CCTC - Computer Science and Technology Center, University of Minho
IBB - Institute for Biotechnology and Bioengineering, Centre of Biological Engineering, University of Minho
SING - Next Generation Computer Systems Group, School of Informatics Engineering, University of Vigo

@Note Biomedical Text Mining Workbench

User Guide of @Note-Basics

Anália Lourenço, Rafael Carreira, Paulo Maia, Sónia Carneiro, Daniel Glez-Peña, Florentino Fdez-Riverola, Eugénio C. Ferreira, Isabel Rocha, Miguel Rocha

2008

Contents

1 Basic concepts and interaction 1

1.1 Introduction . . . 1

1.2 Datatypes and operations . . . 2

1.3 User interaction . . . 2

1.4 Getting started . . . 3

2 Main functionalities of @Note-Basics 4

2.1 @Note Projects . . . 4

2.1.1 Creating a new project . . . 4

2.1.2 Saving a project . . . 6

2.1.3 Loading an existing project . . . 6

2.2 Handling database connections . . . 7

2.2.1 Connecting to an existing database . . . 7

2.2.2 Creating a local database . . . 8

2.2.3 Loading default data . . . 9

2.3 PubMed Searches . . . 9

2.4 Journal Retrieval . . . 10

2.5 Named Entity Recognition . . . 13

2.5.1 Document View . . . 18

2.6 Manual Annotation . . . 20

2.6.1 Annotating a new term . . . 20

2.6.2 Correcting an annotation . . . 20

2.6.3 Removing an annotation . . . 23

2.7 Handling lexical resources . . . 23

2.7.1 Handling dictionaries . . . 24

2.7.2 Loading lookup tables . . . 25

2.8 Project Settings . . . 26

List of Figures

2.1 Creating a new project. 4
2.2 Configuring a new project. 5
2.3 Saving a project. 6
2.4 Loading a project. 7
2.5 Creating a database connection 8
2.6 Connecting to an existing database. 8
2.7 Creating a new local database. 9
2.8 Listing PubMed queries. 10
2.9 Viewing a Result Set. 11
2.10 Checking detailed information about a publication. 12
2.11 Viewing the PDF file of a publication. 13
2.12 Setting the publication's relevance. 14
2.13 Selecting the publication set. 14
2.14 Viewing the Publication Set. 15
2.15 Selecting the Publication Set for NER. 16
2.16 Selecting the dictionary and the classes to annotate. 16
2.17 Annotation options. 17
2.18 NER running operation. 18
2.19 ANoteNerBox view. 19
2.20 Document view. 21
2.21 Adding a NER term annotation. 22
2.22 Dictionary enrichment 23
2.23 Diagram illustrating the options of adding new terms to the dictionary. 24
2.24 Correcting a term annotation. 25
2.25 Changing project settings. 26

Chapter 1

Basic concepts and interaction

1.1 Introduction

@Note is a Biomedical Text Mining workbench that integrates current Biomedical Text Mining (BioTM) methods and provides biologists with intuitive tools capable of supporting their bibliographic searches and further literature curation. The major guidelines of its development were interoperability, extensibility and user-friendly interface. The workbench is meant for both BioTM research and curation. On one hand, it supports regular curation activities, providing an intuitive Graphical User Interface (GUI) interface that does not require any knowledge about workbench or technique implementation. On the other hand, it is also meant for people with programming skills that might wish to extend the workbench capabilities.

@Note is implemented over AIBench, a JAVA framework meant to ease the development of Artificial Intelligence and Data Analysis applications. The main strengths of AIBench are its clear design and available services. Its design is problem-independent, minimum framework-related code is required in order to produce new functionalities. Moreover, it generates GUI code and enforces well-designed MVC code, supporting three main artifacts: operations, data types and views. Operations and data types are used in problem modelling while views display data in a ”friendly” way.

Regarding operations, @Note sustains the general workflow of BioTM, fully covering all activities performed in manual curation. The workbench supports the retrieval, processing and annotation of documents as well as their analysis at different levels. So far, only dictionary and ontology-based annotation are supported as it was considered more important to provide means for the creation of annotated corpora rather than the construction of models based on general biomedical corpora.

This document briefly explains the functionalities of the @Note-Basics application and the way it can presently be used. This application brings together a number of basic components of the full @Note platform in a single application with the basic tools of BioTM, oriented towards the basic needs of biologists. This is still a preliminary version of the documentation.

1.2 Datatypes and operations

Every application built based on AIBench is organized around the concepts of datatypes and operations, defined as follows:

- **Datatypes:** define the types of data that are of interest to a given application. For each data object, one or more visualizers can be defined to show its content to the user in a given perspective. Data objects can have a hierarchy, where a given object A contains objects B and C (these objects are called compound objects).

  The set of data objects and their hierarchy in a given application are shown in a tree (the clipboard), typically appearing in the left side of the screen. In this tree, compound objects can be opened to show their contents (a list of other data objects). When a given data object is selected (double clicked), the available visualizers (if any) are launched in the right area of the screen.

- **Operations:** each operation defines a function that takes zero or more data objects as its inputs and can create as an output zero or more data objects and/or merely change the input data objects. Operations can be accessed through the menu options, being typically grouped in several menus and sub-menus. Operations can also be run from the clipboard, by right clicking a data object (this will show the list of all available operations using that data object as an input).

1.3 User interaction

Since @Note-Basics is an AIBench based application, the user interaction is thought to be as simple as possible. The Model-View-Controller (MVC) architectural pattern has been used in every step of development of AIBench as well as of @Note, resulting in a great deal of decoupling between the operational data and the views.

As referenced before, a View is related to a given Datatype. If there is no View associated with a Datatype, a Default View is launched (a bean inspector). The Views will, by default, be launched on the right side of the work area.

So, the original layout of the components has three major areas: the menus on the top, the clipboard on the left and the view's area on the right.

1.4 Getting started

The concepts within @Note can be overwhelming for a beginner in BioTM. Therefore, we provide a guideline to start using @Note-Basics.

The first two steps that are needed are: (i) to create a new project (see Section 2.1.1) and (ii) to create a new database - see Section 2.2.2 (in alternative, the user can create a connection to an existing database as explained in Section 2.2.1).

From this point on, the user has two alternatives: (i) to use the application with some sample data provided, by using the option **Load sample data** in the **Database** menu; (ii) to start their own case study from scratch.

We would recommend the first option to start with. In this case, the results of a pre-defined query to PubMed are loaded into the catalogue and a dictionary with terms related to the organism *E. coli* is loaded into the database. Alternatively, at this step, a query can be performed by the user (see Section 2.3) and there is also the need to create specific lexical resources (check Section 2.7 for details).

The next logical step will be to select a number of documents and to load them. The user selects the set of interesting documents (or all of them) and chooses if only abstracts or full texts will be used. The first option reduces the time since abstracts are already loaded from PubMed. The latter implies the journal retrieval of the available full texts (see Section 2.4), an operation that can take quite a while if the number of selected documents is big.

The set of selected documents can then be annotated using the available lexical resources (see Section 2.5 for details). A final step is the visualization of the annotation results and the manual curation of the user desires to correct errors and enrich the lexical resources (see Section 2.6).

The next sections will give further detail in the operations mentioned in this brief introduction.

Chapter 2

Main functionalities of @Note-Basics

2.1 @Note Projects

2.1.1 Creating a new project

To create a new project, the user chooses the corresponding operation in the **File** menu, selecting **New Project**.

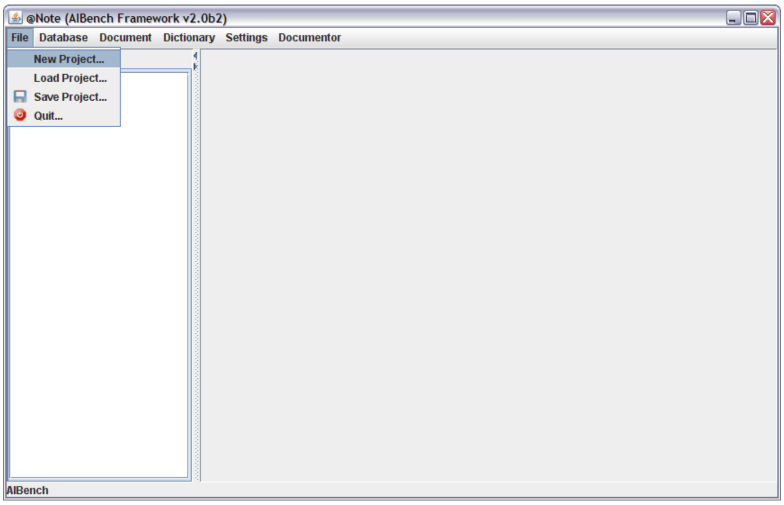
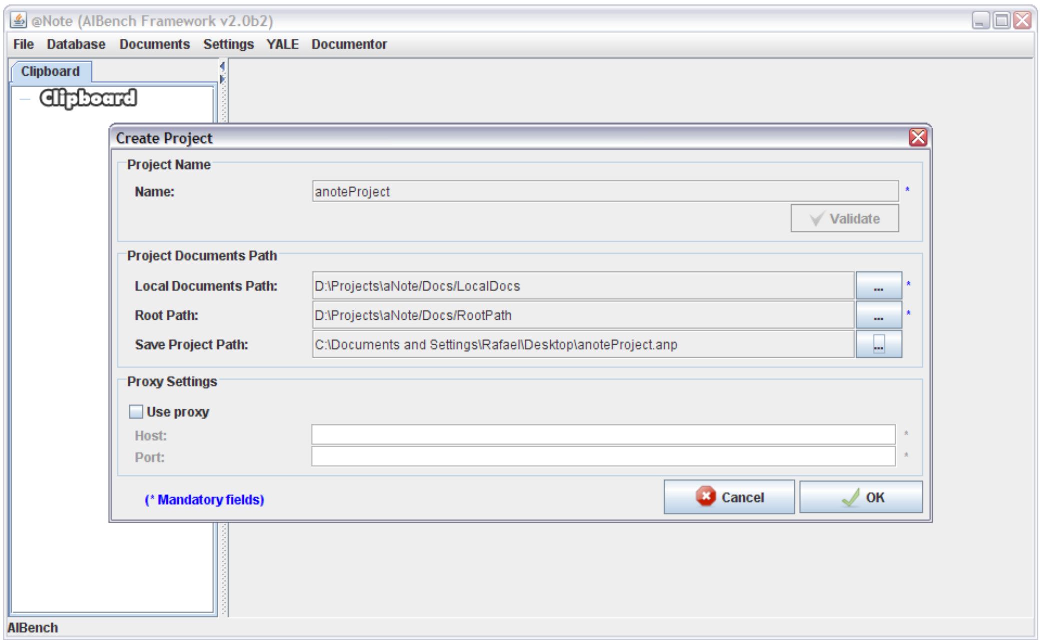


Figure 2.1: Creating a new project.

