科学数据共享工程技术标准

SDS/T 2133—2004

数据模式描述规则和方法

Rule and Method for Description of Data Schema

（征求意见稿）

（本稿完成日期：2005 年 5 月）

××××-××-××发布　　××××-××-××实施

中华人民共和国科学技术部　发布

目　　次

目　　次 I
前　　言 II
引　　言 III
数据模式描述规则和方法 1
1 范围 1
2 规范性引用文件 1
3 术语和缩略语 1
3.1 术语 1
3.2 缩略语 3
4 概述 3
4.1 数据模式简介 3
4.2 数据模式作用 5
4.3 数据模式语言 9
4.4 共享科学数据模式 10
5 数据模式描述规则 12
5.1 概述 12
5.2 UML 图形式描述规则 12
5.3 字典形式描述规则 18
6 数据模式建立与描述方法 19
6.1 建模方法概述 19
6.2 需求收集 20
6.3 数据模式建立与描述 26
6.4 标准协调 31
6.5 标准实现 32
附　录　A 数据模式建立与描述的文档模板 (资料性附录) 34
A.1 概述 34
A.1.1 数据模式标准的需求收集文档模板 34
A.1.2 数据模式标准的草案文档模板 34
A.1.3 数据模式标准的正式文档模板 34
A.1.4 数据模式标准的征求意见文档模板 34
A.1.5 数据模式标准的意见处理文档模板 34
A.2 数据模式标准的需求收集文档(资料性附录) 36
A.3 数据模式标准的草案文档(资料性附录) 45
A.4 数据模式标准的正式文档(资料性附录) 56
A.5 数据模式标准的征求意见文档(资料性附录) 69
A.6 数据模式标准的意见处理文档(资料性附录) 75
附　录　B 反向工程示例 (资料性附录) 81

前　言

本标准为第一次制定。
本标准由中华人民共和国科学技术部基础研究司提出。
本标准由中华人民共和国科学技术部基础研究司归口。
本标准主要起草单位：国家信息中心、中科院地理所。
本标准主要起草人：徐枫、游松财、宦茂盛、林菁、武晋平、石雯雯、吕明。

引　言

根据科学数据共享工程对主体数据库的建设要求，各国家科学数据中心和国家科学数据网已经具备大量的科学数据资源。这些科学数据资源将以主体数据库形式进行建设，以科学数据共享数据集形式进行共享。对科学数据共享数据集的内容进行规范化和标准化描述是真正实现科学数据共享的基本前提。

通过数据模式，各个领域能够准确描述和理解科学数据共享数据集的内容，生产、加工出符合科学数据共享工程需求的数据集，进而保证科学数据共享活动的实现。

本标准明确了各领域共享数据集的描述方式，使数据集制作人员及数据集用户对共享数据集内容有准确而一致的理解。同时，本标准提出了建立数据模式的具体方法。

数据模式描述规则和方法

1 范围

本标准提出了的构建各数据集的数据模式的规范化描述方式、表示和操作的步骤，规范了各个领域里数据模式的制定，使数据集制作人员及数据集用户对共享数据集内容有准确的理解。

本标准适用于各领域制定科学数据共享数据集内容模式时使用，保障数据集生产者及数据集使用者对共享数据集内容能够无歧义的理解。

本标准也可以用于一般数据内容建模。

2 规范性引用文件

下列文件中的条款通过本标准的引用而成为本标准的条款。凡是注日期的引用文件，其随后所有的修改单（不包括勘误的内容）或修订版均不适用于本标准。然而，鼓励根据本标准达成协议的各方，研究是否可使用这些文件的最新版本。凡是不注日期的引用文件，其最新版本适用于本标准。

SDS/T 2132—2004 数据元标准化原则与方法

SDS/T 2134—2004 数据交换格式设计规则

SDS/T 2321—2004 科学数据中心建设规范

SDS/T 2322—2004 科学数据网建设规范

Unified Modeling Language 1.5 统一建模语言**1.5**

3 术语和缩略语

本标准采用下列缩略语和术语定义。

3.1 术语

3.1.1.

数据模式 **Data Schema**

数据的概念、组成、结构、相互关系的总称。

注： 从本质上，数据模式反映的是人类对客观世界的主观认知。在具体内容上，数据模式涉及到数据的描述范围、描述的方式和描述的结果。

3.1.2.

概念数据模式 **Conceptual Data Schema**

通过抽象、归纳、概括、分类等各种方法，对客观世界的现象进行概括性的描述，重点是定义客观世界的各种基本实体，并对它们的相互关系进行描述。

3.1.3.

逻辑数据模式 **Logical Data Schema**

概念数据模式的细化，在逻辑数据模式中，考虑到信息技术实现的因素，需要对概念数据模式进一步分析，并增加各种对象和事件，作为物理数据模式建立的基础，逻辑数据模式和具体实现无关。

3.1.4.

物理数据模式 **Physical Data Schema**

逻辑数据模式集合了具体的实现技术后形成的，它和具体的实现技术紧密相关。

3.1.5.

数据模式语言 **Data Schema Language**

用于对数据模式进行分析、构造、表现和记录的语言。

3.1.6.

实体 **Entity**

任何可以明确的人、地方、事件、概念、事物。

3.1.7.

属性 **Attribute**

描述或标识实体的实体或值。

3.1.8.

值域 **Domain**

属性可以取值的范围。

注： 值域是单独定义的，用于重用，即多个属性可以使用同一个值域。

3.1.9.

关系 **Relation**

实体间的关联。

3.1.10.

主键 **Primary Key**

对取值给出唯一性限制的一种属性。

注： 所有实体实例的该属性取值不会出现重复。通过该键值可以唯一的确定一个实体。在**UML**中通过设定原型**<<PK>>**标识为“主键”。

3.1.11.

外键 **Foreign Key**

由相关实体的实例指定自身实例的一种属性，是实现一个关系的约束。

注： 在**UML**中通过设定原型**<<FK>>**标识为“外键”。

3.1.12.

包 **Package**

在 UML 中，用于表示实体的组织。

3.1.13.

类 **Class**

对拥有相同的属性、操作、方法、关系和语义的一组对象的描述。

注： 在**UML**中类的图形符号是一个矩形框。其中标注出该类的名称，即为对应实体的名称。

3.1.14.

注释 **Comment**

附在实体、关系上的标注文字。

注： 在**UML**中注释不具有语义和限制功能。

3.1.15.

数据集 **Data Set**

可以标识的数据集合。

3.1.16.

数据元 **Data Element**

通过定义、标识、表示和值域等一系列属性描述的一个数据单元。

3.1.17.

国家科学数据中心 **Scientific data center**

属于国家科学数据共享平台的组成部分。以国家部门、行业系统为基础，按不同科学技术领域建立的社会公益型的科学数据主中心以及根据需要设立的科学数据分中心，统称国家科学数据中心；主要负

责国家长期布局的公益性、基础性科学数据的汇交、管理、交换与共享服务。

3.1.18.

国家科学数据网 **Scientific data network**

是国家科学数据共享平台的组成部分。面向国家重大科技计划、重点区域以及基础科学领域，基于因特网连接分布于各科研院所、高等院校和国际组织的相关专业数据库，开展数据组织、加工与服务，所构建的物理上分布、逻辑上统一的科学数据网。

3.1.19.

主体数据库 **Core database**

依据国际标准、国家标准或行业标准分类体系构建的二级学科及其分支学科的科学数据集，并基于计算机系统运行的数据库。

3.1.20.

接口 **Interface**

是被命名的操作的集合，它们表示一个元素行为的特性。

3.1.21.

数据模式字典 **Data Schema Dictionary**

字典形式从名称、定义、英文名称、英文短名、版本标识、状态、来源、注释等多个方面来描述模型中的实体、属性，从而能够严格的对数据模型中的实体和属性进行描述。

3.2 缩略语

3.2.1.

OMG

对象管理组织，Object Management Group。

3.2.2.

UML

统一建模语言，Unified Modeling Language。

3.2.3.

XML

扩展标记语言，Extensible Markup Language。

3.2.4.

W3C

W3C 组织，World Wide Web Consortium。

3.2.5.

SQL

结构化查询语言，Structured Query Language。

4 概述

4.1 数据模式简介

数据是对客观世界认识的一种表示方式，人类对客观世界的各种现象进行抽象，并通过一定的方式进行组织，最终以数据的形式进行记录。

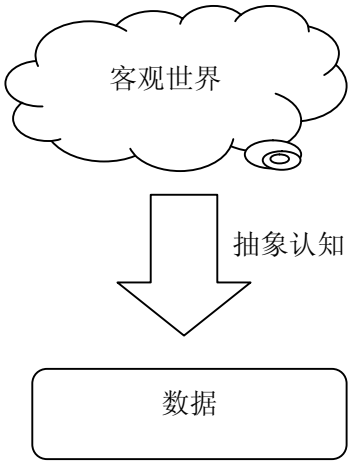


图1　人对世界的认识

数据模式是数据的概念、组成、结构、相互关系的总称。从本质上，数据模式反映的是人类对客观世界的主观认知。在具体内容上，数据模式涉及到数据的描述范围、描述的方式和描述的结果。由于数据模式是人类对客观世界的主观认知，而不同的人群对相同的客观世界的主观认知会有所不同，这就造成了相同领域不同的数据模式存在。在科学数据共享过程中，这种差异对人们进行信息的共享与交换形成了障碍。为了保证能够顺畅的进行信息的共享，对特定领域而言，需要一个统一的数据模式作为数据共享与交换的基础。同时也保证该领域的相关人员对统一的数据模型有准确的、无歧义的理解。

为保证对数据模式的理解一致性，需要分析数据模式的描述层次，进而规定科学数据共享所涉及的数据模式层次。对数据模式的描述规则与方法进行统一的规定，并以此作为基础，进行科学数据各领域数据的共享和交换活动。

分析领域建立数据模式的一般过程，可以明确出数据模式分为三个层次，分别是概念数据模式、逻辑数据模式和物理数据模式。

1) 概念数据模式

主要是通过抽象、归纳、概括、分类等各种方法，对客观世界的现象进行基本的描述，重点是定义客观世界的各种基本实体，并对它们的相互关系进行描述。

2) 逻辑数据模式

是概念数据模式的细化，在逻辑数据模式中，考虑到信息技术实现的因素，需要对概念数据模式进行进一步的分析，并增加各种对象和事件，作为物理数据模式建立的基础，逻辑数据模式和具体实现无关。

3) 物理数据模式

是逻辑数据模式集合了具体的实现技术后形成的，它和具体的实现技术紧密相关。例如关系型数据库技术、**XML**技术等。

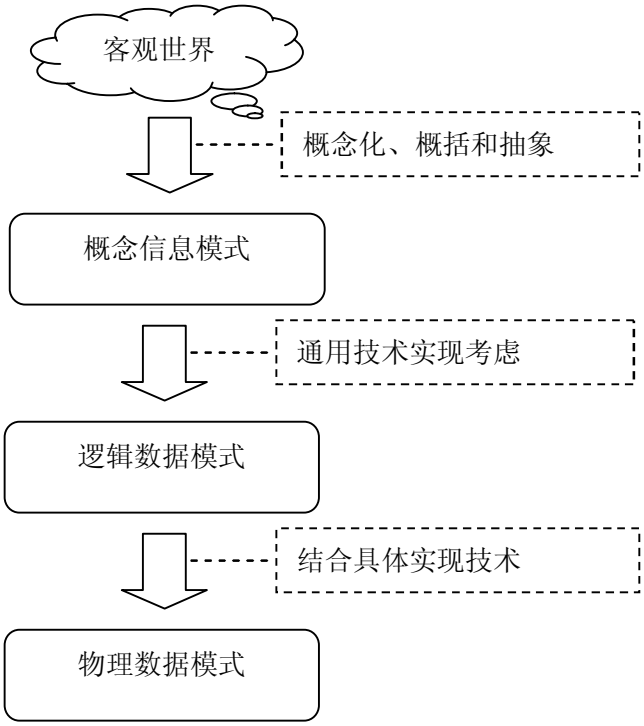


图2　数据模式的三个层次

4.2　数据模式作用

数据模式对于数据集制作,科学数据共享数据服务平台建设以及科学数据共享工程都有非常大的促进作用。

1) 在数据集制作方面的作用

- **n** 有利于标准化数据

  数据模式对数据的内容、组成、结构以及各部分的相互关系都作了相应的规定，相关的领域、部门或者数据集制作者都可以根据数据模式的内容制作出标准化的数据。

- **n** 有利于减少数据冗余

  数据模式中一般对数据的组成进行了精炼，通过数据的各组成部分的复用、关联引用等最大程度了减少的数据冗余。

2) 在进行科学数据数据服务共享平台建设方面的作用

- **n** 有利于减少平台的开发工作量

  统一的数据模式简化了科学数据服务共享平台接口的开发,避免了针对多个不同的数据组织方式开发不同的接口，从而减少了整个科学数据共享数据服务平台的开发工作量。

- **n** 有利于整合的不同的科学数据服务共享平台

  统一的数据组织方式是整合不同数据服务共享平台的基础工作之一。建立统一的数据模式能够保证从各数据服务共享平台中提出的数据有准确而一致的解释。

3) 对于科学数据共享工程的作用

- **n** 有助于扩大科学数据的共享和使用范围。
- **n** 有助于提高科学数据共享的效率。
- **n** 有助于降低科学数据共享的实施风险。
- **n** 有助于减少科学数据共享的成本。

从上面的三个角度看到采用数据模式的优势，可以清晰的体现在示意图中。

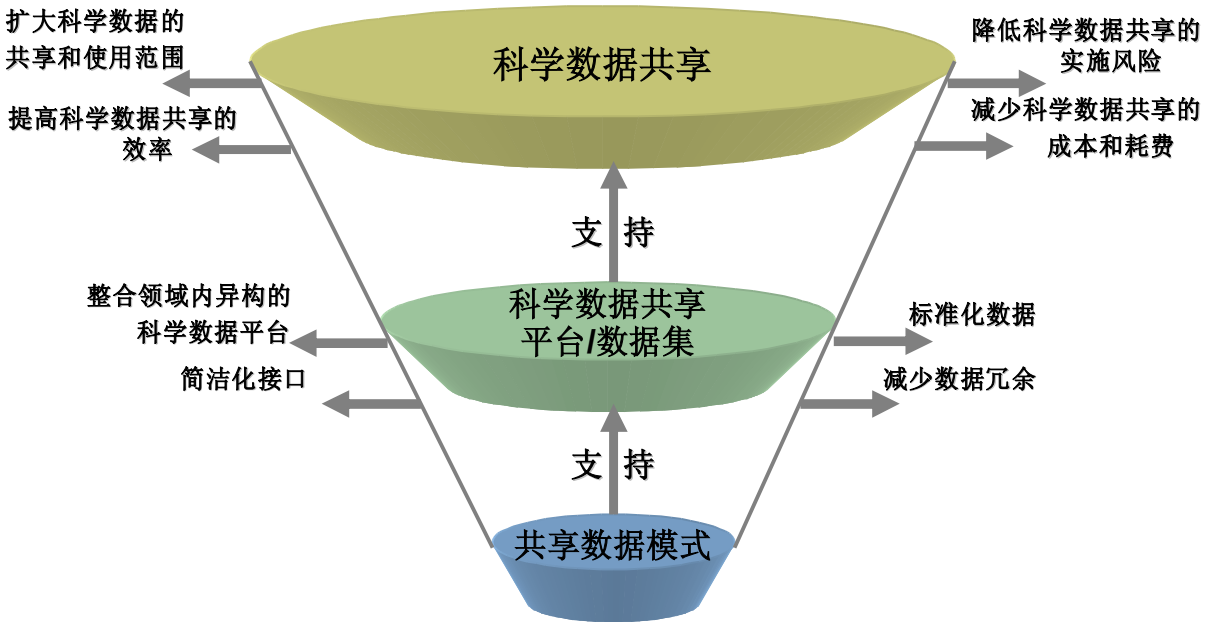


图3　来自数据模式支持的示意图

下面是对于上述优势的具体阐述：

n　利于整合领域内的多个不同的科学数据系统平台

对于本领域内，彼此异构的多个信息系统平台里，科学数据共享数据集的制作和共享过程是十分复杂的。迫切需要领域内通用的共享科学数据模式来提供关于共享数据集的内容、组成和结构信息，从而在本质的逻辑层次上进行理解、整合和分析加工。最终为科学数据共享平台的构建，提供实际可用的共享科学数据集。

考虑领域普遍的现状，针对科学数据共享数据集的内容、组成和结构的描述说明，包括下面的几种情况：

（1）领域内可能已经存在针对于科学数据的交换方面的相关标准、规范（诸如国际标准、国家标准、业界标准、项目规范、单位规范等）含有明确的交换格式。实际上，交换格式隐含了数据的组成和结构的基本方式，但是不适用于直接、完全的应用到本领域的科学数据共享工程建设中。需要针对项目的实际需求，进一步将隐含的数据组成和结构的基本方式确定下来，以此作为基础，吸收到科学数据共享数据集的内容、组成和结构的描述说明体系。

（2）领域内也可能已经存在系统设计、开发类型的文档，包含有实际的字典（诸如表格、文档、文本等），可以描述领域内准备共享的科学数据的组成结构信息，但是可能其描述方式不规范，各种各样的结构方式都有，例如XML SCHEMA、UML图、E-R模型图、摘要文本或者其它的格式。这种只是基于物理实现层次的描述，对于原有的系统内部使用时可能没有问题，但是对于科学数据共享工程来说，在进行交换和共享的时候，科学数据共享数据集的制作者和使用者都需要准确的了解数据的本质内容、组成和结构才能够进行深入的处理、交流和应用。需要先将隐含的数据组成和结构的基本方式明确出来，再对其的描述格式进行重新整理、统一表达，然后筛选、提炼、整合到科学数据共享数据集的内容、组成和结构的描述说明体系里。

（3）领域内还可能包括有历史遗留系统，这些系统是没有字典进行描述的，对于其中的大量数据的查找、使用、驱动是依赖于已经存在的软件系统的运行。对于系统外界的人员是无法直接应用的，特别是对于科学数据共享数据集的制作者和使用者来说，更加无法直接参考和使用。可见这部分数据只是停留在物理层次，更需要规范化，才能使得它们脱离开原有软件系统，独立拥有实际的参考和使用价值；否则，不但不易于理解、分析这部分数据，而且更谈不上对其进行加工、处理和应用。所以，缺乏对于数据内容、组成和结构的描述说明信息，就是

缺少进行科学数据共享的前提、基础，很难进行科学数据共享数据集的制作、交换和共享。需要首先通过相应的工具从这类的软件系统中按照规范化的步骤，提取出必要的数据内容、组织和结构信息，再遵循统一的描述规则进行表达说明、挑选后将必要的信息纳入到科学数据共享数据集的内容、组成和结构的描述说明体系。

n 利于简化科学数据共享平台接口

首先在概念和逻辑层次上，各个领域需要分析科学数据共享平台的需求和整理清楚共享数据集的概念、内容、组成及其结构信息，再建立明确、统一的共享数据模式，才能在物理层次上切实保障实施科学数据共享时有效、合理的简化接口。

