inquiro

Easily collect and exploit your scientific data

Share and collaborate across your organization.

Inquiro version 3.4 – Mars 2020

Toulouse – Mars 2020 : DEXSTR, today announced the availability of Inquiro Version 3.4. This is a new version of our software. It comes with the following major enhancements that makes Inquiro V3 the Insight Engines for Life science.

Inquiro V3.4 New features & enhancements :

Jconnector robustness

The Jconnector has been enhanced in order to improve its robustness:

If, for any reasons, at any moment, the communication between the Jconnector and the Inquiro API, or if MongoDB or SolR is broken, the Jconnector will restart at least once.

In the same way, the jconnectors connected to external documentary systems will restart at least once if the connection is lost.

When the communication between the Jconnector and the other systems is restored, the Jconnector will restart automatically (SysD is required)

In case of communication issue between the Jconnector and RabbitMQ (annotation Engine), the admin user is able to relaunch the annotation on missed files using a dedicated button on the admin interface.

New automatic annotation improvements

Annotation engine has evolved to provided more flexibility and enhanced annotation capabilities:

Adding new dictionaries to the annotation is now feasible without relaunching the annotation engine for all metadata : The engine can be relaunched to consider only some dictionaries, either for all documents, or for a subset of documents .

Generate penalties for the annotation: You can exclude non-significant terms from dictionaries with a script that checks if the terms from target dictionaries are present in a “neutral” corpus. This corpus is constituted of the debates in the Canadian parliament (https://www.isi.edu/natural-language/download/hansard/).

Words context for annotation: boost scores of dictionary terms if some keywords are near detected terms in the text. This configuration is done by dictionary
(example: boost detection of “protein dictionary” terms if the word “protein” is close to the detected term)

TF-IDF: A new coefficient in the annotation formula has been added. TF-IDF (term frequency–inverse document frequency) is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus (https://en.wikipedia.org/wiki/Tf%E2%80%93idf)

API for all Jconnector

Refactoring to uniformize the Jconnector behavior after changes on the remote Jconnector functionality.

Annotation engine configuration

Reminder of automatic annotation functional principe

The dictionaries used for automatic annotation are defined in the templates section, at the template level, in the dynamical field section

For example, on “file” template:

```
        {
                "name": "human_disease_ontology",      -> Metadata field used for
annotation
                "translator": "[string]",              -> Data type (several text)
                "dictionary": [
                "human_disease_ontology"               -> dictionary used for annotation
        ],
        "prediction": true                             -> annotation is used
}
```

This configuration example means that the metadata field “human_disease_ontology” will be automatically filled with several terms of the dictionary “human_disease_ontology” by the annotation engine.

Each time a file is added to Inquiro, the annotation engine will calculate a score for each term of the dictionary “human_disease_ontology”. The calcul of the score depends of the annotation configuration. If the score is above a threshold, the term is added in the corresponding metadata field.

