LexPredict ContraxSuite Documentation

Release Notes and Changelog
Release 1.1.2 - August 1, 2018

Summary

Version String	1.1.2
Major Version	1
Minor Version	1
Increment Number	2
Release Date	August 1, 2018
Release Branch	1.1.2

Release Notes

ContraxSuite Release 1.1.2 is the thirteenth open source release and is generally available on August 1, 2018.

Release 1.1.2 focused on:

- Several updates in API methods.
- Database queries optimization, added simple way to cleanup database.
- Improvements in document field detection..
- Implemented docker deployment with auto-scaling of Celery worker groups.

Detailed Changelog

New in Release 1.1.2

- Lexnlp: reduced run time of get_sentences method.
- API: added ability to export any objects list, f.e. document list in csv or xlsx file.
- Implemented role-based complex sequrity requirements.
- Implemented transferring training data for document field values.
- Optimized most of API sql queries.
- Implemented tables extraction from document.
- Added Review Status Group model to group document/project statuses.
- Allow to authenticate via query parameter in GET request.
- Overall Celery stability has been improved: document loading, field value detection and other long running asynchronous tasks.
- Improvements in document field detection and models:
  - description field added to fild model (Allows storing more human-understandable info on what the field means);
  - support for "object of definition" field logic (If definition terms are entered into a field detector then it will first check if one of them is defined in a sentence and only after this it will apply the regexps).
- Added common Nginx HTTP basic authentication behind Kibana, Flower, Jupyter. Django has its own authentication.
- Implemented support for separate Docker cluster architectures for bigger deployments with autoscaling Celery worker groups and smaller ones with single server used for all components.
- Implemented total_cleanup.sh script which allows fast deleting of all documents, field values, tasks and other data entered in the system but keeps important configuration such as field definitions, field detectors, users/roles.
- Implemented Dirty Field Retraining Process.

New in Release 1.1.1

- Lexnlp:
  - multiple fixes/improvements in get_companies method in nltk_maxent module; in get_titles; in get_currencies
  - standardized currency_type letter case in get_money method to prevent row duplication
  - updated stopword/collocation pickle files, added stopword/collocation scripts
  - improved segmenting sentences to not include page numbers in result.
- Allow to start multiple Locate tasks.
- Used separate Role model fro user roles, define user permissions based on role flags, added management command to install initial roles from fixtures.
- Added “Geography” document field type.
- Added user data in ajax login response.
- Do not store document full text in history to free db.
- Added ReviewStatus.is_active flag to detect active documents, added ReviewStatus model in admin site, added management command to install initial statuses from fixtures.
- Implemented detecting whether document is contract or not, store that value in Document.metadata.
- Improved base methods for sorting/filtering/paginating querysets.
- Implemented own django-celery database backend to easily store celery tasks, track their progress and log info.
- Extract internal nginx to separate docker container and make it routing to all components.
- Implemented text log rotating for all docker services as a separate docker container.
- Implemented auto-scaling for docker containers.
- Workaround for processing tasks left after killing Celery worker.
- Switched to higher logger level (DEBUG→INFO) to optimize logs size.
- Added auto retry for failed celery tasks.

New in Release 1.1.0

- Improved document type detection and text extraction.
  - Types of loaded documents are detected by their contents.
  - Apache Tika is now used for text extraction by default.
- Custom Apache Tika Docker image has been created and published: lexpredict/tika-server.
  The image contains the latest Tika 1.18 and latest Tesseract OCR engine version 4. It allows external Tika configuration and ready for using in Docker Swarm clusters.
- Contraxsuite logging has been switched to FileBeat.
  - Django, Celery and DB logs are first written to files in JSON format. Separate FileBeat Docker container reads them in asynchronous mode and pushes log records to Elasticsearch. Logging system is now unwired from Python modules and will not hang or slow down the application in case of Elasticsearch problems.
  - Internal Nginx logs are now sent to Elasticsearch. Standard FileBeat Kibana dashboards now work and display Nginx access/error data.
  - Logs are written to Elasticsearch indexes containing dates in their names. Old log indexes are deleted by Curator.
- Logging routines of Contraxsuite asynchronous Celery tasks has been refactored:
  - Task logs no longer stored in DB. Elasticsearch is now the primary source of log data. Task logs in UI at Tasks / Admin Tasks are now loaded from Elasticsearch.
  - Task logs provide more detail. Task logs in Kibana can be searched by user, by document name/id.
- MetricBeat now tracks metrics of Contraxsuite clusters.
  MetricBeat has been added to track metrics of Docker containers in Contraxsuite cluster. Standard MetricBeat dashboards now available in Kibana allowing easy tracking of CPU, memory usage, availability and status of different Contraxsuite components.
  Metrics are written to Elasticsearch indexes containing dates in their names. Old log indexes are deleted by Curator.
- Improved project cleanup method to delete all related objects, added "total cleanup" method and UI. Fixed “purge_task” celery task to handle GroupResults.
- Added “Project Creator” non-admin role for access to all but admin interface and admin tasks.
- Included “set_site” management command into deployment script.
- Fixed broken reset password API.
- Redirect after change password to user detail page.
- Improved celery task progress calculation.
- Added celery subtasks logging.
- Better handling exceptions while clustering project documents.
- Added user name into response cookies and json response of login rest API.
- Better handling memory error in training document field model.
- Updated task list view to sort/filter by calculated fields.