In the following popup window, a name for the project has to be chosen. When the name is set, the **Validate** button is pressed. If no project with the chosen name exists, it is accepted and the user is able to proceed with the configuration. At this point, there are two mandatory fields to configure (Figure 2.2):

1. Firstly, the **Local Documents Path** is set, which is the local folder for the **pdf** original documents handled by the project. In this folder, all **pdf** documents captured in Journal Retrieval processes will be saved.
2. Secondly, the **Root Path** is defined. This is the folder where all the project documents processed by @Note will be saved. This includes all the annotated documents, documents created as a result from **pdf** to **txt** processes, among others.

At this stage, the user can also select a path to save the project main file (anp extension) and also define proxy configurations if they are needed by the available internet connection. The configurations carried out in this step, can be changed later in the **Project settings** menu. When the **OK** button is pressed, the project is created and a data item of the type *ANoteProject* is added to the clipboard. This will be the root of all objects of a given working session.



Create Project

Project Name

Name: anoteProject *

Project Documents Path

Local Documents Path: D:\Projects\aNote/Docs/LocalDocs *

Root Path: D:\Projects\aNote/Docs/RootPath *

Save Project Path: C:\Documents and Settings\Rafael\Desktop\anoteProject.anp

Proxy Settings

□ Use proxy

Host: *

Port: *

(* Mandatory fields)

Figure 2.2: Configuring a new project.

When a project is created, it has two different objects under its clipboard tree:

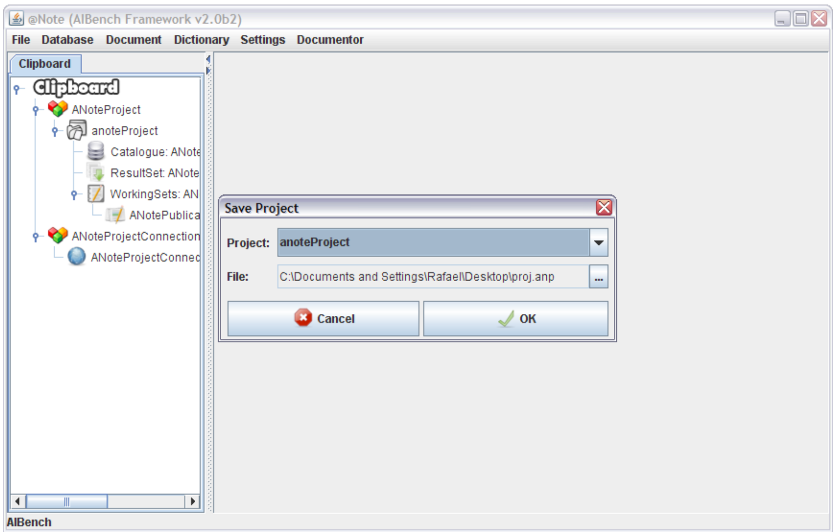
- A **Catalogue**: that represents the object used to perform queries to PubMed and store the results;

- **LexicalResources**, a sub-tree that handles the resources for performing annotation. These include dictionaries and lookup tables. This set is initially empty.

In the project tree, other types of objects will appear as a result of the operations described in the next few sections.

2.1.2 Saving a project

To save a project, the user chooses the **File** menu option, and then **Save Project**. In the popup window, a project and a file to save it are selected (as before the file must have extension ".anp").



Project: anoteProject

File: C:\Documents and Settings\Rafael\Desktop\proj.anp

Figure 2.3: Saving a project.

2.1.3 Loading an existing project

If there are previously saved projects, it is possible to load them. To load a project, the user selects the **File** menu and the option **Load Project**. In the popup window, the user chooses the file where the project was saved (".anp" extension) and clicks **Load** to perform the operation (Figure 2.4). As a result, an **ANoteProject** object is added to the clipboard.

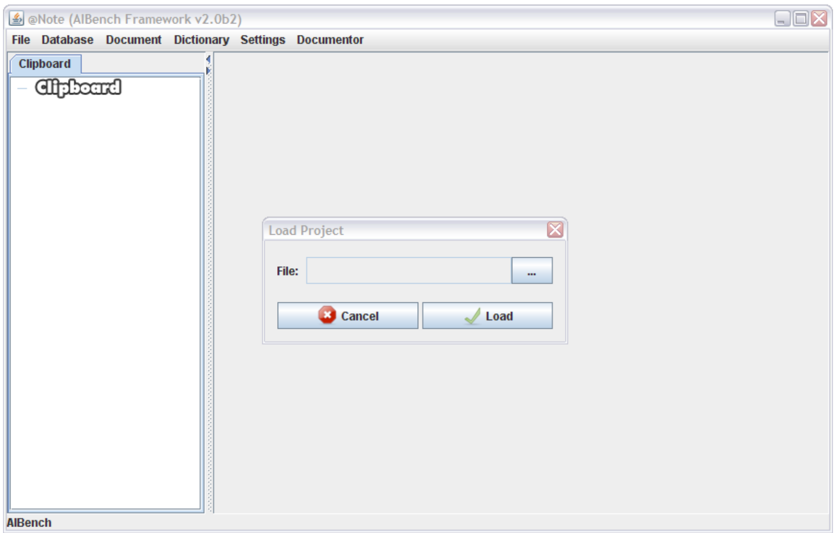


Figure 2.4: Loading a project.

2.2 Handling database connections

An @Note project needs to have a database connection associated with it (the MySQL database engine is used) since many operations work over data in the database. The database connection is created in the context of the Catalogue datatype (Figure 2.5) or under the menu option **Databases**. The user can choose to create a connection to an existing database or to create a new local database.

2.2.1 Connecting to an existing database

To create a connection to a previously existing database, the user selects the option **Create DB Connection**.

In the popup window (Figure 2.6), the user can select previously saved connection parameters and edit them if necessary or define a new connection. The user saves the new configuration by clicking in the **Add** button. The user can also remove a previous connection configuration.

After configuring all the connection fields (host, port, database schema, user and password) the **Connect** button must be pressed. A new item of datatype Database Connection is added to the clipboard and the view for this datatype includes information about the host, port and database name.

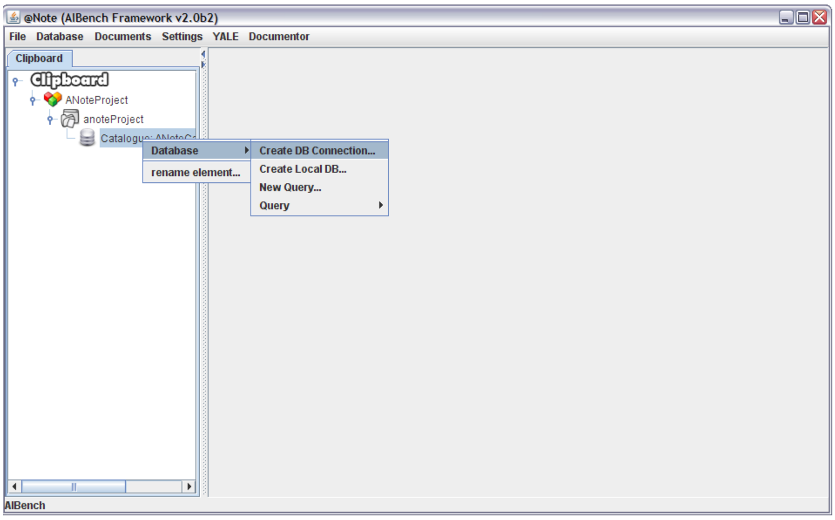
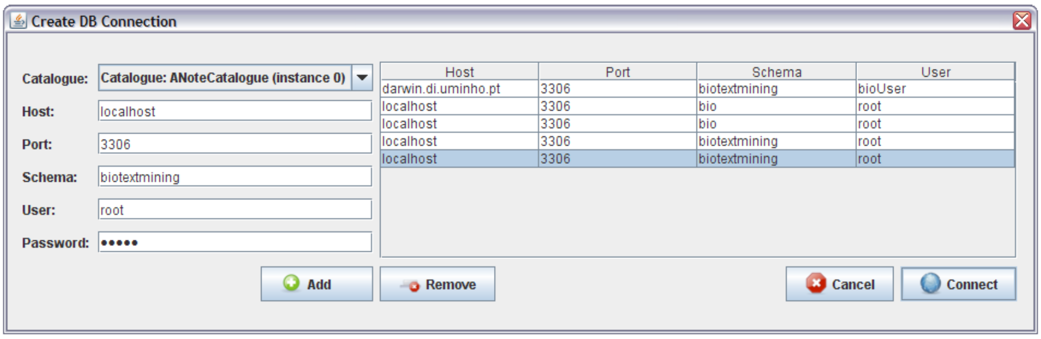


Figure 2.5: Creating a database connection



Catalogue: Catalogue: ANoteCatalogue (instance 0)

Host: localhost

Port: 3306

Schema: biotextmining

User: root

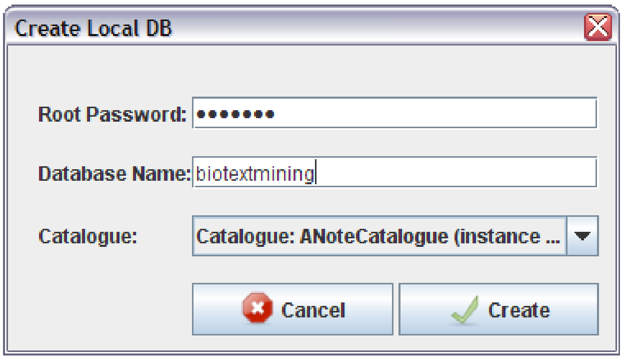
Password: •••••

Host	Port	Schema	User
darwin.di.uminho.pt	3306	biotextmining	bioUser
localhost	3306	bio	root
localhost	3306	bio	root
localhost	3306	biotextmining	root
localhost	3306	biotextmining	root

Figure 2.6: Connecting to an existing database.

2.2.2 Creating a local database

To create a new local database, the user selects the option **Create Local DB**. In the popup window, two fields have be to filled (Figure 2.7). The first is the *MySql* root password on the local host (given in the instalation) and the second is the new database name. When all the fields have been set and the **Create** button is pressed, @Note will create a new database predefined schema. As before an object representing the new connection will be added to the clipboard.



Root Password: •••••••

Database Name: biotextmining

Catalogue: Catalogue: ANoteCatalogue (instance ...

Figure 2.7: Creating a new local database.

