为了治愈，我们选择开放和共享

本体医学术语标准构建

浙江数字医疗卫生技术研究院OMAHA秘书处
中国卫生信息学会健康医疗大数据标准研究院

2019年8月

开放医疗与健康联盟

Open Medical and Healthcare Alliance

NGO 自组织 开放 开源 医疗 健康

Contents

Contents

随着信息的网络化发展，健康医疗信息资源迅猛增加，各种数据分散在不同机构、系统和应用中

医学术语标准是实现不同信息系统间语义互操作的基础支撑，是医疗信息化发展的重要影响因素

HIMSS将互操作能力*程度分为4个等级

*互操作能力：不同信息系统和软件应用之间的通讯能力、数据交换能力、信息使用能力

基础级别 (Foundational)

- 能够满足一个信息系统到另一个系统的数据传输，但对数据的解释能力并不作要求，计算机不关心传输什么，也不处理格式、不解读内容，是单纯的技术上的互操作。

结构级别 (Structural)

- 定义了数据交换的语法规范，确保数据交换的过程中系统可以在字段层面正确理解数据，实现语法互操作。

依赖术语标准

语义级别 (Semantic)

- 可以满足两个或多个系统之间交换信息以及使用这些交换后的信息。语义级别的互操作性同时具有结构级别互操作性和对数据编码的能力，从而实现系统对数据的解读。

组织级别 (Organizational)

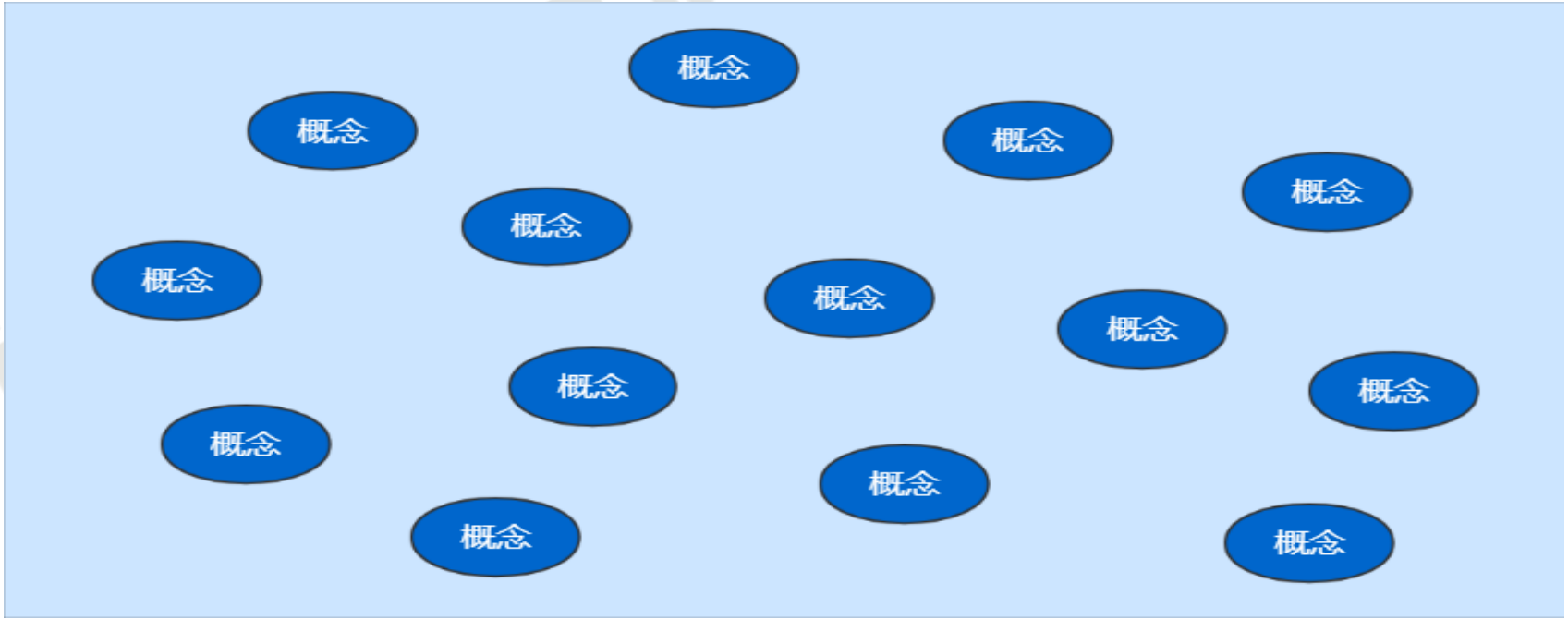
- 包括技术组件以及明确的政策、社会和组织组件。这些组件有助于组织和个人之间安全、无缝和及时地通信和使用数据。

https://www.himss.org/library/interoperability-standards/what-is

医学术语标准按照语义关系强弱、结构化程度和语言受控程度可以分为词表类、分类聚类和关联组类三种类型

Gall H. Systems of knowledge organization for digiM libraries：beyond traditional authority files卟Wash—ington，American：The Digital Library Federation，2003：4-7.