- Several bug fixes related to annotating API.

New in Release 1.0.9

- New address locator based on machine learning.
- Several bug fixes related to annotating API.
- Multiple improvements in Docker image building and installation scripts:
  - TIKA is working as a separate Docker container;
  - built resource limits specification in docker compose scripts;
- Added ReviewStatus model, added status field to Document and Project models related to ReviewStatus.
- Added “assignee” field to Document model.
- Updated historical model records to delete if source object deleted.
- File type detection by content on document loading.
- Bug fixing and stability improvements of celery tasks related to document loading and processing.
- Improved calculated fields in Task model.
- Fixed issue with uploading files with special characters in their names.
- Resolved duplicate key issue in DateUsage extraction.
- Fixed missed template for Top Date Duration list page.
- Enabled server-side pagination for all list views that have large amounts of data
- Created plain html template to speed up rendering for Clustering task form.
- Added Text Unit list by language page.
- Added “Language” field into Document model.
- Fixed language detection while extracting text units from documents.
- Fixed broken filters on Document Cluster list and Text Unit list pages.
- Display document type title instead of uid on all pages which include document data.
- Improved “heavy” SQL queries on Top Geo Entities list, Top Date Duration list, Document Detail and Party Summary pages.
- Added more debug info into task logging.
- Improved currency locator.

New in Release 1.0.8

- Embedded Jupyter Notebooks inside Contraxsuite Docker Image. Contraxsutie Docker image now contains embedded Jupyter having access to all the project code base, Postgres DB and ElasticSearch. If any use case is not covered by the Contraxsuite UI the customer's specialists can use Jupyter for implementing any algorithms and/or reports they require using all the advantages of the Contraxsute code and data extraction libraries.
- Embedded Kibana. Internally Contraxsuite uses ElasticSearch for indexing documents contents for searching purposes. Also now Contraxsuite backend forwards its logs into the embedded ElasticSearch for easier debugging. Now Contraxsuite by default embeds Kibana into its Docker

cluster. Kibana is configured to have access to the embedded ElasticSearch and can be used for accessing the indexed documents and logs, building complex search queries and dashboards.

- Improvements for document field value extraction:
  - Models and machine learning workflow for detecting multiple field values in a single sentence.
  - User-selected value ranking for single-value choice fields.
  When system detects multiple possible variants of value for a single-value choice field, the system selects single value according to the order configured by user for this field.
  - Support for calculated fields.
  Values of some fields can not be directly extracted from the text in many cases. Instead the related data on which this field's value depends can be detected in text and the original field can be easily calculated based on this intermediate data. For such cases we introduce ability to configure a field, specify other fields on which this fields depends and define a formula of calculation.
  Example: contract start date, contract end date and term. Sometimes only start date and term is specified in the text. For this case we can calculate end date based on them. In other cases we can calculate term based on start and end dates.
- Significantly improved extraction precision for dates and persons.
- Added backend (database) storage for project-wide variables.
- Added ipython notebook describing clustering process – see notebook-examples/clustering.ipynb