2.2.3 Loading default data

The option **Load sample data** in the **Database** menu loads some pre-defined data, allowing for the beginner user to get acquainted with the application with a reduced effort and time.

In this case, the results of a pre-defined query to PubMed are loaded into the catalogue (the query uses the keywords *Escherichia coli stringent response*) and a dictionary with terms to the organism *E. coli* is loaded to the database. The data source used for this dictionary is the BioWarehouse integrated repository.

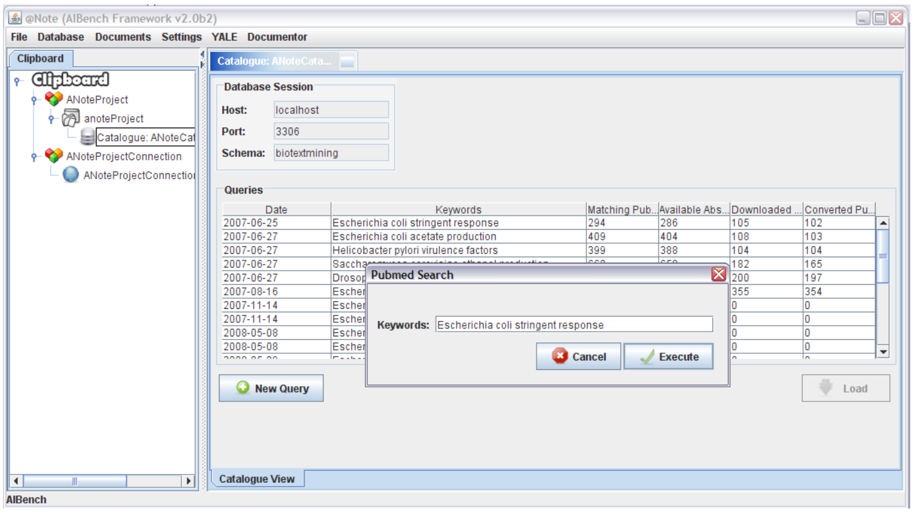
2.3 PubMed Searches

To perform PubMed searches, a database connection has to be previously and successfully established. The user clicks in the project's catalogue, and a view will appear in the working area of the application. A list of the database's existing queries is given.

If none of the listed queries is wanted, the user can add a new query (pressing the **New Query** button). This option is also available in the menu **Database**, option **New Query**.

A new *PubMed* search will be performed using the keywords selected by the user in the popup window. The **Execute** button starts the search process. This new query, if succeed, will be added to the previous list.

A query has an associated list of publications. The user can select the query he intends to work on and click in the **Load** button. This action will load the information about all the publications of the selected query. Information about these publications will be listed on a new datatype item, named **ResultSet** that is loaded into the clipboard. By clicking in the



Database Session

Host: localhost

Port: 3306

Schema: biotextmining

Queries

Date	Keywords	Matching Pub...	Available Abs...	Downloaded ...	Converted Pu...
2007-06-25	Escherichia coli stringent response	294	286	105	102
2007-06-27	Escherichia coli acetate production	409	404	108	103
2007-06-27	Helicobacter pylori virulence factors	399	388	104	104
2007-06-27	Saccha			182	165
2007-06-27	Drosop			200	197
2007-08-16	Escher			355	354
2007-11-14	Escher			0	0
2007-11-14	Escher			0	0
2008-05-08	Escher			0	0
2008-05-08	Escher			0	0

Pubmed Search

Keywords: Escherichia coli stringent response

Figure 2.8: Listing PubMed queries.

ResultSet, the user can analyse the set of loaded publications.

2.4 Journal Retrieval

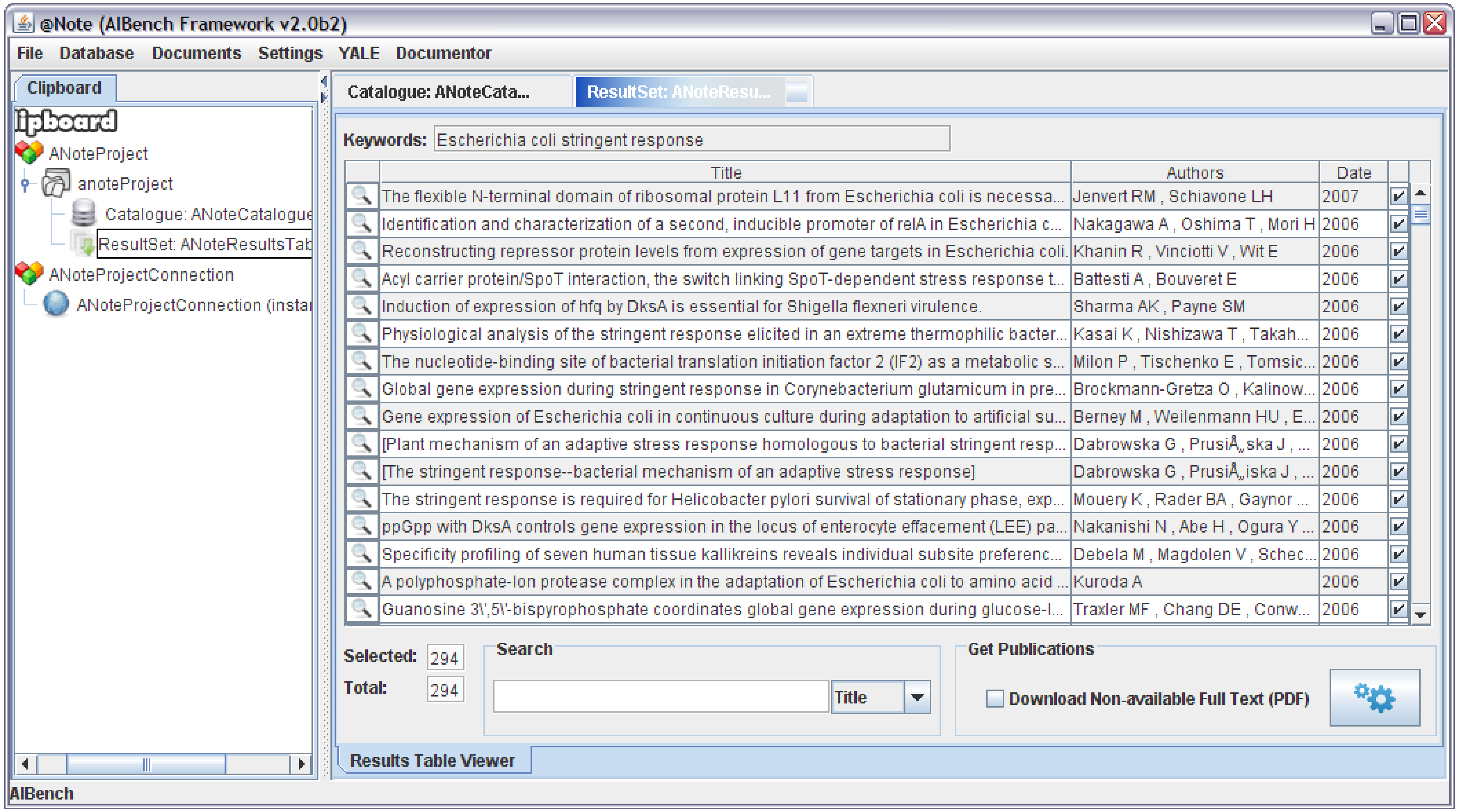
In the **ResultSet** view, the list of publications is presented. This list contains all the publications that were selected from PubMed using the original query. In this step, it is possible to select what are the publications the user really wants to retrieve to the project. Each line of the view's table, corresponds to one publication, and contains the title, author's list and date of the publication.

If this information is not sufficient for the user to decide if he/she wants to get the publication, more detailed information about a publication can be viewed by clicking on the leftmost side button on the publication's row.

The publication information view shows the available data about the publication, and also implements two other features:

- to view the PDF document;
- to view and edit the publication's relevance to the query.

In case the publication holds the respective **PDF** document locally, i.e. the pdf file is in the project local document's folder, it is possible to visualise this PDF. This typically occurs when a previous **Journal Retrieval** (JR) process has been performed. In this case, the **PDF** button will be enabled, and the user can click it to see the document (Figure 2.11).



Keywords: Escherichia coli stringent response

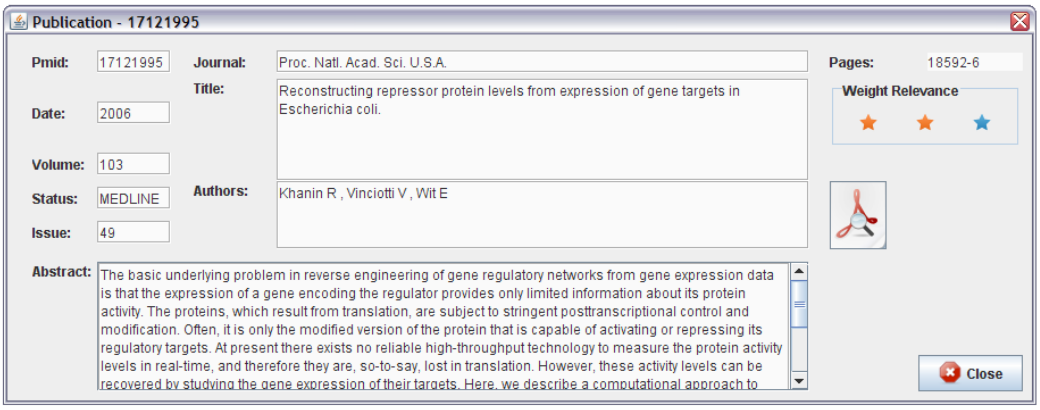
Title	Authors	Date	
The flexible N-terminal domain of ribosomal protein L11 from Escherichia coli is necessa...	Jenvert RM , Schiavone LH	2007	☑
Identification and characterization of a second, inducible promoter of relA in Escherichia c...	Nakagawa A , Oshima T , Mori H	2006	☑
Reconstructing repressor protein levels from expression of gene targets in Escherichia coli.	Khanin R , Vinciotti V , Wit E	2006	☑
Acyl carrier protein/SpoT interaction, the switch linking SpoT-dependent stress response t...	Battesti A , Bouveret E	2006	☑
Induction of expression of hfq by DksA is essential for Shigella flexneri virulence.	Sharma AK , Payne SM	2006	☑
Physiological analysis of the stringent response elicited in an extreme thermophilic bacter...	Kasai K , Nishizawa T , Takah...	2006	☑
The nucleotide-binding site of bacterial translation initiation factor 2 (IF2) as a metabolic s...	Milon P , Tischenko E , Tomsic...	2006	☑
Global gene expression during stringent response in Corynebacterium glutamicum in pre...	Brockmann-Gretza O , Kalinow...	2006	☑
Gene expression of Escherichia coli in continuous culture during adaptation to artificial su...	Berney M , Weilenmann HU , E...	2006	☑
[Plant mechanism of an adaptive stress response homologous to bacterial stringent resp...	Dabrowska G , PrusiÅ„ska J , ...	2006	☑
[The stringent response--bacterial mechanism of an adaptive stress response]	Dabrowska G , PrusiÅ„iska J , ...	2006	☑
The stringent response is required for Helicobacter pylori survival of stationary phase, exp...	Mouery K , Rader BA , Gaynor ...	2006	☑
ppGpp with DksA controls gene expression in the locus of enterocyte effacement (LEE) pa...	Nakanishi N , Abe H , Ogura Y ...	2006	☑
Specificity profiling of seven human tissue kallikreins reveals individual subsite preferenc...	Debela M , Magdolen V , Schec...	2006	☑
A polyphosphate-lon protease complex in the adaptation of Escherichia coli to amino acid ...	Kuroda A	2006	☑
Guanosine 3\',5\'-bispyrophosphate coordinates global gene expression during glucose-l...	Traxler MF , Chang DE , Conw...	2006	☑

Selected: 294

Total: 294

□ Download Non-available Full Text (PDF)

Figure 2.9: Viewing a Result Set.



