



## PRACA DYPLOMOWA MAGISTERSKA

inż. Paweł Szostek

# Metadata extraction from scholarly publications

Opiekun pracy  
prof. dr hab. inż. Piotr Gawrysiak

Ocena: .....

.....

Podpis Przewodniczącego  
Komisji Egzaminu Dyplomowego





Kierunek:	Informatyka
Specjalność:	Inżynieria Systemów Informatycznych
Data urodzenia:	1987.11.27
Data rozpoczęcia studiów:	01.10.2010

### Życiorys

Paweł Szostek uzyskał tytuł inżyniera na Wydziale Elektroniki i Technik informacyjnych na kierunku Informatyka za prace pod tytułem "Analizator ruchu obiektu wykorzystujący czujniki przyspieszenia". Autor jako student zaczął pracę w Europejskim Ośrodku Badan Jądrowych CERN w 2010 gdzie zajmował się narzędziami do symulacji i budowy projektów dla programowalnych układów logicznych (FPGA). W 2012 autor został zatrudniony w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego, gdzie przez rok pracował nad system ekstrakcji metadanych z publikacji naukowych. W 2013 przeniósł się do CERN openlab, gdzie zajmuje się testowaniem najnowszych platform obliczeniowych, monitorowaniem wydajności sprzętu i jego efektywnym wykorzystaniem, optymalizacją oprogramowania używanego w fizyce wysokich energii oraz zarządzaniem klastrem maszyn deweloperskich.

.....  
Podpis studenta

### EGZAMIN DYPLOMOWY

Złożył egzamin dyplomowy w dniu ..... 20\_\_ r

z wynikiem .....

Ogólny wynik studiów: .....

Dodatkowe wnioski i uwagi Komisji: .....

.....

.....



## STRESZCZENIE

Managing and processing collections of scientific literature has become an important aspect of digital libraries. Instead of handling publications by hand, publishers try to process the articles automatically in order to extract relevant metadata used to feed publication search engines. In this work we present a machine learning-based system for automatic metadata extraction from scholarly publications. Furthermore, we present a data set of open-access articles, made of full texts and classification data, used for training of the classifiers. Finally, we report validation results comparable to cutting-edge systems currently available.

Keywords: metadata extraction, digital libraries, machine learning, SVM

---

## EXTRAKCJA METADANYCH Z PUBLIKACJI NAUKOWYCH

Zarządzanie i przetwarzanie zbiorów literatury naukowej stało się ważnym aspektem bibliotek cyfrowych. Zamiast ręcznie obrabiać publikacje ręcznie, wydawcy próbują przetwarzać artykuły automatycznie w celu wyodrębnienia odpowiednich metadanych stosowanych w wyszukiwarkach publikacji. W tej pracy przedstawiamy system do automatycznej ekstrakcji metadanych z publikacji naukowych oparty o techniki uczenia maszynowego. Ponadto prezentujemy zbiór artykułów do wykorzystania w procesie uczenia klasyfikatorów do przetwarzania publikacji naukowych, składający się z pełnych tekstów opublikowanych na zasadach otwartego dostępu, jak i z danych klasyfikacyjnych. W poniższej pracy przytoczone są także wyniki walidacji systemu, porównywalne do najefektywniejszych systemów obecnie dostępnych na świecie.

Słowa kluczowe: ekstrakcja metadanych, biblioteki cyfrowe, uczenie maszynowe, SVM