

OPIS ZAŁOŻEŃ PROJEKTU INFORMATYCZNEGO

Tytuł projektu	Digitalizacja cennych zasobów nauki i udostępnienie danych genomicznych z wykorzystaniem infrastruktury Polskiego Węzła Europejskiego Archiwum Genomów		
Wnioskodawca	Minister Nauki i Szkolnictwa wyższego		
Beneficjent	Uniwersytet Łódzki		
Partnerzy	Pomorski Uniwersytet Medyczny w Szczecinie Instytut Chemii Bioorganicznej Polskiej Akademii Nauk - Poznańskie Centrum Superkomputerowo-Sieciowe Akademia Górniczo-Hutnicza im. S. Staszica w Krakowie, Akademickie Centrum Komputerowe Cyfronet AGH		
Źródło finansowania	Fundusze Europejskie dla Nowej Gospodarki Działanie 2.3 Cyfrowa dostępność i ponowne wykorzystanie informacji - nauka		
Całkowity koszt projektu	37 019 980,00 zł		
Planowany okres realizacji projektu	01-2027 do 12-2029		
Osoba kontaktowa	Błażej Marciniak	blazej.marciniak@biol.uni.lo dz.pl	600936417

1. POWODY PODJĘCIA PROJEKTU

1.1. Identyfikacja problemu i potrzeb

Dane genetyczne stały się nieocenionym zasobem także w medycynie i szeroko pojętej branży ochrony zdrowia. Dla przykładu dzięki danym zgromadzonym w UK Biobank (500 000 genomów) opracowano statystyki populacji UK, system oceny ryzyka wystąpienia jednostki chorobowej na podstawie danych genetycznych (Poligenic Risk Score - PRS). Dzięki temu możliwe jest odpowiednie przygotowanie polityk profilaktyki czy ochrony zdrowia. Od tego czasu systematycznie spadają statystyki zgonów w wyniku chorób nowotworowych czy sercowo naczyniowych. Co więcej dzięki otwartemu dostępowi cały świat ma możliwość analizy tych danych/tej populacji a wyniki usprawniają tylko system ochrony zdrowia UK. Nowością, która powstanie w projekcie będą dane WGS + Epigenom, które jeszcze nie funkcjonują w ekosystemie nauki z racji na wcześniejszy brak technologii.

Sektor AI, tylko dzięki dużej dostępności danych przeżywa ogromny rozkwit – brak danych dotyczących populacji polskiej ogranicza postępy w tym zakresie ograniczając szanse na opracowanie terapii czy metod diagnostycznych – trenowanie modeli w oparciu o inne częstości występowania wariantów czy skalibrowane PRS dla naszej populacji.

Pomimo iż w Polsce realizuje się dużo badań naukowych w trakcie których generowane są sekwencje genetyczne nie trafiają one później do otwartego obiegu i dostęp do nich jest utrudniony, jeśli nie niemożliwy. Zapewne jest to wynikiem stosowania (słusznie) RODO do regulacji obrotu tego typu danymi. Podobnie sprawa wygląda z danymi generowanymi na potrzeby diagnostyczne - również mogłyby zasilić ekosystem nauki – zmiany, które niesie EHDS. Naturalnie dochodzimy do kolejnego wyzwania jakim jest przygotowanie techniczne/infrastrukturalne do obsługi założeń EHDS - którą to lukę może wypełnić silny i skalowany Polski Węzeł FEGA, którego wzmocnienie zaplanowano w projekcie. Brak danych i utraty korzyści z tym

związanych oraz niedobory infrastrukturalne są główną przesłanką do podjęcia realizacji projektu

Interesariusz	Zidentyfikowany problem	Szacowana wielkość grupy
Pracownicy naukowi	<ul style="list-style-type: none"> - brak dostępnych w otwartym dostępie/ bezpłatnych wysokiej jakości danych genomicznych i epigenomicznych umożliwiających prowadzenie badań podstawowych w dziedzinie Genetyki i Epigenetyki populacji polskiej lub medycynie spersonalizowanej - niewystarczająca liczba centrów eksperckich wspierających procesy digitalizacji tego typu danych w Polsce - generowanie nowych danych podnosi koszty i wydłuża czas badań 	<p>Personel badawczy: 72 556 źródło: Raport „Nauka w Polsce 2023”, Ośrodek Przetwarzania Informacji – PIB (OPI)</p> <p>Nauczyciele akademicy: 96 773, źródło: POLON</p> <p>Doktoranci: 19 097, źródło: POLON</p> <p>Łącznie: 188 426</p>
Personel B+R	<ul style="list-style-type: none"> - niewystarczająca liczba centrów eksperckich wspierających procesy digitalizacji tego typu danych w Polsce - brak dostępnych w otwartym dostępie/ darmowych wysokiej jakości danych umożliwiających prowadzenie badań podstawowych w dziedzinie Genetyki populacji polskiej lub medycynie spersonalizowanej - generowanie nowych danych podnosi koszty i wydłuża czas badań 	<p>305 000 źródło: Raport „Nauka w Polsce 2023”, Ośrodek Przetwarzania Informacji – PIB (OPI)</p>
Lekarze	<p>Brak skalibrowanych danych dotyczących genetycznego ryzyka wystąpienia jednostki chorobowej dla polskiej populacji</p>	<p>166 515, źródło: GUS - „Zasoby kadrowe w wybranych zawodach medycznych na podstawie źródeł administracyjnych w 2024 r.”</p>
Przemysł farmaceutyczny	<p>brak dostępnych w otwartym dostępie wysokiej jakości danych umożliwiających prowadzenie badań</p> <p>brak danych służących do budowy modeli oddziaływania leków w populacji polskiej.</p> <p>niewystarczająca liczba centrów eksperckich wspierających procesy digitalizacji tego typu danych w Polsce</p> <p>Innowacyjne terapie opracowywane w innych rejonach świata dla populacji o innym rozkładzie wariantów genetycznych mogą okazać się mniej skuteczne dla polskich pacjentów; podobnie rzecz się ma, jeśli chodzi o skutki uboczne terapii - posiadając takie dane możliwa będzie dokładniejsza prognoza oddziaływania/skuteczności danej terapii.</p>	<p>697 podmiotów, źródło: https://alertmedyczny.pl/nowe-firmy-w-zdrowiu-mniej-w-farmacji-gus-publickuje-dane-o-firmach-z-marca-2025/#wszystkie-podmioty-w-marcu-2025-wg-rejestru-gus</p>