Publication - 17121995

Pmid: 17121995

Journal: Proc. Natl. Acad. Sci. U.S.A.

Pages: 18592-6

Title: Reconstructing repressor protein levels from expression of gene targets in Escherichia coli.

Date: 2006

Volume: 103

Status: MEDLINE

Authors: Khanin R , Vinciotti V , Wit E

Issue: 49

Abstract: The basic underlying problem in reverse engineering of gene regulatory networks from gene expression data is that the expression of a gene encoding the regulator provides only limited information about its protein activity. The proteins, which result from translation, are subject to stringent posttranscriptional control and modification. Often, it is only the modified version of the protein that is capable of activating or repressing its regulatory targets. At present there exists no reliable high-throughput technology to measure the protein activity levels in real-time, and therefore they are, so-to-say, lost in translation. However, these activity levels can be recovered by studying the gene expression of their targets. Here, we describe a computational approach to

Figure 2.10: Checking detailed information about a publication.

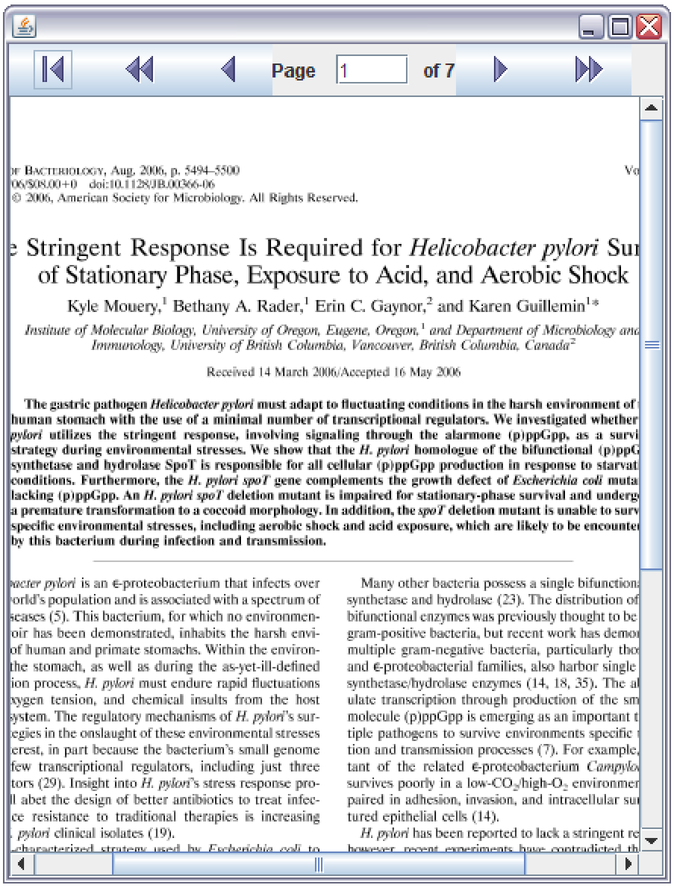
In the publication's details view, a weight relevance measure is presented. If the publication belongs to more than one query, a average of all relevance for those queries is calculated and presented. The user can visualise the actual relevance of the document for each query it belongs to and edit it (Figure 2.13). If the same relevance is pretended for all the queries, it is possible to select all queries and then choose the relevance level.

Let us now define how the user can retrieve the documents from the editor's sites (when these are available according to the user's permissions). By default, all the publications in the Result Set are selected, but the user can select the intended publications. If the **Download Non-available Full Text (PDF)** (bottom right) option is selected, the application will invoke the **Journal Retrieval** operation.

The Journal Retrieval operation will try to find, on the Web, the **PDFs** of the selected publications. For each document found, the application will download it to the project's local directory and this **PDFs** will be available for future work. After downloading all the **PDFs** available, a **pdf** to text conversion will be conducted.

By default, this option is not selected, because this take a few minutes to process. When the user presses the "Get Publications" button, if JR option was selected, the preceding process will be done, and the selected publication will be loaded to the application.

In the end of the process, a window will be presented to the user that will choose the **Publication Set** where the publications will be loaded. It is possible to add new publications to an existing **Publication Set** or to create a new one (Figure 2.13]. A Publication Set can have documents coming from distinct queries and it is also possible to add previously non selected publications from the same query.



ᴏғ Bᴀᴄᴛᴇʀɪᴏʟᴏɢʏ, Aug. 2006, p. 5494–5500 Vo
06/$08.00+0 doi:10.1128/JB.00366-06

e Stringent Response Is Required for *Helicobacter pylori* Su
of Stationary Phase, Exposure to Acid, and Aerobic Shock

Kyle Mouery,[1] Bethany A. Rader,[1] Erin C. Gaynor,[2] and Karen Guillemin[1]*
Institute of Molecular Biology, University of Oregon, Eugene, Oregon,[1] *and Department of Microbiology and*
Immunology, University of British Columbia, Vancouver, British Columbia, Canada[2]

Received 14 March 2006/Accepted 16 May 2006

The gastric pathogen *Helicobacter pylori* must adapt to fluctuating conditions in the harsh environment of
human stomach with the use of a minimal number of transcriptional regulators. We investigated whether
***pylori* utilizes the stringent response, involving signaling through the alarmone (p)ppGpp, as a survi**
strategy during environmental stresses. We show that the *H. pylori* homologue of the bifunctional (p)ppG
synthetase and hydrolase SpoT is responsible for all cellular (p)ppGpp production in response to starvat
conditions. Furthermore, the *H. pylori spoT* gene complements the growth defect of *Escherichia coli* muta
lacking (p)ppGpp. An *H. pylori spoT* deletion mutant is impaired for stationary-phase survival and underg
a premature transformation to a coccoid morphology. In addition, the *spoT* deletion mutant is unable to surv
specific environmental stresses, including aerobic shock and acid exposure, which are likely to be encounte
by this bacterium during infection and transmission.

acter pylori is an ϵ-proteobacterium that infects over
orld's population and is associated with a spectrum of
seases (5). This bacterium, for which no environmen-
oir has been demonstrated, inhabits the harsh envi-
of human and primate stomachs. Within the environ-
the stomach, as well as during the as-yet-ill-defined
ion process, *H. pylori* must endure rapid fluctuations
xygen tension, and chemical insults from the host
ystem. The regulatory mechanisms of *H. pylori*'s sur-
egies in the onslaught of these environmental stresses
erest, in part because the bacterium's small genome
few transcriptional regulators, including just three
tors (29). Insight into *H. pylori*'s stress response pro-
l abet the design of better antibiotics to treat infec-
ce resistance to traditional therapies is increasing
. *pylori* clinical isolates (19).
-characterized strategy used by *Escherichia coli* to

Many other bacteria possess a single bifunctiona
synthetase and hydrolase (23). The distribution of
bifunctional enzymes was previously thought to be
gram-positive bacteria, but recent work has demo
multiple gram-negative bacteria, particularly tho
and ϵ-proteobacterial families, also harbor single
synthetase/hydrolase enzymes (14, 18, 35). The a
ulate transcription through production of the sm
molecule (p)ppGpp is emerging as an important t
tiple pathogens to survive environments specific
tion and transmission processes (7). For example,
tant of the related ϵ-proteobacterium *Campylo*
survives poorly in a low-CO_2/high-O_2 environme
paired in adhesion, invasion, and intracellular su
tured epithelial cells (14).
H. pylori has been reported to lack a stringent re
however, recent experiments have contradicted th

Figure 2.11: Viewing the PDF file of a publication.

If a new Publication Set is selected, a new instance of Publication Set will be added to the clipboard. All the instances of that type will be squat on a root object of the type **WorkingSets**.

2.5 Named Entity Recognition

When there are one or more Publication Sets available in a project, it is possible to execute the **Named Entity Recognition** (NER) operation over one of these sets (right clicking it). When the user clicks on a Publication set, a view is presented with information about the publications added to it (Figure 2.14) and some more information about the sets of processed documents associated to it.

When the "@" button in the view is pressed, the **Txt Structuring and NER** option on the **Document** menu or by right clicking on a **Publication**

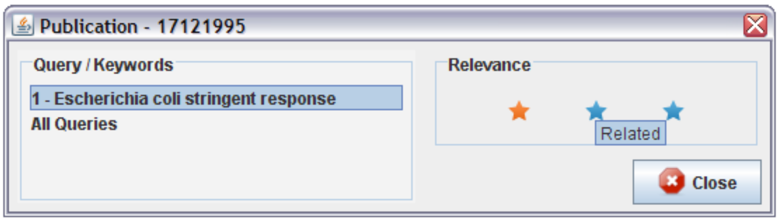


Figure 2.12: Setting the publication's relevance.

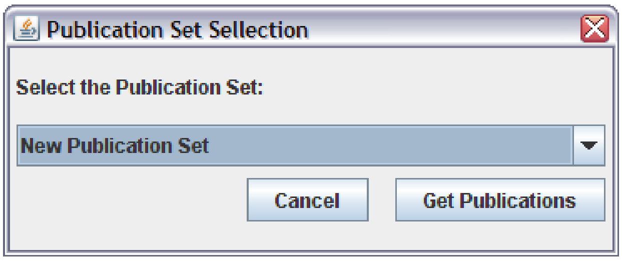


Figure 2.13: Selecting the publication set.

