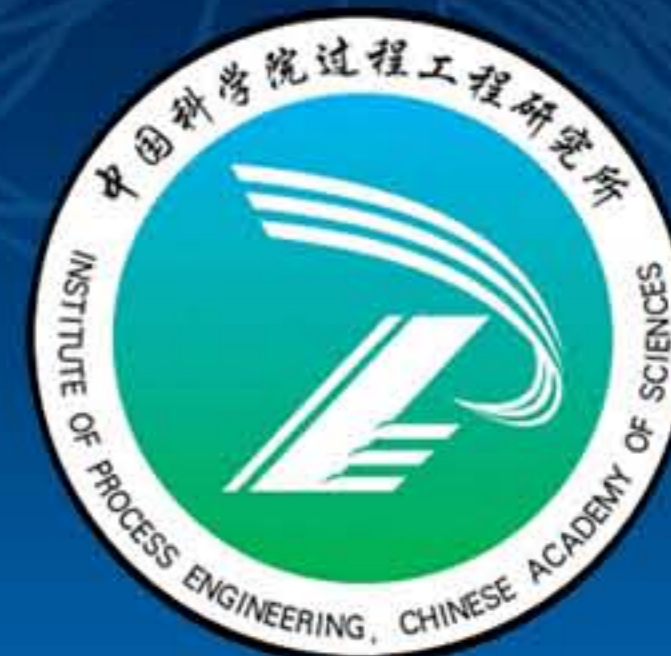




多相复杂系统国家重点实验室

STATE KEY LABORATORY OF MULTIPHASE COMPLEX SYSTEMS



化学深层网搜索引擎

Search Engine for Chemistry Deep Web

化学深层网是分布在网上的各种Web化学数据库的集合，基于链接分析的搜索引擎不能检索这类数据。化学深层网搜索引擎的目标是对这类数据建立索引、实现网络多来源化学数据库的统一检索，以回答两个看似简单、却非常有挑战性的问题：(1) 某个化合物在哪些化学数据库中存在？(2) 该化合物的某项数据存在于哪些化学数据库中？

近年来国际上也在致力于网络多来源化学数据库的统一检索，采用的方法是不同来源的数据库自愿向一个权威（中心）站点提交各自的化合物（结构）索引，由中心站点生成连接这些数据库的化合物索引。最具影响力的是NIH的PubChem。它可以回答问题(1)。

本实验室对化学深层网化学数据的挖掘采用了不同的思路，即自动将一个查询请求提交到分布在网络上的不同化学数据库站点，接收各库返回的检索结果页面，对结果页面中包含的数据进行自动提取，从而解决多来源化学数据库的统一检索问题。这一思路在回答问题(1)的同时，可部分地回答问题(2)。建立的化学深层网搜索引擎**ChemDB Portal** (<http://www.chemdb-portal.cn/>)已上线运行。