Interesariusz	Zidentyfikowany problem	Szacowana wielkość grupy
	Posiadanie odpowiedniej ilości wysokiej jakości danych genomicznych o polskiej populacji pozwoli na wykorzystanie tej wiedzy przy projektowaniu nowoczesnych terapii przez krajowe podmioty	
Branża AI	Brak danych dokładnych danych genomicznych opisujących polską populację gotowych do wykorzystania w przygotowaniu algorytmów AI. Dostarczenie odpowiednich danych pozwoli na przygotowanie bardziej wiarygodnych modeli AI, mogących wesprzeć lekarzy lub system ochrony zdrowia, Dane te mogą stanowić zarówno komponent treningowy zwiększając różnorodność i reprezentacyjność danych wykorzystywanych do treningu modeli AI, jak również niezależny zbiór walidacyjny pozwalający na wiarygodną ewaluację modeli AI	13 000 podmiotów, źródło: https://26pietro.pl/ile-firm-w-polsce-zajmuje-sie-ai-sa-nowe-dane/
System ochrony zdrowia	<p>Brak danych dokładnych danych genomicznych opisujących polską populację.</p> <p>Statystyki wariantów genetycznych występujących w populacji mogą stać się cennym narzędziem w planowaniu profilaktyki - dla przykładu znajomość częstości występowania i rozkładu występowania wariantów predysponujących do ostrego przebiegu COVID pozwoliłaby odpowiednio zaplanować rozmieszczenie miejsc hospitalizacji;</p> <p>Podobnie w przypadku występowania zwiększonego ryzyka poligenicznego wystąpienia danej jednostki chorobowej na jakimś obszarze pozwoli na dobór odpowiednich działań profilaktycznych,</p> <p>Ponadto znajomość częstości występowania wariantów pozwala określić aplikacyjność testów molekularnych przed ich wdrożeniem np. test genetyczny szacujący predyspozycję do danej jednostki chorobowej mający zastosowanie w innych populacjach może być nieskuteczny w populacji PL ze względu na niską częstość występowania danego wariantu</p>	305 620 podmiotów, źródło: https://alertmedyczny.pl/nowe-firmy-w-zdrowiu-mniej-w-farmacji-gus-publicuje-dane-o-firmach-z-marca-2025/#wszystkie-podmioty-w-marcu-2025-wg-rejestru-gus

1.2. Opis stanu obecnego

digitalizacja materiałów biologicznych odbywa się w trakcie procesu sekwencjonowania - generowane są cyfrowe zapisy odczytanych sekwencji DNA. Następnie w trakcie analiz bioinformatycznych generowany jest produkt docelowy - sekwencja pełnego genomu WGS. Infrastruktura Polskiego Węzła Europejskiego Archiwum Genomów (PL-FEGA) zintegrowana jest z Centralnym węzłem EGA za pomocą interfejsów API i brokerów wiadomości RabbitMQ. Wrażliwe dane dotyczące polskiej populacji przechowywane są w infrastrukturze PL-FEGA w formie szyfrogramu wygenerowanego przez algorytm Crypt4GH. Zarządzanie użytkownikami i przechowywanie niewrażliwych w rozumieniu RODO metadanych realizowane jest w Centralnym EGA, odpowiedzialnym za dostarczenie oprogramowania do zarządzania dostępem do danych - Portal Komitatu Dostępu do Danych (DAC Portal). Infrastruktura jest dostępna produkcyjnie od końca 2023 roku. Skuteczność infrastruktury może potwierdzić historia pobrań danych a bezpieczeństwo przeprowadzone w 2025 roku testy penetracyjne. Obecnie dane stanowią podstawę rozwoju nowoczesnej gospodarki, dane genomiczne nowoczesnej medycyny i profilaktyki zdrowia. UE stawia na zwiększenie innowacyjności (raport Draghi) również w tym obszarze uruchamiając ogromne projekty GDI, "Genome of Europe" (GoE) - jako część większego programu 1+MG, którego celem jest milion WGS mieszkańców Europy. Zgodnie z dokumentacją GoE aby należycie opisać polską populację należy dysponować 50-70 tys. Genomów. Obecnie w otwartym dostępie z wykorzystaniem PL-FEGA udostępniono 400 WGS. Dane wygenerowane w projektach naukowych nie stają się publicznie dostępne. Dla porównania UK dysponuje 500 tys a ZEA 800 tys WGS. W Polsce równolegle trwa projekt G4PL produkujący również dane genomiczne - projekt zawiera komponent komercyjny i inny zakres danych i wytworzenie narzędzi i usług. Choć obie inicjatywy się dopełniają, w tym projekcie kładziony jest nacisk na nową technologię i unikatowe obecnie dane WGS+Epigenom.

