NCIt 数据结构及构建模式分析*

冀玉静

（中国医学科学院医学信息研究所 北京 100020）

〔摘要〕 从树状层级结构、子集、数据来源、术语类型、概念属性、语义关系等方面简要描述 NCI 叙词表（NCI thesaurus, NCIt）的数据结构，继而探讨其构建模式。

〔关键词〕 NCIt；元数据；属性；语义关系；本体；构建模式

Analysis on Data Structure and Constructing Mode of National Cancer Institute Thesaurus (NCIt) *JI Yu-jing, Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing100020, China*

〔**Abstract**〕 The paper briefly introduces the data structure of NCI thesaurus (NCIt) from the aspects of tree structure, sub-set, data sources, term type, concept features, semantic relation and so on, then discusses its constructing mode.

〔**Keywords**〕 NCIt; Metadata; Properties; Semantic relations; Ontology; Construction mode

1 引言

NCI 叙词表（National Cancer Institute thesaurus, NCIt）是一部由美国国立癌症研究所（National Cancer Institute, NCI）编制的参考术语表和生物医学本体，作为一种公认的生物医学编码和参考标准，应用范围越来越广。对其数据结构、构建模式进行深入分析，对于知识组织领域的相关研究具有重要的参考价值。NCIt 的收词范围很广，包括与癌症相关的临床医护、转化研究、基础研究及公共信息和管理活动等。NCIt 提供近 10 000 种癌症和相关疾病、8 000 种单药和联合治疗方案及其他有关癌症的生物医学研究主题的定义、同义词和其他信息，每月更新发布，新增记录大约 900 条。NCIt 具有以下特色：每个生物医学概念都有稳定而唯一的代码；针对每个概念提供优选词、同义词、定义、研究代码、外部源代码和其他信息；提供到 NCI Metathesaurus 和其他信息源的链接；提供 200 000 多条概念间的交叉链接，许多概念还提供正规的基于逻辑的定义；整合了来自 NCI 和其他合作者的广泛内容，其中许多作为单独的 NCIt 子集可供获取；多学科主题专家小组经常对其进行更新维护，数据质量可靠[1,2]。

2 NCIt 数据结构

2.1 树状层级结构

NCIt 概念间通过上下位关系，形成树状层级结构。NCIt 允许同一概念在树状层级结构中处于不同位置。NCIt 树状层级结构的 20 个顶级类目，见图 1。

〔收稿日期〕 2012-03-01

〔作者简介〕 冀玉静，编辑，发表论文数篇。

〔基金项目〕 国家“十二五”科技支撑计划项目“面向外文科技文献的超级科技词表和本体建设”（项目编号：2011BAH10B01）。

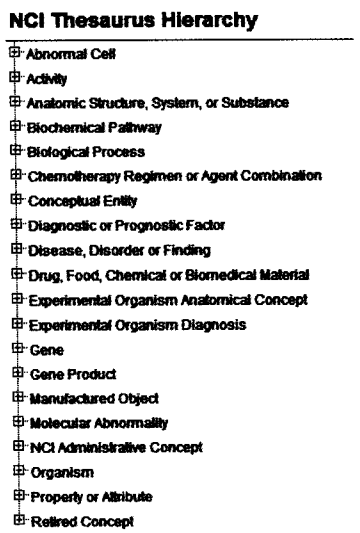

图1 NCIt 树状层级结构一级类目

2.2 NCIt 子集[3]

NCIt 包括 100 多个术语子集，每个子集对应一个 NCIt 概念，定义该子集的实质。这些子集对应的 NCIt 概念都按照等级组织在概念“Terminology Subset (C54443)”之下，见图2，子集概念通过 20 000 多个关系（concept in subset/subset includes concept）和子集所含具体概念建立关联。

