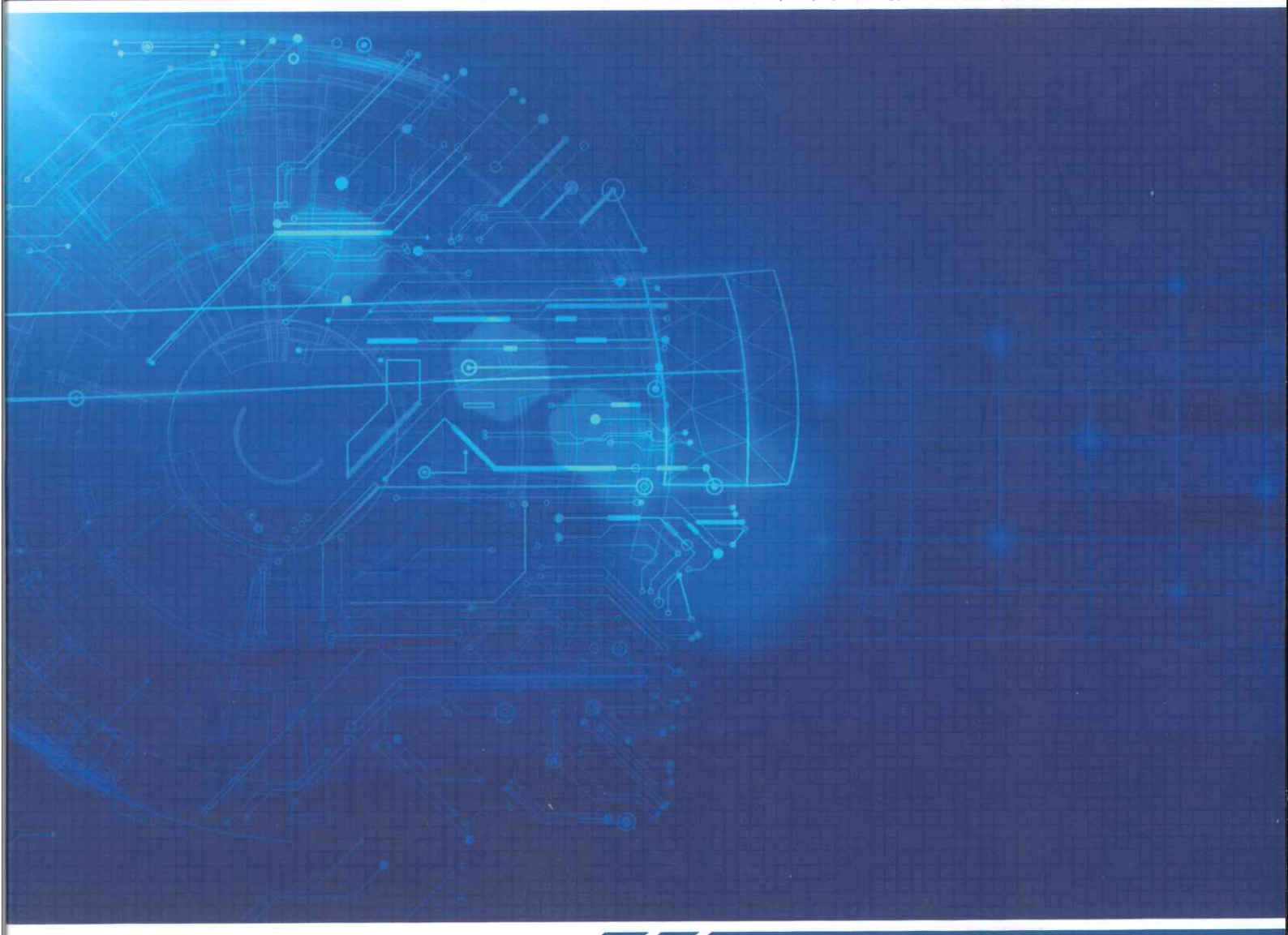
ISSN 1671-2064
CN11-4650/N

12下
2022
总第396期

中国科技纵横

中华人民共和国科学技术部主管



ISSN 1671-2064

万方数据数字化期刊群收录期刊
龙源国际期刊网全文收录期刊
中国核心期刊(遴选)数据库收录期刊
中文科技期刊数据库全文收录

能源科技

科技创新驱动

节能环保与生态建设

信息技术与应用

工艺设计改造及检测检修

工程设计施工与管理

XML 解析及数据校验系统研发设计与实现

李真[1] 孟晓[2] 黄江平[3]

