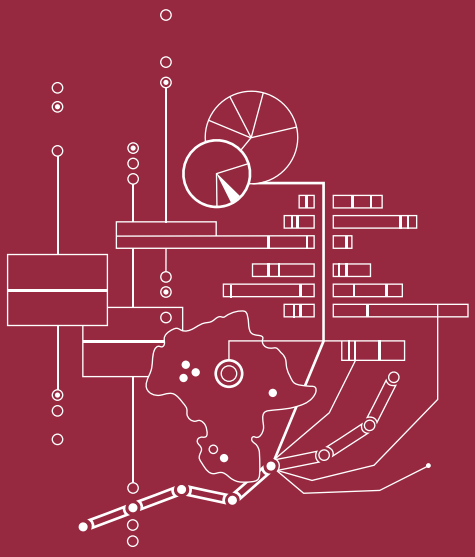
OŚRODEK
PRZETWARZANIA
INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

RAD-on

Raporty
Analizy
Dane



redakcja naukowa
Aldona Tomczyńska
Anna Knapińska
Sylwia Ostrowska

Systemy informatyczne
wspierające naukę i szkolnictwo wyższe

RAD-on:
Raporty, analizy, dane

Redakcja naukowa:
dr Aldona Tomczyńska
dr Anna Knapińska
Sylwia Ostrowska

OŚRODEK PRZETWARZANIA INFORMACJI
PAŃSTWOWY INSTYTUT BADAWCZY

Systemy informatyczne wspierające naukę i szkolnictwo wyższe
RAD-on: Raporty, analizy, dane

Redakcja naukowa:
dr Aldona Tomczyńska
dr Anna Knapińska
Sylwia Ostrowska

Redakcja techniczna:
Michał Tomaszewski

Autorzy:
Marcin Białas: rozdział 2, 5
Łukasz Błaszczyk: rozdział 4
dr inż. Sławomir Dadas: rozdział 3
dr Anna Knapińska: rozdział 2, 3
dr inż. Marcin Mirończuk: rozdział 5
Sylwia Ostrowska: rozdział 1, 2, 4
Emil Podwysocki: rozdział 4
dr Aldona Tomczyńska: rozdział 1, 3

Recenzent:
dr hab. Zenon A. Sosnowski, prof. Politechniki Białostockiej

Wydawca:
Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy
al. Niepodległości 188b
00-608 Warszawa
e-mail: opi@opi.org.pl
www.opi.org.pl

ISBN 978-83-63060-25-1

Projekt okładki i opracowanie graficzne:
Karina Maszewska

Skład i łamanie:
dr Jakub Kierzkowski

Druk:
Lenis Przemysław Dubrzyński
ul. Kosynierów 27, 04-641 Warszawa

Szanowni Państwo,

we współczesnym społeczeństwie, w którym informacja odgrywa bardzo ważną rolę, dostęp do danych ma kluczowe znaczenie dla rozwoju kraju. Ważne jednak, aby te dane były aktualne i przede wszystkim wiarygodne. Powinny być one także pozyskane z zaufanych źródeł i przetwarzane w odpowiedzialny sposób. Źródłem takich danych dla środowiska akademickiego i całej nauki polskiej jest portal RAD-on – prowadzony przez Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy (OPI) na zlecenie Ministerstwa Edukacji i Nauki.

Monografia, którą Państwo otrzymują, w szczegółowy sposób przedstawia serwis RAD-on. Zawiera również wiele ciekawych informacji dotyczących m.in. idei integracji i przetwarzania danych czy też narodowych systemów o nauce i szkolnictwie wyższym. To niezwykle interesująca pozycja wydawnicza zarówno dla całego sektora nauki, informatyków odpowiedzialnych za tworzenie nowoczesnych systemów informatycznych, jak również decydentów, którzy wdrażają decyzje polityczne oparte na wiarygodnych danych.

Zachęcam Państwa nie tylko do lektury monografii, ale także do korzystania z portalu RAD-on. Umożliwia on analizowanie trendów w polskim szkolnictwie wyższym i wgląd w dane dotyczące sektora nauki i szkolnictwa wyższego. To nowoczesna, stale aktualizowana baza wiedzy z dostępem do wielu przydatnych funkcji. Narzędzie opracowane przez ekspertów OPI zawiera nie tylko same dane, ale także dużą liczbę merytorycznych raportów i analiz.

Platforma oferuje kompleksowe rozwiązania informatyczne, dzięki czemu dostęp do danych jest prosty i szybki. Dane można filtrować na wiele sposobów, dostosowując przeszukiwanie do indywidualnych potrzeb. Warto dodać, że narzędzie OPI pozwala na bezpłatne pobranie i wykorzystywanie zasobów w aplikacjach, bazach danych, programach i usługach informatycznych. Ponadto portal posiada REST API, co umożliwia otwarty dostęp do danych.

Gratuluję ekspertom OPI przygotowania wartościowej monografii naukowej i dziękuję za wdrożenie oraz stałą modernizację portalu RAD-on.

Przemysław Czarnek

Minister Edukacji i Nauki

Szanowni Państwo,

żyjemy w czasach czwartej rewolucji przemysłowej, zwanej również rewolucją cyfrową. Informacja stała się bardzo cennym zasobem. To niesamowite, że cały „świat informacji" mieści się w dłoni – jest dostępny praktycznie dla każdego z poziomu małego urządzenia. Żyjąc w XXI wieku, możemy w łatwy sposób dotrzeć do wszelkich zasobów informacyjnych, które do tej pory zgromadzono.

W tworzeniu społeczeństwa informacyjnego, opartego na wiarygodnych źródłach wiedzy, szczególnie cenne są takie inicjatywy i projekty, jak portal RAD-on. To nowoczesne narzędzie IT jest wynikiem partnerstwa Ministerstwa Edukacji i Nauki z Ośrodkiem Przetwarzania Informacji – Państwowym Instytutem Badawczym (OPI PIB), które zainicjowało w ramach projektu „ZSUN II. Zintegrowany System Usług dla Nauki. Etap II". Portal RAD-on gwarantuje dostęp do rzetelnych informacji pochodzących nie tylko ze skarbnicy wiedzy, jaką niewątpliwie jest Zintegrowany System Informacji o Szkolnictwie Wyższym i Nauce POL-on, ale z wielu innych centralnych systemów z obszaru szkolnictwa wyższego i nauki. To nowoczesny portal umożliwiający wykonywanie różnych raportów i analiz, dostosowanych do potrzeb użytkownika. RAD-on efektywnie realizuje założenia polityki otwartej nauki, udostępniając jeden z największych zasobów danych o szkolnictwie wyższym i nauce w Europie (ok. 11,5 TB danych). Portal RAD-on służy nie tylko środowisku akademickiemu, studentom, przedstawicielom przemysłu, decydentom politycznym, ale także wszystkim, którzy chcą pozyskać kompleksowe, rzetelne i aktualne informacje o instytucjach tworzących system szkolnictwa wyższego i nauki – w tym, co warto podkreślić – prowadzonej przez nie działalności dydaktycznej i naukowej.

Niniejsza publikacja opisuje w szczegółowy sposób proces powstawania i doskonalenia portalu RAD-on. Prezentuje jego możliwości i potencjał do dalszego rozwoju. Jestem przekonany, że lektura monografii przekona Państwa nie tylko do korzystania z rzetelnej wiedzy dostępnej na portalu, ale także zachęci do współpracy i ciągłego rozwijania jego funkcjonalności.

Łukasz Wawer

Zastępca dyrektora Centrum Transformacji Cyfrowej

Ministerstwo Edukacji i Nauki

SPIS TREŚCI

Rozdział 1. Idea i cele RAD-onu **13**

dr Aldona Tomczyńska, Sylwia Ostrowska

1.1. O idei integracji i udostępniania danych ... 13
1.1.1. Decyzje polityczne na podstawie danych ... 13
1.1.2. Otwarte dane rządowe ... 14
1.1.3. Dane FAIR ... 15
1.2. Narodowe systemy informacji o nauce i szkolnictwie wyższym ... 16
1.3. Polityka naukowa a proces informatyzacji sektora nauki i szkolnictwa wyższego w Polsce ... 18
1.4. Potrzeby interesariuszy ... 20
1.5. Cele projektu ... 21
1.6. Dane udostępniane przez RAD-on ... 22
1.7. Realizacja projektu i badania użyteczności ... 23
1.8. Wskaźniki projektu i przykłady wykorzystania systemu RAD-on ... 24

Rozdział 2. Portal **27**

Sylwia Ostrowska, Marcin Białas, dr Anna Knapińska

2.1. Użytkownicy i usługi portalu ... 27
2.1.1. Dane ... 28
2.1.2. Raporty ... 31
2.1.3. Analizy ... 31
2.1.4. Wyszukiwarka ... 32
2.1.5. Konto użytkownika ... 33
2.2. Architektura portalu ... 37

Rozdział 3. Raporty **39**

dr Anna Knapińska, dr Aldona Tomczyńska, dr inż. Sławomir Dadas

3.1. Narzędzia BI a platformy analityczne ... 39
3.2. Zaplecze platformy analitycznej RAD-onu ... 41
3.2.1. Warstwy architektury systemu ... 42
3.2.2. Warstwa raportów ... 43

3.2.3. Rodzaje sekcji 44
3.3. Raporty w RAD-onie jako narzędzie rozpowszechniania informacji i podejmowania decyzji 45
3.3.1. Wyzwania w tworzeniu raportów 48

Rozdział 4. Hurtownia i model wymiany danych 51
Łukasz Błaszczyk, Sylwia Ostrowska, Emil Podwysocki

4.1. Model wymiany danych 51
4.1.1. Technologia 51
4.1.2. Założenia modelu wymiany danych 52
4.1.3. Zastosowanie modelu wymiany danych 55
4.2. Hurtownia danych 64
4.2.1. Architektura hurtowni danych 64
4.2.2. Wdrożenie narzędzia klasy BI 66
4.2.3. Zarządzenie danymi 67
4.2.4. Rola hurtowni danych w architekturze systemów informatycznych OPI PIB 69

Rozdział 5. API 71
Marcin Białas, dr inż. Marcin Mirończuk

5.1. Geneza API 71
5.2. Architektura systemu w ujęciu ogólnym 73
5.3. Architektura systemu w ujęciu szczegółowym 76
5.3.1. Zarządzanie strumieniami 77
5.3.2. Menadżer Zarządzania Indeksami (MZI) 79
5.3.3. Obsługa zapytań z API 81
5.3.4. System logowania zdarzeń 82
5.3.5. System prezentacji danych 83
5.4. Technologie 84

Spis ilustracji 87

Bibliografia 89

WSTĘP

Serwis RAD-on jest częścią systemu informatycznego prezentującego raporty, analizy oraz dane o nauce i szkolnictwie wyższym w Polsce. Został opracowany przez Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy (OPI PIB) we współpracy z Ministerstwem Edukacji i Nauki (MEiN) w ramach projektu Zintegrowanej Sieci Informacji o Nauce i Szkolnictwie Wyższym.

