数据标准建设研究及其应用

摘要：智慧校园数据中心建设以苏州经贸职业技术学院《教育事业“十三五”发展规划（2016-2020）》为统领，坚持推动信息技术与教育教学的深度融合。本文主要介绍高校数据中心建设实现过程中数据标准建设的主要理论和实现方法。我校在数据标准建设进行研究与探索，把各个独立的业务系统数据通过采集，数据进行梳理和质量初检，数据标准核对与数据补充核准，数据清洗和整合交换，建立共享数据中心，数据管理与质量评估，可用服务接口设计，数据价值模型建立于展示；本文通过 Oracle ODI 和应用系统建立标准化的数据，利用数据交换工具构建数据中心，实现数据平台标准化的建设、管理与维护。

关键词：数据元；OracleODI；元数据；目标代码；主数据

致谢

https://mp.weixin.qq.com/s/inPBG30I2VQASVbnYmlJDg

作者 吴艳伟

本文大量文字参考以上文章，对此向作者表示衷心感谢，后期文档会不断完善，学习作者的思想和方法，完成我们的具体项目

0 引言

智慧校园建设发展到今天，很多业务都已经信息化建设完成或都正在建设，大多数系统都是在某些业务需求基础上建立，没有考虑与其它系统的功能重复和数据重复，数据一致性和可用性矛盾突出。具体表现为：（1） 数据需求缺乏规范，造成数据对象多份存储，存储结构各异，严重影响数据共享；（2） 数据标准依据各异，造成统计口径无法匹配；（3） 业务口径不统一，造成沟通困难，发生歧义。

学校数据标准体系建设已经到了迫切需要的的时候，数据标准的建设可以为以后大数据应用提供建设性的依据，中华人民共和国教育行业标准 JY/T 1006—2012《教育管理信息 高等学校管理信息》也已经颁布多年，我校数据标准体系架构在最近 2 年已经开始建设，建设的具体表现主要包括元数据(标准数据)管理、代码标准管理和主数据管理。元数据管理主要是指教育部部标和教育行业的行业标准,这里主要涉及 11 个数据集应用域；代码标准主要是指元数据管理中数据的基本单元，例如：性别、宗教等，在代码标准中予以定义；主数据管理，主要是实现校标的指定，校级数据标准定义主要来源与元数据+代码标准+学校自定义标准完成。数据标准的制定可以在业务、技术和管理多个方面给提供支撑。业务方面可以提升规范性和提升数据对业务分析支持度，通过数据标准，实现数据信息统一一致，使得数据更容易在各业务部门之间流转；技术方面首先，相同结构的数据，才更容易实现共享和交换，其次，相同的数据标准，减少大量的转换、清洗工作，极大的提升数据处理效率；管理方面，数据标准更多的是能提供完整、及时、准确、高质量的数据，为决策支持、精细化管理等提供支撑。

本文根据实际工作中的 2 个案例，通过原人事系统教职工基本信息表数据标准化，学生教务系统完成数据标准化，完成实际工作中数据标准建设。

1 数据分析

1.1 人事数据和学生数据采集

学校人事数据标准平台建设之前建设的一个业务系统，有独立的服务器系统，采用的是

SqlServer 数据库，通过与原开发商沟通，在原系统上指定用户上获取视图接口，采集到的数据是没有经过。

学校学生数据标准平台建设之前建设的一个业务系统，有独立的服务器系统，采用的是Oracle 数据库，通过与原开发商沟通，在原系统上指定用户上获取视图接口。

1.2 元数据数据分析

元数据是描述具体的信息资源对象的数据，并能对该对象进行识别和管理，实现信息资源的有效发现与获取。元数据定义均来自于数据国家或者行业标准，学校需要结合自身情况进行维护更新。

人事标准数据元 T_JZG 数据组成包括规定了教职工的个人基本数据项。

学校学生数据元 T_BZKS 数据组成包括本专科生的基本信息、学籍信息和来源信息。

元数据标准涉及到的分类信息包括：学校概况、人员基本信息、人事管理、科研管理、学工管理、教务管理、资产管理、财务管理、党组织管理、统战工作管理、学生组织管理、外事管理、办公管理、体育卫生管理、档案管理、校友管理和统一身份管理等组成。

1.3 目标代码分析

目标代码主要是对元数据部分字段代码的详细说明，目标代码和数据元定义比较类似通过定义、标识、表示以及允许值等一系列属性描述的数据单元，在特定的语义环境中是不可再分的最小数据单元。