词表类的术语标准强调概念的定义和术语表达，不涉及概念之间的语义关系



1
2
3

词表类术语标准最常见，以各种词典、规范文档为主

卫生领域的词汇表类术语体系

规范和统一医学名词，减少或消除社会上名词的混乱使用。

词汇表&字典

- 全国自然科学名词审定委员会：《医学名词》
- 人民卫生出版社：《汉英医学大辞典》
- 国家药典委员会：《中国药典》

规范文档

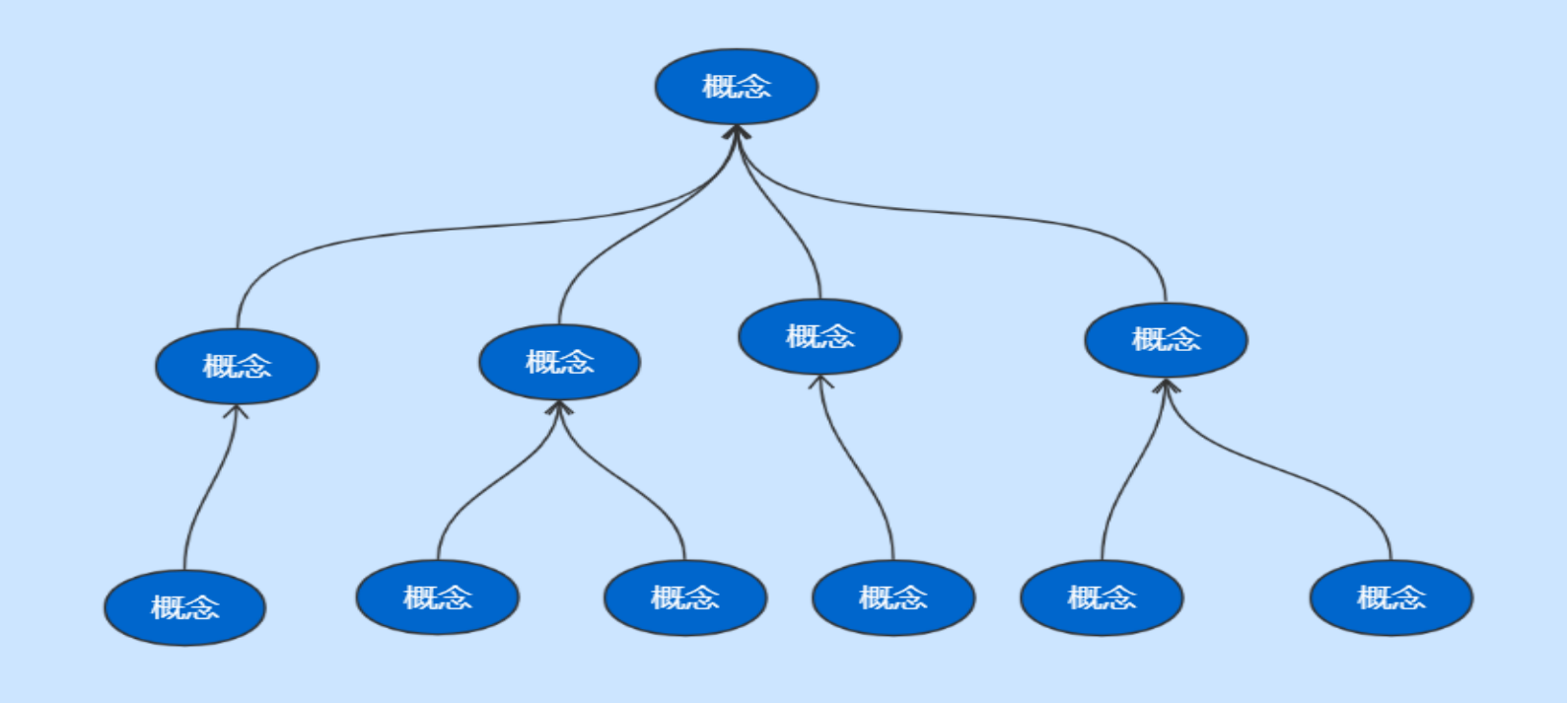
- 卫健委：《WS/T 203 - 2001输血医学常用术语》
- 卫健委：《WS 375-2012 疾病控制基本数据集》