New in Release 1.0.7

- Project has been switched to Continuous Integration of Deployment process.
- Jenkins build server has been set up. It is continuously building Contraxsuite project and deploying it to LexPredict internal servers as well as to the public demo site - https://demo.contraxsuite.com.
- Contraxsuite now provides Docker image and set of scripts for fast and easy project deployment on clean Ubuntu machine.
- Docker images are build with Jenkins server on every change in Contraxsuite git repository. They are deployed to LexPredict DockerHub repository (https://hub.docker.com/r/lexpredict/lexpredict-contraxsuite/). Latest installation scripts are maintained automatically at https://demo.contraxsuite.com/files/contraxsuite-deploy.tar.gz
- Added REST API for advanced users. List or REST API urls (version #1) available for superusers under /api url.
- Internal support and APIs for user-defined document fields has been implemented.
- User defined customizable document fields will be one of the core concepts of the future Contraxsuite. Internal models and REST APIs for user defined fields has been implemented.

- Value extraction rules for the fields are totally customizable for client admins via regular expressions and set of configuration preferences - for extracting the initial values.
- After the initial values are set the system will use machine learning algorithms to train itself based on the users' modifications to the field values and allow more and more accurate extraction of field values on other documents. Currently these workflows are available via REST API only. User interface is in progress of being created.

New in Release 1.0.6

- Created "Employment" custom application - application which deals with Employment Agreements.
- Created "Lease" custom application - application for Lease Agreements.
- Added "Development Guide.md" - a guide for developers, small “how to”.
- Improved LexNLP - added wrapper to get sentence ranges in addition to sentence texts.

New in Release 1.0.5

- Added links to result list from admin task list page
- Added "description" columns in admin task list table for task details.
- Added Geo Entity list page.
- Make priority column editable in Geo Entity list view to allow a user to reorder priority for geo entities.
- Fixed "purge task" issue for admin tasks.
- Gzip html, enable django-pipeline to decrease traffic / loading time for pages.
- Added ability to Cluster by currency value and currency name.
- Added ability to Cluster by date duration.
- Used amount of days as weighted value for clustering by dates.
- Allow clustering by courts.
- Allow clustering by document metadata.
- Added ability to plugin custom applications (see "Development Guide.md")
- Made "Autologin" configurable via web app (see "Application Settings" page)

New in Release 1.0.4

- Simplified web application requirements for deployment and licensing.
- Improved UI for navigation and analysis.
- Improved locator workflow in admin tasks with “locate all” flow.
- Increased flexibility for clustering and classification dimensions with dates.
- Implemented non-administrative application configuration menu.
- Implemented default locator configuration through application configuration menu.
- Refactored distributed task engine for pluggable application architecture.
- Refactored presentation layer for pluggable application architecture.
- Added favicon configuration for web application and admin screens.
- Improved data model and database details on statistics page.
- Integrated LexNLP URL locator into web application.
- Integrated LexNLP copyright locator into web application.
- Integrated LexNLP trademark locator into web application.
- Integrated LexNLP title locator into web application with document metadata.
- Implemented LexNLP title locator.
- Implemented additional LexNLP transforms for skipgrams and n-grams.
- Improved LexNLP handling for parties with abbreviations and other cases.

- Improved LexNLP handling for amounts with mixed alpha and numeric characters.
- Improved LexNLP unit test coverage.
- Improved knowledge sets for US regulators and real estate concepts.
- Preparation for open source example applications for employment and leasing use cases.
- Updated source code license headers.

New in Release 1.0.3

- Improved UI for navigation.
- Improved UI and ranking for search results.
- Increased flexibility for clustering and classification dimensions.
- Refactored unit test framework into CSV-based formats.
- Improve unit test framework handling for language and locales.
- Fixed issue with HTML file extension whitelists for web application.
- Fixed issue with snippet display characters in jQWidgets tables.
- Implemented method and input-level CPU and memory benchmarking for unit tests.
- Migrated all unit tests to 60 separate CSV files.
- Added over 1,000 new unit tests for most LexNLP methods.
- Reduced memory usage for paragraph and section segmenters.
- Improved handling of brackets and parentheses within noun phrases.
- Added URL locator to LexNLP.
- Added trademark locator to LexNLP.
- Added copyright locator to LexNLP.
- Standardization of lower/uppercase for party names in presentation layer.
- Enhanced translations of common scientific and chemical terms in knowledge sets.
- Improved default Punkt sentence boundary detection.
- Added custom sentence boundary training methods.
- Added common acronyms for Australian agencies to knowledge sets.
- Added list of common real estate terms to knowledge sets.
- Improved handling of US court names when informally referenced.
- Improved handling of multilingual text, especially around geopolitical entities.
- Improved default handling of party names with non-standard characters.
- Enhanced metadata related to party type in LexNLP.
- Improved continuous integration for public repositories.

