OPIS ZAŁOŻEŃ PROJEKTU INFORMATYCZNEGO

Tytuł projektu	Artificial Intelligence Repository „AIR”		
Wnioskodawca	Minister Nauki i Szkolnictwa Wyższego		
Beneficjent	Politechnika Gdańska, Centrum Transferu Wiedzy i Technologii		
Partnerzy	brak		
Źródło finansowania	Budżet państwa, część 27 - Informatyzacja Program Operacyjny Polska Cyfrowa Poddziałanie 2.3.1 „Cyfrowe udostępnienie informacji sektora publicznego ze źródeł administracyjnych i zasobów nauki” (Typ II projektu: Cyfrowe udostępnienie zasobów nauki)		
Całkowity koszt projektu	6 859 475,00 zł		
Planowany okres realizacji projektu	12-2019 do 01-2022		
Osoba kontaktowa	Jerzy Buszke	jerzy.buszke@pg.edu.pl	583486632

1. POWODY PODJĘCIA PROJEKTU

1.1. Identyfikacja problemu i potrzeb

W krajach będących liderami wzrostu gospodarczego, liczba firm budujących rozwiązania oparte o sztuczną inteligencję podwaja się co roku. Szacuje się, że gospodarki krajów, w których działają podmioty aktywnie badające i wdrażające sztuczną inteligencję rosną o 1,5% PKB szybciej od krajów, które tego nie robią. Aby polskie firmy i jednostki naukowe budowały, a nie tylko były biernymi użytkownikami rozwiązań opartych na sztucznej inteligencji, potrzebne są zdecydowane działania na wielu polach. Jednym z nich jest budowa specjalistycznych repozytoriów danych.
Rozwój technologii wykorzystujących sztuczną inteligencję wymaga szerokiego dostępu do specjalistycznych danych. Dane te powinny spełniać kilka kryteriów, dotyczących rozmiaru, jakości, opisu, struktury, anonimizacji, relacji, użyteczności.
Wskazane jest, aby oprócz danych, dostępny był także zbiór modeli (ang. model zoo), wytrenowanych na udostępnionych danych w celu ich bezpośredniego wykorzystania w zagadnieniach związanych z danym zbiorem lub zagadnieniach pokrewnych (ang. transfer learning).
Obecnie istniejące i budowane repozytoria danych w większości nie spełniają w pełni powyższych warunków.
W efekcie braku dostępu do odpowiednich danych, firmy oraz środowisko naukowe wykorzystujące i budujące rozwiązania oparte o AI nie mają możliwości projektowania nowych metod i konkurencyjnych narzędzi, które mogłyby potencjalnie rozwiązywać wiele społecznie i gospodarczo użytecznych zagadnień, np. wspomaganie obrazowania medycznego, tworzenie nowych materiałów, czy syntetyzowanie nowych leków. Zakłada się, że dostęp do odpowiednich danych zainicjuje powstanie wielu mikroprzedsiębiorstw oraz zespołów naukowych wyspecjalizowanych w wykorzystywaniu i budowaniu rozwiązań opartych o AI.
W związku z powyższym istnieje potrzeba zidentyfikowania, przygotowania i udostępnienia odpowiednio opracowanych danych, znajdujących się w posiadaniu naukowców na Politechnice Gdańskiej.

Interesariusz	Zidentyfikowany problem	Szacowana wielkość grupy

Interesariusz	Zidentyfikowany problem	Szacowana wielkość grupy
Firmy wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję	Słaba dostępność do odpowiednio opisanych i dobrej jakości źródłowych danych i modeli przydatnych do uczenia maszynowego	Około 50 firm (Polska) ponad 3000 firm (świat)
Środowisko naukowe wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję	Słaba dostępność do odpowiednio opisanych i dobrej jakości źródłowych danych i modeli przydatnych do uczenia maszynowego	Około 500 naukowców (Polska) kilkaset tysięcy (świat)

1.2. Opis stanu obecnego

W związku z zakrojonymi na szeroką skalę działaniami związanymi z udostępnianiem danych przez administrację publiczną i środowisko naukowe, powstaje w Polsce i zagranicą szereg repozytoriów danych (np. w ramach realizowanego przez konsorcjum Politechniki Gdańskiej, Gdańskiego Uniwersytetu Medycznego oraz Uniwersytetu Gdańskiego projektu MOST DANYCH, realizowanego w ramach Działania 2.3.1 POPC). Dane te będą miały z pewnością szereg zastosowań związanych z analizą statystyczną, wnioskowaniem statystycznym, przewidywaniem trendów, diagnozowaniem potrzeb itp. W zdecydowanej większości nie będą jednak selekcjonowane i opracowywane pod kątem wykorzystania ich do trenowania rozwiązań wykorzystujących sztuczną inteligencję.
Najbardziej rozpoznawalne repozytoriom danych dla uczenia maszynowego, tzw. UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php) zawiera prawie 500 zestawów danych z różnych dziedzin nauki, ale wiele z udostępnionych zestawów danych zawiera skrótowy lub niekompletny opis atrybutów, a ponadto część z nich zawiera bardzo małą liczbę obserwacji. Dodatkowo, otwartych i interdyscyplinarnych repozytoriów tego typu jest wciąż bardzo niewiele.
Na uczelniach, w szczególności technicznych i medycznych, znajdują się bardzo duże zasoby danych badawczych, gromadzone przez dziesięciolecia przez pracowników naukowych. Tylko niewielka część tych danych reprezentuje dużą wartość dla podmiotów zajmujących się budowaniem rozwiązań opartych o AI.
Politechnika Gdańska, jako beneficjent naboru nr POPC.02.03.01-IP.01-00-006/17, buduje platformę informatyczną pozwalającą na gromadzenie, wyszukiwanie, analizowanie i udostępnianie danych badawczych. Będzie więc posiadała odpowiednie doświadczenie w budowaniu rozwiązań sprzętowo-programowych do składowania specjalistycznych danych dla AI.
Duże zasoby przydatnych dla AI danych znajdują się w firmach. Jednak biznes nie chce udostępniać swoich danych ze względu na ryzyko utracenia przewagi konkurencyjnej.