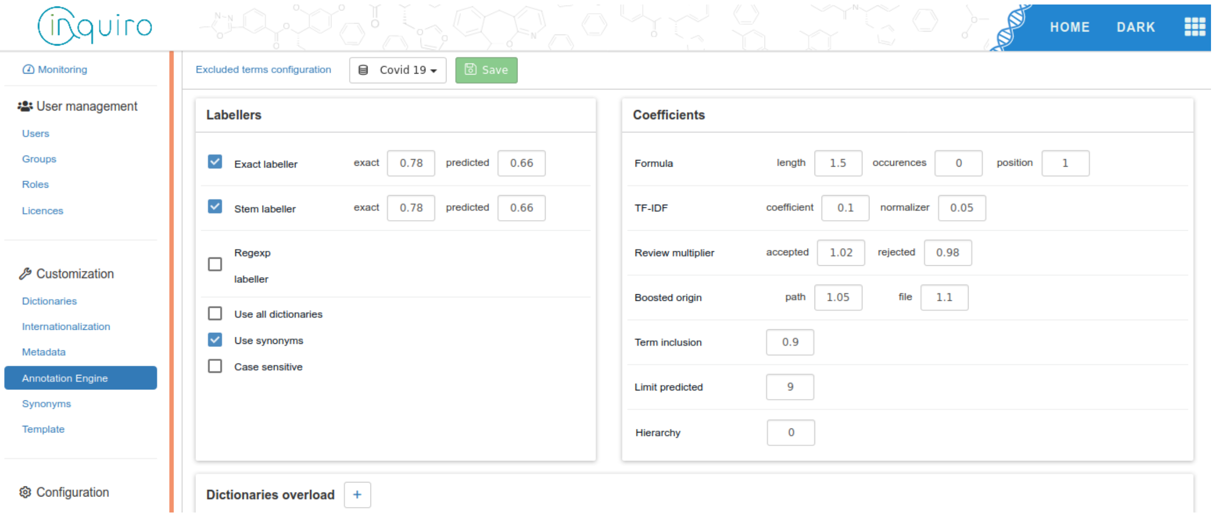
Annotation engine configuration

Since Inquiro v3.0, the annotation engine has evolved to be more relevant. As an example, we introduced some additional parameters to the score calculation in v3.0, such as:

- Length of the term
- Nb of occurrence of the term
- Position of the term in the document

The annotation engine continues to evolve and to ease the way it can be configured, we introduce a configuration interface in version 3.4.

This interface is accessible from the admin menu to users who have admin privileges on the annotation engine configuration for each instance.



Labellers

☑ Exact labeller	exact	0.78	predicted	0.66
☑ Stem labeller	exact	0.78	predicted	0.66
☐ Regexp labeller				

☐ Use all dictionaries
☑ Use synonyms
☐ Case sensitive

Coefficients

Formula	length	1.5	occurences	0	position	1
TF-IDF	coefficient	0.1	normalizer	0.05		
Review multiplier	accepted	1.02	rejected	0.98		
Boosted origin	path	1.05	file	1.1		
Term inclusion	0.9					
Limit predicted	9					
Hierarchy	0					

Dictionaries overload +

This section describes how to use this new interface

General section

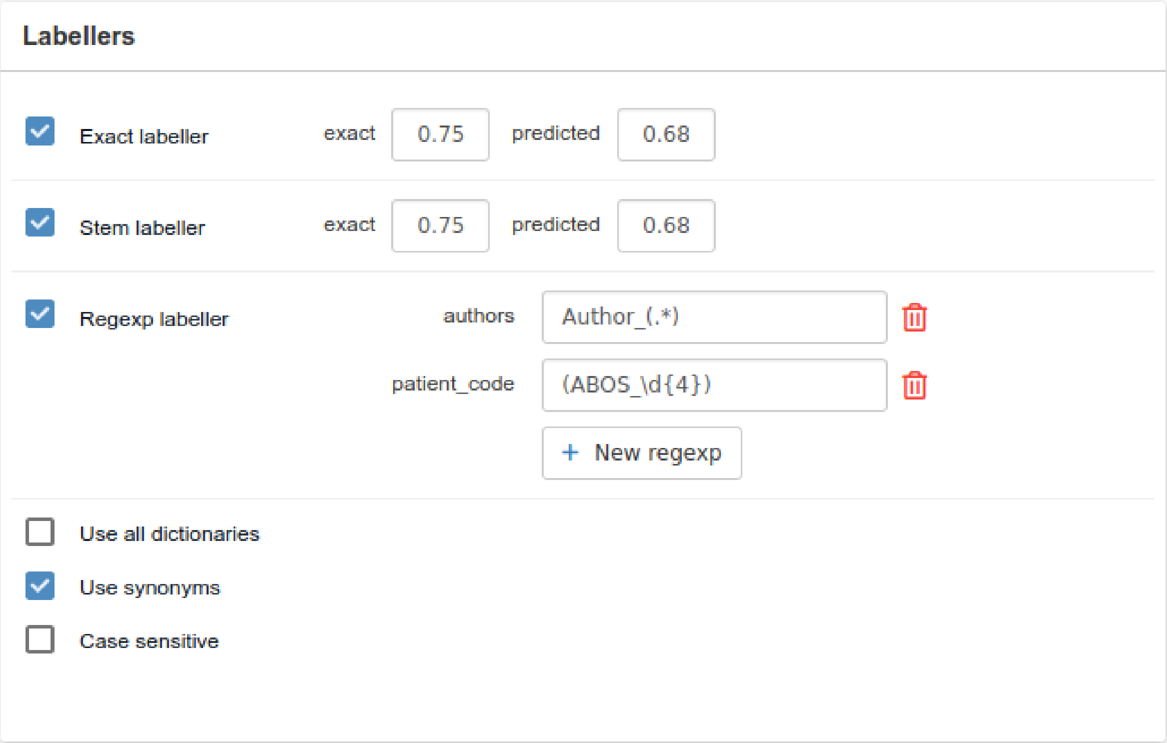
The annotation engine configuration is applied to the selected instance on top of the panel

