



中华人民共和国国家标准

GB/T 19101—2003

建立术语语料库的一般原则与方法

General principles and methods of establishing terminology corpus

2003-05-14 发布

2003-12-01 实施

中华人民共和国
国家质量监督检验检疫总局 发布

目 次

前言	Ⅲ
1 范围	1
2 规范性引用文件	1
3 术语和定义	1
4 基本要求	1
4.1 语料的要求	1
4.2 置标语言的要求	2
4.3 术语语料库系统的要求	2
5 术语语料库的加工和组织	2
5.1 术语语料库的加工层次	2
5.2 术语语料库的加工流程	2
5.3 术语语料库的组织	4
6 术语语料库系统的建立和功能设计	4
6.1 术语语料库系统的建立	4
6.2 术语语料库系统的功能设计	4
6.3 术语语料库系统的服务方式	4
7 术语语料库系统的管理与维护	4
附录 A(资料性附录) 建立术语语料库的有关国家标准	5

前 言

本标准是术语数据库的系列国家标准之一。

已经发布的系列标准有：

GB/T 13726—1992 术语与辞书条目的记录交换用磁带格式

GB/T 16785—1997 术语工作 概念与术语的协调

GB/T 16786—1997 术语工作 计算机应用 数据类目

GB/T 17532—1998 术语工作 计算机应用 词汇

GB/T 18155—2000 术语工作 计算机应用 机器可读术语交换格式(MARTIF) 协商交换

GB/T 13725—2001 建立术语数据库的一般原则与方法

GB/T 15387.1—2001 术语数据库开发文件编制指南

GB/T 15387.2—2001 术语数据库开发指南

GB/T 15625—2001 术语数据库技术评价指南

GB/T 19102—2003 术语部件库的信息描述规范

本标准的附录 A 是资料性附录。

本标准由全国术语标准化技术委员会提出。

本标准由中国标准研究中心归口。

本标准由中国标准研究中心、北京大学计算语言学研究所等单位起草。

本标准主要起草人：陈玉忠、宋敏、何燕、叶盛、穗志方、程永红、肖玉敬。

建立术语语料库的一般原则与方法

1 范围

本标准规定了建立术语语料库的一般原则与方法。

本标准适用于术语语料库的研究、开发、维护及有关管理工作。其他涉及语料库建设的工作也可参照使用。

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。

GB/T 13725—2001 建立术语数据库的一般原则与方法

GB/T 15237.1—2000 术语工作 词汇 第1部分：理论与应用(eqv ISO 1087-1:2000)

3 术语和定义

GB/T 15237.1 确立的术语和定义适用于本标准。为了便于使用，本标准重复了其中的部分术语和定义。

3.1

术语 term

在特定专业领域中一般概念的词语指称。[GB/T 15237.1—2000,3.4.3]

3.2

语料库 corpus

集中起来供分析用的语言数据集合。[GB/T 15237.1—2000,3.6.9]

3.3

术语语料库 terminology corpus

分析和研究术语用的语料库。

3.4

术语语料库系统 terminology corpus system

包含管理框架的术语语料库。

注：改自 GB/T 17532—1998,7.7。

4 基本要求

4.1 语料的要求

4.1.1 一致性

入库语料应是格式一致、有效的。

4.1.2 适用性

语料应从正式出版物或从权威性网站发布的相关专业文献中选择。

4.1.3 忠实性

语料中原来使用的标题、摘要、关键字以及参考文献等各类原始信息和篇章结构应保持完整。