Organizing Health Data Standards Based on Knowledge Map

Liu Shen, Panpan Deng, Haixia Sun

Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences & Peking Union Medical College

Abstract

To facilitate easy use of health data standards, we collected health data standards and parsed them into more fine-grained knowledge units. Reference and inclusion relations among these standards were constructed into knowledge map. Until now there are 156 standards collected and 4796 reference relations linked within those standards. Besides, an interface was built to enable users to easily get one standard's main information without referring to numerous PDF documents.

Keywords:

ICD Codes; Data Linkage; Knowledge Bases

Introduction

The increasing volume of data in healthcare makes it a challenge to use these health data correctly and effectively. The standardization of the health data is the precondition to perform analysis methods [1]. National Health Commission of the People's Republic of China released many health data standards covering many fields in medical management activities [2]. Besides, some commercial literature databases also include some health data standards. But data in health data standards are difficult to capture and are not systematically collected and linked in those databases, usually as they are "trapped" in PDF documents that can be difficult to parse into structured fields computationally. For example, to know what data item and their allowable values in a laboratory test report would be a time-consuming task, for several databases and documents should be searched and checked. Health data standards will be collected, parsed and organized in a formalized, editable and searchable way. Also based on the reference relations among standards, linkage among those health information standards will be built into a knowledge map thus researchers, as well as clinicians, can easily get well-organized knowledge from the national health data standards.

Methods

Data Overview and Preprocessing

The health data standards can be divided into 3 main collections[2]. One part is health data elements which consist of 6 properties and focused on one subject. Take the standard 'Health data element dictionary-Part 12: Medical plan and intervention' as an example, abdominal pain degree code is one of the data elements in it. As listed in table 1, the six properties are: data element identifier; data element name; definition; data type of data element value; presentation format; and the allowable values. In that data element, the data type is S3, meaning they referred to other code systems and the presentation format is N..3, which means the code is a number

with at most 3 characters, which were all defined in the primary health data standards. Sometimes the allowable values may be an enumerated type, or sometimes they may be referring to other resources like other code systems or other code value domain standards.

Table 1–Data Element for Administration Route Code

Items	Values
Identifier	DE06.00.134.00
Name	Administration route code
Definition	Indicate the route of medication in a particular coding system
Data Type	S3
Presentation Format	N..3
Allowable Values	Refer to the standard "WS 364.12-2011 Classification and coding for value domain of health data element-part 12: Medical plan and intervention", one of the common value sets CV06.00.102 Administration route code

Other important parts are standards for medical activity records, like electronic medical records and residents' health records. These standards have specifications for the document of different medical activity scenarios, like outpatient and emergency medical record and consultation record. They defined what items should be included in those standards, how these data should be organized and what standards should be referred to or adopted. Besides, standards for coding value domains are collected which consist of allowable values for some of the data element. Most of the health data collected standards documents are in a non-editable format. To parse the knowledge units in them and digitize and create linkage among health data standards, we conducted Optical Character Recognition (OCR) process to all the collected health data standards and put them in a structured database.