The sections “labeller section” and “coefficient section” are the base configuration for all the dictionaries used in selected instances.

Note: the dictionaries used for the annotation engine in an instance are configured in “Template section”

Labellers section

Labellers are used to calculate a score for each dictionary term used for the annotation. The score ranges from 0 (not found) to 1 (found with the maximum certainty).



Labellers

☑ Exact labeller — exact 0.75 — predicted 0.68

☑ Stem labeller — exact 0.75 — predicted 0.68

☑ Regexp labeller — authors Author_(.*) — patient_code (ABOS_\d{4})

☐ Use all dictionaries

☑ Use synonyms

☐ Case sensitive

- **Exact labeller**: if checked, it indicates that the exact labeller is used to predict. If used, the admin must assign a threshold for automatic annotation and term suggestion ranging from 0 to 0.99.
  - Exact : a floating number representing the threshold for the term to be displayed as “Annotated metadata”
  - Predicted**:** a floating number representing the threshold for the term to be displayed as “Suggested metadata”
- **Stem labeller** : if checked, it indicates that stem labeller must be used to predict. This labeller uses the stemming (root of a word) for the annotation engine. If used, the admin can assign a threshold for automatic annotation and term suggestion ranging from 0 to 0.99
  - Exact : a floating number representing the threshold for the term to be displayed as “Annotated metadata”
  - Predicted**:** a floating number representing the threshold for the term to be displayed as “Suggested metadata”
- **Regexp labeller**: permits to define a regular expression for a particular field. Field to annotate value must be multivalued.
  If term corresponding the regexp are detected in the document, they are added to the metadata.
  Be careful with greedy matches (".*"), it will capture all the characters until an end of line.
  For each regexp, 2 parameters must be set:
  - The key: the field of interest
  - The values: the pattern to be found

Example 1 :

Field = project

Value = (Project\d*)

⇨ will match Project 11111

Example 2 :

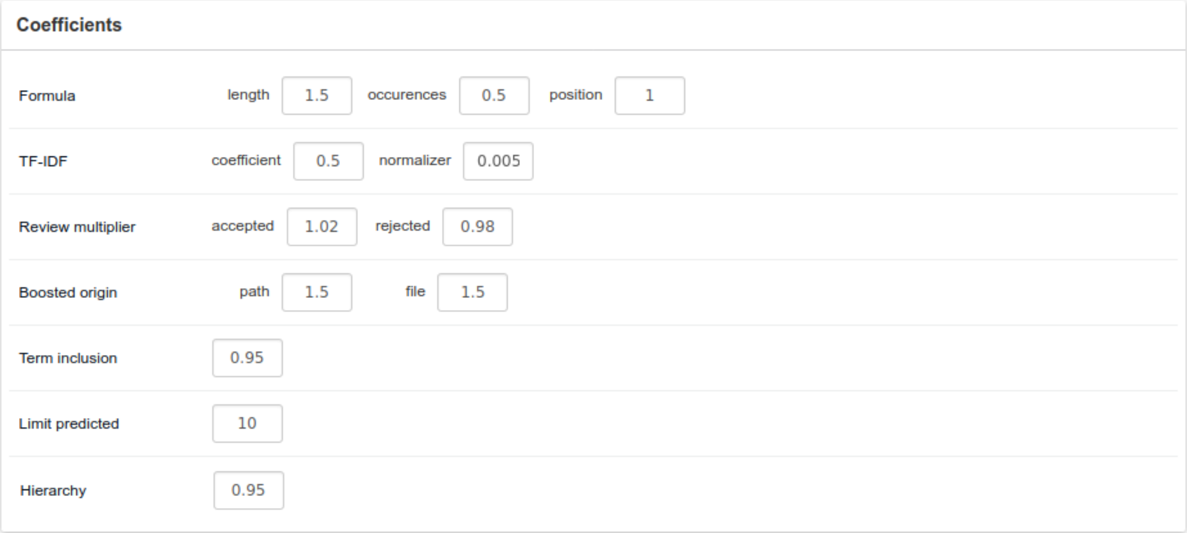
Field = Author
Value = Author: *(.*)

