|  |
| --- |
| 浙江大学CCNT实验室 |
| 中医药知识服务平台 |
| 项目实施方案 |

|  |
| --- |
| 于彤  2013/6/6 |

目录

[要点 3](#_Toc358662061)

[实施计划 4](#_Toc358662062)

[中医药知识服务平台描述 5](#_Toc358662063)

[中医药科学数据库 5](#_Toc358662064)

[中医药术语系统 6](#_Toc358662065)

[TCMLS 6](#_Toc358662066)

[TCMCT 6](#_Toc358662067)

[中医药文本挖掘工具集 6](#_Toc358662068)

[DartGrid工具集 6](#_Toc358662069)

[DartOnto 6](#_Toc358662070)

[DartAPI 6](#_Toc358662071)

[DartMapping 7](#_Toc358662072)

[DartQuery 7](#_Toc358662073)

[DartReasoner 8](#_Toc358662074)

[DartSearch 8](#_Toc358662075)

[DartWiki 8](#_Toc358662076)

[DartSpora 9](#_Toc358662077)

[DartMashup 9](#_Toc358662078)

[参考文献 9](#_Toc358662079)

# 要点

1. 在2013年6月底搭建起一个中医药知识服务平台。
2. 可将DartWiki和TCMSearch作为恢复中医药知识服务平台的重点。
3. 为了恢复DartWiki和TCMSearch这两个系统，还需要测试和恢复查询重写引擎DartCore（或者叫DartGrid）。
4. 为了恢复TCMSearch以及其他的系统，还需要测试和恢复我们原来开发的文本挖掘工具，尤其是其中的分词工具和概念识别工具（另外，基于文本的语义关系发现工具也是很有价值的）。
5. 需要搭建一个Oracle数据库，将这两个系统所需的后台数据导进去；最好这两个系统能够支持Access数据库，这样便于在笔记本电脑上演示（或搭建其他类型的数据库，将相关数据转换后导入）。
6. 我（于彤）会在中医科学院搭建一个开发环境。
7. 我（于彤）会将代码上传到Github上。
8. 我（于彤）会调研一下需要哪些数据库，将对应的数据导出。
9. 在浙大那边是否还有中医科学院的数据？是以什么形式存储的（Oracle数据库？）？
10. 是否可以在GitHub的TCM-ZJU上再建一个项目,用于管理中医药数据库的说明文档以及一些小型的数据库和数据表(如Access,RDF,Excel等)？
11. 如何创建一个类似于TCM-ZJU的组织？其中的

# 实施计划

在2013年6月底恢复DartWiki和TCMSearch的运行，搭建起一个中医药知识服务平台。

|  |  |  |
| --- | --- | --- |
| 时间 | 主要任务 | 备注 |
| 2013/6/10-  2013/6/16 | 1. 上传DartWiki 2. 上传TCMSearch 3. 阅读陶金火的程序 | 在北京 |
| 2013/6/17-  2013/6/23 | 1. Flex开发 2. GWT开发 3. 讨论陶金火的程序 | 在杭州 |
| 2013/6/24-  2013/6/30 | 1. 完成数据导入 2. 恢复DartWiki运行 3. 恢复TCMSearch运行 | 在北京 |

# 中医药知识服务平台描述

中医药知识服务平台的核心组件以及它们之间的依赖关系如图 1所示。该平台的核心技术是DartGrid工具集以及中医药文本挖掘工具集，其数据资源包括浙大和信息所合作开发的文献库、药理库等中医药科学数据库，以及TCMLS等术语系统。