科学数据共享工程中，共享数据模式作为简化科学数据共享平台接口的基础和前提，是本领域内所有共享数据的概念、内容、组成及其结构的描述信息的有机组合。它包含有统一的描述规则和规范化的构建方法，来清晰、科学的描述本领域的共享数据集的概念、内容、组成及其结构信息。

需要明确在逻辑层次上可以对科学数据共享数据集的内容、组成及其结构信息，进行合理的、规范的、本质上的说明和描述。共享数据模式的分析和构建工作应该重点放在逻辑层次上的。在逻辑层次上进行描述说明，不但可以基于概念数据模式，具有本质化、清晰化和概念化的特点，而且结合了信息技术实现的因素，进一步的分析并增加各种对象、事件和其关系。同时又能够高于具体的物理实现，拥有更大的灵活性、抽象性和概括性，是作为物理数据模式建立的基础。

实施科学数据共享时，基于统一的共享科学数据模式所制作的科学数据共享数据集，更加利于构建高效、简洁的物理数据模式，从而保障了简化科学数据共享平台接口。

如果科学数据共享数据集是基于彼此异构的多个数据模式，容易导致构建本质上相同的“重复”众多共享数据集，就需要构建、开发繁琐的接口来支持不同数据模式之间的映射和转换。

n 利于减少数据冗余

减少数据的冗余，通常是根据数据本质的内容组成和结构信息，抽取出公共的通用性数据元素作为共用的可复用数据元素。那么在制作的科学数据共享数据集时，就可以按照数据的内在组成信息，分离出共用的可复用数据元素，再建立一组关联到其对应的独特数据元素，从而降低了数据的冗余。

共享数据模式是对于领域中科学数据共享数据集的内容、组成及其结构的规范化描述。它包含有统一的描述规则和规范化的构建方法，来清晰、科学的描述本领域的共享数据集的概念、内容、组成及其结构信息。重点是在逻辑数据模式层次上对科学数据共享数据集的内容、组成及其结构信息，进行合理的、规范的、本质上的说明和描述。不但可以基于概念数据模式，具有本质化、清晰化和概念化的特点，而且结合了信息技术实现的因素，进一步的分析并增加各种对象、事件和其关系。同时又能够高于具体的物理实现，拥有更大的灵活性、抽象性和概括性，是作为物理数据模式建立的基础。

领域内采用共享数据模式可以有效的避免造成同样的事物有错误的互异的数据内容结构描述，从而合理的抽取出在本质上相同的“公共”数据集合元素，同时也有利于识别出潜在的可重用的数据集合元素。最终使得科学数据共享数据集的数据冗余得到降低。

n 利于标准化数据

领域内依据通用的共享数据模式标准，来规范、构建科学数据共享数据集，得到的科学数据共享数据集里的数据必然是标准化的。

各个领域共享的内容是需要标准、规范化的科学数据共享数据集。科学数据共享工程对主体数据库的建设要求，将各科学数据中心和科学数据网已经具备大量的科学数据资源，以主体数据库形式进行建设，形成科学数据共享数据集进行共享。

需要从已有的业务数据系统中，查找、抽取、筛选、制作、最终形成符合科学数据共享工程要求的切实可用的科学数据共享数据集。实际的各个领域内的科学数据是多种多样的。可能存在于数据库、文件、二进制数据、文本数据、图像数据、图形数据、声音、视频、多媒体数据等多种形式。

在制作共享的科学数据集时，如果拥有针对数据内容组成和结构的描述说明，就可以协助科学数据共享数据集的制作者，产出所需的标准化科学数据集。

- 有助于扩大科学数据的共享和使用范围

同样重要的是使用科学数据共享数据集时，对于科学数据共享数据集的用户，来自完整的共享科学数据模式的数据集，是更加有实用价值的。这样的科学数据共享数据集易于准确理解，便于有效使用。尤其是在进行深入的再开发、再处理分析时，就会迫切需要对于数据集本身内容、组成和结构的描述说明信息。

否则没有针对科学数据共享数据集的内容、组成和结构进行说明和描述，用来共享的科学数据集就只是一堆数据，必然会限制它的共享和使用。对于数据集制作人员及数据集的用户来说，不能够进行数据的处理和分析，意味着没有参考和使用价值，必然会限制科学共享数据的共享和使用范围。

- 有助于提高科学数据共享的效率

更为有效、合理、科学的完成从复杂多样的异构业务数据系统到通用的科学数据共享系统的重整、转换、新建、发布过程。

科学数据共享工程中，共享数据模式是本领域内所有共享数据的概念、内容、组成及其结构的描述信息的有机组合。它包含有统一的描述规则和规范化的构建方法，来清晰、科学的描述本领域的共享数据集的概念、内容、组成及其结构信息。

特别加强在逻辑数据模式上对科学数据共享数据集的内容、组成及其结构信息，进行合理的、规范的、本质上的说明和描述。在逻辑层次上进行描述说明，可以基于概念数据模式，具有本质化、清晰化和概念化的特点。是结合了信息技术实现的因素，进一步的分析后增加各种对象、事件和其关系。同时又能够高于具体的物理实现，拥有更大的灵活性、抽象性和概括性，是作为物理数据模式建立的基础。

- 有助于降低科学数据共享的实施风险

由于缺乏数据模式会给科学数据的共享工程带来的很大的实施风险。

有缺陷的数据模式标准会导致科学数据共享数据集的制作有可能不规范，存在错误。缺乏数据模式，易于造成同样的事物有错误的互异的数据内容结构描述，进而构建了本质上相同的“重复”数据集合元素，也可能没有识别出潜在的可重用的数据集合元素。提炼不当的数据模型，也会限制到基于它的科学数据共享数据集的共享和应用范围。

缺乏数据模式会导致科学数据共享数据集的用户拿到数据后无法使用。科学数据共享数据集的用户在使用获取到的共享数据集时，通常是再加工、再处理、再开发的应用过程。没有对于该科学数据共享数据集的内容组织和结构的描述信息，将会使得这个过程隐含了许多风险、隐患和不必要的大量系统资源的消耗。

- 有助于减少科学数据共享的成本和耗费

缺乏数据模式或制定的是错误的数据模式，将会导致进行科学数据共享和交换时，会消耗掉大量宝贵的资源，远超出其实际上的需求量。

科学数据共享工程中，共享数据模式是本领域内所有共享数据的概念、内容、组成及其结构的描述信息的有机组合。它包含有统一的描述规则和规范化的构建方法，来清晰、科学的描述本领域

的共享数据集的概念、内容、组成及其结构信息。

共享数据模式的分析和构建工作的重点是在逻辑层次上，进行合理的、规范的、本质上的说明和描述科学数据共享数据集的内容、组成及其结构信息。在逻辑层次上进行描述说明，不但可以基于概念数据模式，具有本质化、清晰化和概念化的特点，而且结合了信息技术实现的因素，进一步的分析并增加各种对象、事件和其关系。同时又能够高于具体的物理实现，拥有更大的灵活性、抽象性和概括性，是作为物理数据模式建立的基础。

总之，数据模式提供高质量的数据定义和数据格式，依据该数据模式实现的信息系统，可以完成对于领域内科学数据集的交换和使用。根据相同的数据结构和统一的含义理解，进行数据的存储和访问，使得领域内不同的应用系统间合理、有效的共享科学数据集。

4.3 数据模式语言

数据模式语言是用于对数据模式进行分析、构造、表现和记录的语言。在建立各数据集的数据模式的过程中，需要对客观世界的实体进行分析和抽象，定义各种实体及相互关系，通过实体描述整个客观世界的组成，并使用图形、文字等方式来表现和记录。

为对数据模式建立过程、成果形成一致性的理解，必须有一种通用语言来描述数据模式，它能够客观的、无歧义的描述数据模式。这种语言既可以是自然语言，也可以是机器语言，既可以是文字，也可以是图形，或者是文字或图形的混合体。

目前，国际上比较通用的数据模式语言有**UML**，它是目前大多数国际组织和工业联盟采用的数据模式描述语言。科学数据共享工程的数据模式定义应当使用**UML**作为标准的数据模式语言。

UML（统一建模语言，**Unified Modeling Language**）是一种通用的建模语言，它由**Grady Booch**，**James Rumbaugh**和**Ivar Jaccobson**共同提出，该语言由**OMG**（对象管理组织，**Object Management Organization**）采纳作为业界标准。目前，**OMG**正准备将该标准提交到**ISO**，使其成为信息技术方面的国际标准。

UML是一种通用的可视化建模语言，它主要用于理解、设计、浏览、配置、维护以及控制系统的信息。**UML**易于使用、表达能力强，可升级，具有很强的适用性和可用性。**UML**不是编程语言，但它可应用于任何编程语言和工具平台。

作为一种建模语言，**UML**的定义包括**UML**语义和**UML**表示法两个部分。

1) **UML**语义

描述基于**UML**的精确元模型定义。元模型为**UML**的所有元素在语法和语义上提供了简单、一致、通用的定义性说明，使开发者能在语义上取得一致，消除了因人而异的最佳表达方法所造成的影响。此外**UML**还支持对元模型的扩展定义。

2) **UML**表示法

定义**UML**符号的表示法，为开发者或开发工具使用这些图形符号和文本语法为系统建模提供了标准。这些图形符号和文字所表达的是应用级的模型，在语义上它是**UML**元模型的实例。

标准建模语言**UML**的内容主要通过各种视图（**View**）来定义，**UML**的视图可以分为三个层次：

1) 结构性分类视图

用于描述系统中事物之间的关系。分类包括类、用例、构件和节点。分类视图包括静态视图、用例视图和实现视图。

- 静态视图（Static View）包括类图、对象图和包图。其中类图描述系统中类的静态结构。不仅定义系统中的类，表示类之间的关系（如关联、依赖、聚合等），也包括类的内部结构(类的属性和操作)。对象图是类图的实例，使用与类图几乎完全相同的标识。他们的不同点在于对象图显示类的多个对象实例，而不是实际的类。一个对象图是类图的一个实例。包由子包或类组成，用于描述系统的分层结构。
- 用例图（Use Case View）从用户角度描述系统功能，并指出各功能的操作者。
- 实现图（Implementation View）描述代码部件的物理结构及各部件之间的依赖关系。一个

部件可能是一个资源代码部件、一个二进制部件或一个可执行部件。它包含逻辑类或实现类的有关信息。部件图有助于分析和理解部件之间的相互影响程度。

2) 动态行为视图：描述系统在时间上的行为。包括状态机图、活动图和交互图。

- 状态图描述类的对象所有可能的状态以及事件发生时状态的转移条件。通常，状态机图是对类图的补充。
- 活动图描述满足用例要求所要进行的活动以及活动间的约束关系，有利于识别并行活动。
- 交互图描述对象间的交互关系，包括顺序图和合作图两种图。其中顺序图显示对象之间的动态合作关系，它强调对象之间消息发送的顺序，同时显示对象之间的交互。合作图描述对象间的协作关系，合作图跟顺序图相似，显示对象间的动态合作关系。除显示信息交换外，合作图还显示对象以及它们之间的关系。如果强调时间和顺序，则使用顺序图;如果强调上下级关系，则选择合作图。

3) 模型管理：描述了用层次式的单元对模型自身的组织。包是模型的通用组织单元。

科学数据共享数据模式主要使用结构性分类视图中的静态视图，即主要是类图、对象图和包图三种视图。

由于科学数据共享数据模式要求有统一的的描述规则和规范化的构建方法，来清晰、科学的描述本领域的共享数据集的概念、内容、组成及其结构信息。在客观上需要统一建模语言UML中的静态视图，特别是类图、对象图和包图三种视图，来描述科学数据共享数据集的概念、内容、组成及其结构的描述信息。

有关UML的详细定义及内容可以参考OMG的UML规范内容。

4.4 共享科学数据模式

为了对科学数据进行共享和交换，各科学领域需要建立共享科学数据模式。首先需要明确三个关键内容。

1) 领域共享的内容是科学数据共享数据集。

2) 共享科学数据模式是对本领域内的科学数据共享数据集的内容、组成及其结构的规范化描述。

3) 本标准将主要基于逻辑层次，来规范共享科学数据模式的建立和描述。

下面对于上述的三个方面进行具体说明：

依据科学数据共享工程对主体数据库的建设要求,将各科学数据中心和科学数据网已经具备大量的科学数据资源，以主体数据库形式进行建设，形成科学数据共享数据集形式进行共享。这样需要从已有的业务数据系统中，查找、抽取、筛选、制作、最终形成符合科学数据共享工程要求的切实可用的科学数据共享数据集。实际上，各个领域内的科学数据是多种多样的。可能存在于数据库、文件、二进制数据、文本数据、图像数据、图形数据、声音、视频、多媒体数据等多种形式。

在制作共享的科学数据集时，如果拥有针对数据内容组成和结构的描述说明，就可以协助科学数据共享数据集的制作者，更为有效、合理、科学的完成从复杂多样的异构业务数据系统到通用的科学数据共享系统的重整、转换、新建、发布过程。

同样重要的是使用科学数据共享数据集时，对于科学数据共享数据集的用户，来自完整的共享科学数据模式的数据集，是更加有实用价值的。这样的科学数据共享数据集易于准确理解，便于有效使用。尤其是在进行深入的再开发、再处理分析时，就会迫切需要对于数据集本身内容、组成和结构的描述说明信息。

总而言之，需要对科学数据共享数据集的内容进行规范化和标准化描述，才可能真正实现科学数据共享。通过数据模式，各个领域能够准确描述和理解科学数据共享数据集的内容，生产、加工出符合科学数据共享工程需求的数据集，进而保证科学数据共享活动的实现。否则，用来共享的科学数据集就只是一堆数据，必然会限制它的共享和使用。所以对于数据集制作人员及数据集用户来说，不能够进行数据的处理和分析，就意味着没有参考和使用价值。

考虑领域普遍的现状，针对科学数据共享数据集的内容、组成和结构的描述说明，包括下面的三种情况：

1) 领域内可能已经存在针对于科学数据的交换方面的相关标准、规范（诸如国际标准、国家标准、业界标准、项目规范、单位规范等）含有明确的交换格式。

实际上，交换格式隐含了数据的组成和结构的基本方式，但是不适用于直接、完全的应用到本领域的科学数据共享工程建设中。

需要针对项目的实际需求，进一步将隐含的数据组成和结构的基本方式确定下来，以此作为基础，吸收到科学数据共享数据集的内容、组成和结构的描述说明体系。

2) 领域内也可能已经存在系统设计、开发类型的文档。

包含有实际的字典（诸如表格、文档、文本等），可以描述领域内准备共享的科学数据的组成结构信息，但是可能其描述方式不规范，有着各种各样的结构方式。例如**XML SCHEMA**、**UML**图、**E-R**模型图、摘要文本或者其它的格式。

这种只是基于物理实现层次的描述，对于原有的系统内部使用时可能没有问题，但是对于科学数据共享工程来说，在进行交换和共享的时候，科学数据共享数据集的制作者和使用者都需要准确的了解数据的本质内容、组成和结构才能够进行深入的处理、交流和应用。

需要先将隐含的数据组成和结构的基本方式确定下来，再对其的描述格式进行重新整理、统一表达，然后筛选、提炼、整合到科学数据共享数据集的内容、组成和结构的描述说明体系里。

3) 领域内还可能包括有历史遗留系统。

这些系统是没有字典进行描述的，对于其中的大量数据的查找、使用、驱动是依赖于已经存在的软件系统的运行。那么，系统外界的人员是无法直接应用的，特别是对于科学数据共享数据集的制作者和使用者来说，更加无法直接参考和使用。可见这部分数据只是停留在物理层次，更需要规范化，才能使得它们脱离开原有软件系统，独立拥有实际的参考和使用价值；否则，不但不易于理解、分析这部分数据，而且更谈不上对其进行加工、处理和应用。所以，缺乏对于数据内容、组成和结构的描述说明信息，就是缺少进行科学数据共享的前提、基础，很难进行科学数据共享数据集的制作、交换和共享。

需要首先通过相应的工具从这类的软件系统中按照规范化的步骤，提取出必要的数据内容、组织和结构信息，再遵循统一的描述规则进行表达说明、挑选后将必要的信息纳入到科学数据共享数据集的内容、组成和结构的描述说明体系。

上述三种情况的阐述，说明在科学数据共享工程中，共享数据模式是本领域内所有共享数据的概念、内容、组成及其结构的描述信息的有机组合。它包含有统一的的描述规则和规范化的构建方法，来清晰、科学的描述本领域的共享数据集的概念、内容、组成及其结构信息。

需要考虑清楚在什么层次上可以对科学数据共享数据集的内容、组成及其结构信息，进行合理的、规范的、本质上的说明和描述。

分析和构建共享数据模式的工作重点应该放在逻辑层次上的。在逻辑层次上进行描述说明，不但可以基于概念数据模式，具有本质化、清晰化和概念化的特点，而且结合了信息技术实现的因素，进一步的分析并增加各种对象、事件和其关系。同时又能够高于具体的物理实现，拥有更大的灵活性、抽象性和概括性，是作为物理数据模式建立的基础。明确针对逻辑层次上，进行共享数据模型的标准化的描述和表达，能够有力的支持科学数据的共享和使用。本标准将基于逻辑层次，来规范共享科学数据模式的建立和描述。

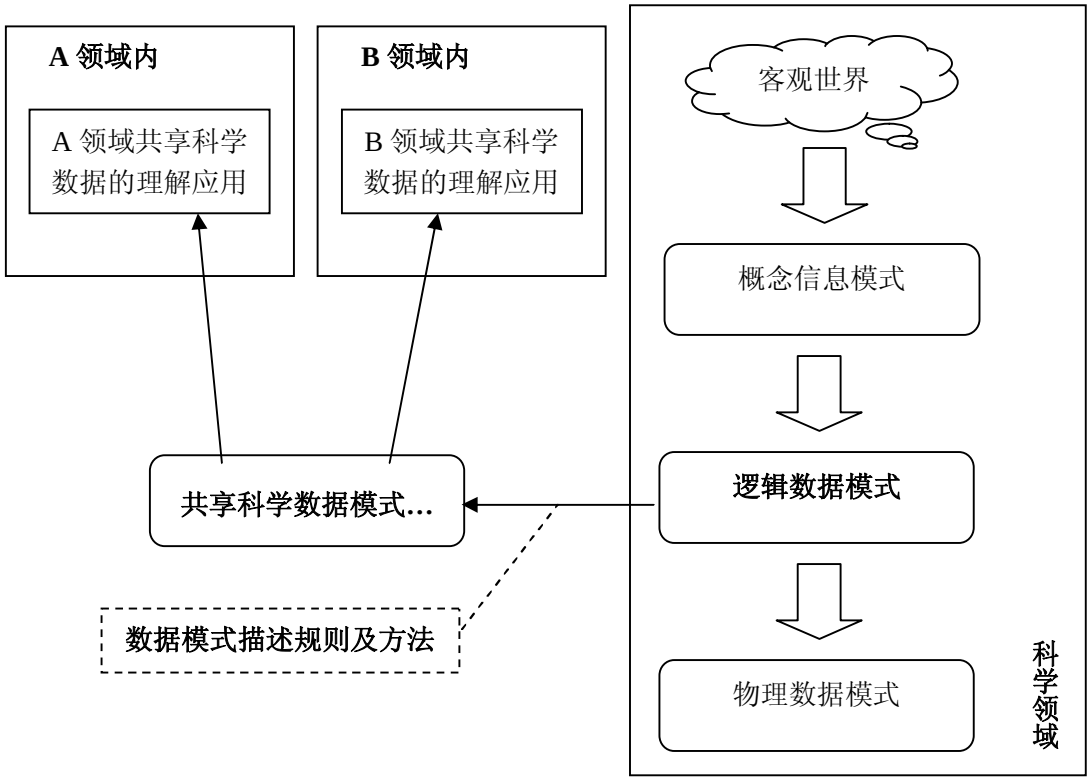


图4　共享科学数据模式

5　数据模式描述规则

5.1　概述

E-R模型图，就是用图解的方法描述实体与联系以及它们的一些性质。现实世界中任何可以明确识别的事件和事物都称之为实体。实体与实体之间可以存在某种联系。例如“张老师”、“李学生”等都是实体，老师和学生之间存在“教学”的关系，老师教学生，学生向老师学习。