（1.中国软件评测中心软件与信息系统测评工程技术中心，北京 100048；2.中国软件评测中心软件与信息系统测评工程技术中心，北京 100048；3.中国软件评测中心软件与信息系统测评工程技术中心，北京 100048）

摘 要：XML 指可扩展标记语言，可用来存储和传输数据。各行业基于 XML 语言制订一系列标准，然而，目前对于依据标准生成 XML 文档的标准符合性并没有通用工具可进行数据验证。本文对 XML 数据校验进行行业需求分析，对各行业 XML 数据规则进行提炼，描述 XML 解析及数据校验系统关键点设计，分析了系统技术架构和 XML 测试流程。

关键词：XML；Schema；XML 校验；规则

中图分类号：TP314 文献标识码：N 文章编号：1671-2064(2022)24-0048-04

1. 系统需求分析

随着标准表示格式的丰富，XML 文档作为系统间数据传输和数据互通的载体，被各行各业广泛应用于标准制定[1]。XML 格式数据的高可读性使其不仅适合用于网络中结构化数据传输，更方便程序员读写[2]。

XML 文档校验中一个很重要的角色是 XML Schema。XML Schema 主要用于描述 XML 文档的结构，同时也可用于对 XML 文档校验[3]。一般用户方提供的 Schema 可能因描述颗粒度不同，未涵盖全部业务规则要求。通过编写代码可实现最细致的 XML Schema，能够将 XML 结构和数据规则均进行定义。但代码实现的 Schema 不仅非常复杂，而且较为死板，无法达到界面可视化灵活可配。而大部分 Schema 可能仅仅是框架描述。除此之外，业务规则如节点或数据之间的约束关系（排他、多选、外键等）一般未在 Schema 中体现。例如，医疗行业标准《WS/T 500.32-2016 电子病历共享文档规范 第 32 部分：住院病案首页》[4] 定义诊断记录章节基数为 1..1，意为诊断记录章节必须存在且只能出现一次，而没有一个现成工具能对这些约束关系进行校验。因此，虽然 XML Spy 等工具可对 XML 文档格式进行校验，但如果 XML Schema 本身不是最细致的表达，将导致验证被测目标是否满足规范要求存在覆盖不全面、数据标准需要人工逐一比对验证等问题[5]。

我们曾开展医疗行业标准符合性测试（如《WS 445-2014（所有部分）电子病历基本数据集》和《WS/T 500-2016（所有部分）电子病历共享文档规范》）[4,6,7]，具有丰富的测试经验。但曾用于开展测试的工具仅为定制化工具，无法应用于所有行业。基于业务对数据格式校验的需求，我们在此基础上研究提取各行业 XML 标准中所涵盖通用规则，设计开发一套通用于多行业的 XML 解析及数据校验系统。该系统主要实现以下功能：

（1）实现针对不同数据标准规则，对 XML 文档中数据项进行标准符合性测试。

（2）界面可视化实现标准规则可定制化、灵活配置。

（3）仅有 XML 示例模板而无 Schema 的情况下，可反向生成 Schema 文件。

该系统是一套满足校验各行业 XML 文档结构的解析及数据标准校验系统。该系统方便用户操作同时可灵活配置，通过配置业务规则即可满足测试需求。该工具将适用于更多行业，且拥有完全自主知识产权。

2. 系统关键点设计

2.1 格式校验配置

2.1.1 约束条件设置

某些行业在制订标准的同时会编制 Schema 文件，此时可将 Schema 文件导入系统，通过配置基线（节点出现的最大值和最小值）实现对文档中章节和节点的循环设置。基线对应标准约束条件，如医疗行业标准《WS/T 500.32-2016 电子病历共享文档规范 第 32 部分：住院病案首页》定义诊断记录章节基数为 1..1。此处基数即为该章节约束条件，可通过系统界面化实现基线配置功能，灵活设置 Schema 文件约束条件。

2.1.2 生成 Schema

通过总结多种行业标准，我们发现并非每个行业制定 XML 标准及模板后，均会编写 XML Schema。研究该需求缺口，我们开发通过导入 XML 示例模板并进行简单编辑的方式生成 Schema 文件功能。提交 XML 示例模板至系统，根据行业标准规定配置 XML 示例模版相关约束和条件，可生成相应 Schema 文件，再对 XML 文档进行校验。系统根据 XML 示例模版生成的 Schema 文件将放在项目目录 Schema 下，数据库中同时保存该文件完整目录，如此可针对两种来源保持一致对外接口。

