**数据自治组件设计方案**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **文档信息** | | | | | | | | | |
| 主题词 | |  | | | | 文档版本 | | 2.0 | |
| 编写单位 | | 上海金仕达卫宁软件股份有限公司 | | | | | | | |
| 原作者 | | 马路 | | 文档类别 | | | | 技术文档 | |
| 审核 | |  | | 文档性质 | | | | 一般 | |
| 批准 | |  | | 发布日期 | | | |  | |
| 主送 | |  | | 存档日期 | | | |  | |
| 抄送 | |  | | 发行份数 | | | |  | |
| **签收信息** | | | | | | | | | |
| 提交单位 | |  | | 接收单位 | | | |  | |
| 提交人员 | |  | | 接收人员 | | | |  | |
| 提交日期 | |  | | 接收日期 | | | |  | |
| **文档变更纪录** | | | | | | | | | |
| 版本号 | 变更内容 | | 变更人 | | 变更日期 | | 确认人 | | 确认日期 |
| 1.0 | 创建文档 | | 马路 | |  | |  | |  |
|  |  | |  | |  | |  | |  |
|  |  | |  | |  | |  | |  |
|  |  | |  | |  | |  | |  |
|  |  | |  | |  | |  | |  |

**目录**

第1章 引言 3

1.1 编写目的 3

1.2 背景 3

1.3 术语定义。 3

1.4 参考资料 4

第2章 清洗维度框架 5

2.1 完整性 5

2.1.1 关联性 5

2.1.2 业务逻辑关系约束性 9

2.1.3 字段语义约束性 9

2.2 规范性 9

2.2.1 编码符合性 9

2.2.2 .非空性 10

2.2.3 合法性 10

2.3 适时性 10

第3章 规则(知识)库 11

第4章 算法库 12

第5章 推理引擎 13

第6章 清洗框架分析 14

6.1 整体架构 14

6.2 用户特点 15

6.3 假定和约束 16

第7章 技术框架分析 17

第8章 核心实现技术 18

第9章 整体功能分析 19

# 引言

## 背景

医疗临床大数据采集平台是医疗大数据平台的核心组件之一，主要负责采集，预处理，整合来自各个机构及各个平台海量的结构化和非结构化的临床数据。临床大数据中心数据根据不同类型的服务需求，需要提供实时或者离线的数据采集方式。数据采集平台全流程涵盖了数据准备阶段，检测阶段，定位阶段，修正阶段，验证阶段。对数据交换过程，校验过程，质量评估过程进行了端到端的跟踪和监管。产品具有以下特色

* 实现了端到端的全流程数据跟踪和监管。
* 采用了ＰＤＬＭＶ清洗框架清洗框架，在架构了保证了流程设计的合理性。
* 逻辑架构上采用了三段式工作驱动的方式，质量维度监控评估，不依赖于知识库的非规则性清洗，依赖于应用逻辑规则的规则清洗。三个维度按照层次分类，彼此独立，不相互依赖。
* 提供自定义规则的数据预处理清洗框架。
* 全路径排错功能，方便监管人员动态，全面的掌握数据校验的过程和结果。
* 分类汇总知识库管理体系，以保证各个现场不同的校验规则可以集中汇总，统一分享。

## 目标

随着区域信息平台信息化进程的不断深入，从各个医疗机构及相关平台采集的数据的规模越来越大，如何保证有质有量的数据进入到平台系统就成了首要的关键性问题，做到正确的时间，正确的地点将正确的数据提供给正确的人员来进行决策，执行业务则是数据质量交换监控系统首要完成的任务，也是本方案设计的初衷。

## 术语定义。

|  |  |  |
| --- | --- | --- |
| 序号 | 术语名称 | 术语定义 |
| 1 | EHR区域信息平台 | 连接区域内基本医疗系统数据交换和共享的平台。是不同系统间信息整合的基础和载体。 |
| 2 | 数据清洗 | 数据清洗是将数据库精简以除去重复记录，并将剩余部分转换成符合标准的过程，使数据实现准确性，完整性，一致性，适时性，有效性。以适应后续操作的过程，从提高数据质量的角度，凡是有助于提高数据质量的数据处理过程，都可以称之为数据清洗。 |
| 3 | 数据清洗框架 | 数据清洗框架是由准备，检测，定位，修正，验证5个部分组成，简称PDLMV。 |