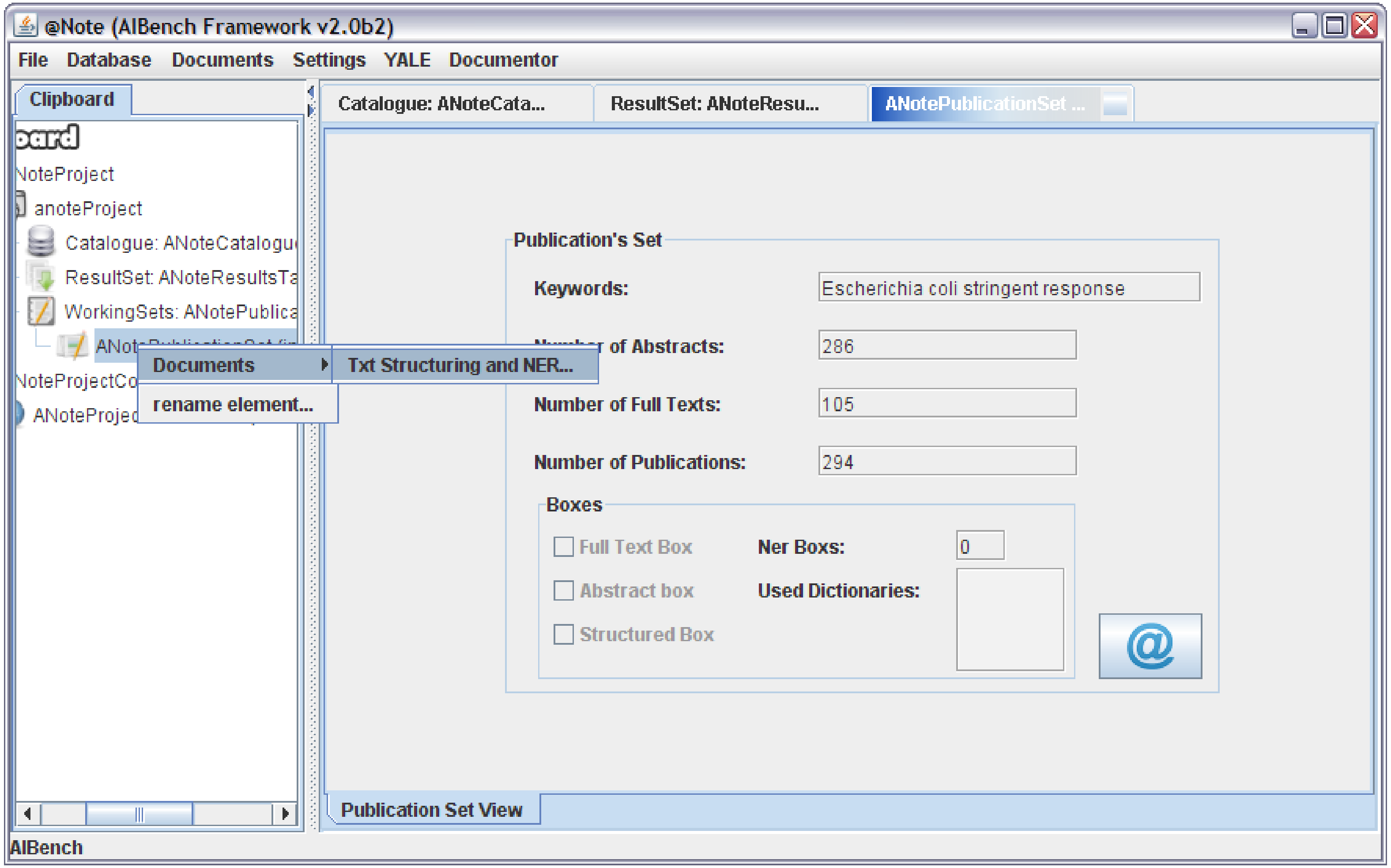
Set item of the clipboard, a wizard will be presented. This allows to configure the NER process. The first step is to select the **Publication Set** over which the **NER** will be performed (Figure 2.15). When the desired **Publication Set** is selected, the **Next** button is pressed.

In the next step, a dictionary must be selected for the **NER**. Here, a new dictionary can be imported (how to import dictionaries will be described later in this document). After the dictionary has been chosen, the list of possible classes will be presented. The user selects the classes to annotate by moving them from the left to the right list.

In the next step (Figure 2.17), a set of complementary classes that the user can choose to be annotated are presented. Those are classes which are given by lists of terms manually compiled. The available options are:

- Biology-related Verbs;
- Laboratory Techniques;
- Physiological States;
- Predefined Expert Hand Rules.

In the same window the user defines if he decides to annotate abstracts or full texts.



Publication's Set

Keywords: Escherichia coli stringent response

Number of Abstracts: 286

Number of Full Texts: 105

Number of Publications: 294

Boxes

☐ Full Text Box

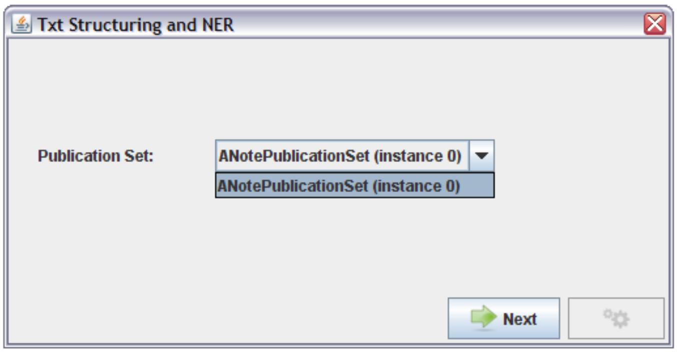
☐ Abstract box

☐ Structured Box

Ner Boxs: 0

Used Dictionaries:

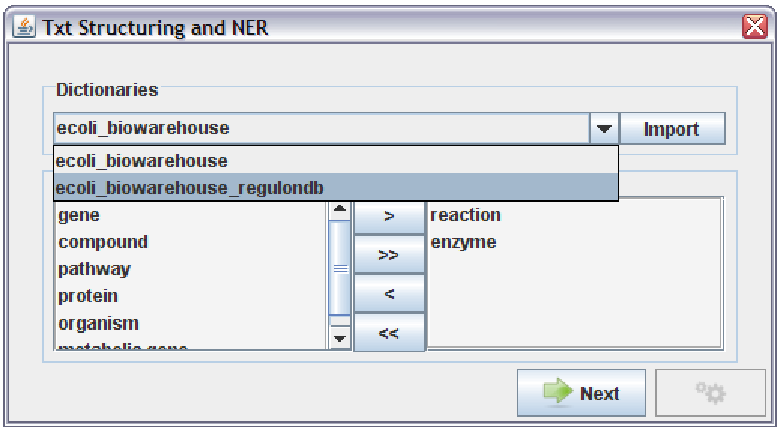
Figure 2.14: Viewing the Publication Set.



Publication Set: ANotePublicationSet (instance 0)

ANotePublicationSet (instance 0)

Figure 2.15: Selecting the Publication Set for NER.



ecoli_biowarehouse

ecoli_biowarehouse
ecoli_biowarehouse_regulondb

gene
compound
pathway
protein
organism

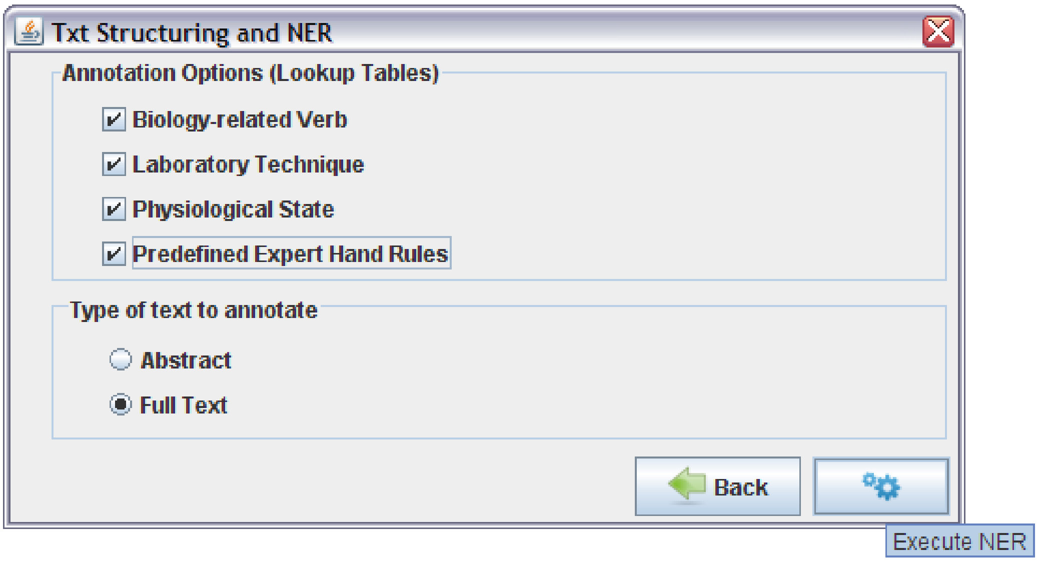
reaction
enzyme

Figure 2.16: Selecting the dictionary and the classes to annotate.

After all the configurations have been made, the **Execute** button (gear icon) has to be pressed. When the button is pressed, the **NER** operation will start and a small window will appear, indicating the execution of the operation (Figure 2.18). The **NER** operation will take a few minutes.

When the process is finished, a new **Ner Box List** object will be added to the clipboard. This object contains a list of items of the datatype **ANoteNerBox**, each being the result of a **NER** operation. The **Ner Box List** exists because it is possible to create different kinds of configurations to **NER** (e.g. distinct dictionaries), and each configuration yields a distinct **NerBox**.

By clicking on a **NERBox** in the clipboard, the respective view window is presented (Figure 2.19). In the upper part of this window the keywords that originated the original **Publication Set** are given. The used dictionary, the annotated entities, the number of publications annotated and all the



Txt Structuring and NER

Annotation Options (Lookup Tables)

☑ Biology-related Verb

☑ Laboratory Technique

☑ Physiological State

☑ Predefined Expert Hand Rules

Type of text to annotate

○ Abstract

● Full Text

Figure 2.17: Annotation options.

annotation options are also presented. In the bottom part of the window, there are two sections. The **Search** section allows to search a publication in the list. A search can be done by different contents that can be selected in the list at the right hand side of the search's text field. If there are matches between the text typed in the text field and the document's selected content, the matched publications will be highlighted.