New in Release 1.0.2

- Improved documentation for installation and configuration.
- Automated Canvas theme installation for single-line installer.
- Automated jq package installation for single-line installer.
- Added new visualization/report functionality.
- Added “export to calendar” functionality for dates.
- Refactored and integrate core extraction into separate LexNLP package.
- Released nearly 200 unit tests with over 500 real-world test cases in LexNLP.
- Improved definition, date, and financial amount locators for corner cases.
- Integrated PII locator for phone numbers, SSNs, and names from LexNLP.
- Integrated ratio locator from LexNLP.
- Integrated percent locator from LexNLP.
- Integrated regulatory locator from LexNLP.
- Integrated distance locator from LexNLP.
- Integrated case citation locator from LexNLP.

- Improved geopolitical locator to allow non-master-data entity location.
- Improved party locator to allow configuration and better handle corner cases.
- Refactored English term locator for improved scalability and database compatibility.
- Resolved URL issue in embedded document viewer.
- Releasing common legal term set for top 1000 terms based on 100K contract sample.
- Added geopolitical subdivisions for Spain, China, and England.
- Improved list of US Federal and State regulators.
- Improved list of US Federal and State courts.
- Improved error message for locators when Court master data is missing.
- Improved UI to prevent multiple submission of admin tasks.
- Releasing word embedding model for credit/loan agreements.
- Releasing word embedding model for real estate and leasing contracts.
- Releasing word embedding model for operating agreements.
- Releasing word embedding model for labor and employment agreements.
- Releasing word embedding model for service and consulting agreements.
- Releasing word embedding model for generic agreements.
- Releasing pre-trained document type classifier.

New in Release 1.0.1

- Added deployment automation for superuser credentials and creation.
- Improved documentation for passwordless SSH for deployment automation.
- Changed from git to HTTPS protocol for deployment automation.
- Added two-factor authentication (2FA) for TOTP and HOTP.
- Added “Search by Party” to default UI.
- Added “Currencies” tab to default UI.
- Added additional metrics to global statistics page.
- Improved audit on Document and Text Unit Property data models.
- Changed default UI for editing Properties on Document Detail page.
- Improved default UI for DocumentTag lists and detail.
- Decreased default similarity threshold to 75% for Document and Text Unit task.
- Added knowledge set loading from lexpredict-legal-dictionary repository.
- Improved Court data model for names and abbreviations.
- Fixed “empty” cluster issue for non-DBSCAN clustering tasks.
- Improved auto-complete result order by corpus frequency.
- Fixed error on global statistics page when no Projects or Queues exist.
- Fixed path issues for ES, git, and celery in deployment automation.
- Improved “Add to task queue” form for Cluster workflow.
- Fixed “Class Name” issue for new Classifier workflow.
- Allow clustering by term or model dimensions.
- Added new semi-supervised classification method (LabelSpreading).
- Added new Quick Start Installation Guide for Linux.
- Added new Administration Guide for Linux.
- Updated Installation and Configuration Guide.
- Updated Software and Data Dependencies.
- Updated Technical FAQ.
- Updated Security FAQ.
- Updated Data Model Diagrams.
- Updated Architecture Diagram.

- Separated Knowledge Set documentation from Dependency document.
- Updated Knowledge Set documentation.
- Refactored Knowledge Set structure and naming.
- Added UK GAAP Accounting terms.
- Added US FASB Accounting terms.
- Added US State regulators.
- Added limited English, French, and Spanish translations for German courts.
- Translated geopolitical entities for English, Spanish, German, and French.
- Added geopolitical relationships.
- Translated chemical elements and compounds for English, Spanish, German, and French.
- Added word2vec models for employment agreements.
- Added word2vec models for leases.
- Added US hazardous waste.
- Added 300 new software license samples.
- Added 100 new construction agreement samples.
- Added 500 new credit agreement samples.
- Added 200 new severance agreement samples.