PROCEEDINGS OF SCIENCE

A Portal Dedicated to Higgs Bosons for Experts and the General Public

André Sopczak[a,b,*]

[a] *Czech Technical University in Prague, Institute of Experimental and Applied Physics,*
Husova 240/5, CZ-110 00, Prague 1, Czechia

[b] *on behalf of the International Particle Physics Outreach Group*

E-mail: andre.sopczak@cern.ch

As an educational aid and source for expert information, a web portal dedicated to Higgs boson research is presented. A database is created with more than 1000 relevant articles using CERN Document Server API and web scraping methods. The database is automatically updated when new results on the Higgs boson become available. Using artificial intelligence and natural language processing, the articles are categorized according to properties of the Higgs boson and other criteria. The process of designing and implementing the Higgs Boson Portal (HBP) is described. The components of the HBP are deployed to CERN Web Services using the OpenShift cloud platform. The HBP is accessible within the Czech Particle Physics Project (CPPP) at http://cern.ch/cppp and directly at http://cern.ch/higgs.

The Tenth Annual Conference on Large Hadron Collider Physics - LHCP2022
16-20 May 2022
online

[*] Speaker

PoS(LHCP2022)353

1. Introduction

Over the past decades, the search for the Higgs boson of the Standard Model and Higgs bosons in extended models have been at the forefront of research in particle physics. Ten years ago the discovery of a Higgs boson with Standard Model properties was established. The Higgs boson research has led to more than 1000 experimental publications on the Higgs boson search and the measurement of its properties. Many aspects of the Higgs boson research have been addressed, including the search for Higgs bosons beyond the Standard Model. For the benefit of experts and the general public interested in this exciting field of research, a Higgs Boson Portal (HBP) is implemented as part of the Czech Particle Physics Project together with other modules [1].

The HBP uses Artificial Intelligence (AI) and Natural Language Processing (NLP) for categorization of the Higgs boson publications in a user-accessible list with the following categorizations:

- Publication stage (preprint, journal accepted, published)
- Year of public release
- Experiment
- Luminosity
- Higgs boson decay mode
- Higgs boson production mode
- Higgs boson model (Standard Model or Beyond Standard Model)

The detailed categorization of the production and decay modes follows the review [2].

The database is automatically updated when new results on the Higgs boson become available, and the components of the HBP are deployed to CERN Web Services. The HBP also includes a visualisation of some developments of limits and precision.

Following a feasibility study [3], implementation [4] and presentation [5], the HBP is accessible on http://cern.ch/higgs.

2. Sources of Publications

The publication information is extracted using web scraping of the Fermilab web sites [6, 7] and with the CERN Document Server API [8]. Publications of the following collaborations are included:

- 1989 – 2000: CERN – Large Electron-Positron Collider (LEP)
  - ALEPH, DELPHI, L3, OPAL
- 1987 – 2011 Fermilab – Tevatron Collider
  - CDF, D0
- 2010 – present: CERN – Large Hadron Collider (LHC)
  - ATLAS, CMS

3. Natural Language Processing

An expert in the field of Higgs boson research recognizes immediately the category of an article. However, this identification is not a trivial task for an automated computer categorization. Therefore, a complex solution is chosen applying Artificial Intelligence (AI) and Natural Language Processing (NLP). Details and further references are given in Ref. [3]. A probability for a publication is assigned for all the predefined classes. The class with the highest probability is chosen for the categorization.

A publication text is expressed as a vector of words. Before classification, it is beneficial to pre-process the text by removing stopwords, which are the most frequently used words in the given language. English examples are "the", "and" and "from". These words do not play any role in classifying an article. The remaining words undergo stemming, which involves reducing the words to their root form, e.g. "decaying" to "decay". Although the context of individual words is lost during the pre-processing, this method can still be used to classify the text as a whole. It is computationally ineffective to apply NLP to the full text of the article. Therefore, for the HBP, only the title and the abstract of an article are selected for processing.

The HBP utilises a naive Bayes classifier to distinguish between articles studying the Standard Model and articles searching beyond the Standard Model. These two directions of research generally use slightly different wording, which helps the classification process to be accurate. For example, the word "search" appears more often in publications beyond the Standard Model.