2. EFEKTY PROJEKTU

2.1. Cele i korzysci wynikające z projektu

Cel - 1	Zbudowanie w Polsce silnych kompetencji w zakresie rozwoju rozwiązań opartych o sztuczną inteligencję, poprzez zapewnienie firmom zajmującym się korzystaniem i budowaniem narzędzi w oparciu o AI dostępu do odpowiednio wyselekcjonowanych, precyzyjnie opisanych i dobrej jakości danych, w tym źródłowych danych i modeli, przydatnych do uczenia maszynowego. Dane źródłowe powinny spełniać kryteria:

	1. powinny być możliwe do wykorzystania w uczeniu maszynowym, przykładowo: powinny uwzględniać podział na klasy, posiadać zdefiniowaną liczbę powtórzeń, zawierać częstotliwość próbkowania, powinny posiadać wejściowy zbiór zmiennych oraz pożądaną odpowiedź; 2. powinny być opisane we właściwy sposób, tj. posiadać odpowiednio opracowane metadane; 3. rozmiar i jakość danych powinien umożliwiać wytrenowanie modeli przy użyciu metod uczenia maszynowego na satysfakcjonującym poziomie; 4. dane powinny być odpowiednio przygotowane lub przeprocesowane w celu możliwie łatwego i szybkiego wykorzystania w rozwiązaniach bazujących na AI; 5. powinny być zanonimizowane przy jednoczesnym zachowaniu relacji pomiędzy powiązanymi obiektami; 6. powinny charakteryzować się użytecznością, tj. dotyczyć zagadnień pożądanych do rozwiązania przez firmy i środowisko naukowe. Modele: są to modele uczone i stosowane do rozwiązywania specyficznych problemów, obejmujące pełen wachlarz zagadnień z obszaru sztucznej inteligencji.. Przetrenowane modele można wykorzystać w tzw. „transfer learning”,co pozwoli na wykorzystanie wcześniej wytrenowanego modelu do rozwiązania nowego (ale zbliżonego) problemu. Modele zapisane są w odpowiednich plikach - np. dokumenty z architekturą sieci neuronowej (np. pliki JSON, YAML), dokumenty z wagami sieci neuronowej (np. pliki HDF5), dokumenty z modelami zapisanymi w plikach binarnych (np. pliki SAV)
Cel strategiczny	Cel projektu wpisuje się w następujące strategie i programy: 1. PROGRAM OPERACYJNY POLSKA CYFROWA Przedmiotowy Projekt wykazuje zgodność z celem głównym Programu Operacyjnego Polska Cyfrowa (POPC), jakim jest wzmocnienie cyfrowych fundamentów dla rozwoju kraju. 2. ZAŁOŻENIA DO STRATEGII AI W POLSCE Dokument wskazuje, że celem strategicznym Polski jest bycie w top 20%-25% krajów budujących AI. Oznacza to, że rynek budowy rozwiązań opartych o AI musi do 2025 roku wzrosnąć przynajmniej 24-krotnie do poziomu ~8,3 mld zł przychodów z tworzenia rozwiązań opartych o AI. Oznacza to również, że do tego czasu w Polsce powinno funkcjonować około 720 firm budujących takie rozwiązania. 3. SPRAWNE PAŃSTWO 2020 W ramach celu nr 2.3.2 Efektywne wykorzystanie nowoczesnych technologii cyfrowych przyjęto założenie, że wzrastać będzie wykorzystanie technologii teleinformacyjno-komunikacyjnych (ICT) w sektorze publicznym, w tym w sektorze szkolnictwa wyższego. 4. STRATEGIA INNOWACYJNOŚCI I EFEKTYWNOŚCI GOSPODARKI Zgodność projektu oraz Strategii wynika z faktu, że w ramach jego realizacji powstanie nowoczesna infrastruktura umożliwiająca dostęp do najnowszych zasobów naukowych. 5. PROGRAM ZINTEGROWANEJ INFORMATYZACJI PAŃSTWA W dokumencie wskazano na konieczność podjęcia działań w ramach wsparcia nowoczesnych form podnoszenia umiejętności cyfrowych w ramach edukacji formalnej i nieformalnej. Wśród nich znajduje się zwiększenie dostępności zasobów naukowych . 6. STRATEGIA NA RZECZ ODPOWIEDZIALNEGO ROZWOJU Dostęp do unikalnych danych dla AI pozwoli na szybszy obieg wiedzy,

	wciągnięcie w międzynarodowy obieg nauki, a w konsekwencji wzrost innowacyjności polskiego sektora badań i rozwoju. 7. KRAJOWA STRATEGIA ROZWOJU REGIONALNEGO REGIONY – MIASTA – OBSZARY WIEJSKIE 8. STRATEGIA ROZWOJU KAPITAŁU SPOŁECZNEGO 9. STRATEGIA ROZWOJU POLITECHNIKI GDAŃSKIEJ Celem strategicznym projektu jest przyczynienie się do powstania lub rozbudowy firm oferujących rozwiązania w obszarze AI.
Korzyść:	Dostęp do odpowiednio wyselekcjonowanych, precyzyjnie opisanych i dobrej jakości danych, może zainicjować powstanie startupów nastawionych na budowanie narzędzi informatycznych wykorzystujących sztuczną inteligencję. Dodatkowo, w oparciu o udostępnione dane, firmy będą mogły budować narzędzia przydatne do osiągania przewagi konkurencyjnej.
KPI:	Liczba podmiotów, które udostępniły on-line informacje sektora publicznego
Wartość aktualna i docelowa KPI:	wartość aktualna: 0 (zero) wartość docelowa: 1 (jeden)
Metoda pomiaru KPI	Raport z realizacji projektu AIR. Częstotliwość pomiaru: 1 raz na rok, na zakończenie roku kalendarzowego
Cel - 2	Zbudowanie w Polsce silnych kompetencji w zakresie rozwoju rozwiązań opartych o sztuczną inteligencję, poprzez zapewnienie środowisku naukowemu zajmującemu się korzystaniem i budowaniem rozwiązań i narzędzi w oparciu o AI dostępu do odpowiednio wyselekcjonowanych, precyzyjnie opisanych i dobrej jakości danych, w tym źródłowych danych i modeli, przydatnych do uczenia maszynowego.
Cel strategiczny	Cel projektu wpisuje się w następujące strategie i programy: 1. PROGRAM OPERACYJNY POLSKA CYFROWA Przedmiotowy Projekt wykazuje zgodność z celem głównym Programu Operacyjnego Polska Cyfrowa (POPC), jakim jest wzmocnienie cyfrowych fundamentów dla rozwoju kraju. 2. ZAŁOŻENIA DO STRATEGII AI W POLSCE Dokument wskazuje, że celem strategicznym Polski jest bycie w top 20%-25% krajów budujących AI. Oznacza to, że rynek budowy rozwiązań opartych o AI musi do 2025 roku wzrosnąć przynajmniej 24-krotnie do poziomu ~8,3 mld zł przychodów z tworzenia rozwiązań opartych o AI. Oznacza to również, że do tego czasu w Polsce powinno funkcjonować około 720 firm budujących takie rozwiązania. 3. SPRAWNE PAŃSTWO 2020 W ramach celu nr 2.3.2 Efektywne wykorzystanie nowoczesnych technologii cyfrowych przyjęto założenie, że wzrastać będzie wykorzystanie technologii teleinformacyjno-komunikacyjnych (ICT) w sektorze publicznym, w tym w sektorze szkolnictwa wyższego. 4. STRATEGIA INNOWACYJNOŚCI I EFEKTYWNOŚCI GOSPODARKI Zgodność projektu oraz Strategii wynika z faktu, że w ramach jego realizacji powstanie nowoczesna infrastruktura umożliwiająca dostęp do najnowszych zasobów naukowych. 5. PROGRAM ZINTEGROWANEJ INFORMATYZACJI PAŃSTWA 6. STRATEGIA NA RZECZ ODPOWIEDZIALNEGO ROZWOJU 7. KRAJOWA STRATEGIA ROZWOJU REGIONALNEGO REGIONY – MIASTA –

