Machine Learning and Data-Driven Tools for Automatic Evaluation of RADNEXT Experiments Proposals

Jaroslaw Szumega, Lamine Bougueroua, Blerina Gkotse, Pierre Jouvelot, Federico Ravotti

▶ **To cite this version:**

Jaroslaw Szumega, Lamine Bougueroua, Blerina Gkotse, Pierre Jouvelot, Federico Ravotti. Machine Learning and Data-Driven Tools for Automatic Evaluation of RADNEXT Experiments Proposals. The Workshop for Industry on Radiation Hardness Testing of Semiconductor Devices and Systems at the RADNEXT Facilities (G-RADNEXT Workshop 2023), Nov 2023, Meyrin, Switzerland. hal-04283238

HAL Id: hal-04283238

https://minesparis-psl.hal.science/hal-04283238v1

Submitted on 13 Nov 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Machine Learning and Data-Driven Tools for Automatic Evaluation of RADNEXT Experiments Proposals

Jaroslaw Szumega[1,2], *Lamine Bougueroua*[3], *Blerina Gkotse*[1,2], *Pierre Jouvelot*[2], *Federico Ravotti*[1]

[1] Experimental Physics Department, CERN, Geneva, Switzerland
[2] Mines Paris, PSL University, Paris, France
[3] Efrei, Université Paris-Panthéon-Assas, Paris, France

ABSTRACT

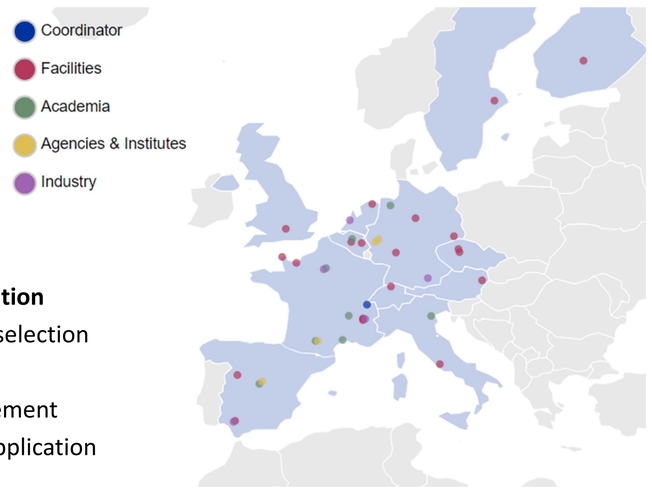
In the framework of **RADNEXT Work Package 3 (WP3)**, the Transnational Access (TA) portal was created for the management of the submission process of experimental proposals. One of the WP3 research activities, and the topic of an ongoing PhD project, is to support the assessment of TA requests. Using **Natural Language Processing** (NLP) techniques, we aim to provide automatic assistance for all the interested stakeholders. That includes support for both the users during the submission process and reviewers and User Selection Panel (USP) members during the project-selection period. This poster introduces the machine learning and **data-driven** tools used for these goals.
Some initial experiments, accomplished tasks and created software are presented in order to notify the RADNEXT network about the current status and advances of this research and provide a baseline for subsequent discussion related to such activities. We take advantage of the **Open Peer Review** (OPR) movement to gather High-Energy Physics (HEP)-related data to build and train custom **Machine Learning** (ML) **models** able to provide initial evaluation of experimental proposals.
The presented research is highly innovative, since NLP-based processing is mostly used in the field of computer science and human sciences - and not necessarily High-Energy Physics. Future plans for the use of ML methods are presented - both in the framework of RADNEXT-related activities and resulting PhD thesis research.

RADNEXT TRANSNATIONAL ACCESS

RADNEXT Transnational Access

- Funded by EU Horizon 2020
- **6000 beam hours** to be awarded
- 20 different facilities in Europe and beyond
- Eligibility to academia and industry