Goal - to answer the following questions

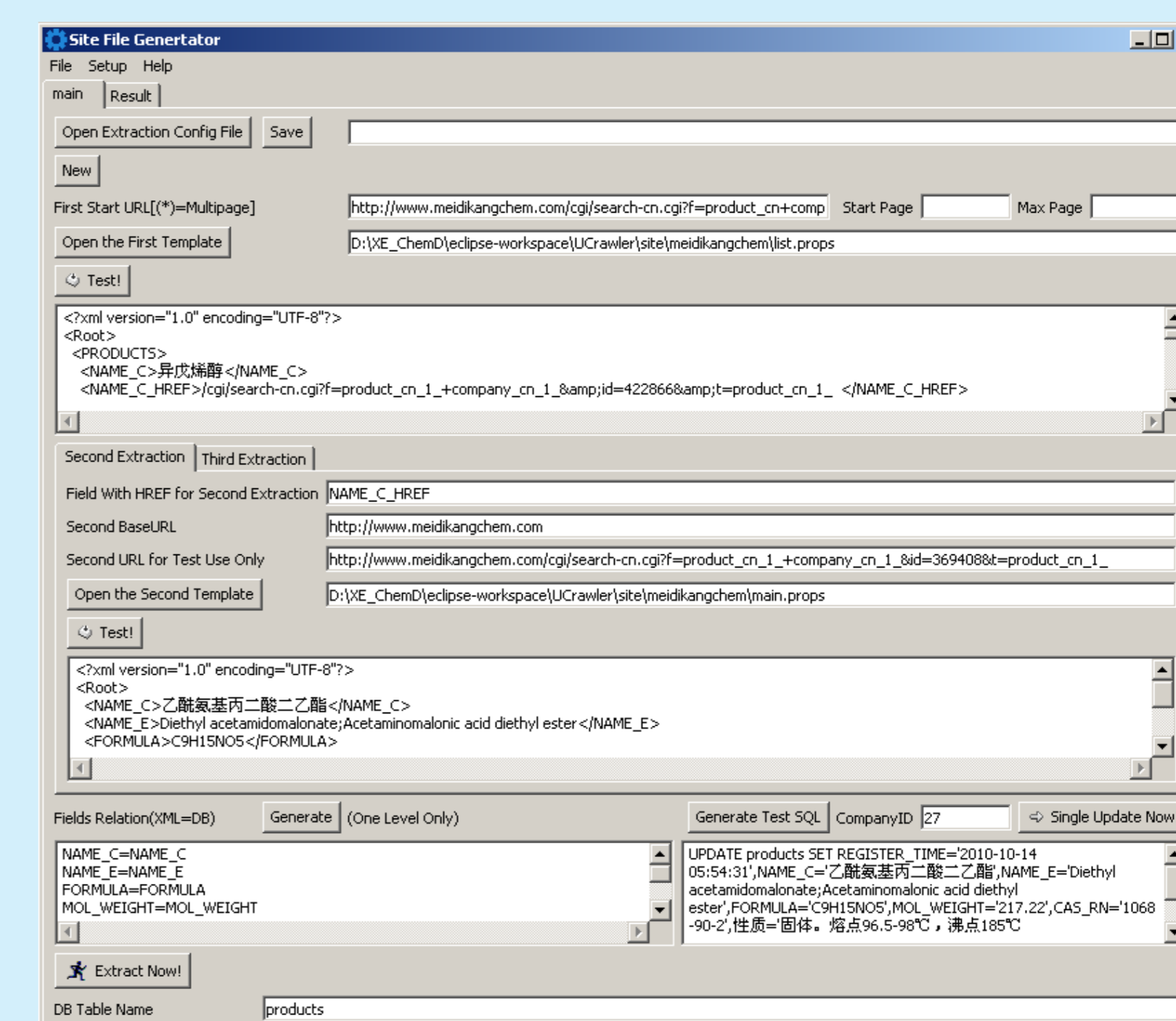