	OBSZARY WIEJSKIE 8. STRATEGIA ROZWOJU KAPITAŁU SPOŁECZNEGO 9. STRATEGIA ROZWOJU POLITECHNIKI GDAŃSKIEJ Na polskich uczelniach muszą zostać zbudowane nowe kadry zdolne do tworzenia rozwiązań opartych o AI, nauczania zagadnień związanych z AI, aplikowania o dedykowane międzynarodowe programy badawcze. Celem strategicznym projektu jest przyczynienie się do umocnienia i powstania zespołów badawczych działających w obszarze AI.
Korzyść:	Dostęp do odpowiednio wyselekcjonowanych, precyzyjnie opisanych i dobrej jakości danych powinien przyczynić się do wzmocnienia polskiego środowiska naukowego działającego w obszarze AI, zapobieżenia wyjazdom za granicę najlepszych naukowców, zbudowania silnej pozycji naukowej Polski w międzynarodowych programach badawczych. Dodatkowo, dzięki dostępowi do specjalistycznych danych i budowaniu narzędzi opartych o te dane, część naukowców może zdecydować się na założenie spółek typu spin-off.
KPI:	Liczba pobrań/odtworzeń dokumentów zawierających informacje sektora publicznego
Wartość aktualna i docelowa KPI:	wartość aktualna: 0 (zero) wartość docelowa: 250 szt./rok
Metoda pomiaru KPI	Pomiar będzie polegał na sumowaniu wystąpień faktu pobrania lub odtworzenia próbki danych źródłowych lub modelu (dokumentu), tj. wyświetlenia danego zasobu na ekranie użytkownika bezpośrednio (zasób odtworzony przez przeglądarkę internetową przez platformę AIR) lub pośrednio (zasób jest udostępniony w sieci na innym serwerze lub w chmurze poprzez API). Częstotliwość pomiaru: 6 razy na rok
Cel - 3	Zapewnienie firmom i środowisku naukowemu zajmującemu się korzystaniem i budowaniem rozwiązań i narzędzi w oparciu o AI możliwości zestawiania konektorów pozwalających na budowanie relacji oraz fuzji danych źródłowych z różnych źródeł, w celu utworzenia nowych, dedykowanych zbiorów danych źródłowych, dostosowanie ich do potrzeb Interesariuszy i konwersję na dane przydatne do uczenia maszynowego
Cel strategiczny	Cel projektu wpisuje się w następujące strategie i programy: 1. PROGRAM OPERACYJNY POLSKA CYFROWA Przedmiotowy Projekt wykazuje zgodność z celem głównym Programu Operacyjnego Polska Cyfrowa (POPC), jakim jest wzmocnienie cyfrowych fundamentów dla rozwoju kraju. 2. ZAŁOŻENIA DO STRATEGII AI W POLSCE Dokument wskazuje, że celem strategicznym Polski jest bycie w top 20%-25% krajów budujących AI. Oznacza to, że rynek budowy rozwiązań opartych o AI musi do 2025 roku wzrosnąć przynajmniej 24-krotnie do poziomu ~8,3 mld zł przychodów z tworzenia rozwiązań opartych o AI. Oznacza to również, że do tego czasu w Polsce powinno funkcjonować około 720 firm budujących takie rozwiązania. 3. SPRAWNE PAŃSTWO 2020 W ramach celu nr 2.3.2 Efektywne wykorzystanie nowoczesnych technologii cyfrowych przyjęto założenie, że wzrastać będzie wykorzystanie technologii teleinformacyjno-komunikacyjnych (ICT) w sektorze publicznym, w tym w

	sektorze szkolnictwa wyższego. 4. STRATEGIA INNOWACYJNOŚCI I EFEKTYWNOŚCI GOSPODARKI Zgodność projektu oraz Strategii wynika z faktu, że w ramach jego realizacji powstanie nowoczesna infrastruktura umożliwiająca dostęp do najnowszych zasobów naukowych. 5. PROGRAM ZINTEGROWANEJ INFORMATYZACJI PAŃSTWA W dokumencie wskazano na konieczność podjęcia działań w ramach wsparcia nowoczesnych form podnoszenia umiejętności cyfrowych w ramach edukacji formalnej i nieformalnej. Wśród nich znajduje się zwiększenie dostępności zasobów naukowych . 6. STRATEGIA NA RZECZ ODPOWIEDZIALNEGO ROZWOJU Dostęp do unikalnych danych dla AI pozwoli na szybszy obieg wiedzy, wciągnięcie w międzynarodowy obieg nauki, a w konsekwencji wzrost innowacyjności polskiego sektora badań i rozwoju. 7. KRAJOWA STRATEGIA ROZWOJU REGIONALNEGO REGIONY – MIASTA – OBSZARY WIEJSKIE 8. STRATEGIA ROZWOJU KAPITAŁU SPOŁECZNEGO 9. STRATEGIA ROZWOJU PG Dostarczając specyficzne dane na konkretne potrzeby Interesariuszy, projekt przyczyni się do powstania i umocnienia firm i wybranych zespołów naukowych.
Korzyść:	Zarówno firmy, jak i zespoły naukowe zajmujące się AI mają potrzebę dostępu do bardzo specyficznych danych łączących dane z różnych repozytoriów. Takie unikalne i dedykowane zbiory danych będą mogły być tworzone za pomocą konektorów, zaprojektowanych w ramach projektu
KPI:	Liczba udostępnionych on-line dokumentów zawierających informacje sektora publicznego [szt.]
Wartość aktualna i docelowa KPI:	wartość aktualna: 0 (zero) wartość docelowa: 500
Metoda pomiaru KPI	Pomiar będzie polegał na określeniu liczby dokumentów znajdujących się w repozytorium w ramach platformy AIR. Częstotliwość pomiaru: 1 raz na 3 miesiące