Jest to największy pod względem zakresu zbieranych i udostępnianych danych narodowy system informatyczny z obszaru nauki i szkolnictwa wyższego w Europie. RAD-on zawiera informacje o niemal wszystkich instytucjach naukowych w Polsce i zatrudnionych w nich naukowcach, w tym nauczycielach akademickich i innych osobach prowadzących zajęcia na uczelniach. Dodatkowo RAD-on udostępnia szczegółowe wykazy publikacji naukowych, patentów krajowych i zagranicznych, projektów badawczych finansowanych z różnych źródeł, a także informacje o inwestycjach instytucji. Wreszcie, w RAD-onie znajdują się obszerne informacje dotyczące studentów i absolwentów studiów, w tym o kierunkach studiów, kształceniu cudzoziemców, doktorantów i możliwych formach odbywania studiów.

Najważniejsze usługi systemu obejmują:

- bazę wiedzy, na którą składają się aktualne i oficjalne dane o instytucjach naukowych, działalności naukowej i artystycznej, pracownikach akademickich oraz postępowaniach awansowych w nauce;
- interaktywne, przekrojowe raporty o uczelniach, studentach, absolwentach i kadrze akademickiej, a także badaniach prowadzonych w Polsce i na świecie;
- repozytorium rzetelnych opracowań zagadnień z zakresu nauki i szkolnictwa wyższego, zawierających rozbudowane komentarze;
- interfejs programowania aplikacji (*Application Programming Interface*, API), który umożliwia wolny dostęp do zasobów RAD-onu. Dzięki API użytkownicy mogą budować własne rozwiązania i aplikacje wymagające dostępu do danych z zakresu szkolnictwa wyższego;
- konto użytkownika służące do przeglądania danych gromadzonych w systemie RAD-on i jego systemach źródłowych. Zalogowani użytkownicy mogą aktualizować i weryfikować dane o sobie, co zapewnia wysoką jakość przedstawianych w RAD-onie informacji.

Większość wymienionych usług dostępna jest dla wszystkich za pośrednictwem serwisu internetowego RAD-on, a niewielka ich część tylko dla użytkowników zweryfikowanych. System jest nadal rozwijany, co oznacza, że zakres udostępnianych w nim danych będzie się systematycznie zwiększał, a użyteczność poszczególnych usług będzie

rosła. Niniejsza monografia prezentuje funkcjonalności tego systemu, jego analityczne i informatyczne podstawy, a także praktyczne zastosowania.

Celem monografii jest ukazanie systemu RAD-on jako unikatowego na skalę europejską rozwiązania, wpisującego się w trzy ważne trendy polityki naukowej w Polsce i na świecie.

Pierwszym z nich jest idea otwartych danych rządowych (*open government data*, OGD). Jest ona osadzona w dyskursie o konieczności zapewnienia transparentności działania organów publicznych. Udostępnianie danych wszystkim obywatelom na równych zasadach jest jedną ze strategii zwiększania zaangażowania społecznego. Zapewniając narzędzia służące gromadzeniu, udostępnianiu i analizie danych pochodzących z oficjalnych źródeł, rządy państw przyczyniają się do tworzenia społeczeństwa obywatelskiego oraz promują innowacyjność [61, 39].

Drugim trendem, w który wpisuje się system RAD-on, jest wspieranie procesów podejmowania decyzji politycznych na podstawie danych (*data-driven policy making*). Zadaniem decydentów politycznych jest rozwiązywanie problemów społecznych i ekonomicznych poprzez formułowanie i wdrażanie przepisów, zasad i wytycznych. Częścią procesu politycznego jest, oprócz tworzenia, także monitorowanie tych polityk. Cykl polityczny składa się z wielu różnych faz, takich jak ustalanie agendy, formułowanie polityki, podejmowanie decyzji, wdrażanie i ocena. Systemy informatyczne mogą być wykorzystywane na każdym z tych etapów, między innymi poprzez gromadzenie danych, usprawnianie ich przepływu, wreszcie przetwarzanie danych w informację i prognozowanie zjawisk [57].

Trzecią ideą wpisaną w system RAD-on jest projektowanie systemów udostępniania danych w taki sposób, by dane były łatwe do znalezienia przez wszystkich, interoperacyjne i wielokrotnego użytku (*findable, accessible, interoperable, reusable*, FAIR). Idea FAIR odnosi się przede wszystkim do zbiorów danych naukowych, wykorzystywanych w badaniach. RAD-on nie stanowi repozytorium tego typu danych, ale jego zasoby służyć mogą prowadzeniu analiz szkolnictwa wyższego i nauki. W tym aspekcie spełniają założenia FAIR [27, 62].

Trzy wymienione powyżej koncepcje zostały przez nas wzięte pod uwagę w procesie projektowania systemu RAD-on, dlatego opowiemy o nich szerzej już w pierwszym rozdziale monografii. Przybliży on politykę naukową Polski w odniesieniu do gromadzenia danych na temat aktywności wszystkich studentów i naukowców zatrudnionych w instytucjach nauki. Przedstawienie uwarunkowań systemowych pozwoli wykazać, że powstanie RAD-onu jest naturalną konsekwencją prowadzenia polityki opartej na danych w obszarze nauki. W rozdziale tym porównamy także nasz serwis z innymi systemami informatycznymi o podobnych funkcjach w Europie. Ponieważ zależy nam, aby monografia miała także zastosowanie praktyczne, podzielimy się z Czytelnikami naszymi doświadczeniami projektowymi, w tym wyzwaniami, z jakimi zmierzyliśmy się w trakcie realizacji tego przedsięwzięcia.

W drugim rozdziale skupimy się na usługach RAD-onu udostępnianych za pośrednictwem otwartego portalu obywatelskiego, który stanowi główny rezultat projektu. Opiszemy funkcjonalności portalu, które nie tylko ułatwiają dostęp do danych, ale także zapewniają ich aktualność i kompleksowość, takie jak usługa „Dane obywatela".

Trzeci rozdział przedstawi narzędzia zaimplementowane w systemie RAD-on, które zastępują komercyjne systemy analityki biznesowej (*business intelligence*, BI) i wspierają procesy przekształcania danych w informacje i wiedzę. Pokażemy zbudowany przez nas system do tworzenia raportów przyjaznych użytkownikom. Rozdział ten ma na celu pokazanie dylematów, jakie stoją przed analitykami danych opracowującymi statystyki na potrzeby niehomogenicznej grupy odbiorców.

RAD-on to przede wszystkim dane. O skali naszych działań związanych z zarządzaniem danymi i budową modelu ich wymiany opowiemy w rozdziale czwartym. W ramach projektu powstała hurtownia danych, integrująca niemal wszystkie systemy informatyczne stworzone w Ośrodku Przetwarzania Informacji – Państwowym Instytucie Badawczym w ciągu ostatniej dekady. Opiszemy problemy związane z ujednoliceniem formatu danych i nasze strategie ich rozwiązywania.

Rozdział piąty w całości poświęcimy interfejsowi programowania aplikacji, czyli usłudze, która jest niezbędnym elementem współczesnych systemów informatycznych. Jej adresatem są programiści zatrudniani przez instytucje naukowe i inne podmioty, zainteresowane maszynowym przetwarzaniem udostępnianych w systemie danych.

Liczymy, że tak opracowane zagadnienia spotkają się z zainteresowaniem Czytelników, którzy zaangażowani są w procesy wytwórcze oprogramowania wspierającego gromadzenie i przetwarzanie danych na zlecenie lub z udziałem podmiotów publicznych. Uważamy, że zawarta w tym opracowaniu wiedza pozwoli także zrozumieć specyfikę danych z zakresu szkolnictwa wyższego i nauki. Publikacja ta jest także zaproszeniem do dyskusji o przyszłości systemu RAD-on. Liczymy, że upowszechnienie się wiedzy o tym systemie zachęci wszystkich interesariuszy projektu do współtworzenia portalu i zwiększania jego użyteczności.

ROZDZIAŁ 1

IDEA I CELE RAD-ONU

dr Aldona Tomczyńska
Sylwia Ostrowska

1.1. O idei integracji i udostępniania danych

Jak zapowiedzieliśmy we wstępie, celem monografii jest ukazanie systemu RAD-on jako unikatowego na skalę europejską rozwiązania, wpisującego się w trzy trendy polityki naukowej w Polsce i na świecie. W tym miejscu opiszemy te idee szerzej.