⇨ will match Author: John Lennon, Paul McCartney

- **Use all dictionaries**: If checked, use all the dictionaries that are not already used for automatic annotation. The predicted terms of these dictionaries are entered in the field “custom tags”
- **Use synonyms**: if checked, synonyms are used in exact and stemmed labeller score calculation
- **Case sensitive**: If checked, the term detection is case sensitive.

Coefficient section

Refers to parameters that are used to calculate the score assigned to each annotation term



Coefficients

Formula	length	1.5	occurences	0.5	position	1
TF-IDF	coefficient	0.5	normalizer	0.005		
Review multiplier	accepted	1.02	rejected	0.98		
Boosted origin	path	1.5	file	1.5		
Term inclusion	0.95					
Limit predicted	10					
Hierarchy	0.95					

- **Formula**
  - Length: impact of the length of the term
    The lower the value is, the lower the impact of the length parameter on the score calculated.
  - Occurrence : impact of the number of times the term is found in the document
    The lower the value is, the lower the impact of the number of occurrence on the score calculated.
  - Position : impact of the position of the word in the document
    The lower the value is, the lower the impact of the position of the term.
- **TF-IDF** : relevance of term by considering the frequency of the term in the whole corpus (https://fr.wikipedia.org/wiki/TF-IDF).

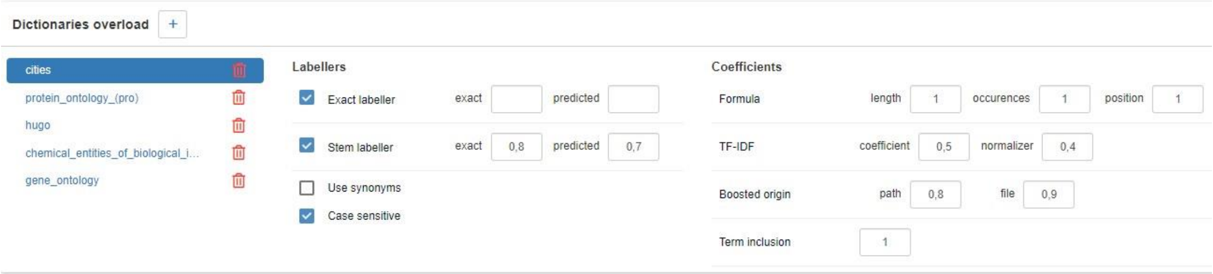
- ○ Coefficient: Impact of the TF-IDF The lower the value is, the lower the impact of the TF-IDF on the score calculated.
- ○ Normalizer: to normalize the TF-IDF score comparing to the other parameters (length, occurence or position), this value corresponds to the threshold limit of a “good” TF-IDF score. Usually, the recommended value for this parameter is 0.03

The values of those 4 parameters (length, occurrences, position and TF-IDF coefficient) are relevant only relative to each other.
I.e., if they are all set to the same value (1 by default), they are equally important for score calculation.
If one parameter is set to 0, it will not be considered for score calculation.
For example, this setting: length: 2, occurrences: 1, position: 1, TF-IDF: 0
means that the length of the detected term is 2 times more important than the number of occurrences or the position of the term in the document, and TF-IDF is not considered

- **Review Multiplier**: multiplier coefficients to boost or penalize terms score after manual curation: if a dictionary term is accepted or rejected by Inquiro users, a coefficient is applied to this term to improve its future automatic detection. (1 is neutral, >1 to boost, <1 to penalize)
  Default Review Multiplier is 1.02 for accepted terms and 0.98 for rejected terms
- **Boosted origin**: multiplier coefficient to boost the score of a term detected in the the file name or in the path (if the coefficient is > 1, it will boost the detected term)
- **Term inclusion**: A multiplicative coefficient to decrease the score of predicted terms that are included in another predicted term. this is to favorize detection of more precise terms. (1 is neutral, <1 to penalize)
  For example: "cancer" score will be decreased if "breast cancer" is also detected
- **Limit predicted**: Maximum number of predicted/suggested metadata for each field. -1 for no limit. Default to -1

Dictionaries overload

This section permits to overload the annotation configuration for any dictionary.