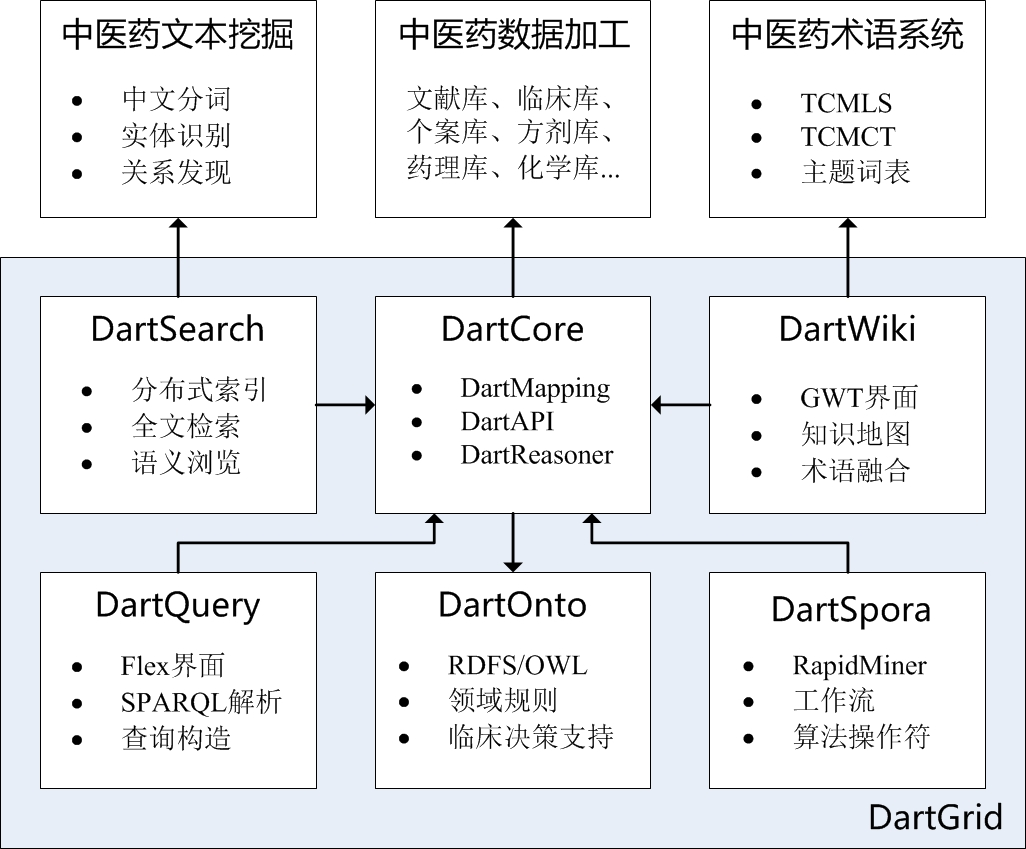


图 1 中医药知识服务平台的核心组件以及它们之间的依赖关系

下面对这些组件进行简要介绍。

## 中医药科学数据库

## 中医药术语系统

### TCMLS

### TCMCT

## 中医药文本挖掘工具集

陶金火（2011）实现了基于TCMLS的中文分词和语义关系发现方法。

## DartGrid工具集

数据库网格中包含大量高度异构、广泛分布、自治管理的数据库，这就需要数据库工具和中间件能够在语义层次上处理模式调节和查询处理。浙江大学研究小组自主研发一个成熟的数据驱动的网格计算平台DartGrid，它能够有效地整合大规模的异质异构数据，实现了网格环境下数据资源的语义集成、动态管理和分布式查询。该系统利用了资源描述框架（RDF）来实现一组语义工具和网格服务，包括语义浏览器、语义映射工具、本体服务、语义查询服务和语义注册服务。目前，该研究成果已经应用于北京中国中医研究院的中医药科技数据库群管理和浙江省公安厅的科技信息管理，并取得了一定的社会经济效应。下面介绍DartGrid 3.0版本的核心组件和模块。

### DartOnto

DartOnto是基于Web框架（如Spring等）构建的本体协同加工组件，在分布式加工，性能优化以及中文支持等问题上提供重点支持（宓金华，2010）。DartOnto还支持多术语系统的融合功能。

#### 功能

1. RDFS本体的协作开发。

#### 核心技术

1. Flex技术
2. Java语言

#### 评价

1. 价值中等：其功能可被Protégé等工具替代。主要支持RDFS,尚未提供对OWL的支持。
2. 代码质量较高。

### DartAPI

DartAPI为解决广域环境下的大规模科学数据共享问题，并重点针对中医药应用领域而设计；它实现了一个基于语义的虚拟组织模型，支持在广域环境下数据库的动态化的语义注册、分布式的语义查询与语义浏览。该组件包括如下服务：语义查询服务；语义注册服务；查询分配服务；数据库注册服务；以及资源层数据库服务。