1.1.1. Decyzje polityczne na podstawie danych

Państwa Zachodu, w tym członkowie Unii Europejskiej, zmierzają do możliwie szerokiego wykorzystania danych w procesach decyzyjnych. Trend ten nie jest nowy. Jego początków upatrywać można w latach dziewięćdziesiątych XX wieku [51].

W 1994 roku wydana została książka *The new production of knowledge* [22], poruszająca temat interakcji naukowców z obywatelami. Odbiła się ona szerokim echem w kręgach akademickich i pobudziła dyskusję o zmianach, jakie dotykają naukę pod wpływem procesów zachodzących w społeczeństwach Zachodu. W 2001 roku ukazała się publikacja będąca swoistą kontynuacją tej z 1994 roku – *Re-thinking science: Knowledge and the public in an age of uncertainty* [38]. Autorzy przyglądali się w niej roli społeczeństwa w procesie tworzenia wiedzy. Ich zdaniem nie tylko naukowcy, ale także obywatele zaangażowani są w proces badawczy, stawiając pytania i oczekując rzetelnych odpowiedzi opartych na danych. Tym samym, granice pomiędzy nauką a gospodarką i polityką, dotychczas wyraźnie wytyczone, zacierają się.

Te teoretyczne rozważania na temat roli nauki można powiązać z dyskursem o zmianach, jakie przynoszą kolejne rewolucje technologiczne. Zainteresowanie wiedzą płynącą z danych rośnie, w miarę jak zmniejszają się koszty i czasochłonność procesów analitycznych. Rozwój polityki opartej na danych można zatem odnieść do trzech zjawisk, które obserwujemy [45]:

- postępu technologii, w tym infrastruktury i oprogramowania służącego przetwarzaniu dużych zbiorów danych (*big data*). Siłą napędową tego rozwoju jest przede

wszystkim sektor biznesu, który docenia komercyjny potencjał analizy danych pochodzących z wielu rozległych źródeł;
- pojawienia się nowych metod analizy danych. Obszar zaawansowanej analityki danych (*data science*) obejmuje statystykę i wiedzę z obszaru informatyki, pozwalającą na uczenie maszynowe, w tym przetwarzanie języka naturalnego, obrazów, materiałów audio i wideo;
- większego zainteresowania narzędziami informatycznymi wśród osób niebędących ekspertami, w tym wzrostu poziomu kompetencji cyfrowych w społeczeństwach państw rozwiniętych gospodarczo;
- zrozumienia znaczenia badań naukowych i korzyści z zastosowania wiedzy naukowej w życiu codziennym.

1.1.2. Otwarte dane rządowe

Wymienione trendy stopniowo zwiększają zainteresowanie narzędziami zapewniającymi dostęp do danych nieobjętych ograniczeniami prywatności czy klauzulą poufności [48]. Dobrym przykładem na to jest popularność serwisu Google Trends, dającego wgląd w dane o frazach najczęściej wyszukiwanych w sieci internetowej[1]. Jest on wykorzystywany już nie tylko jako źródło wiedzy dla specjalistów od marketingu internetowego, ale także badaczy, w tym ekonomistów oraz socjologów, i to mimo uzasadnionych wątpliwości co do jakości tych danych [11].

Na tle danych zbieranych przez sektor biznesu, wyróżniają się dane udostępniane przez organy państwowe, pochodzące z oficjalnych rejestrów i sprawozdań. Otwarte dane rządowe (*open government data*, OGD) to wiarygodne źródła informacji nieodpłatnie udostępniane przez administrację publiczną lub podmioty jej podległe. Trend otwartości w odniesieniu do oficjalnych danych trwa w Unii Europejskiej i Stanach Zjednoczonych Ameryki już od ponad dekady [20]. Wiele państw zdecydowało się na powołanie organów odpowiedzialnych za cyfryzację różnych usług publicznych, dzięki czemu stopniowo powstają coraz bardziej złożone i wzajemnie powiązane systemy informatyczne. W Polsce funkcjonuje strona internetowa Otwarte Dane[2], zapewniająca dostęp do informacji pochodzących głównie z organów administracji rządowej i samorządowej. Podobne serwisy funkcjonują w innych państwach UE.

W miarę postępu prac nad OGD, pojawiają się nowe wyzwania. Na obecnym etapie rozwoju, konieczne jest zapewnianie interoperacyjności baz danych. Jak pokazują badania [17], to, jak chętnie OGD wykorzystywane są w procesach decyzyjnych, jest ściśle powiązane ze sposobem udostępniania danych. Użytkownicy systemów informatycznych niechętnie wykorzystują dane w postaci surowej [60], gdyż ich czyszczenie wymaga kompetencji analitycznych. W związku z tym istotnym elementem tworzenia infrastruktury OGD jest przetwarzanie informacji. Dodatkowo, jak podkreślają

[1] trends.google.com (dostęp: 02.03.2023)
[2] dane.gov.pl (dostęp: 02.03.2023)

Patricia Huston, Victoria L. Edge i Erica Bernie [29], największą korzyść przynosi łączenie danych pochodzących z wielu źródeł, wchodzących w skład systemu OGD. Innymi słowy, najlepszy system to taki, który w miejsce surowych, obszernych zbiorów danych oferuje dostęp do powiązanych ze sobą i odpowiednio przetworzonych informacji. Tylko takie OGD nadają się do interpretacji i bezpośredniego wykorzystania w procesach analitycznych.

1.1.3. Dane FAIR

Aby zaoferować użytkownikom kontekstową, dogłębną analizę danych, warto zatroszczyć się o spełnienie założeń standardu FAIR. Został on po raz pierwszy przedstawiony w artykule *The FAIR Guiding Principles for scientific data management and stewardship* [62] w czasopiśmie „Nature" w 2016 roku i od tej pory rozwijany jest między innymi w ramach Europejskiej Przestrzeni Badawczej.

Zgodnie z tym standardem dane powinny być łatwe do znalezienia i dostępne dla wszystkich, interoperacyjne i wielokrotnego użytku (*findable, accessible, interoperable, reusable*, FAIR). Idea FAIR wpisuje się w szerszy ruch otwartej nauki i kładzie szczególny nacisk na repozytoria danych służących badaniom naukowym, takich jak wyniki eksperymentów. Do idei FAIR można, jak przyznają sami autorzy, podejść szerzej i objąć nią wszystkie zdigitalizowane zasoby badawcze – od danych po procesy analityczne. Warto udostępniać informacje zebrane na wszystkich etapach procesu badawczego, aby zapewnić przejrzystość, odtwarzalność i replikację wyników badań. Zmiany w zakresie przechowywania danych powinny przynieść korzyści zarówno ludziom, jak i wspomagającym ich pracę maszynom. Z tej przyczyny w ramach FAIR szczególny nacisk kładziony jest na tworzenie metadanych zbiorów oraz instrukcji ich przetwarzania, które są zrozumiałe dla człowieka i programów komputerowych.

W obszarze nauki i szkolnictwa wyższego idea FAIR jest wyrażana w standardzie CERIF (*Common European Research Information Format*). Ten rekomendowany przez Unię Europejską model służy przechowywaniu informacji o działalności naukowej oraz pozwala na swobodną wymianę danych pomiędzy systemami informacji naukowej. CERIF jest rozwijany przez Międzynarodową Organizację do spraw Informacji Naukowej EuroCRIS[3].

[3] eurocris.org (dostęp: 02.03.2023)

1.2. Narodowe systemy informacji o nauce i szkolnictwie wyższym

Zarówno w Polsce, jak i za granicą coraz powszechniejsze staje się wykorzystanie systemów informatycznych do organizacji danych gromadzonych przez instytucje naukowe. Oprócz systemów zarządzania danymi o studentach (*student management systems*, SMS), popularne są także systemy informacji naukowej (*research information systems*, RIS). Dostęp do niektórych z nich jest otwarty, czego przykładem może być repozytoryjny system DSpace-CRIS[4], opracowany w Stanach Zjednoczonych Ameryki przez Massachusetts Institute of Technology oraz firmę Hewlett-Packard. W Polsce w 2012 roku na Politechnice Warszawskiej rozpoczęły się prace nad systemem Omega-PSIR, który również pełni rolę repozytorium oraz zapewnia funkcjonalności wspomagające pracę władz uczelni i pracowników naukowych. Licencja wraz z otwartym kodem źródłowym jest udostępniana bezpłatnie, dzięki czemu w 2022 roku z systemu korzystało prawie czterdzieści instytucji naukowych, w tym uczelni i instytutów badawczych[5] .

O ile rozwiązania takie są stosunkowo popularne na poziomie poszczególnych uczelni lub ich stowarzyszeń, o tyle wciąż niewiele państw posiada systemy SMS i RIS na poziomie centralnym. W zakresie danych o studentach i pracownikach sektora badań i rozwoju najpopularniejsze systemy tworzą organizacje międzynarodowe: Bank Światowy, Organizacja Narodów Zjednoczonych do spraw Oświaty, Nauki i Kultury (UNESCO), Organizacja Współpracy Gospodarczej i Rozwoju (OECD) czy też Unia Europejska (za pośrednictwem serwisu Eurostat, czyli Europejskiego Urzędu Statystycznego). Dane w nich zawarte dotyczą podstawowych wskaźników, takich jak liczba studentów czy pracowników sektora B+R w poszczególnych państwach członkowskich i pochodzą z ich urzędów statystycznych. Dane zagregowane są najczęściej na poziomie kraju, przez co szczegółowe informacje o systemie szkolnictwa wyższego nie są w nich dostępne. Od 2014 roku tworzony jest Europejski Rejestr Szkolnictwa Wyższego (European Tertiary Education Register, ETER), którego celem jest dostarczenie danych o instytucjach szkolnictwa wyższego funkcjonujących w państwach Europy. Zawiera on dane o studentach, kadrze akademickiej i finansach uczelni. Dane zbierane są za pośrednictwem urzędów statystycznych i ministerstw w poszczególnych państwach. W procesie tym nie są wykorzystywane systemy informatyczne. Rejestr ETER posiada wiele braków danych, szczególnie w odniesieniu do mniej popularnych wskaźników[6].