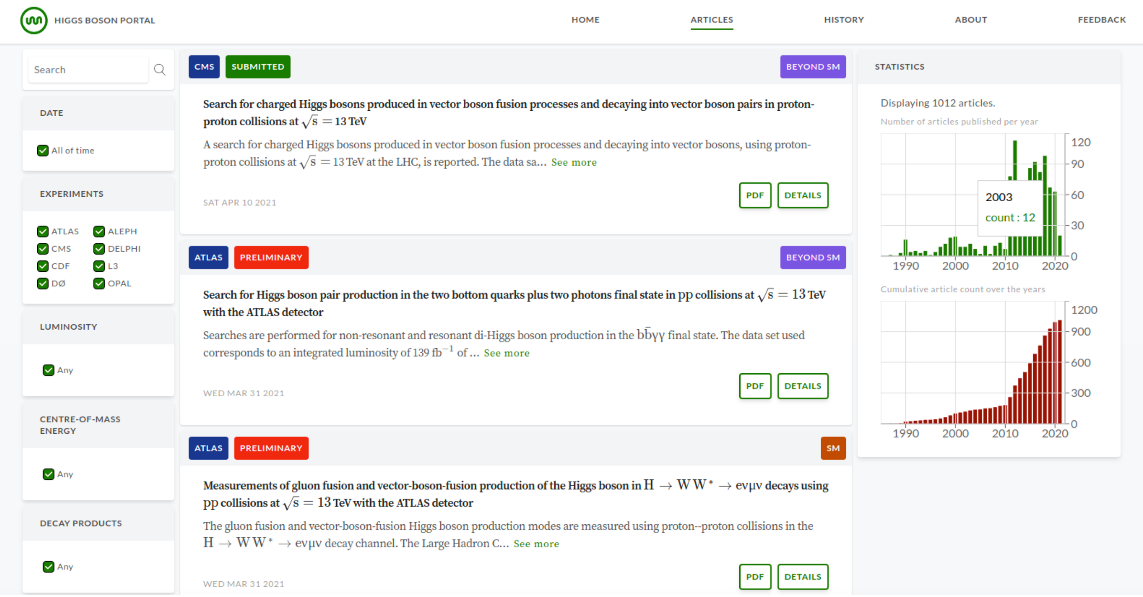
In order to extract more concrete information from the articles, e.g. integrated luminosity, centre-of-mass energy, decay products or production modes, a different method is applied. These entities are dependent on the context in which they are used. For example [9], in the title: "Measurements of gluon fusion and vector-boson-fusion production of the Higgs boson in $H \to WW^* \to e\nu\mu\nu$ decays using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", two production modes are mentioned – gluon fusion and vector boson fusion. The decay mode is expressed using the notation "$H \to WW^*$". The centre-of-mass energy ($\sqrt{s}$) is given as the number "13" combined with the "TeV" unit. The process of automatically recognising and extracting meaningful words, phrases or numeric values is called Named Entity Recognition (NER). NER has to recognize syntactic structures of the text, and therefore is language specific. Most scientific articles are written in English, therefore the HBP uses a NER model pre-trained on the English language. The training can be extended for specific needs of the domain. After extracting entities from the text, they have to be parsed or categorized. When numeric values are expected, the number and unit are parsed algorithmically using a set of predetermined rules. In case of the decay mode and the production mode, the category is decided by identifying keywords and special characters in the extracted named entity. The performance of the NER categorization is given in Table 1.

Table 1: NER categorization performance. True Positives (TP), False Positives (FP), False Negatives (FN).

	TP	FP	FN	Precision (%)	Recall (%)	F_1-score (%)
Luminosity	51	2	7	96.2	87.9	91.9
Centre-of-mass energy	51	0	9	100	85.0	91.9
Production mode	52	8	9	86.7	85.3	86.0
Decay mode	71	17	19	80.7	78.9	79.8

4. User interface

The user interface is shown in Fig. 1.



HIGGS BOSON PORTAL

HOME ARTICLES HISTORY ABOUT FEEDBACK

Search

DATE

☑ All of time

EXPERIMENTS

☑ ATLAS ☑ ALEPH
☑ CMS ☑ DELPHI
☑ CDF ☑ L3
☑ DØ ☑ OPAL

LUMINOSITY

☑ Any

CENTRE-OF-MASS ENERGY

☑ Any

DECAY PRODUCTS

☑ Any

CMS SUBMITTED BEYOND SM

Search for charged Higgs bosons produced in vector boson fusion processes and decaying into vector boson pairs in proton-proton collisions at $\sqrt{s} = 13$ TeV

A search for charged Higgs bosons produced in vector boson fusion processes and decaying into vector bosons, using proton-proton collisions at $\sqrt{s} = 13$ TeV at the LHC, is reported. The data sa... See more

SAT APR 10 2021

PDF DETAILS

ATLAS PRELIMINARY BEYOND SM

Search for Higgs boson pair production in the two bottom quarks plus two photons final state in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector

Searches are performed for non-resonant and resonant di-Higgs boson production in the $b\bar{b}\gamma\gamma$ final state. The data set used corresponds to an integrated luminosity of 139 fb^{-1} of ... See more

WED MAR 31 2021

PDF DETAILS

ATLAS PRELIMINARY SM

Measurements of gluon fusion and vector-boson-fusion production of the Higgs boson in $H \rightarrow WW^* \rightarrow e\nu\mu\nu$ decays using pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector

The gluon fusion and vector-boson-fusion Higgs boson production modes are measured using proton--proton collisions in the $H \rightarrow WW^* \rightarrow e\nu\mu\nu$ decay channel. The Large Hadron C... See more

WED MAR 31 2021

PDF DETAILS

STATISTICS

Displaying 1012 articles.