2.2. Udostępnione e-usługi

Lp.	Nazwa e-usługi	Typ	Zakres oddziaływania	Poziom dojrzałosci e-usługi
1	Udostępnienie firmom zajmującym się korzystaniem i budowaniem narzędzi w oparciu o AI wyselekcjonowanych, precyzyjnie opisanych i dobrej jakości źródłowych danych i modeli przydatnych do uczenia maszynowego.	A2B	Firmy wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję (rocznie ok 100 transakcji)	Transakcja

Lp.	Nazwa e-usługi	Typ	Zakres oddziaływania	Poziom dojrzałosci e-usługi
2	Udostępnienie środowisku naukowemu zajmującemu się korzystaniem i budowaniem rozwiązań i narzędzi w oparciu o AI wyselekcjonowanych, precyzyjnie opisanych i dobrej jakości źródłowych danych i modeli przydatnych do uczenia maszynowego.	A2A	Środowisko naukowe wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję (rocznie ok 100 transakcji)	Transakcja
3	Umożliwienie interesariuszom (np. innym uczelniom, podmiotom publicznym) udostępniania swoich danych dla trenowania sztucznej inteligencji na podstawie wypracowanego zestawu procedur, obejmujących analizę prawną, analizę przydatności, standardy opisu, anonimizację, model biznesowy, zagadnienia związane z przechowywaniem i bezpieczeństwem danych.	A2B	Firmy wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję Środowisko naukowe wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję (rocznie ok 1 transakcji)	Transakcja
4	Zapewnienie interesariuszom dostępu do danych tworzonych na życzenie z dostępnych w systemie źródeł danych (ang. data on demand), przydatnych do uczenia maszynowego	A2B	Firmy wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję Środowisko naukowe wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję (rocznie ok 1 transakcji)	Transakcja
5	Udostępnienie interesariuszom zbioru modeli (ang. model zoo) wytrenowanych na danych w celu ich bezpośredniego wykorzystania w zagadnieniach uczenia maszynowego związanych z danym zbiorem	A2B	Firmy wykorzystujące i budujące rozwiązania oparte o	Transakcja

Lp.	Nazwa e-usługi	Typ	Zakres oddziaływania	Poziom dojrzałosci e-usługi
	danych lub zagadnieniach pokrewnych (ang. transfer learning)		sztuczną inteligencję Środowisko naukowe wykorzystujące i budujące rozwiązania oparte o sztuczną inteligencję (rocznie ok 50 transakcji)	

2.3. Udostępnione informacje sektora publicznego i zdigitalizowane zasoby

Rodzaj informacji/zasobów	Planowana data udostępnienia	Szacowana liczba obiektów objętych digitalizacją (udostępnianiem informacji)
Źródłowe dane dla uczenia maszynowego	31-01-2022	W ramach projektu nie jest przewidziana digitalizacja zasobów. Dane znajdujące się w dyspozycji naukowców na Politechnice Gdańskiej mają już postać cyfrową. Planowana liczba udostępnionych obiektów (zestaw dokumentów: nagłówek zbioru z danymi + zbiór z danymi + zbiór z metadanymi): 150
Wstępnie przetrenowane modele.	31-01-2022	350

Czy wszystkie zdigitalizowane zasoby objęte projektem będą udostępniane bezpłatnie?
~~TAK~~/NIE
Liczba zasobów jaka nie zostanie udostępniona bezpłatnie: 30 %

2.4. Produkty końcowe projektu

Nazwa produktu	Planowana data wdrożenia
Platforma sprzętowo-programowa do udostępniania źródłowych danych i modeli	01-2022
Konektory do różnych zbiorów danych źródłowych, m.in. umieszczonych na platformie MOST DANYCH, pozwalające na ich dostosowanie do potrzeb Interesariuszy i konwersję na dane przydatne do uczenia maszynowego (np. automatyczne pobieranie, selekcja, preprocessing, etykietowanie, kategoryzowanie lub anonimizacja).	12-2021

3. KAMIENIE MILOWE

Kamienie milowe	Planowany termin osiągnięcia
Ustanowiony kompletny zespół projektowy	2020-02-29
Rozstrzygnięcie postępowania przetargowego na zakup infrastruktury teleinformatycznej	2020-08-31
Zgromadzenie 50% wartości docelowej danych	2021-01-31
Prototyp platformy AIR	2021-03-31
Przygotowana i przetestowana docelowa wersja platformy AIR	2021-09-30
Zgromadzenie 100% wartości docelowej danych	2021-11-30
Zestaw usług: gromadzenie danych, wyszukiwanie/udostępnianie danych, aktualizowanie danych, usuwanie danych	2021-09-30
Opracowany mechanizm tworzenia konektorów do różnych źródeł danych	2021-12-31
Odbiór i uruchomienie platformy sprzętowo-programowej AIR do udostępniania danych do AI	2022-01-31

4. KOSZTY

4.1. Koszty ogólne projektu wraz ze sposobem finansowania

Całkowity koszt projektu (netto oraz brutto), w tym	Netto 6 706 755,00 zł Brutto 6 859 475,00 zł	
Procent dofinansowania ze środków UE (brutto)	84,63%	
Procent środków z budżetu państwa (brutto)	15,37%	
Podział całkowitego kosztu projektu na poszczególna lata (netto oraz brutto)	2019	Netto 29 700,00 zł Brutto 29 700,00 zł
	2020	Netto 3 064 600,00 zł Brutto 3 166 260,00 zł
	2021	Netto 3 232 900,00 zł Brutto 3 283 960,00 zł
	2022	Netto 379 555,00 zł Brutto 379 555,00 zł

