# Personalized Semantic Assistance for the Curation of Biochemical Literature

Fedor Bakalov*, Marie-Jean Meurs†,‡, Birgitta König-Ries*, Bahar Sateli†, René Witte†,
Greg Butler†,‡, and Adrian Tsang‡,¶
**Institute for Computer Science, Friedrich Schiller University of Jena, Germany*
†*Department of Computer Science and Software Engineering, Concordia University, Montréal, QC, Canada*
‡*Centre for Structural and Functional Genomics, Concordia University, Montréal, QC, Canada*
¶*Department of Biology, Concordia University, Montréal, QC, Canada*

***Abstract*—The number of scientific publications available in multiple repositories is huge and rapidly growing. Accessing this information is of critical importance to conducting research and designing experiments. However, retrieving data of particular interest for a specific research field in such a large volume of publications is often like looking for a needle in a haystack. We present a web platform that supports researchers in navigating and curating biochemical literature. Our platform provides a single-point of access to abstracts of publications harvested from multiple databases and supports further analysis of these abstracts. It also allows users to obtain a personalized view of the literature and its semantic analysis results.**

***Keywords*-biocuration, natural language processing, text mining, personalization, scientific literature portal**

## I. Introduction

The number of scientific publications in the biomedical and life science bibliome reached 21 million articles literature in May 2012, as indexed by PubMed [1]. However, retrieving data of particular interest for a specific research field in such a large volume of publications is often like looking for a needle in a haystack. A researcher querying various biological bibliographic databases typically collects a long list of potentially relevant papers. Reading all the abstracts and full-texts of these papers to locate relevant information is an unavoidable step in the literature curation process. Unfortunately, since the task is highly time-consuming and error-prone, it is a bottleneck in the knowledge discovery workflow [2]. Researchers, curators and experimenters are querying different resources with different goals in mind, and often are looking for different kinds of information. To break the curation bottleneck, the biological research community needs flexible and user-centric tools.

We present a web platform that supports researchers in the curation of biochemical literature. It provides personalized access to abstracts of scientific publications harvested from multiple databases and supports further semantic analysis. More specifically, it allows users to process content with a number of *Semantic Assistants* (SAs), e.g., for extracting named entities (such as organisms, enzymes, genes, or substrates), generating summaries, and indexing literature. Additionally, the platform allows users to obtain a personalized view of the literature and the semantic assistants' results. It can sort abstracts either chronologically or according to the relevance to a users' interest profile, which the platform builds unobtrusively based on his browsing history. It can foreground the most relevant abstracts or fragments of abstracts that match items in the user profile. It is important to note that users can access their interest profiles generated by the platform and override the system beliefs in an efficient and easy-to-use way. They can also use their interest profiles to directly access abstracts relevant to their interests.

## II. Related Work

Natural Language Processing (NLP) and semantic web approaches are increasingly being adopted in biomedical research [3], [4], [5]. During the last decade, several systems combining text mining and semantic processing have been developed to help life sciences researchers in extracting knowledge from the literature. For instance, Textpresso [6] enables the user to search for categories of biological concepts and classes relating two objects and/or keywords within an entire literature set. GoPubMed [7] supports the arrangement of the abstracts returned from a PubMed query. More visual than the two aforementioned systems are Bio-Jigsaw [8], a visual analytics system highlighting connections between biological entities or concepts grounded in the biomedical literature and Reflect [9], a Firefox plugin that tags gene, protein and small molecule names in any Web page. In the NLP community, the research efforts made for the last decades to develop such tools have been mainly focused on the capabilities of the systems to achieve text mining tasks. Examples of such tasks are (bio-)entity recognition, linkage to reference database entries or relationship extraction between entities of interest. Open challenges [10], [11] dedicated to specific tasks have provided the community with intrinsic evaluation of systems in a reproducible context. However, few studies report on the ease of user-system interactions or a system's effectiveness in supporting the user's needs.