Większej kompletności danych o sytuacji poszczególnych instytucji naukowych można się spodziewać w centralnych systemach tworzonych przez poszczególne państwa. W zakresie danych o instytucjach i studentach na wyróżnienie zasługuje Agencja Statystyczna Szkolnictwa Wyższego w Wielkiej Brytanii (Higher Education Statistics Agency, HESA). Udostępniany przez nią serwis zapewnia otwarty dostęp do danych o szkolnictwie wyższym w Walii, Anglii, Szkocji i Irlandii Północnej, w tym dane

[4] 4science.com/dspace-cris (dostęp: 02.03.2023)
[5] omegapsir.io/pl (dostęp: 02.03.2023)
[6] eter-project.com (dostęp: 02.03.2023)

dotyczące studentów, pracowników i absolwentów [10]. System jest stale modernizowany i obecnie oferuje nie tylko dostęp do danych, ale także systemy analityki biznesowej umożliwiające ich filtrowanie i wizualizację. System nie zawiera danych z obszaru nauki i usług maszynowego przetwarzania[7]. Wybrane informacje o sektorze szkolnictwa wyższego publikuje w ten sposób również austriackie Ministerstwo Edukacji, Nauki i Badań Austrii (Bundesministerium für Bildung, Wissenschaft und Forschung, BMBWF) czy Nordycki Instytut Badań nad Innowacjami, Badaniami i Edukacją z Norwegii (Nordisk institutt for studier av innovasjon, forskning og utdanning, NIFU)[8].

W ostatnich latach w wielu państwach trwają intensywne prace nad centralnymi systemami informacji naukowej, które mają gromadzić dane o naukowcach i prowadzonych przez nich badaniach oraz pozwalać na przepływ tych informacji pomiędzy usługami. Z prezentacji i publikacji udostępnianych przez EuroCRIS wynika, że systemy takie stworzyło lub właśnie buduje co najmniej trzydzieści państw[9]. Jednym z najnowszych jest Research Portal Denmark, uruchomiony w 2022 roku. Zawiera on dane o nauce pochodzące z wielu źródeł, w tym od uczelni i instytucji finansujących naukę. System tworzony jest z wykorzystaniem środków rządowych ministerstwa do spraw edukacji i nauki oraz Duńskiej Agencji Szkolnictwa Wyższego i Nauki. Co ważne, portal oferuje dostęp nie tylko do państwowych źródeł danych, ale także do źródeł międzynarodowych, w tym zbieranych przez podmioty komercyjne: Clarivate, Elsevier i Digital Science[10]. W momencie powstawania monografii, portal oferował głównie dane o publikacjach duńskich naukowców i korzystał z integracji z wymienionymi bazami bibliograficzno-abstraktowymi.

Na tle systemów rozwijanych w innych państwach europejskich, RAD-on wyróżnia:

- szeroki zakres udostępnianych danych: zarówno z obszaru szkolnictwa wyższego, jak i nauki. RAD-on integruje informacje o dydaktyce i badaniach prowadzonych przez instytucje naukowe w Polsce;
- duży stopień dezagregacji danych i ich interoperacyjności. Niektóre z rejestrów RAD-onu, na przykład rejestr nauczycieli akademickich, dostarcza informacji o konkretnych osobach zatrudnionych w sektorze;
- rozległe portfolio oferowanych usług, w tym maszynowe przetwarzanie danych, narzędzia służące wizualizacji i pobieraniu danych już przetworzonych, opracowanych przez ekspertów i zawierających odpowiedni komentarz, czy też narzędzia umożliwiające wgląd i korektę danych przez osoby, których one dotyczą.

[7] hesa.ac.uk (dostęp: 02.03.2023)
[8] Z zakresem informacji udostępnianych przez trzy systemy można zapoznać się na następujących stronach: hesa.ac.uk/data-and-analysis (dostęp: 02.03.2023), unidata.gv.at/Pages/auswertungen.aspx (dostęp: 02.03.2023), nifu.no/fou-statistiske/fou-statistikk (dostęp: 02.03.2023)
[9] dspacecris.eurocris.org (dostęp: 02.03.2023)
[10] forskningsportal.dk (dostęp: 02.03.2023)

1.3. Polityka naukowa a proces informatyzacji sektora nauki i szkolnictwa wyższego w Polsce

Systemy informacji naukowej i dydaktycznej mogą wspierać bardzo wiele procesów, spośród których najważniejszym wydaje się być podział środków finansowych pomiędzy instytucje naukowe. Aby lepiej zrozumieć rolę RAD-onu w polityce naukowej, warto prześledzić cały proces rozwoju systemu nauki i szkolnictwa wyższego w Polsce. Jednym z najważniejszych aspektów tego procesu było stopniowe zwiększanie wagi przykładanej do danych naukometrycznych.

Julita Jabłecka i Benedetto Lepori [31] wyróżnili trzy fazy rozwoju systemu nauki w Polsce:

- 1989–1991: radykalna zmiana w krótkim czasie. W 1990 roku, wkrótce po zmianie ustrojowej i odejściu od socjalizmu, powrócono do zachodniej idei autonomii uniwersytetów i wolności badawczej;
- 1991–2000: znacząca stabilizacja. Od 1991 roku politykę naukową prowadził Komitet Badań Naukowych (KBN), który zajmował się między innymi podziałem środków na badania;
- 2000–2007: systematyczne zmiany prowadzące do dalszej restrukturyzacji polityki naukowej i systemu finansowania nauki. W 2005 roku KBN został włączony do resortu nauki. Politykę naukową kształtowało od tej pory Ministerstwo Nauki i Informatyzacji, a następnie Ministerstwo Edukacji i Nauki (MEiN).

W 2007 roku rozpoczął się proces, który doprowadził do powstania systemu nauki w jego dzisiejszym kształcie. Na początek powołano pierwszą niezależną agencję finansującą badania stosowane na wzór agencji funkcjonujących w państwach Unii Europejskiej czy też w Stanach Zjednoczonych Ameryki: Narodowe Centrum Badań i Rozwoju (NCBR)[11]. Do jego misji należy wspieranie innowacyjności, między innymi poprzez ułatwianie współpracy nauki i biznesu. Następnie, w 2010 roku, rozpoczęło działalność Narodowe Centrum Nauki (NCN)[12], czyli agencja finansująca badania podstawowe. W 2017 roku dołączyła do nich Narodowa Agencja Wymiany Akademickiej (NAWA)[13], która działa na rzecz umiędzynarodowienia polskiej nauki poprzez wspieranie i stymulowanie współpracy badawczej i wymiany akademickiej z zagranicą.

Trzy wymienione agencje dysponują środkami krajowymi i europejskimi na działalność naukową i innowacyjną. Rozdzielają je w konkursach na projekty realizowane przez indywidualnych naukowców lub zespoły badawcze zatrudniane przez instytucje naukowe i innowacyjne przedsiębiorstwa.

[11] ncbr.gov.pl (dostęp: 02.03.2023)
[12] ncn.gov.pl (dostęp: 02.03.2023)
[13] nawa.gov.pl (dostęp: 02.03.2023)

Jednocześnie z rozwojem projektowego systemu finansowania nauki, w Polsce stale udoskonalano system rozdziału środków zapewniających ciągłość działania instytucji naukowych. Uczelnie i pozostałe podmioty systemu nauki, oprócz funduszy na działalność dydaktyczną, mogą ubiegać się o publiczne środki na badania, rozdzielane w procesie ewaluacji instytucjonalnej. W odróżnieniu od finansowania projektowego, instytucjonalne nie jest przyznawane na konkretne inicjatywy badawcze. Decyzja o tym, co zostanie sfinansowane z wykorzystaniem tych środków, leży w gestii samych instytucji.

Pierwszą ewaluację działalności naukowej, w wyniku której rozdysponowano środki na badania prowadzone przez instytucje, przeprowadził KBN w 1991 roku. Oceny potencjału badawczego miały charakter ekspercki i zmierzały do kategoryzacji jednostek naukowych (takich jak wydziały uczelni) przez pryzmat tak zwanej doskonałości naukowej (*scientific excellence*). Kolejne ewaluacje w większym stopniu opierały się na wskaźnikach naukometrycznych, pozwalających na ilościowe ujęcie produktywności badawczej [32]. Produktywność jest pojęciem odnoszącym się do intensywności, z jaką naukowcy publikują, uczestniczą w projektach badawczych czy też angażują się w komercjalizację wyników swojej pracy naukowej [43]. Wykorzystanie naukometrii w ewaluacji miało z jednej strony służyć jeszcze większej obiektywizacji tego procesu, a z drugiej było powiązane z pojawiającymi się możliwościami informatyzacji sektora nauki i szkolnictwa wyższego. Na potrzeby funkcjonowania systemów rozdzielania środków zaczęto gromadzić coraz więcej danych. Ich przetwarzanie stanowiło wyzwanie legislacyjne, organizacyjne i technologiczne.

Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy (OPI PIB), który został utworzony w 1991 roku, od początku swojego istnienia wspierał od strony informatycznej i analitycznej omówione przemiany systemu nauki. Ośrodek czynił to, zapewniając infrastrukturę informatyczną obsługującą podział środków na granty, na przykład za pośrednictwem systemu Obsługa Strumieni Finansowych (OSF) oraz podział środków instytucjonalnych w ramach ewaluacji, poprzez System Ewaluacji Działalności Naukowej (SEDN). Instytut opracował również inne systemy bazodanowe, które gromadziły informacje o różnych aspektach działalności badawczej i dydaktycznej uczelni. Już w 2010 roku w OPI PIB rozpoczęły się prace nad systemem POL-on, w ramach którego, na podstawie obowiązujących przepisów prawa, uczelnie i inne instytucje naukowe, takie jak instytuty badawcze, instytuty Polskiej Akademii Nauk, międzynarodowe instytuty badawcze oraz instytuty Sieci Badawczej Łukasiewicz zobowiązane są zamieszczać różnorodne informacje. Wśród danych zbieranych w systemie POL-on znajdują się na przykład dane o działalności dydaktycznej i naukowej tych instytucji oraz ich dane finansowe [36].

OPI PIB zaangażowany był także w rozwój pozostałych systemów informatycznych resortu nauki i szkolnictwa wyższego, na przykład bazy Polska Bibliografia Naukowa (PBN), która jest – w uproszczeniu – polskim odpowiednikiem komercyjnych baz bibliograficzno-abstraktowych, takich jak Scopus. Poszczególne systemy tworzone były na różnych etapach działalności OPI PIB, co skutkowało tym, że różniły się one od siebie w aspekcie technologicznym i merytorycznym. Systemy te opiszemy szerzej w kolejnym podrozdziale.

RAD-on został opracowany we współpracy z MEiN. Nie jest to system obsługujący konkretny proces sprawozdawczy, finansowy czy ewaluacyjny, lecz miejsce udostępniające dane sprawozdawane przez instytucje naukowe do licznych systemów źródłowych, tworzonych w OPI PIB.

1.4. Potrzeby interesariuszy

Tworzeniu systemu RAD-on przyświecała idea integracji i udostępniania danych gromadzonych w systemach informatycznych OPI PIB w celu tworzenia polityki naukowej opartej na danych. Eksperci OPI PIB oraz MEiN na wstępie zdiagnozowali następujące problemy, których rozwiązania dostarczyć miał nowy system:

- rozproszenie danych o nauce i szkolnictwie wyższym w Polsce pomiędzy odseparowanymi od siebie systemami;
- niewystarczający poziom technicznej i metodologicznej integracji danych;
- brak centralnego miejsca do zarządzania danymi przetwarzanymi w poszczególnych systemach;
- niski stopień otwarcia danych.

Na podstawie wstępnej diagnozy zaprojektowaliśmy usługi, których założenia skonfrontowaliśmy z oczekiwaniami środowiska naukowego. Badanie jakościowe potrzeb interesariuszy RAD-onu przeprowadzono od 31 sierpnia do 25 września 2015 roku w Laboratorium Analiz Statystycznych i Ewaluacji pod kierownictwem dr Marzeny Feldy. Wykorzystano metodę wywiadów pogłębionych, umożliwiającą intensywne sondowanie i zebranie szczegółowych informacji. Łącznie w ramach badania odbyło się 19 spotkań i przebadano 29 osób. Każdy wywiad przebiegał zgodnie z przygotowanym wcześniej scenariuszem, składającym się z dwóch części ukierunkowanych na: 1) pozyskanie informacji o aktualnych i prognozowanych potrzebach w zakresie dostępu do informacji o nauce i szkolnictwie wyższym oraz sposobach ich zaspokajania oraz 2) zebranie opinii na temat usług i modułów planowanych w projekcie.

Rozmówcami byli interesariusze systemów i baz danych wytwarzanych przez OPI PIB: pracownicy naukowi, przedstawiciele administracji uczelni, kadra zarządzająca i pracownicy Ministerstwa Nauki i Szkolnictwa Wyższego (obecnie MEiN) oraz Ministerstwa Administracji i Cyfryzacji (obecnie Ministerstwo Cyfryzacji), członkowie Polskiej Komisji Akredytacyjnej (PKA), pracownicy agencji finansujących, czyli NCN i NCBR, a także dziennikarze i przedstawiciele organizacji pozarządowych z obszaru edukacji.

Jako uzupełnienie informacji pozyskanych w ramach wywiadów dokonano przeglądu potrzeb zgłaszanych do działu pomocy technicznej systemu POL-on w okresie od października 2013 do sierpnia 2015 roku.

Rozmówcy zwracali uwagę, że rozproszenie danych o sektorze w różnych bazach i portalach utrudnia ich codzienną pracę. Poszczególne instytucje wprowadzały dane do baz OPI PIB, jednak nie były w stanie z powrotem zasilić nimi wewnętrznych syste-

mów. Z ich punktu widzenia bazy centralne tworzyły jedynie nowe wymogi sprawozdawcze. Instytucje naukowe, agencje finansujące badania oraz PKA oczekiwały, że RAD-on przyczyni się do redukcji biurokracji. Ważnym elementem systemu, na który zwracali uwagę, miała być usługa zdalnej sprawozdawczości i automatycznego zasilania rejestrów szkolnictwa wyższego, zapewniająca terminowe wywiązywanie się z obowiązków sprawozdawczych. Badani chcieli ujednolicenia sposobu gromadzenia danych oraz zapewnienia ich stabilności i jakości. Newralgiczną kwestią pozostawało zadbanie o kompatybilność systemu centralnego z wewnętrznymi bazami instytucji oraz zapobieżenie dublowaniu się pracy w związku z synchronizacją różnych systemów.

Indywidualni pracownicy naukowi chcieli z kolei zwiększenia ich uprawnień do zarządzania własnymi danymi osobowymi zgromadzonymi w istniejących bazach. Jako korzyść z integracji rozproszonych systemów rozmówcy wymieniali skrócenie czasu potrzebnego na wyszukanie informacji naukowej oraz ograniczenie konieczności zapamiętywania wielu różnych kodów dostępu.

Obserwacje interesariuszy RAD-onu pomogły w doprecyzowaniu zakresu usług oferowanych przez system, które szczegółowo opisujemy w dalszych rozdziałach.

1.5. Cele projektu

Biorąc pod uwagę wyniki badań potrzeb użytkowników systemów OPI PIB, uznaliśmy, że system RAD-on powinien być systemem komplementarnym do ekosystemu POL-on [42]. RAD-on miał uzupełnić ofertę usług systemów OPI PIB o możliwości automatycznego przetwarzania i wizualizacji danych zagregowanych, a także zaoferować wsparcie w podejmowaniu decyzji w obszarze szkolnictwa wyższego, badań i rozwoju (B+R) oraz innowacyjności. Większość systemów OPI PIB gromadzi dane, tymczasem RAD-on miał zapewnić ich właściwe przetwarzanie i udostępnianie.

Tym samym cele projektu zdefiniowano jako:

- integracja danych o szkolnictwie wyższym i nauce w Polsce pochodzących z różnych autonomicznych baz danych;
- zapewnienie otwartego dostępu do najbardziej aktualnych i wiarygodnych danych dotyczących B+R, nauki i szkolnictwa wyższego;
- ograniczenie biurokracji między innymi poprzez odciążenie użytkowników w obszarze zasilania baz danymi;
- wspieranie procesów decyzyjnych poprzez dostarczenie właściwych narzędzi informatycznych umożliwiających analizę i interpretację udostępnianych danych.

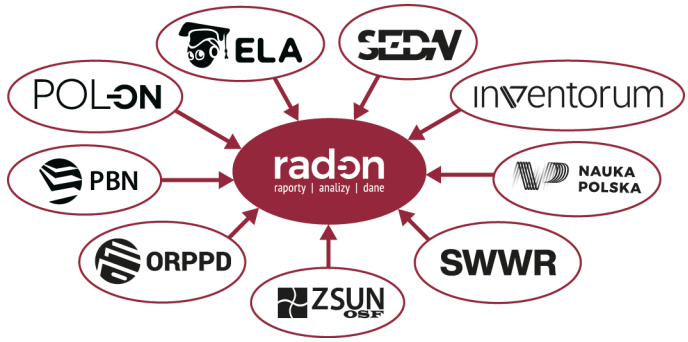
1.6. Dane udostępniane przez RAD-on

Aby osiągnąć tak sformułowane cele, w pierwszej kolejności należało zidentyfikować systemy, których integrację zapewnić miał RAD-on (Ilustracja 1.1). Ich lista przedstawia się następująco:

- **Zintegrowany System Informacji o Nauce i Szkolnictwie Wyższym POL-on**[14]**:** system gromadzący dane na podstawie ustawy Prawo o szkolnictwie wyższym [56] oraz rozporządzenia w sprawie systemu POL-on [47]. Obejmuje moduły danych o instytucjach nauki i szkolnictwa wyższego, studentach, nauczycielach akademickich, doktorantach, pracownikach naukowych i nauczycielach akademickich, prowadzonych postępowaniach awansowych w nauce, kierunkach studiów, inwestycjach, osiągnięciach naukowych, sprawozdaniach finansowych, sprawozdaniach z rekrutacji do szkół wyższych i inne;
- **Polska Bibliografia Naukowa (PBN)**[15]**:** system zawierający dane o dorobku publikacyjnym polskich naukowców i instytucji naukowych oraz o czasopismach;
- **Ogólnopolskie Repozytorium Pisemnych Prac Dyplomowych (ORPPD):** repozytorium, w którym od 2009 roku składowane są pisemne prace dyplomowe polskich absolwentów;
- **Ekonomiczne Losy Absolwentów (ELA)**[16]**:** portal prezentujący wyniki realizowanych co roku badań losów studentów, absolwentów i doktorantów, dotyczących ich sytuacji na rynku pracy. Badanie bazuje na anonimizowanych danych z systemu POL-on oraz Zakładu Ubezpieczeń Społecznych. Wyniki prezentowane są w formie interaktywnych raportów i infografik, pomagających studentom w wyborze kierunku kształcenia;
- **Inventorum**[17]**:** system, który ułatwia współpracę sektora gospodarki ze sferą nauki. Jako system aktywnie rekomendujący dostarcza dopasowanych informacji o innowacjach, projektach, innowacyjnych przedsiębiorstwach, naukowcach i instytucjach naukowych;
- **Nauka Polska**[18]**:** najstarsza baza danych OPI PIB, od 1999 roku udostępniana bezpłatnie w sieci internetowej. Zawiera dane o nauce polskiej w zakresie instytucji, osób, prac naukowych, publikacji i konferencji. Od roku 2018 jest częścią systemu POL-on, w wymiarze archiwalnym;
- **System Wspomagania Wyboru Recenzentów (SWWR)**[19]**:** adaptacyjna baza wiedzy o potencjalnych recenzentach, na przykład wniosków o granty. Rekomenduje ekspertów według zadanych kryteriów lub na podstawie tekstu wniosku;
- **System Ewaluacji Dorobku Naukowego (SEDN)**[20]**:** system obsługujący proces ewaluacji podmiotów nauki i szkolnictwa wyższego w Polsce, wykorzystujący dane z systemu POL-on oraz PBN;

[14] polon2.opi.org.pl (dostęp: 02.03.2023)
[15] pbn.nauka.gov.pl (dostęp: 02.03.2023)
[16] ela.nauka.gov.pl (dostęp: 02.03.2023)
[17] inventorum.opi.org.pl (dostęp: 02.03.2023)
[18] nauka-polska.pl (dostęp: 02.03.2023)
[19] recenzenci.opi.org.pl (dostęp: 02.03.2023)
[20] sedn.opi.org.pl (dostęp: 02.03.2023)

- **Obsługa Strumieni Finansowania / Zintegrowany System Usług dla Nauki (OSF)**[21]**:** system do rejestrowania i obsługi wniosków o finansowanie projektów i badań z obszaru B+R przyznawanych przez MEiN, NCN i NCBR, wymieniający z systemem POL-on informacje na temat instytucji nauki i szkolnictwa wyższego, projektów naukowych i publikacji (więcej o systemie w [6]).



Ilustracja 1.1. Systemy zintegrowane w ramach RAD-onu

1.7. Realizacja projektu i badania użyteczności

Oficjalna nazwa projektu RAD-on to „ZSUN II. Zintegrowany System Usług dla Nauki. Etap II". Projekt uzyskał dofinansowanie ze środków Programu Operacyjnego Polska Cyfrowa. Realizacja projektu rozpoczęła się w listopadzie 2017, a zakończyła w styczniu 2021 roku. Całkowity budżet wyniósł 27,6 miliona złotych. System jest nadal rozwijany i promowany w środowisku naukowym.

W trakcie tworzenia systemu RAD-on korzystano z połączenia metodyk zwinnych i zarządzania kaskadowego. Obecnie większość przedsięwzięć IT jest zarządzana w sposób zwinny, co oznacza dużą elastyczność w podejściu do wymagań produktowych oraz przyrostowe dostarczanie kolejnych usług lub ich fragmentów [13]. Organizacje finansujące badania ze środków publicznych preferują jednak tradycyjne podejścia kaskadowe, umożliwiające większą kontrolę nad budżetem i harmonogramem projektu. Oba podejścia mają istotne zalety, ale zasadniczo się od siebie różnią [55]. Na potrzeby projektu zdecydowaliśmy, że prace zespołu deweloperskiego prowadzone w metodyce SCRUM dostosujemy do wymaganego przez instytucję finansującą modelu zarządzania. Ponieważ projekt był współfinansowany ze środków Unii Europejskiej, budżet, harmonogram i zakres projektu, zatwierdzone na etapie ubiegania się o dofinansowanie, nie mogły ulec zmianie w czasie jego trwania. Połączenie metodyk zwinnych i kaskadowych generowało wiele wyzwań, których dokładne omówienie przekracza ramy niniejszej monografii.

[21] osf.opi.org.pl (dostęp: 02.03.2023)

W tym miejscu omówimy jedynie podjęte przez nas wysiłki zmierzające do zapewnienia interesariuszom systemu zadowolenia z dostarczonych usług. Ze względu na długi okres realizacji projektu, konieczne okazało się elastyczne podejście do produktów końcowych, uwzględniające zmieniające się oczekiwania użytkowników.

Aby móc właściwie ocenić jakość usług oferowanych przez RAD-on, po każdej kluczowej fazie rozwoju systemu zespół Laboratorium Interaktywnych Technologii OPI PIB przeprowadzał kompleksowe badania użyteczności. Łącznie było ich cztery. Wykorzystywały one metodę wywiadów pogłębionych. Wśród rozmówców wyróżniliśmy grupy interesariuszy systemu, zdefiniowane już na etapie wspomnianego wcześniej pierwszego badania potrzeb. Aby system spełniał wyśrubowane normy dostępności, konieczne okazało się uwzględnienie użytkowników z wadami wzroku oraz osób starszych.

Badaniami użyteczności objęliśmy na początku makietę systemu RAD-on, a w dalszej kolejności wybrane usługi i szatę graficzną. Po każdym badaniu powstał raport, którego rekomendacje modyfikowały zakres lub formę oferowanych usług. Powodowało to istotne zmiany w planie prac deweloperskich, ale przełożyło się na wyższe wskaźniki satysfakcji użytkowników.

Wskaźniki satysfakcji użytkowników monitorowane są z kolei z wykorzystaniem ankiety, dostępnej na portalu. Uwzględnia ona następujące aspekty: 1) ogólne zadowolenie z raportów, analiz i zintegrowanego dostępu do danych; 2) satysfakcję z jakości danych oraz 3) zadowolenie z usług maszynowego przetwarzania danych.

1.8. Wskaźniki projektu i przykłady wykorzystania systemu RAD-on

Aby zapewnić transparentność działań, zdecydowaliśmy się również na publiczny monitoring wskaźników projektu. System zbierający dane o popularności poszczególnych usług jest obecnie dostępny w serwisie RAD-on[22].

Od początku realizacji projektu za pośrednictwem RAD-onu udostępnionych zostało 11,45 terabajtów danych z obszaru nauki i szkolnictwa wyższego. W 2021 roku liczba pobrań lub odtworzeń dokumentów wyniosła ponad 150 milionów, co oznacza istotny, około pięciokrotny wzrost w stosunku do roku 2020.

Znaczące zwiększenie wykorzystania RAD-onu wynikało ze wzrostu zakresu danych dostępnych w systemie. Na przykład pomiędzy rokiem 2020 a 2021 zwiększył się on o około jedną czwartą. Dodatkowo, dzięki promocji projektu i rozpowszechnianiu informacji o stale poszerzanej ofercie usług, wzrosła liczba indywidualnych użytkowników i podmiotów, które regularnie wykorzystują udostępniane dane w swoich systemach.

[22] radon.nauka.gov.pl/o-systemie/statystyki (dostęp: 02.03.2023)

W okresie od grudnia 2021 do listopada 2022 roku do najpopularniejszych usług API, za pomocą których można pobrać oficjalne dane, należały:

- instytucje naukowe (POL-on): 105 004 023 pobrania;
- publikacje (PBN): 455 962 pobrania;
- pracownicy (POL-on): 193 142 pobrania;
- kierunki studiów (POL-on): 106 414 pobrań.

Aby przybliżyć Czytelnikom, jakie może być praktyczne zastosowanie udostępnianych danych, poniżej znajdują się przykłady wykorzystania usług RAD-onu przez podmioty zewnętrzne, a także w ramach integracji systemów administrowanych i rozwijanych przez OPI PIB.

Z RAD-onu korzysta Narodowa Agencja Wymiany Akademickiej, czyli agencja odpowiedzialna m.in. za upowszechnianie informacji o polskim systemie szkolnictwa wyższego i nauki na świecie. Pobiera ona informacje o instytucjach systemu szkolnictwa wyższego w Polsce oraz o prowadzonych na uczelniach kierunkach studiów. Dostęp do oficjalnych i regularnie odświeżanych danych umożliwia Agencji sprawną organizację procesu międzynarodowej uznawalności kształcenia. Dane z RAD-onu NAWA przekazuje przedstawicielom zagranicznych uczelni, ośrodków uznawalności wykształcenia czy organizacjom studenckim.

Inną instytucją sektora nauki wykorzystującą RAD-on jest Instytut Badań Edukacyjnych (IBE)[23] prowadzący interdyscyplinarne badania naukowe nad funkcjonowaniem i efektywnością systemu edukacji w Polsce. Dzięki usłudze maszynowego udostępniania publicznych danych (API RAD-onu), IBE może w łatwy i zautomatyzowany sposób zasilać danymi o kierunkach studiów prowadzony przez siebie rejestr kwalifikacji[24] nadawanych przez instytucje sektora szkolnictwa wyższego.

Wśród użytkowników API RAD-onu znajdują się także podmioty zagraniczne. W ramach projektu The Network4Growth organizacja UNICO pobiera dane o naukowcach z Polski, w tym o ich publikacjach, patentach oraz projektach. Inicjatywa ma na celu wzmacnianie potencjału transferu technologii w Polsce, Słowacji, Czechach, na Węgrzech oraz w Gruzji i Armenii[25].