临床医学指南

行业应用

- 制作成医学专业词典或工具，用于查阅和学习
- 作为垂直领域的语料库

分类聚类体系医学术语标准强调概念之间的层级、类别关系，不包更复杂的关系，且分类一般具有排他性，缺乏灵活性



分类聚类体系术语标准目前在我国医疗信息化行业中最常用，例如用ICD-10解决疾病诊断的标准化，但有局限性

分类体系术语集举例：ICD-10

标准版本

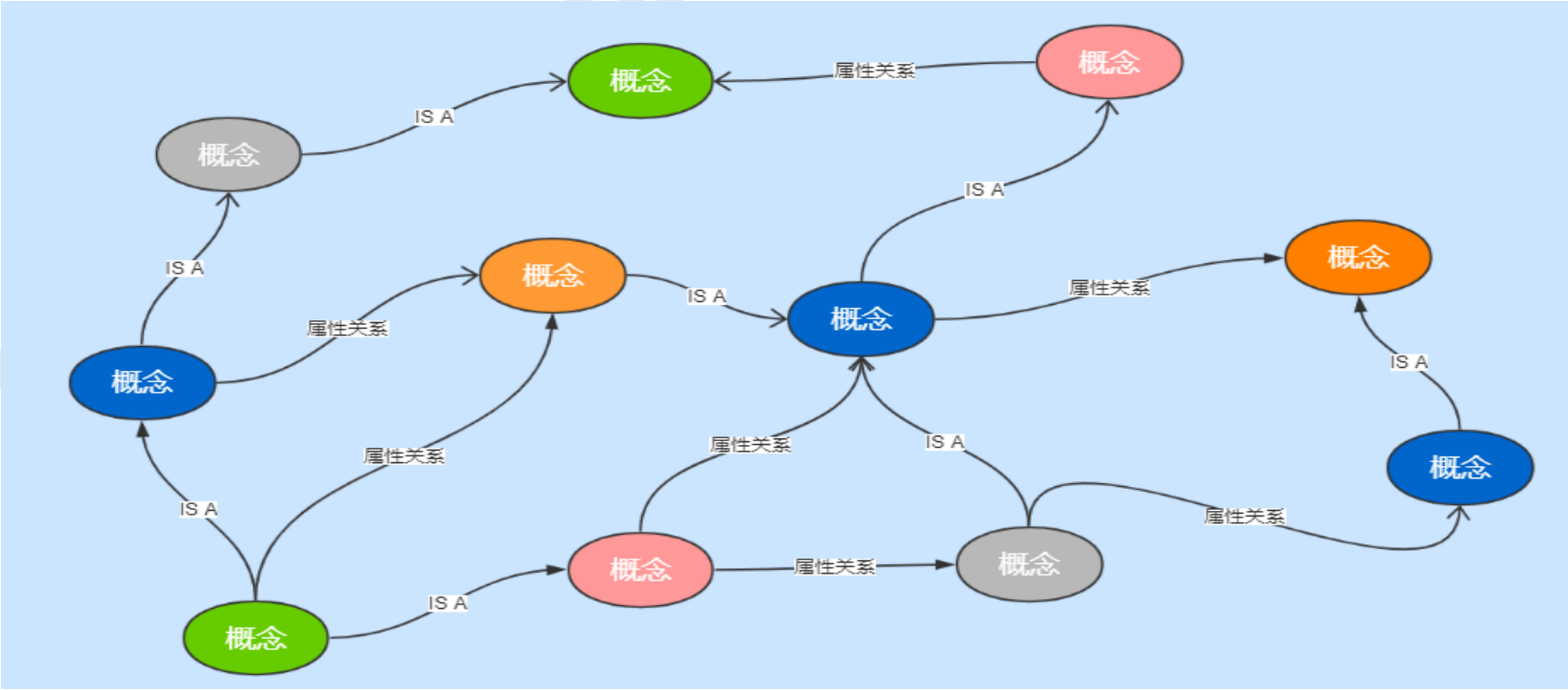
- **全国**：GB/T 14396-2016疾病分类与代码
- **全国**：《疾病分类与代码（修订版）》国家临床版2.0
- **全国**：《疾病分类与代码（修订版）》全国2011版
- **北京**：住院病案首页疾病诊断名称与代码标准V6.01版
- **上海**：2013上海卫生局_病案首页_疾病分类与代码_ICD-10_更新版
- …

行业应用

非标准版本

- HIT厂商系统中的字典，作为各医疗机构的内码库
  - 临床科室版本：书写病历用
  - 病案室版本：满足卫生统计上报用
  - 医保办版本：满足医疗保险报销用
- 应用于科研、管理等

关联组类术语标准强调概念的表达、概念间各种复杂关系的揭示，语义关系强，将知识组织成网状结构



关联组类医学术语标准不仅能够高效解决不同系统间语义互操作问题，还可以赋能大数据分析，医疗AI等新技术

关联组类术语体系举例

国外

医学系统命名法-临床术语（SNOMED-CT）、解剖学本体（FMA）、基因本体（GO）、NCI叙词表、医学主题词表（MeSH）、药物标准术语表（RxNorm）、生物医学顶层本体（BFO）、一体化语言系统（UMLS）、ICD-11等