## III. System Architecture

The ultimate goal of this research is to develop a web platform that assists scientists in keeping track of relevant literature available in multiple scientific databases. To fulfil its purpose, the platform needs a user model providing

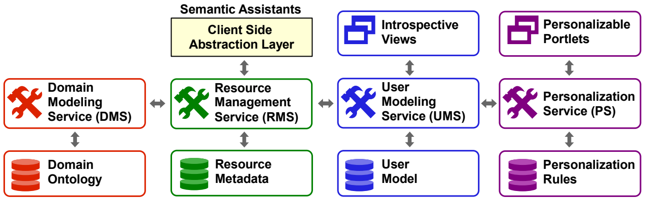

Semantic Assistants
Client Side Abstraction Layer
Introspective Views
Personalizable Portlets
Domain Modeling Service (DMS)
Resource Management Service (RMS)
User Modeling Service (UMS)
Personalization Service (PS)
Domain Ontology
Resource Metadata
User Model
Personalization Rules

Figure 1. System architecture

information about scientific interests of individual users. It also needs a repository providing metadata of literature being published in scientific databases. To achieve automatic selection of publications that match user interests, both the user model and metadata must use the same vocabulary. This vocabulary must provide semantics of the domain knowledge represented in a machine-processable way. Finally, the platform needs personalization rules that govern how and when relevant literature must be delivered for the given user.

We present a portal framework that fulfils the aforementioned requirements. The framework consists of four units as shown in Figure 1: The *Domain Modeling Unit* encapsulates the components responsible for storing, accessing, and managing the domain model, which is represented as an ontology, formalized in the Web Ontology Language (OWL)[1]. The *Resource Management Unit* is responsible for harvesting and annotating literature from multiple scientific databases. A *User Modeling Unit* provides the models, mechanisms, and interfaces for managing information about users required for adaptation. Finally, our *Personalization Unit* stores personalization rules and provides mechanisms for performing personalization in the portal.

**Resource Management.** The method for gathering and annotating content in the framework follows our approach to semantic enrichment of resources for adaptation proposed in our earlier work [12]. It consists of three operations: In the first step, the resource management service (RMS) fetches content of publications from predefined databases, e.g., PubMed. Using Application Programming Interfaces (API) provided by the databases, the service retrieves publications matching queries defined by users in their individual query lists. For a retrieved publication, the service checks if the metadata repository already contains a record for the publication. If no record is found, the service annotates the publication. For annotation, it uses NLP pipelines provided by the Semantic Assistants framework [13]. For instance, for annotation of biochemical literature related to lignocellulosic degradation, it uses the mycoMINE pipeline [14]. For every extracted entity type, the service identifies an appropriate concept of the domain ontology based on semantic type mapping, which maps the ontology scheme of the NLP tools to the scheme of the domain ontology. Once the domain concept is known, the service invokes the domain modeling service (DMS) to check the domain ontology for the existence of a matching instance. To mitigate the problem of ambiguity, the service leverages the entity name, its type, and properties. If a corresponding instance is found, DMS checks whether it can update it using the results of semantic processing (e.g. update one of the properties) and then uses the Unified Resource Identifier (URI) of that instance to include it in the document metadata. If no instance is found, RMS calls DMS to insert a new instance using the entity name, ontology concept, and other attribute values or relations returned by the NLP tools. After the instance is inserted, its URI is used to include it in the document metadata. For each document, the output of the annotation step is a set of URIs referring to the instances from the domain ontology. These instances represent the semantic entities identified in the document. Also, depending on the NLP tool that is used for entity extraction, the output may include the relevance of identified entities to the given document. For each identified instance, RMS generates an annotation record and writes it into the metadata repository.

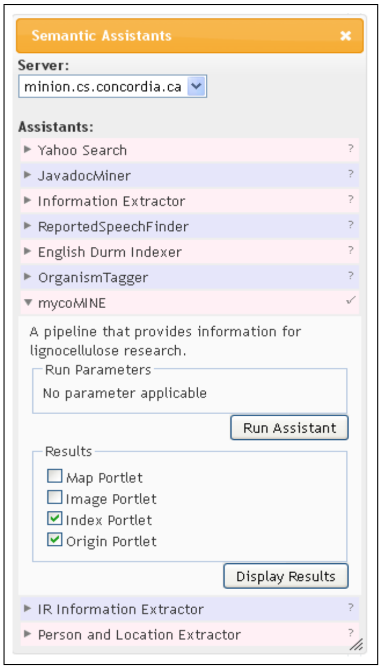
**User Modeling.** The user model stores information about interests of individual scientists. It is designed as an overlay model, i.e., user interests are represented as an overlay of domain concepts defined in the domain ontology. For each concept, the model stores information about the exact *degree* to which the user is interested in it. The degree of interest is defined by an *interest weight*, represented by a real number in the range from 0 to 1. Additionally, the model stores an approximated semantic interpretation of the interest weight as *interest status*. The model supports four statuses of interest, namely interesting, partially interesting, uninteresting, and blocked. A blocked concept is a concept that the user explicitly blocked from being monitored and used for personalization. It means that the system ignores any evidence of the user activity on this concept and does not recommend any content related to this concept.