## 参考资料

1. 基于健康档案区域信息平台建设技术解决方案
2. 数据清洗研究综述\_王曰芬
3. 大数据环境下面向知识服务的数据清洗研究\_蒋勋
4. 数据清洗及其一般性系统框架\_曹建军
5. 数据质量和数据清洗研究综述\_郭志懋

# 数据采集流程

## 交换接口方式

1. **实时批量数据采集接口(主动抓取)**

实时采集文档类型的数据，主要用于实时的业务数据的采集。交换技术采用flume + kafka +spark框架实现。

1. **离线批量数据采集接口(主动抓取，被动收集)**

离线两点之间发送批量数据，主要用于非实时数据传输方式，用户批量数据上传到数据中心或者从数据中心下载各类表格信息等业务场景。交换技术采用ETL中间件实现。

1. **服务接口方式**

用于向别的系统提供数据，接受别的系统发出的数据请求，返回本地的业务数据，通常用于为其它的系统提供服务接口或者已建系统的接口改造工作。

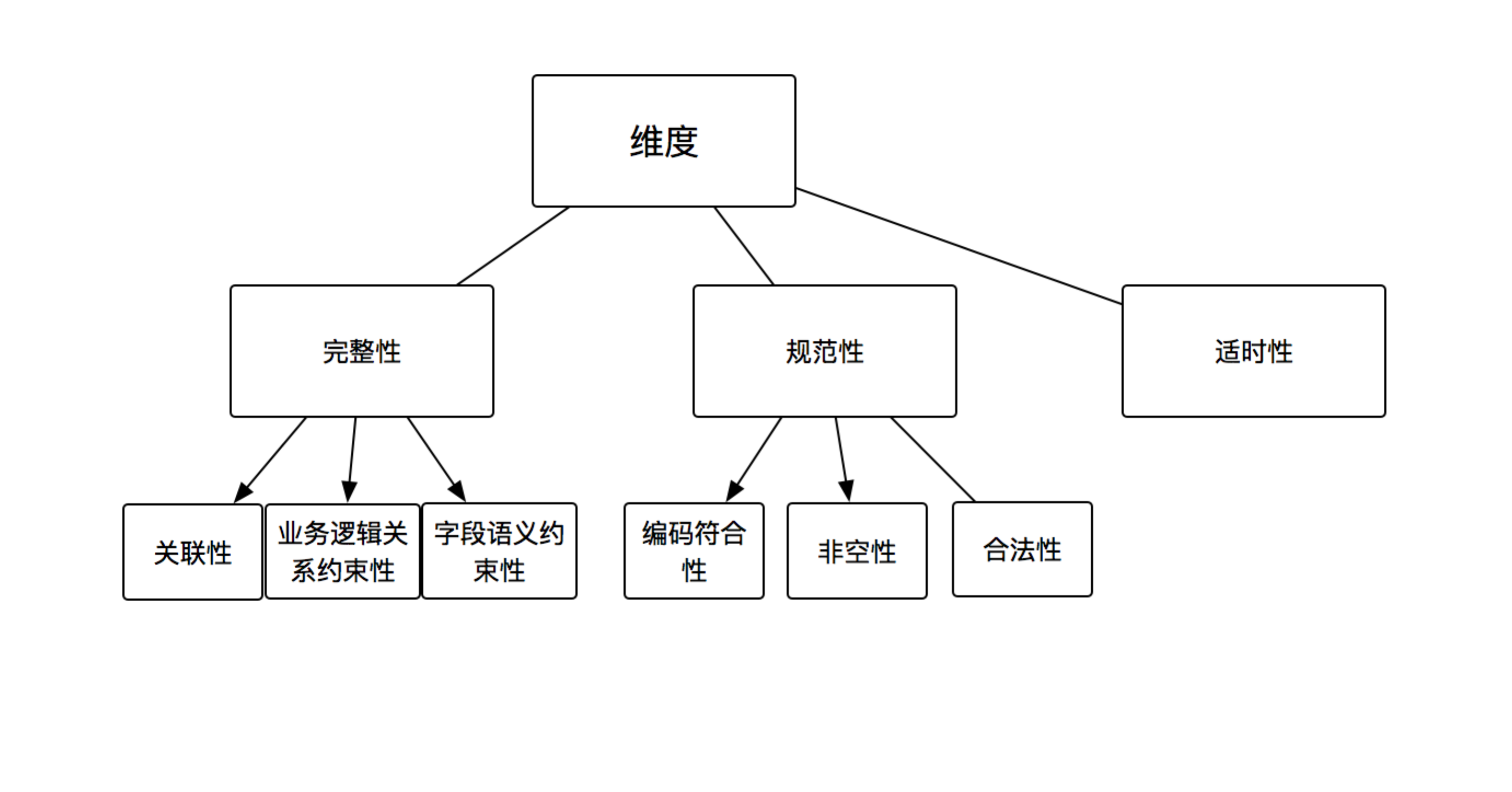
1. **发布订阅接口方式**

订阅某个主题的数据。当主题有信息的消息的时候，可以自动获取订阅信息。用于一对多的广播式数据纷发。通常用于各种监测服务，当监测指标发生变化的时候通知订阅系统。

## 数据交换技术方案介绍

## 数据交换模型

# 清洗维度框架



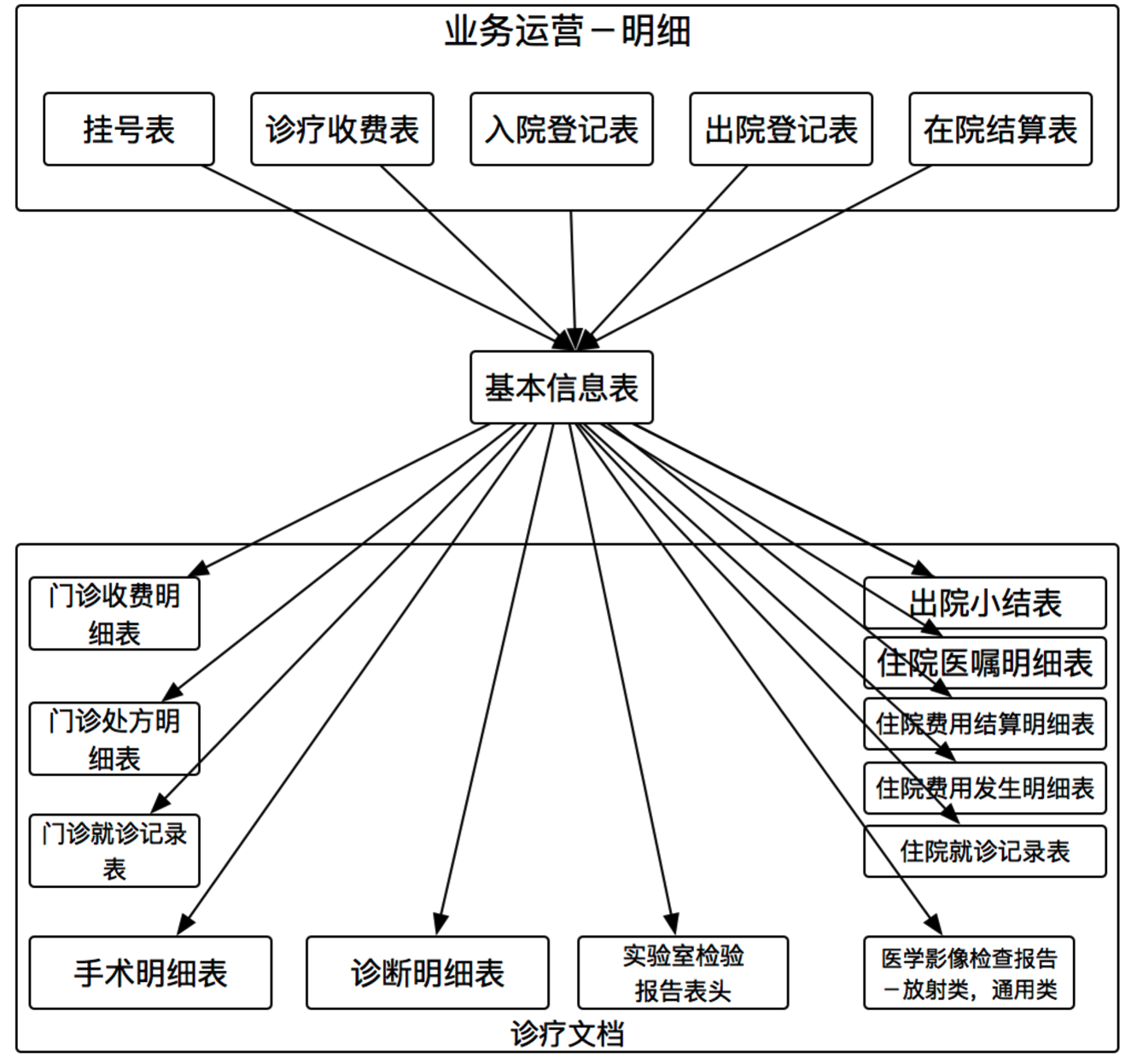
## 完整性

为了监控各个机构上传到平台数据完整性，提出了数据完整性的要求，完整性又分为关联性，业务约束性，统计约束性三个子纬度。

### 关联性

关联性源于参照完整性约束，业务逻辑上要求业务子事件报告一定可以追溯到业务父事件报告,在业务逻辑规则上，这是一种弱和强的依赖关系。隶属记录级别的错误。分为以下两种情况