对于系统中已存在的 Schema，可通过关联查看对应 XML 示例模版内容，同时选择 Schema 名称也可查看 Schema 内容，即 XSD 文件内容。若各行业标准因行业发展产生版本变更，则可根据标准版本变动情况对 Schema 进行修改。

2.2 数据标准校验设置

2.2.1 数据规则灵活可配置

数据元规则表达式的解析和对数据的校验计算均来自行业扩展规则，此时涉及两方面问题，一个是解析，另一个是计算。

规则解析是为了将一个完整的规则表达式转换成其最小子规则，也就是系统只需实现每一个最小子规则动态可配置，即可实现完整规则表达式动态可配置。根据对医疗、水利、交通等行业的规则分析，最小子规则主要包含两部分内容：数据格式和数据长度。实现方式如下。

数据格式一般可通过正则表达式实现。

数据长度是计算后的结果。计算过程将涉及输入、计算函数和输出。输出是我们需要的结果，输入是该数据元配置 XPath 所指定的元素值，因此需给每个最小子规则配置计算函数并能动态执行即可实现规则的灵活可配置。

动态计算引擎一般和规则引擎技术关联起来，因此系统引入 Aviator 规则引擎。如一个简单的固定长度验证表达式：string.length(*s*) == long(*s*1)，Aviator 可动态转换上述表达式进行计算，返回计算结果，这里 *s* 和 *s*1 为传入参数，剩下的问题为如何获取参数。以 AN3 为例，*s* 参数为该数据元 XPath 所定位的元素值，*s*1 参数为“3”，因此实现获取 3 这个值即可。这里同样采用正则表达式来实现，以该表达式为例，提取 3 这个值的正则表达式为 ^AN((？=[^a-zA-Z])|(？=$))，同样正则表达式也可配置可动态修改，因此在规则部分即可实现灵活可配置。

2.2.2 数据规则执行

最小子规则实现灵活可配置，余下工作就是组合各个最小子规则。

这里引入链表的概念，一个规则表达式会被解析为多个最小子规则节点，这些节点首先按优先级高低排序并按顺序执行，同时每个节点有一个 next 标志位，next 为 true 则代表本节点执行完成后可继续执行下一节点，next 为 false 则代表本节点执行完成后结束整个校验流程。

测试执行流程图如图 1 所示。

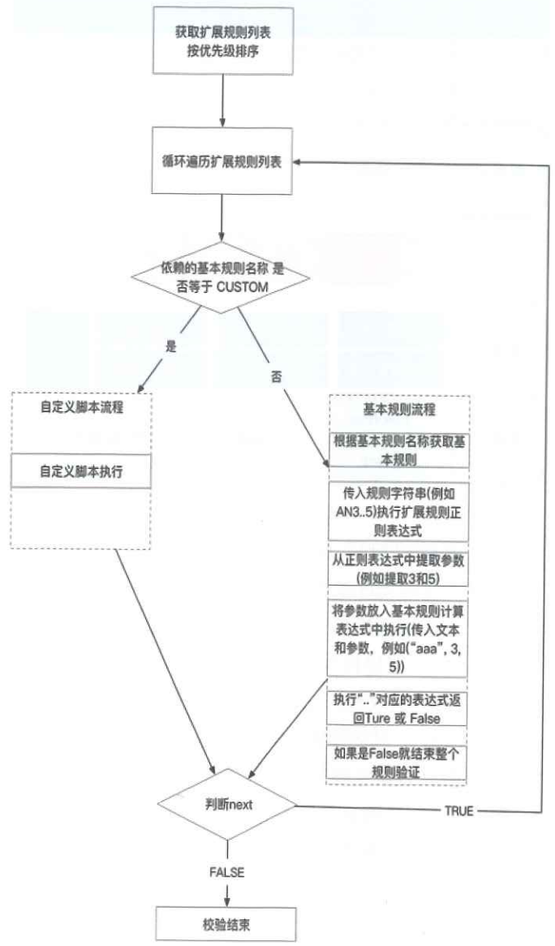


图1 测试执行流程图

2.2.3 获取数据元值

数据元中的规则表达式可动态解析和执行，值域有自己的配置表保存和查询，因此，目前最主要问题为如何将数据从 XML 文档中取出。XML 文档中数据定位一般使用 XPath 定位和查询元素，它是一门在 XML 文档中查找信息的语言，拥有自己的语法。因此为数据元配置 XPath 即可。