国内

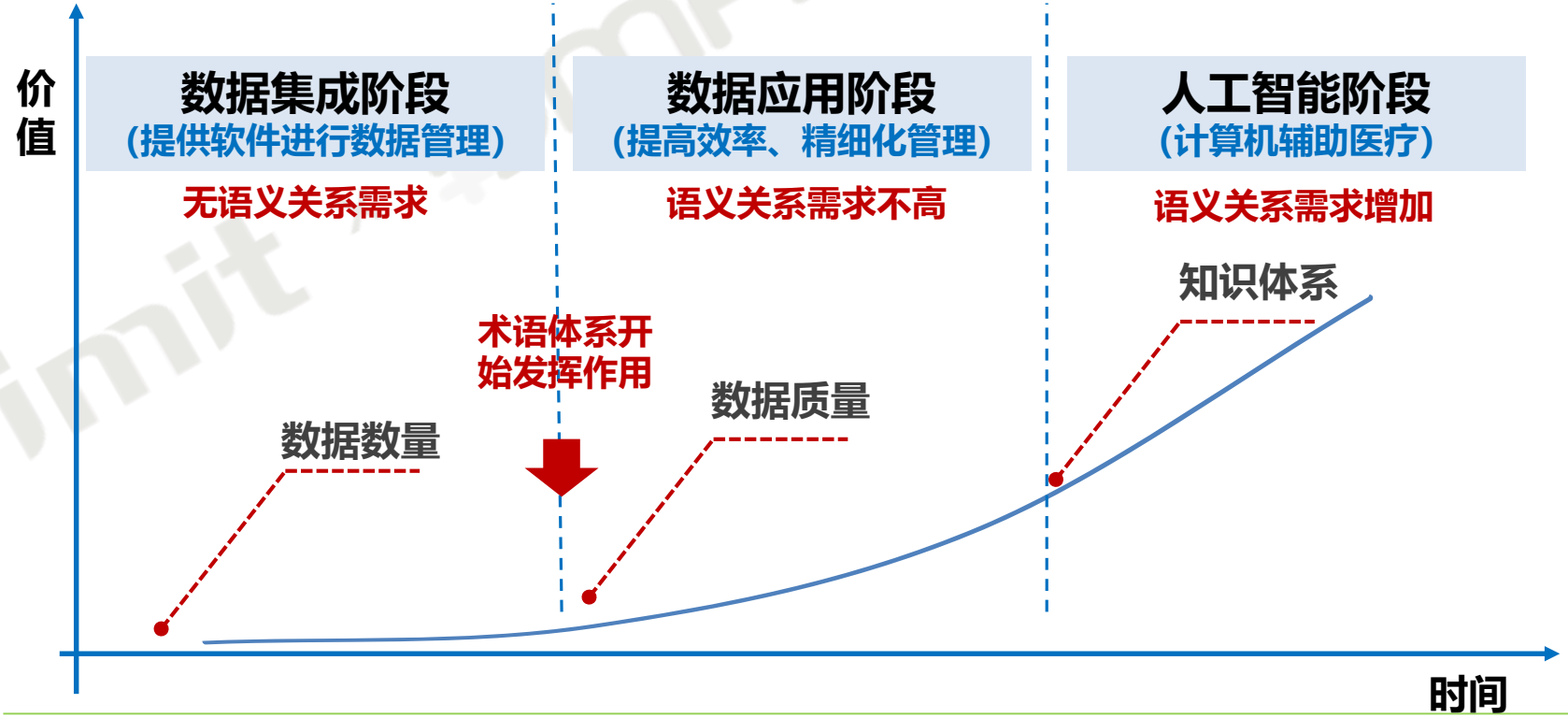
- 中文版医学主题词表（CMeSH）
- 中文一体化医学语言系统（CUMLS）
- 中国中医药学主题词表
- 中医药一体化语言系统（TCMLS）
- OMAHA七巧板医学术语集

行业应用

- 医疗数据互操作
- 电子病历结构化处理，协助医嘱录入
- 基于语义的信息检索
- 数据挖掘、统计分析
- 知识库构建，临床决策支持

Contents

医疗AI时代对“让计算机理解人类语言”提出需求，具有强语义关系的关联组类医学术语标准正在成为必需品



以“本体论”建立的医学术语标准，已被充分应用于知识工程、人工智能、语义网等相关领域

本体：是共享概念模型的明确的形式化的规范说明（Studer, 1998）。本体因其可以使人与人、人与机器、机器与机器的交流建立在对领域知识共识基础上，是当前国际上主流的语义知识表示模型。

本体的优势

1. 本体具有良好的概念关系、层次结构和对逻辑推理的支持能力。
2. 本体的概念模型是对领域中公认知识的逻辑抽象，是领域约定的“范式”，可以很好地共享和重用，避免重复领域知识的分析。
3. 本体统一的概念减少了歧义，可以使不同的知识组织体系、软件之间进行映射，以实现不同信息系统间的语义互操作。
4. 本体作为语义网（Semantic Web）技术栈的第四层，是语义网的核心，提供了资源的语义模型，实现了资源语义的表述，解决资源的语义异构，实现网上信息资源在语义层上的全方位的互联，从而可以对异构、分布的网络信息进行有效检索和访问。