### DartMapping

DartMapping是辅助用户编辑语义映射的可视化工具，其基本功能是建立关系数据库模式与本体模式之间的映射（王恒，2006）。具体而言，就是数据库中表与本体、表中字段与本体属性之间映射关系的编辑，并且使用图形化技术实现整个编辑过程 。语义映射可视化编辑工具除了要完成以上的基本功能，还要根据内核应用添加一个智能化的新功能点。语义映射可视化编辑工具可以图形化显示关系数据库中表、表中字段的信息，还可以显示关系数据库中表之间的外键关系，我们就可以更根据表之间的外键关系与本体模式之间的相关性来进行智能化的推测和将这种推测结果写入xml文件，以便于基于语义的数据库集成系统核心部分更加方便的应用。

### DartQuery

DartQuery是基于分布式数据库实现综合性语义查询的引擎，其基本原理是基于语义映射实现从SPARQL到SQL的查询重写（唐晶明，2007）。该系统相较传统的查询系统，具有如下技术特点：利用本体中定义的同义词或异名词对检索进行扩展；利用语义本体中所定义的词汇作为查询检索的辅助词汇；支持相关概念检索：依据语义本体中定义的与当前检索词的语义关联关系，提供给用户一组相关概念与语义关联词表，用户可进一步扩展查询和查找相关信息；集成异构数据库模式：语义本体还承担着屏蔽数据模式异质异构性的作用。在该系统中，所有的数据依然存储在传统的关系型数据库中，所有的关系数据库表都通过一个语义映射工具映射到语义本体中的相关类上。在搜索到结果时，所有的结果集都依据语义映射翻译成RDF/OWL的格式。语义查询被定义到语义本体上，这一语义查询将被系统进一步翻译成一组关系数据库查询来完成实际的查询计算。

#### 功能

1. SPARQL查询构造界面
2. SPARQL查询处理
3. 从SPARQL到SQL的查询重写

#### 核心技术

1. SPARQL技术
2. Java语言

#### 评价

1. 价值高
2. 代码难度高

### DartReasoner

郑耀文（2012）基于MapReduce框架实现了面向海量数据的语义推理引擎DartReasoner。

### DartSearch

DartSearch为基于基于DartGrid基础框架开发的全文语义搜索系统（付志宏，2010；杨克特，2010）。DartSearch作为一个全文检索系统，在中文分词、索引、排序的等方面结合关系数据库的特点，已经做了不少有益的尝试。尤其是为面向互联网环境下异质异构数据库的集成查询提供一种实际可行的、卓有成效的解决方案。此外，DartSearch还逐渐积累了一些相关构件工具模块，如CNKI关联检索、语义相关词、VML语义图工具包、相关图文聚合工具等。

#### 功能

1. 中医药文献的分布式索引
2. 中医药文献的语义搜索

#### 核心技术

1. 中文分词
2. 分布式索引
3. 语义排序

#### 评价

1. 价值高。
2. 代码质量高。

### DartWiki

DartWiki是基于语义Web的Wiki工具，它除了支持文档的编辑之外，还支持结构性知识的共享（宓金华，2010）。DartWiki的前身是“中医药多术语融合平台”，该平台集成了“中医药学语言系统”，“中医临床术语集”，“中医药学主题词表”和“中医药学名词”等四个中医药研究领域的计算机数据源。

#### 功能

1. 多个中医药术语系统的融合。
2. 中医药概念信息的展示。
3. 知识地图：实现语义网络的可视化浏览。

#### 核心技术

1. Google Web Toolkit

#### 总体评价

1. 价值高。
2. 代码质量较高。

#### 依赖关系

1. 依赖于DartCore的数据库集成功能。
2. 依赖于TCMLS等中医药术语系统。

#### 恢复计划

为了恢复DartWiki，需要掌握Java Web开发方法，以及Google Web Toolkit工具，还需要先恢复DartCore的数据库集成功能，所需数据包括TCMLS等4个中医药术语系统。