Knowledge Linkage Design

To construct the knowledge map, we digitized each standards' data elements, code values and document content compositions and specifications. We defined three main relations in the knowledge map: has part; refers data model and refers data value. Has part means the inclusion relationship. Refers data value indicates the value range will be based on the other standards'. Refers data model means one standard may adopt the other standards' data elements which provide not only data values but also data types and formats. Thus linkage can be constructed based on the inclusion relationships and reference information. The health data standards knowledge map linked among standards are listed below in figure 1. Besides referred
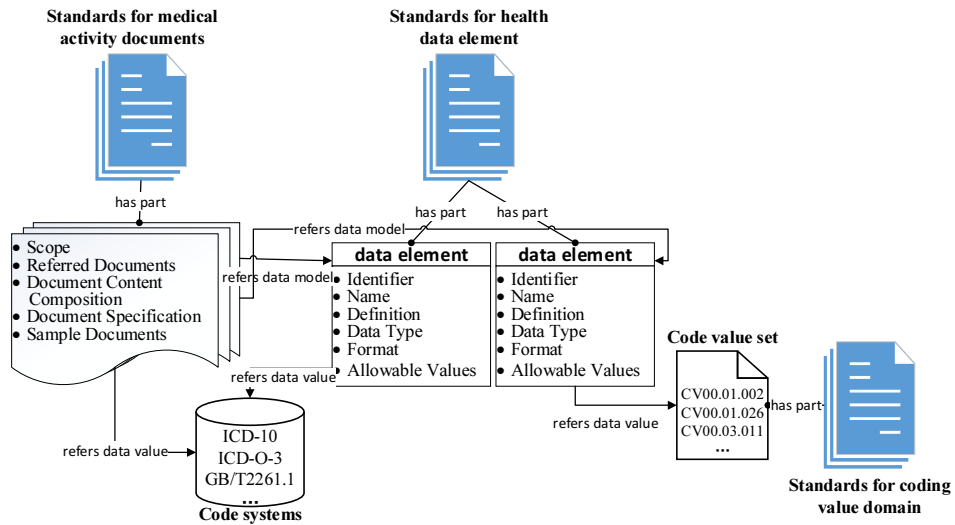
Figure 1–Health Data Standards Knowledge Map

to other standards' code values, some data elements' allowable value range may also adopt some common international code system's standard, like International Classification of Diseases (ICD) and national recommended standards. In this way, we linked all those three collections of health data standards and they can be easily reached and explored.

Results

This work provided a more fine-grained and searchable health data standards database, which currently consists of 156 health data standards, of which 37 are mainly standards for medical activity documents and 119 are mainly standards for health data elements and code values. There are 622 code value sets and 1748 data elements who will provide references for other standards. Up to the submission, there are 4796 reference relations. Besides, we also provide users an interface to search and check standards, more finely grained data elements and the reference map. The health data standards were linked by the data elements, code value sets and other code systems, and both the standards and data elements can be browsed or searched in a user-friendly interface. All of the data have been curated and standardized. Here is an example of data linkage based on the knowledge map.
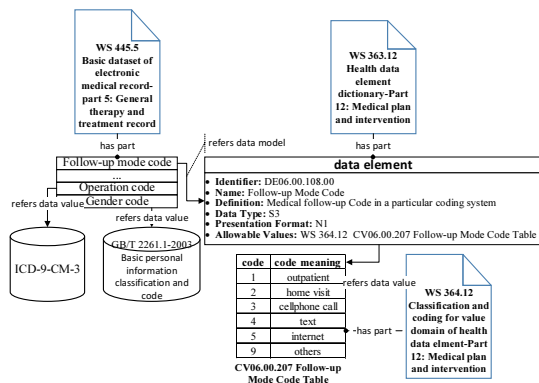

Figure 2– Data Linkage examples

Conclusions

By digitizing and connecting health data standards in a knowledge map, the detailed items and referrention relations can be more directly obtained thus in the future could improve the efficiency for researchers and medical staffs. Compared in searching abilities with WANFANG DATA and CNKI, two of the most commonly used commercial literature databases in China, and Chinese Health Information Standard Portal, our system can provide data element retrieval function within hundreds of health data standards. Properties of data elements will be easily accessed, especially the allowable value which may be the most commonly used ones. Besides, reference relations like which standards have referred to one same data element are showed clearly. However, there are still some places needed to be further explored since we haven't opened this interface to researchers and medical staffs to facilitate their work and get their feedbacks and evaluations.

Acknowledgements

References

[1] D. Liu et al., Harmonization of health data at national level: A pilot study in China, *Int J Med Inform* **79** (2010):450-458.

[2] National Health Commission of the People's Republic of China. Health Data Standard [EB/OL] http://www.nhc.gov.cn/wjw/s9497/wsbz.shtml [accessed March 23, 2019]

Address for Correspondence

NAME: Haixia Sun, EMAIL: sun.haixia@imicams.ac.cn