直接本地化国外本体术语集有局限性，构建具有自主知识产权且适用中文临床环境的本体医学术语集才是正确路径

以SNOMED CT为例：

本地化SNOMED CT的局限性

术语本地化后的冗余

由于中英文表达的差异，术语不适用于中文语境

- 如SCT“头痛”概念的的7种描述（Headache, Cephalalgia, Cephalgia, Cephalodynia, HA-headache, Head pain, Pain in head）本地化后只得到1个有效中文描述

概念不适用

很多概念在中国不适用

- 如解剖结构采用“SEP”理论模型搭建，头部这一解剖部位有三个概念： Head structure（头部）、Entire head（整个头部）、Head part（部分头部），然而中国解剖学并非采用这个理论架构，中文临床环境下，头部已经包含了整个头部的含义，即“整个头部”的概念不适用

概念缺失

中国很多常用的概念，在SCT里面找不到

- 如“高血压1级”、“高血压2级”、“高血压3级”等中文常用的医学概念缺失

费用、产权和管理

国家层面的顾虑

- 费用问题：几千万美金/年。
- 产权问题

管理团队

- 翻译团队的能力
- 持续维护更新
- 经费来源问题
- 不同的术语集需要有不同的管理团队

Contents

设计本体术语模型，收集和整合现行医学术语标准和真实语料，持续维护更新是形成中文本体医学术语标准的合理规划

数研院医学术语标准的构建规划

初期阶段

- 学习国外成熟术语体系的结构和顶层设计
- 收集已有中文医学词汇资料
- 收集已有行业术语标准

形成阶段

- 设计符合中文临床环境的本体术语模型
- 对中文医学术语进行整合梳理
- 采用众包/协作方式标记语义类型、同义关系、上下位关系、属性关系等
- 建立维护更新机制

