ICS 35.240.80
C 07

团　　体　　标　　准

T/CHIA 17.2-2020

健康医疗大数据信息资源目录体系
第 2 部分：技术要求

Health big data information resource catalog system
Part 2：Technical requirement

2020-11-16 发布　　2020-12-1 实施

中国卫生信息与健康医疗大数据学会　发布

目　次

前　言

T/CHIA17-2020《健康医疗大数据资源目录体系》分为以下五个部分：

——第1部分：总体框架；

——第2部分：技术要求；

——第3部分：基本元数据；

——第4部分：资源分类；

——第5部分：资源标识符编码规则。

本部分为T/CHIA17-2020的第2部分。

本部分按照GB/T 1.1-2020给出的规则起草。

本部分由华中科技大学提出并归口。

本部分主要起草单位：华中科技大学、国家卫生健康委统计信息中心、空军军医大学。

本部分主要起草人：马敬东、李岳峰、胡建平、董方杰、沈丽宁。

健康医疗大数据信息资源目录体系

第 2 部分：技术要求

1 范围

本部分规定了健康医疗大数据资源目录体系的基本技术要求。

本部分适用于健康医疗大数据信息资源目录管理系统的建设。

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。但是，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。

T/CHIA17.1-2020 健康医疗大数据信息资源目录体系 第 1 部分：总体框架

T/CHIA17.3-2020 健康医疗大数据信息资源目录体系 第 3 部分：基本元数据

T/CHIA17.4-2020 健康医疗大数据信息资源目录体系 第 4 部分：信息资源分类

T/CHIA17.5-2020 健康医疗大数据信息资源目录体系 第 5 部分：资源标识符编码规则

3 术语和定义

T/CHIA17.1-2020 中规定的与以下术语和定义适用于本部分。

3.1

目录服务器 catalog server

按照目录服务器接口的要求，提供健康医疗大数据信息资源发现和目录管理的计算机服务程序。

3.2

元数据库 metadata database

存储元数据的逻辑数据库。

3.3

结果集 resultset

根据目录检索请求在服务器端间的查询结果集合。

4 技术要求

4.1 概述

健康医疗大数据资源目录的整体功能实现依靠各个计算机系统来实现。根据每个系统

所实现的功能不同可细分为编目管理系统、目录内容管理系统、目录内容服务系统、元数据管理系统和目录功能系统。

4.2 编目管理系统

编目管理系统根据各个健康医疗大数据资源的内容，提取其基本特征，按照相关标注实现元数据赋值，形成目录内容。

编目管理系统应遵循以下技术要求：

a) 编目对象是具体的健康医疗大数据资源，其内容包括各个行业、部门日常长期以来形成的海量数据。具体形式可以是数据库、图片、文档等各自类型的数据。

b) 编目管理系统应该支持自动、机辅方式完成元数据元素的赋值。

c) 唯一标识符管理功能：支持唯一标识符的分配和赋值，包括支持后段码的自动生成和管理。

d) 标准符合性检查功能：支持政务信息资源元数据和标准一致性检查，元数据完整性检查的主要目的是保证所有必选的元素据实体和元数据元素已经赋值，标准一致性检查的主要目标是保证已填写好的元数据实体和元素据元素的取值符合编目管理系统的相关规定。

e) 信息资源分类：按照大数据信息资源分类标准，实现对共享大数据信息资源的分类。

4.3 目录内容管理系统

健康医疗大数据资源目录管理系统包含外部管理系统和内部管理系统。外部管理系统就是要建立外部网站门户，对外实现健康医疗大数据资源注册、发布、查询、调阅、推送等功能；内部管理系统就是要对内实现健康医疗大数据资源编目、目录维护、主题统计、共享监测等功能。

4.4 目录内容服务系统

健康医疗大数据资源目录服务系统是健康医疗大数据资源目录管理系统的子系统。可细分为三个服务系统：资源共享服务系统、公共信息服务系统、辅助决策服务系统。

4.4.1 资源共享服务系统

资源共享服务系统是指通过资源注册和查询，实现单一信息源对其他机构、部门的信息资源共享，从而解决信息的完整性和一致性问题。

资源共享服务系统应具备的基本功能包括：

a) 共享资源注册：各类健康医疗数据能够通过资源共享服务系统实现注册，从而达到以规范的方式对各级各类医疗卫生机构产生的各种信息资源进行标准化编目，对注册的资源目录元数据进行集中管理，促进跨机构、跨地域健康医疗大数据资源的共享、开放与应用。

b) 共享资源查询：能够通过资源共享服务系统查询健康医疗大数据资源目录中所包含的数据，从而实现资源的共享与利用。

4.4.2 公共信息服务系统

公共信息服务系统是指通过资源查询与推送，实现对授权人提供完整个人健康医疗大数据信息或对社会公众提供公共健康医疗信息，从而解决信息的可及性和公开性问题。资源共享服务系统主要实现的功能是信息推送，即向授权人提供完整个人健康医疗大数据信息或对社会公众提供公共健康医疗信息。