The user model is updated following our hybrid approach to user modeling proposed in [15]. This approach supports three types of user model updates: log-based updates, inference-based updates, user explicit updates. Log-based updates are performed based on the evidence of publications that the user has accessed in the portal. Inference-based updates are performed based on semantic relations among the instances in the domain ontology, i.e., the user interest is propagated from one item to another via the object properties existing between the items. Log-based and inference-based updates are performed automatically by the user modeling service. User explicit updates are made by the user through the *IntrospectiveViews* interface presented in Section IV.

**Personalization.** As described in detail in the next section, the portal content is delivered to users through personalizable portlets[2]. Users can view portlets in standard or personalized

[1] http://www.w3.org/TR/owl-features/

[2] A portlet is a pluggable user interface component of web portals that provides a specific piece of content or an application.



Semantic Assistants ✕

Server:

minion.cs.concordia.ca

Assistants:

- ▸ Yahoo Search
- ▸ JavadocMiner
- ▸ Information Extractor
- ▸ ReportedSpeechFinder
- ▸ English Durm Indexer
- ▸ OrganismTagger
- ▾ mycoMINE

A pipeline that provides information for lignocellulose research.

Run Parameters

No parameter applicable

Run Assistant

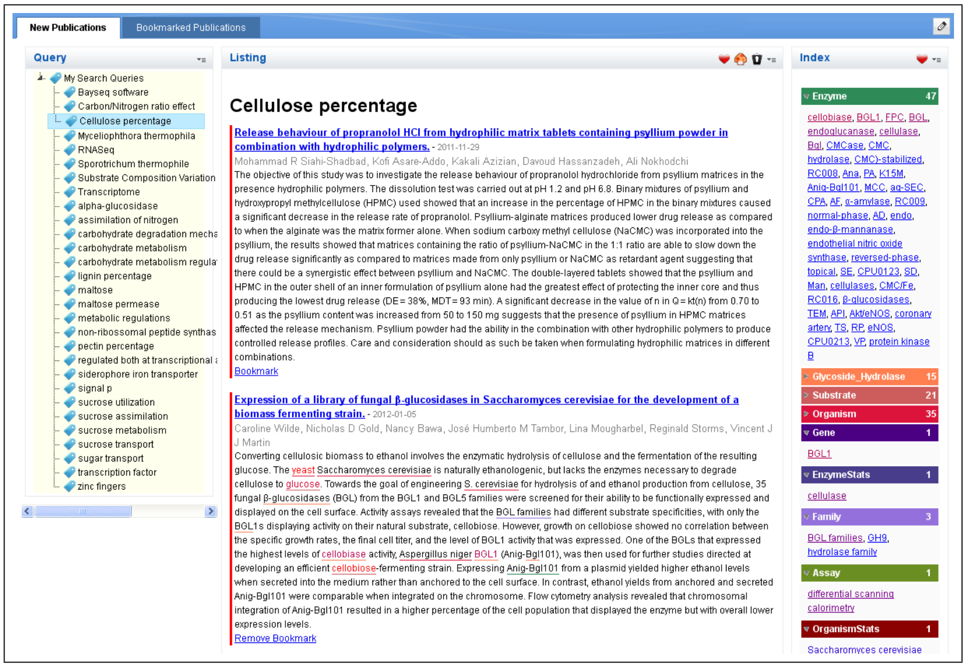
Results

☐ Map Portlet
☐ Image Portlet
☑ Index Portlet
☑ Origin Portlet

Display Results

- ▸ IR Information Extractor
- ▸ Person and Location Extractor

Figure 2. A menu for SAs



New Publications | Bookmarked Publications

**Query**

- My Search Queries
  - Bayseq software
  - Carbon/Nitrogen ratio effect
  - Cellulose percentage
  - Myceliophthora thermophila
  - RNASeq
  - Sporotrichum thermophile
  - Substrate Composition Variation
  - Transcriptome
  - alpha-glucosidase
  - assimilation of nitrogen
  - carbohydrate degradation mecha
  - carbohydrate metabolism
  - carbohydrate metabolism regula
  - lignin percentage
  - maltose
  - maltose permease
  - metabolic regulations
  - non-ribossomal peptide synthas
  - pectin percentage
  - regulated both at transcriptional a
  - siderophore iron transporter
  - signal p
  - sucrose utilization
  - sucrose assimilation
  - sucrose metabolism
  - sucrose transport
  - sugar transport
  - transcription factor
  - zinc fingers