Labellers

☑ Exact labeller exact predicted

☑ Stem labeller exact 0,8 predicted 0,7

□ Use synonyms

☑ Case sensitive

Coefficients

Formula length 1 occurences 1 position 1

TF-IDF coefficient 0,5 normalizer 0,4

Boosted origin path 0,8 file 0,9

Term inclusion 1

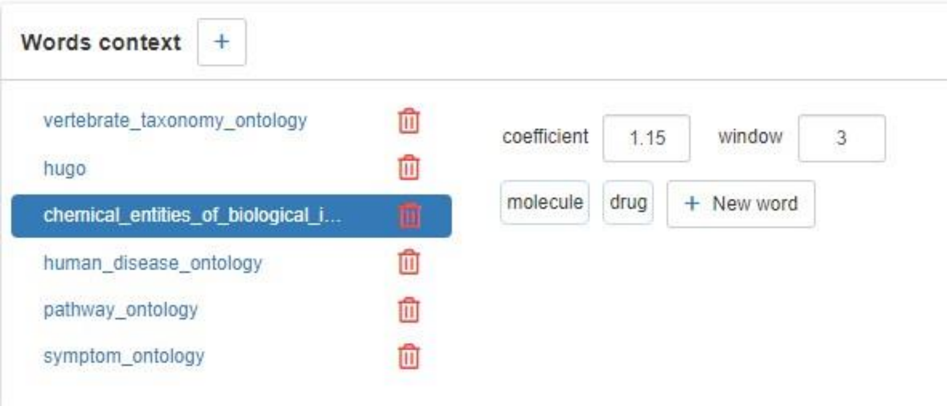
It is limited to the following parameters

- ○ Exact labeler
- ○ Stem labeler
- ○ Use synonyms
- ○ Case sensitive
- ○ Formula
- ○ TF-IDF

- Boosted origin
- Term inclusion

Words context

This feature boosts the scores of dictionary terms if some keywords are near detected terms in the text. Configuration is done by dictionary



Words context +

vertebrate_taxonomy_ontology	🗑
hugo	🗑
chemical_entities_of_biological_i...	🗑
human_disease_ontology	🗑
pathway_ontology	🗑
symptom_ontology	🗑