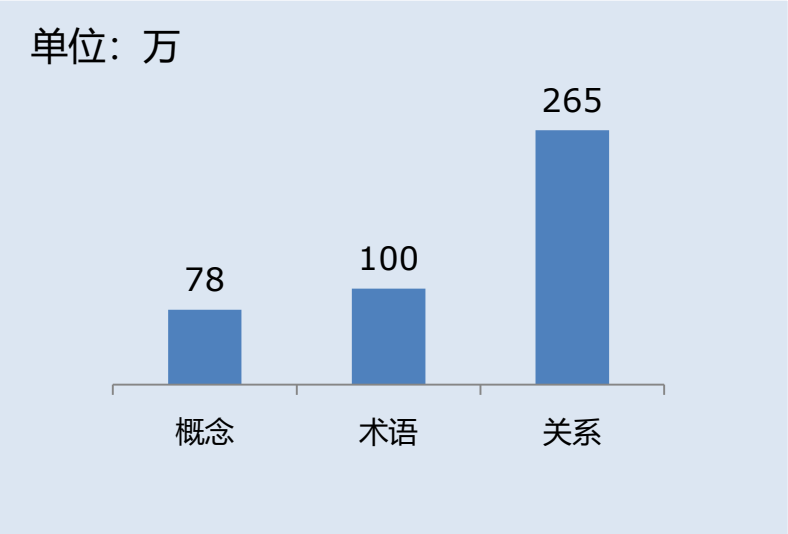
维护阶段

- 持续更新和维护，并定期进行发布
- 保证术语集的正确性、临床适用性
- 不断完善维护更新机制

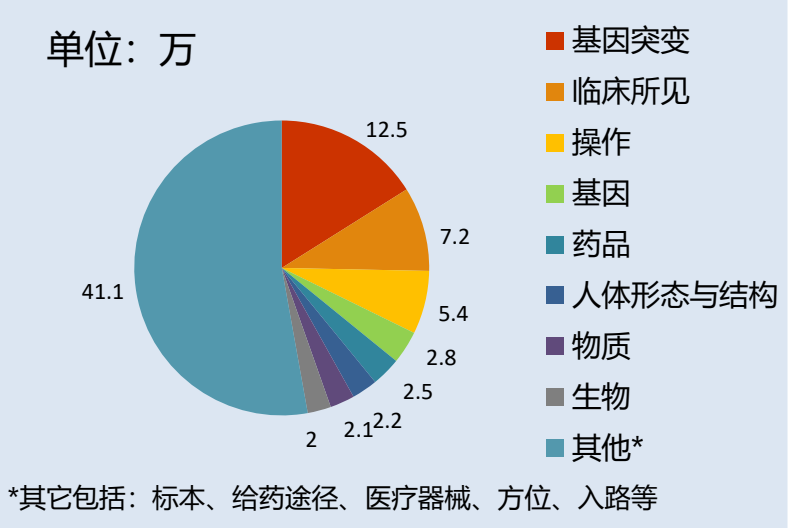
目前进度

术语集目前收录78万概念，对应100万术语和265万关系，主要有疾病、症状体征、手术操作、药品、基因等内容

概念、术语和关系的数量



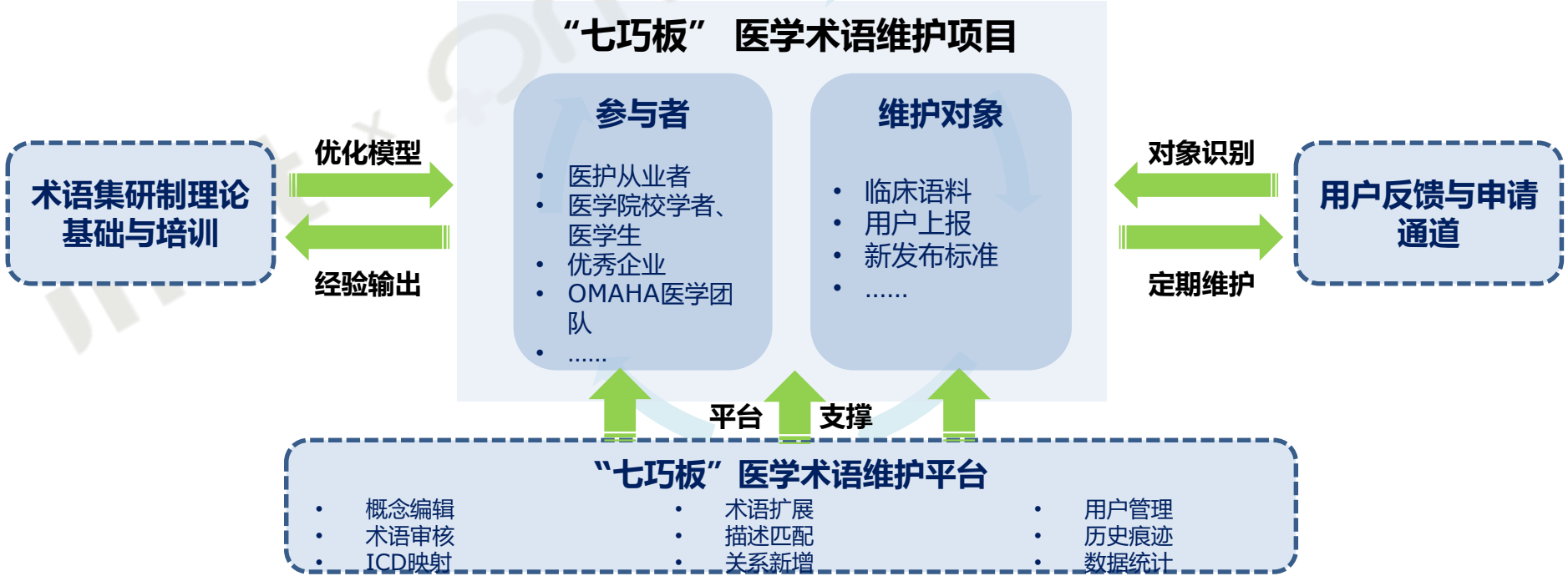
概念的语义类型分布



*其它包括：标本、给药途径、医疗器械、方位、入路等

行业内相关从业者充分参与协作，共同开展术语集构建，参与者即是使用者，形成良性闭环

持续动态维护术语集，按季度发布更新版本



Contents

- 医学术语标准所解决的问题
- 行业需求现状和应对方法
- 数研院医学术语标准研制进展
- **基于本体的医学术语集构建**
- 医学人工智能应用研究配套术语集编制流程

构建能够适用于医疗人工智能应用的本体医学术语集，离不开模型设计、人才、规则、维护更新机制等关键支撑

构建中文医学本体术语标准的关键成功因素

1. 术语模型设计
2. 人才配备
3. 编制规则
4. 维护更新机制

构建能够适用于医疗人工智能应用的本体医学术语集，离不开模型设计、人才、规则、维护更新机制等关键支撑

构建中文医学本体术语标准的关键成功因素

1. 术语模型设计
2. 人才配备
3. 编制规则
4. 维护更新机制

斯坦福大学Noy和McGuinness提出领域本体构建七步法是相对成熟且易操作的本体构建方法

本体构建七步法

1. 确定本体的专业领域
2. 考虑已有的本体重用
3. 理出本体中重要的术语
4. 定义概念以及概念层级结构
5. 描述定义概念的属性
6. 定义属性的分面
7. 依据概念来创建实例

从实际使用需求出发确定本体范围，不要盲目建立。

建议从已有行业术语标准、教科书、指南、期刊、专家共识、临床语料等出发整理。

建议在充分学习已有本体术语体系基础上结合实际需求进行设计。

本体建模元语包括概念（类），关系，函数，公理和实例等5种，实际建立过程中可以按需选择建立

√ **类（classes）或概念(concepts)**

- 指任何事务，如工作描述、功能、行为、策略和推理过程。从语义上讲，它表示的是对象的集合，其定义一般采用框架（ frame） 结构，包括概念的名称，与其他概念之间的关系的集合，以及用自然语言对概念的描述。

√ **关系（relations）**

- 在领域中概念之间的交互作用。从语义上讲，基本的关系共有4种：part of, kind of, instance of, attribute of，在实际建模过程中，概念之间的关系可以根据领域的具体情况再增加相应的关系。