**Listing**

Cellulose percentage

Release behaviour of propranolol HCl from hydrophilic matrix tablets containing psyllium powder in combination with hydrophilic polymers. - 2011-11-29

Mohammad R Siahi-Shadbad, Kofi Asare-Addo, Kakali Azizian, Davoud Hassanzadeh, Ali Nokhodchi

The objective of this study was to investigate the release behaviour of propranolol hydrochloride from psyllium matrices in the presence hydrophilic polymers. The dissolution test was carried out at pH 1.2 and pH 6.8. Binary mixtures of psyllium and hydroxypropyl methylcellulose (HPMC) used showed that an increase in the percentage of HPMC in the binary mixtures caused a significant decrease in the release rate of propranolol. Psyllium-alginate matrices produced lower drug release as compared to when the alginate was the matrix former alone. When sodium carboxy methyl cellulose (NaCMC) was incorporated into the psyllium, the results showed that matrices containing the ratio of psyllium-NaCMC in the 1:1 ratio are able to slow down the drug release significantly as compared to matrices made from only psyllium or NaCMC as retardant agent suggesting that there could be a synergistic effect between psyllium and NaCMC. The double-layered tablets showed that the psyllium and HPMC in the outer shell of an inner formulation of psyllium alone had the greatest effect of protecting the inner core and thus producing the lowest drug release (DE = 38%, MDT = 93 min). A significant decrease in the value of n in Q = kt(n) from 0.70 to 0.51 as the psyllium content was increased from 50 to 150 mg suggests that the presence of psyllium in HPMC matrices affected the release mechanism. Psyllium powder had the ability in the combination with other hydrophilic polymers to produce controlled release profiles. Care and consideration should as such be taken when formulating hydrophilic matrices in different combinations.

Bookmark

Expression of a library of fungal β-glucosidases in Saccharomyces cerevisiae for the development of a biomass fermenting strain. - 2012-01-05

Caroline Wilde, Nicholas D Gold, Nancy Bawa, José Humberto M Tambor, Lina Mougharbel, Reginald Storms, Vincent J J Martin

Converting cellulosic biomass to ethanol involves the enzymatic hydrolysis of cellulose and the fermentation of the resulting glucose. The yeast Saccharomyces cerevisiae is naturally ethanologenic, but lacks the enzymes necessary to degrade cellulose to glucose. Towards the goal of engineering S. cerevisiae for hydrolysis of and ethanol production from cellulose, 35 fungal β-glucosidases (BGL) from the BGL1 and BGL5 families were screened for their ability to be functionally expressed and displayed on the cell surface. Activity assays revealed that the BGL families had different substrate specificities, with only the BGL1s displaying activity on their natural substrate, cellobiose. However, growth on cellobiose showed no correlation between the specific growth rates, the final cell titer, and the level of BGL1 activity that was expressed. One of the BGLs that expressed the highest levels of cellobiase activity, Aspergillus niger BGL1 (Anig-Bgl101), was then used for further studies directed at developing an efficient cellobiose-fermenting strain. Expressing Anig-Bgl101 from a plasmid yielded higher ethanol levels when secreted into the medium rather than anchored to the cell surface. In contrast, ethanol yields from anchored and secreted Anig-Bgl101 were comparable when integrated on the chromosome. Flow cytometry analysis revealed that chromosomal integration of Anig-Bgl101 resulted in a higher percentage of the cell population that displayed the enzyme but with overall lower expression levels.

Remove Bookmark

**Index**

▾ Enzyme 47

cellobiase, BGL1, FPC, BGL, endoglucanase, cellulase, Bgl, CMCase, CMC, hydrolase, CMC)-stabilized, RC008, Ana, PA, K15M, Anig-Bgl101, MCC, aq-SEC, CPA, AF, α-amylase, RC009, normal-phase, AD, endo, endo-β-mannanase, endothelial nitric oxide synthase, reversed-phase, topical, SE, CPU0123, SD, Man, cellulases, CMC/Fe, RC016, β-glucosidases, TEM, API, Akt/eNOS, coronary artery, TS, RP, eNOS, CPU0213, VP, protein kinase B