Date of next TA call: January 2024



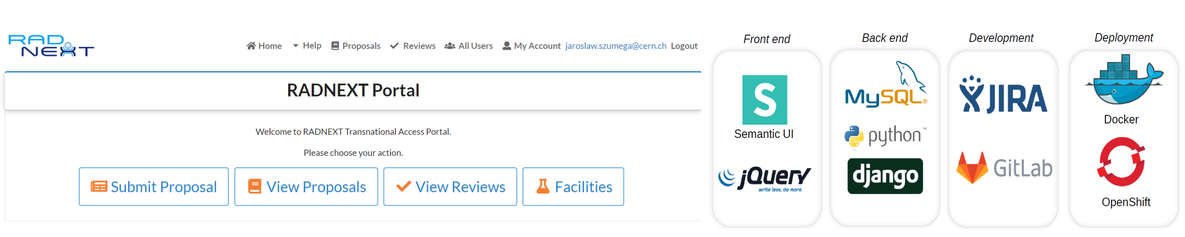
RADNEXT Work Package 3

Transnational access management and harmonization

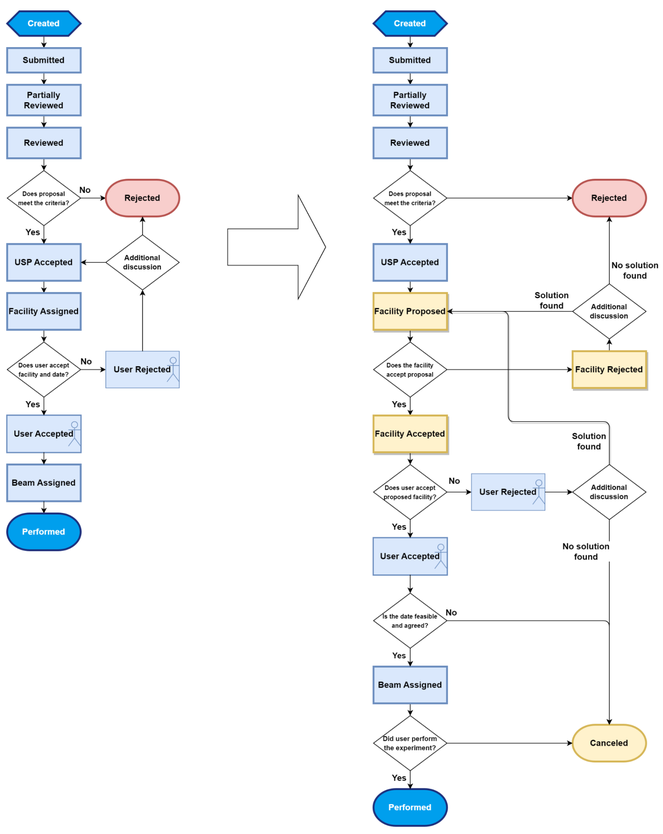
- Web portal for TA proposal management and selection
- Harmonisation of access procedures
- Defining a data model for information management
- Exploration of NLP techniques to support application procedures

RADNEXT TA PORTAL

RADNEXT Portal was created to support TA activities, including the application and selection processes. Currently, over **200 experiment proposals** were submitted via the RADNEXT TA portal and evaluated. **49 experiments** are already performed and **10** have their assigned and agreed-upon beam time.



The RADNEXT portal webpage provides functionalities for USP members, reviewers and prospective users. A modern software stack supports the operation at every stage – starting from development and deployment up to the website-rendering for users.



Feedback from users and experts responsible for proposal evaluation helped improve the portal.

The proposal workflow changed over time thanks to the feedback from WP leaders and USP members.

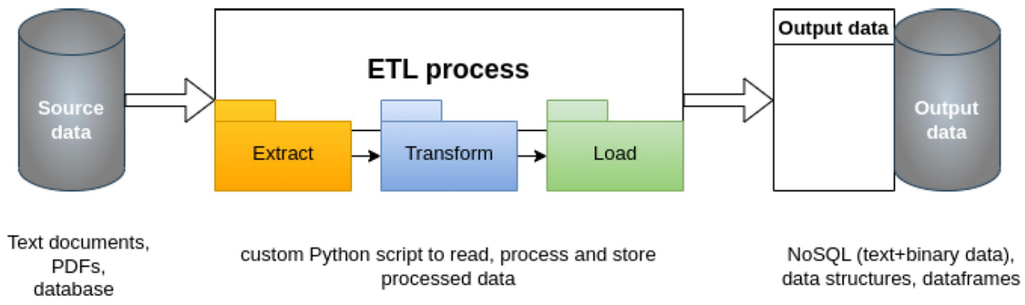
This improved version gave room for more discussion and clarification, and new statuses help to improve the internal management of proposals.