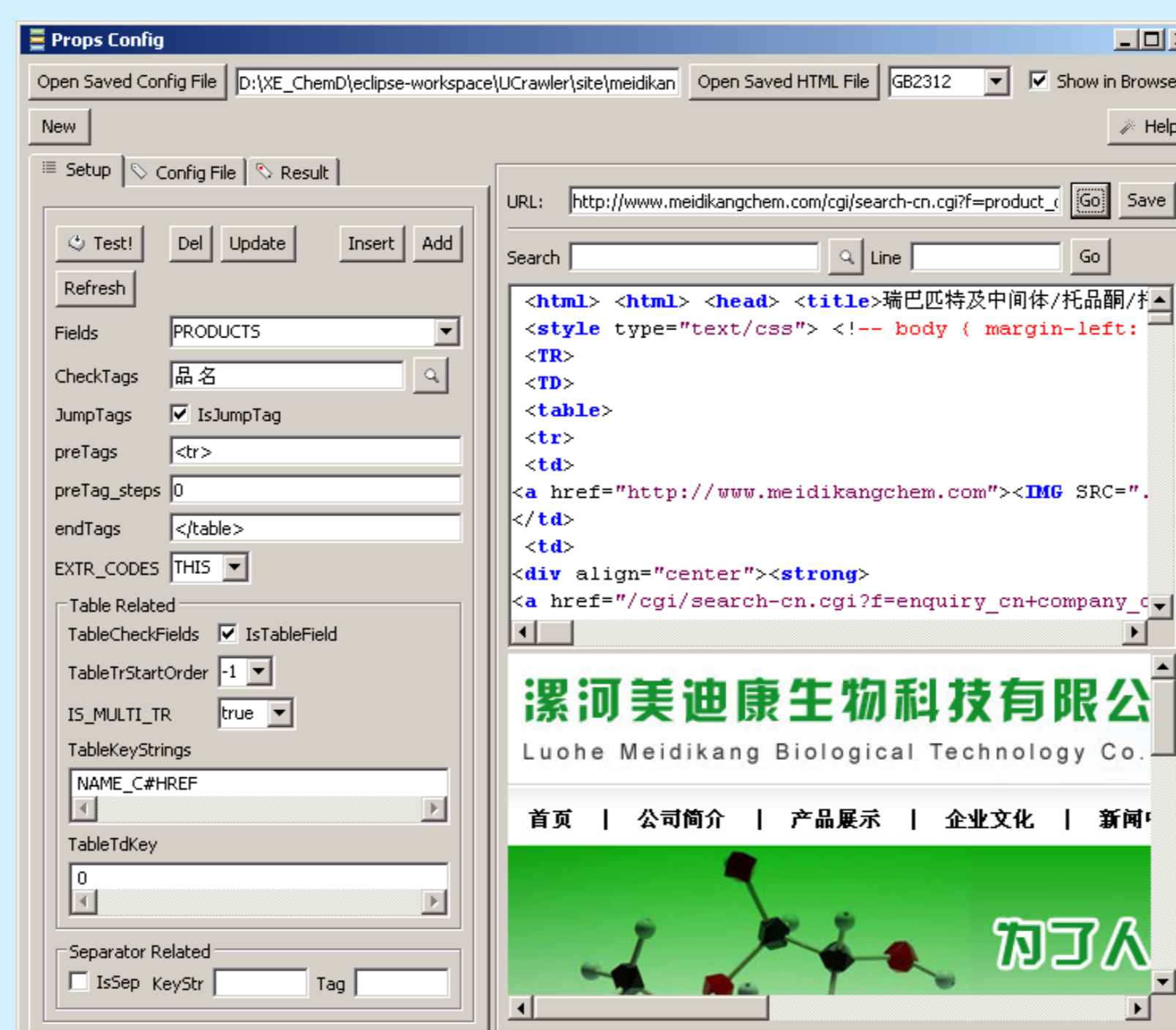
- Q1. Is a chemical in a Web DB?
- Q2. Any data of the chemical in a Web DB?