▸ Glycoside_Hydrolase 15

▸ Substrate 21

▸ Organism 35

▾ Gene 1

BGL1

▾ EnzymeStats 1

cellulase

▾ Family 3

BGL families, GH9, hydrolase family

▾ Assay 1

differential scanning calorimetry

▾ OrganismStats 1

Saccharomyces cerevisiae

Figure 3. A portal page with personalizable portlets

states. In a personalized state, users can choose between several personalization effects. For example, they can choose whether publications must be sorted by interest or chronologically. If a portlet is requested in a personalized view, the portlet invokes the personalization service. This service retrieves user personalization preferences to determine what personalization effects must be generated for the given portlet and user. It also retrieves metadata of the requested content and interests of the given user. Based on the metadata and user interests and personalization preferences, it personalizes the content and passes it to the requesting portlet.

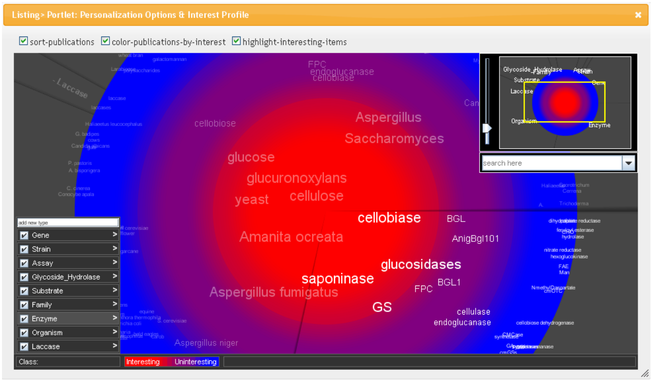
## IV. User Interface

The proposed framework was deployed on an IBM WebSphere Portal Server[3]. Figure 3 displays a personalized page that users see after they have logged into the portal. This page consists of a number of portlets providing different types of content and functions. The *Query* portlet on the left displays a list of user search queries, which are used by the portal to retrieve publications from scientific databases. This portlet allows users to add, edit, and delete queries and organize them hierarchically. Upon a mouse click on a query, the portal will display a list of matching publications in the *Listing* portlet. The *Listing* portlet allows users to request various types of semantic assistance. Users can view a list of named entities extracted from the publications, their summaries, or an index. All types of assistance supported by the portlet can be seen in the *Semantic Assistants* menu (Figure 2). In this menu, users can choose an assistant they want and set desired view options for the assistant's results. Depending on the type of assistant, its results can be displayed in the source text, as an index, a map, or a text in a side portlet. For instance, Figure 3 displays results of the mycoMINE assistant [14], which extracts entities and facts related to fungal enzymes involved in lignocellulose degradation, such as enzymes, organisms, genes, substrates, pH, temperature or activity assay conditions. The entities extracted by the assistant, as selected in Figure 2, are underlined in the text of publications listed in the origin portlet and displayed as an index in a side portlet.

Portal content and results of semantic assistants can be personalized by users. For some portlets, users can select whether they want to see the content in a personalized or standard view. In the personalized view, users can instruct how the portlet content should be personalized. Users can switch between personalized and standard views using a personalization menu. From this menu, users can also request a window displaying the portlet personalization options and the user's interest profile (Figure 4). The personalization options vary from portlet to portlet. For example, the *Listing* portlet displaying a list of new publications supports three personalization effects: (1) publications can be sorted according to the user interest profile; (2) the most interesting publications for the user can be highlighted by a color marker; (3) mentions of items from the user interest profile can be highlighted in the publications list. By selecting corresponding checkboxes users can achieve desired personalization effects in the portlet. User changes on the personalization options are immediately projected onto the portlet content.

For the visualization of user profiles, we leverage the

[3] http://www.ibm.com/software/websphere/portal/



Listing> Portlet: Personalization Options & Interest Profile

☑ sort-publications ☑ color-publications-by-interest ☑ highlight-interesting-items

Figure 4. Personalization options and user interest profile. Color screenshots and screencasts are available at http://www.minerva-portals.de/research/introspective-views/v.3

IntrospectiveViews interface proposed in our earlier work on scrutable user modeling in [16]. The interface visualizes user interests using a metaphor of circular zones partitioned into slices, where each zone represents items of certain interest degree and each slice represents items of a specific type. The hot zone in the center displays items that users are strongly interested in. The cold zone at the circle edge displays items that users are not interested in. Items are grouped into circular sectors by type. The profile shown in Figure 4 displays items of such types as enzyme, gene, organism, strain, and some others. The interface provides functions for getting an overview, zooming, filtering, navigation, and search. It also displays relevant content and semantic relations among items. For example, by clicking an item, it will show a list of all publications where mentions of this item were found.

