Jak przeczytać milion artykułów?

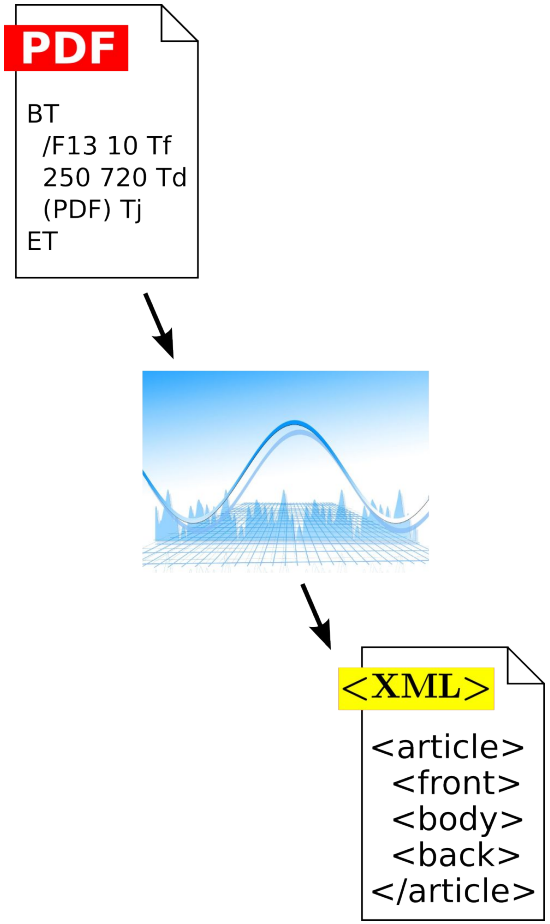
Dominika Tkaczyk
ICM, Uniwersytet Warszawski

10.05.2017

CERMINE

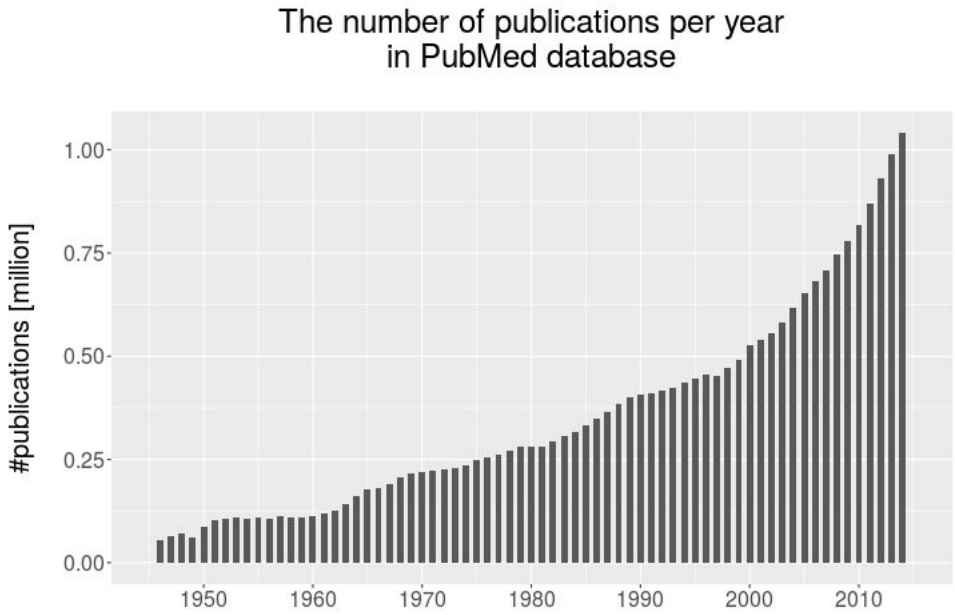
CERMINE to narzędzie przeznaczone do analizy dokumentów naukowych w celu **pozyskania metadanych** w postaci **czytelnej dla maszyn**:

- tytuł dokumentu, lista autorów, ich afiliacje, streszczenie, słowa kluczowe, daty, itp.
- lista odnośników bibliograficznych wraz z metadanymi
- pełen tekst dokumentu z podziałem na sekcje