2. EFEKTY PROJEKTU

2.1. Cele i korzyści wynikające z projektu

Cel - 1	Digitalizacja i udostępnienie zasobów nauki
Cel strategiczny	<p>Bardziej konkurencyjna i inteligentna Europa dzięki wspieraniu innowacyjnej i inteligentnej transformacji gospodarczej oraz regionalnej łączności cyfrowej – Cel działania FERC</p> <p>Polityka rozwoju sztucznej inteligencji w Polsce do 2030 – otwarte dane</p> <p>Polityka naukowa Państwa – otwarte dane, otwarta nauka</p> <p>Krajowej Strategii Rozwoju Regionalnego 2030 - wzmacnianie potencjałów badawczych poszczególnych regionów.</p> <p>Genome of Europe - tworzenie europejskiej bazy danych genomicznych. 1+ Million Genomes (1+MG) - dostarczenie miliona WGS pochodzących od populacji europejskiej</p> <p>Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2025/327 z dnia 11 lutego 2025 r. w sprawie europejskiej przestrzeni danych dotyczących zdrowia oraz zmiany dyrektywy 2011/24/UE i rozporządzenia (UE) 2024/2847 - EHDS - ponowne (secondary use) wykorzystanie danych</p> <p>Agenda Europejskiej Przestrzeni Badawczej - otwarty dostęp do danych oraz transgraniczny dostęp do danych</p>

Korzyść:	Zwiększenie potencjału badawczego polskich uczelni, obniżenie kosztów prowadzenia prac badawczo rozwojowych, badań podstawowych, możliwość prowadzenia badań nad skutecznością leków w odniesieniu do populacji polskiej, rozwój biotechnologii, dane służące rozwojowi sztucznej inteligencji
KPI:	1 - Rozmiar wytworzonych danych [TB]; 2 - Rozmiar udostępnionych danych [TB]; 3 - ilość zdigitalizowanych dokumentów [szt.]; 4- ilość udostępnionych dokumentów [szt.]
Wartość aktualna i docelowa KPI:	1 - 0; 2 - 0; 3 - 0; 4 - 0; 1 - 60 TB; 2 - 60 TB; 3 - 1400 szt.; 4 - 1400 szt.
Metoda pomiaru KPI	<p>Metoda:</p> <p>1 - Zsumowanie objętości wszystkich zdigitalizowanych zasobów (informacji sektora publicznego) wyrażonej w TB, Sposób pomiaru: narzędzie systemowe szacujące objętość plików np. polecenie dh lub na podstawie protokołów odbioru produktów projektu zaakceptowane przez Kierownika projektu. Częstotliwość pomiaru zgodnie z harmonogramem raportów okresowych - zgodnie z DIP;</p> <p>2 - Zsumowanie objętości wszystkich udostępnionych on-line cyfrowych zasobów (informacji sektora publicznego) wyrażonej w TB, Sposób pomiaru: narzędzie systemowe szacujące objętość plików np. polecenie dh lub na podstawie protokołów odbioru produktów projektu zaakceptowane przez Kierownika projektu. Częstotliwość pomiaru zgodnie z harmonogramem raportów okresowych - zgodnie z DIP;</p> <p>3 - suma ilości zdigitalizowanych zasobów nauki wyrażonej w TB, sposób pomiaru suma ilości zdigitalizowanych zasobów liczona na podstawie zaakceptowanych przez kierownika projektu protokołów odbioru produktów</p> <p>4 - suma ilości udostępnionych cyfrowych wersji zasobów z poziomu systemu Sposób pomiaru: suma dokumentów udostępnionych w systemie lub suma ilości dokumentów przeznaczonych do udostępnienia na podstawie protokołów odbioru produktów zaakceptowanych przez kierownika projektu.</p>
Cel - 2	Utworzenie rozproszonej i skalowalnej infrastruktury IT pozwalającej na udostępnienie danych o Ludzkim DNA z terytorium RP w bezpieczny sposób
Cel strategiczny	<p>Bardziej konkurencyjna i inteligentna Europa dzięki wspieraniu innowacyjnej i inteligentnej transformacji gospodarczej oraz regionalnej łączności cyfrowej – Cel działania FERC</p> <p>Polityka naukowa Państwa – budowa infrastruktury udostępniania danych (repozytoria) - Krajowej Strategii Rozwoju Regionalnego 2030 - wzmacnianie potencjałów badawczych poszczególnych regionów.</p> <p>Genome of Europe - tworzenie europejskiej bazy danych genomicznych w ramach inicjatywy 1+MG.</p> <p>1+ Million Genomes (1+MG) - dostarczenie miliona WGS pochodzących od</p>

	<p>populacji europejskiej</p> <p>Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2025/327 z dnia 11 lutego 2025 r. w sprawie europejskiej przestrzeni danych dotyczących zdrowia oraz zmiany dyrektywy 2011/24/UE i rozporządzenia (UE) 2024/2847 - EHDS - ponowne (secondary use) wykorzystanie danych</p> <p>Agenda Europejskiej Przestrzeni Badawczej - otwarty dostęp do danych oraz transgraniczny dostęp do danych</p>
Korzyść:	<p>Budowa gospodarki opartej na wiedzy</p> <p>Skalowalna infrastruktura udostępniania danych</p>
KPI:	<p>Liczba Instytucji publicznych otrzymujących wsparcie na opracowywanie usług, produktów i procesów cyfrowych</p>
Wartość aktualna i docelowa KPI:	<p>4</p>
Metoda pomiaru KPI	<p>Oszacowanie liczby jednostek objętych wsparciem odbywać się będzie na podstawie zawartych umów - umowy o dofinansowanie projektu oraz umowy konsorcjum i potwierdzone przez kierownika projektu</p>
Cel - 3	<p>Udostępnienie wysokiej jakości danych cyfrowych na potrzeby polskiej nauki i sektora B+R, AI lub systemu ochrony zdrowia</p>
Cel strategiczny	<p>Bardziej konkurencyjna i inteligentna Europa dzięki wspieraniu innowacyjnej i inteligentnej transformacji gospodarczej oraz regionalnej łączności cyfrowej – Cel działania FERC</p> <p>Polityka rozwoju sztucznej inteligencji w Polsce do 2030 – otwarte dane</p> <p>Polityka naukowa Państwa – otwarte dane, otwarta nauka</p> <p>Krajowej Strategii Rozwoju Regionalnego 2030 - wzmacnianie potencjałów badawczych poszczególnych regionów.</p> <p>Genome of Europe - tworzenie europejskiej bazy danych genomicznych.</p> <p>1+ Million Genomes (1+MG) - dostarczenie miliona WGS pochodzących od populacji europejskiej</p> <p>Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2025/327 z dnia 11 lutego 2025 r. w sprawie europejskiej przestrzeni danych dotyczących zdrowia oraz zmiany dyrektywy 2011/24/UE i rozporządzenia (UE) 2024/2847 - EHDS - ponowne (secondary use) wykorzystanie danych</p> <p>Agenda Europejskiej Przestrzeni Badawczej - otwarty dostęp do danych oraz transgraniczny dostęp do danych</p>
Korzyść:	<p>Budowa gospodarki opartej na wiedzy, możliwość doboru skuteczniejszych leków/terapii, obniżenie kosztów prac B+R.</p> <p>Zwiększenie dostępności wysokiej jakości danych na potrzeby interesariuszy</p>
KPI:	<p>1 - Rozmiar udostępnionych danych; 2 - ilość udostępnionych dokumentów</p>
Wartość	<p>1 - 0; 2 - 0</p>