对科学数据的建模使用**E-R**模型，并使用**UML**图表达**E-R**模型，并用摘要形式说明每个实体。**UML**图便于表示模型中的实体和相互间关系，而摘要形式则侧重于描述模型中的单个的实体、属性。

本章将介绍**E-R**模型中的组成要素、使用规则以及如何使用**UML**和摘要形式来表达。虽然每个组成要素高度的相互关联，但这里是分别对每个组成要素进行描述的，并没有考虑实际建模时的顺序问题。对科学数据建模时，模型包含多个数据单元（通常会表达为**UML**视图），数据单元中包含多个实体及其属性（以及属性的值域）。

对科学数据建模时，使用到的**E-R**组成要素包括：实体，属性，值域，主键和外键，实体间的泛化关系、包含关系、一般关系、依赖关系，注释，数据单元。

5.2　UML 图形式描述规则

5.2.1　实体

5.2.1.1　定义

一个实体代表一组真实的或抽象的事物（人、物体、地点、事件、想法、多个事物的结合体等），它们拥有共同的属性和特征。

该组中的一个叫做该实体的一个实例。

一个实体对应于现实世界中的一个物体，或者是一个抽象的概念。

例如：

人：学生、老师、公司员工、购买者……

地方：城市、部门……
对象：机器、建筑物、汽车……
事件：财产清查、订单……
概念：劳动、课程、帐户……

真实世界中一个事物可以由一个数据单元中的多个实体来表示。例如：小张既可以是“公司员工”，又可以是“顾客”。而且，一个实体实例可以代表真实世界对象的混合体，如：张氏夫妇可以是实体“夫妻”的一个实例。

一个实体可以是独立确定的，也可以是存在依赖关系的。对立的实体是指可以独立确定而不需要决定与其它实体关系的实体；而依赖性的实体需要确定与其它实体间的关系。例如：“人”这个实体是独立确定的，对它的理解不需要依赖其它实体；对“班主任”这个实体的理解还需要“班级”实体，因为只有知道了该班主任所负责的班级，才能最终确定这个班主任。

每一个实体必须有唯一的名字，意味着只要是相同名字就拥有相同的含义。反过来，同样含义就需要相同的名字，使用别名时例外。比如一个命名为“班主任”的实体，在整个模型中的任何地方使用都有同样的含义，指负责一至两个班级日常工作的教师，用到“班级日常工作的教师”这个概念时，必须使用“班主任”这个实体名称。

一个实体拥有任意多的属性。关于属性，请参见后面的章节。

一个实体可以拥有任意数量的关系。一个实体与其它实体间可以存在多种关系（任意多个）；也可以是一个完全独立的实体，不与其它实体发生关系。

模型中的实体相当于数据元标准中的对象类。

5.2.1.2 图形符号

实体使用**UML**中的类（**Class**）来表示，同时设定该类的原型为“实体”。类的图形符号是一个矩形框，其中标注出实体的名称。

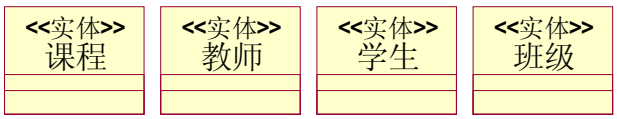


图5

示例中展示了四个实体，它们分别与现实生活中的课程、老师、学生、班集体对应，是这四种事物的在模型中的代表。

实体命名可以使用中文或英文，如果是英文，则首字母必须大写。

5.2.2 属性

5.2.2.1 定义

一个实体所关联的值域称为该实体的属性。

一个属性代表一个特征，或一个与真实或抽象事物关联的特性。通常一个实体包含多个属性，一个属性实例包括特征类型及其值。比如“学生”这个实体包含有“学生编号”、“生源”、“所在班级”等多个属性。

实体实例的相互不同是通过各个属性的取值而不同的。

一个属性实例就是具体的特征。如：小张是“学生”这个实体的一个实例，他的属性“姓名”的值为张××。通常，一个实体的实例的每一个属性都有一个明确的值。有可能这样，实例的属性没有值，比如：一个“办主任”实体的实例有“所带班级”属性，在所带班级刚毕业且还未接新班级时，该值为空。

属性是实体到值域的映射。在一个实体中，属性必须有唯一的名字，同样的属性名字代表同样的意义。反过来，同样含义就需要相同的属性名字，使用别名时例外。一个实体可以拥有任意数量的属性。一个属性由属性名称或其别名来标注。

数据模型中实体的属性相当于数据元中的特性和表示。同实体关系类的数据模型相比，模型中的实体相当于数据元中的对象类，而实体的属性相当于数据元中的特性和表示。同实体关系类的数据模型相比，模型中的实体相当于数据元中的对象类，而实体的属性相当于数据元中的特性和表示。

5.2.2.2 图形符号

属性使用UML中的属性（Attribute）来表示。

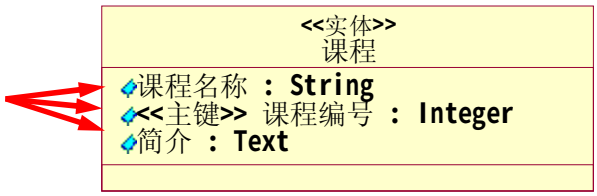


图6

示例中，实体“课程”包含有三个属性：课程名称、课程编号、课程简介。因为是对数据进行建模，因此UML中属性的可见性没有意义，因此实体的每个属性都认为是公开（Public）的。

5.2.3 值域

5.2.3.1 定义

一个值域代表一组命名并定义好的值，属性（Attribute）从中取值。值域的单独定义，为的是可以重用，即多个属性可以使用同一个值域。

值域被视为一个拥有确定的数量无限实例的类。例如：“状态码”可以被认为是一个值域，任何允许的值都必须满足它的定义，“状态码”包含三个枚举型的值，那么该值域的取值只有三种可能。另一个例子，“姓名”可以有无穷的实例，但必须由汉字和字母a-z，A-Z组成。值域被认为是固定不变的的类，它的值不随时间改变。相反，实体是随时间变化的，实例的值可以被修改和维护。值域的实例用一个唯一的值。一般的，通常包含如下几种值域：float、double、integer、string、Date、Time、boolen等。根据需要建立数据模式时，可以自己定义相应的值域。

定义一个值域通常有两种方式：定义值列表和给出取值范围规则。定义值列表方式会给出所有允许取的值，一个属性取的值只有在列表中出现才是有效的。如：“性别”。给出取值范围规则方式通常会给出取值的下边界和/或上边界。如：方位角取值必须在-360°和+360°之间。

5.2.3.2 图形符号

值域没有对应的UML图形符号。而通过设定属性的类型来实现。

5.2.4 主键和外键

5.2.4.1 定义

主键是一种特殊的属性，对该实体属性取值作出了唯一性的限制。即，所有实体实例的该属性取值不会出现重复。通过该键值可以唯一的确定一个实体。

实体实例的相互不同是通过各个属性的取值而不同的。为了通过某个属性而唯一的标识出各个实体，这是采用主键。

例如：“居民”的身份证号定义为主键，这样，每一个公民的身份证号互不相同，不会出现重复现象。

还有一个概念叫“候选键（Candidate Key）”，就是通过一个或多个属性唯一的标识出实体。有时候，一个实体的任何一个属性都不能唯一的确定实体，而通过几个属性的联合则可以，这时候采用候选键。例如：通过“数据库标识”和“记录标识”两个属性可以唯一的确定一条数据记录。

外键也是一种属性，它的实例由相关的实体的实例指定。例如：在学校的图书管理系统中存在“借阅管理”实体和“借阅人”实体，“借阅管理”实体中包含属性“借阅人”、“所借书编号”等属性，“借阅人”的取值应从“借阅人”的实例中取得。“借阅人”中有五个实例：小张、小王、小李、小赵、小武，那么“借阅管理”的“借阅人”属性取值只能是这五个。随着借阅人的实例的增加，“借阅管理”的“借阅人”属性所允许的取值也不断增加。

5.2.4.2 图形符号

在UML中通过设定属性的原型（StereoType）为“主键”。

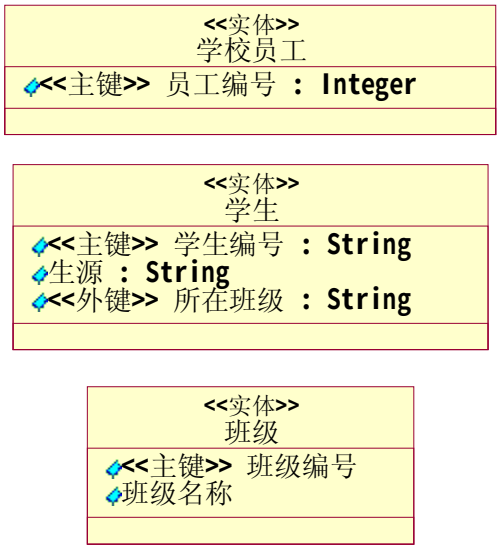


图7

5.2.5 实体间关系

5.2.5.1 继承关系

5.2.5.1.1 定义

继承关系，表示若干数据实体继承自某个数据实体，并因此具有该实体的部分属性；

5.2.5.1.2 图形符号

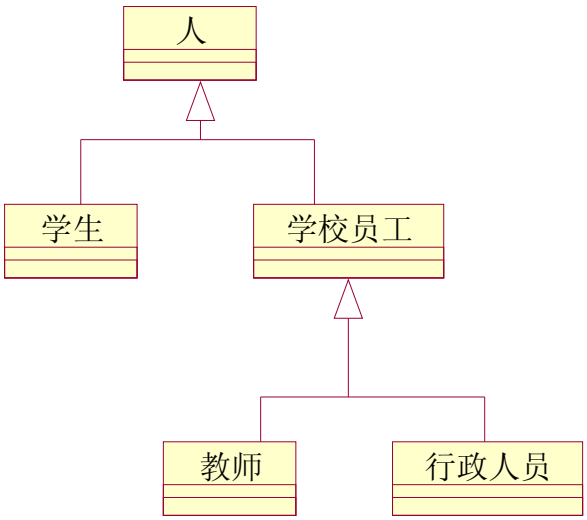


图8

5.2.5.2 包含关系

5.2.5.2.1 定义

包含关系，表示实体（数据单元）包含其它的实体（数据单元），被包含的实体（数据单元）是前者的一个组成部分。例如：一个人包括头、躯干、四肢，一辆汽车包含发动机、车身、轮胎等。

指定包含关系同时要指定包含的数量。如一部汽车饱含一个车架、四个轮胎、一台发动机。一个班级包含一至多个学生。

包含数量通常包括：

0..1　　表示没有，也可以有并最多有一个。

0..n　　表示有，也可以没有，最多有无穷多个。

1　　表示必须有，且只能有一个。

1..n　　表示必须有，最多可以有无穷多个。

n　　表示必须有指定数量的。

包含关系有两种：一种是聚集，另一种是强势聚集。

聚集表示一个实体由一组实体所组成。强势聚集表示每一组成部分是不可分割的，组成部分与整体是“终身关系”，同时建立和清除。

5.2.5.2.2　图形符号

菱形符号出现在整体一侧。

包含数量要在关联线的靠近组成部分的位置上标识。

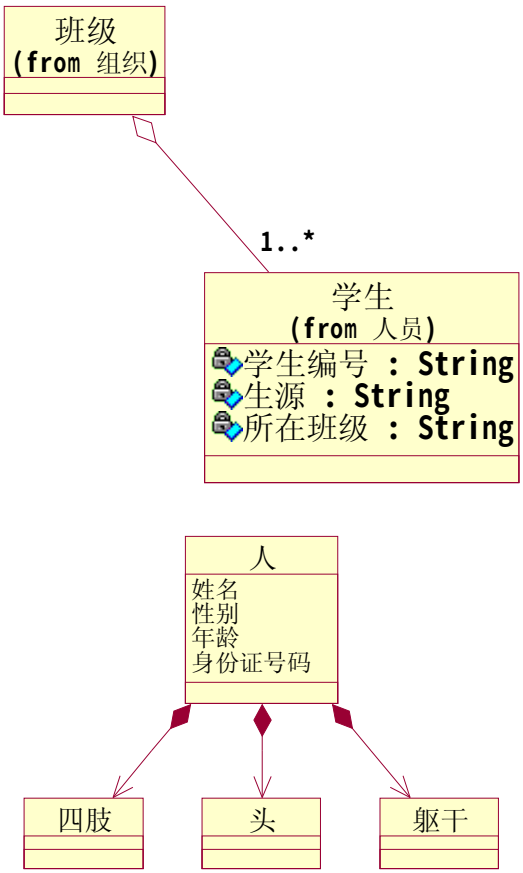


图9

5.2.5.3　依赖关系

5.2.5.3.1　定义

依赖关系，表示对实体（数据单元）的理解、使用等依赖其它的实体（数据单元）。

5.2.5.3.2　图形符号

依赖关系使用带箭头的虚线表示。箭头的方向表示依赖方向。

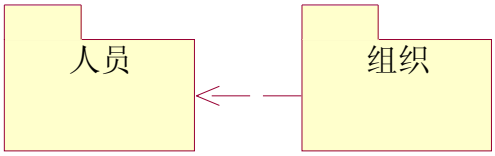


图10

5.2.5.4 关联关系

5.2.5.4.1 定义

用于表示除泛化、包含、依赖关系以外的数据实体之间的关系。

5.2.5.4.2 图形符号

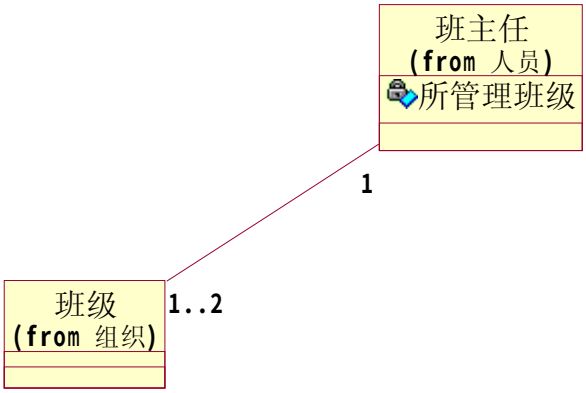


图11

5.2.6 注释

5.2.6.1 定义

注释是自然的文档，对模型的有机组成部分作出限制。

注释通常出现在数据单元中。

5.2.6.2 图形符号

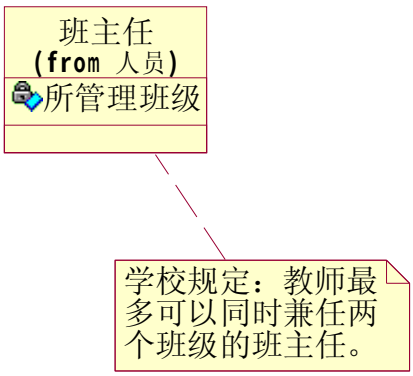


图12

5.2.7 实体的组织

5.2.7.1 定义

出于实体的组织是出于某种目的，由若干个实体和指定的域（属性）组装而成的集合。

一个数据模型通常包含一至多个数据单元。数据单元的之间表现同时也暗示了数据单元之间的关系。如：数据单元“课程”是包含在数据单元“学校”中的，暗示了“课程”是“学校”的逻辑上一部分。

数据单元之间的关系包括：依赖关系和包含关系。依赖关系通过**5.2.5.4**种定义的**UML**图形符号表示。包含关系通过**UML**模型本身的层次关系确定。

5.2.7.2 图形符号

使用**UML**中的包（**Package**）表示。

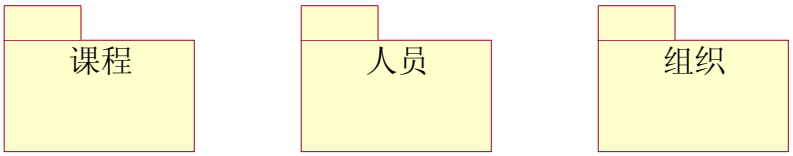


图13

5.3 字典形式描述规则

字典形式从名称、定义、英文名称、英文短名、版本标识、状态、来源、注释等多个方面来描述模型中的实体、属性，从而能够严格的对数据模型中的实体和属性作出描述。

5.3.1 实体

字典形式通过中文名称、中文别名、英文名称、英文短名、定义、注释、版本标识、状态、来源、安全说明几个方面对实体作出详细的描述。

表1 实体的描述项

属性	约束	属性定义
中文名称	必选	实体的标识，一般使用名词表达，通常名称都能反映出实体的属性和特征。
中文别名	可选	实体的别名，一般使用名词表达。
英文名称	必选	实体的英文全称
英文短名	必选	实体的英文名称缩写
定义	必选	实体定义的详细描述。
注释	可选	和实体相关的其它信息。
版本标识	必选	用于实体的配置管理和控制。
状态	必选	0：讨论版本 1：正式版本
来源	可选	说明实体定义的来源，来源包括已有的数据模式标准、已有的信息系统以及其它来源。
安全说明	必选	说明该属性的安全限制信息，包括访问和使用限制等。

5.3.2 属性

字典形式通过中文名称、中文别名、英文名称、英文短名、定义、数据类型、SQL数据类型、注释、版本标识、状态、来源、值域、安全说明几个方面对属性作出详细的描述。

表2 属性的描述项

属性	约束	属性定义
中文名称	必选	属性的标识，一般使用名词表达，通常名称都能反映出属性的属性和特征。
中文别名	可选	属性的别名，一般使用名词表达。
英文名称	必选	属性的英文全称
英文短名	必选	属性的英文名称缩写
定义	必选	属性定义的详细描述。
数据类型	必选	属性的数据类型，例如字符型、数值型、逻辑型、日期型等。
SQL数据类型	可选	该属性在关系型数据库中的数据类型，按照结构化查询语言的数据类型表达方式进行描述，例如varchar(100)，代表可变长字符串，最大长度单位100个字符。
键	必选	是否为主键、外键。如果是，则详细说明。
注释	可选	和属性相关的其它信息。
版本标识	必选	用于属性的配置管理和控制。
状态	必选	0：讨论版本