The **View** section shows the types of documents that is possible to choose. The types are:

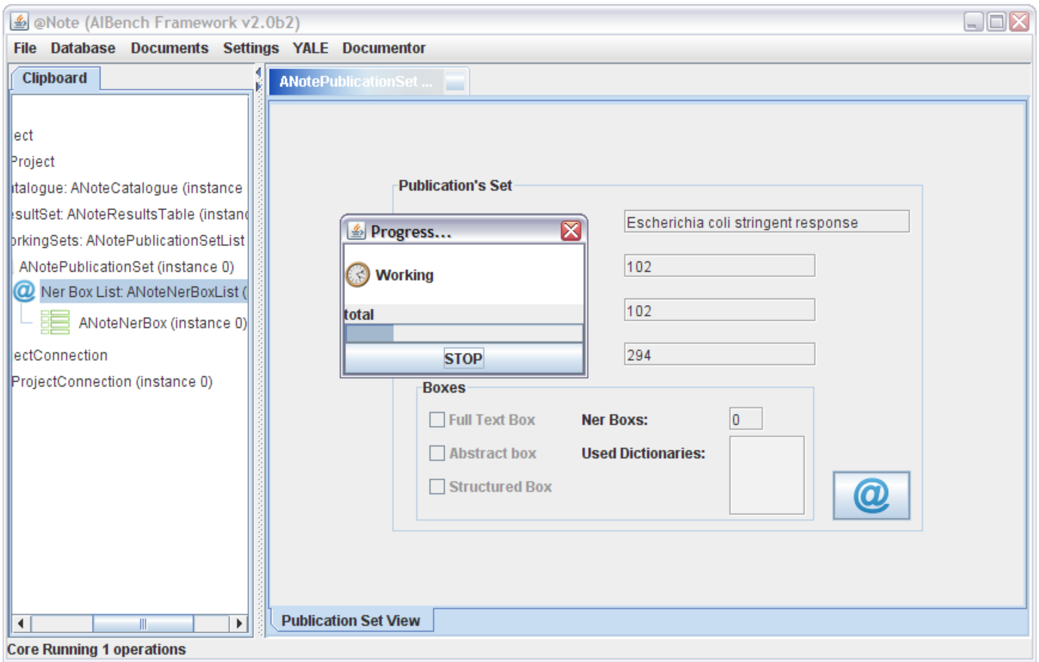
- Abstract: the publication's abstract, without any annotation;

- Full Text: the unstructured full text of the publication without any annotation, this is the direct result of the **PdfToTxt** operation;

- Structured: the entire text of the publication without annotations, but with a base structuring, i.e., the text is split in the areas containing the title, authors, abstract, paper sections and others;

- NER: in case of **NER** been made to abstracts, this shows the publication's annotated abstract, if the **NER** was made over full texts, it shows the entire annotated document.

To view a document, the user just has to click in the right publication's row button. The type of document that will be opened is the selected in the **View** section.



ect
Project
talogue: ANoteCatalogue (instance
sultSet: ANoteResultsTable (instanc
rkingSets: ANotePublicationSetList
ANotePublicationSet (instance 0)
Ner Box List: ANoteNerBoxList (
ANoteNerBox (instance 0)
ectConnection
ProjectConnection (instance 0)

Escherichia coli stringent response
102
102
294

□ Full Text Box
□ Abstract box
□ Structured Box

Ner Boxs: 0
Used Dictionaries:

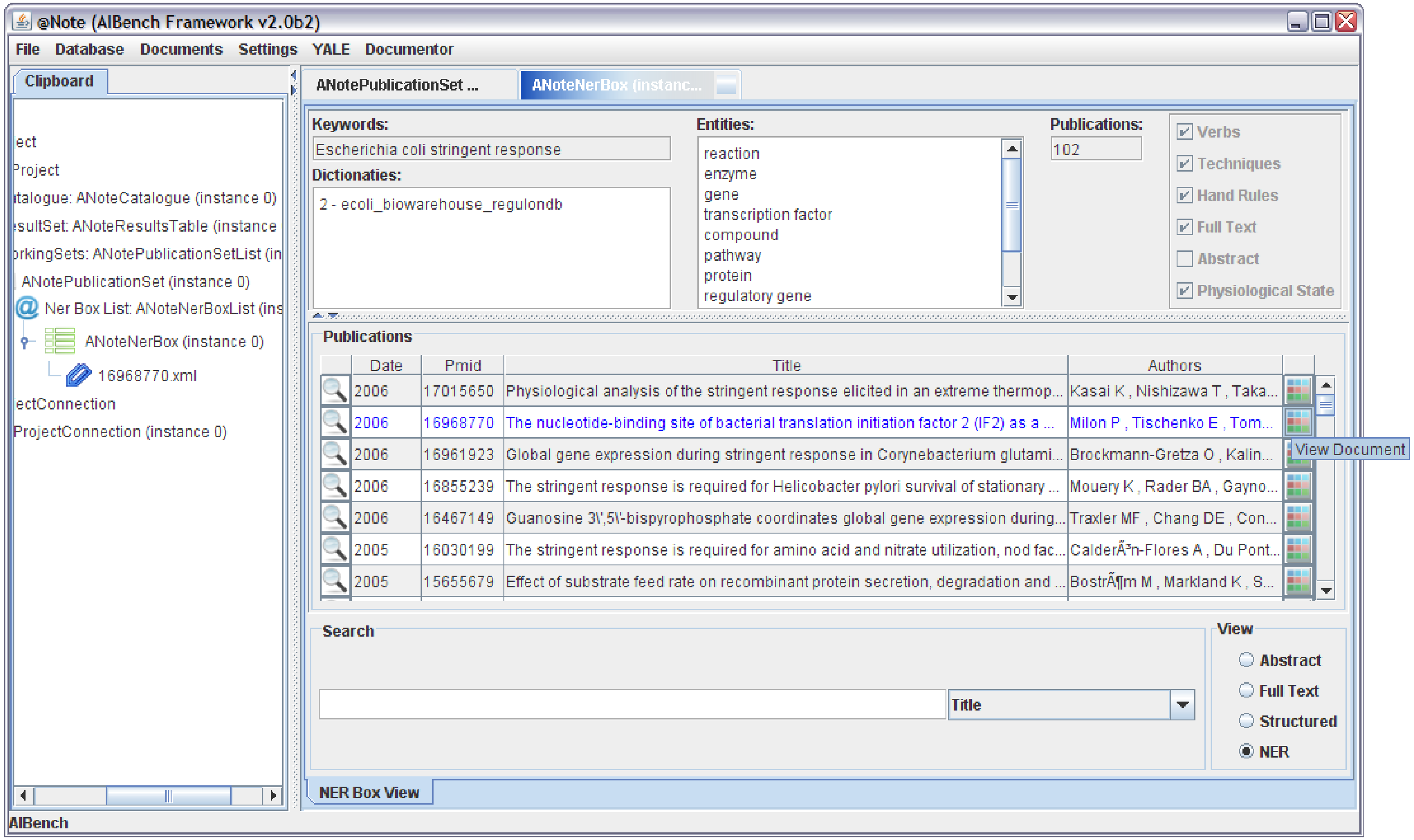
Figure 2.18: NER running operation.

2.5.1 Document View

When the user selects a document, an item representing this document is added to the clipboard, under the tree of the respective **ANoteNerBox** represented by its name. The **PublicationSet** item on the clipboard will have nested boxes of documents. There are four types of boxes that a **PublicationSet** can enclose, namely:

- ANoteNerBox: box with abstracts or fulltexts curated by **NER**;
- Structured Text Box: box with structured documents, but without annotations;
- Full Text Box: box with unstructured documents and without annotations;
- Abstract Box: box with just the abstracts without annotations.

To view the document, the user has to click on the document's item in the clipboard and a view will be opened. The document's view is structured in the following sections:



@Note (AIBench Framework v2.0b2)
File Database Documents Settings YALE Documentor
Clipboard
ect
Project
talogue: ANoteCatalogue (instance 0)
sultSet: ANoteResultsTable (instance
rkingSets: ANotePublicationSetList (in
ANotePublicationSet (instance 0)
Ner Box List: ANoteNerBoxList (ins
ANoteNerBox (instance 0)
16968770.xml
ectConnection
ProjectConnection (instance 0)
ANotePublicationSet ...
ANoteNerBox (instanc...
View Document
NER Box View
AIBench

Keywords: Escherichia coli stringent response

Dictionaties: 2 - ecoli_biowarehouse_regulondb

Entities: reaction, enzyme, gene, transcription factor, compound, pathway, protein, regulatory gene

Publications: 102

☑ Verbs
☑ Techniques
☑ Hand Rules
☑ Full Text
☐ Abstract
☑ Physiological State

Publications

Date	Pmid	Title	Authors
2006	17015650	Physiological analysis of the stringent response elicited in an extreme thermop...	Kasai K , Nishizawa T , Taka...
2006	16968770	The nucleotide-binding site of bacterial translation initiation factor 2 (IF2) as a ...	Milon P , Tischenko E , Tom...
2006	16961923	Global gene expression during stringent response in Corynebacterium glutami...	Brockmann-Gretza O , Kalin...
2006	16855239	The stringent response is required for Helicobacter pylori survival of stationary ...	Mouery K , Rader BA , Gayno...
2006	16467149	Guanosine 3\',5\'-bispyrophosphate coordinates global gene expression during...	Traxler MF , Chang DE , Con...
2005	16030199	The stringent response is required for amino acid and nitrate utilization, nod fac...	CalderÃ³n-Flores A , Du Pont...
2005	15655679	Effect of substrate feed rate on recombinant protein secretion, degradation and ...	BostrÃ¶m M , Markland K , S...

Search

Title ▾

View
○ Abstract
○ Full Text
○ Structured
● NER

Figure 2.19: ANoteNerBox view.

- A section with buttons to save the changes done in the document, doing zoon in and zoon out in the text, undo the last change carried out and a field to search text's excerpt in the document;
- A section to change the colours of annotated entities;
- A section with the annotated classes, and the terms of each class;
- A section with the structure of the text, the user can click on a section and skip to the respective section in the text;
- A central section with the text.

2.6 Manual Annotation

A number of options are available to the user under the document's view described in the previous section. These allow the manual curation of the automatic NER annotations.