4.2. Wykaz poszczególnych pozycji kosztowych

Nazwa pozycji kosztowej		Przewidywany koszt brutto	Uzasadnienie pozycji kosztowej (przeznaczenie)
Oprogramowanie	Wynagrodzenia na opracowanie platformy programowej do udostępniania danych.	1 758 500,00 zł	Opracowanie i wytworzenie platformy programowej do udostępniania źródłowych danych i modeli, uwzględniającej założenia modelu biznesowego polegające na częściowej odpłatności za dane.
Infrastruktura	Macierz dyskowa, serwer, serwer do prac analitycznych, opracowywania i testowania modeli.	270 600,00 zł	Zapewnienie infrastruktury do składowania źródłowych danych i modeli oraz do prac analitycznych.
Koszty UX i grafiki			
Bezpieczeństwo			
Wydajność rozwiązań			
Szkolenia			
Działania informacyjno-promocyjne	Działania informacyjno-promocyjne	369 000,00 zł	Działania niezbędne w celu wypełnienia obowiązków informacyjno-promocyjnych beneficjenta oraz skutecznej promocji projektu wśród interesariuszy
Koszty zarządzania i	Wynagrodzenia	4 461 375,00 z	Koszt związany z opracowaniem

Nazwa pozycji kosztowej		Przewidywany koszt brutto	Uzasadnienie pozycji kosztowej (przeznaczenie)
wsparcia (w tym wynagrodzenia personelu wspomagającego)	zespołu projektowego – zaangażowanego w realizację prac projektowych w obszarze merytorycznym (koordynatorzy merytoryczni, analitycy, naukowcy, którzy wytworzyli dane, prawnik), zarządczym (kierownik projektu, komitet biznesowy), wsparcia (specjaliści ds. zamówień publicznych, promocji, rozliczeń) bezpośrednio związanych z głównymi celami i produktami projektu.	ł	danych pod kątem wykorzystania w uczeniu maszynowym. W opracowaniu muszą wziąć udział naukowcy, którzy wytworzyli dane, analitycy, koordynatorzy merytoryczni, przydatność danych konsultowana jest z komitetem biznesowym, kwestie prawne z prawnikiem. Personel wspomagający to także specjalista ds. rozliczeń, promocji, zamówień publicznych.

4.3. Koszty ogólne utrzymania wraz ze sposobem finansowania (okres 5 lat)

Całkowity koszt utrzymania trwałości projektu (brutto)	90 000,00 zł		Źródło finansowania
Podział całkowitego kosztu utrzymania trwałości projektu na poszczególna lata (netto oraz brutto)	2022	16 500,00 zł (brutto) (16 500,00 zł netto)	środki prywatne
	2023	18 000,00 zł (brutto) (18 000,00 zł netto)	środki prywatne
	2024	18 000,00 zł (brutto) (18 000,00 zł netto)	środki prywatne
	2025	18 000,00 zł (brutto) (18 000,00 zł netto)	środki prywatne
	2026	18 000,00 zł (brutto) (18 000,00 zł netto)	środki prywatne
	2027	1 500,00 zł (brutto) (1 500,00 zł netto)	środki prywatne

4.4. Planowane koszty ogólne realizacji (w przypadku projektu współfinansowanego – wkład krajowy z budżetu państwa) oraz koszty utrzymania projektu:

- zostaną pokryte w ramach budżetów odpowiednich dysponentów części budżetowych bez konieczności występowania o dodatkowe środki z budżetu państwa
~~- będą powodować konieczność przyznania dodatkowych kwot~~

5. GŁÓWNE RYZYKA

5.1. Ryzyka wpływające na realizację projektu

Nazwa ryzyka	Siła oddziaływania	Prawdopodobieństwo wystąpienia ryzyka	Sposób zarzadzania ryzykiem
Trudność w pozyskiwaniu kolejnych zestawów źródłowych danych i modeli	Duża	Średnie	Działania promocyjne skierowane do naukowców, którzy wytworzyli dane. Zabezpieczenie w budżecie projektu odpowiednich wydatków na wynagrodzenie dla naukowców, którzy wytworzyli dane. Bieżący monitoring zamieszczanych zasobów on-line.
Rotacja personelu	Średnia	Średnie	Zapewnienie

Nazwa ryzyka	Siła oddziaływania	Prawdopodobieństwo wystąpienia ryzyka	Sposób zarzadzania ryzykiem
uczestniczącego w projekcie			odpowiedniej liczby pracowników wnioskodawcy zaangażowanych w realizację projektu. Zabezpieczenie w budżecie projektu odpowiednich wydatków na wynagrodzenia osób zaangażowanych w projekt. Odpowiednia motywacja pracowników.
Opóźnienie w realizacji zamówień publicznych	Średnia	Średnie	Przygotowanie procedur przetargowych z odpowiednim wyprzedzeniem czasowym. Zaangażowanie bardzo doświadczonego zespołu zamówień publicznych.
Wzrost nakładów inwestycyjnych	Średnia	Średnie	Zawarcie w umowach pracowniczych odpowiednich klauzul i zapisów regulujących kwestie dot. wynagrodzeń.
Niezawodność funkcjonowania platformy AIR	Mała	Niskie	Bieżący monitoring infrastruktury sprzętowo-programowej.
Małe zainteresowanie interesariuszy korzystaniem z usług platformy AIR	Duża	Niskie	Zabezpieczenie w budżecie projektu odpowiednich wydatków na podjęcie działań promocyjnych skierowanych do poszczególnych grup odbiorców. Bieżący monitoring pobrań i odtworzeń danych udostępnionych on-line.
Opóźnienie w realizacji inwestycji	Duża	Niskie	Dobór osób do poszczególnych grup będzie następował w

Nazwa ryzyka	Siła oddziaływania	Prawdopodobieństwo wystąpienia ryzyka	Sposób zarzadzania ryzykiem
			oparciu o posiadane kwalifikacje i doświadczenie, tak aby zapewnić jak najwyższy poziom kompetencji personalnych. Prace poszczególnych zespołów będą na bieżąco monitorowane przez Kierownika i Koordynatorów projektu, tak aby ewentualność wystąpienia opóźnienia można było wykryć z wyprzedzeniem i aby możliwe było podjęcie działań zapobiegawczych
Problemy z integracją platformy AIR z platformami źródłowymi (na których przechowywane są dane źródłowe, nieprzetworzone na potrzeby wykorzystania do uczenia maszynowego)	Duża	Niskie	Przygotowanie dokumentacji technicznej platformy AIR, zabezpieczenie w budżecie projektu odpowiednich wydatków na wynagrodzenia osób zaangażowanych w projekt.
Trudności w opracowaniu konektorów odpowiedzialnych za automatyczną konwersję danych surowych na dane przydatne do uczenia maszynowego	Średnia	Niskie	Zabezpieczenie w budżecie projektu odpowiednich wydatków na wynagrodzenia osób zaangażowanych w projekt.