人事数据标准数据元依赖政治面貌代码表和部门对应表（部门对应表根据情况设置为校标），由于原人事系统提供数据的局限性，导致更多的其他相关的目标代码没有办法进行匹配，待以后新的系统对接上线可以予以定义和限制。

学生数据标准数据元依赖民族代码表、部门院系对应表、班级信息表、专业信息表等，由于原教务系统提供数据的局限性，导致更多的其他相关的目标代码没有办法进行匹配，待以后新的系统对接上线可以予以定义和限制。

1.4 主数据分析

在完成了元数据分析和目标代码分析以后，就可以构建学校的主数据标准，这项工作是一项要求极其规范的工作，是一项长期的工作。原则上为了使学校内外部使用和交换的数据是一致和准确的，经协商一致制定并由相关主管机构批准，共同使用和重复使用的一种规范性文件。主数据规范又不仅仅是一套规范，而是一套由管理规范、管控流程、技术工具共同组成的体系，是通过这套体系逐步实现信息标准化的过程。主数据规范标准化是通过一整套的数据规范、管控流程和技术工具来确保的各种重要信息，例如学生、教职工、机构、教学等在全学校内外的使用和交换都是一致、准确的过程。另外，主数据规范标准也不仅仅是技术或者业务一个部门的事情，它是在数据层面上对重要业务主题的统一规范，也是业务规范在数据层面上的实现。数据标准实施依赖于业务部门之间的共识，以及业务和技术之间的配合。

主数据规范由元数据规范+目标代码规范+学校自主定义数据规范组成。主数据规范根据不同的数据域可以分为基础类主数据、分析类主数据、专有类主数据。与元数据标准分类相似。主数据建设的最终目标将直接决定数据标准建设的成败。

2 数据标准化实现方案

2.1 Oracle ODI 实现关键过程

ODI（Oracle Data Integrator）是 Oracle 公司提供的一种数据集成工具，能高效地实现批量数据的抽取、转换和加载。ODI 可以实现当今大多数的主流关系型数据库（Oracle、DB2、SQL Server、MySQL、SyBase）的集成。

ODI 提供了图形化客户端和 agent 运行程序。客户端软件主要用于对整个数据集成服务的设计，包括创建对数据源的连接架构、创建模型及反向表结构、创建接口、生成方案和计划等。Agent 运行程序是通过命令行方式在 ODI 服务器上启动的服务，对 agent 下的执行计划周期性地执行。

我校数据化标准建设就是依托于 Oracle ODI 完成。具体的数据标准化实现工具在 ODI 里面主要包括以下步骤：

拓扑管理器 Topology Manager,，

（1） 创建数据服务器和物理架构

（2） 创建逻辑架构

（3） 创建代理

设计和操作 Designer 和 Operator：

Designer 定义数据转换和数据一致性的规则以及数据的过滤条件，Operator 主要用于对生产数据处理过程进行监控。

（1） 创建模型

（2） 创建项目

（3） 创建接口

（4） 创建包

（5） 生成方案、计划

Agent 代理，ODI 的 Agent 是一个能作为 TCP/IP 监听端口的 JAVA 服务，agent 服务下包括一些预先设定时间的方案，当 agent 处于运行时，它会根据方案设定的时间和周期自动执行。

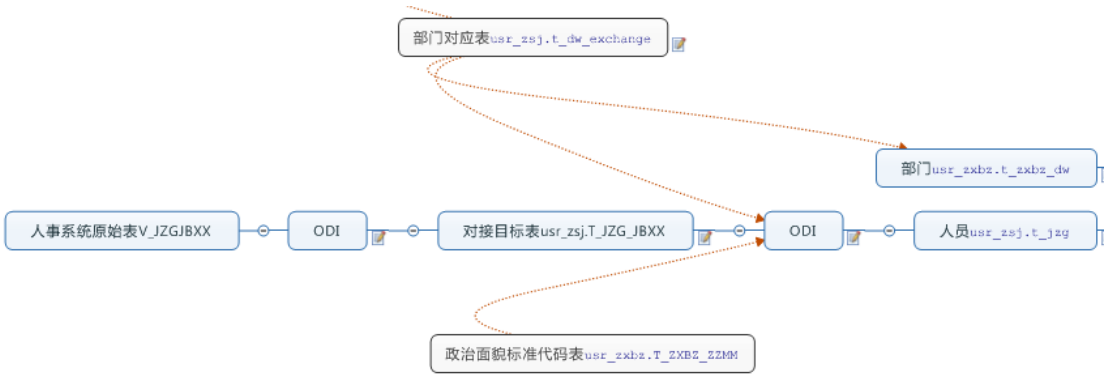
具体详细的操作这里不在详细说明。

2.2 人员主数据标准化实现方案

人员的主数据标准化实现过程主要包括:

A 通过 ODI 对接原人事系统数据 V_JZGJBXX 到数据中心 T_JZG_JBXX；

B 对接数据与目标代码表进行清洗比对，通过 ODI 写入 T_JZG 标准表；



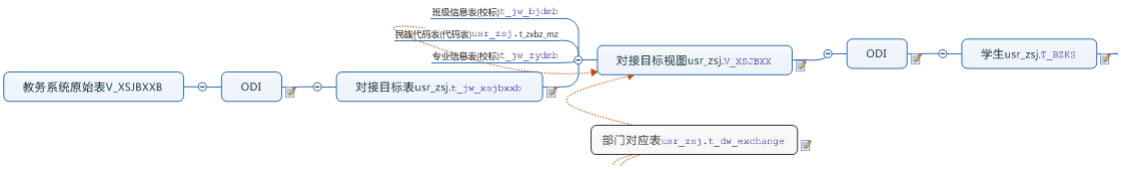
2.3 学生主数据标准化实现方案

学生的主数据标准化实现过程主要包括:

A 通过 ODI 对接原教务系统数据 V_XSJBXXB 到数据中心 T_JW_XSJBXXB；

B 对接数据与目标代码表进行清洗比对，生成视图 V_XSJBXX；

C 通过 ODI 完成视图 V_XSJBXX 写入 T_BZKS 标准表；



2.4 数据标准对核心数据的定义

首先，标准不是模型，标准是可以落地的核心元素；

其次，针对核心数据标准主题选择要多维度考虑。

从数据标准对业务影响度、系统管理度和可实施性等三个方面做分析

A 针对业务影响度，可以通过组织集中讲解、面谈解答以及调查问卷等多种调研活动；获得主题涉及的问题数量、问题影响业务数量、问题影响业务的重要性；

B 应用系统关联度，可以通过分析各部门关注次数、各系统和系统模块使用次数；并通过对应用系统功能梳理，提炼相关实体；以及对相关实体，进行数据主题归结，形成主题在系统中的分布情况；

C 可实施分析，可以通过产品手册、各业务部门体系文件，获得主题定义和分类，以及信息项情况；分析获得数据差异性；获得数据定义不一致程度、业务规则整合难度。

通过分析，每个主题关系的业务系统数量不同，业务关注程度也不同，可实施程度不同（差异量，技术等），最终形成主题选择分析图表。

2.5 数据标准要包括技术与业务两种属性

(1) 数据标准主要是针对业务，学校很多业务的语义十分依赖业务人员的人工梳理，难度大效率低，很可能出现因为梳理人员没有及时梳理，而造成业务语义难以被及时发现和管理的问题。

未来学校将会面临数字化转型，从非结构化的文档中，将大部分业务语义抽取出来，并统一管理，成为未来的发展趋势，这种能力可以通过自然语言分析技术来实现，学校可以通过综合多个材料中对同一业务的描述，分析出最新与最广泛认可的业务定义，由业务人员确认之后，识别出业务语义，这样大大减少了业务人员的工作量，提升了业务人员梳理业务语义的积极性。

(2) 在学校数据治理中，任何一个数据标准，如果没有对应的技术手段，都将难以落地，所以学校建立数据标准时，需要加入信息项的英文名称，来和实际数据库表中的字段相对应。

在数据标准中加入信息项的英文名称能给学校数据治理带来两方面的好处：

在做模型设计的时候，标准可以直接与模型设计工具集成，设计模型时就可以直接引用标准。

对已有系统，标准能够通过英文名称直接和应用系统的相关字段对应，自动发现与不

符合标准的字段，并通过元数据直接通知给相应的系统。

(3) 标准中有了技术和业务信息，还需要有效的关联才能发挥效用。对于学校数据管理来说，技术能弄懂业务的前提是技术与业务之间要有对应，这种对应不能靠大量的人工梳理完成，否则业务部门负担很重，积极性不高。需要能够通过技术手段，利用数据治理工具提供商的行业实践积累，形成业务与技术的自动关联库，自动完成业务与技术对应，将能大大减少业务人员的工作量，同时提升技术与业务关联的准确度，消除业务与技术之间的鸿沟。