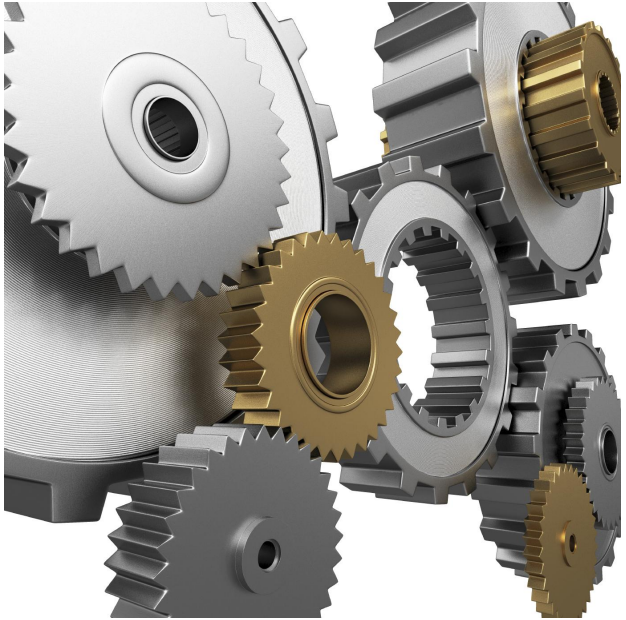
Motywacja

- **automatyczne odzyskiwanie metadanych** z kolekcji dokumentów
- **inteligentne interfejsy** wprowadzania danych