函数（functions）

- 一类特殊的关系。

√ **公理（axioms）**

- 表永真断言，如概念乙属于概念甲的范围。

√ **实例（instances）**

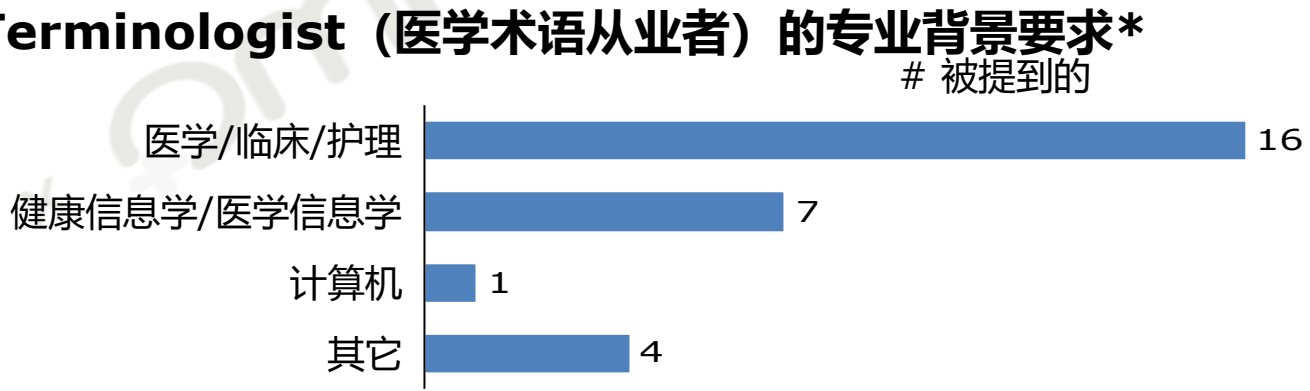
- 代表元素，从语义上讲实例表示的就是对象。

构建能够适用于医疗人工智能应用的本体医学术语集，离不开模型设计、人才、规则、维护更新机制等关键支撑

构建中文医学本体术语标准的关键成功因素

1. 术语模型设计
2. 人才配备
3. 编制规则
4. 维护更新机制

国外医学术语从业者通常来自于临床或医学相关专业，或是具有医学信息学的背景



主要发现：

- 医学术语从业者主要由医学或临床相关、医学信息学两种类型的人员背景构成
- 其中，医学或临床专业是这个岗位的首选，医学信息或健康信息等跨界专业也是此岗位目标群体
- 国外医学信息学相关专业一般在研究生阶段设置，因此，这类人才已经具体了临床或者计算机的专业背景，而跨领域的专业人才正是国内最缺乏的部分

