

数据集标识符与数据元内部标识符现状和应用研究

冯牧青 贺云

高雪建

(四川华软慧医科技有限公司 成都 610000)

(成都智物云科技有限公司 成都 610072)

严梅

冯昌琪

侯光辉

(四川华软慧医科技有限公司
成都 610000)

(四川天府健康产业研究院
成都 610000)

(宜宾市第二人民医院
宜宾 644000)

〔摘要〕 分析数据元内部标识符应用现状、存在的问题及原因, 阐述建立数据集内部标识符同步机制的路径和方法、数据集内部标识符潜在用途, 为相关研究提供参考。

〔关键词〕 数据集标识; 数据子集标识符; 数据元内部标识符; 应用研究

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2022.08.008

Study on the Status and Application of Data Set Identifier and Internal Identifier of Data Element FENG Muqing, HE Yun, Sichuan Huaruan Huiyi Technology Co. Ltd., Chengdu 610000, China; GAO Xuejian, Chengdu Zhiwu Cloud Technology Co. Ltd., Chengdu 610072, China; YAN Mei, Sichuan Huaruan Huiyi Technology Co. Ltd., Chengdu 610000, China; FENG Changqi, Sichuan Tianfu Health Industry Research Institute, Chengdu 610000, China; HOU Guanghui, Yibin Second People's Hospital, Yibin 644000, China

〔Abstract〕 The paper analyzes the application status, existing problems and causes of internal identifier of data element, expounds the path and methods of establishing synchronization mechanism of internal identifier of data set, and the potential use of internal identifier of data set, and provides references for relevant study.

〔Keywords〕 data set identification; data subset identifier; internal identifier of data element; application study

1 引言

根据《卫生信息基本数据集编制规范 (WS370-2012)》(以下简称 WS370-2012), 数据元内部

标识符是由数据集(数据子集)标识符和数据元在数据集(数据子集)中的顺序号组成^[1-2], 见图 1。数据集(数据子集)标识符是数据元内部标识符的前 9 位(含 1 位小数点), 数据元内部标识符是在数据集标识符基础上增加了 4 位(含 1 位小数点, 数据元的顺序号 3 位), 共 13 位。数据元内部标识符具有数据集、数据子集的汇聚作用, 在数据组织、存储和传输方面具有潜在价值。数据元内部标识符完整, 分类和编码科学, 对数据组织、存储、

〔修回日期〕 2022-04-12

〔作者简介〕 冯牧青, 助理工程师, 发表论文 3 篇; 通信作者: 贺云, 技术总监。

传输和分析意义较大,可以促进大数据和区块链技术在卫生健康领域的应用。WS370-2012 的发布对卫生信息化建设具有里程碑意义,为卫生信息(数据)组织、存储和传输提供更安全可靠的思路。但目前在医疗卫生信息化建设中,数据集中数据元内部标识符的作用远没有达到 WS370-2012 制订的目的,应用现状不佳。



图 1 数据集标识符与数据元内部标识符的关系和结构

2 数据元内部标识符应用现状

研究医疗卫生信息系统中数据库的数据表发现,目前使用的医疗卫生信息化产品中,数据元内部标识对信息系统发挥的作用不大。在医疗卫生信息系统中,数据元内部标识符在数据表中有以下几种情况:在数据表中没有内部标识符字段;有内部标识符字段但没有内部标识符;有内部标识符字段但仅部分有内部标识符(抄取已分布的数据集),新增数据元都没有内部标识符;数据集有内部标识符但较为混乱,内部标识符编码不规律。基于以上情况,数据元内部标识在信息系统中无法发挥作用,对卫生健康信息化发展极为不利。

3 数据元内部标识符应用存在的问题及原因

3.1 信息标准供方

3.1.1 数据集分类代码体系不完整 自 2011 年至今已经发布数十项卫生信息基本数据集标准^[3-7],几乎覆盖了大部分医疗卫生业务,但在医疗卫生信息化系统(软件)中应用不多、执行不到位。尽管 WS370-2012 附录列出了卫生信息基本数据集 1 级和 2 级类目的分类名称及代码基本框架,但仅给出

A-D 这 4 个 1 级业务类目名称和代码,卫生服务 2 级类目中的 5 个 2 级类目分类及代码,见表 1。面对不断增长的业务需求,医疗卫生信息系统(软件)厂商不会增加 1 级和 2 级类目,更无法编码,数据集中数据元标识符只能为空。

表 1 卫生信息基本数据集分类代码^[2]

1 级类目名称	1 级类目代码	2 级类目名称	2 级类目代码
基本信息	A	-	00
卫生服务	B	儿童保健	01
		妇女服务	02
		疾病控制	03
		疾病管理	04
		医疗服务	05
卫生管理	C	-	00
卫生综合	D	-	00

