6-15-2010

Semantically Annotated RESTful Services for Large-scale Metabolomics Data Analysis

Ashwin Manjunatha  
*Wright State University - Main Campus*

Paul E. Anderson  
*Wright State University - Main Campus*

Satya S. Sahoo  
*Wright State University - Main Campus*

Ajith H. Ranabahu  
*Wright State University - Main Campus*

Michael L. Raymer  
*Wright State University - Main Campus, michael.raymer@wright.edu*

*See next page for additional authors*

Follow this and additional works at: [https://corescholar.libraries.wright.edu/knoesis](https://corescholar.libraries.wright.edu/knoesis)

Part of the [Bioinformatics Commons](https://corescholar.libraries.wright.edu/knoesis/), [Communication Technology and New Media Commons](https://corescholar.libraries.wright.edu/knoesis/), [Databases and Information Systems Commons](https://corescholar.libraries.wright.edu/knoesis/), [OS and Networks Commons](https://corescholar.libraries.wright.edu/knoesis/), and the [Science and Technology Studies Commons](https://corescholar.libraries.wright.edu/knoesis/)

Repository Citation

[https://corescholar.libraries.wright.edu/knoesis/96](https://corescholar.libraries.wright.edu/knoesis/96)

This Presentation is brought to you for free and open access by the The Ohio Center of Excellence in Knowledge-Enabled Computing (Kno.e.sis) at CORE Scholar. It has been accepted for inclusion in Kno.e.sis Publications by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu, library-corescholar@wright.edu.
## 1. Introduction

### Metabolomics

The term metabolomics is defined as a comprehensive analysis in which metabolites of a biological system are identified and quantified. Any technique that can quantify metabolites can be used for metabolomics, but there are two primary techniques seen in the literature: nuclear magnetic resonance (NMR) and mass spectrometry with a prior on-line separation step such as high-performance liquid chromatography (HPLC) or gas chromatography (GC).

While neither technique is strictly superior, each technique has its own advantages and disadvantages. Existing applications include the identification of biomarkers associated with responses to toxins and pathophysiological changes, associated with toxic exposures, large scale sample classification and the study of genetic disorders.

### NMR Spectrometer

NMR spectroscopy is an experimental technique that exploits the properties of an atom's nucleus. It can be used to obtain information about the concentration and structure of molecules. NMR studies magnetic nuclei by applying a static magnetic field followed by applying a second oscillating magnetic field. Specifically, only nuclei with an odd number of protons or neutrons can be measured using NMR; however, the two most common atoms studied using NMR; however, the two most common atoms studied are H and 13C.

### Toxicology

Toxicology is the branch of pharmacology that deals with poisons and their effects on plant, animal and human life. The term metabolomics is defined as a comprehensive analysis in which metabolites of a biological system are identified and quantified.

## 2. What is the problem?

- **Large Data sets**
- **Standard post-instrumental processing**
- **Quantification of spectral features**
- **Normalization**
- **Scaling**
- **Multivariate statistical modeling**
- **All computationally intensive processes**
- **Variety of algorithms for each step**

### Need a robust and flexible analysis platform

- **Move to a Service based Architecture!**
  - Provide Web Services for each algorithm
  - Assemble workflows as required!

### Taverna – an open source family of tools for designing and executing workflows

http://www.taverna.org.uk/

## 3. How about Scalability?

### Computing Cloud

Shared hardware resources, software and information are provided to computers and other devices on demand.

- **Many vendors**
- **Uses the map-reduce computational paradigm**
- **Runs off Computing Clouds**

### Hadoop

An open-source software framework for reliable, scalable, distributed computing.

- [http://hadoop.apache.org](http://hadoop.apache.org)

## 4. Annotation and SA-REST

### SA-REST

W3C member submission on Semantic Annotation of RESTful services [3].

### Three basic properties

**domain-rel**: mark the top level domain of a document e.g. Nucleotides

**sem-rel**: mark the domain of a linked document

**sem-class**: mark the meaning of a selected word

## 5. Advantages of Annotation

### Faceted Search

Technique for accessing a collection of information represented using a faceted classification, allowing users to explore by filtering available information. When annotated with richer models, the indexing software can easily create faceted indexes to support a fine grained search. Even the regular keyword search can be improved.

1. **Query by concept – not by keyword**
   - Search for “NCI:FASTA” instead just FASTA. Yields documents that indicate the term FASTA as defined by the NCI Thesaurus.

2. **Filter by multiple facets**
   - Issue queries indicating many facets, say “type: soap binding:java include:NCI:FASTA” to look for service descriptions that are SOAP services with Java bindings including mentions about NCI:FASTA.

### Semi-Automated Composition

When service interface documents are annotated service composition can be done more **intelligently**.

1. **A composition tool can warn the creator of incompatible connections**: Output of Service A cannot be input to Service B!

2. **Supplement transformations by suggesting matching elements**: Create transformations or suggest the difficulty of transformation to the human (see **Mediatability(1)**)

## 6. What is the bottom line for the Biologist?

### Better Search for Biological Web Services

Services can be searched with more precise terms and concepts. Search by ontology concept and add facets to make precise filtering.

### Conveniences in Creating Workflows

Find and mash services together with ease. The tools can suggest the degree of match and also create data mappings. The workflows can be made graphically and then executed by just a point and click. There is no need to download, install and configure a number of applications.

### Faster processing and result generation

The backend services can be Cloud based providing results much faster than any single computer.

### No need for heavy in-house computing facilities

Use services that are hosted on Clouds and avoid the equipment costs and all the hassle of hardware maintenance. Pay per use pricing model is convenient for sporadic usage.

## 7. Tools

### Firefox Plug-in

Annotate web pages inside the browser and submit them to the index

### Indexing/Search framework

1. Built using the technology made for faceted classification of Web APIs [2].
2. Multiple Apache Lucene indexes in the back-end

## 8. References


## 9. Conclusion

Better Search for Biological Web Services

Services can be searched with more precise terms and concepts. Search by ontology concept and add facets to make precise filtering.

### Conveniences in Creating Workflows

Find and mash services together with ease. The tools can suggest the degree of match and also create data mappings. The workflows can be made graphically and then executed by just a point and click. There is no need to download, install and configure a number of applications.

### Faster processing and result generation

The backend services can be Cloud based providing results much faster than any single computer.

### No need for heavy in-house computing facilities

Use services that are hosted on Clouds and avoid the equipment costs and all the hassle of hardware maintenance. Pay per use pricing model is convenient for sporadic usage.