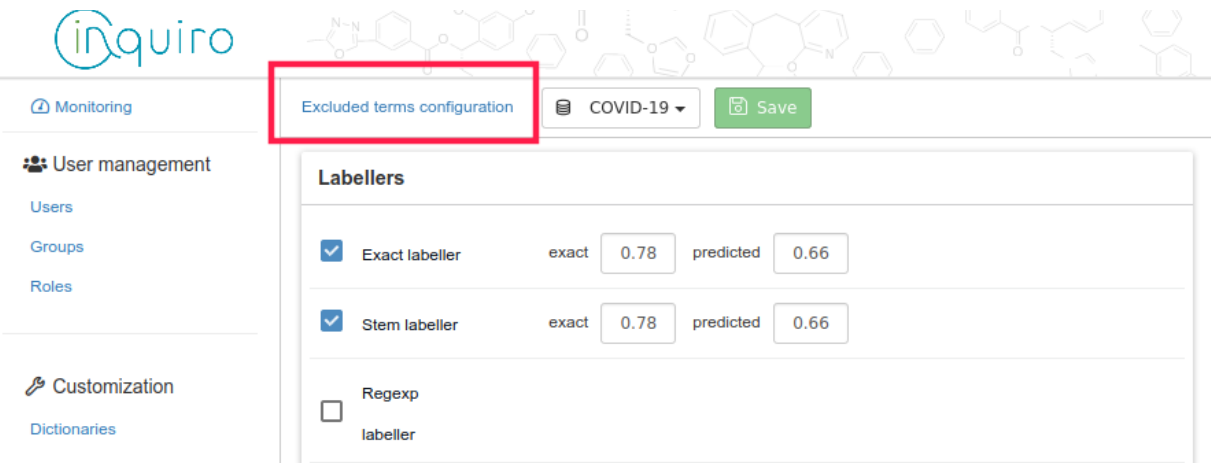
coefficient 1.15 window 3

molecule drug + New word

- **Words:** a list of keywords. If a keywords is close to a dictionary detected term in the document, the score of the term is boosted The keywords must not be composed (single word terms)
- **coefficient**: the multiplicative coefficient to boost the score of the value (1 is neutral, >1 to boost)
- **window**: detection distance in number of words between the dictionary term and the keyword

For the example : term of the dictionary “chemical_entities_of_biological_interest” that are presents in the document 3 words before or after “molecule” or “drug” have their score multiplied by 1.15

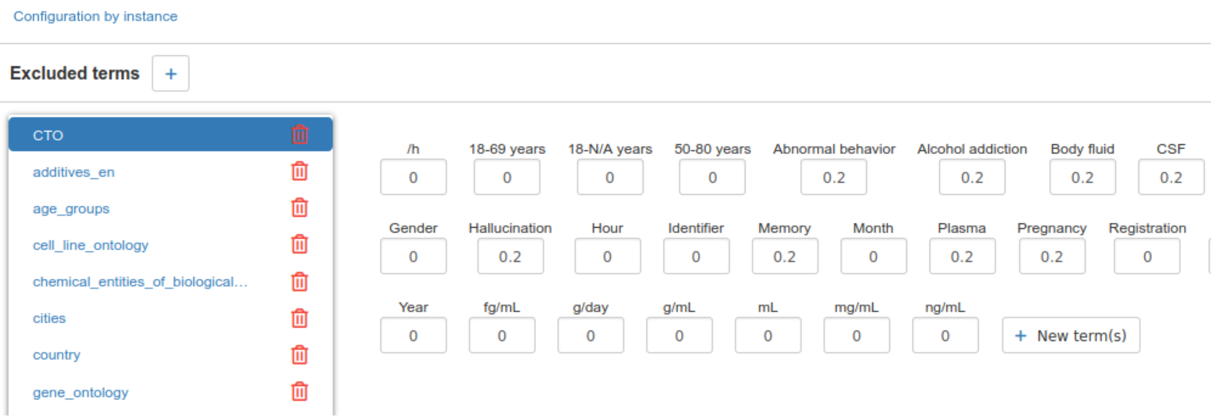
Excluded terms configuration



Labellers

☑	Exact labeller	exact	0.78	predicted	0.66
☑	Stem labeller	exact	0.78	predicted	0.66
□	Regexp labeller				

This section allows to boost or penalize prediction of specific terms from dictionaries for all instances.



Configuration by instance

Excluded terms +

- CTO
- additives_en
- age_groups
- cell_line_ontology
- chemical_entities_of_biological…
- cities
- country
- gene_ontology

/h	18-69 years	18-N/A years	50-80 years	Abnormal behavior	Alcohol addiction	Body fluid	CSF
0	0	0	0	0.2	0.2	0.2	0.2

Gender	Hallucination	Hour	Identifier	Memory	Month	Plasma	Pregnancy	Registration
0	0.2	0	0	0.2	0	0.2	0.2	0