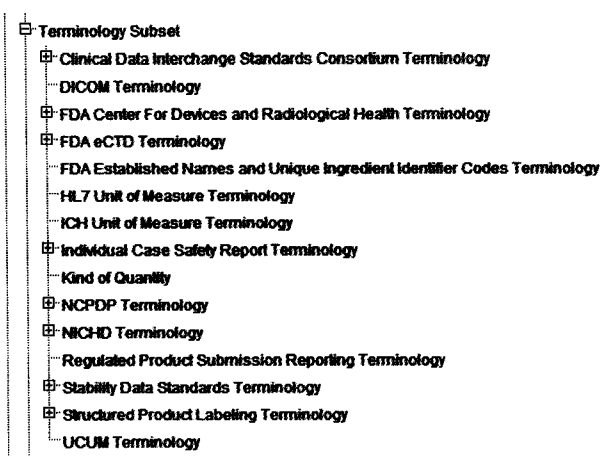

图2 “Terminology Subset”概念下的各个具体子集

这些子集大部分是美国和国际编码标准，由 NCI 和 FDA、CDISC 等其他合作者共同维护。目前这些子集大部分都能以表格形式的文档在 http://evs.nci.nih.gov/ftp1 上分布式下载。以下是其中比较重要的4 个子集集合。（1）临床数据交换标准协会术语（Clinical Data Interchange Standards Consortium Terminology）。（2）美国食品和药品管理术语表（U.S. Food and Drug Administration Terminology）。（3）联邦药物治疗术语表（Federal Medication Terminologies, FMT）。（4）美国国立处方药项目委员会术语表（National Council for Prescription Drug Programs Terminology）。

2.3 NCIt 数据来源[4]

NCIt 中有些特别术语、代码和定义，带有表1 所列的数据来源标签。这些数据来源主要包括以下几类：（1）一些数据来源，如 FDA 和 CDISC 是出于可调和标准化目的而研制多学科标签术语表子集的重要合作者。（2）CTCAE，DCP，DTP 和 NCI - GLOSS 是已经和 NCIt 概念建立全部或部分链接的独立 NCI 术语表。（3）几个外部数据源，如 CRCH 和 JAX 是在特定领域对 NCIt 有贡献的术语表。（4）其他数据来源则以多种更有限的方式提供相关链接和信息。

表1 NCIt 数据源缩写和全称

数据源缩写	数据源全称
BioCarta	BioCarta 在线分子路径图谱，改编版（用于 NCI）
CDC	美国疾病控制和预防中心
CDISC	临床数据交换标准协议
COH	希望城市
CRCH	癌症研究中心夏威夷营养术语表
CTCAE	不良反应通用术语表标准
DCP	NCI 癌症预防项目部
DICOM	医学数字成像交换
DTP	NCI 发展治疗学项目
FDA	美国食品药品管理
ICH	国际协调会议
JAX	Jackson 实验室小鼠术语表，改编版（用于 NCI）
KEGG	KEGG Pathway Database KEGG 路径数据库
NCI	国立癌症研究所叙词表
NCI - GLOSS	NCI 癌症术语词典
NCI - HL7	NCI 卫生标准 7
NICHD	国立儿童健康和人类发展研究所
RENI	注册命名信息系统
UCUM	统一计量单位代码
Zebrafish	斑马鱼模型有机体数据库

2.4 NCIt 术语类型[5]

NCIt 术语类型是用2 或3 个字母缩写来表示的某概念相关术语的种类。表2 是具体术语类型及其含义。

表2 NCIt 术语类型及其含义

术语类型缩写	术语类型含义
AB	缩写
AD	形容词形式（和语法的其他部分）
AQ*	旧版优选词
AQS	旧版术语，用于一个概念有旧版同义词时
BR	美国商标名称
CA2	ISO3166 α-2 位国家代码
CA3	ISO3166 α-3 位国家代码
CNU	ISO3166 数字国家代码
CI	IO 国家代码
CN	药品研究代码
CS	美国国务院国家代码
DN	显示名称
FB	外国商标名
HD*	标题（概念组，不用于编码数据）
PT*	优选词
SN	化学结构名称
SY	同义词