Note: 表单参考了领英招聘网站中，10国外家企业发布的11个相关的岗位的招聘信息，表单中是招聘信息中岗位需求被提及到有关“专业背景”信息。

构建能够适用于医疗人工智能应用的本体医学术语集，离不开模型设计、人才、编制规则、维护更新机制等关键支撑

构建中文医学本体术语标准的关键成功因素

1. 术语模型设计
2. 人才配备
3. 编制规则
4. 维护更新机制

编制规则指导术语集编制的全过程，是术语集编制的核心工作指南

一般编制规则

理解性
- 必须能够被一般的医疗工作者理解并用于交流，不涉及到难以理解的、隐蔽的或私人的知识。

适用性
- 对于一般的临床工作者具有使用意义。

复用性
- 必须是可以被多人共同理解并以相同的方式进行重复使用。

核心组件编制规则

概念
- 单义性、唯一性
- 新增、失活、复活、层级调整等规则

术语
- 术语类型确定，如首选术语、许用术语
- 英文缩写添加、人名命名术语添加、标点符号限制等规则
- 新增、失活、复活等规则

关系
- 除了顶层概念外，每个概念至少有一条层级关系，层级关系不能成环
- 属性关系使用的限制
- 新增、失活、复活等规则

构建能够适用于医疗人工智能应用的本体医学术语集，离不开模型设计、人才、规则、维护更新机制等关键支撑

构建中文医学本体术语标准的关键成功因素

1. 术语模型设计
2. 人才配备
3. 编制规则
4. 维护更新机制

对编制形成的医学术语集持续进行维护更新，是保证术语集正确性、临床适用性的关键

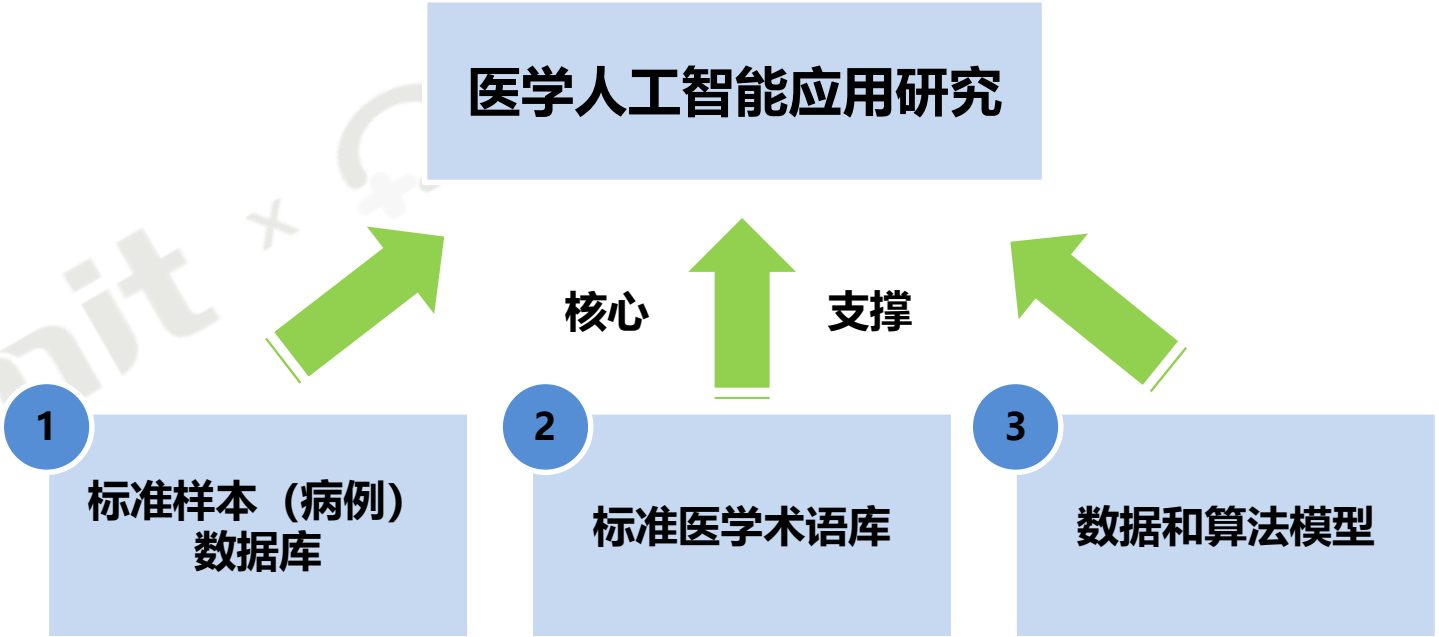
合理的维护更新机制主要内容

专人负责	• 由专门的人负责统筹管理维护更新的所有工作。
更新频率	• 合理的更新频率，如每季度/每半年至少更新一次，并及时发布更新后成果。
流程清晰	• 发现错误或需要新增、优化，由谁具体执行，增、删、改的操作步骤，由谁最终审核，发布等所有工作必须制定清晰的工作流程和管理规范。
逻辑验证	• 建立自动的逻辑验证规则，借助机器自动发现可能存在的错误，如概念下没有术语，同一术语在多个概念下存在，层级关系成环，关系多余等
模型优化	• 术语模型不是一成不变的，需要根据实际使用需求和应用反馈不断进行优化调整。
反馈通道	• 建立用户反馈通道，从实际应用中发现术语集缺陷或错误，及时增补或优化。

Contents

- 医学术语标准所解决的问题
- 行业需求现状和应对方法
- 数研院医学术语标准研制进展
- 基于本体的医学术语集构建
- **医学人工智能应用研究配套术语集编制流程**

医学人工智能应用研究开展过程中，本体医学术语集的编制是一项基础且非常重要的工作



术语集编制工作需要大家通力配合，遵循统一的编制要求和审定要求，才能保证编制成果质量

编制阶段	工作内容	负责团队
阶段一：准备	1. 完成编制文档，内容包括研制目的、研制规则、覆盖范围、构建方法、术语框架体系、术语构建流程、术语维护流程、与现有术语系统（含国际主流术语体系）的衔接和兼容等。 2. 团队组建。 3. 术语集编制平台搭建。	术语集编制单位
	对编制文档进行审定。	浙江数字医疗卫生技术研究院
阶段二：编制	进行术语集编制。	术语集编制单位
	对编制形成的术语集进行审定，审定内容包括术语范围、术语方法、术语内容、术语来源等。	浙江数字医疗卫生技术研究院
阶段三：发布、使用	经过审定的术语集可以发布、使用。	术语集编制单位
阶段四：维护更新	对编制形成的术语集持续进行维护更新，并定期提交维护更新后的成果至统筹管理单位。	术语集编制单位
阶段五：成果统筹	对各编制单位形成的术语集进行统筹管理。	浙江数字医疗卫生技术研究院

imit
浙江数字医疗卫生技术研究院
Institute of Medical-care Information Technology

OMAHA
Open Medical and Healthcare Alliance
开放医疗与健康联盟

THANK YOU!

关注我们

加入我们

地 址：浙江省杭州市余杭区良渚文化村随园嘉树风情大道8号
网 站：www.omaha.org.cn
微信公众号：china-omaha
联系电话：0571-88983625
联系我们：us@omaha.org.cn