参考咨询系统中的统计数据进行评价。
- 用户访问情况：根据对参考型数据库用户访问情况的统计分析数据进行评价。
- 服务案例数量：根据数据中心门户系统服务案例管理子系统中的数据进行评价。

关于参考型数据库服务质量的其他详细规定，见数据应用环境建设和服务项目发布的《共享服务评价指标体系》。

附录 A（规范性附录）规范实施一致性测试

为了保证本规范在中国科学院数据应用环境建设与服务项目建设中的实施，充分发挥其服务参考型数据库建设的重要作用，特别是参考型数据库内容组织、资源建设、运维和服务方面各项具体要求的落实，下文特明确规范实施一致性测试之具体内容，满足本测试所有项目者即视为参考型数据库建设工程中贯彻实施了本规范。

本规范从内容组织、资源建设、接口规范、运维与服务等方面规范了中国科学院数据应用环境建设与服务项目中参考型数据库的建设，规范实施一致性测试也针对上述内容分别予以明确。

A.1 内容组织

A.1.1 数据集名称及标识符

数据集名称是项目内识别建设任务和目标的主要依据，所以应该保证名称在项目内的统一，不应出现“一库多名”的现象，所以在数据集标识符注册的过程应明确各类数据集的名称，并在各类应用中使用同一名称和标识符。具体测试内容如下：

- 经数据中心认证，成功注册参考型数据库，确认数据库名称，并获取其唯一标识符；
- 确定子库名称，根据唯一标识符命名规则[3]，在数据中心为参考型数据库各级子库注册唯一标识符；
- 在项目建设和服务过程中，完全使用所注册的参考型数据库及其子库名称、唯一标识符一致，并保持持久性；
- 参考型数据库及其子库名称、唯一标识符在项目内具体实现和服务的情况将通过有关工具软件自动检测，如存在不一致视情况定论。

[3] 数据集唯一标识符命名原则

1. 标识字符组成：
   - 26 个英文字母，不区分大小写；
   - "0，1，2，3，4，5，6，7，8，9"十个数字 ；
   - 字符“-”、“_”、“~”。
2. 对于本项目内部所有数据集的标识，统一规范为：

   DatasetURI：＝cn.csdb.<三级域名> [.<子库实体标识>]

   其中，cn.csdb.<三级域名>为各建库单位在数据中心注册所得，与各建库单位承担的数据库建设的主体任务是一一对应的，且其确定后将相对固定，用户不得自行更改；而后续字符串为用户自主命名，需保证其内部唯一。

A.1.2 数据库建立

参考型数据库应立足定位，从学科领域对参考型数据资源的需求出发建设数据库，此间应该实现如下目标：

- 在兼容领域内相关国家或国际规范，国内外权威数据数据规范等的基础上建立参考型数据库数据模型，并将国内外同学科领域标准或数据规范的分析形成文档（也可包含在《数据应用环境建设和服务 数据库元数据需求规格书》中），并提交数据中心备案；
- 数据库建设所所遵循的标准规范必须在元数据、数据服务系统页面上明确说明；
- 建设统一的数据编码标准，统一的数据录入、存储、输出标准；
- 应向数据中心门户系统开放符合 6.3 节“接口规范”的接口。

文档

在建设过程中，关系型数据建设除了通过技术实现资源的整合和数据入库外，还应完成以下文档：

- 《数据应用环境建设和服务 数据库需求说明书》；
- 《数据应用环境建设和服务 数据库元数据需求规格书》，包括对领域内已有数据标准规范、相似的权威性数据库和建库相关学科背景知识等内容的分析；
- 《数据应用环境建设和服务 数据库设计说明书》。

元数据

为参考型数据库及其各级子库著录符合核心元数据、系统元数据的元数据，并按照要求向数据中心门户系统开放符合“6.3 接口规范”要求的接口，以便实现元数据的注册和同步，支持核心元数据和领域元数据通过参考型数据库服务网站和数据中心门户为用户服务，且不限制非注册用户对元数据的直接访问；而系统元数据应按照数据中心的要求提供给数据中心，以便数据中心门户可以通过该元数据实现对数据的直接访问。