		1：正式版本
来源	可选	说明属性定义的来源，来源包括已有的数据模式标准、已有的信息系统、数据元以及其它来源。
值域	必选	属性的取值范围
安全说明	必选	说明该属性的安全限制信息，包括访问和使用限制等。

5.3.3 值域

字典形式通过中文名称、中文别名、英文名称、英文短名、定义、值、注释、版本标识、状态、来源、几个方面对属性作出详细的描述。

表3 表 XX 值域的描述项

属性	约束	属性定义
中文名称	必选	属性的标识,一般使用名词表达,通常名称都能反映出属性的属性和特征。
中文别名	可选	属性的别名，一般使用名词表达。
英文名称	必选	属性的英文全称
英文短名	必选	属性的英文名称缩写
定义	必选	属性定义的详细描述。
值	必选	值域中值的限制条件、列举等。
注释	可选	和属性相关的其它信息。
版本标识	必选	用于属性的配置管理和控制。
状态	必选	0：讨论版本 1：正式版本
来源	可选	说明属性定义的来源，来源包括已有的国内国际标准、已有的信息系统、数据元以及其它来源。

6 数据模式建立与描述方法

6.1 建模方法概述

根据科学数据共享工程对主体数据库的建设要求,各科学数据中心和科学数据网已经具备大量的科学数据资源，这些科学数据资源将以主体数据库形式进行建设，以科学数据共享数据集形式进行共享。对科学数据共享数据集内容进行规范化和标准化描述是真正实现科学数据共享的基本前提,通过数据模式，才能够准确描述和理解科学数据共享数据集的内容，才能够保证科学数据共享活动的实现。

目前各领域都已经积累了大量的科学数据资源。合理、规范地建立领域的共享数据模式标准，就可以有效地进行科学数据共享数据集资源的共享和交换，从而充分利用现有的这些科学数据资源。

本章描述了各个领域在已有的工作基础上建立科学数据共享逻辑数据模式的过程，包括需求收集、数据模式建立与表达、标准协调和标准实现四个阶段。同时，本章还描述了各阶段的主要工作内容及具体要求。为了更加方便的完成逻辑数据模式标准的制定，本标准还给出了一套完整的工作文档模板，这些模板可用于协助完成各阶段的工作。

数据模式标准化可以分为四个阶段进行，如下图所示：

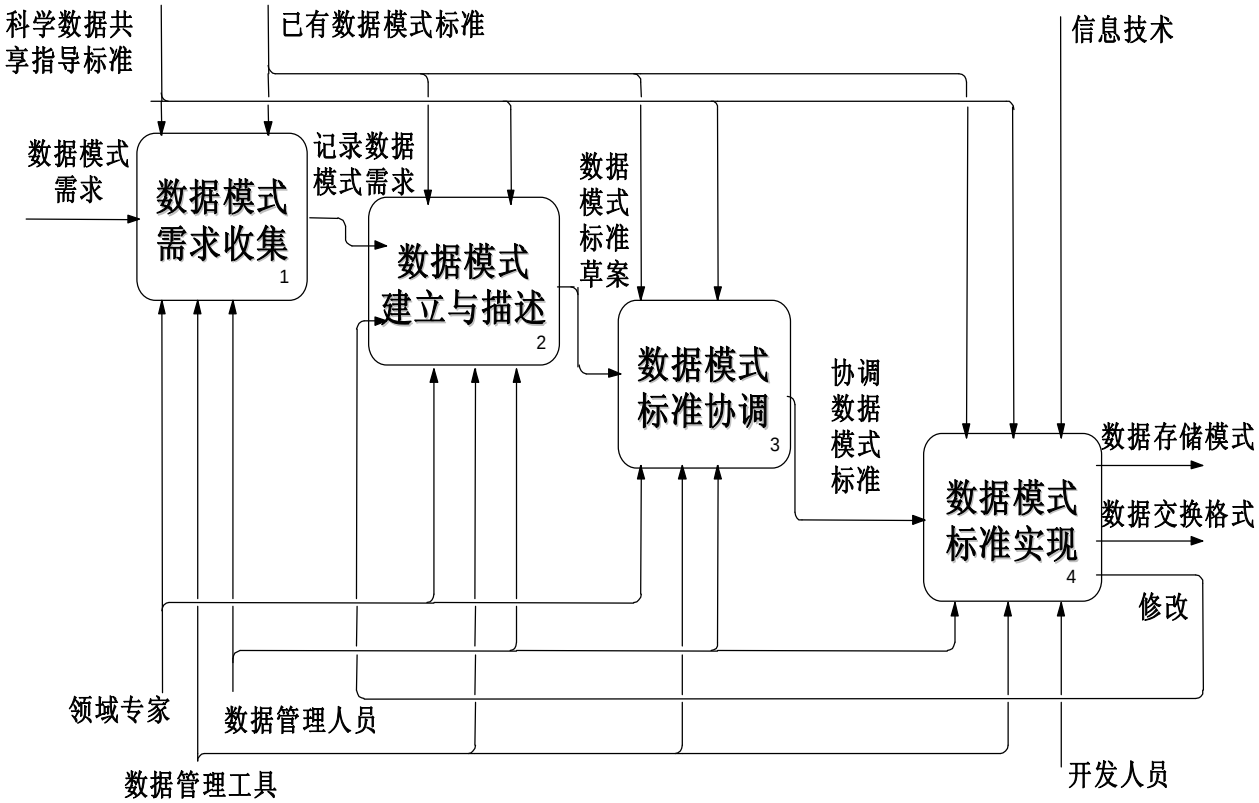


图14　数据模式建立方法流程示意图

1) 需求收集阶段

确立目标与范围的基础上，收集相关资料与需求，然后依据实际情况的限制，进行数据集组成结构信息的分析提取（必要时通过相应的工具软件来进行分析、捕捉）。同时考察本领域内现有的数据标准，最终形成确定为数据模式标准需求收集文档。

2) 数据模式建立与描述阶段

根据前面阶段所产生的文档，建立本领域共享数据集的数据模式。建模时，首先是对于收集的需求进一步提炼出实体、属性和关系的信息，再定义实体、属性、值域、主外键和关系；然后进行规范化的表达描述，生成准备征求意见的数据模式标准草案。

3) 标准协调阶段

结合需求，从信息内容、功能需求、可操作性等方面对已建立的数据模式标准进行协调。协调的范围包括项目组层次和领域层次。输出的阶段性成果是数据模式标准正式文档、数据模式标准征求意见文档、数据模式标准意见处理文档。

4) 标准实现阶段

在已建立的数据模式标准基础上，结合特定的信息技术，建立数据的存储模式和交换格式。

6.2　需求收集

需求收集是在确立数据模式标准服务目标与范围的基础上，广泛的收集相关资料，并依据实际情况的限制，进行数据集组成结构信息的分析提取，必要时通过相应的工具软件来进行分析、捕捉。同时考查本领域现有的数据标准，最后确定出数据模式标准需求文档。分析需求的工作可以分为三个步骤进行：

1) 确立目标与范围

明确提出总体目标、共享范围和共享数据集。

2) 收集数据模式需求

从已有的数据（含有交换格式）内容标准，信息资源规划、在建信息系统和已运行的数据信息系统中收集整理数据模式的需求。

3) 收集现有数据模式标准

收集、考察领域现有的数据模式相关标准，分析其内容组成和结构。

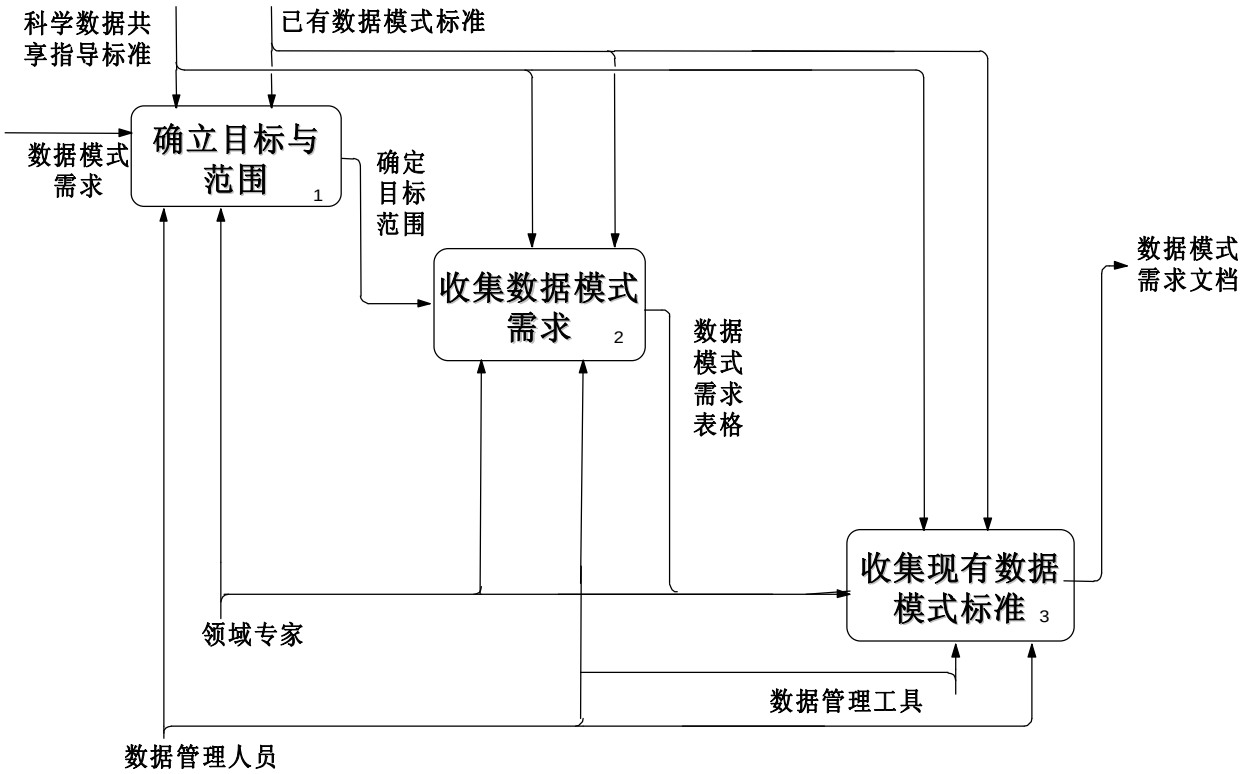


图15　数据模式标准需求收集的流程示意图

6.2.1　确立目标与范围

建立数据模式标准是为了规范化科学数据生产、加工、存储、交换的数据内容，满足科学数据共享的需求。在此基础上，尽量考虑各领域相关的数据模式需求，制定出科学、合理的数据模式标准。

1) 任务是明确提出总体目标是为了科学研究进行的数据共享和交换。

2) 任务是划定初步的共享范围和共享数据集。

3) 任务是明确出共享范围和共享数据集。

下面是上述三个执行任务的具体描述：

明确提出总体目标是为了科学研究进行的数据共享和交换：

- **n** 描述共享数据的逻辑数据模型，是为了规范化共享的数据内容。
- **n** 为确切表达和理解共享内容，不会发生理解上的错误、歧异。
- **n** 准备建立科学数据共享数据集。

下面是上述三个需要达成共识的详细内容说明：

① 描述共享数据的逻辑数据模型，是为了规范化共享的数据内容。

在逻辑层次上可以对科学数据共享数据集的内容、组成及其结构信息，进行合理的、规范的、本质上的说明和描述。不但可以基于概念数据模式，具有本质化、清晰化和概念化的特点，又结合了信息技术实现的因素，进一步的分析并增加各种对象、事件和其关系。同时又能够高于具体的物理实现，拥有更大的灵活性、抽象性和概括性，是作为物理数据模式建立的基础。数据模式正是属于逻辑层次上的，是数据的概念、内容、组成及其结构的描述信息的有机组合。它包含有统一的的描述规则和规范化的构建方法，来清晰、科学

的描述本领域的共享数据集的概念、内容、组成及其结构信息。保障了形成的科学数据共享数据集的数据内容和格式的规范化和标准化。

② 为确切表达和理解共享内容，不会发生理解上的歧义。

基于逻辑数据模型所构建的科学数据共享数据集，客观上具备了确切表达和理解共享内容的基础和前提条件。统一而通用的共享科学数据模式提供了科学数据共享数据集的内容组成和其结构的描述信息，保障了理解的准确性和一致性，避免了歧义。

③ 准备建立科学数据共享数据集。

按照定义，科学数据共享数据集是可以标识的数据集合。数据集在物理上可以是更大数据集的比较小的数据组。从理论上讲，数据集可以小到更大数据集内的单个要素或要素属性（某个特征或特征属性）。一张硬拷贝地图或图表均可以被认为是一个数据集。本标准所指的数据集是指不可再细分的数据集，即可以用一个字典能够唯一描述的数据集合。数据集实际存在的多种多样的物理形式，主要考虑的两种情况如下：

n 文件形式

需要共享的数据集是以文件的形式存放，准备获取到其的内容组织和结构说明性信息，记录归档到“数据文件说明表格”，参见格式为附录**A.1** 数据模式标准的需求收集文档中的数据文件说明。特别是有些用于交换的数据集，是以公认的交换格式作基础，交换格式可以是已有的标准的部分内容，也可以不是存在于现有发布标准。需要在下面的操作中，专门从交换格式里抽取有效的内容组织和结构说明性信息，再进行归档到“相关数据标准采标情况说明表格”。参见附录**A.1** 数据模式标准的需求收集文档中的相关数据标准采标情况说明。有关交换格式的设计是对数据模式的物理实现，可以参考第**4**阶段“标准实现”内容。

n 数据库形式

需要共享的数据集是在数据库表内存放，通常符合**ER**模型，反向获取到其它的内容组织和结构说明性信息，记录归档到“数据库说明表格”和“数据表说明表格”，参见格式为附录**A.1** 数据模式标准的需求收集文档中的数据库说明。

划定初步的共享范围和共享数据集：

根据科学数据共享工程的领域共享需求，研究现有的数据集，重点是规范出共享的范围。先遵照本领域的共享需求，考察研究准备用于共享和交换的业务数据。以业务数据为基础和前提，在业务数据上经过加工、制作和规范化后的数据，就是科学数据共享数据集。当然，属于研究内容本身的管理信息属于共享数据的内容，而不包括有关业务管理方面的信息。

然后划定范围。依据科学数据共享工程对主体数据库的建设要求，对于前面步骤的研究成果进行确认，划定初步的范围。注意不是业务数据交换共享，必须划清楚界限。同时也讲清楚和业务数据共享的关系：业务数据是共享数据的基础和前提，共享数据是在业务数据的基础上经过加工和综合的规范化数据，它不包括有关业务管理方面的信息。

明确出共享范围和共享数据集：

根据各试点本身的试点建设申请书中的建设内容，确定共享的信息内容的范围和初步的共享数据集，建议进行归档记录到“共享的信息内容的范围说明表格”和“共享数据集说明表格” （部分字段的填写）。开展讨论、征求意见，判断前面列举的共享数据集是否采用，无关的内容可以从收集成果范围中删除。填写完整“共享数据集说明表格”，表明已经作过工作的内容（注意说明不采用的理由）。这个表格在本标准中给出，参见附录**A.1**数据模式标准的需求收集文档中的共享信息内容的范围说明和数据集说明。

本节应该输出的具体成果是：整理后的“共享的信息内容的范围说明表格”和“共享数据集说明表格”，参见附录**A.1**数据模式标准的需求收集文档共享信息内容的范围说明和数据集说明部分。

6.2.2 收集数据模式需求

收集数据模式需求是建立数据模式标准的基础。应按照服务于科学数据共享的目标，结合各领域具体的信息化程度和信息资源建设情况，从已有的数据（含有交换格式）内容标准，信息资源规划、在建信息系统和已运行的数据信息系统中收集整理数据模式的需求。

根据前面阶段的工作成果是“共享的信息内容的范围说明表格”和“共享数据集说明表格”，准备展开对于数据模式需求的收集步骤（逐步填写附录A.1数据模式标准的需求收集文档）。基本可以从下面图示的三个方面，展开收集数据模式需求的工作。

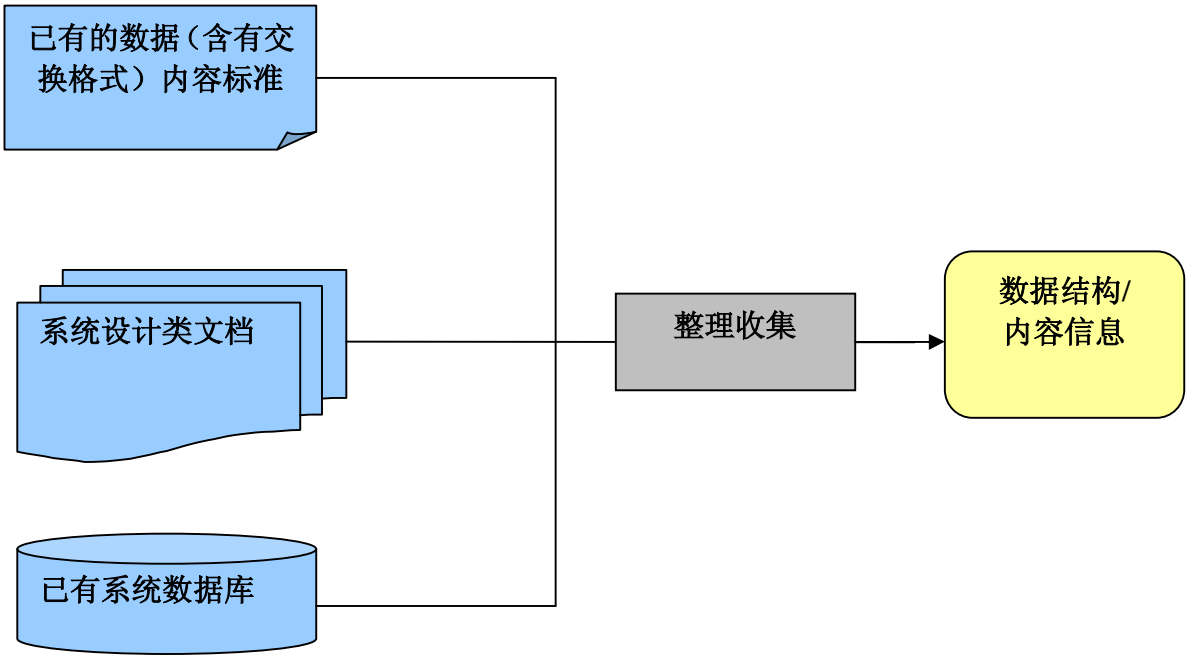


图16 收集数据模式需求的示意图

下面是上述三个工作方面的具体阐述：

1) 对于领域内已有的数据（含有交换格式）内容标准的收集。

领域内可能已经存在针对于科学数据的交换方面的相关标准、规范含有明确的交换格式。实际上交换格式隐含了数据的组成和结构的基本方式，但是不适用于直接、完全的应用到本领域的科学数据共享工程建设中。需要针对项目的实际需求，进一步将隐含的数据组成和结构的基本方式确定下来，以此作为基础，吸收到科学数据共享数据集的内容、组成和结构的描述说明体系。