2.6 数据标准要持续更新

对于学校数据治理来说，有很多数据标准建立以后，往往只是一套书，没有根据学校业务发展及时做出更新，时间长了就成为了摆设，实际上，数据标准是需要随着学校的业务变化而不断进行修订的，比如在学校新业务的时候，需要增加相应的标准进去，对于没有价值的标准，也要及时废弃。只有这样，才能保证数据标准一直能适应业务发展需要，促进标准落地。

3 数据标准体系结构

数据标准内容根据不同的数据域，主要包括基础类数据、分析类数据、和专有类数据；

3.1 基础类数据

基础类数据是学校日常业务开展过程中所产生的具有共同业务特征的基础性数据；

根据 JYT1006_高等学校管理信息规范，针对基础类数标，可以教育行业经常用的数据标准主题模型。该模型是以主题组织数据，包学校情况、人员基本信息、人事管理、科研管理、学工管理、教务管理、资产管理、财务管理、党组织管理、统战工作管理、学生组织管理、外事管理、办公管理、体育卫生管理、档案管理、校友管理和统一身份管理等主题。

3.2 专有类数据

属于基础类数据，高职院校专有数据子集和学校自定义数据属于专有类数据；

3.3 分析类数据

分析类数据是为满足学校内部管理需要及外部监管要求，在基础性数据基础上按一定统计、分析规则加工后的数据，属于以后的大数据分析范畴不在本文讨论之内；

3.4 建设数据标准步骤

根据学校的实际情况，建设数据标准包括制定、落地、维护等过程。其中制定过程包括规划、调研、设计；落地过程通过映射、标准执行等实现；维护过程保证了数据标准的持续更新。

A 首先，在标准制定过程中的第一个阶段，标准规划阶段，要根据业界经验和学校实际情况确定实施范围，并根据优先级和难易度制定计划。

B、接下来，在调研阶段，通过制定调查问卷、安排现场访谈、收集文档资料等手段，针对各个业务系统以及应用系统进行调研，了解跟标准相关的内容，包括现有定义、使用习惯、数据分布、数据流向、业务规则、服务部门等，形成调研报告，分析问题，并讨论解决

方案。

C、标准设计工作，在方法论指导下，完成数据标准设计和定义工作，包括数据业务描述定义(业务属性)、类型长度定义（技术属性）、其他标准信息定义。设计出定义与分类、信息项、标准码等文档，并通过各部门的评审验证。最终达成一致，形成校级标准。

D、接下来主要是标准如何落地工作。把已定义的数据标准与业务系统、业务应用进行映射，标明标准和现状的关系以及可能影响到的应用。

标准落地一般通过两种方式：

1）新系统建设，直接参考数据标准；

2）旧系统通过标准映射，实现数据关系转换，以及指导后续数据平台建设。

E、做完数据标准映射，接下了就是标准落地执行。

这个过程一般需要借助专业的工具实现标准落地检查。标准执行一般有两个过程

1）第一步分析出来现有问题，例如数据缺失、数据不一致等；

2）第二步修正，例如补录数据、修改系统、新建系统等。

通过这些措施，逐步规范数据建设过程，实现数据标准的落地。

F、数据标准也不是一成不变的，随着业务发展，有些标准需要不断的修订和完善。因此数据标准还有一个关键的管理环节，那就是需要能持续维护改进。

G、在数据标准维护阶段，需要有相应的需求收集、需求评审、变更评审、发布等多个步骤，并能对所有的修订做版本管理，以方便将来问题查找。

4 结论

一般学校数据标准建设完，只停留在册子和书本上，缺乏落地手段，不能有效执行；另外，针对数据标准本身缺乏管理，不能有效适应新业务发展。)

学校数据管理建设思路侧重于事前预防，将各领域数据管理的要求融入到系统研发当中，从需求编写和需求分析等数据产生源头进行管理。严格按照数据标准进行需求编写，结合数据质量管理、元数据管理串联整个软件生命周期。同时在这个过程中，不断的验证和修订数据标准，使得数据标准一直能够适应新业务的发展需要。

通过项目实施：

(1) 借助技术手段实现了数据标准的实施落地。在需求、开发、上线等各阶段都会有数标检查，实现全生命周期数据管控；

(2) 通过系统管理，推进了数标的持续更新，保持了数据标准生命力。