5.2. Ryzyka wpływające na utrzymanie efektów

Nazwa ryzyka	Siła oddziaływania	Prawdopodobieństwo wystąpienia ryzyka	Sposób zarzadzania ryzykiem
Utrata bezpieczeństwa i stabilności platformy AIR	Duża	Niskie	Bieżący monitoring działania platformy. Zapewnienie stałej kadry przeznaczonej do obsługi

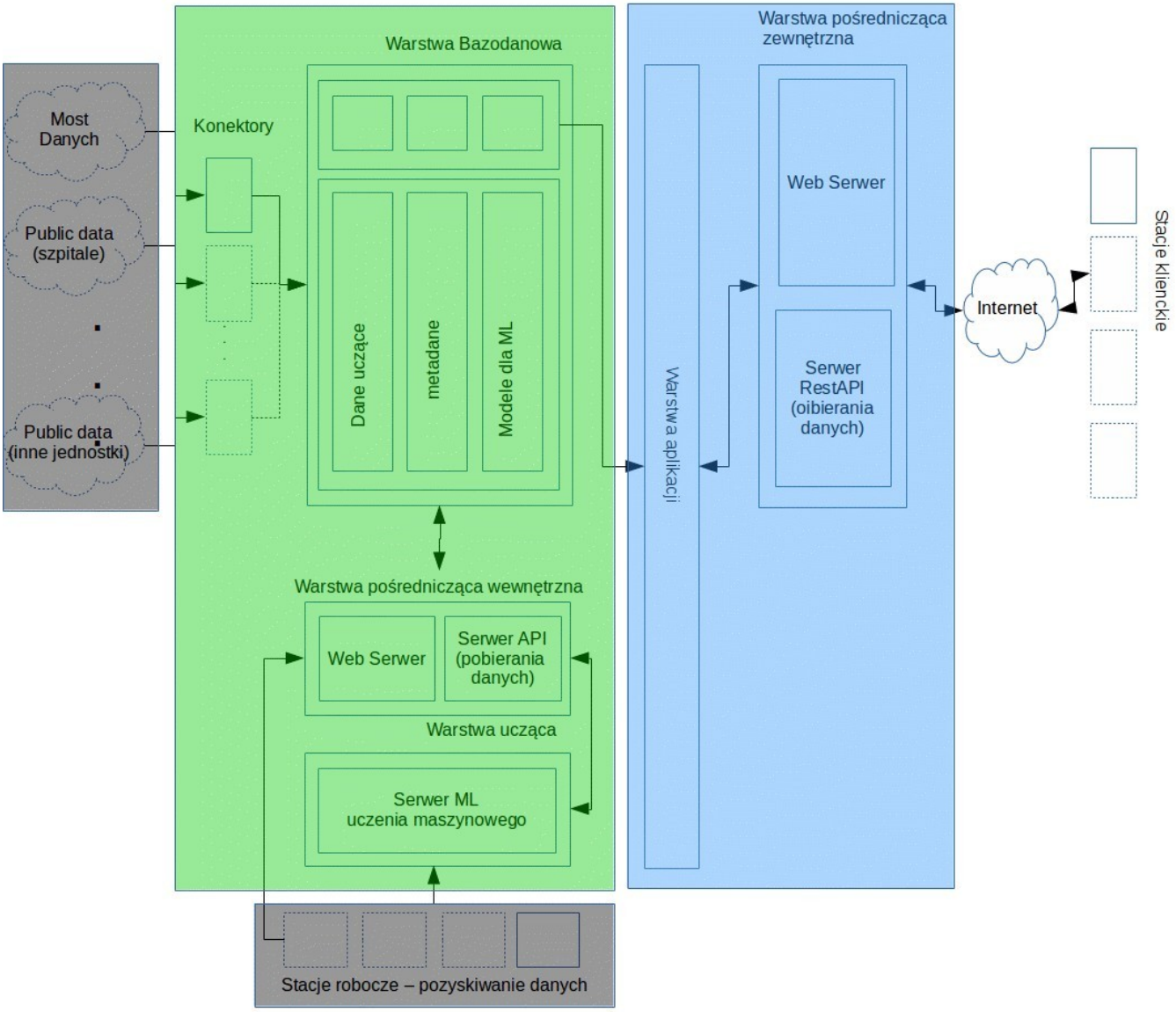
Nazwa ryzyka	Siła oddziaływania	Prawdopodobieństwo wystąpienia ryzyka	Sposób zarzadzania ryzykiem
			bezpieczeństwa i stabilności platformy
Brak zainteresowania naukowców, którzy wytworzyli dane, do opracowywania nowych zestawów danych	Mała	Średnie	Promocja korzyści z udostępniania danych. Przeznaczenie części opłat za udostępniania danych na wynagrodzenia za opracowanie nowych zestawów danych.
Ryzyko rotacji personelu utrzymującego produkty projektu	Średnia	Średnie	Zapewnienie stałej, redundantnej kadry przeznaczonej do utrzymania produktów projektu. Przeznaczenie części opłat za udostępniania danych na wynagrodzenia za opracowanie nowych zestawów danych.

