# Application for DBDBD 2012

**Name:** Brend Wanders <b.wanders@utwente.nl>
**Affiliation:** Databases Group, Faculty of Electrical Engineering, Mathematics & Computer Science, University of Twente

**Title of Talk:** Pay-as-you-go data integration for bio-informatics

## Abstract

Scientific research in bio-informatics is often data-driven and supported by numerous biological databases. A biological database contains factual information collected from scientific experiments and computational analyses about areas including genomics, proteomics, metabolomics, microarray gene expression and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

In a growing number of research projects, bio-informatics researchers like to ask combined questions, i.e., questions that require the combination of information from more than one database. We have observed that most bio-informatics papers do not go into detail on the integration of different databases. It has been observed that roughly 30% of all tasks in bio-informatics workflows are data transformation tasks, a lot of time is used to integrate these databases (shown by [1]).

As data sources are created and evolve, many design decisions made by their creators. Not all of these choices are documented. Some of such choices are made implicitly based on experience or preference of the creator. Other choices are mandated by the purpose of the data source, as well as inherent data quality issues such as imprecision in measurements, or ongoing scientific debates. Integrating multiple data sources can be difficult.

We propose to approach the time-consuming problem of integrating multiple biological databases through the principles of 'pay-as-you-go' and 'good-is-good-enough'. By assisting the user in defining a knowledge base of data mapping rules, schema alignment, trust information and other evidence we allow the user to focus on the work, and put in as little effort as is necessary for the integration to serve the purposes of the user. By using user feedback on query results and trust assessments, the integration can be improved upon over time.

The research will be guided by a set of use cases. As the research is in its early stages, we have determined three use cases:

Homologues, the representation and integration of groupings. Homology is the relationship between two characteristics that have descended, usually with divergence, from a common ancestral characteristic. A characteristic can be any genic, structural or behavioural feature of an organism

Metabolomics integration, with a focus on the TCA cycle. The TCA cycle (also known as the citric acid cycle, or Krebs cycle) is used by aerobic organism to generate energy from the oxidation of carbohydrates, fats and proteins.

Bibliography integration and improvement, the correction and expansion of citation databases.

[1] I. Wassink. *Work flows in life science.* PhD thesis, University of Twente, Enschede, January 2010.