具体过程如下：

将数据元配置 XPath 路径（这里设计为可以有多个 XPath，同一元素可能在不同结构体中）；

根据 XPath 将 XML 文档节点 Node 取出；

获取 Node 的 Text，该 Text 即为要获取的数据；

收稿日期：2022-07-07

课题项目：2018 年工业互联网创新发展工程——汽车行业工业互联网平台试验测试项目（2018 年工业互联网创新发展工程）

作者简介：李真（1989—），女，满族，河北邯郸人，硕士研究生，工程师，研究方向：软件测试研究。

分别对 Text 数据进行规则校验和值域校验。

3. 系统技术架构

XML 解析及数据校验系统技术架构如图 2 所示，系统分为基础层、支撑层、应用层、展示层和用户层。

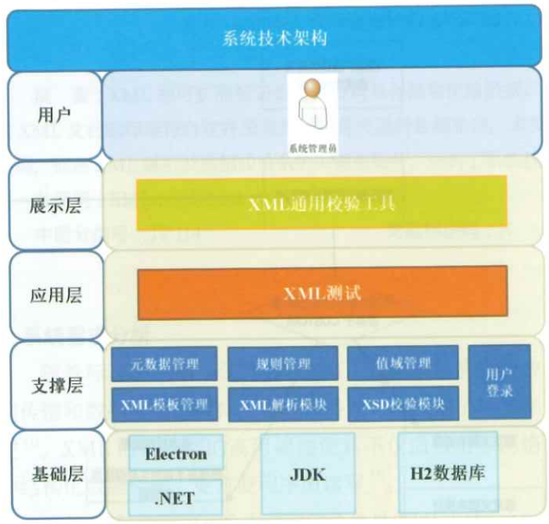


图2 系统技术架构

（1）基础层包括 .NET、JDK 和数据库等基础结构。主要实现底层数据库逻辑。

（2）支撑层包括 Service 层和 DAO 层。

DAO 层可直接操作数据库代码，该层仅负责使对应每个表完成增删改查。DAO 层同样需先创建 DAO 接口，再在配置文件中定义该接口的实现类，配置数据源和数据库连接参数。该层主要负责数据持久化存储。

Service 层负责管理具体功能实现，因此该层主要为一些实现具体业务功能类，称为 Business 类。Controller 层可以调用该层接口，处理业务逻辑应用。Service 层负责处理 Controller 层传递过来的数据，再将处理后数据传给 DAO 层，用于链接数据库，方便 DAO 层进行增删改查。同时，Service 层负责处理 DAO 层传递过来的数据，将其进行封装，如封装成 JavaBean。系统主要业务逻辑也在该层进行实现，我们改为先设计接口，再创建需要实现的类，然后在配置文件中配置实现的关联。这样封装好 Service 层业务逻辑并进一步划分，便于业务逻辑的独立和可再现，增加该层可维护性。

支撑层实现基础配置及管理功能，包括数据元管理、XML 模板管理、规则管理、值域管理和 XML 解析模块等。

（3）应用层为 Controller 层，即为控制器，负责管理业务调度和管理跳转。主要通过获取数据后调用 Service 层接口，包括 Service 层的处理业务逻辑，然后返回数据来实现控制具体流程的功能。如具体业务功能由何种类来实现，实现结果通过何种途径显示等，均由 Controller 层决定。同时 Controller 层需负责与展示层和 Service 层的通信，通信过程需借助一些 Bean 类进行信息传递。从控制层功能上来说，因为该系统业务逻辑并不复杂，所以该层代码编写并未如其他复杂业务逻辑的系统繁重和复杂。

应用层主要实现具体测试功能。

（4）展示层是与客户的交互层，主要接收用户提交的请求，将用户提交请求和数据传递给下一层，并将后台响应结果返回给客户层。

展示层为系统的展示界面。

本系统采用基于 C/S 模式开发，有如下特点。

1）服务端基于 Java 的 Springboot 框架开发，创建独立的 Spring 应用程序，并且直接嵌入 Tomcat，无需其他配置，主要负责业务逻辑处理。服务端优点是仅仅依赖 JDK 运行环境，不受外部操作系统环境影响，一次编译，多处运行。