3.1.2 新增数据元未及时更新发布 随着我国卫生健康事业的发展,新业务不断产生。信息系统中新数据元和新数据集不断出现。截至目前卫生行业基本数据集(子集)标准正式发布多达数十项^[3-7],但由于新数据集标准没有得到及时共享,后来者在不知情的情况下又另行编制相同或部分相同的数据集,导致相同业务信息系统出现多个基本数据集,相同业务信息系统不同厂家的产品数据汇聚后,数据集中数据元内部标识符仍然不一致,需花费大量时间、人力和财力对照做接口。

3.1.3 卫生信息标准监督执法不力 部分卫生信息标准是强制性标准,受相关法规约束必须强制执行,否则执法部门可追究法律责任,见表 2。但目前没有明确针对卫生信息标准的执法主体,其监督评审基本由业务部门或标准主管部门组织实施,如医疗健康信息医院信息互联互通标准化成熟度测评、电子病历系统应用水平分级评价标准(试行)、医院智慧服务分级评估标准体系(试行)等^[8-10],这些测评还没有上升到法律法规层面,因此卫生信息强制性标准未能很好地执行,导致信息系统存在问题。

表 2 部分卫生信息强制执行标准示例

序号	分类	标准名称	标准号	发布时间 (年)
1	强制	《卫生信息数据元目录》	WS363. 1 – 17	2011
2	强制	《卫生信息数据元值域代码》	WS364. 1. 17	2011
3	强制	《城乡居民健康档案基本数据集》	WS365	2011
4	强制	《电子病历基本数据集》	WS445	2014
5	强制	《基本信息基本数据集》	WS371	2012
6	强制	《疾病管理基本数据集》	WS372	2012
7	强制	《医疗服务基本数据集》	WS373	2012
8	强制	《卫生管理基本数据集》	WS374	2012
9	强制	《疾病控制基本数据集》	WS375	2012
10	强制	《儿童保健基本数据集》	WS376	2012
11	强制	《妇女保健基本数据集》	WS377	2012

3.2 信息标准需方

3.2.1 缺乏主动执行标准意识 医疗卫生信息系统（软件）生产厂商是卫生信息标准的最大需方。卫生信息标准是否执行直接关系到产品质量，但生产厂商的设计开发人员在短时间内掌握信息标准和医疗卫生业务知识，充分理解数据元内部标识符的经济价值是较困难的，同时厂商在相应领域的投入积极性不高。这造成医疗卫生信息系统（软件）生产厂商设计和开发人员缺乏卫生信息标准和医疗卫生业务基础知识，主观上缺乏执行卫生信息标准意识，导致了市场上大多数医疗卫生信息系统（软件）产品在卫生信息标准执行和标准应用上存在较多缺陷。例如未利用标准的数据元目录构造标准的数据表和标准的数据集（子集），未科学利用数据集中数据元内部标识符去组织、应用、存储和传输数据，导致数据不准确、数据无法合并归档、统计指标口径不统一等问题，直接影响信息系统（软件）产品质量。

3.2.2 厂商被迫应急应对测评 当前医疗卫生信息化建设和应用水平的国家级测评有 3 个^[8-10]，部分省份还有省级测评^[11]。一方面医院建设之初提出测评的达标目标，将达标目标作为信息化建设项目

的验收标准之一，而且时间要求非常高；另一方面承建商为了满足医院相关需求，在时间紧、自身信息标准和业务理解不深入的情况下有针对性地按测评要求开发局部信息标准内容或功能，即测什么做什么，导致承建商在执行信息标准方面缺乏全盘思考，“偷工减料”“以点盖面”，无法发挥数据集中数据元内部标识符的积极作用。

3.2.3 标准管理手段落后 标准制修订工作中手动建立新数据集（子集）时数据元查重非常困难，难以保证数据集标识符和数据元内部标识符的一致性；发布新数据集标准时，因厂家不知晓而无法及时更新其信息标准体系；项目急需时，厂家因不了解数据集标识符和数据元内部标识符结构和规则，数据集标识符和数据元内部标识符只好空缺。从标准的生产到标准的使用过程中，信息化和自动化程度较低，未能做到信息标准的生产和更新同步。

4 建立数据集内部标识符同步机制

4.1 概述

卫生信息标准服务和云服务的研究为解决数据集中数据元内部标识符应用问题提供了业内认同的理念和可行的技术路线^[12-14]。数据集中数据元内部标识符同步首先要解决应用架构、函数模型、程序流程 3 方面的问题。