aktualna i docelowa KPI:	1 - 60 TB; 2 - 1400 szt.
Metoda pomiaru KPI	<p>1 - Zsumowanie objętości wszystkich udostępnionych on-line cyfrowych zasobów (informacji sektora publicznego) wyrażonej w TB, Sposób pomiaru: narzędzie systemowe szacujące objętość plików np. polecenie dh lub na podstawie protokołów odbioru produktów projektu zaakceptowane przez Kierownika projektu. Częstotliwość pomiaru zgodnie z harmonogramem raportów okresowych - zgodnie z DIP;</p> <p>2 - suma ilości udostępnionych cyfrowych wersji zasobów z poziomu systemu Sposób pomiaru: suma dokumentów udostępnionych w systemie lub suma ilości dokumentów przeznaczonych do udostępnienia na podstawie protokołów odbioru produktów zaakceptowanych przez kierownika projektu.</p>

2.2. Udostępnione e-usługi

Lp.	Nazwa e-usługi	Typ	Zakres oddziaływania	Poziom dojrzałości e-usługi

2.3. Udostępnione informacje sektora publicznego i zdigitalizowane zasoby

Rodzaj informacji/zasobów	Planowana data udostępnienia	Szacowana liczba obiektów objętych digitalizacją (udostępnianiem informacji)
<p>Kolekcja materiału biologicznego Generatio-A jest to kolekcja zebrana przy okazji realizowanego badania populacyjnego mieszkańców miasta Aleksandrów Łódzki. Kolekcja jest jedną z nielicznych w Polsce kolekcji rodzinnych. Dzięki czemu możliwe jest prowadzenie badań nad dziedziczeniem poszczególnych genów. Z uwagi na planowaną unikatową jeszcze w skali kraju i europy metodę digitalizacji zasobów Genom + Epigenom, możliwe będzie prowadzenie badań</p>	30-06-2028	400 Genomów (WGS)

Rodzaj informacji/zasobów	Planowana data udostępnienia	Szacowana liczba obiektów objętych digitalizacją (udostępnianiem informacji)
powiązanych z chorobami neurodegeneracyjnymi, badanie interakcji pomiędzy genomem i epigenomem, zasób będzie nieoceniony jako niezależny zestaw danych walidacyjnych dla innych badań naukowych.		
Kolekcja Materiału biologicznego POPULOUS. Digitalizacji podlegać będzie fragment jednej z większych kolekcji materiału biologicznego w kraju. Kolekcja odzwierciedla strukturę zasiedlenia kraju z uwzględnieniem podziału na województwa, tereny miejskie, wiejskie i duże aglomeracje.	30-06-2029	1000 Genomów (WGS)

Czy wszystkie zdigitalizowane zasoby objęte projektem będą udostępniane bezpłatnie?
TAK/NIE

2.4. Produkty końcowe projektu

Nazwa produktu	Planowana data wdrożenia
Rozproszona infrastruktura Polskiego Węzła Europejskiego Archiwum Genomów	12-2029
Zdigitalizowane zasoby WGS + Epigenom (400 szt.)	06-2028
Zdigitalizowane zasoby WGS (1000 szt.)	06-2029

3. KAMIENIE MIŁOWE

Kamienie milowe	Planowany termin osiągnięcia
Opracowanie koncepcji architektury rozproszonego krajowego węzła FEGA wraz z analizą wymagań technologicznych, bezpieczeństwa i interoperacyjności oraz przygotowaniem infrastruktury pod wdrożenie w PCSS i Cyfronet	2027-03-31

Kamienie milowe	Planowany termin osiągnięcia
Wdrożenie i uruchomienie preprodukcyjnych instancji środowisk usługowych i narzędziowych LocalEGA w PCSS i Cyfronet, obejmujących środowiska testowe oraz komponenty niezbędne do walidacji procesów deponowania, przechowywania, katalogowania i kontrolowanego udostępniania danych	2027-09-30
Produkcyjne uruchomienie platform LocalEGA w PCSS i Cyfronet, poprzedzone testami funkcjonalnymi, integracyjnymi, wydajnościowymi i bezpieczeństwa oraz przygotowaniem dokumentacji technicznej i procedur operacyjnych	2027-12-31
Synchronizacja pierwotnych zasobów danych krajowego węzła FEGA do środowisk w PCSS i Cyfronet, obejmująca walidację kompletności, integralności i zgodności metadanych	2028-03-31
Wdrożenie i walidacja mechanizmów automatycznej replikacji danych, wysokiej dostępności oraz procedur awaryjnego przełączania usług pomiędzy ośrodkami PCSS i Cyfronet	2028-09-30
Wyłoniony dostawca na odczytniki do digitalizacji zasobów	2027-06-30
Zdigitalizowane próbki kolekcji GEN-A (WGS + Epigenom)	2028-06-30
Zdigitalizowane próbki kolekcji POPULOUS (WGS)	2029-06-30