*注：关于 NCIt 术语类型 PT、HD 和 AQ 的特殊规则：每个概念之下都有一个并且只有一个术语被赋予这 3 种类型之一。NCI 通常是从多个 NCI 术语中选择一个作为优选词，赋予术语类型 PT。但是某些特殊情况下，一个概念没有优选词 PT，取而代之的是标题词 HD 或旧版优选词 AQ。HT 或 AQ 被计算机软件认为等同于 PT，这就意味着一个概念只能有一个术语的类型是 PT、HT 或 AQ 中的一种。当一个旧版概念有多个旧版术语时，这个旧版概念的类型就是旧版优选词 AQ，其余为旧版术语 AQS。

2.5 NCIt 属性（properties）

NCIt 作为一个叙词表，同时还具有某些类似于本体的特点。关于某个概念的大部分信息都存储在属性（Properties）中。NCIt 共有 74 个属性，表 3 选取其中有代表性的一部分属性进行展示。

2.6 NCIt 语义关系[6]

2.6.1 Roles Roles 是 NCIt 概念对间的双向关系，共 114 种，详见 http://evs.nci.nih.gov/ftp1/ThesaurusSemantics/Roles.xls。Roles 的特点，也是区别于 Associations 之处在于其可继承性，即上位概念如果有某种 Roles 关系，则其下位概念也可有这种 Roles 关系。并非所有 114 种 Roles 关系都有实例，有的关系虽然存在，但目前还没有用于限制任何概念。每种 Roles 关系都对应一个领域（Domain）和一个范围（Range）。领域指的是当前概念所属的种类（Kind），范围则是与当前概念有关系的另一个概念所属的种类。每个概念都属于并且只能属于一个种类，不同的概念种类互不交叉，是排他的。NCIt 部分关系名称及对应的领域和范围，见图 3，NCIt 涉及的概念种类及其含义，见图 4。

表3 NCIt 属性及其含义（节选）

属性名称	属性含义
Preferred_ Name	NCI 优先使用的某概念的单词或词组
Synonym	NCI 概念优选名称的有效替代形式
Definition	NCI 给出的某概念的英文定义。长度限制为 1 024 个字符。可包含该定义的来源或归因
Neoplastic_ Status	包含与单个癌症概念癌症状态有关的信息：恶性、良性、癌前病变、未知恶性可能和不确定
Chemical_ Formula	给出构成某种化合物的简明原子表达式
Accepted_ Therapeutic_ Use_ For	指明化学物或药物公认用于治疗某种疾病或状况
In_ Clinical_ Trial_ For	指明一种药当前正处于治疗某种疾病的临床试验阶段
PDQ_ Closed_ Trial_ Search_ ID	NCI 网站用于检索涉及特定临床试验药物的 PDQ 封闭临床试验数据的标识符
Related_ MedDRA_ Code	编码 CTEP 临床试验中用到的 MedDRA/CTEP 疾病名称最近似匹配的 MedDRA 代码
Essential_ Amino_ Acid	说明某种氨基酸是必需的，在膳食中必须包括
Micronutrient	指明人体健康所需的一类食品组分，特点是消耗毫克或微克
Tolerable_ Level	指明美国国家科学院食品营养部制订的不会对对绝大多数美国人的健康不会产生不利影响的食物组分每日最高摄入量
Unit	一种营养素或食品组分常用的计量单位，不包括克、毫克、微克焦耳、千焦