1. 就诊报告(包含患者唯一标示)必须能够关联到患者信息表

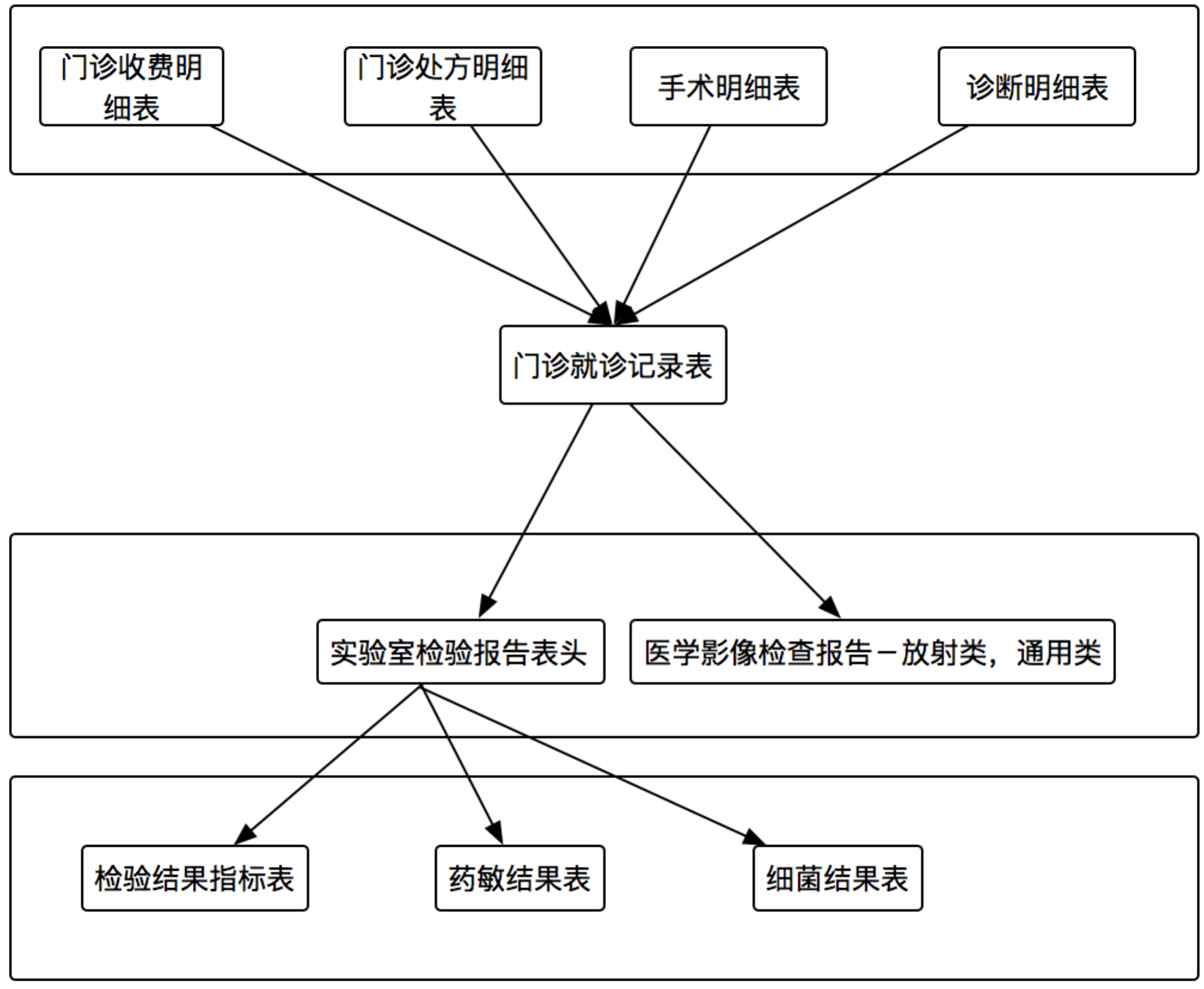


范例如下

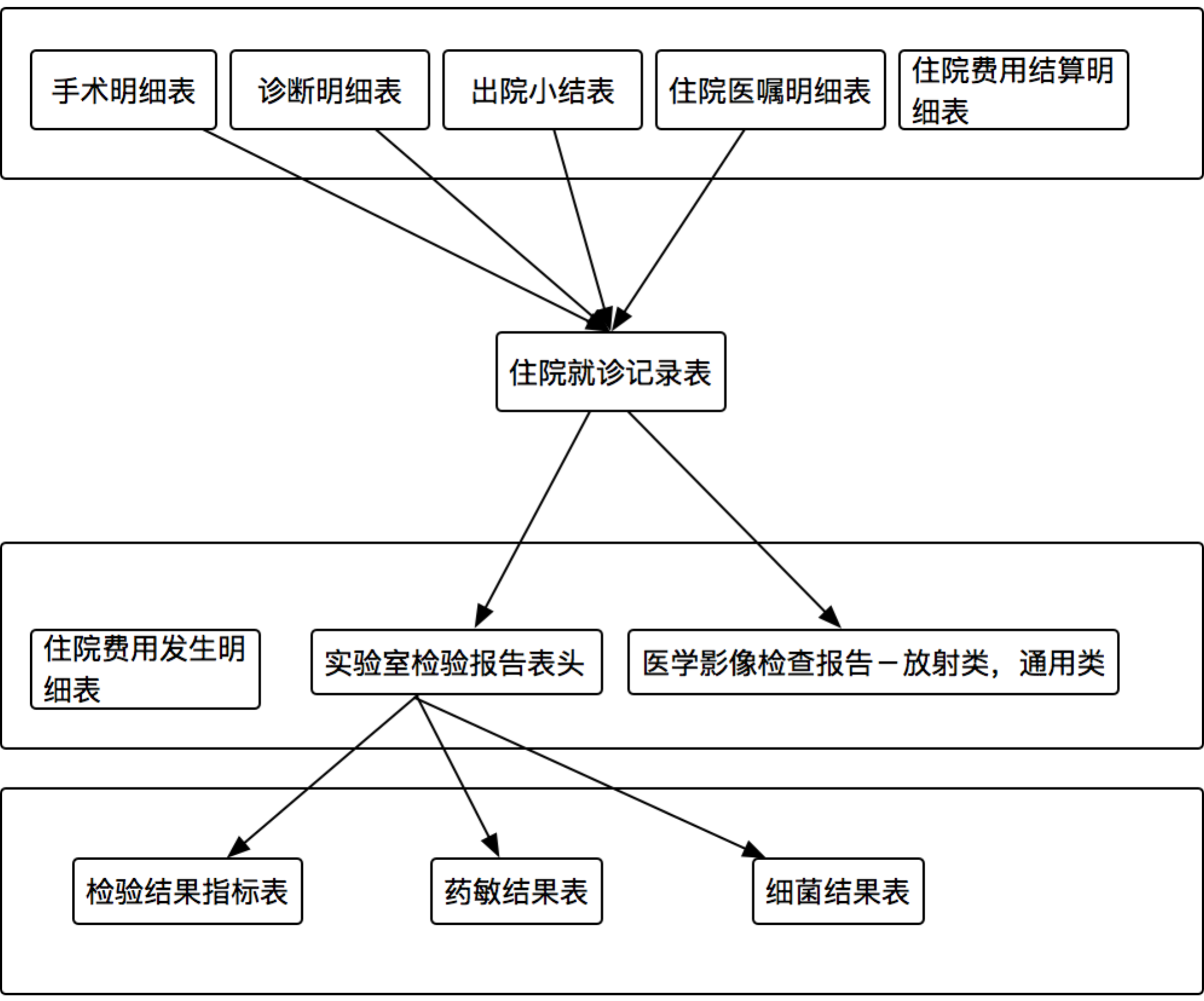
｀

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 完整性－关联性－1 | 10001 | 挂号表中患者唯一标示(就诊卡)是否能与患者信息表关联。 |
| …………… | ………… | ……….. |

1. 就诊子事件报告必须能够关联到就诊事件主报告



**门诊业务关联**



**住院业务关联**

范例如下

｀

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 完整性－关联性－n | 10301 | 已出院患者是否具有出院小结(出院患者-出院小结) |
| …………… | ………… | ……….. |

### 业务逻辑关系约束性

业务统计约束性源于模式层面的统计约束性,一致性。监测各区县平台/机构上传的统计指标是否与明细报告汇总值相符，提出了统计约束性要求。隶属记录级别的错误。统计约束性分为如下几类

1. 业务运营（明细报告）与业务运营（统计报告）是否符合。

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 完整性－统计约束性－1 | **20001** | 挂号表中统计出的门诊人次与业务量统计表中填报值是否一致 |

1. 诊疗文档与业务运营（明细报告）是否相符