4. KOSZTY

4.1. Koszty ogólne projektu wraz ze sposobem finansowania

Całkowity koszt projektu (netto oraz brutto), w tym	Netto 34 917 887,52 zł Brutto 37 019 980,00 zł	
Procent dofinansowania ze środków UE (brutto)	79,71%	
Procent środków z budżetu państwa (brutto)	20,29%	
Podział całkowitego kosztu projektu na poszczególne lata (netto oraz brutto)	2027	Netto 17 547 529,20 zł Brutto 19 313 821,68 zł
	2028	Netto 11 034 648,32 zł Brutto 11 324 448,32 zł
	2029	Netto 6 335 710,00 zł Brutto 6 381 710,00 zł

4.2. Wykaz poszczególnych pozycji kosztowych

Nazwa pozycji kosztowej		Przewidywany koszt brutto	Uzasadnienie pozycji kosztowej (przeznaczenie)
Oprogramowanie	System informatyczny przeznaczony do katalogowania i zarządzania analogowymi kolekcjami przeznaczonymi do digitalizacji, licencje oprogramowania do digitalizacji genomów	1 242 300,00 zł	Systemy niezbędne do udostępniania danych zostały zaimplementowane w trakcie realizacji poprzedniego projektu w ramach Programu Operacyjnego Polska Cyfrowa. W niniejszym projekcie planowany jest zakup systemu wspierającego katalogowanie i zarządzanie zasobami przeznaczonymi do digitalizacji. Dzięki m.in. implementacji elementów AR wspierających pracę personelu system usprawni proces digitalizacji i zarządzania zasobami - Szacowany koszt: 996300 Oprogramowanie Dragen jest wykorzystywane w procesie digitalizacji przez Beneficjenta - szacowany koszt: 246 000
Infrastruktura	Zakup serwera realizującego backup w lokalne przestrzeni składowania danych	492 000,00 zł	Od momentu zsekwencjonowania danych do wytworzenia gotowego do udostępniania produktu cyfrowego/dokumentu niezbędne jest wykonanie ciągu analiz. Dodatkowo w docelowym repozytorium umieszczane będą kompletne paczki danych. DO czasu ulokowanie danych w docelowym repozytorium należy je zabezpieczyć w lokalnej infrastrukturze. Wycena zawiera rezerwę z uwagi na aktualną wzrostą dynamikę cen infrastruktury IT
Koszty UX i grafiki			
Bezpieczeństwo	Kontrakty serwisowe na urządzenia wykorzystywane w procesie digitalizacji	2 282 520,00 zł	Pozyskanie pełnogenomowej sekwencji DNA jest procesem kosztownym i czasochłonnym. Beneficjent i Partnerzy minimalizują ryzyko w sposób wcześniej sprawdzony. Mianowicie poprzez zakup kontraktu serwisowego. Takie rozwiązanie przenosi znaczną część ryzyka na dostawcę technologii. W ramach takiego zabezpieczenia dostawca nie tylko odpowiedzialny jest za

Nazwa pozycji kosztowej		Przewidywany koszt brutto	Uzasadnienie pozycji kosztowej (przeznaczenie)
			naprawę urządzenia, ale również za zwrot materiałów zużywalnych i odczynników jeśli sekwencjonowanie nie powiodło się w wyniku awarii.
Wydajność rozwiązań			
Szkolenia			
Działania informacyjno-promocyjne	Działania pinformacyjno-promocyjne	615 000,00 zł	Przekazanie interesariuszom informacji o funkcjonowaniu i zasadach korzystania z Repozytorium jest kluczowa dla korzystania z zasobu przez grupy odbiorców. W ramach dziania przewidziano m.in. udział w konferencjach naukowych i branżowych w celu prezentacji funkcjonalności, organizację hackathonów z wykorzystaniem dostępnych danych, produkcję materiałów informacyjnych i szkoleniowych
Koszty zarządzania i wsparcia (w tym wynagrodzenia personelu wspomagającego)	Pozycja zawiera koszty personelu, materiałów zużywalnych i odczynników, koszty aparatury niezbędnej do digitalizacji, koszty usług obliczeniowych realizowanych w środowisku chmurowym dostawcy technologii oraz koszty pośrednie	32 388 160,00 zł	<p>Materiały zużywalne i odczynniki są niezbędne do przeprowadzenia procesu digitalizacji - szacunkowy koszt: 6 157 305,00</p> <p>Sprzęt laboratoryjny niezbędny w procesie digitalizacji - szacunkowy koszt: 430 500,00</p> <p>usługi obliczeniowe realizowane w środowisku chmurowym dostawcy technologii - szacunkowy koszt: 22 000</p> <p>koszt wynagrodzeń - szacunkowy koszt: 22 415 961,00</p> <p>koszty pośrednie - szacunkowy koszt: 3 362 394,00</p> <p>Dużą część budżetu stanowią koszty personelu, jest to wynikiem niskich kosztów infrastruktury, która została przygotowana w trakcie realizacji wcześniejszego projektu. Zostanie ona teraz rozproszona pomiędzy akademickie centra komputerowe tak aby uzyskała większą niezawodność i</p>

Nazwa pozycji kosztowej		Przewidywany koszt brutto	Uzasadnienie pozycji kosztowej (przeznaczenie)
			skalowalność. Personel projektu składa się z ekspertów biotechnologicznych i informatycznych. Poziom wynagrodzeń starają się odzwierciedlać stawki rynkowe, aby wysokiej klasy specjaliści nie opuszczali środowiska akademickiego. Chodzi o zminimalizowanie ryzyka odejścia kluczowego personelu przed końcem realizacji projektu i zapewnienie sprawnej jego realizacji. Z drugiej strony jest zgodne z Polityką Naukową Państwa - ochrona środowiska naukowego przed drenażem mózgów.