### DartSpora

DartSpora是一个面向中医药领域的数据挖掘平台（吴毅挺，2008；秘中凯，2010）。DartSpora基于AJAX技术以及Rapid Miner等开源项目实现，并提供了针对语义Web数据的数据挖掘能力。刘洋（2010）基于MapReduce实现了并行数据挖掘服务，从而保证了面向语义Web数据进行知识发现的效率。

#### 功能

1. 多个中医药术语系统的融合。
2. 中医药概念信息的展示。
3. 知识地图：实现语义网络的可视化浏览。

#### 核心技术

1. Rapid Miner

#### 评价

1. 价值高。
2. 代码质量较高。

#### 依赖关系

1. 依赖于中医药科学数据库。

### DartMashup

DartMashup是基于语义Web服务的混搭（Mashup）工具（卢宾，2010；王俊健，2010）。它支持用户通过浏览器以可视化的方式来定制混搭应用，并将这些应用在Web上发布 。DartMashup在中医药领域获得了成功的应用，协助中医利用各种数据源来快速创建更加丰富的应用，从而满足个性化的需求（王俊健，2010）。

# 参考文献

范宽. 2006. 基于语义本体的中医药科学数据共建工程[D]. 浙江大学计算机科学与技术学院.

冯叶磊. 2011. 基于稳语义的中医药文献搜索引擎[D]. 浙江大学计算机科学与技术学院.

付志宏. 2010. 面向中医药的分布式语义搜索系统[D].浙江大学计算机科学与技术学院.

方青. 2004. 基于本体论的中医药一体化语言系统[D]. 浙江大学计算机科学与技术学院.

景鲲. 2006. 基于语义的中医药专家综合决策系统DartSurvey[D]. 浙江大学计算机科学与技术学院.

梁欣颖2012. 基于OWL的中医理论语义模型及应用[D]. 浙江大学计算机科学与技术学院.

刘洋. 2010. 基于MapReduce的中医药并行数据挖掘服务[D].浙江大学计算机科学与技术学院.

卢宾. 2010. 基于语义的混搭系统[D]. 浙江大学计算机科学与技术学院.

宓金华. 2010. 中医药知识工程应用[D]. 浙江大学计算机科学与技术学院.

秘中凯. 2010. 中医药数据挖掘平台与服务[D]. 浙江大学计算机科学与技术学院.

盛浩. 2011. 基于超链数据的中医药语义查询系统[D]. 浙江大学计算机科学与技术学院.

汤萌芽. 2007. 中医药本体工程及相关应用[D]. 浙江大学计算机科学与技术学院.

唐晶明. 2007. 基于SPARQL的分布式语义查询处理[D]. 浙江大学计算机科学与技术学院.

陶金火. 2011. 基于语义的中医药数据采集工程及应用平台[D]. 浙江大学计算机科学与技术学院.

王超. 2012. 基于知识集成引擎的中医药知识服务社区[D]. 浙江大学计算机科学与技术学院.

王恒. 2006. DartGridV3语义数据库网格的设计与核心模块实现[D]. 浙江大学计算机科学与技术学院.

王俊健. 2010. 中医药共享平台与Mashup[D]. 浙江大学计算机科学与技术学院.

吴毅挺. 2008. DartSpora数据挖掘平台的构建[D]. 浙江大学计算机科学与技术学院.

吴振宇. 2007. 基于语义标注的中医药数据加工平台[D]. 浙江大学计算机科学与技术学院.

谢骋超. 2006. 基于语义的数据库全文检索系统[D]. 浙江大学计算机科学与技术学院.

杨克特. 2010. 面向中医药的多元语义搜索引擎[D]. 浙江大学计算机科学与技术学院.

于彤. 2006. 中医临床数据仓库的设计与构建[D]. 浙江大学计算机科学与技术学院.

张慧敏. 2008. 面向领域的关系数据库全文检索系统的优化设计[D]. 浙江大学计算机科学与技术学院.

张小刚. 2010. 基于中医药本体的语义关系发现及验证方法[D]. 浙江大学计算机科学与技术学院.

郑清照. 2010. 基于Linked Open Data的语义关联发现及其应用[D]. 浙江大学计算机科学与技术学院.

郑耀文. 2012. DartReasoner面向海量数据的语义推理[D]. 浙江大学计算机科学与技术学院.