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 完整性－统计约束性－n | **20101** | 门诊就诊记录表与挂号表各自统计出的门诊人次是否一致 |

1. 3. 医保内容与诊疗文档是否相符

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 完整性－统计约束性－n | **20201** | 医保文件中统计出的门急诊医保医疗收入与业务收入统计表中填报值是否一致 |

### 字段语义约束性

字段语义约束性源于模式层面的函数依赖性。隶属字段级别的错误。

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 完整性－字段语义约束性－表名－1 |  | 如果卡类型为医保卡，那么卡号的长度应该为12. |

## 规范性

为更好的满足基础服务、统计分析等数据利用需求，区县平台/机构上传的数据需满足接口中定义的数据规范（准确）性要求。依据接口表单是否存在业务时间字段，将规范（准确）性监测进一步细分为基础信息监测和业务信息监测两类。隶属于数据源级别错误。规则性错误分为一下几类

### 编码符合性

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 规范性－编码符合－表名－1 | **31001** | 修改标志（XGBZ）编码符合率 |
| …………… | ………… | ……….. |

### .非空性

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 规范性－非空性－表名－1 |  | 药品名称字段值不能为空。 |
| …………… | ………… | ……….. |

### 合法性

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 规范性－合法性－表名－1 |  | 调价日期超出合法日期范围。 |
| …………… | ………… | ……….. |

## 适时性

为更好的满足基础服务、统计分析等数据利用需求，区县平台/机构上传的数据需在业务数据产生后及时上传。

适时性监测主要记录各机构上传的业务表单数据中，最后一次上传时间与业务发生时间的日期差距。日期差不足1天记为1天，大于或者等于1天截取整数部分。

|  |  |  |
| --- | --- | --- |
| 规则编号 | 别名 | 规则含义 |
| 适时性－1 | 40001 | 业务量统计表(TB\_STAT\_YWL\_Report)上传及时性 |
| …………… | ………… | ……….. |

# 规则(知识)库

# 清洗功能及算法

# 推理引擎

# 清洗框架分析

## 整体架构



1. 准备阶段

第一阶段的主要任务是获取清洗任务，

包括需求分析、信息环境分析、规则定义、算法定义以及基于以上工作获得数据清洗方案等。通过需求分析明确信息系统的数据清洗需求，通过信息环境分析明确数据所处的信息环境特点，任务定义要明确具体的数据清洗任务目标，方法定义确定合适的数据清洗方法，基本配置完成数据接口等的配置，要形成完整的数据清洗方案，并整理归档。确定数据清洗对象。

# 技术框架分析

# 核心实现技术

# 整体功能分析