已有的内容相关标准的收集范围如下所示：（凡是涉及到对数据内容进行描述和规范的标准都应当在资料的收集范围之内）

- 国际标准，包括ISO、相关国际标准化组织、
- 国家主管部门、行业的标准，
- 单位自己制定的标准。

已有的内容相关标准的收集选择原则如下所示：

- 标准本身就是有关数据内容的规范，包括概念信息模型、逻辑数据模型、业务数据建库规范、数据元（素）标准。
- 标准内容包含有关信息交换、通讯标准中对交换内容的规范，在通讯的报文协议中，请求或响应中会对交换内容进行定义。
- 标准本身是对共享内容的初步分类和描述。

输出阶段成果是：填写完整的“已有内容相关类标准的采标情况说明表格”，参见附录A.1 数据模式标准的需求收集文档中的已有内容相关类标准的采标情况说明。

2) 对于领域内已有的系统设计、开发类型的文档的收集。

领域内也可能已经存在系统设计、开发类型的文档。包含有实际的字典（诸如表格、文档、文本等），可以描述领域内准备共享的科学数据的组成结构信息，但是可能其描述方式不规范，各种各样的结构方式都有，例如**XML SCHEMA**、**UML**图、**ER**表、摘要文本或者其它的格式。这种只是基于物理实现层次的描述，对于原有的系统内部使用时可能没有问题，但是对于科学数据共享工程来说，在进行交换和共享的时候，科学数据共享数据集的制作者和使用者都需要准确的了解数据的本质内容、组成和结构才能够进行深入的处理、交流和应用。需要对其的描述格式进行重新整理、统一表达，再筛选、提炼、整合到科学数据共享数据集的内容、组成和结构的描述说明体系里。

在系统设计类文档中收集，需要包括的内容如下所述：

- 数据库设计说明（填入数据库说明表格）
- 交换数据格式设计说明（填入已有内容相关类标准的采标情况说明表格）
- 系统内部和系统之间通讯协议对交换内容的说明（填入已有内容相关类标准的采标情况说明表格）

输出阶段成果是：填写完整的“数据库说明表格”和“已有内容相关类标准的采标情况说明表格”。参见附录**A.2**数据模式标准的需求收集文档中的第七章数据库说明和第九章已有内容相关类标准的采标情况说明。

3) 对于领域内已有的系统数据库的收集。

这些系统是没有字典进行描述的，对于其中的大量数据的查找、使用、驱动是依赖于已经存在的软件系统的运行。那么，系统外界的人员是无法直接应用的，特别是对于科学数据共享数据集的制作者和使用者来说，更加无法直接参考和使用。

可见这部分数据只是停留在物理层次，更需要规范化，才能使得它们脱离开原有软件系统，独立拥有实际的参考和使用价值；否则，不但不易于理解、分析这部分数据，而且更谈不上对其进行加工、处理和应用。所以，缺乏对于数据内容、组成和结构的描述说明信息，就是缺少进行科学数据共享的前提、基础，很难进行科学数据共享数据集的制作、交换和共享。

需要首先通过相应的工具从这类的软件系统中按照规范化的步骤，提取出必要的数据内容、组织和结构信息，再遵循统一的描述规则进行表达说明、挑选后将必要的信息纳入到科学数据共享数据集的内容、组成和结构的描述说明体系。注意的原则是应当先找系统设计文档，如果确实没有文档或者文档内容不完整再通过数据库。

下面的反向工程主要针对数据库。需要借助相应的数据文件的反向工具，同样可以进行数据模式需求的提取。

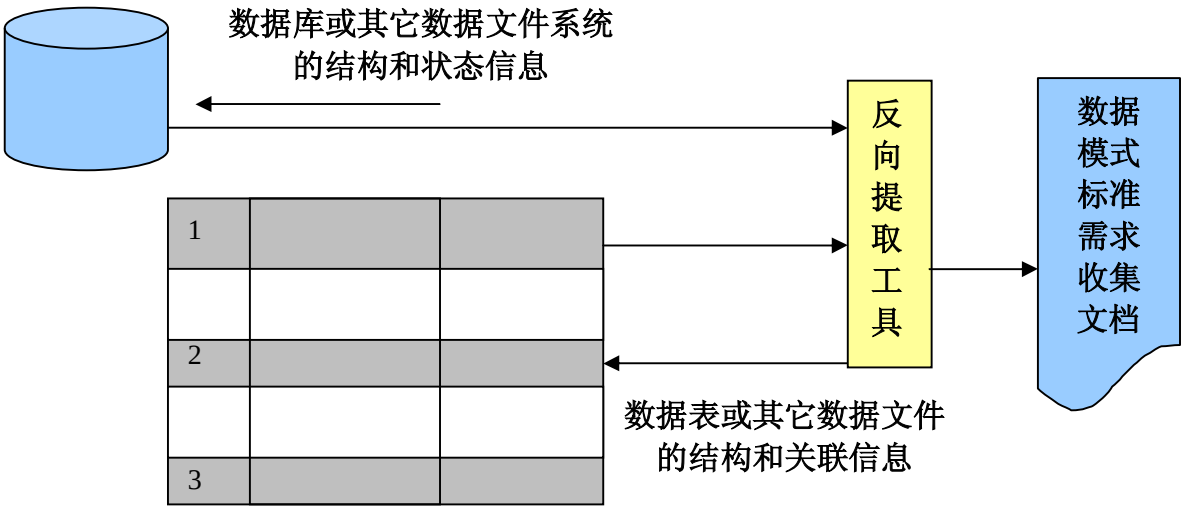


图17　反向工程的流程示意图

描述反向工程的步骤说明，逆向提取需要的数据库表结构。要生成数据库和表的内容与结构信息表格，请执行如下操作步骤。

① 根据前面确定的目标和范围，从已有数据库中，挑选需要的数据库表，并且记录归档到“数据库说明表格”。参见附录**A.2.1**数据模式标准的需求收集文档中的数据库说明部分。

② 选择适合工具，按照对应帮助文档，逐步从反方向对上面所选择的数据库表进行操作，提取相应数据结构和数据内容说明。

提取的内容包括：

- 数据表结构的提取
- 数据表关联（关系、索引、主键和外键）的提取
- 数据库结构和状态信息的提取

反向工程的使用工具说明：

- 数据库管理系统（如 ORACLE，SQL SERVER, SYBASE 等的库管理工具）。
- 现在的数据库管理系统一般都会自带上对于自身数据库表结构、关联和状态信息的反向提取功能。通常以下面的形式给出，诸如显示所选对象的 DDL（即为数据定义语言 Data Define Language），显示所选对象的相关性、显示所选对象的状态信息报告等。这些输出的结果，基本上符合 ER 模式（关系型数据模式），可以读取后填写到“数据库说明表格”和“数据文件说明表格”。参见附录 A.1 数据模式标准的需求收集文档中的数据库说明。
- 专用数据工具，例如 Rose, PowerDesigner 等。

收集分析后形成的字典的内容应当包括：

- 数据库内容的说明。
- 数据库的组成结构说明（包括多少各表，各表之间的联系）。
- 数据表内容的说明。
- 数据表构成结构的说明。

③ 依据上面的数据库结构和内容的描述，提取、提炼出数据模式的需求，记录后归档到“数据表说明表格”。参见附录**A.1**数据模式标准的需求收集文档中的数据库说明部分。

④ 反向工程的阶段性成果是，在资料收集工作完成后，最终填写数据模式需求文档。注意在数据模式需求文档中，记录每个数据需求和其引用的权威资料来源。

输出阶段成果是：填写完整的“数据库说明表格”和“数据文件说明表格”。参见附录**A.1**数据模式标准的需求收集文档的数据库说明和数据文件说明。

6.2.3　收集现有数据模式标准

收集、考察领域现有的数据模式相关标准，分析其内容组成。这些标准中大都会有相关实体、属性的定义和说明。根据已定义的共享科学数据的实体与属性定义，按照不同的复用程度分别处理。注意区分开，前面提到的数据相关标准，多数是隐含的需要提炼加工的，少有直接的相关实体、属性的定义和说明。

1) 根据领域共享的需求，收集已有的相关标准：

- 国际标准，包括**ISO**、相关国际标准化组织。
- 国家主管部门、行业的标准。
- 单位自己制定的标准。

2) 分析现有标准的复用程度，并且记录归档。输出阶段成果是：填写完整的“相关数据标准采标情况说明表格”，参照附录**A.1**数据模式标准的需求收集文档的相关数据标准采标情况说明。

- 完全采标：已有的标准可以直接满足科学数据共享的需求，且描述规范，可直接引用该标准作为科学数据共享的数据模式标准；如果描述不规范，需要按照标准化描述方法进行文

档处理。或者已有标准本身就是领域共享交换标准，且描述规范。如果描述不规范，需要按照标准化描述方法进行文档处理。

- n 部分采标：已有的标准只能部分满足科学数据共享的需求，挑选出符合科学数据共享需求的内容，在此基础上进行修订并增加内容，形成科学数据共享的数据模式标准。
- n 不符合科学数据共享需求的标准不采用，但是需要记录归档不予采纳的原因，作为数据模式标准协调阶段的参照资料。

完成全部的需求收集阶段的文档，准备带入到下面的数据模式需求分析中，根据结合考虑分析现有标准的使用程度，然后整合成为完整的共享数据模式。

6.3 数据模式建立与描述

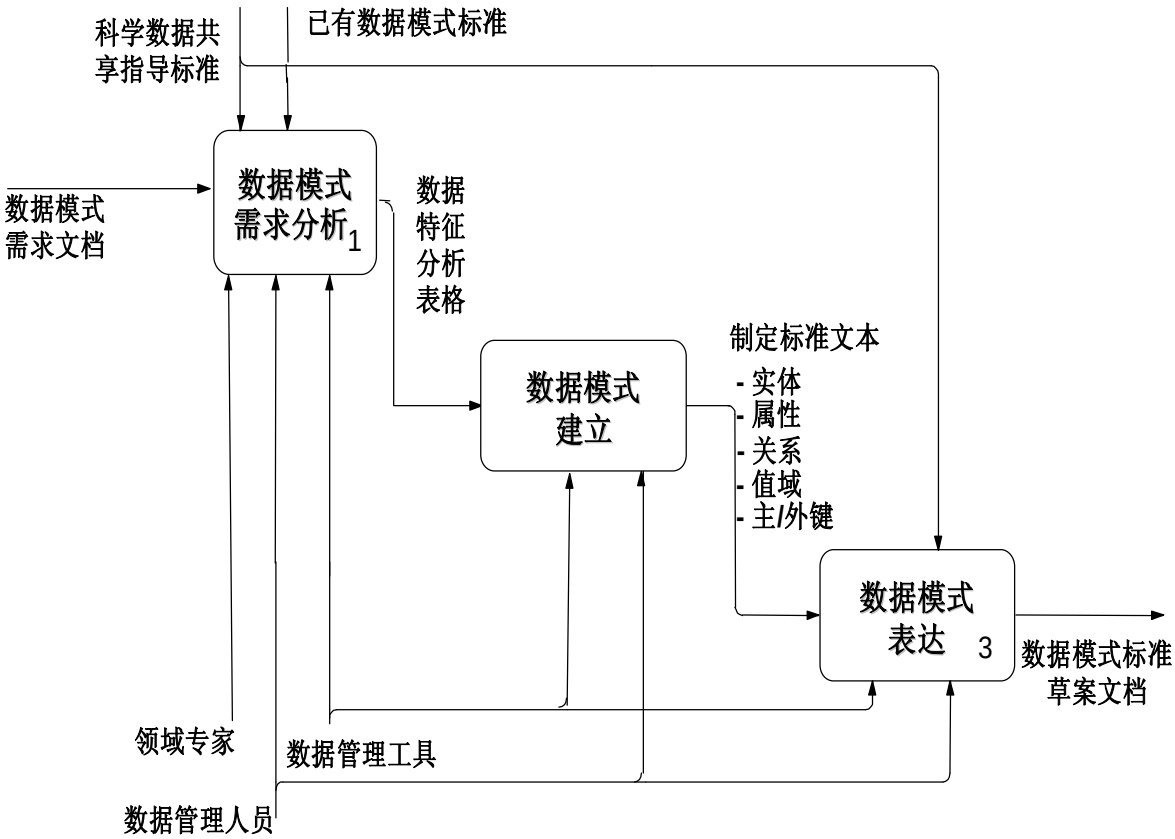


图18 数据模式标准建立与表达的流程图

数据模式建立与描述阶段包括三个主要步骤：

1) 数据模式需求分析

在前期资料收集和标准调研的基础上，通过分析、概括、抽象等工作确定数据模式标准所包括的实体、属性和关系。

2) 数据模式建立

结合领域的共享需求，准确鉴别和描述各实体、属性及其之间的关系，建立领域通用的共享数据模式。

3) 数据模式描述

按照数据模式描述规则与描述方法，规范化描述和表达实体定义及其相互之间的关系，并按照国家相关有关标准文本规范，书写数据模式标准。

6.3.1 数据模式需求分析

在前面阶段收集的需求收集文档基础上，基于实体关系模型，针对每项数据模式需求进一步提取其特征信息，捕捉数据模式需求的特征（捕捉数据模式需求的特征主要从实体层次和属性层次进行）。

前面产出的阶段性数据模式需求收集文档包含有两个层次的内容：

- 逻辑数据模型层次

  主要是指已发布、存在的数据模式内容标准、系统级设计文档中描述的数据内容组织和结构的描述、说明信息。注意，这不是软件设计级别描述的内容，软件设计级别只是基于物理实现层次的描述，逻辑层次上的数据模型对于科学数据共享工程来说，在的时候，科学数据共享数据集的制作者和使用者都可以了解到的所进行交换和共享数据的本质内容、组成和结构。

  需要针对科学数据共享项目的实际需求，进一步将相关的数据内容组成和结构的描述定义记录确定下来，统一表达，筛选后以此作为基础，构建科学数据共享数据集的内容、组成和结构的描述说明体系。最后还需要进行迭代循环的验证构建的共享数据模式。

- 物理数据模型层次

  主要是指已发布、存在的数据交换格式标准、软件设计类文档或者是数据库和其它数据文件系统，根据反向工程生成的对物理存储内容的说明信息。隐含了数据的组成和结构的基本方式，但是不适合直接、完全的应用到本领域的科学数据共享工程建设中来。可见这部分数据只是停留在物理层次，更需要规范化，才能使得它们脱离开原有软件系统，独立拥有实际的参考和使用价值；否则，不但不易于理解、分析这部分数据，而且更谈不上对其进行加工、处理和应用。所以，缺乏对于数据内容、组成和结构的描述说明信息，就是缺少进行科学数据共享的前提、基础，很难进行科学数据共享数据集的制作、交换和共享。

  需要首先通过相应的工具从这类的软件系统中按照规范化的步骤，提取出必要的物理存储信息，再提炼出数据内容、组织和结构信息，然后遵循统一的描述规则进行表达说明、进行挑选后将必要的信息纳入到科学数据共享数据集的内容、组成和结构的描述说明体系。

进行数据模式需求的分析工作包含如下四个步骤：

1) 如果存在逻辑数据模型层次上的标准和文档，直接填写到“共享数据集组成描述表格”和“实体列表”、“属性列表”。否则，进入到步骤二。
2) 对物理数据模型层次的内容，进行抽象、重组。初步确定信息层次上的实体和属性。
3) 在已完成的实体和属性两个表的基本信息填写后，按照前面的“共享范围和需求” 进行取舍挑选。
4) 循环迭代地再分析原有的数据模式标准的使用程度。将逻辑数据模型层次和物理数据模型层次上的各自的共享数据模式需求整合起来，成为有机的整体。

   下面是对上述步骤的具体操作描述：

6.3.1.1 步骤一

如果存在逻辑数据模型层次上的已发布的数据模式内容标准、系统级设计文档，直接填写到“共享数据集组成描述表格”和“实体列表”、“属性列表”。否则，进入到步骤二。

当存在逻辑数据模型层次上的有效文档，就可以依据科学数据共享项目的实际需求，分析考察前面阶段收集的需求收集文档的对应部分内容（相关数据标准采标情况说明）。针对其中的每项数据模式需求提取其特征信息，捕捉数据模式需求的特征。捕捉数据模式需求的特征主要从两个层次进行（即实体层次和属性层次），填写到“共享数据集组成描述表格”和“实体列表”、“属性列表”（注意参照附录A.2 数据模式标准的草案文档中的第五章共享数据集组成描述）。完成针对实体、属性列表的信息填写，作为共享数据模式的基础。

6.3.1.2 步骤二

对物理数据模型层次的内容，进行抽象、重组。在信息层次上描述内容，初步确定实体和属性。

根据科学数据共享项目的需求。从已发布、存在的数据交换格式标准、软件设计类文档，或者是数据库和其它数据文件系统经过提取，得到的科学数据共享数据集的物理存储内容的说明文档。就是前面

产生的数据模式需求收集和分析文档中的对应部分内容（数据库说明、数据文件说明和相关数据标准采标情况说明）。针对每项数据模式需求进一步提取其特征信息，捕捉数据模式需求的特征。根据该数据字段的内容、组成、关联和结构信息，初步判断后，分别添加到“实体描述字典表格”或者“ 属性描述字典表格” （注意参照附录**A.2** 数据模式标准的草案文档中的共享数据集组成描述）。

具体操作如下：

1) 从已有的数据模式需求收集和分析文档整理和抽象出相关实体定义，进行实体和其关系的处理：依赖、继承、关联和包含。

按照实体描述字典的内容进行规范化的实体描述信息填写“实体描述字典表格” 。（参见附录**A.2** 数据模式标准的草案文档中的第五章共享数据集组成描述的实体描述字典和表格）

2) 从已有的需求文档整理和抽象出相关属性定义，进行属性的处理。

按照属性描述字典的内容进行规范化的属性描述信息填写“属性描述字典表格”。（参见附录**A.2** 数据模式标准的草案文档中的第五章共享数据集组成描述的属性描述字典和表格）

6.3.1.3 步骤三

在已完成的实体和属性两个表的基本信息填写后，按照前面的“共享范围和需求”（参见附录**A.1** 数据模式标准的需求收集文档中的主体数据库的基本信息列表的要求） 进行取舍挑选。对需要的实体和属性说明用途、来源和定义，完善其详细信息的填写。

