Leibniz Bioactives Cloud

Transkript, 2018-04-24, Frank Broda (fbroda@ipb-halle.de)

Welcome everybody to my talk about the Leibniz Bioactives Cloud!

My talk will be structured as follows:

I'll start by giving you an overview, on how our project fits into the big picture. I'll continue with an introduction to the Leibniz Bioactives Cloud project and report on concepts, ideas and recent accomplishments. The talk will be concluded with outlook to plans and tasks for the near future.

The Leibniz association is composed of over 80 independent member institutions distributed all over Germany. These member institutions are very diverse with respect to size, legal form and area of work.

The Leibniz Association aims at scientific excellence as well as national and international competitiveness. Given the diversity of the member institutes and the strong competition - from Max Planck society and Helmholtz Association to name just two - promoting competitiveness is not an easy task.

The Leibniz Association has responded to this challenge by inventing several new funding instruments. One of these instruments are the Leibniz Research Alliances. In these alliances several Leibniz institutes cooperate on a certain topic. This cooperation is important to obtain the capability to tackle more complex scientific questions and thus to strengthen competitiveness.

In the beginning the petitioners of the Leibniz Research Alliance "Bioactive Compounds and Biotechnology" probably had questions such as

"What might be the best strategies to initiate cooperation?"

And they came up with more than one answer:

- The Leibniz conferences series on bioactive compounds
- The awarding of seed money to project studies an pilot projects
- A software solution, the "Leibniz Bioactives Cloud" which is the topic of my talk.

So we started our project with ideas outlined on roughly two pages in the funding proposal. Essentially it boiled down to these five points: We're supposed to develop a distributed research infrastructure which will be accessible via web browser. The users from the participating institutes are supposed to upload data which will be analyzed with semantic web technology. You've heard about text analysis in the previous talk and I'll make an additional comment on the next slide. Additionally the solution should provide mining and analytics strategies to generate knowledge from data. As the research alliance is about bioactives, the solution should facilitate a compound driven interaction. That means for instance to bring together compound providers and scientists who test them. We coined the term "chemistry aware" to describe that our software should be able to recognize and work with chemical structures.

Semantic web technology is the automatic supplementation of texts with additional facts. These additional facts come from ontologies, which could be seen as fact databases. A lot of these ontologies on different topics are freely available from the community. I want to illustrate this with an example. Here we have a text from the scientific report of the IPB. A useful ontology for this text might contain the fact, that the organism *Phytophtora infestans* belongs to the taxonomic class Oomycota. An additional ontology might bring in some facts about chemistry. After this text has been annotated, one would be able to search for the keywords Oomycota and fatty acid although these terms do not occur in the text.

One of the first things we did, was to look around, if there were related projects from which we could learn or draw code. Indeed we found quite a lot of interesting projects. All of them are somehow web based. Some of them also had quite interesting features for semantic annotation or big data analytics. All of them however follow a centralized approach with either a single instance or a single master node dominating its subsidiary nodes. And none of the solutions came with chemistry awareness.

So we had to come up with our own concept. From a developer's point of view, a centralized approach would be the most convenient way. However, it as some drawbacks. For instance there are funding issues – someone has to pay for the centralized resources. Or liability issues, who is responsible if data gets lost or disclosed. Finally, a centralized data store is always attractive to attackers.

Our approach is therefore decentralized. Each participating institute hosts – and pays for – its own node. The node and all the data on it is always under the control of the institute. All these nodes are connected via encrypted channels ti form the cloud. Users interact with their own node. Search queries or download requests are delivered via encrypted channels.

Within this architecture we will realize a smart data warehouse. In a first stage it will handle the sharing of documents. Here we think of scientific reports, stale project proposals or master theses. In general: grey literature which would be hard to find otherwise.

Later on we want to extend this by more advanced features. We think of semantic annotation using ontologies, adding of chemistry awareness, machine learning algorithms but also data mining and analytics functions to generate knowledge from research data.

To foster communication and cooperation we plan to create a social network for scientists. It would enable them to highlight their expertise and technical capabilities. It would thus facilitate the establishing of contacts among scientists from different institutions. In the same way a market place would foster cooperation by allowing

- the advertisement of compounds for testing,
- the offering of services like newly established assays or
- the sharing of research infrastructure.

My next slide gives you an impression of the current state of development. The application currently looks pretty much like Google or any other search engine. You have an input field for your search terms. If corresponding documents are found somewhere in the cloud, links for downloading these documents are displayed. If you were logged in, an upload button would show up which allows you to add documents to your specific node.

So our accomplishments so far are as follows: We have a distributed network of nodes, which is capable of performing the basic document handling operations. These are uploading and indexing of new documents, searching of documents and download of search results. Our application supports basic user authentication and authorization. It comes with automated installation and management procedures to reduce the amount of work for system administrators. Last but not least we have documented our work to build up trust because operating networked services is always sensitive.

Now I want to introduce you to one of the advanced features we plan for the Leibniz Bioactives Cloud: Distributed Structure Activity Relationships. This feature would allow you to perform a virtual screening of your compounds against a set of established assays. Let’s suppose we have three participating institutes which have data for their established assays. For instance on inflammation, cytotoxicity and a protease assay. These institutes will then use their data to train individual artificial neural networks. These trained neural networks will be made available within the Leibniz Bioactives Cloud. If you were a compound provider, you would transform the structure of your chemical compound into a set of descriptors which is then submitted to each of the neural networks. The neural networks would then predict, whether your substance would be active in one of the assays. It is important to note, that the chemical structure at this stage is not revealed to the other institutes.

The performance of our machine learning approach is superior compared to established techniques. After some learning cycles it reached a coefficient of determination of greater than 0.75 for our test data. This is displayed in the left diagram which monitors the learning curve for the training data and independently also evaluates the test data . The right diagram plots the actual activities versus the predicted values for both training and test data – both axes on a logarithmic scale. We are optimistic to be able to integrate this feature into the web portal during the next few months. And we believe that this feature could help in bringing together researchers who provide compounds and the ones who are testing them.

Our other ideas for the Leibniz Bioactives Cloud are summarized on this slide. One major task is "Becoming smart", that is the introduction of sophisticated text processing algorithms. These would allow for instance the clustering of documents by topic, the generation of knowledge or the use of fuzzy search terms. We have some algorithms and ideas we may want to try. Among them are ontologies and machine learning, but we are at an early stage here. Another idea we're following is advanced visualization which would allow the visualization of trends or relationships. An example would be an interactive word cloud which would allow you to visually refine your search results.

And of course we'll have to come up with implementations of a market place and – as far as data protection permits – a social network for scientists.

Finally the decentralized approach requires your support. We need your user demand to press your IT department to install a cloud node. Installation is straight forward. Starting from scratch it requires only a few hours. The current hardware requirements are really moderate: a single CPU, two gigabytes random access memory and approximately 20 gigabytes disk space.

Once established, we need you to feed documents and data into the cloud. Without data, the whole project wouldn't make any sense.

And last but not least, we could use some test drivers who are willing to dedicate a few hours to testing and maybe conceptual work.

I’m now at the end of my talk. I want to thank the IPB and the research alliance for funding our project and I want to thank you for listening. I’d be happy to answer your questions.