6. OTOCZENIE PRAWNE

Lp.	Tytuł aktu prawnego	Czy wymaga zmian	Opis zmian (jeśli dotyczy)	Etap prac legislacyjnych (jeśli dotyczy)
1	Projekt może zostać zrealizowany w obecnym stanie prawnym, jednak modyfikacja otoczenia prawnego pozwoli zmultiplikować efekt skali i zwiększyć pozytywne efekty projektu. Lista krajowych inteligentnych specjalizacji	TAK/~~NIE~~	Pożądane zmiany: wprowadzenie nowej krajowej inteligentnej specjalizacji: budowanie technologii sztucznej inteligencji. Obecne KIS zakładają w wielu miejscach wykorzystanie sztucznej inteligencji do opracowywania nowych rozwiązań, ale brak jednoznacznego wskazania rozwijania technologii sztucznej inteligencji.	Uzgodnienia wewnętrzne

7. ARCHITEKTURA

7.1. Widok kooperacji aplikacji



Lista systemów wykorzystywanych w projekcie

Lp.	Nazwa systemu	Gestor systemu	Opis systemu	Status	Krótki opis ewentualnej zmiany
1	Warstwa bazodanowa	Politechnika Gdańska	Zespół współpracujących aplikacji i usług (udostępnianych w chmurze oraz na serwerach lokalnych) umożliwiający składowanie i akwizycję danych w tym danych uczących, danych	Planowany	

Lp.	Nazwa systemu	Gestor systemu	Opis systemu	Status	Krótki opis ewentualnej zmiany
			dotyczących opisu eksperymentu, metadanych oraz modeli.		
2	Warstwa pośrednicząca wewnętrzna	Politechnika Gdańska	Bazuje na serwerze aplikacyjnym i kompatybilnym frameworku. Jej zadaniem będzie umożliwienie komunikacji pomiędzy odpowiednimi komponentami całego systemu (dzięki wykorzystaniu technologii np. takich jak REST, JSON, ODBC, JDBC, PHP...)	Planowany	
3	Warstwa ucząca	Politechnika Gdańska	System zbudowany na potrzeby przeprowadzania weryfikacji algorytmów i zbiorów uczących oraz tworzenia modeli opartych na danych. Wykorzystujący technologię opensource (np. system z rodziny UNIX) wraz ze wsparciem dla kart graficznych (np. sterowniki cuDNN, CUDA) oraz przygotowanymi frameworkami pozwalającymi na uczenie maszynowe (np. TensorFlow, Keras, PyToarch, scikit-learn).	Planowany	
4	Warstwa aplikacji	Politechnika Gdańska	Oparta jest o sprawdzone w ramach wcześniejszych prac Centrum Usług Informatycznych rozwiązanie. Obecnie jest to stos technologiczny PHP oraz Nginx. Dostęp i serwowanie danych statycznych jest oparte o serwer WWW Nginx, natomiast wybór silnika obsługującego język PHP padł na rozwiązanie PHP-FPM, które jest aktualnie jedną z najbardziej	Modyfikowany	Zakłada się wykorzystanie innej technologii niż NGINX czy PHP dla Web Serwer i Serwera restAPI

Lp.	Nazwa systemu	Gestor systemu	Opis systemu	Status	Krótki opis ewentualnej zmiany
			wydajnych metod przetwarzania skryptów PHP. Sam serwis wykorzystuje Framework Symfony oraz Bootstrap. Rozwiązanie w sprawny i wygodny sposób pozwala na tworzenie responsywnych stron webowych, dostosowanych do najnowszych standardów.		
5	Warstwa pośrednicząca zewnętrzna	Politechnika Gdańska	Oparta jest o sprawdzone w ramach wcześniejszych prac Centrum Usług Informatycznych rozwiązanie, czyli stos technologiczny PHP oraz Nginx. Dostęp i serwowanie danych statycznych jest oparte o serwer WWW Nginx, natomiast wybór silnika obsługującego język PHP padł na rozwiązanie PHP-FPM, które jest aktualnie jedną z najbardziej wydajnych metod przetwarzania skryptów PHP. Sam serwis wykorzystuje Framework Symfony oraz Bootstrap. Rozwiązanie w sprawny i wygodny sposób pozwala na tworzenie responsywnych stron webowych, dostosowanych do najnowszych standardów.	Modyfikowany	Zakłada się wykorzystanie innej technologii niż NGINX czy PHP dla Web Serwer i Serwera restAPI
6	MOST DANYCH	Politechnika Gdańska	Repozytorium danych badawczych	Istniejący	
7	Public Data	Dane publiczne	Repozytoria danych finansowanych z funduszy publicznych (np. bazy danych w szpitalach, centrum zarządzania ruchem). System poza zakresem projektu, zaprezentowany w celu	Istniejący	

Lp.	Nazwa systemu	Gestor systemu	Opis systemu	Status	Krótki opis ewentualnej zmiany
			demonstracji oczekiwanej i docelowej architektury systemu.		
8	Stacje robocze przesyłania danych	Politechnika Gdańska	Stanowiska robocze naukowców opracowujących eksperymenty dla pozyskiwania danych dla uczenia maszynowego lub trenowania modeli.	Istniejący	
9	Konektory	Politechnika Gdańska	Element odpowiedzialny za budowanie relacji oraz fuzji danych źródłowych z różnych repozytoriów w celu utworzenia nowego, dedykowanego zbioru danych.	Planowany	
10	Stacje klienckie	Użytkownicy (np. Interesariusze)	Stacje klienckie (komputery PC, urządzenia mobilne)	Istniejący	