Number of articles published per year
120
90
60
30
0
1990
2000
2010
2020
2003
count : 12
Cumulative article count over the years
1200
900
600
300
0
1990
2000
2010
2020

Figure 1: User interface Higgs Boson Portal (HBP).

5. Developments of Limits and Measurements

The "History" link in the user interface gives examples of the developments of limits and measurements in the Higgs boson research (Fig. 2). Clicking on a data point leads to the corresponding publication, and the development of the upper Higgs boson mass limit is taken from Ref. [10].

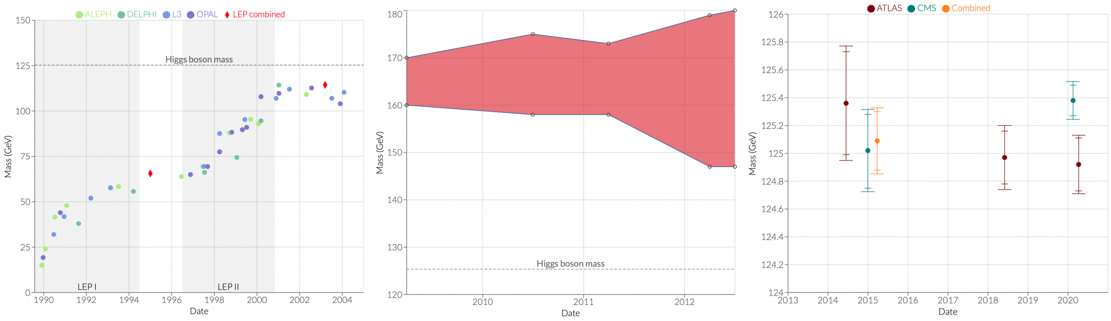


Figure 2: Developments of limits and measurements in the Higgs boson research.

6. Administration and Maintenance

The user interface allows administration login. The administrator can adjust the NLP categorization. Categorized publications are stored in a Mongo database. Updates are performed daily with Python cron jobs.

7. Conclusion

A Higgs Boson Portal has been established with more than 1000 categorized publications and daily updates. It serves as an educational aid and data base for experts.

8. Acknowledgements

The author would like to thank the International Particle Physics Outreach Group (IPPOG) for fruitful discussions, and the students Martin Kupka, Antoine Vauterin and Peter Zacik for their dedication. The grant support from the "Fondu celoškolských aktivit pro rok 2022" Czech Technical University in Prague is gratefully acknowledged. The project is also supported by the Ministry of Education, Youth and Sports of the Czech Republic under the project number LTT 17018.

References

[1] A. Sopczak, *Outreach Modules for New Particle Searches Using the ATLAS Forward Proton Detector and for Higgs Boson Physics*, *PoS* (2022). these proceedings.

[2] A. Sopczak, *Precision measurements in the Higgs sector at ATLAS and CMS*, *PoS* **FFK2019** (2020) 006, [arXiv:2001.0592].

[3] M. Kupka, *Feasibility Study of Portal to Provide Knowledge about Higgs Boson to General Public and Experts*, June, 2020. https://cds.cern.ch/record/2722144, presented 23 June 2020.

[4] P. Zacik, *Implementation of a Portal Dedicated to Higgs Bosons for Experts and the General Public*, May, 2021. https://cds.cern.ch/record/2774895, presented 15 June 2021.

[5] A. Vauterin and A. Sopczak. 22nd IPPOG meeting, 17-19 Nov. 2021 Available online https://indico.cern.ch/event/1084892, 2021.

[6] CDF Collaboration. Available online https://www-cdf.fnal.gov/physics/new/hdg/Published.html.

[7] D0 Collaboration. Available online https://www-d0.fnal.gov/d0_publications/d0_pubs_list_bydate.html.

[8] CERN Document Server. Search Engine API. Available online https://cds.cern.ch/help/hacking/search-engine-api?ln=en.

[9] ATLAS Collaboration, *Measurements of gluon-gluon fusion and vector-boson fusion Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Lett. B* **789** (2018) 508–529, [arXiv:1808.0905].

[10] A. Sopczak, *Status of Higgs boson searches at the beginning of the LHC era*, *Journal of Physics G: Nuclear and Particle Physics* **39** (2012) 113001.