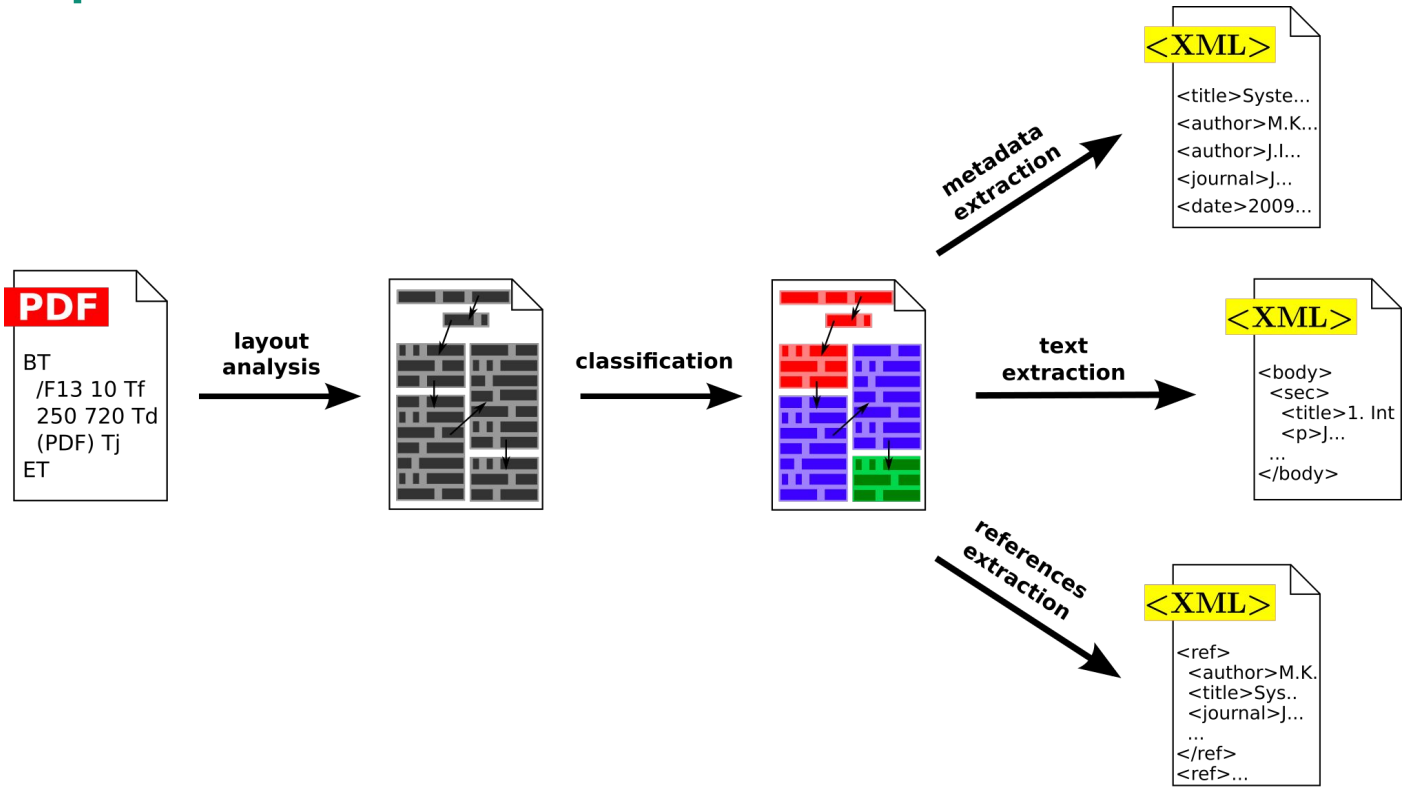


Cechy systemu

- analiza jest w pełni **automatyczna**
- system działa w oparciu o **modularny potok przetwarzania**
- wykorzystywane są techniki **uczenia maszynowego** z nadzorem i bez
- algorytmy analizują **geometryczne własności tekstu**



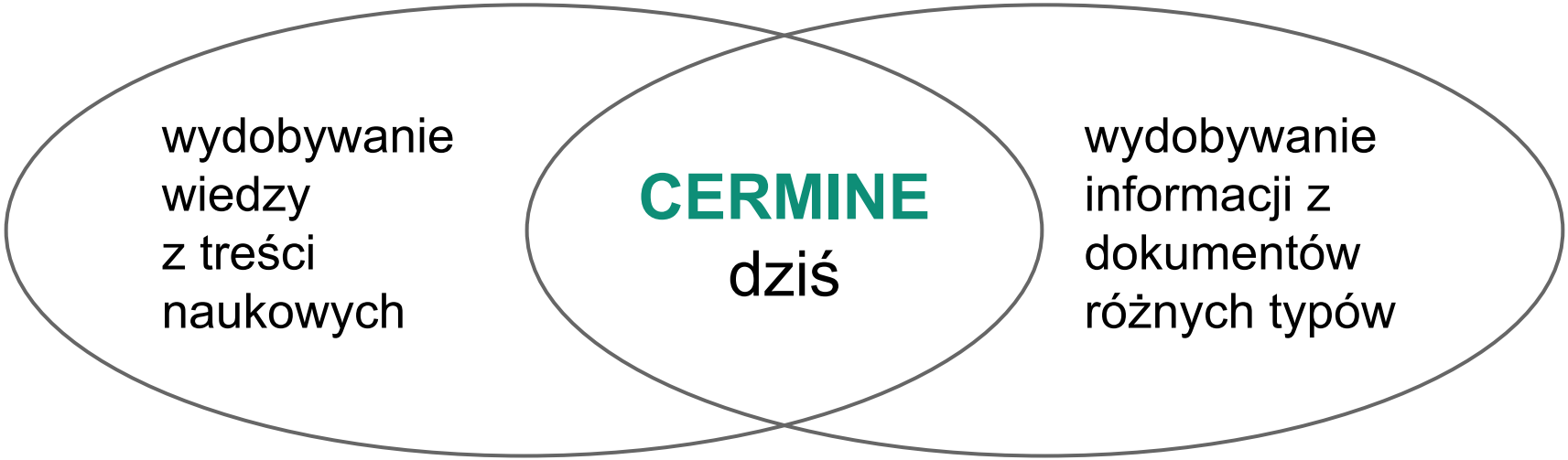
Potok przetwarzania



Efekty projektu

- serwis webowy **cermine.ceon.pl**
- kod źródłowy dostępny na GitHub
- wdrożenia w kilku międzynarodowych projektach
- ESWC 2015 Semantic Publishing Challenge **Best Performing Approach Award**

Co dalej?


wydobywanie
wiedzy
z treści
naukowych
CERMINE
dziś
wydobywanie
informacji z
dokumentów
różnych typów

Dziękuję!

Dominika Tkaczyk
d.tkaczyk@icm.edu.pl