Dane udostępniane w ramach RAD-onu są wykorzystywane w procesie składania wniosków o finansowanie badań naukowych w systemie Obsługa Strumieni Finansowych (OSF). OSF jest zintegrowany z usługą API RAD-onu zwracającą oficjalne dane instytucji systemu nauki i szkolnictwa wyższego z systemu POL-on. Dzięki temu podmioty składające wnioski nie muszą ponownie wprowadzać danych, które są dostępne w publicznym rejestrze. API RAD-onu wykorzystywane jest też przez systemy wewnętrzne dużych uczelni w Polsce, w szczególności w odniesieniu do danych o publikacjach.

[23] ibe.edu.pl (dostęp: 02.03.2023)
[24] kwalifikacje.gov.pl (dostęp: 02.03.2023)
[25] v4transfer.com (dostęp: 02.03.2023)

ROZDZIAŁ 2

PORTAL

Sylwia Ostrowska
Marcin Białas
dr Anna Knapińska

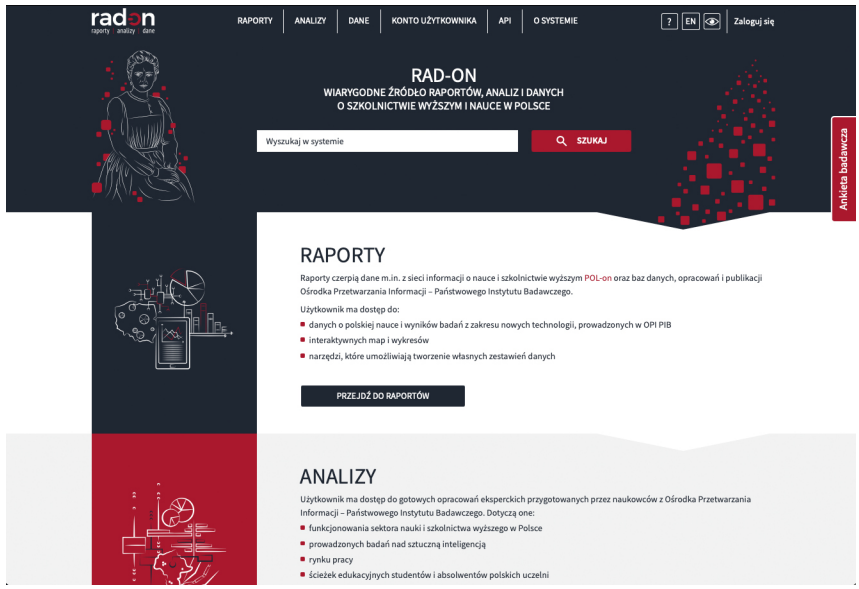
2.1. Użytkownicy i usługi portalu

System RAD-on był tworzony z myślą o naukowcach, studentach, przedsiębiorcach, pracownikach administracji publicznej oraz dziennikarzach poszukujących rzetelnych i aktualnych informacji o instytucjach nauki i szkolnictwa wyższego. Przed powstaniem systemu, informacje te udostępniane były przez różne, rozproszone bazy i portale. Mnogość źródeł danych znacząco utrudniała osobom zainteresowanym sektorem nauki w Polsce szybkie dotarcie do informacji. Jak pokazały badania potrzeb, opisane w poprzednim rozdziale 1.4, niektórzy interesariusze systemu nauki nie mieli wiedzy o istnieniu poszczególnych baz danych, stanowiących potencjalne źródło wartościowych informacji.

Projekt RAD-on obejmował tworzenie infrastruktury koniecznej do efektywnego zarządzania danymi o nauce i szkolnictwie wyższym. Od początku realizacji projektu wiedzieliśmy, że jego najważniejszym produktem będzie portal obywatelski, stanowiący otwarty, zintegrowany punkt dostępu do danych.

Opisując jego poszczególne funkcjonalności, chcemy zobrazować różnorodność form udostępniania danych. Pokażemy także, że system oferuje usługi dostosowane do potrzeb grona użytkowników zróżnicowanego pod względem umiejętności analitycznych i zakresu wiedzy o sektorze nauki w Polsce.

Portal RAD-on jest dostępny pod adresem radon.nauka.gov.pl. Strona główna portalu jest przedstawiona na Ilustracji 2.1.



Ilustracja 2.1. Strona główna portalu RAD-on

Usługi udostępniane na portalu RAD-on możemy podzielić na dostępne publicznie oraz wymagające uwierzytelnienia. Usługi dostępne w sposób otwarty wspólnie tworzą zintegrowaną bazę wiedzy dla sektora nauki i szkolnictwa wyższego. Należą do nich:

1. **Raporty:** interaktywne wizualizacje danych, pochodzących między innymi z systemu POL-on oraz baz danych, opracowań i publikacji OPI PIB.
2. **Analizy:** gotowe opracowania przygotowane przez ekspertów OPI PIB, dotyczące funkcjonowania sektora nauki i szkolnictwa wyższego w Polsce.
3. **Dane:** zestawienia danych o instytucjach, ich działalności naukowej i artystycznej, pracownikach czy postępowaniach w staraniach o awanse naukowe.

Przeszukiwanie zasobów nauki i szkolnictwa wyższego, udostępnianych w ramach wszystkich powyższych usług umożliwia **wyszukiwarka pełnotekstowa**.

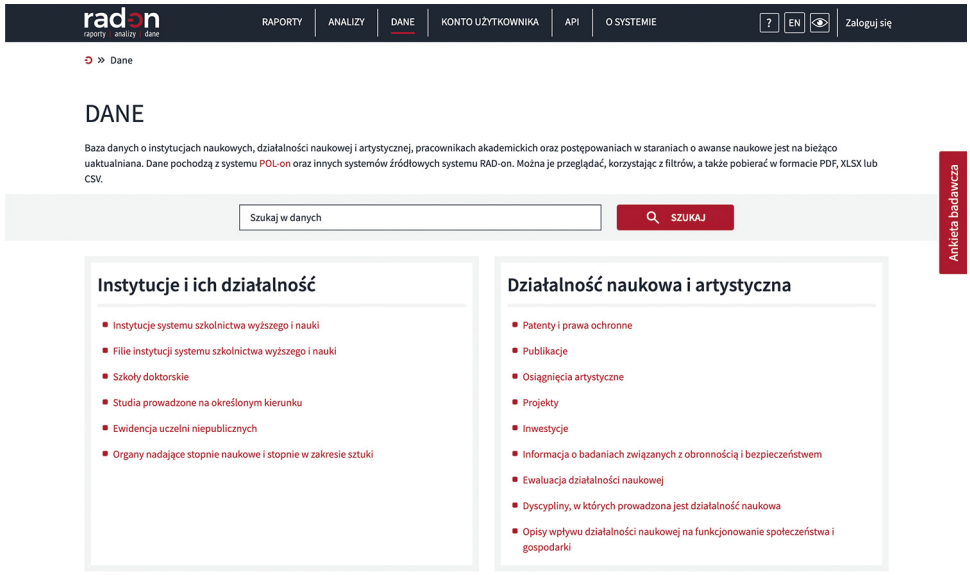
Katalog usług dostępnych dla użytkowników portalu uzupełnia **konto użytkownika** – jedyna funkcjonalność, która wymaga logowania. Zabezpieczenie usługi jest konieczne, ponieważ pozwala ona obywatelom na dostęp do swoich danych, w tym danych osobowych i innych niepublicznych informacji.

2.1.1. Dane

Moduł zawierający dane to podstawowa usługa, zapewniająca łatwy, bezpośredni dostęp do oficjalnych danych z wielu systemów źródłowych. W tym przypadku zależało nam, aby dane były prezentowanie w prostej formie, przyjaznej dla szerokiego grona

odbiorców i nie wymagającej wysokiego poziomu umiejętności analitycznych. Moduł ten obejmuje zestawienia z różnych kategorii tematycznych, które można przeglądać za pomocą filtrów i pobierać w różnych formatach. Obecnie dostępnych jest kilkanaście zestawień danych, zaprezentowanych na Ilustracji 2.2. Są one podzielone na następujące kategorie:

1. **Instytucje i ich działalność:** instytucje systemu nauki i szkolnictwa wyższego w Polsce, ich filie, prowadzone kierunki studiów, szkoły doktorskie oraz organy nadające stopnie i tytuły naukowe.
2. **Działalność naukowa i artystyczna:** patenty i prawa ochronne, publikacje, projekty, osiągnięcia artystyczne, inwestycje i informacje związane z ewaluacją jakości działalności naukowej podmiotów.
3. **Pracownicy:** nauczyciele akademiccy, inne osoby prowadzące zajęcia, osoby prowadzące działalność naukową oraz osoby biorące udział w jej prowadzeniu.
4. **Postępowania awansowe:** wykaz postępowań dotyczących nadania stopnia doktora, doktora habilitowanego oraz tytułu profesora wraz z wyszukiwarką plików z postępowań awansowych (rozprawy, streszczenia i opisy, recenzje i opinie, wnioski).



Ilustracja 2.2. Moduł „Dane" portalu RAD-on

Dane pochodzą z oficjalnych źródeł (takich jak POL-on czy PBN), dla których zakres danych, zasady ich wprowadzania, aktualizacji i dostępu są zdefiniowane w aktach prawnych [56, 47]. Informacje są wprowadzane do systemu wyłącznie przez uprawnione podmioty (np. uczelnie, instytucje naukowe, ministerstwa), co gwarantuje ich jakość i rzetelność. Na portalu RAD-on udostępniane są wyłącznie te dane, które zostały wskazane w aktach prawnych jako publicznie dostępne.