ETL PROCESS

Extract, Transform, Load (ETL): A standardized workflow of data pre-processing, starting from its raw form up to a structure suitable for actual processing. This complex process is broken up into three steps that:

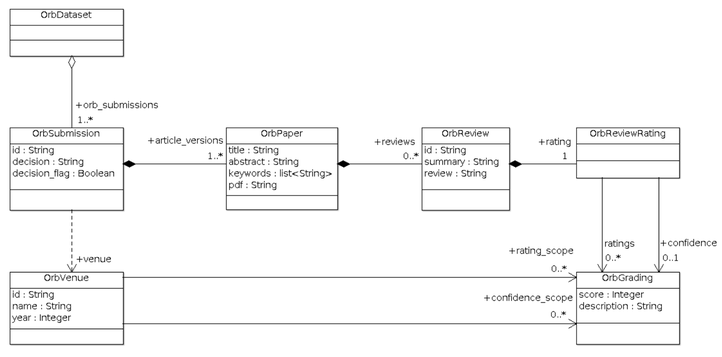
- **extract** the raw data from the initial storage;
- **transform** it into a useful format;
- **load** it into memory.

This modular architecture allows to process the multimodal data coming from heterogenous sources.



OPEN PEER REVIEW DATA

Open Peer Review (OPR) is a new peer-review model that provides transparency in the process of scientific publication assessment.
We aim to build upon this data to create a knowledge representation that will support RADNEXT TA in various ways, including support to users at the time when they are submitting their own experimental proposal up to the entire evaluation process.



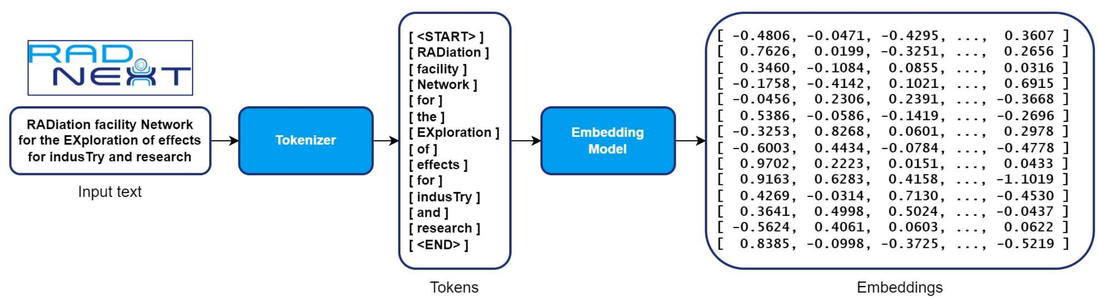
A new OPR dataset was created in order to facilitate future work in the domain of automatic assessment of scientific experiment proposals.

This highly structured data is meant to provide a reusable resource that will be accessible and usable for various tasks in the rapidly evolving field of NLP.

NATURAL LANGUAGE PROCESSING TECHNIQUES

Natural Language Processing (NLP) is a subfield of Machine Learning (ML) and linguistics. It provides the tools that enable the processing of natural language – human-readable texts.

In **WP 3,** we deal with a large number of documents – scientific texts of experiment proposals. To provide assistance to users and support to experts, semantic relationships need to be properly represented and processed. NLP techniques are the core of such processing.



The first stage of the ML pipeline uses NLP to transform texts into "embeddings", i.e., vectors of numbers. Once computed, they are processed further, as numerical representations are the only ones suitable for ML tasks.

SUMMARY AND FUTURE WORK

Summary

- RADNEXT WP 3 is dedicated to a wide scope of activities aimed at process management and harmonization of the procedures for facility access.
- RADNEXT TA Portal was created to support TA activities, improving the application process.
- Various software packages were used to create the tools that improve the procedures related to the application process.
- Building on the OPR process and novel NLP techniques, we plan to provide an objective and transparent way to support the application process and the proposals' assessment.

Future Work

- Continuous improvement of the RADNEXT TA procedures and web portal
- Design of a ML model that helps analyse scientific proposals and support their assessment
- Using NLP techniques, support to users during the application procedure, e.g., via automatically generated suggestions regarding their experiment description

www.radnext.web.cern.ch

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101008126.