	Roles (restriction relationships) declared in the NCI Thesaurus, as of August 21, 2008		
2	Role Name	Domain	Range
3	Allele_Absent_From_Wild-type_Chromosomal_Location	Gene_Kind	Anatomy_Kind
4	Allele_Associated_With_Disease	Gene_Kind	Findings_and_Disorders_Kind
5	Allele_Causes_Function_In_Pathway	Gene_Kind	Pathway_Kind
6	Allele_Has_Abnormality	Gene_Kind	Molecular_Abnormality_Kind
7	Allele_Has_Activity	Gene_Kind	Properties_or_Attributes_Kind
8	Allele_In_Chromosomal_Location	Gene_Kind	Anatomy_Kind
9	Allele_Is_Cancer_Related_Type	Gene_Kind	Gene_Kind
10	Allele_Not_Associated_With_Abnormality	Gene_Kind	Molecular_Abnormality_Kind
11	Allele_Not_Associated_With_Disease	Gene_Kind	Findings_and_Disorders_Kind
12	Allele_Plays_Altered_Role_In_Process	Gene_Kind	Biological_Process_Kind
13	Allele_Plays_Role_In_Metabolism_Of_Chemical_Or_Drug	Gene_Kind	Chemicals_and_Drugs_Kind
14	Anatomic_Structure_Has_Location	Anatomy_Kind	Anatomy_Kind
15	Anatomic_Structure_Is_Physical_Part_of	Anatomy_Kind	Anatomy_Kind
16	Biological_Process_Has_Associated_Location	Biological_Process_Kind	Anatomy_Kind
17	Biological_Process_Has_Initiator_Chemical_or_Drug	Biological_Process_Kind	Chemicals_and_Drugs_Kind
18	Biological_Process_Has_Initiator_Process	Biological_Process_Kind	Biological_Process_Kind
19	Biological_Process_Has_Result_Anatomy	Biological_Process_Kind	Anatomy_Kind
20	Biological_Process_Has_Result_Biological_Process	Biological_Process_Kind	Biological_Process_Kind
21	Biological_Process_Has_Result_Chemical_or_Drug	Biological_Process_Kind	Chemicals_and_Drugs_Kind
22	Biological_Process_Is_Part_of_Process	Biological_Process_Kind	Biological_Process_Kind
23	Chemical_Or_Drug_Affects_Abnormal_Cell	Chemicals_and_Drugs_Kind	Abnormal_Cell_Kind
24	Chemical_or_Drug_Affects_Cell_Type_or_Tissue	Chemicals_and_Drugs_Kind	Anatomy_Kind
25	Chemical_Or_Drug_Affects_Gene_Product	Chemicals_and_Drugs_Kind	Gene_Product_Kind

图3 NCIt 语义关系 Roles 的名称及其对应的领域（Domain）和范围（Range）（节选）

Kind	Description of Kind's Coverage
Gene	Any definable DNA sequence capable of being transcribed and having biological significance. "
Gene_Products	Endogenous RNAs, proteins, protein complexes and riboprotein complexes. Excludes exogenous chemicals.
NCI	Conceptual entities required by NCI operations and systems. Includes administrative, financial, organizational and quasi-scientific concepts .
Findings and Disorders	Classification of human conditions that are relevant to cancer. Includes observations, test results, history and other concepts relevant to characterization of human cancer-related conditions. Includes non-neoplastic conditions of special interest
Anatomy	Naturally occurring human biological structures, fluids, and substances. Includes embryonic, gross and micro anatomic structures and surgically created structures, including cellular *organelles but excluding single molecules.*
EO_Anatomy	Naturally occurring non-human biological structures. Includes embryological, gross and micro anatomic structures in all species used as models of human cancer. Excludes structures smaller than can be visualized by light microscopy.
EO_Findings_and_Disorders	Classification of non-human conditions that are relevant to cancer. Includes observations, test results, history and other concepts relevant to characterization of cancer-related conditions in species used as models of human cancer.
Abnormal_Cell	An enumeration of abnormal cell types that occur in human cells and in cells of experimental models of human cancer.