6.3.1.4 步骤四

循环迭代地再分析原有的数据模式标准的使用程度。将逻辑数据模型层次和物理数据模型层次上的各自的共享数据模式需求整合起来，成为有机的整体。

具体操作如下：

1) 内容的比较映射包括实体、属性的比较。

实体的比较包括实体定义、描述等。属性比较同样包括属性的定义、值域、类型等。记录归档到下面的模板表格里的相应字段项。

表4 实体比较信息列表

共享实体名称	共享实体定义	原实体名称	原实体定义	原标准来源	选择结果	说明

注：“共享实体名称” 是指从物理层次的资料中（在前面的步骤）提炼而新产生的共享数据模式里实体的名称。

注：“共享实体定义” 是指提炼而新产生的该共享数据模式里实体定义的描述信息。

注：“原实体名称” 是指原有、已经发布的数据模式标准的实体名称。

注：“原实体定义” 是指原有、已经发布的数据模式标准里实体定义的相关描述信息。

注：“原标准来源” 是指原有或已经发布的数据模式标准的来源（名称和内容）信息。

注：“选择结果”是填写最后选择的结论。例如共享实体名称，共享实体定义，原实体名称，原实体定义。

注：“说明”填写的是进行比较两种实体的名称和定义时，主要的判断理由。

表5 属性比较信息列表

共享属性名称	共享属性定义	原属性名称	原属性定义	原标准来源	选择结果	说明

注：“共享属性名称” 是指从物理层次的资料中（在前面的步骤）提炼而新产生的共享数据模式里属性的名称。

注：“共享属性定义” 是指提炼而新产生的该共享数据模式里属性定义的描述信息。

注：“原属性名称” 是指原有、已经发布的数据模式标准的属性名称。

注：“原属性定义” 是指原有、已经发布的数据模式标准里属性定义的相关描述信息。

注：“原标准来源” 是指原有、已经发布的数据模式标准的来源（名称和内容）信息。

注：“选择结果”是填写最后选择的结论。例如“共享属性名称”，“共享属性定义”，“原属性名称”，“原属性定义”。

注：“说明”填写的是进行比较两种属性的名称和定义时，主要的判断理由。

2) 根据比较的结果，判断是否采用已有的标准。

下面的两种情况下可以直接采用已有标准：

- 已有标准本身就是领域共享交换标准，且描述规范。如果描述不规范，需要按照标准化描述方法进行文档处理；
- 已有数据标准可以直接满足需求的，且描述规范。如果描述不规范，需要按照标准化描述方法进行文档处理；

6.3.2 数据模式建立

基于前面阶段产生的数据模式需求分析结果，共享数据模式建立工作的内容包括如下的五个主要部分：

1) 定义实体

分析和识别所有实体，给出各实体的准确定义。

2) 定义属性

根据各实体的定义，按照实体的特性和行为，确定实体的属性。

3) 确立值域

确定实体中各属性值的取值范围。

4) 定义主键

确立实体的唯一标识属性。

5) 定义关系

确定实体和实体之间的相互关系。

6.3.2.1 定义实体

实体是现实世界的事物或现象的抽象表示。它既能够表示实际事物，也能够表示抽象的概念。同类实体的成员具备相同的属性和特征。在数据模式中，实体是基本的数据组织单元。分析和确定实体是建立数据模式的第一步工作。然后确认实体的定义和命名。

6.3.2.1.1 实体判别

首先，基于前面阶段产生的如同**A.2** 数据模式标准的草案文档中的共享数据集组成描述的实体描述字典，开始逐个对于每个记录归档的字段项信息进行鉴别和审核，判断共享数据模式的真正实体。是，就保留该实体，进入到下面的实体定义。

否则，就是不符合科学数据共享项目需求的实体，需要添加“不符合的原因”到模板表格的“备注”字段内，作为删除的标记，同时建议将挑拣出来形成不符合共享数据模式的集合，单独的记录文档资料。同时注意，判断是否为实体的属性或者关系。

实体可以从以下几个角度进行判断：

- 该事物能否被描述

  所有实体必须能够被准确的描述。对于那些不能够被准确描述的对象，可以确定它们不是实体。

- 该事物是否具有多个实例

  实体是同类事物的抽象表示，因此它能够被实例化。即在现实或抽象世界中，存在一组具备相同属性和特性的事物或现象，它们具备有不同的取值，同时他们属于同一个实体。不能够被实例化的肯定不是实体。

- 该事物的一个实例是否能与其他实例区别开来

  同一个实体的各个不同实体是相互独立的，虽然它们有相同的特征和属性。

- 该事物是否还能描述其他事物

  如果该事物还可以用于描述其它事物，那么它是一个属性而不是实体。

6.3.2.1.2 定义实体

接下来的工作是进行实体的定义，包括三部分内容：实体命名、实体定义、实体别名。对于确认后的实体信息，分别记录归档到**A.2** 数据模式标准的草案文档中的第五章共享数据集组成描述的实体描述字典。

1) 实体命名

在逐个定义每个实体的名称时，需要注意：

- 实体名称能够唯一标识实体。

  不能有重复的名称，每一个实体必须有唯一的名字，意味着只要是相同名字就拥有相同的含义。反过来，同样含义就需要相同的名字，使用别名时例外。

- 实体名称具有一定的含义。

  实体对应于现实世界中的一个物体，或者是一个抽象的概念。实体的名称最好具有代表性，是可以表明它内涵的一个名词。

2) 实体定义

准确描述实体的定义，包括对实体的内容、范围、使用等多方面的描述。由于实体代表一组真实的或抽象的事物（人、物体、地点、事件、想法、多个事物的结合体等），它们拥有共同的属性和特征。那么在定义实体时，必须注意它的定义的内涵和外延是否足够概括这一类型的所有事物，是属于逻辑层次上的实体定义。

3) 实体别名

在现实生活中大量存在相同的事物拥有不同名称的情况，为了使人们能够更加方面的了解实体的含义，可以根据不同的使用习惯，给出该实体的别名。

6.3.2.2 定义关系

实体之间存在着相互联系，这种联系需要通过实体之间的关系定义来表达。实体之间的关系类型包括：关联关系、继承关系、包含关系和依赖关系。

实体之间的关系定义，需要包括的主要内容：

1) 相关性描述（确定关系类型）、

2) 关系名称确定、

3) 关系内容描述。

在确定关系名称时，需要确定的三个原则是：

- 关系名称是确定的。
- 关系名称是准确的。
- 关系名称是有意义的。

6.3.2.3 定义主键

实体可具备多个实例，不同实例之间通过主键取值的不同来区分。主键也是实体的一个属性。

在泛化关系中，主键可以而且必须继承。即泛化关系中的父类定义的主键同时也必将是子类的主键或子类的主键成员。

在定义主键值域时，不能包含空值。

6.3.2.4 定义属性

属性是实体的基本特征的描述。定义属性主要是定义非主键属性，主键属性已在前一部分单独定义。

属性的定义内容包括：

1) 属性名称：用于标识属性，不同的属性具备不同的属性名称。属性名称应当具备一个的含义。

2) 属性别名：属性的其它名称，可以方便人们快速了解属性的含义。

3) 属性定义：对属性的含义进行详细的解释。

6.3.2.5 确立值域

确立值域就是确定属性的数据类型和取值范围。

6.3.3 数据模式描述

共享数据模式建立完成后，形成了完整的“实体－关系”定义。对实体－关系的准确描述可采用本标准上一章所介绍的“数据模式描述规则”进行规范化描述。

输出的阶段性成果为：整理完整的“数据模式标准的草案文档和其相应的UML图”参照附录A.3 数据模式标准的草案文档和其相应的附录（数据模式标准的草案UML图）。

6.4 标准协调

针对已完成的数据模式标准草案文本和UML图，先在标准项目组范围内进行功能和技术上的审核、验证，再在领域范围内广泛征求意见并对标准进行修订，最终形成正式的数据模式标准文档。

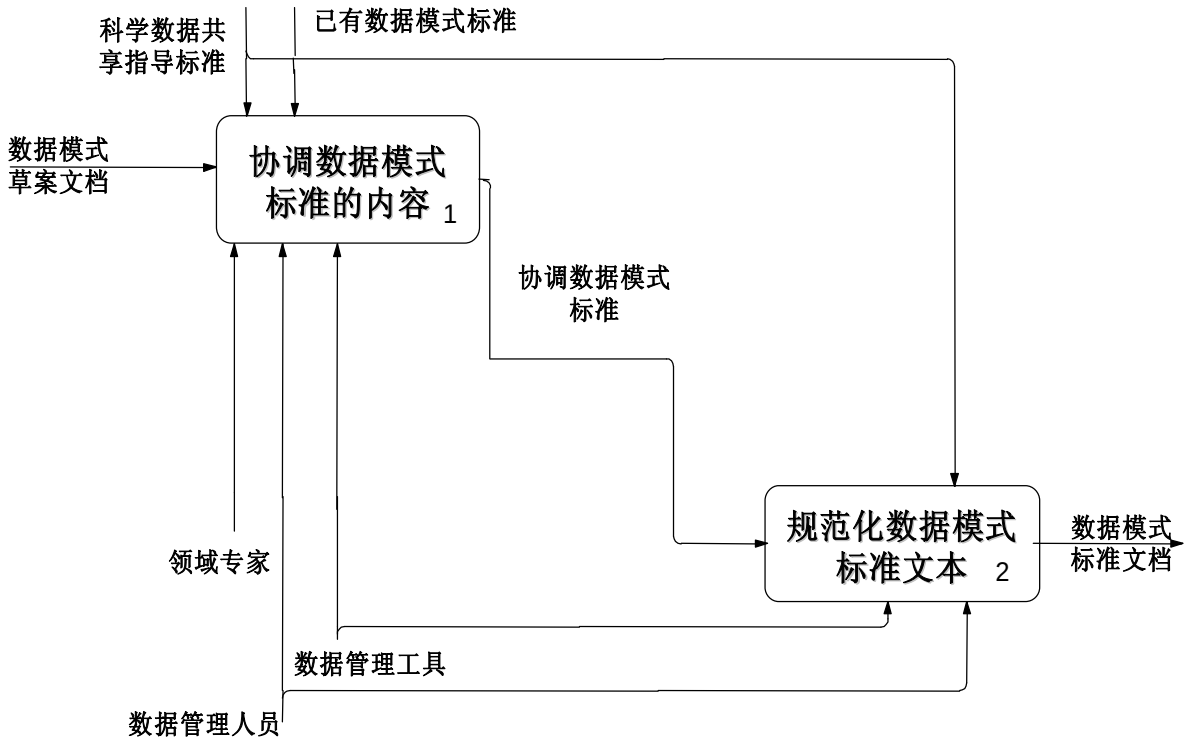


图19 协调数据模式标准的流程图

标准协调阶段包括两个主要步骤：

1) 协调数据模式标准的内容

针对已完成的数据模式标准草案文本和UML图，进行意见的广泛征求工作。

2) 规范化数据模式标准文本

根据上个步骤提交的数据模式标准文本，遵照规定的体例书写，最后提交正式的数据模式标准文档和该标准相应的附录。

6.4.1 协调数据模式标准的内容

针对已完成的数据模式标准草案文本和UML图，在标准项目组范围内，分别从功能角度、技术角度进行协调，然后在领域范围内，进行意见的广泛征求工作。

1) 功能协调

从应用的角度，考察标准内容对科学研究活动的支持程度，并根据考察结果，对数据模式标准文本进行修订。

2) 技术协调

组织领域信息化专家，针对数据模式模型的实体、属性、关系以及整个数据模式模型，进行技术方面的审核。技术审核的目标：

- 确保提交的数据模式标准与已有的相关的本标准不冲突。
- 验证并整合提交的数据模式标准与正在使用的原有数据模式标准版本。
- 确保实体属性的命名定义，满足需求列表，符合本标准的要求。
- 检验关系的命名。
- 检验功能角度的实现。

3) 征求意见

完成对于征求意见的数据模式标准文本的修改，并返回本领域的范围，广泛征求意见。要求填写相应的“数据模式标准的征求意见表格”，参见本文档附录**A.5.** 数据模式标准的征求意见文档(资料性附录)说明部分。

根据反馈意见，对数据模式标准内容进行修订，对于不采纳的意见要说明理由。需要填写相应的“数据模式标准的意见处理表格”，参见本文档附录**A.6.** 数据模式标准的意见处理文档(资料性附录)说明部分。

6.4.2 规范化数据模式标准文本

根据上个步骤提交的数据模式标准文本，遵照**GB/T 1.1-2000**规定的体例书写规范，最后提交正式的数据模式标准文档和其相应的附录。

输出的阶段性成果为：修订确认后的“数据模式标准正式文档和其相应的附录”，参照附录 A.3 数据模式标准文档和该文档相应的附录（数据模式标准的 UML 图附录，数据模式标准实现的 SQL 示例附录，数据模式标准实现的 XML 示例附录）

6.5 标准实现

如图所示，根据制定出的数据模式标准文档和其它标准（科学数据共享标准和其它标准），考虑具体的信息技术，建立数据存储模式，并形成数据交换格式。

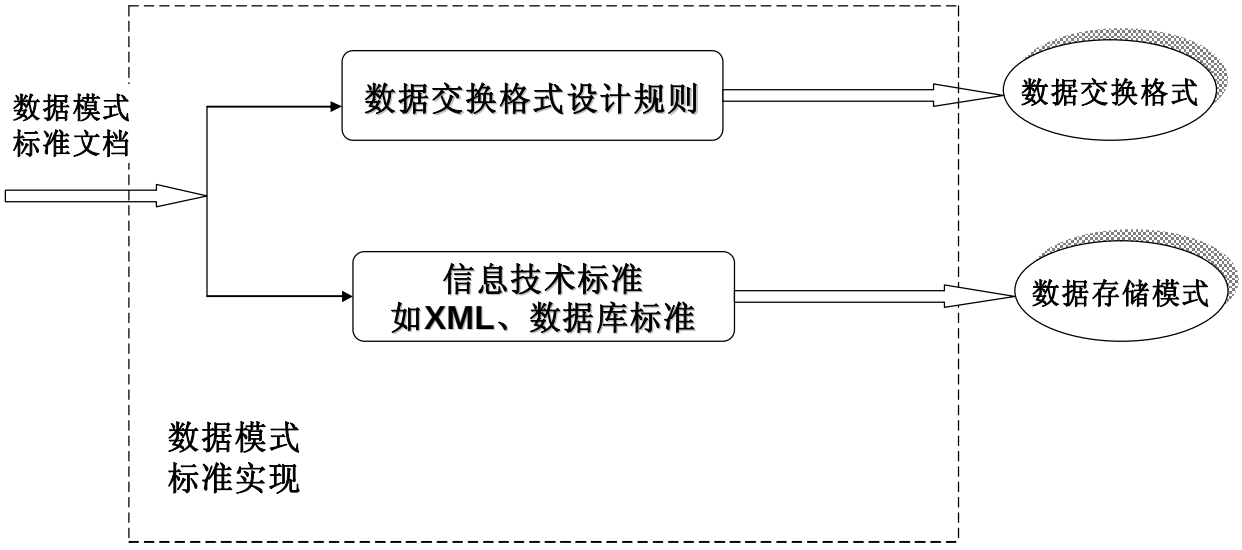


图20 实现数据模式标准的示意图

数据模式本身与具体平台没有关系，可以映射到各个平台的（例如数据库、**XML**、其它形式的数据系统等）。建立数据模式的工作重点在于逻辑层次上，是数据的概念、内容、组成及其结构的描述信息的有机组合。它包含有统一的的描述规则和规范化的构建方法，来清晰、科学的描述本领域的共享数据集的概念、内容、组成及其结构信息。在逻辑层次上进行描述说明。不但可以基于概念数据模式，具有本质化、清晰化和概念化的特点，又结合了信息技术实现的因素，进一步的分析并增加各种对象和其关

系。同时又能够高于具体的物理实现，拥有更大的灵活性、抽象性和概括性，是作为物理数据模式建立的基础。

1) 确定领域数据交换格式

利用制定出的正式的数据模式标准，以及数据交换格式设计规则，便可制定出该领域数据的交换格式，应用于科学数据共享范围（在多个领域间和领域内）进行数据交换活动。

2) 制定领域数据存储模式

利用制定出的正式的数据模式标准和信息技术中已有的标准（如**XML**、关系数据库技术），可以制定出该领域数据的存储模式。

3) 其它

开发标准的支持工具，从而能够支持数据从逻辑模式到物理模式的转换或映射，如向**XML**模式、数据库关系模式的转换。

附 录 A
数据模式建立与描述的文档模板
(资料性附录)

A.1 概述

本附录为数据模式建立与描述过程中四个阶段产出的全部文档提供了具体模板,可以在实际的数据模式建立与描述工作中应用和参考。

A.1.1 数据模式标准的需求收集文档模板

在确立数据模式标准服务目标与范围的基础上，广泛的收集相关资料，并依据实际情况的限制，进行数据集组成结构信息的分析提取，必要时通过相应的工具软件来进行分析、捕捉。同时考查本领域现有的数据标准，最后确定出数据模式标准需求文档。该文档模板主要包含内容，分别如下所示。

- 概述主体数据库的基本内容。
- 概述数据集的基本内容，并说明组织方式（以数据集为单位描述其相关的数据库、数据文件和标准）。

A.1.2 数据模式标准的草案文档模板

对于建立完成的共享数据模式（形成了完整的“实体－关系”定义），采用本数据模式描述规则和方法标准所介绍的“数据模式描述规则”进行规范化描述；最后形成的数据模式标准的草案文档。该文档模板主要包含内容，分别如下所示。

- 概述描述该数据集的基本信息，包括数据集内容、来源、数据量、数据更新方式等。
- 概述介绍数据集数据模式的基本内容，并说明组织方式（以数据集为单位，描述其数据模式包含的实体、属性和相互关系）。

A.1.3 数据模式标准的正式文档模板

针对已完成的数据模式标准草案文本和UML图，先在标准项目组范围内进行功能和技术上的审核、验证，再在领域范围内广泛征求意见并对标准进行修订，最终形成正式的数据模式标准文档。该文档模板主要包含内容，分别如下所示。

- 概述描述该数据集的基本信息，包括数据集内容、来源、数据量、数据更新方式等。
- 概述介绍数据集数据模式的基本内容，并说明组织方式（以数据集为单位，描述其数据模式包含的实体、属性和相互关系）。
- 概述一致性测试的基本内容。
- 数据模式标准实现的SQL示例附录部分。
- 数据模式标准实现的XML示例附录部分。

A.1.4 数据模式标准的征求意见文档模板

针对已完成的数据模式标准草案文本和UML图，先在标准项目组范围内进行功能和技术上的审核、验证，再在领域范围内广泛征求意见并对标准进行修订。遵照项目需求和约定，建议填写到本征求意见文档模板。

- 概括介绍此次征求意见的数据模式标准的基本内容和重点问题。
- 说明此次征求意见的范围。

A.1.5 数据模式标准的意见处理文档模板

针对已完成的数据模式标准的征求意见文档模板，采集到的广泛意见。按照要求填写本意见处理文档模板，准备进行对数据模式标准的修订工作，生成最终的数据模式标准的正式文档和相应附录。

- 说明此次数据模式标准的意见征求工作的范围。
- 说明此次数据模式标准的意见征求工作的反馈情况。
- 说明此次意见征求工作对于数据模式标准修订的作用。

A.2 数据模式标准的需求收集文档(资料性附录)

科学数据共享工程

<项目名称>

数据模式标准需求收集文档

文档版本：1.0

文档编号：XXX

（本稿完成日期：**XX** 年 **XX** 月）

修订版历史

日期	版本	说明	作者
<日/月/年>	<x.x>	<详细信息>	<姓名>

目录

[本小节应完整地列出此**数据模式标准需求收集文档**中所有的标题及其对应页码。]

1. 范围……页码
2. 参考文档……页码
3. 术语与缩写……页码
    3.1 术语……页码
    3.2 缩写……页码
4. 概述……页码
5. 主体数据库说明……页码
6. 数据集说明……页码
    6.1.数据集概述……页码
    6.2.XX 数据集详细说明……页码

数据模式标准需求收集

1. 范围

[本小节说明文档的内容和适用范围。]

- 主要内容、作用和意义等等。
- 适用范围（包括领域、部门单位、项目、阶段、人员等等）。

2. 参考文档

[本小节列出所有引用到的文档。]

列举所有参考文档的作者（发布单位）、标题、版本、日期等信息。

示例：

GB/T 18793—2002　信息技术　可扩展置标语言（XML）1.0

3. 术语与缩略语

[本小节给出了相关术语定义和缩略语的含义。]

3.1 术语

[本小节给出了相关术语定义。]

示例：

*类 **Class***

对拥有相同的属性、操作、方法、关系和语义的一组对象的描述。

在 UML 中类的图形符号是一个矩形框。其中标注出该类的名称，即为对应实体的名称。

3.2 缩略语

[本小节给出了相关缩略语的含义。]

示例：

UML

统一建模语言，Unified Modeling Language。

4. 概述

[本小节概述第五章和第六章的主要内容以及组织方式。]

- 概述第五章：主体数据库的基本内容。
- 概述第六章：数据集的基本内容，并说明第六章的组织方式（以数据集为单位描述其相关的数

据库、数据文件和标准）。

5. 主体数据库说明

[本小节描述了主体数据库的基本信息。]

主体数据库（Core database）是依据国际标准、国家标准或行业标准分类体系构建的二级学科及其分支学科的科学数据集，并基于计算机系统运行的数据库。

表6 主体数据库列表

序号	名称	内容
1.		
2.		