2.6.1 Annotating a new term

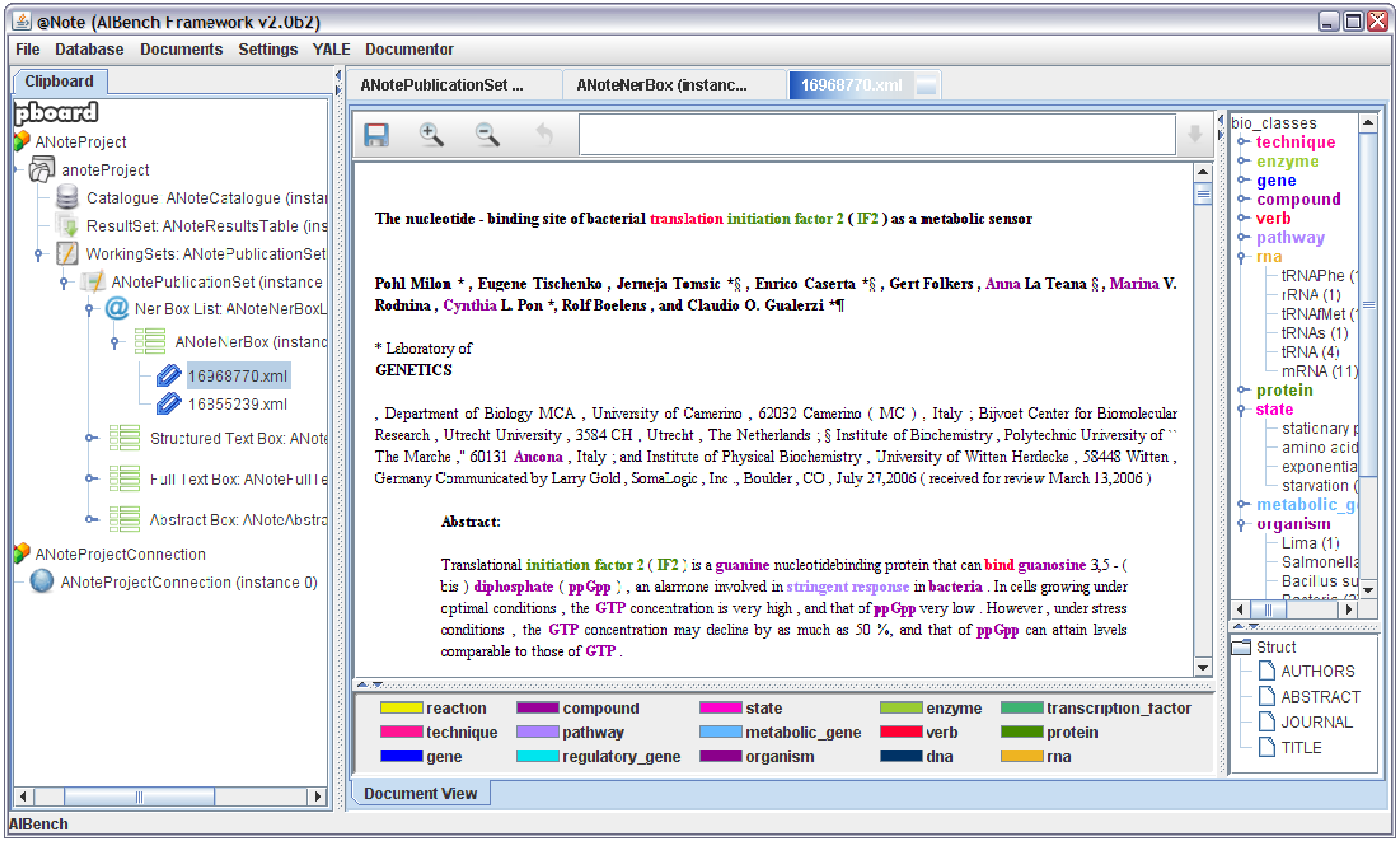
To annotate a new term, the user must select the term and a popup window will appear with the possible options, i.e. biological classes. The **Add Tag** option must be chosen and the intended class is selected. If the selected term is already annotated, it can't be annotated again.

After adding a new annotation, the new term can be added to the currently used dictionary if that is intended by the user. The changes will be made in the underlying database supporting the dictionary and can therefore be used to annotate other documents in the future.

The diagram depicted in Figure 2.23 explains how this option works and the effects of the user's choices over the dictionary.

2.6.2 Correcting an annotation

It is also possible for the user to correct an annotation. This correction can be done in the lexical form of the term or in the class that the term is annotated. To do so, the user selects the term and chooses the **Correct Tag** option in the popup displayed. When this option is selected, a window appears where the user can correct the annotation. The window contains the current class of the term, the new term, initially identical to the selected term and a list with all the classes that the user can choose to the term.



The nucleotide - binding site of bacterial translation initiation factor 2 (IF2) as a metabolic sensor

Pohl Milon * , Eugene Tischenko , Jerneja Tomsic *§ , Enrico Caserta *§ , Gert Folkers , Anna La Teana § , Marina V. Rodnina , Cynthia L. Pon * , Rolf Boelens , and Claudio O. Gualerzi *¶

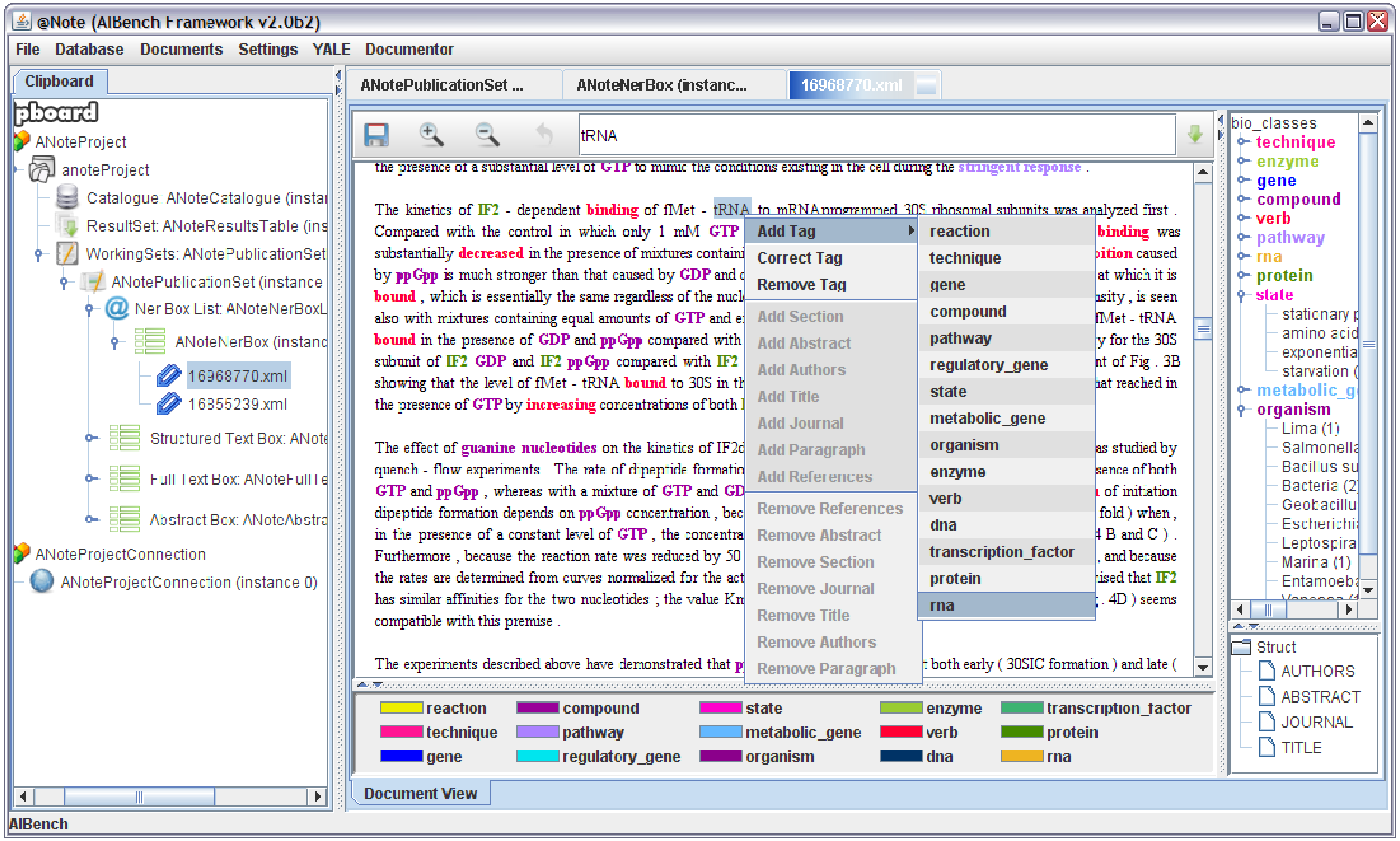
* Laboratory of
GENETICS

, Department of Biology MCA , University of Camerino , 62032 Camerino (MC) , Italy ; Bijvoet Center for Biomolecular Research , Utrecht University , 3584 CH , Utrecht , The Netherlands ; § Institute of Biochemistry , Polytechnic University of `` The Marche ,'' 60131 Ancona , Italy ; and Institute of Physical Biochemistry , University of Witten Herdecke , 58448 Witten , Germany Communicated by Larry Gold , SomaLogic , Inc ., Boulder , CO , July 27,2006 (received for review March 13,2006)

Abstract:

Translational initiation factor 2 (IF2) is a guanine nucleotidebinding protein that can bind guanosine 3,5 - (bis) diphosphate (ppGpp) , an alarmone involved in stringent response in bacteria . In cells growing under optimal conditions , the GTP concentration is very high , and that of ppGpp very low . However , under stress conditions , the GTP concentration may decline by as much as 50 %, and that of ppGpp can attain levels comparable to those of GTP .

Figure 2.20: Document view.



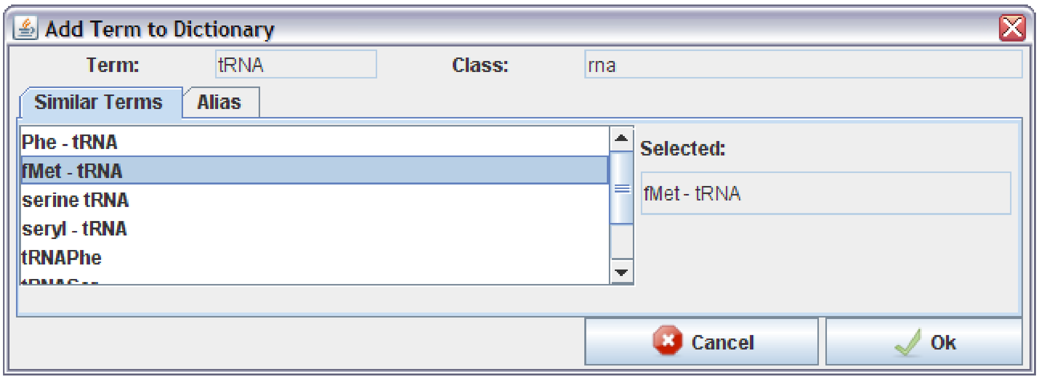
the presence of a substantial level of GTP to mimic the conditions existing in the cell during the stringent response .