4.2 应用架构

构建数据集标准自动服务，即面向医疗卫生信息系统（软件）开发厂商建立为企业产品（项目）服务的数据集标准自动化服务系统，使数据集标准的生产和使用始终保持同步，这样就可保证厂商在第一时间获得最新发布的数据集标准、数据集标识符和数据元内部标识符。数据集标准生产方负责建立数据元内部标识符的生产和发送系统，使用方负责接收并且与企业信息标准体系同步更新，见图 2。



图 2 数据集标准应用架构

4.3 构建函数

依照 WS370 – 2012 编码规则与编码结构可事先构建数据元内部标识符编码函数，由程序调用函数引导生成数据集数据元内部标识符，其函数模型如下：

$$f(x) = \text{"HDS"} + \mathcal{I}! DC \in \{A, B, C \cdots Z\} + \mathcal{I}! SC \in \{01, 02, 03 \cdots 99\} + \mathcal{I}! SSN \in \{01, 02, 03 \cdots 99\} + \mathcal{I}! ON \in \{001, 002, 003 \cdots 999\}^{[15]}$$

$f(x)$ 代表待定数据元内部标识符，HDS 表示卫生信息领域^[2]，DC (DomainCode) 代表业务域 1 级类目，SC (SubClasses) 代表业务域 2 级类目，SSN (Subset Sequence Number) 代表数据子集序号，ON (OrderNumber) 代表数据元在数据集 (子集) 顺序号。

4.4 程序流程

由程序引导完成数据集中数据元内部标识符编制可最大限度降低人为因素 (知识欠缺、操作笔误、前后矛盾等) 影响，保证数据集中数据元内部标识符质量，见图 3。程序引导生成数据元内部标识符过程中，手动选择 1 级和 2 级业务域类目名称 (DC $\in \{A, B, C, D \cdots Z\}$ 、SC $\in \{01, 02, 03 \cdots 99\}$) 时，需由具备一定医疗卫生基础和分类知识的专人完成；选定 1 级类目名称就有 1 位字母代码自左至右自动迭加，生成数据集标识符；选定 2 级业务类目名称后 (SSN $\in \{01, 02, 03 \cdots 99\}$) 有 2 位数字代码由程序自左至右自动迭加，生成数据子集标识符；SC $\in \{001, 002, 003 \cdots$

999} 顺序号无需人工操作，由程序在上一个数据元序号后加 1 构成。

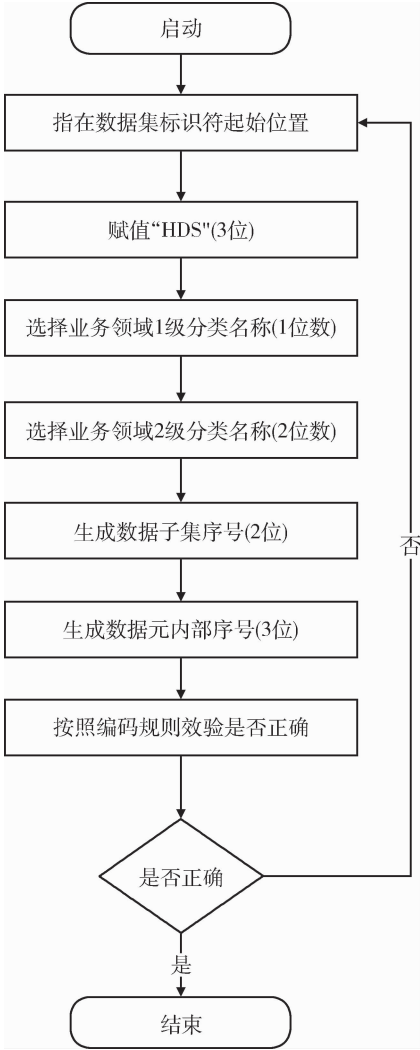


图 3 数据元内部标识符程序流程

5 数据集内部标识符潜在应用价值巨大

5.1 数据组织更科学灵活

利用数据集标识符和数据元内部标识符可科学地组织数据，方便快捷地建立主题数据库，增强数据使用的灵活性和利用效率。如临床科研需要在数据中心提取数据时，可利用数据元内部标识符的前 6 位代码提取 1 级类目 (主题) 数据，利用前 9 位代码提取 2 级类目 (亚主题) 数据，利用前 13 位代码提取该数据元所有数据。由计算机程序实现非常简单。

5.2 存储传输数据更安全

在数据传输和存储过程中,可以数据元为最小单位而不是以数据表,用特定算法打乱数据元在数据集(子集)中的顺序进行传输或存储,即便被局部截获也不会泄漏完整数据结构,数据接收方收到数据后,通过特定算法的逆运算还原数据,为数据传输和存储提供新的安全手段。