表 1 中各项的含义如下：

- **序号**：为唯一标识主体数据库而分配的序号。
- **名称**：主体数据库的名称。
- **内容**：内容包括两部分，一是主体数据库内容的概括性描述，二是主体数据库所包含的数据集的简单说明，有关数据集的详细情况将在第 3 章说明。

6. 数据集说明

6.1 数据集概述

[本小节描述了数据集的详细情况。]

数据集（Data Set）是可以标识的数据集合。

表7 数据集列表

序号	名称	内容	来源	限制信息
1.				
2.				

表 2 中各项的含义如下：

- **序号**：为唯一标识数据集而分配的序号。
- **名称**：数据集的名称。
- **内容**：简要描述该数据集的内容信息。
- **来源**： 简要描述该数据集和业务系统或者科研项目的对应关系，并列出该数据集对应的数据库、数据文件、相关标准的名称。
- **限制信息**：包括三方面的限制信息，一是出于安全保密而对数据集施加的限制，二是出于知识产权保护对数据集施加的限制，三是由于数据本身的质量和精度而造成的适用性局限。

6.2 XX 数据集详细说明

[本小节描述了单个数据集的详细情况，包括数据集对应的数据库、数据文件、相关标准的具体说明。]

本节内容可以根据需要而循环阐述。

6.2.1 数据库说明

6.2.1.1 数据库概述

[简要描述了该 XX 数据集对应的数据库基本情况。]

表8 数据库列表

序号	中文名称	英文名称	内容
1.			
2.			

表 3 中各项的含义如下：

- **序号**：为唯一标识数据库而分配的序号。
- **中文名称**：数据库的中文名称。
- **英文名称**：数据库的英文名称。
- **内容**：简要描述数据库的内容信息。

6.2.1.2 XX 数据库详细说明

[具体描述该数据库的结构情况（其所包含的数据表信息列表）。可在此加入数据库的实体－关系图]

本节内容可以根据需要的数据库数量而循环阐述。

表9 XX 数据库结构

序号	表中文名称	表英文名称	表内容

表 4 中各项的含义如下：

- **序号**：为唯一标识数据库所包含的数据表而分配的序号。
- **表中文名称**：数据库所包含的数据表的中文名称。
- **表英文名称**：数据库所包含的数据表的英文名称。

- **表内容**：简要描述该数据库所包含的数据表的内容信息。

6.2.1.2.1 XX 表结构说明

[本小节包含该 XX 数据库所包含的各个表的定义。]

本节内容可以根据需要的表数量而循环阐述。

表10 XX 表结构

序号	字段中文名称	字段英文名称	字段类型	字段含义	值域	主/外键	备注

表 5 中各项的含义如下：

- **序号**：为唯一标识该数据表内所有的字段而分配的序号。
- **字段中文名称**：该数据表内字段的中文名称。
- **字段英文名称**：该数据表内字段的英文名称。
- **字段类型**：该数据表内字段的数据类型。
- **字段含义**：简要描述该数据表所有的字段的含义。
- **值域**：该数据表内字段的值域范围。
- **主/外键**：该字段的是否为主/外键。
- **备注**：该字段的备注信息。

6.2.2 数据文件说明

6.2.2.1 数据文件概述

[本小节概括性的描述该数据集所含的数据文件的信息列表：基本内容、来源、数据量、数据更新方式等相关信息。]

表11 数据文件列表

序号	中文名称	英文名称	内容

表 6 中各项的含义如下：

- **序号**：为唯一标识数据文件而分配的序号。

- **中文名称**：数据文件的中文名称。
- **英文名称**：数据文件的英文名称。
- **内容**：简要描述数据文件的内容信息。

6.2.2.2 XX 数据文件详细说明

[具体描述该（XX）数据文件的组成结构情况。]

本节内容可以根据需要的数据文件数量而循环阐述。

6.2.3 相关标准说明

6.2.3.1 相关标准概述

[简要描述该数据集对应的相关标准，包括：数据元标准，概念数据模型标准，逻辑数据模型标准，交换格式标准，指标体系标准等等。]

6.2.3.2 相关标准详细说明

[本小节列出相关数据标准采标情况说明信息。]

表12 相关标准列表

序号	名称	发布单位	数据集名称	采用内容	采标情况

表 7 中各项的含义如下：

- **序号**：唯一标识相关标准所分配的序号。
- **名称**：该标准的具体名称。
- **发布单位**： 该标准的发布单位。
- **数据集名称**： 该标准所包含的数据集名称。
- **采用内容信息**： 该标准的具体采用内容和其位置信息。
- **采标情况**：示例如下三种取值。

完全采标：*已有的标准可以直接满足科学数据共享的需求，且描述规范，可直接引用该标准作为科学数据共享的数据模式标准；如果描述不规范，需要按照标准化描述方法进行文档处理。或者已有标准本身就是领域共享交换标准，且描述规范。如果描述不规范，需要按照标准化描*

述方法进行文档处理；

***部分采标**：已有的标准只能部分满足科学数据共享的需求，挑选出符合科学数据共享需求的内容，在此基础上进行修订并增加内容，形成科学数据共享的数据模式标准。*

***不采用**：不符合科学数据共享需求的标准，但是需要记录归档不予采纳的原因，作为数据模式标准协调阶段的参照资料。*

A.3 数据模式标准的草案文档(资料性附录)

科学数据共享工程

<项目名称>

数据模式标准的草案文档

文档版本：**1.0**

文档编号：**XXX**

（本稿完成日期：**XX** 年 **XX** 月）

修订版历史

日期	版本	说明	作者
<日/月/年>	<x.x>	<详细信息>	<姓名>

目录

[本小节应完整地列出此**数据模式标准草案文档**中所有的标题及其页码。]

1. 范围……页码
2. 参考文档……页码
3. 术语与缩写……页码
   3.1 术语……页码
   3.2 缩写……页码
4. 概述……页码
5. 数据集数据模式……页码
   5．1 数据模式整体框架……页码
   5．2 XX 实体……页码

数据模式草案标准

1. 范围

[本小节说明文档的内容和适用范围。]

- 主要内容、作用和意义等等。
- 适用范围（包括领域、部门单位、项目、阶段、人员等等）。

2. 参考文档

[本小节列出所有引用到的文档。]

列举所有参考文档的作者（发布单位）、标题、版本、日期等信息。

示例：

GB/T 18793—2002　信息技术　可扩展置标语言（XML）1.0

3. 术语与缩略语

[本小节给出了相关术语定义和缩略语的含义。]

3.1 术语

[本小节给出了相关术语定义。]

示例：

*类 **Class***

对拥有相同的属性、操作、方法、关系和语义的一组对象的描述。

在 UML 中类的图形符号是一个矩形框。其中标注出该类的名称，即为对应实体的名称。

3.2 缩略语

[本小节给出了相关缩略语的含义。]

示例：

UML

统一建模语言，Unified Modeling Language。

4. 概述

[本小节概述数据集的基本信息，同时说明了本文档的组织结构和各个章的主要内容。]

（1） 概述描述该数据集的基本信息，包括数据集内容、来源、数据量、数据更新方式等。

表13 数据集信息列表

序号	字段名称	定义	填写信息
1	数据集标识符	数据集的唯一标识符	
2	数据集语种	数据集使用的语言	
3	数据集联系方	对数据集信息负责的单位或个人	
4	数据集创建日期	数据集创建的日期	
5	数据集名称	数据集名称	
6	数据集版本	数据集版本	
7	数据集标识信息	数据集描述的资源的基本信息	
8	数据集内容信息	提供数据内容特征的描述信息	
9	数据集分发信息	提供获取资源所需的分发者和分发方式的信息	
10	数据集数据质量信息	提供资源质量的总体评价信息	
11	数据集数据表现形式信息	数据集信息的数据表示形式	
12	数据集应用模式信息	提供有关数据集概念模式的信息	
13	数据集限制信息	提供访问和使用数据集的限制信息	
14	数据集维护信息	提供有关数据集的更新频率及更新范围的信息	

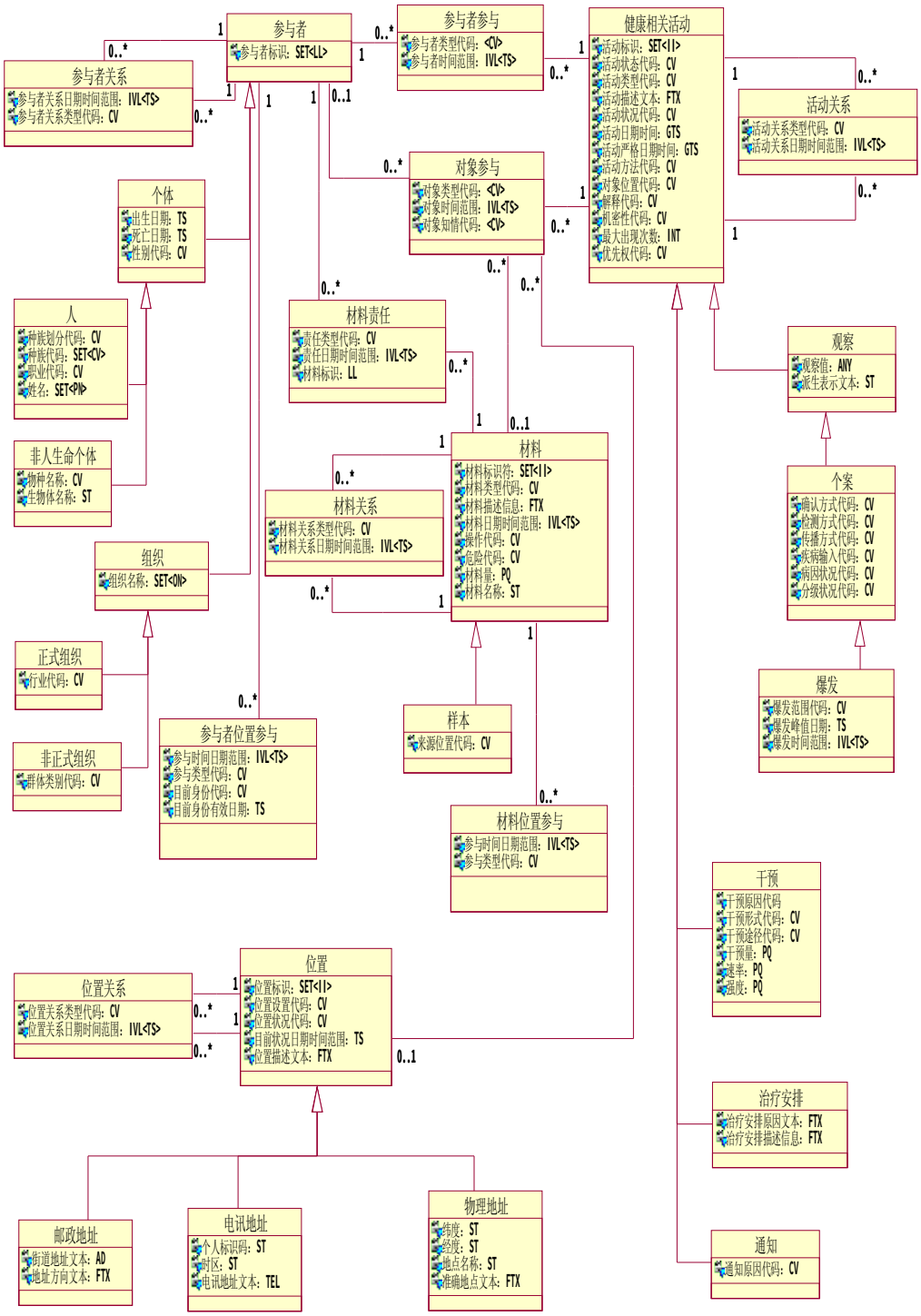
（2） 概述介绍第五章：数据集数据模式的基本内容，并说明第五章的组织方式（以数据集为单位，描述其数据模式包含的实体、属性和相互关系）。

5. 数据集数据模式

5.1 数据模式整体框架

[本小节描述数据模式构成，采用 UML 静态类图形式和文字结合的方式进行描述。]

（1） 绘制数据模式构成的整体框架图



（2） 分类描述数据模式的组成内容。

（3） 列出数据模式的所有实体。

表14 实体列表

序号	实体名称	定义
3.		
4.		
5.		

表 1 中各项的含义如下：

- **序号**：为唯一标识数据集而分配的序号。
- **实体名称**：简要描述该数据模式所含实体名称。
- **定义**： 简要描述该实体的定义。

5.2 XX 实体

[本小节描述 XX 实体的详细信息。]

本节内容可以根据需要而循环阐述。

5.2.1 实体信息

[本小节依据该实体的定义和说明信息，填写对应的实体描述字典。]

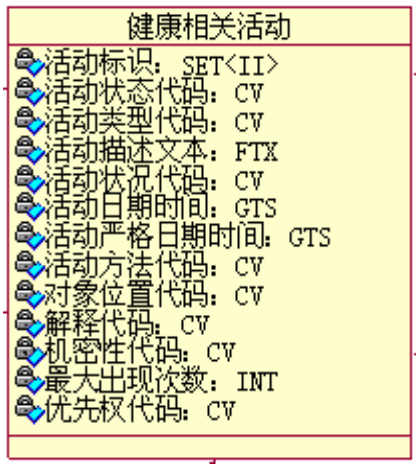
表15　实体描述字典

字段	可选性	描述	填写信息
实体名称	必选	实体的标识，一般使用名词表达，通常名称都能反映出实体的属性和特征。	
别名	可选	实体的别名，一般使用名词表达。	
英文名称	必选	实体的英文全称	
短名	必选	实体的英文名称缩写	
定义	必选	实体定义的详细描述	
备注	可选	和实体相关的其它信息	
版本标识	必选	用于实体的配置管理和控制。	
状态	必选	0：讨论版本 1：正式版本	
实体来源	可选	说明实体定义的来源，来源包括已有的数据模式草案标准、已有的信息系统和其它来源。	

5.2.2 实体构成说明

[本小节描述该实体构成，采用 UML 静态类图形式和文字结合的方式进行描述。]

（1） 绘制该实体的 UML 静态类图。



（2） 列出该实体的所有属性。

表16 属性列表

序号	属性名称	定义
6.		
7.		
8.		

表 1 中各项的含义如下：

- **序号**：为唯一标识该属性而分配的序号。
- **属性名称**：简要描述该实体所含的属性名称。
- **定义**： 简要描述该属性的定义。

5.2.2.1 XX 属性信息

[本小节依据该属性的定义和说明信息，填写对应的属性描述字典。]

本节内容可以根据需要而循环阐述。

表17　属性描述字典

属性名称	可选性	描述	填写信息
数据元素名称	必选	属性标识，一般使用名词表达，属性名称能够反映属性的特征。	
别名	比选	属性的别名	
英文名称	必选	属性的英文全称	
短名	必选	属性的英文缩写	
定义	必选	实体定义的详细描述	
备注	可选	和实体相关的其它信息	
版本标识	必选	用于实体的配置管理和控制。	
状态	必选	0：讨论版本 1：正式版本	
属性来源	可选	说明属性定义的来源，来源包括已有的数据模式草案标准、已有的信息系统和其它来源。	
数据类型名称	必选	属性的数据类型，例如字符型、数值型、逻辑型等。	
SQL数据类型	可选	该属性在关系型数据库中的数据类型，按照结构化查询语言的数据类型表达方式进行描述，例如varchar(100)，代表可变长字符串，最大长度单位100个字符。	
值域	必选	属性的取值范围	

安全说明	必选	说明该属性的安全限制信息，包括访问和使用限制等。	

A.4 数据模式标准的正式文档(资料性附录)

科学数据共享工程

<项目名称>

数据模式标准文档

文档版本：**1.0**

文档编号：**XXX**

（本稿完成日期：**XX** 年 **XX** 月）

修订版历史

日期	版本	说明	作者
<日/月/年>	<x.x>	<详细信息>	<姓名>

目录

[本小节应完整地列出此**数据模式标准草案文档**中所有的标题及其页码。]

1. 范围……页码
2. 参考文档……页码
3. 术语与缩写……页码
    3.1 术语……页码
    3.2 缩写……页码
4. 概述……页码
5. 数据集数据模式……页码
    5．1 数据模式整体框架……页码
    5．2 XX 实体……页码
6. 一致性检测……页码
附录 A1 数据模式标准实现的 SQL 示例 (资料性附录) ……页码
附录 A2 数据模式标准实现的 XML 示例 (资料性附录) ……页码

数据模式标准

1. 范围

[本小节说明文档的内容和适用范围。]

- 主要内容、作用和意义等等。
- 适用范围（包括领域、部门单位、项目、阶段、人员等等）。

2. 参考文档

[本小节列出所有引用到的文档。]

列举所有参考文档的作者（发布单位）、标题、版本、日期等信息。

示例：

GB/T 18793—2002　信息技术　可扩展置标语言（XML）1.0

3. 术语与缩略语

[本小节给出了相关术语定义和缩略语的含义。]

3.1 术语

[本小节给出了相关术语定义。]

示例：

*类 **Class***

对拥有相同的属性、操作、方法、关系和语义的一组对象的描述。

在 UML 中类的图形符号是一个矩形框。其中标注出该类的名称，即为对应实体的名称。

3.2 缩略语

[本小节给出了相关缩略语的含义。]