The kinetics of IF2 - dependent binding of fMet - tRNA to mRNAprogrammed 30S ribosomal subunits was analyzed first .
Compared with the control in which only 1 mM GTP binding was
substantially decreased in the presence of mixtures containi ition caused
by ppGpp is much stronger than that caused by GDP and c at which it is
bound , which is essentially the same regardless of the nucl sity , is seen
also with mixtures containing equal amounts of GTP and e fMet - tRNA
bound in the presence of GDP and ppGpp compared with y for the 30S
subunit of IF2 GDP and IF2 ppGpp compared with IF2 nt of Fig . 3B
showing that the level of fMet - tRNA bound to 30S in th hat reached in
the presence of GTP by increasing concentrations of both I

The effect of guanine nucleotides on the kinetics of IF2d as studied by
quench - flow experiments . The rate of dipeptide formatio sence of both
GTP and ppGpp , whereas with a mixture of GTP and GD of initiation
dipeptide formation depends on ppGpp concentration , bec fold) when ,
in the presence of a constant level of GTP , the concentra 4 B and C) .
Furthermore , because the reaction rate was reduced by 50 , and because
the rates are determined from curves normalized for the act ised that IF2
has similar affinities for the two nucleotides ; the value Kn . 4D) seems
compatible with this premise .

The experiments described above have demonstrated that p t both early (30SIC formation) and late (

Figure 2.21: Adding a NER term annotation.



Term: tRNA

Class: rna

Phe - tRNA
fMet - tRNA
serine tRNA
seryl - tRNA
tRNAPhe

Selected: fMet - tRNA

Figure 2.22: Dictionary enrichment

To change the class of the term, the user just has to select one of the classes in the given list. If the user wants to correct the term, she/he can edit the term in the **New Text** field; if not, she/he just clicks the **Apply** button without editing the term.

When the **Apply** button is pressed, the changes will be made and the window to add a term to the dictionary will appear. The process of adding a term to the dictionary is the same as described above.

2.6.3 Removing an annotation

If the user knows that the term's annotation is incorrect and that the term should not be annotated with any of the possible classes, she/he can remove the annotation of that term. To do that, the user has just to select the term and choose the option **Remove Tag** from the popup. This action will only remove the annotation but not the term from the dictionary.

2.7 Handling lexical resources

The menu **Lexical resources** contains a number of operations to manage the lexical resources of a project, namely dictionaries and lookup tables. Both the set of dictionaries and lookup tables are represented by clipboard objects and the current state can be viewed by clicking on each of these datatypes.

The sub-menus **Dictionaries** and **Lookup tables** handle the operations regarding each resource type. The options for each case are given below.

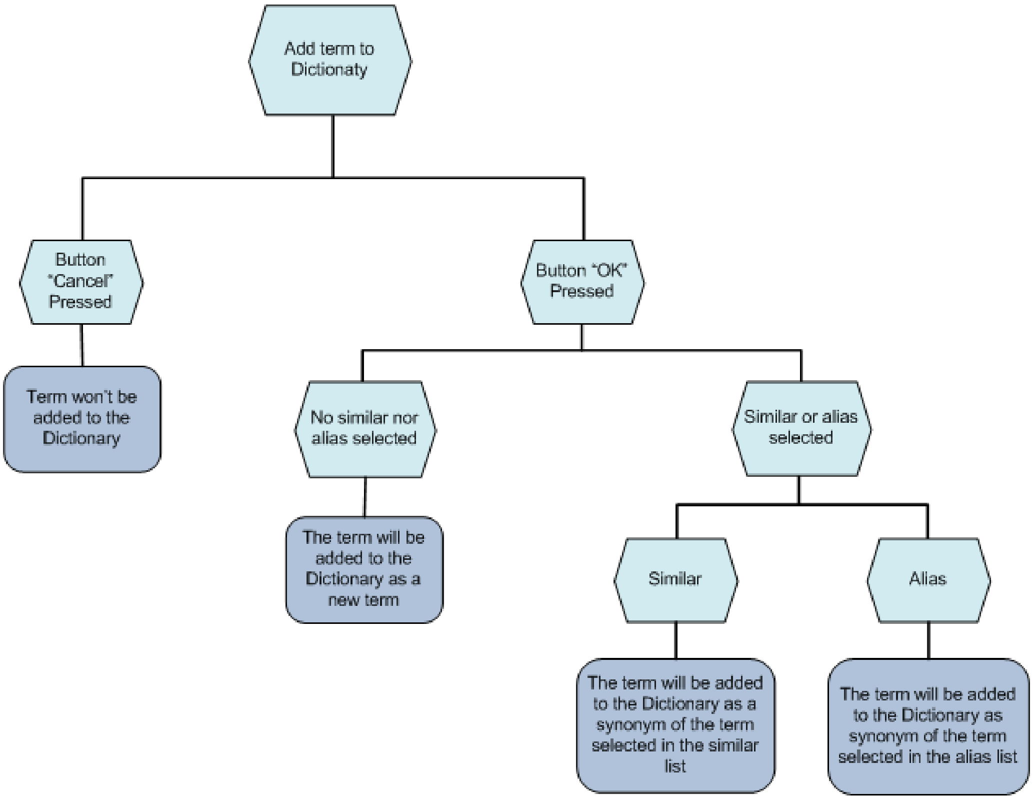

Add term to Dictionaty
Button "Cancel" Pressed
Button "OK" Pressed
Term won't be added to the Dictionary
No similar nor alias selected
Similar or alias selected
The term will be added to the Dictionary as a new term
Similar
Alias
The term will be added to the Dictionary as a synonym of the term selected in the similar list
The term will be added to the Dictionary as synonym of the term selected in the alias list

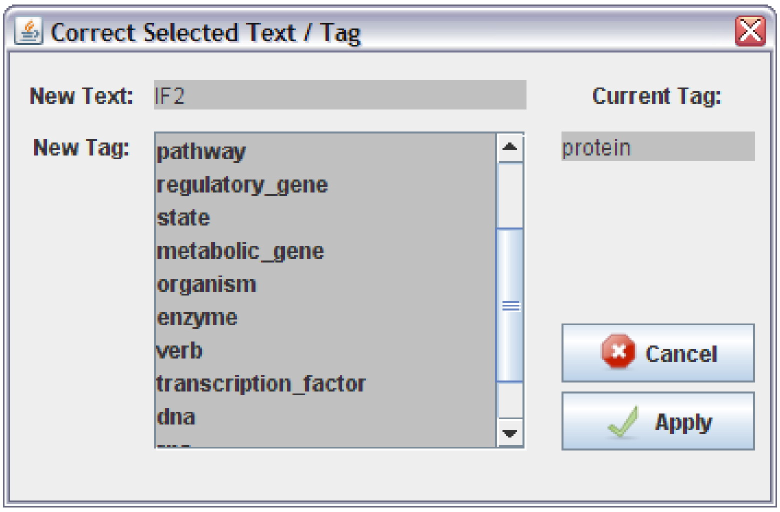
Figure 2.23: Diagram illustrating the options of adding new terms to the dictionary.

2.7.1 Handling dictionaries

Three distinct operations can be performed in dictionary management, that are given by three options in the **Dictionaries** sub-menu (or by right clicking a dictionary object):

- **New dictionary**: creates a new (empty!) dictionary in the project.
- **Dictionary contents**: allows the user to add contents to a dictionary, which can come from several sources.
- **Merging dictionaries**: allows the user to merge the contents of several dictionaries into a new one (only allowed for dictionaries where the sets of classes do not overlap!).

The second option (adding contents) deserves a more complete explanation. The process starts by the selection of the dictionary where the contents



New Text: IF2

Current Tag: protein

New Tag: pathway, regulatory_gene, state, metabolic_gene, organism, enzyme, verb, transcription_factor, dna

Figure 2.24: Correcting a term annotation.

will be added. In the bottom part, the data source is configured. Currently, the system supports the following sources:

- BioWarehouse integrated databases (`http://biowarehouse.ai.sri.com/`);
- BioCyc flatfile(`http://biocyc.org/download.shtml`) ;
- ChEBI flatfiles (`http://www.ebi.ac.uk/chebi/downloadsForward.do`);
- NCBI Taxnonomy flatfiles (`ftp://ftp.ncbi.nih.gov/pub/taxonomy/`);
- UniProtKB/Swiss-Prot flatfiles (`http://www.uniprot.org/downloads`);
- MGI Entrez Genes flatfiles (`ftp://ftp.informatics.jax.org/pub/reports/index.html`).

The user may choose to upload all contents available at a given source or, according to source specifications, he can restrict data upload to a given organism and a subset of the embraced classes.

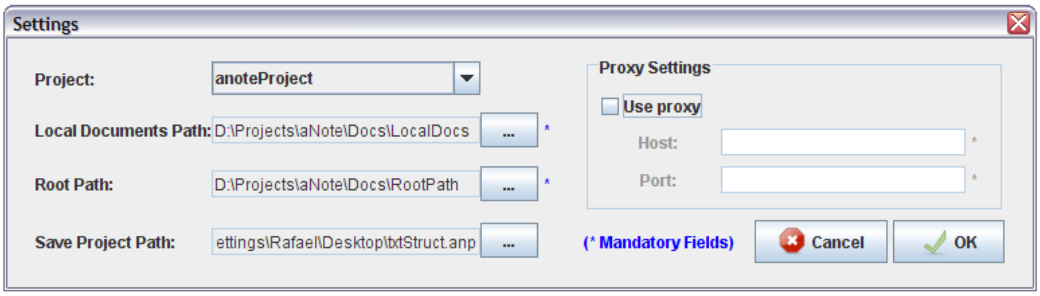
2.7.2 Loading lookup tables

This option allows to load pre-defined lookup tables for a number of biological classes. These are, at current version, only available for three classes: biological related verbs, physiological states and experimental techniques.

2.8 Project Settings

It is possible to change the settings of a project after it has been created. To do so, the user should select the **Settings** option on the menu bar and then **Project Settings**, or click with the right button over a project on the clipboard.

In the popup window (Figure 2.25) it is possible to change the location of the project's local documents, the root directory of the documents, the file to where the project could be saved and editing the proxy's configuration. To define the host and port for the proxy, the user has to activate the option **Use proxy** and then type the host and port.



Settings

Project: anoteProject

Local Documents Path: D:\Projects\aNote\Docs\LocalDocs *

Root Path: D:\Projects\aNote\Docs\RootPath *

Save Project Path: ettings\Rafael\Desktop\txtStruct.anp

Proxy Settings

□ Use proxy

Host: *

Port: *

(* Mandatory Fields)

Figure 2.25: Changing project settings.