A.1.3 数据处理

对参考型数据库而言，数据处理是实现高质量保证的重要一个环节，通过映射、清洗、转化等实现资源的净化处理，在该过程中应将数据资源整理过程及其关键内容（节点、事项、方法等）形成数据组织的文档，其中包括数据映射关系的详细记录，文档记录内容应以所有活动的可重复再现为基本目标。

文档需提交数据中心备案。

A.1.4数据质量

数据质量指标能够实现数据质量状况综合反映，参考型数据库建设单位应该与本学科领域专家一起确定质量指标，特别应从“5.4.1 数据质量指标”中选用指标，并根据资源的特点确定质量控制、评价方法，并将包括上述内容在内的数据质量活动完整记录到质量报告的文档中，用户基于质量文档应能够判断数据对其需求的满足程度，质量文档应遵循以本学科领域内公认的内容和格式，亦可参照与数据中心商定的质量文档模版，文档记录内容应以所有活动的可重复再现为基本目标。数据质量报告至少应包含以下内容：

- 数据质量指标
- 数据质量控制方法及执行情况
- 数据质量评审方法极结果

参考型数据库质量报告需提交数据中心备案，也应将报告中相关主要内容作为元数据发布。

参考型数据库应在人工抽查、工具软件检查或二者结合检查时达到如下基本要求：

- 准确率
  - 不低于 99%
  - 完整性
  - 数据库的核心属性（学科领域专家讨论确定）不能空，非关系型文件的元数据库的核心元素不能为空
  - 关系数据库的填充率不能低于 70%
- 重复率
  - 不超过 1%
  - 此外，数据质量相关的具体内容和方法，执行数据应用环境建设和服务项目发布的《数据质量评测方法与指标体系》、《共享服务评价指标体系》。

A.2 技术架构与接口规范

基于 “6.3 接口规范”，实现各级数据组织架构的功能、系统建设，并开放符合本规范的各类数据接口，支持彼此间的数据交互。

特别在技术实现数据服务系统的过程中，应完成以下文档：

- 《数据应用环境建设和服务 数据库软件概要设计说明书》
- 《数据应用环境建设和服务 数据库软件详细设计说明书》
- 《数据应用环境建设和服务 数据库软件开发卷宗》

A.3 服务

按照 7.2 之“服务方式与要求”、7.4 之“其他服务要求”实现参考型数据库的应用服务环境建设。

至少配备一名专业的数据服务人员，并在数据中心门户系统中注册为咨询员，为用户提供与参考型数据库的应用有关的咨询服务及更高层的支持，其工作时间每周累计至少 1.5 天，且对用户非实时咨询的响应时间不超过两天。

按照服务案例模板（7.6）积累和整理服务案例，并在数据中心门户系统的服务案例管理子系统中填写和发布。公开发布的服务案例应是得到客户认可并同意发布的；每年公开发布案例的个数应不少于 2 项。

A.4 共享

原则上参考型数据库应面向个人用户和数据库应用提供完全免费的公开共享，具体数据的分发和共享应遵照《中国科学院数据应用环境建设和服务数据共享办法》的要求执行。特别应实现：

- 各类元数据完全公开，支持非注册用户的访问；
- 完全公开共享的数据中，应有 20%以上的资源允许非注册用户的直接访问和获取。

A.5 运行维护

运行维护是数据资源建设和服务的保证，应实现以下要求：

- 组建专业的运维人员队伍（至少配备一名以上工作人员），定人、定责和定规，承担参考型数据库的各项运维工作；应安排一名总负责人，并作为数据中心与专业数据库在运维工作方面的联络人注册到在数据中心部署的“科学数据库站点服务状态监控系统”；
- 构建与 8.2 所明确的参考型数据库基础运行环境；
- 在线服务网站应保证 7×24 小时运行，全年因故中断运行时间不得大于 5%；
- 按照数据中心的部署上传（www、FTP）日志文件或其它资源访问或使用记录（离线服务、应用案例等），保障数据中心及时获得用户访问和应用情况；
- 按照第 8 章“参考型数据库的运行、安全保障和故障处理、备份和恢复”实现参考型数据库的运行、安全保障和故障处理、数据备份和系统镜像；
- 完成《数据应用环境建设和服务 数据库运行维护记录》文档并提交至数据中心。