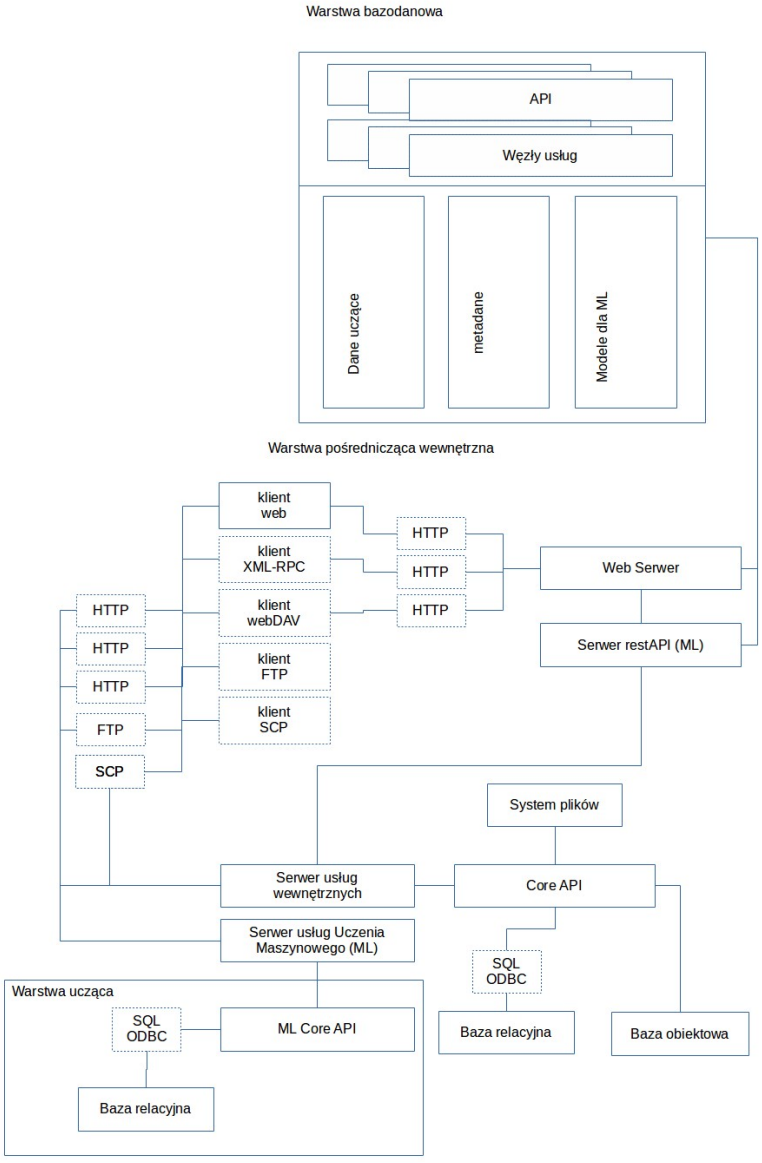
Lista przepływów

Lp.	System źródłowy	System docelowy	Zakres wymienianych danych	Sposób wymiany danych	Typ modyfikacji	Typ interfejsu
1	MOST DANYCH	Warstwa bazodanowa	Dane źródłowe, niemodyfikowane	Kopiowanie danych	Normalizacja i przekształcanie danych realizowane przez dziedzinowe konektory opracowane i zaimplementowane w toku realizacji projektu. Konektory są elementami krytycznymi dla realizacji projektu.	Np. HTTP / XML lub inny dziedzinowo odpowiedni do możliwości technicznych źródła danych (po stronie MOSTU DANYCH).
2	Warstwa bazodano	Warstwa aplikacji	Dane źródłowe (dane uczące),	Odwołania bezpośrednie	Dane nie podlegają	

Lp.	System źródłowy	System docelowy	Zakres wymienianych danych	Sposób wymiany danych	Typ modyfikacji	Typ interfejsu
	wa		metadane oraz modele dla ML		modyfikacji	
3	Warstwa bazodanowa	Warstwa pośrednicząca wewnętrzna	Dane źródłowe (dane uczące), metadane oraz modele dla ML	Kopiowanie danych	obowiązkowy	SCP, HTML, SSH, lub inny dziedzinowo odpowiedni protokół transmisji i komunikacji
4	Warstwa pośrednicząca wewnętrzna	Warstwa ucząca	Dane źródłowe (dane uczące), metadane oraz modele dla ML	Odwołania bezpośrednie lub kopiowanie danych	obowiązkowy	SCP, HTML, SSH, lub inny dziedzinowo odpowiedni protokół transmisji i komunikacji
5	Warstwa ucząca	Warstwa pośrednicząca wewnętrzna	Metadane oraz wytrenowane, nowe modele lub modele douczone	Kopiowanie danych	obowiązkowy	SCP, HTML, SSH, lub inny dziedzinowo odpowiedni protokół transmisji i komunikacji
6	Warstwa aplikacji	Warstwa pośrednicząca zewnętrzna	Dane źródłowe (dane uczące), metadane oraz modele dla ML	Odwołania bezpośrednie lub kopiowanie danych	obowiązkowy	SCP, HTML, SSH, lub inny dziedzinowo odpowiedni protokół transmisji i komunikacji
7	Stacje robocze pozyskiwania danych	Warstwa ucząca	Dane źródłowe (dane uczące), metadane oraz modele dla ML oraz skrypty	Odwołania bezpośrednie lub kopiowanie danych	obowiązkowy	np. SCP, HTML, SSH, lub inny dziedzinowo odpowiedni protokół transmisji i komunikacji
8	Stacje robocze pozyskiwania danych	Warstwa pośrednicząca wewnętrzna	Dane źródłowe (dane uczące), metadane oraz modele dla ML	Odwołania bezpośrednie lub kopiowanie danych	obowiązkowy	np. SCP, HTML, SSH, XML, lub inny dziedzinowo odpowiedni protokół transmisji i komunikacji
9	Warstwa	Warstwa	Dane źródłowe	Kopiowanie	obowiązkowy	np. SCP,

Lp.	System źródłowy	System docelowy	Zakres wymienianych danych	Sposób wymiany danych	Typ modyfikacji	Typ interfejsu
	pośrednicząca wewnętrzna	bazodanowa	(dane uczące), metadane, modele dla ML	danych		HTML, SSH, lub inny dziedzinowo odpowiedni protokół transmisji i komunikacji

7.2. Kluczowe komponenty architektury rozwiązania



7.3. Przyjęte założenia technologiczne

Lp.	Obszar	Założenie technologiczne
1.	Infrastruktura	
2.	Sieć i bezpieczeństwo	
3.	Standardy wymiany danych	
4.	Systemy operacyjne serwerowe	
5.	Bazy danych	
6.	Serwery aplikacji	
7.	Portale	
8.	Inne	

7.4. Opis zasobów danych przetwarzanych w planowanym rozwiązaniu

Czy nowy system będzie tworzył zasoby danych o charakterze rejestru publicznego?
~~TAK~~/NIE
Czy nowy system będzie przetwarzał (używał, zmieniał) zawartość innych rejestrów publicznych?
~~TAK~~/NIE

7.5. Bezpieczeństwo

Planowany poziom zapewnienia bezpieczeństwa (w rozumieniu przepisów §20 rozporządzenia Rady Ministrów z dnia 12 kwietnia 2012 r. w sprawie Krajowych Ram Interoperacyjności […] (Dz. U. 2012, poz. 526 z późn. zm.) w zakresie dot. systemu zarządzania bezpieczeństwem informacji:
~~- system nie podlega rygorom KRI – należy wyjaśnić czy istnieją inne normy bezpieczeństwa, które będą spełnione przez system zgodnie z wymogami KRI~~
~~- dodatkowe zabezpieczenia powyżej wymogów KRI: należy wskazać uzasadnienie~~