Kluczowym problemem utrudniającym odkrywanie, przypisywanie i ponowne wykorzystywanie otwartego oprogramowania badawczego jest to, że odniesienia do niego często pozostają ukryte w manuskryptach prac badawczych. Aby te zasoby mogły stać się pierwszorzędnymi rekordami bibliograficznymi, muszą być najpierw zidentyfikowane, a następnie zarejestrowane za pomocą trwałych identyfikatorów (PIDs), aby stały się FAIR (Findable, Accessible, Interoperable and Reusable).

Do dziś znaczna część otwartego oprogramowania badawczego nie spełnia zasad FAIR, a zasoby oprogramowania w większości nie są jednoznacznie powiązane z manuskryptami, które się do nich odnoszą.

Projekt ten rozszerzy możliwości kluczowych i szeroko stosowanych otwartych infrastruktur naukowych (CORE, Software Heritage, HAL) i narzędzi (GROBID) obsługiwanych przez partnerów konsorcjum, dostarczając i implementując efektywne rozwiązania do zarządzania cyklem życia oprogramowania badawczego, w tym: 1) identyfikacji zasobów oprogramowania badawczego z manuskryptów prac naukowych wspomaganego przez ML, 2) walidację zidentyfikowanych zasobów przez autorów, 3) rejestrację zasobów oprogramowania za pomocą PIDs i ich archiwizację.

Rozwiązanie będzie zoptymalizowane do zastosowania w otwartych treściach dostępnych poprzez globalną sieć otwartych repozytoriów agregowanych przez CORE (core.ac.uk), która obejmuje ponad 32 mln pełnych tekstów i 250 mln rekordów metadanych z ponad 10 tys. repozytoriów, co stanowi obecnie największą na świecie kolekcję otwartych dokumentów.

Nasze oprogramowanie ML do ekstrakcji i dezambiguacji oprogramowania zostanie zrealizowane jako rozszerzenie najnowszego narzędzia GROBID. Będziemy bazować na uznanych protokołach, takich jak np. OpenAIRE Guidelines v4.0, RIOXX v3 i Codemeta, aby zakodować informacje o zasobach oprogramowania i ich powiązania z manuskryptami badawczymi, tworząc interoperacyjny i rozszerzalny workflow łączący otwarte repozytoria (reprezentowane przez HAL), agregatory (reprezentowane przez CORE) i archiwa oprogramowania (reprezentowane przez Software Heritage).

Skuteczność opracowanych narzędzi i workflows zostanie zweryfikowana w trzech przypadkach użycia: 1) demonstrator dla nauk przyrodniczych (dla Europe PMC), 2) demonstrator wielodyscyplinarny dla instytucjonalnych repozytoriów (reprezentowanych przez HAL) oraz 3) studium przypadku cyfrowej humanistyki (z powiązaniami z DARIAH i EOSC).