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

Całkowity koszt utrzymania trwałości projektu (brutto)	1 936 000,00 zł		Źródło finansowania
Podział całkowitego kosztu utrzymania trwałości projektu na poszczególne lata (netto oraz brutto)	2030	372 000,00 zł (brutto) (372 000,00 zł netto)	krajowe środki publiczne - budżet państwa
	2031	372 000,00 zł (brutto) (372 000,00 zł netto)	krajowe środki publiczne - budżet państwa
	2032	372 000,00 zł (brutto) (372 000,00 zł netto)	krajowe środki publiczne - budżet państwa
	2033	410 000,00 zł (brutto) (410 000,00 zł netto)	krajowe środki publiczne - budżet państwa
	2034	410 000,00 zł (brutto) (410 000,00 zł netto)	krajowe środki publiczne - budżet państwa

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 zarządzania ryzykiem
Nieuwzględnienie w analizie istotnych wymagań funkcjonalnych	Średnia	Niskie	Unikanie ryzyka - Przegląd wymagań, analiza dostępnych technologii.
Niedoszacowanie trudności realizacji poszczególnych funkcjonalności	Średnia	Niskie	Unikanie ryzyka - Wstępne oszacowanie pracochłonności pod kątem kosztów i czasochłonności w oparciu o analizę obecnie dostępnych technologii i rozwiązań. Ciągłe monitorowanie postępów i dokonywanie korekt w planach rozwoju.
Poważna zmiana w dostępnych technologiach	Duża	Niskie	Redukowanie ryzyka - Śledzenie tendencji rozwoju technologii i standardów.
Problemy integracyjne z istniejącym środowiskiem infrastrukturalnym	Duża	Średnie	Unikanie ryzyka - Wczesne testy integracyjne, przygotowanie środowiska testowego oraz stała współpraca zespołów odpowiedzialnych za utrzymanie infrastruktury.
Problemy z integracją mechanizmów replikacji danych pomiędzy partnerami infrastrukturalnym i	Duża	Średnie	Unikanie ryzyka - Wspólne uzgodnienie architektury replikacji w oparciu o standardy i dobre praktyki oraz przeprowadzenie testów interoperacyjności.
Opóźnienie łańcucha dostaw w wyniku niestabilnej sytuacji geopolitycznej	Średnia	Średnie	Redukowanie ryzyka - Zawarcie długofalowych umów na dostawy (sukcesywna dostawa), przeniesienie części ryzyka na dostawcę
Z uwagi na	Mała	Średnie	Redukowanie ryzyka – przy zakupie

Nazwa ryzyka	Siła oddziaływania	Prawdopodobieństwo wystąpienia ryzyka	Sposób zarządzania ryzykiem
trwający 36 miesięcy okres realizacji projektu mogą ulec zmianie ceny usług, środków trwałych i materiałów niezbędnych do realizacji projektu.			materiałów i usług strategicznych z punktu widzenia realizacji projektu Benefcjent stosuje umowy gwarantujące stałą cenę dla sukcesywnych dostaw lub dokonuje jednorazowego zakupu.
Ryzyka kursowe USD/PLN - ceny odczynników i kontraktów serwisowych wyrażone są w USD.	Duża	Średnie	Budżet projektu planowany jest z marginesem błędu, kontrakty z dostawcami zawierane będą w PLN - przeniesienie części ryzyka na dostawców, w wycenach starano się zachować odpowiednie marginesy

5.2. Ryzyka wpływające na utrzymanie efektów

Nazwa ryzyka	Siła oddziaływania	Prawdopodobieństwo wystąpienia ryzyka	Sposób zarządzania ryzykiem
Zmiana przepisów prawa uniemożliwiająca udostępnianie danych genomicznych na obecnych zasadach	Duża	Niskie	Akceptacja ryzyka - Obserwacja aktualnego otoczenia prawnego, konsultowanie aktów prawnych w procesie legislacyjnym na etapie konsultacji społecznych
Problemy z kompatybilnością rozwiązań po aktualizacjach istniejącej infrastruktury lub infrastruktury partnerskiej	Średnia	Średnie	Unikanie ryzyka - Stosowanie otwartych standardów i kompatybilnych technologii, wersjonowania komponentów oraz przeprowadzanie testów kompatybilności przed wdrożeniem aktualizacji w środowisku produkcyjnym.
Przerwy w dostępności usług wynikające z awarii infrastruktury lub	Duża	Znikome	Redukcja ryzyka - Wdrożenie mechanizmów redundancji i monitoringu usług, regularne testy procedur odtworzeniowych oraz utrzymywanie kopii zapasowych. Ponadto każdy z