图4 NCIt 概念的种类（Kind）及含义（节选）

2.6.2 Associations Associations 也是 NCIt 概念间的双向关系，特点是不可继承，共9 种，详见表4。语义上仅仅表示特定两个概念间的关系，这种关系可能并不适用于这两个概念的下位概念，因此不能继承。

3 NCIt 的构建模式[6]

3.1 构建目的和程序

NCIt 作为一个受控的叙词表，所有术语按照已知顺序和结构组织在一起，术语间的不同关系得以清晰呈现和说明，同时具有某些类似于本体的特点。其构建目的是在与癌症相关的不同专业学科和数据来源之间搭建一个可供沟通的渠道。NCIt 是利用 Apelon 公司的 Ontylog 语言构建的，Ontylog 语言是描述逻辑（Description Logic，DL）的一种，用于构建和维护大规模知识库。描述逻辑一直以来在复杂学科领域的知识组织及词表和分类表构建方面颇有建树[7]。Ontylog 在构建生物医学术语表方面应用广泛，除了 NCIt 外，SNOMED/RT，SNOMED/CT 也是基于 Ontylog 构建的。

表4 NCIt 语义关系 Associations 的名称及含义

名称	含义
Concept_ In_ Subset	揭示某术语子集所含概念与该术语子集概念间的包含关系
Has_ CDRH_ Parent	指出在 CDRH 中的上位概念
Has_ Free_ Acid_ Or_ Base_ Form	说明一个药物自由酸或碱形式
Has_ Salt_ Form	说明一个药物有不同盐形式
Has_ Target	揭示一个药物或其他物质具有临床上有意义的分子靶标。靶标可以是一个基因、基因产物、解剖结构、生物学过程或其他概念。药物对靶标的作用应和一种疾病过程具有治疗、诊断或其他临床相关性。这种作用最常见的是直接的、分子水平上的
Is_ Related_ To_ Endogenous_ Product	揭示治疗性部分和内源性基因产物相关
Role_ Has_ Domain	指明关系 role 对应的领域
Role_ Has_ Parent	指明关系 role 在关系层级表中的上位关系
Role_ Has_ Range	指明关系 role 对应的范围

3.2 概念为基础

在 Ontylog 语言中，以概念“Concept”为基础，每个概念都是抽象的类，如果只有一个特例，就不会生成相应的概念。基于 Ontylog 构建的 NCIt 也是以概念“Concept”为基础，每个概念或者是原始的，或者是经过定义的。原始概念是通过“定义超级概念”人工确定的，虽然只有基本的描述，但却是整个词表的基础；而定义概念，除了具有基本的描述，还有更详细更完整的描述，则是通过“直接上位概念”在聚类过程中基于算法自动生成的。最初 NCIt 中所有概念都是原始的，随着该叙词表的发展成熟，定义概

念的比例不断增加，但是其主体仍然是原始概念。此外顶级概念用于推测定义概念的含义，因此会始终保持原始状态，从而确保 NCIt 树状层级结构的顶级类目能够很好地构建。NCIt 关于某个概念的大部分信息都以字符串的形式存储在多个不同的属性（Properties）中，属性在一定程度上类似于通常所说的字段。

许多描述逻辑（DL）都对类概念（用于描述集合）和个例概念（用于描述集合中的个例）加以区分，将整个词表分成 T 盒（T - box，都是类概念）和 A 盒（A - box，都是个例概念）。但是基于 Ontylog 语言的 NCIt 并不支持个例概念，原因在于：首先，Ontylog 有一个强制语义学算法，确保叙词表不存在特例情况，因而不需要个例概念与之对应；其次，NCIt 主要用于实时系统以支持基础、转化和临床研究，这些实时系统中的实例是一些有特色的实验数据或研究主题记录，并存储在数据库中，而处理语义学和其他核心系统关系才是最重要的，同时对个例概念的推理类型用于推理类概念更简便，因此没有必要再去推理确定个例概念。

3.3 概念聚类

依照描述逻辑（DL）的惯例，概念聚类是通过在以每个概念为节点的非循环结构图中进行的包含测试完成的。在非循环结构图中，每个概念看作一个节点，节点间连线就是概念间的语义关系。每个概念代表一个语义单元。所有概念构成一个树状层级表，并在概念聚类过程中不断校验、调整，同一概念可以在树状层级结构中处于不同位置。Ontylog 语言的简明语义学算法符号，见图 5。NCIt 构建过程中用到了 Ontylog 中除了“modal restriction”和“right identity”之外的所有语义算法。