4.4.3 辅助决策服务系统

辅助决策服务系统是指通过资源查询与调阅，实现多渠道健康医疗信息的采集、汇总、分析与综合应用，为行政管理部门提供多样、科学的决策信息。

辅助决策服务系统应具备的基本功能包括：

a) 信息资源的查询与调阅；

b) 信息资源的汇总与分析；

c) 信息资源的综合应用。

4.5 元数据管理系统

元数据管理系统的目的是实现对元数据的管理，应提供元数据元素管理、元数据实体管理、元数据集管理、元数据版本管理等功能。

其具体功能包括：

a) 添加、修改元数据库；

b) 添加、修改元数据库信息；

c) 添加、修改元数据表；

d) 添加、修改元数据表信息。

4.6 目录功能系统

目录功能系统是实现健康医疗大数据资源目录功能的主要系统。其应具备的基本功能包括：

a) 健康医疗大数据资源注册：注册的资源应该符合健康医疗大数据资源目录的要求，并且给每个注册的资源赋予唯一标识符；

b) 健康医疗大数据资源发布：将注册的信息资源加入到资源目录体系中，并且在门户网站上发布；

c) 健康医疗大数据资源查询：根据查询请求对目录内容信息进行查询，并返回查询结果；

d) 健康医疗大数据资源调阅：实现健康医疗大数据调阅功能，方便查询者浏览信息资源；

e) 健康医疗大数据资源目录维护；

f) 健康医疗大数据资源主体统计；

g) 健康医疗大数据资源共享监测。

5 关键技术方法

5.1 数据库技术

构建健康医疗大数据资源目录体系的关键在于通过构建元数据库来实现元数据管理。

元数据是关于数据的数据，又称中介数据、中继数据，主要是描述数据属性的信息，用来支持如指示存储位置、历史数据、资源查找、文件记录等功能。元数据库是按照数据结构来组织、存储和管理数据的数据仓库。使用元数据目的在于：识别资源；评价资源；追踪资源在使用过程中的变化；实现简单高效地管理大量网络化数据；实现信息资源的有效发现、查找、一体化组织和对使用资源的有效管理。

元数据库的基本结构可以分为三层：

a）物理数据层：数据库的最内层，是物理存贮设备上实际存储的数据的集合。这些数据是原始数据，是用户加工的对象，由内部模式描述的指令操作处理的位串、字符和字组成。

b）概念数据层：数据库的中间一层，是数据库的整体逻辑表示。指出了每个数据的逻辑定义及数据间的逻辑联系，是存贮记录的集合。它所涉及的是数据库所有对象的逻辑关系，而不是它们的物理情况，是数据库管理员概念下的数据库。

c）用户数据层：用户所看到和使用的数据库，表示了一个或一些特定用户使用的数据集合，即逻辑记录的集合。数据库不同层次之间的联系是通过映射进行转换的。

根据我国健康医疗大数据信息资源管理、应用需求，综合国外大数据资源目录体系构建元素，参考国内其他行业大数据资源归类方法，从资源内容、资源表示、资源管理、资源责任和资源获取等5个维度构建我国健康医疗大数据信息资源目录元数据库。

5.2 资源目录分类模型

构建健康医疗大数据资源目录体系的第二个关键问题在于如何实现资源的分类。因此需要构建分类模型来实现健康医疗大数据资源的分类。依据相关学者的研究，健康医疗大数据资源分类由类目、亚目和细目等3个层次组成。类目共划分为8个大类：依据国家全民健康信息化框架中对6大业务应用和3大数据库的总结，分为公共卫生、计划生育、医疗服务、医疗保障、药品管理和综合管理等6个基本业务类，将全员人口信息数据库、电子病历数据库和健康档案数据库归为基础信息类，考虑到移动通讯、云计算、物联网和人工智能等新兴技术在健康医疗领域的广泛应用，增加新兴业态类。亚目是根据各类目领域的特点，按照业务内容的组成部分或业务流程的先后顺序进行分类，并对每个类目都增加“其他”项作为兜底项。

5.3 分类编码

编码是标识信息资源的关键方法，依据我国《卫生信息标识体系对象标识符编号规则》和《卫生信息标识体系对象标识符管理注册管理规程》，我国健康医疗大数据资源根目录为2.16.156.10011.2.100，并分别对类目（2位码）、亚目（2位码）、细目（4位码）和信息资源（10位码）分别进行编码，从而实现对健康医疗大数据资源的标识。