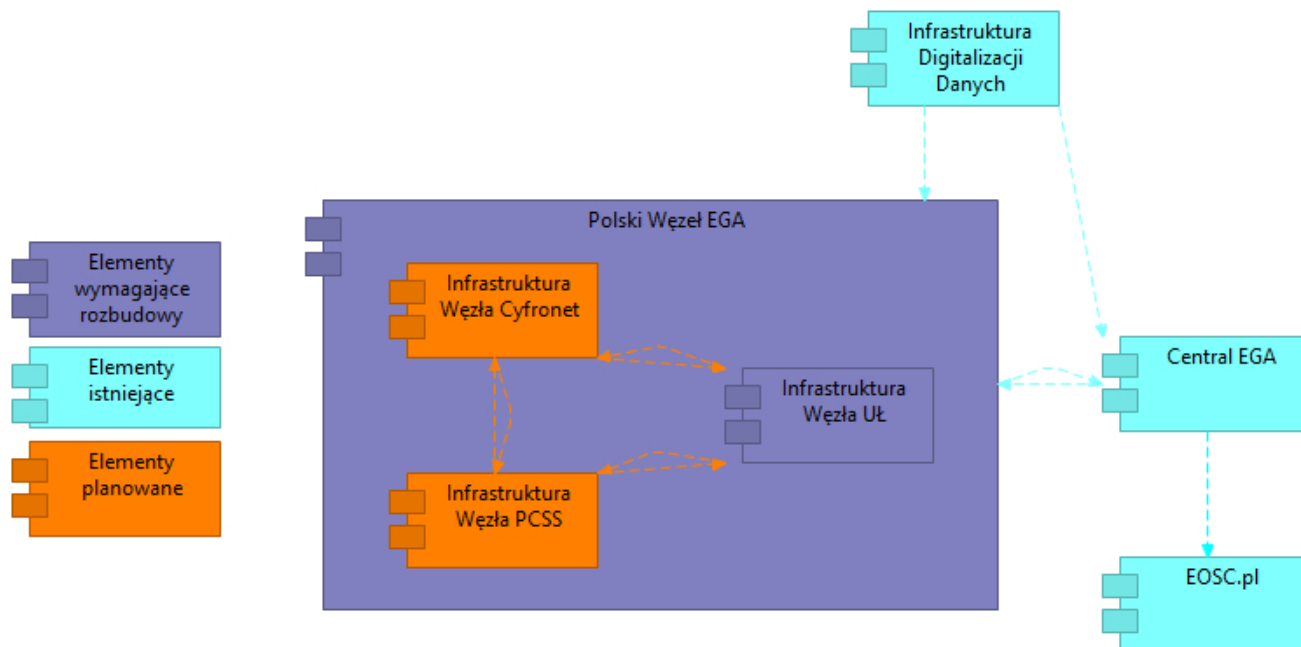
Nazwa ryzyka	Siła oddziaływania	Prawdopodobieństwo wystąpienia ryzyka	Sposób zarządzania ryzykiem
usług zależnych			Partnerów implementuje redundancje na poziomie własnego centrum danych. Właśnie ograniczenie tego ryzyka spowodowało podjęcie realizacji tego projektu.
Utrata synchronizacji danych pomiędzy węzłami redundantnymi	Duża	Niskie	Unikanie ryzyka - Stałe monitorowanie procesów replikacji, automatyczne alertowanie o błędach synchronizacji oraz okresowe testy odtwarzania danych.
Niewystarczająca wydajność mechanizmów replikacji przy rosnącym wolumenie danych	Średnia	Średnie	Redukcja ryzyka - Regularne monitorowanie wydajności mechanizmów replikacji, dostosowywanie konfiguracji synchronizacji danych do aktualnego obciążenia systemu.

6. OTOCZENIE PRAWNE

Lp.	Tytuł aktu prawnego	Czy wymaga zmian	Opis zmian (jeśli dotyczy)	Etap prac legislacyjnych (jeśli dotyczy)
1	Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych)	TAK/NIE		
2	Ustawa z dnia 5 lipca 2018 r. o krajowym systemie cyberbezpieczeństwa	TAK/NIE		
3	Ustawa z dnia 11 sierpnia 2021 r. o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego	TAK/NIE		
4	Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2025/327 z dnia 11 lutego 2025 r. w sprawie europejskiej przestrzeni danych dotyczących zdrowia oraz zmiany dyrektywy 2011/24/UE i rozporządzenia (UE) 2024/2847.	TAK/NIE		

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	Centralny Węzeł Europejskiego Archiwum Genomów	Europejski Instytut Bioinformatyki i Centrum Regulacji Genomicznych w Barcelonie	European Genome Archive – repozytorium danych genetycznych przechowujących i udostępniających dane dotyczące ludzi	Istniejący	
2	EOSC	EOSC Association / Polski Węzeł EOSC	Europejska chmura danych	Istniejący	
3	Polski Węzeł Europejskiego Archiwum Genomów	Uniwersytet Łódzki	Lokalne repozytorium danych genomicznych - danych wymagające szczególnej ochrony wg.	Modyfikowany	Istniejąca infrastruktura Polskiego Węzła