示例：

UML

统一建模语言，Unified Modeling Language。

4. 概述

[本小节概述数据集的基本信息，同时说明了本文档的组织结构和各个章的主要内容。]

（3） 概述描述该数据集的基本信息，包括数据集内容、来源、数据量、数据更新方式等。

表18 数据集信息列表

序号	字段名称	定义	填写信息
15	数据集标识符	数据集的唯一标识符	
16	数据集语种	数据集使用的语言	
17	数据集联系方	对数据集信息负责的单位或个人	
18	数据集创建日期	数据集创建的日期	
19	数据集名称	数据集名称	
20	数据集版本	数据集版本	
21	数据集标识信息	数据集描述的资源的基本信息	
22	数据集内容信息	提供数据内容特征的描述信息	
23	数据集分发信息	提供获取资源所需的分发者和分发方式的信息	
24	数据集数据质量信息	提供资源质量的总体评价信息	
25	数据集数据表现形式信息	数据集信息的数据表示形式	
26	数据集应用模式信息	提供有关数据集概念模式的信息	
27	数据集限制信息	提供访问和使用数据集的限制信息	
28	数据集维护信息	提供有关数据集的更新频率及更新范围的信息	

（4） 概述介绍第五章：数据集数据模式的基本内容，并说明第五章的组织方式（以数据集为单位，描述其数据模式包含的实体、属性和相互关系）。

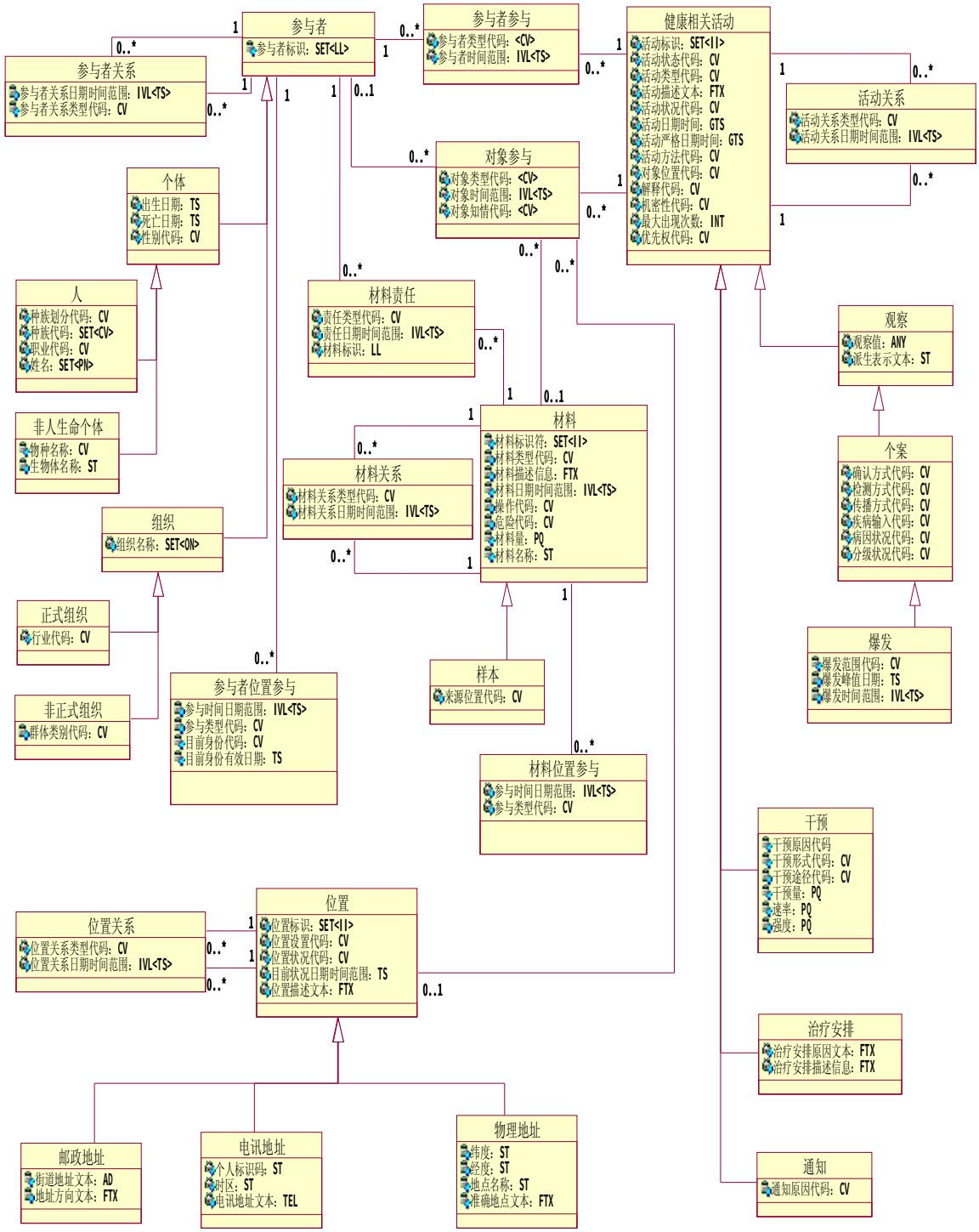
（5） 概述第六章一致性测试的基本内容。

5. 数据集数据模式

5.1 数据模式整体框架

[本小节描述数据模式构成，采用 UML 静态类图形式和文字结合的方式进行描述。]

（4） 绘制数据模式构成的整体框架图



（5） 分类描述数据模式的组成内容。

（6） 列出数据模式的所有实体。

表19 实体列表

序号	实体名称	定义
9.		
10.		
11.		

表 1 中各项的含义如下：

- **序号**：为唯一标识数据集而分配的序号。
- **实体名称**：简要描述该数据模式所含实体名称。
- **定义**： 简要描述该实体的定义。

5.2 XX 实体

[本小节描述 XX 实体的详细信息。]

本节内容可以根据需要而循环阐述。

5.2.1 实体信息

[本小节依据该实体的定义和说明信息，填写对应的实体描述字典。]

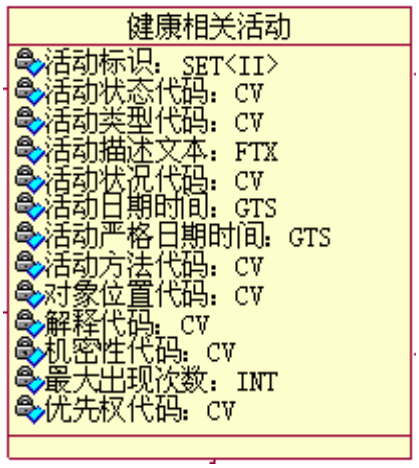
表20　实体描述字典

字段	可选性	描述	填写信息
实体名称	必选	实体的标识，一般使用名词表达，通常名称都能反映出实体的属性和特征。	
别名	可选	实体的别名，一般使用名词表达。	
英文名称	必选	实体的英文全称	
短名	必选	实体的英文名称缩写	
定义	必选	实体定义的详细描述	
备注	可选	和实体相关的其它信息	
版本标识	必选	用于实体的配置管理和控制。	
状态	必选	0：讨论版本 1：正式版本	
实体来源	可选	说明实体定义的来源，来源包括已有的数据模式草案标准、已有的信息系统和其它来源。	

5.2.2 实体构成说明

[本小节描述该实体构成，采用 UML 静态类图形式和文字结合的方式进行描述。]

（3） 绘制该实体的 UML 静态类图。



（4） 列出该实体的所有属性。

表21 属性列表

序号	属性名称	定义
12.		
13.		
14.		

表 1 中各项的含义如下：

- **序号**：为唯一标识该属性而分配的序号。
- **属性名称**：简要描述该实体所含的属性名称。
- **定义**： 简要描述该属性的定义。

5.2.2.1 XX 属性信息

[本小节依据该属性的定义和说明信息，填写对应的属性描述字典。]

本节内容可以根据需要而循环阐述。

表22　属性描述字典

属性名称	可选性	描述	填写信息
数据元素名称	必选	属性标识，一般使用名词表达，属性名称能够反映属性的特征。	
别名	比选	属性的别名	
英文名称	必选	属性的英文全称	
短名	必选	属性的英文缩写	
定义	必选	实体定义的详细描述	
备注	可选	和实体相关的其它信息	
版本标识	必选	用于实体的配置管理和控制。	
状态	必选	0：讨论版本 1：正式版本	
属性来源	可选	说明属性定义的来源，来源包括已有的数据模式草案标准、已有的信息系统和其它来源。	
数据类型名称	必选	属性的数据类型，例如字符型、数值型、逻辑型等。	
SQL数据类型	可选	该属性在关系型数据库中的数据类型，按照结构化查询语言的数据类型表达方式进行描述，例如varchar(100)，代表可变长字符串，最大长度单位100个字符。	
值域	必选	属性的取值范围	

安全说明	必选	说明该属性的安全限制信息，包括访问和使用限制等。	

6. 一致性测试

[描述数据模式一致性测试的框架、方法和达到标准一致性要求的条件。]

根据具体条件，可以提供以下内容辅助一致性测试，包括：

- 标准一致性测试工具软件包，该部分内容是纯的软件工具。
- 标准测试用例库，该部分内容包括一个完成的测试用例序列，每一个用例都是针对标准一致性的某一个方面而设计。
- 标准测试指南，主要说明测试目标、测试方法、标准测试流程及操作说明、标准一致性指标内容及其解释。

附录 A 1
数据模式标准的 SQL 实现示例
(资料性附录)

1．概述

[本小节概括说明数据模式标准实现的 SQL 示例基本内容和情况。]

2．SQL 实现示例

[本小节具体给出实现数据模式的 SQL 示例。]

附录 A2
数据模式标准的 XML 实现示例
(资料性附录)

1．概述

[本小节概括说明数据模式标准实现的 XML 示例基本内容和情况。]

2．XML 实现示例

[本小节具体给出实现数据模式的 XML 示例。]

A.5 数据模式标准的征求意见文档(资料性附录)

科学数据共享工程

<项目名称>

数据模式标准的征求意见文档

文档版本：**1.0**

文档编号：**XXX**

（本稿完成日期：**XX** 年 **XX** 月）

修订版历史

日期	版本	说明	作者
<日/月/年>	<x.x>	<详细信息>	<姓名>

目录

[本小节应完整地列出此**数据模式标准的征求意见文档**中所有的标题及其页码。]

1. 范围……页码
2. 参考文档……页码
3. 术语与缩写……页码
   3.1 术语……页码
   3.2 缩写……页码
4. 概述……页码
5. 征求意见表的说明……页码
附录 A 科学数据共享数据模式标准的征求意见表格(资料性附录) ……页码

数据模式标准征求意见

1. 范围

[本小节说明文档的内容和适用范围。]

- 主要内容、作用和意义等等。
- 适用范围（包括领域、部门单位、项目、阶段、人员等等）。

2. 参考文档

[本小节列出所有引用到的文档。]

列举所有参考文档的作者（发布单位）、标题、版本、日期等信息。

示例：

GB/T 18793—2002　信息技术　可扩展置标语言（XML）1.0

3. 术语与缩略语

[本小节给出了相关术语定义和缩略语的含义。]

3.1 术语

[本小节给出了相关术语定义。]

示例：

*类 **Class***

对拥有相同的属性、操作、方法、关系和语义的一组对象的描述。

在 UML 中类的图形符号是一个矩形框。其中标注出该类的名称，即为对应实体的名称。

3.2 缩略语

[本小节给出了相关缩略语的含义。]

示例：

UML

统一建模语言，Unified Modeling Language。

4. 概述

[本小节概述文档的主要内容和组织方式。]

- 概括介绍此次征求意见的数据模式标准的基本内容和重点问题。
- 说明此次征求意见的范围。

5. 数据模式标准的征求意见表说明

[本小节描述该数据模式的征求意见表格内容说明。]

附录 A

科学数据共享数据模式标准的征求意见表

(资料性附录)

<table>
<tr><td colspan="2">标准名称</td><td colspan="6"></td></tr>
<tr><td>姓名</td><td></td><td>电话</td><td></td><td>传　真</td><td></td><td>E-mail</td><td></td></tr>
<tr><td>单位</td><td colspan="3"></td><td>通信地址</td><td colspan="2"></td><td>邮编</td><td></td></tr>
<tr><td colspan="2">总体意见</td><td colspan="7"></td></tr>
<tr><td>章、条号</td><td colspan="3">修改建议</td><td colspan="5">修改理由</td></tr>
<tr><td></td><td colspan="3"></td><td colspan="5"></td></tr>
</table>

A.6　数据模式标准的意见处理文档(资料性附录)

科学数据共享工程

<项目名称>

数据模式标准的意见处理文档

文档版本：1.0

文档编号：XXX

（本稿完成日期：XX年XX月）

修订版历史

日期	版本	说明	作者
<日/月/年>	<x.x>	<详细信息>	<姓名>

目录

[本小节应完整地列出此**数据模式标准的意见处理文档**中所有的标题及其页码。]

1. 范围……页码
2. 参考文档……页码
3. 术语与缩写……页码
   3.1 术语……页码
   3.2 缩写……页码
4. 概述……页码
5. 意见汇总处理表的说明……页码

附录 A 科学数据共享数据模式标准的意见汇总处理表(资料性附录) ……页码

1. 范围

[本小节说明文档的内容和适用范围。]

- 主要内容、作用和意义等等。
- 适用范围（包括领域、部门单位、项目、阶段、人员等等）。

2. 参考文档

[本小节列出所有引用到的文档。]

列举所有参考文档的作者（发布单位）、标题、版本、日期等信息。

示例：

GB/T 18793—2002　信息技术　可扩展置标语言（XML）1.0

3. 术语与缩略语

[本小节给出了相关术语定义和缩略语的含义。]

3.1 术语

[本小节给出了相关术语定义。]

示例：

*类 **Class***

对拥有相同的属性、操作、方法、关系和语义的一组对象的描述。

在 UML 中类的图形符号是一个矩形框。其中标注出该类的名称，即为对应实体的名称。

3.2 缩略语

[本小节给出了相关缩略语的含义。]

示例：

UML

统一建模语言，Unified Modeling Language。

4. 概述

[本小节概述文档的主要内容和组织方式。]

- 说明此次数据模式标准的意见征求工作的范围。
- 说明此次数据模式标准的意见征求工作的反馈情况。

- 说明此次意见征求工作对于数据模式标准修订的作用。

5. 数据模式标准的意见汇总处理表说明

[本小节描述标准主要修订意见以及总体处理的方式。]

附录 A

科学数据共享数据模式标准的意见汇总处理表

(资料性附录)

共　　页　第　　页

标准名称　　负责起草单位

年　　月　　日　　填写

序号	标准章条编号	意　见　内　容	提出单位	处理意见	备注

说明：① 发送《征求意见稿》的单位数：　　个。

② 收到《征求意见稿》后，回函的单位数：　　个。

③ 收到《征求意见稿》后，回函并有建议或意见的单位数：　　个。

④ 没有回函的单位数：　　个。

附 录 B
反向工程示例
(资料性附录)

以**ORACLE**为例，描述反向工程的步骤说明，逆向提取需要的数据库表结构。

步骤一

根据前面确定的目标和范围，从已有数据库中，挑选需要的数据库表，并且记录归档到“数据库说明表格”。参见附录**A.2.1**数据模式标准的需求收集文档中的第七章数据库说明部分。

示例为，选择相关的交换任务队列的数据库数据库“**EXCHGSYS**”和表“**TASKQUEUE**“，添加到各自的表格。

格式如同下面的两个表格：

表 符合科学数据共享工程需要的数据库的列表

序号	名称	来源	内容	所有表组成信息	采用说明	所属范围名称	所属共享数据集名称	负责人员

注释：

- “序号”是为了唯一标识所需要数据库而自行分配的序号。
- “名称”是需要填写所选择的数据库的名称。
- “来源” 是需要填写所选择的数据库的来源信息。
- “内容”是需要填写所选择的数据库的内容。
- “所有表组成信息” 是需要填写该数据库内所有表的组成信息，例如本数据库包括多少个表，组成关系如何。
- “采用说明” 是需要填写该数据库被选择、采用的原因说明。
- “所属范围名称”是需要填写该数据库所属于的共享信息内容范围的名称。
- “所属共享数据集名称”是需要填写该数据库所对应的共享数据集的名称。

表 符合科学数据共享工程需要的数据表的列表

序号	名称	来源	内容	结构	采用说明	所属范围名称	所属共享数据集名称	负责人员

注释：

- “序号”是为了唯一标识所需要数据表而自行分配的序号。
- “名称”是需要填写所选择的数据表的名称。
- “来源” 是需要填写所选择的数据表的来源信息。
- “内容”是需要填写所选择的数据表的内容。
- “结构” 是需要填写该数据表的结构信息，例如本数据表包括多少个字段及其定义信息。
- “采用说明” 是需要填写该数据表被选择、采用的原因说明。
- “所属范围名称”是需要填写该数据表对应的所属于的共享信息内容范围的名称。
- “所属共享数据集名称”是需要填写该数据表对应的所属于的共享数据集的名称。

步骤二

选择适合工具，按照对应步骤和帮助文档，从反方向对上面选择数据库表进行操作，提取相应数据结构和数据内容说明。

示例：考察、比较同类工具后，决定使用**ORACLE**自带的反向工具。

步骤三

依据上面的数据库结构和内容的描述，提取、提炼出数据模式的需求，记录后归档到“数据表说明表格”。参见附录**A.1**数据模式标准的需求收集文档中的第七章数据库说明部分。

数据表结构“**TASKQUEUE**“的提取过程如下：

- 在指定的表上，单击右键，弹出下面的对话框，单击 “显示对象 DDL” 生成表结构的 XML 脚本。
- ORACLE 会自动从反方向提取数据表的存储结构。
- 可以单击“另存为”，导出脚本*.XML。

数据表关联“**TASKQUEUE**“的提取过程如下：

- 在指定的表上，单击右键，弹出下面的对话框，单击 “显示对象相关性”。
- 单击“生成报告”，出现系统对话框。
- 可以选择，将其另存为 HTML 或者文本形式输出。

数据库结构和状态信息“**EXCHGSYS**”的提取过程如下：

- 在指定的数据库上，单击右键，弹出下面的对话框（可以根据需要来定制汇报的内容和形式），单击 “显示对象 DDL” 生成需要的报告。
- ORACLE 会自动从反方向提取数据库的存储结构。
- 可以选择，将其另存为 HTML 或者文本形式输出。

收集分析后形成的字典的内容应当包括：

- 数据库内容的说明。
- 数据库的组成结构说明（包括多少各表，各表之间的联系）。
- 数据表内容的说明。
- 数据表构成结构的说明。

步骤四

反向工程的阶段性成果是，在资料收集工作完成后，最终形成数据模式需求文档。注意在数据模式需求文档中，记录每个数据需求和其引用的权威资料来源。

输出阶段成果是：填写完整的“数据库说明表格”和“数据文件说明表格”。参见附录**A.2.1**数据模式标准的需求收集文档第七章数据库说明和第八章数据文件说明。