5.3 分析数据更全面精准

通过数据集标识符、数据子集标准符和数据元内部标识符可准确定位大数据分析所需要的数据集、数据子集、数据元。在用 Selecte 语句提取数据时,避免同一数据元因为数据集不同或数据元内部标识符重码而导致的数据误差。数据元内部标识符的唯一性使得数据分析精度更高,大数据逻辑链更长、更有实用价值。

6 结语

数据元内部标识符具有数据集、数据子集的汇聚作用,其在数据组织、存储和传输方面具有巨大应用价值。但目前医疗卫生信息化建设中,数据集标识符和数据元内部标识符的作用未能发挥,这缘于信息标准供方和信息标准需方两方面原因。信息标准服务亟需实现自动化,只有建立自动化信息标准同步机制才能保证标准供方和标准需方的信息实时同步,数据集(子集)标识符和数据元标识符才能在数据组织、存储、传输和分析中发挥作用。

参考文献

- 1 原卫生部. 卫生信息数据集分类与编规则: WS/T306 - 2009 [EB/OL]. [2021 - 01 - 22]. <http://www.nhc.gov.cn/wjw/s9497/200909/42764/files/6bf81697b96e4fd284e920225a66b09d.pdf>.
- 2 原卫生部. 卫生信息基本数据集编制规范: WS370 - 2012 [EB/OL]. [2021 - 03 - 15]. <http://www.nhc.gov.cn/ewebeditor/uploadfile/2014/11/20141124175958601.PDF>.
- 3 原卫生部. 城乡居民健康档案基本数据集: WS365 -

- 2011 [EB/OL]. [2021 - 08 - 02]. <http://www.nhc.gov.cn/wjw/s9497/201108/52775/files/fa747f4fc6204368bf3777a926a9624.pdf>.
- 4 原国家卫生计生委. 电子病历基本数据集: WS445 - 2014 [EB/OL]. [2021 - 05 - 30]. <http://www.nhc.gov.cn/ewebeditor/uploadfile/2014/06/20140620111008557.PDF>.
- 5 原卫生部. 疾病管理基本数据集: WS372 - 2012 [EB/OL]. [2021 - 03 - 15]. <http://www.nhc.gov.cn/wjw/s9497/202204/d9073cac740e43188cbcc160dd362019/files/d09d5e4e5f3a445f8e87bf8bd4b7f7a7.PDF>.
- 6 原卫生部. 疾病控制基本数据集: WS375 - 2012 [EB/OL]. [2021 - 03 - 15]. <http://www.nhc.gov.cn/ewebeditor/uploadfile/2014/11/20141124180434232.PDF>.
- 7 原国家卫生计生委. 妇女保健基本数据集: WS377 - 2013 [EB/OL]. [2021 - 12 - 27]. <http://www.nhc.gov.cn/ewebeditor/uploadfile/2014/10/20141010171441556.pdf>.
- 8 中华人民共和国国家卫生标准委员会. 医院信息互联互通标准化成熟度测评指标体系 [EB/OL]. [2021 - 06 - 26]. <https://wenku.baidu.com/view/efbc625e0508763231121298.html?fixfr=BiOKT9Z1nATEkzKL3%252F8zFQ%253D%253D&fr=income1-aladdin-search>.
- 9 原国家卫生计生委医院管理研究所. 《电子病历系统应用水平分级评价标准(试行)》[EB/OL]. [2021 - 09 - 26]. <https://wenku.baidu.com/view/c3eece5a66e58fafab069dc5022aaea998f41be.html>.
- 10 国家卫生健康委. 《医院智慧服务分级评估标准体系(试行)》[EB/OL]. [2021 - 03 - 05]. <https://wenku.baidu.com/view/1afee78e2e60ddccda38376baf1ffc4ff47e20c.html>.
- 11 四川省卫生健康委. 四川省智慧医院评审标准 [EB/OL]. [2020 - 05 - 30]. <https://www.doc88.com/p-26116026943696.html>.
- 12 冯牧青, 贺云, 高雪建. 卫生信息标准服务平台设计[J]. 医学信息学杂志, 2021, 42 (7): 48 - 52.
- 13 林德南, 郑静, 王爽, 等. 深圳市全民健康信息标准体系研制的探讨[J]. 中国卫生信息管理杂志, 2019, 16 (4): 418 - 422.
- 14 杨雪蓉, 王俊淑. 卫生信息标准管理系统设计与应用[J]. 中国卫生信息管理杂志, 2020, 17 (1): 30 - 33.
- 15 屈婉玲, 耿素云, 张立昂. 离散数学(第2版)[M]. 北京: 高等教育出版社, 2015.