In addition to viewing, IntrospectiveViews allows editing information in the model. It allows adding and deleting items, changing interest degree, organizing items by type, defining user-specific types, and creating semantic relations among items. To change the interest degree of an item, the user needs to drag the item to the corresponding interest zone. Dragging to the center increases interest and dragging to the edge decreases interest. New items can be added to the profile simply by double clicking on an empty space on the circular surface. Items can be blocked from personalization by dragging them onto the recycle bin. Similarly to changes of personalization options, all changes in user interest profiles made by users through IntrospectiveViews are immediately projected onto the personalized content.

## V. Conclusions and Future Work

We presented a web platform that supports researchers in the curation of biochemical literature. The platform provides a single-point of access to a bibliography of publications harvested from multiple scientific databases. It allows users to process and analyze publications using a number of semantic assistants, e.g., named entity extractors, summarizers, and indexers. Additionally, it allows users to obtain a personalized view of publications and results of semantic assistants.

In our future work, we plan to perform a detailed evaluation of the platform through a user study.

## Acknowledgements

We would like to thank Justin Powlowski for his expert advice on the biology content. Funding for part of this work was provided by Genome Canada and Génome Québec. Part of this research was sponsored by the IBM Ph.D. Fellowship Awards Program and carried out in the framework of the Minerva Portals project in cooperation with IBM Deutschland Research & Development GmbH.

## References

[1] E. Sayers et al., “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Research*, vol. 38, no. suppl 1, pp. D5–D16, 2009.

[2] Doug Howe et al., “Big data: The future of biocuration,” *Nature*, vol. 455, pp. 47–50, 2008.

[3] L. Hirschman et al., “Text mining for the biocuration workflow,” *Database*, vol. 2012, 2012.

[4] C. J. O. Baker and K.-H. Cheung, Eds., *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*. Springer, 2007.

[5] N. Shadbolt, T. Berners-Lee, and W. Hall, “The Semantic Web Revisited,” *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, 2006.

[6] H.-M. Müller, E. E. Kenny, and P. W. Sternberg, “Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature,” *PLoS Biol*, vol. 2, no. 11, p. e309, 09 2004.

[7] A. Doms and M. Schroeder, “GoPubMed: exploring PubMed with the Gene Ontology,” *Nucleic Acids Research*, vol. 33, no. suppl 2, pp. W783–W786, 2005.

[8] C. Görg et al., “Visualization and Language Processing for Supporting Analysis across the Biomedical Literature,” in *Knowledge-Based and Intelligent Information and Engineering Systems*, ser. LNCS. Springer, 2010, vol. 6279, pp. 420–429.

[9] E. Pafilis et al., “Reflect: augmented browsing for the life scientist,” *Nature Biotechnology*, vol. 27, pp. 508–510, 2009.

[10] C. Arighi, P. Roberts, S. Argawal, and et al., “BioCreative III Interactive Task: an Overview,” *BMC Bioinformatics;12 Suppl 8:S1*, 2011.

[11] W. Hersh and E. Voorhees, “TREC genomics special issue overview,” *Inf. Retr.*, vol. 12, no. 1, pp. 1–15, Feb. 2009.

[12] O. Schimratzki, F. Bakalov, A. Knoth, and B. König-Ries, “Semantic enrichment of social media resources for adaptation,” in *Workshop on Adaptation in Social and Semantic Web held in conj. with the 18th Int. Conf. on User Modeling, Adaptation, and Personalization*, 2010.

[13] R. Witte and T. Gitzinger, “Semantic Assistants – User-Centric Natural Language Processing Services for Desktop Clients,” in *3rd Asian Semantic Web Conference (ASWC 2008)*, ser. LNCS, vol. 5367. Springer, 2008, pp. 360–374.

[14] M-J. Meurs et al., “Semantic text mining support for lignocellulose research,” *BMC Medical Informatics and Decision Making, Vol 12 Suppl 1*, 2012.

[15] F. Bakalov, B. König-Ries, A. Nauerz, and M. Welsch, “A Hybrid Approach to Identifying User Interests in Web Portals,” in *Int. Conf. on Innovative Internet Community Systems*, 2009.

[16] ——, “Introspectiveviews: An interface for scrutinizing semantic user models,” in *Int. Conf. on User Modeling, Adaptation, and Personalization*, 2010.