Year	fg/mL	g/day	g/mL	mL	mg/mL	ng/mL
0	0	0	0	0	0	0

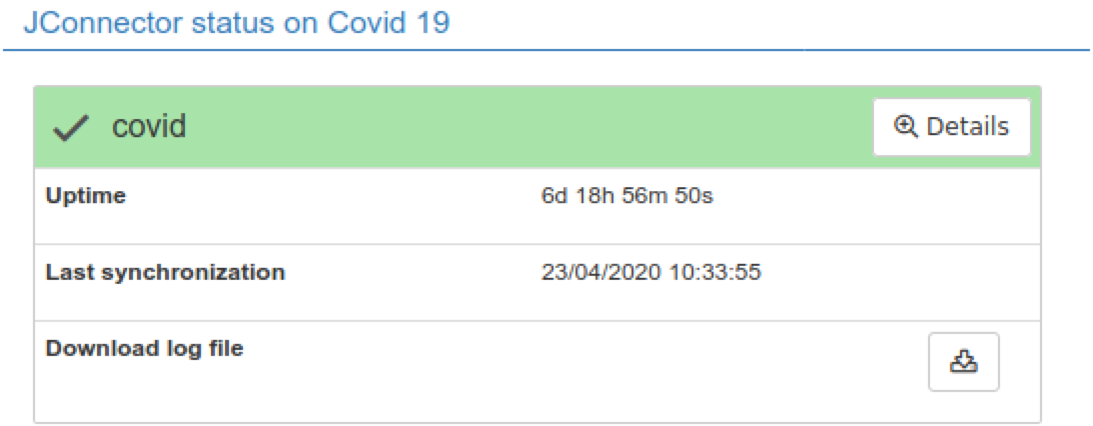
For each dictionary: terms can be entered and a multiplicator coefficient is always applied to the term score when it is detected by the prediction engine.

if the coefficient is set to 0, the term will never be predicted.

Remote JConnector logs/monitoring

The Jconnector could be installed on a different server than Inquiro. This can be useful for indexing data that is not available on the same network than Inquiro.

This new architecture required a change in the way the status of the Jconnectors are sent to the API.
To ease the administrator work, a button to download the logs of the Jconnector has been added. This button appears after that a Jconnector error occurs, and is removed after that the Jconnector resumes its normal operation, and that the logs have been downloaded.



JConnector status on Covid 19

✓ covid

Uptime	6d 18h 56m 50s
Last synchronization	23/04/2020 10:33:55
Download log file	

Various improvements

"All metadata" is now selectable on heatmap after search

Add jconnectors default configuration on docker
CMIS systems : the administrator can choose to open files either in the CMIS system (ex: Alfresco) or directly in Inquiro
Possibility to edit restrictions on the root of connector container
Change the automatic annotation logging in order to follow a container id through the annotation process
Possibility to paste a list to fill a dictionary field in the advanced search's querybuilder
Improved the automatic annotation performance for files with a lot of numeric values
Search with quotes for exact search
The minimal template should have the field "name" displayed by default

Bug corrections

Template cache not emptied for remote jconnector
upload center closes when uploading new versions/increment filenames
Missing i18n keys
Uninformative error toaster when there is no annotation engine
Missing a link to go back to the explorer from the trash bin
Querybuilder reset button wrongly positioned
No cross button to close review metadata modals
Can't modify text colors in HTML editor
Missing migration script for existing cmis containers before 3.3.9
When searching the exact term in a dictionary, it disappears
Relaunch prediction is launching prediction twice
Remove the link "Click here to access the advanced help." at the bottom of search help pop-up
Instance Selector on explorer page wrongly placed when resolution is low
Cannot connect multiple folders at same time
Error with local cmis connector when using API

Error with local openlabeln connector when using API
Eworkbook connector - icon is not good
Dashlet “cumulated Upload” too long
Dashboard : My last upload dashlet is slow
No API log for a LDAP locked user
Memory issue with Libreoffice
Relaunch annotation do nothing if the previous annotation ends with an error
Limit to 300 facets is causing trouble
Annotation cleaning is too long (exceed time limit for a labeller)
Import LDAP fields causing an error
Too many calls to restrictions functions on container page
MongoDB is overwhelmed by long queries sent by the API on large instance
Remains if remote connector container is sent to trash before being disconnected
cy-walk not included in release
Can't import synonyms of multiple dictionary at the same time
Annotation boost path does not look at tokens
Fuse dynamic views broken
hiddenAtView template option not working
i18n comes and goes on explorer
Reannotate on folder fail and kill tomcat if use on large folder
RegexpLabeller error when no text
Bad display of metadata
Add pwalk plugin exclusion in default jconnector configuration
When a template has errors, the warning icon is missing
Missing index for tracking
Duplicated annotated term when rename folder
Deleting molecule_name in templates does not remove it from items

Annotation stayed in progress forever