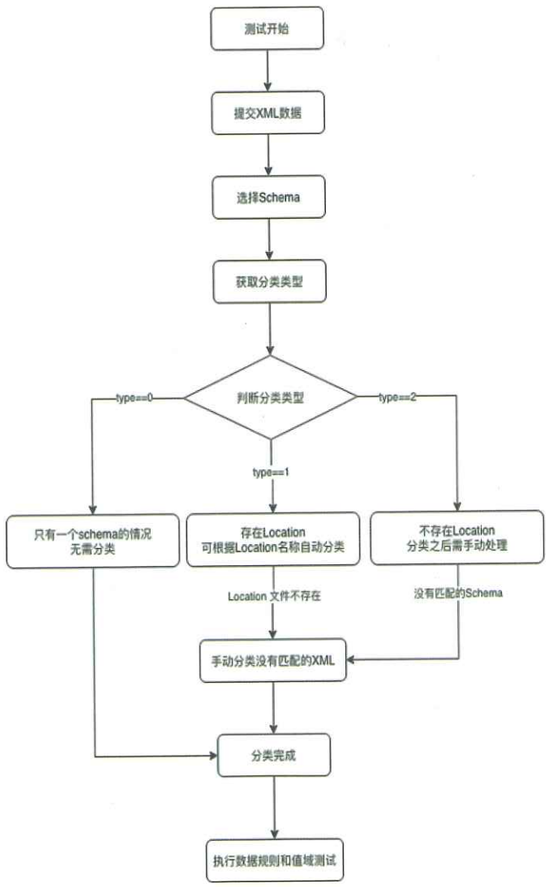


图3 XML文档校验测试流程图

2）客户端基于 Electron 框架开发，主要负责界面展示、控制服务端后台进程启停，优点是 Electron 打包之后的客户端本身就是一个浏览器，屏蔽客户机器由于使用不同浏览器而产生的兼容问题，同时还支持一键安装，部署过程和安装客户端软件同样便宜。

3）数据库 H2，相比其他数据库系统，H2 数据库轻便，仅有一个 jar 文件，无需安装，同时还可嵌入服务端一起打包发布，方便存储系统产生的结构化数据，随着 Java 服务端启动而启动，和服务端共用一个 JVM 内存空间。

4. 测试流程

XML 文档校验完整测试流程如图 3 所示，此处以提交医疗行业互联互通测评电子病历共享文档（电子病历共享文档规范 第 2 部分：门（急）诊病历）为例。

首先，需提交 XML 文档到 XML 解析及数据校验系统，在此需要说明的是 XML 文档结构必须符合 XML 格式，若不符合基本 XML 格式，将无法进入后续校验流程。

系统采用 Schema 对文档结构进行标准化校验，验证 XML 文档结构是否符合标准规定。在系统对一批 XML 文档进行 Schema 结构校验时，要求必须将每一个 XML 文档均指定对应的 Schema 文件，即需要基于 Schema 对 XML 文档进行分类校验。系统针对 XML 文档分类进行自动化处理，有如下两种情况：

（1）XML 文档头标签存在 Location 属性，可根据 Location 名称，自动匹配 Schema 文件。

（2）XML 文档头标签不存在 Location 属性，则自动分类失败，此时需要将该类文件手动指定一个 Schema 文件或引入新 Schema 文件，则该分类下所有 XML 文档均会自动归类到该 Schema 分类下。

具体流程如下：

（1）在系统中导入 / 提交门（急）诊病历 .xml 文档。

（2）点击下一步，进行 Schema 文件的选择。

（3）系统对提交的 XML 文档自动匹配相同种类 Schema 文件，即门（急）诊病历 Schema 文件。

（4）若此时自动匹配 Schema 文件名称有误，则需要选择系统中与该 XML 文档种类相同的 Schema 文件，进行手动分类。

（5）点击测试执行，系统会对提交的门（急）诊病历 .xml 文档进行测试，对文档中数据源进行校验。

（6）执行结束后，系统会返回执行结果，并提供导出结果功能。

参考文献

[1] 逯喜林.XML文档有效性验证系统[D].南京:南开大学,2011.

[2] 汪洋,徐建芬,王海平.基于XML的自动测试信息交换标准研究综述[J].电子测量与仪器学报.2008,22(5):1-6.

[3] 余双,曹东磊,戴蓓洁,等.高效XML验证技术的实现[J].计算机工程与设计,2008,29(4):937-941.

[4] 中华人民共和国国家卫生和计划生育委员会. 电子病历共享文档:WS/T 500-2016[S].卫生健康标准网,2016.

[5] 吴楠.XML语义验证算法的研究与应用[D].北京:首都经济贸易大学,2007.

[6] 中华人民共和国国家卫生和计划生育委员会.电子病历基本数据集:WS 445-2014[S].卫生健康标准网,2014.

[7] 中华人民共和国卫生部.卫生信息数据元目录:WS 363-2011[S].卫生健康标准网,2011.