5.4 元数据采集与存储技术

元数据采集技术包括元数据的自动采集技术和手工采集技术。自动采集技术一般和业务系统或者健康医疗大数据资源生产系统结合比较紧密。无论是元数据的自动采集还是手工采集，其基本核心包括两方面的内容：

一是对元数据内容标准的支持。不同的健康医疗大数据资源类型、不同的应用需求所需要的元数据内容是不同的。而且越是复杂的元数据内容标准，其内部的结构和相互关系就越复杂。因此，元数据采集应当支持对不同元数据内容标准的元数据进行采集，同时能够对采集的元数据进行数据完整性和逻辑一致性的检查。数据完整性主要指的是元数据内容标准中所规定的必选必填内容是否都已有值，逻辑一致性指元数据的实体和元数据元素的相互关系是否符合元数据内容的规定。

二是对元数据输出格式的支持。元数据采集完成后，必须首先输出再建立相应的存储。因此，元数据的输出必须采用成熟、主流的数据编码技术进行编码，方便元数据的输出和交换。目前，网络数据交换一般使用扩展标记语言（XML）进行编码，就目前阶段而言，支持XML格式的元数据内容输出是必要的。

元数据存储是目录体系的重要内容。元数据建库就是建立已经采集完毕的元数据的存储。目前，主流的信息存储是采用关系型数据库管理系统对信息进行存储管理，它具有工业化程度高、经济高效的特点。因此，健康医疗大数据资源元数据的存储需要尽量使用已有的关系型数据库进行存储。

元数据是层次型数据，在存储到关系型数据库时，需要进行层次型到关系型的模型转换。如果直接针对元数据实体和元数据元素建立字段，无论对存储结构的稳定性和系统的效率来讲都是不可接受的。元数据的关系型存储的核心需要解决两个问题：一是存储的模式不会随元数据标准的变换而变化，需要在关系型数据库中建立元数据的数据字典描述元数据结构；二是要建立高效率的索引机制保证对元数据内容的有效检索。

5.5 目录服务与应用技术

目录服务技术是指通过网络查询健康医疗大数据资源元数据以得到相关信息的技术。

目录应用技术是向用户展现目录的技术。目录应用技术的核心是元数据的展现技术。目前，元数据一般采用XML进行编码，因此XML编码元数据的展现是目录应用技术需要重点考虑的问题。一般在XML元数据的展现方面有基于级联样式表（CSS）的技术，也有基于DOM和SAX的解析技术。我们需要基于此类技术，并结合相关的应用开发环境，建立各种健康医疗大数据资源目录应用。

6 健康医疗大数据资源目录服务接口

健康医疗大数据目录服务包括发现和管理两种基本功能，发现功能用于对元数据进行检索，管理功能实现元数据的管理。

健康医疗大数据资源目录体系包含五种接口，即基础接口、发现接口、管理接口、交换接口和应用接口。其中基础接口是将发现接口和管理接口中基础性的操作定义成一个公共接口，基础接口和发现接口是必选实现，管理接口、交换接口、应用接口为可选实现，有利于接口的扩展性。这五类接口实现大数据资源的发现功能和管理功能。

a) 基础接口：提供会话管理功能和服务自描述功能，包含有目录服务初始化接口、目录服务终止接口和服务自描述接口；

b) 发现接口：提供元数据检索功能和元数据检索结果提取功能，包含有目录检索接口以及目录检索结果提取接口，这些接口本身不提供资源，而是提供资源基本信息和如何去获得这些资源的元数据；

c) 管理接口：提供元数据管理的功能，包含元数据管理接口；

d) 交换接口：实现数据资源的交换和传输；

e）应用接口：提高目录内容管理系统的可扩展性。

健康医疗大数据资源目录服务是基于超文本传输协议（HTTP）的POST方式实现，协议消息适用XML编码。目录服务各个接口操作均是通过客户端和服务器之间传递的请求/响应消息对来实现。请求消息和响应消息是一一对应的，即对每一个请求消息有且只有一个响应消息产生。目录服务的客户端和服务器通讯建立在会话的基础上，会话通过请求消息和响应消息对来完成，每一个请求消息都有相对应的响应消息。

健康医疗大数据资源目录服务支持核心元数据及其扩展内容的查询，该元数据需要符合健康医疗大数据资源目录体系对于核心元数据的要求，目录服务支持对多个元数据库的查询，元数据一般按照描述大数据资源的内容分别建立。

健康医疗大数据资源目录服务可是集中式的也可以是分布式的。集中式是指元数据库和目录服务在物理上和逻辑上都部署在一个节点，分布式指的是元数据库物理分布分散、目录服务逻辑集中。