3.4 版本转换

因为 Ontylog 不能与很多免费开源软件兼容，因此将 NCIt 的 Ontylog 版本转成 OWL 版本可供免费下载。OWL（Web Ontology Language）是一种用于在语义 Web 上发布和共享本体的语义置标语言，它代表了面向 Web 的本体表示语言的最新发展趋势。它面向 Web，相对于 XML，RDF 和 RDF Schema 拥有更多的机制来表达语义，而又与它们兼容。OWL 能够被用来清晰地表达词表中的词汇含义以及这些词汇之间的关系，并具备良好的扩展性[8]。

Ontylog Language[1]

Constructor	*Syntax*	*Semantics*
Concept name	C	C^I (where $C^I \subseteq \Delta^I$)
Top	$\top$	Δ^I
Bottom	$\bot$	$\varnothing$
Conjunction	$C \sqcap D$	$C^I \cap D^I$
Disjunction	$C \sqcup D$	$C^I \cup D^I$
Universal restriction	$\forall R.C$	$\{x \mid \forall y : R^I(x,y) \rightarrow C^I(y)\}$
Existential restriction	$\exists R.C$	$\{x \mid \exists y : R^I(x,y) \wedge C^I(y)\}$
Modal restriction	$\Diamond R.C$	$\{x \mid Pr(\exists y : R^I(x,y) \wedge C^I(y)) > 0\}$
Role name	R	R^I (where $R^I \subseteq \Delta^I \times \Delta^I$)

Definitional or Axiomatic Constraint	*Syntax*	*Semantic Constraint*
Concept definition	$C \doteq D$	$C^I \equiv D^I$
Concept subsumption axiom	$C \sqsubseteq D$	$C^I \subseteq D^I$
Role subsumption axiom	$R \sqsubseteq S$	$R^I \subseteq S^I$
Right identity axiom	$R \circ S \doteq R$	$(R \circ S)^I \equiv R^I$

[1] *Note that disjunction, a feature currently under development, can only be used in role values*

图 5 Ontylog 的语义符号

4 结语

NCIt 作为一种重要的生物医学叙词表和本体，分析其数据结构、探讨其构建模式的特点，可以为知识组织领域的相关研究提供重要参考。

参考文献

1 http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources [EB/OL]. [2012 - 01 - 15].

2 http://ncit.nci.nih.gov/ [EB/OL]. [2012 - 01 - 15].

3 http://ncit.nci.nih.gov/ncitbrowser/pages/subset.jsf [EB/OL]. [2012 - 01 - 15].

4 http://ncit.nci.nih.gov/ncitbrowser/pages/source_help_info.jsf? dictionary = NCI Thesaurus [EB/OL]. [2012 - 01 - 15].

5 http://ncit.nci.nih.gov/ncitbrowser/pages/term_type_help_info.jsf? dictionary = NCI Thesaurus [EB/OL]. [2012 - 01 - 15].

6 http://evs.nci.nih.gov/ftp1/ThesaurusSemantics/NCI%20Thesaurus%20Semantics.pdf [EB/OL]. [2012 - 01 - 15].

7 https://github.com/bdionne/bitstore [EB/OL]. [2012 - 01 - 15].

8 曾新红. 中文叙词表本体——叙词表与本体的融合 [J]. 现代图书情报技术，2009，(1)：36.

NCIt数据结构及构建模式分析

作者： 冀玉静
作者单位： 中国医学科学院医学信息研究所,北京,100200
刊名： 医学信息学杂志
英文刊名： Journal of Medical Informatics
年，卷(期)： 2012,33(6)

参考文献(8条)

1. 查看详情 2012

2. 查看详情 2012

3. 查看详情 2012

4. 查看详情 2012

5. 查看详情 2012

6. 查看详情 2012

7. 查看详情 2012

8. 曾新红 中文叙词表本体--叙词表与本体的融合 2009(01)