Repository Approach: Can answer Q1 but Q2

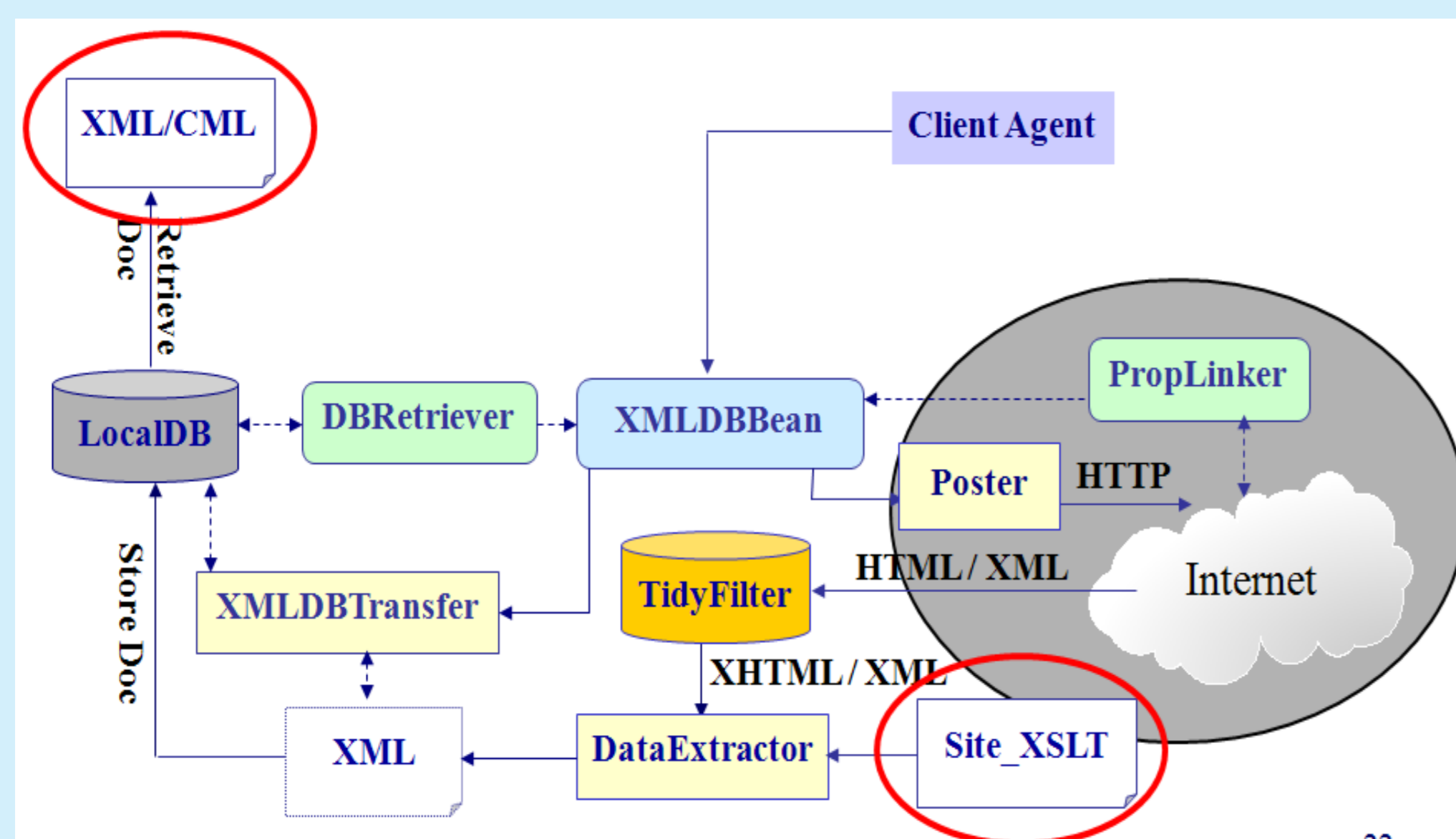
Method: Center site + index submitted by various DBs
The best: PubChem (2005 -, NIH)
Limitation: Submitted DB index required

Our Approach: Can answer Q1 and partly answer Q2

Method: Automated search multi-source DBs + Data extraction
Problems: BD index not known, data structure may be lost
Challenges: Semi structural data extraction? Scalability if remote DB changes?



基于局部特征的移动窗口式数据提取模版生成工具
 Flexible slide window based data extraction template generator



基于XML的化学深层网数据提取
 XML based data extraction for chemistry Deep Web



ChemDB Portal (<http://www.chemdb-portal.cn/>) - 化学深层网搜索引擎
 (a chemistry Deep Web search engine for substance data)

ChemDB Portal data extraction template generator

Data transformation by data extraction template

DB search result pages (HTML)

Extracted data in XML

Semi-auto template generator



联系人: 李晓霞

E-mail: xxia@home.ipe.ac.cn