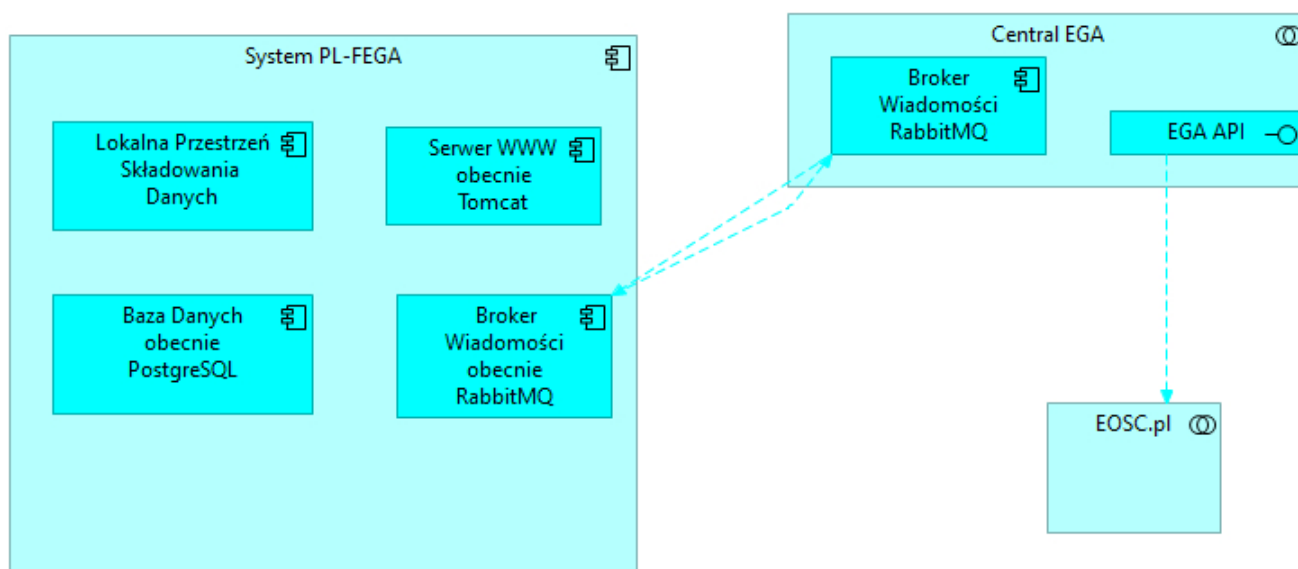
Lp.	Nazwa systemu	Gestor systemu	Opis systemu	Status	Krótki opis ewentualnej zmiany
			RODO. Dane lokalne dane przechowywane są w zgodzie z lokalną jurysdykcją.		zostanie rozproszona pomiędzy Partnerów projektu - powstaną trzy lokalizacje w których zostaną składowane zasoby niezbędne do działania systemu. Dzięki temu rozwiązaniu Polski Węzeł uzyska większą niezawodność i skalowalność.

Lista przepływów

Lp.	System źródłowy	System docelowy	Zakres wymienianych danych	Sposób wymiany danych	Typ modyfikacji	Typ interfejsu
1	Centralny Węzeł Europejskiego Archiwum Genomów	Polski Węzeł Europejskiego Archiwum Genomów	Po pozytywnej weryfikacji przez Komitet Dostępu do Danych żądania dostępu do danych, przekazywane jest żądanie nadania uprawnień do zbioru danych konkretnemu użytkownikowi	Tryb odwołań bezpośrednich	Brak	RabbitMQ
2	Polski Węzeł Europejskiego	Centralny Węzeł Europejskiego	Potwierdzenie poprawnego zdeponowania przez	Tryb Odwołań Bezpośrednich	Brak	RabbitMQ

Lp.	System źródłowy	System docelowy	Zakres wymienianych danych	Sposób wymiany danych	Typ modyfikacji	Typ interfejsu
	Archiwum Genomów	Archiwum Genomów	użytkownika danych w Polskim Węźle Europejskiego Archiwum Genomów			
3	Centralny Węzeł Europejskiego Archiwum Genomów	EOSC	Metadane opisujące zdeponowane w Polskim Węźle Europejskiego Archiwum Genomów zbiory danych	Kopiowanie	Brak	API

7.2. Kluczowe komponenty architektury rozwiązania



7.3. Przyjęte założenia technologiczne

Lp.	Obszar	Założenie technologiczne
1.	Infrastruktura	W projekcie zostanie wykorzystana w całości już istniejąca infrastruktura informatyczna FEGA. W ramach projektu istniejąca instancja Polskiego Węzła FEGA zostanie rozproszona pomiędzy

Lp.	Obszar	Założenie technologiczne
		trzy jednostki, zyskując tym samym większą odporność, skalowalność i niezawodność. Partnerzy projektu są jednostkami publicznymi PCSS i Cyfronet są dużymi publicznie dostępnymi Data Center realizujące krajowe projekty infrastruktury takich jak: KMD, Fabryki AI,
2.	Sieć i bezpieczeństwo	Komunikacja z planowanymi do wdrożenia aplikacjami webowymi odbywać się będzie z wykorzystaniem protokołu TLS, SSH lub SFTP/FTPS
3.	Standardy wymiany danych	Wymiana danych odbywać się będzie w oparciu o standardy wypracowane przez Europejskie Archiwum Genomów
4.	Systemy operacyjne serwerowe	Preferowane do stosowania przez Beneficjenta, Partnerów i twórców rozwiązania są systemy operacyjne klasy Linux
5.	Bazy danych	Obecnie stosowany system wykorzystuje silnik bazy danych PostgreSQL, Partnerzy projektu pragną pozostać przy rozwiązaniach open source.
6.	Serwery aplikacji	System FEAGA został wytworzony w języku JAVA i wykorzystuje framework Spring. Uruchamiany jest na serwerze Tomcat. Wykorzystane technologie są szeroko wykorzystywane w systemach klasy biznes zapewniając bezpieczeństwo danych oraz wydajność 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

Beneficjent jest operatorem Węzła EGA. System został opracowany przez Europejski Instytut Bioinformatyki jako rozwiązanie paneuropejskie.

Obecnie nie ma wymogów dotyczących zabezpieczania danych objętych projektem. Mechanizmy zabezpieczeń wynikają z wymogów RODO i zostały dobrane na podstawie analizy ryzyka zarówno na poziomie Węzła jak i Całego EGA. W chwili obecnej trwają prace nad spójną polityką bezpieczeństwa infrastruktury FEAGA jako całości. Kluczowe jest szybkie przywrócenie systemu po awarii/incydencie - zastosowano dwie redundantne geograficznie kopie. Dane wrażliwe przechowywane są w formie szyfrogramu. Komunikacja odbywa się poprzez zaszyfrowane łącza

a poszczególne serwisy muszą uwierzytelnić się certyfikatem. Wdrożone zostały mechanizmy bezpieczeństwa wypracowane przez Europejski Instytut Bioinformatyki. Bezpieczeństwo systemu potwierdzone zostało testami penetracyjnymi wykonanymi przez podmiot trzeci.

-dodatkowe zabezpieczenia powyżej wymogów KRI: należy wskazać uzasadnienie