Spośród udostępnianych rejestrów szczególną rolę pełni wykaz instytucji nauki i szkolnictwa wyższego, obejmujący następujące podmioty wskazane w ustawie [56][26]:

- uczelnie publiczne;
- uczelnie niepubliczne;
- uczelnie kościelne;
- instytucje naukowe;
- federacje instytucji.

Każda instytucja posiada swój profil, na którym prezentowane są jej podstawowe dane (np. nazwa, numery identyfikacyjne, osoba kierująca, organ nadzorujący, status) wraz z historią zmian, dostępną dla wybranych pól. Użytkownicy mogą także przeglądać szczegółowe dane, korzystając z linkowania do podstron prezentujących wybrane aspekty działalności określonej instytucji. Profil każdej instytucji zawiera odnośniki do wszelkich powiązanych z nią informacji, w tym dotyczących:

- filii uczelni, w których prowadzi działalność poza swoją główną siedzibą (np. kierunki studiów);
- prowadzonych kierunków studiów i szkół doktorskich;
- pracowników;
- działalności naukowej i artystycznej (publikacje, osiągnięcia artystyczne, patenty, projekty);
- inwestycji związanych między innymi z kształceniem i działalnością naukową, w tym w aparaturę naukowo-badawczą oraz infrastrukturę informatyczną o wartości przekraczającej pół miliona złotych;
- prowadzonych postępowań awansowych, dotyczących nadania stopnia doktora i doktora habilitowanego;
- ewaluacji jakości działalności naukowej podmiotu, o której mowa w rozdziale 1.

Niezależnie od możliwości, jakie daje linkowanie między zasobami, użytkownicy mogą przeglądać odrębnie każde z zestawień, korzystając z rozbudowanej sekcji filtrów. Dodatkowo mogą pobierać dane do wybranego formatu (XLSX, CSV, PDF), zachowując przy tym ustawienia filtrowania. Dane z zestawień są również dostępne do pobrania maszynowego za pomocą usług interfejsu programowania aplikacji (*Application Programming Interface*, API), opisanych w rozdziale 5.

Ułatwienie dostępu do danych, poprzez zgromadzenie ich w jednym miejscu oraz zastosowanie rozwiązań dostosowanych do potrzeb i umiejętności analitycznych użytkowników, w dłuższej perspektywie przyczynia się również do zwiększenia ich jakości i aktualności. Wraz ze wzrostem liczby użytkowników, zwiększa się także poziom kontroli nad danymi. Na przykład nauczyciele akademiccy mogą weryfikować, jakie dane na ich temat zostały wprowadzone przez uczelnię do systemu źródłowego POL-on, a następnie opublikowane na portalu RAD-on. W razie wykrycia nieścisłości mogą zgłosić

[26] Art. 7 ust. 1 pkt 1, 2, 4–8

instytucji konieczność poprawy danych. Z kolei instytucjom, których dane są publikowane na portalu, zależy na ich kompletności, co może motywować do częstszej aktualizacji danych w systemach źródłowych. Większość zestawień danych jest aktualizowana raz dziennie, a wybrane rejestry nawet co godzinę. Dzięki temu użytkownicy mają szybki dostęp do poprawionych lub uzupełnionych danych.

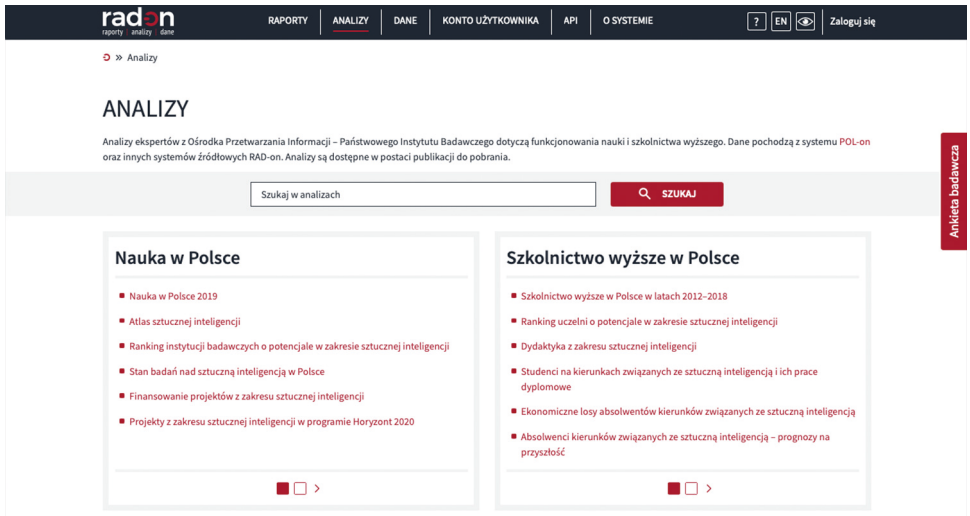
2.1.2. Raporty

Całościowy proces tworzenia interaktywnych raportów, z wykorzystaniem autorskiego systemu, szeroko omówimy w kolejnym rozdziale 3. W tym miejscu warto nadmienić, że pulpit nawigacyjny zawierający raporty jest na bieżąco aktualizowany i obecnie składa się z trzynastu folderów – kategorii. Znajdują się w nich tematyczne raporty o uczelniach, rekrutacji na studia, studentach, kierunkach studiów, absolwentach, doktorantach, nauczycielach akademickich oraz o badaniach nad sztuczną inteligencją (Ilustracja 3.2 przedstawia widok niektórych folderów).

2.1.3. Analizy

Opracowane przez ekspertów naukowych OPI PIB analizy, udostępniane jako możliwe do pobrania publikacje w formacie PDF, poświęcone są różnym aspektom funkcjonowania sektora szkolnictwa wyższego i nauki. Niektóre z nich stanowią obszerne, przekrojowe opracowania statystyczne. Obecnie udostępnionych jest kilkanaście analiz (Ilustracja 2.3) podzielonych na następujące foldery – kategorie: 1) nauka w Polsce; 2) szkolnictwo wyższe w Polsce; 3) ludzie nauki w Polsce. Z poszczególnych analiz czytelnicy dowiedzą się między innymi:

- jak wygląda finansowanie sfery badawczo-rozwojowej w Polsce,
- które ośrodki naukowe przodują w badaniach nad sztuczną inteligencją,
- jak przedstawiają się finanse uczelni,
- jaki jest stan nauk humanistycznych w Polsce,
- z jakich państw pochodzą cudzoziemcy odbywający studia w Polsce,
- jaki jest poziom umiędzynarodowienia kadry akademickiej,
- ile kobiet studiuje na kierunkach technicznych i informatycznych.



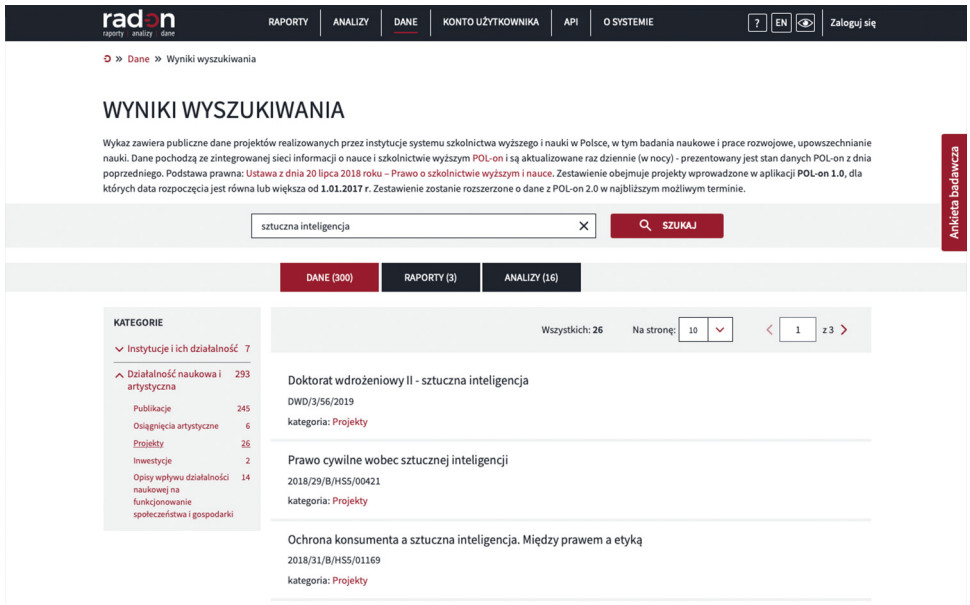
Ilustracja 2.3. Moduł „Analizy" portalu RAD-on

2.1.4. Wyszukiwarka

Wyszukiwarka pełnotekstowa oparta na narzędziu Elasticsearch (ES) [3] analizuje źródłowe bazy danych i pozwala na odkrywanie korelacji semantycznych między danymi.

Użytkownik może przeglądać wyniki wyszukiwania z modułów zawierających dane, raporty i analizy oraz wybrać konkretną kategorię lub podkategorię informacji, którą jest zainteresowany (np. kategoria: działalność naukowa i artystyczna, podkategoria: projekty – przykład na Ilustracji 2.4).

Z poziomu wyników wyszukiwania istnieje możliwość przejścia do szczegółów, na przykład profilu projektu, instytucji czy raportu dotyczącego badań w określonym obszarze. Dodatkowo każda sekcja posiada własną wyszukiwarkę pozwalającą na przeszukiwanie zasobów w wybranym obszarze.



Ilustracja 2.4. Wyniki wyszukiwania na portalu RAD-on

2.1.5. Konto użytkownika

W procesie otwierania lub udostępniania danych należy szczególnie zadbać o ochronę danych osobowych. Aby zachować zgodność z Rozporządzeniem o Ochronie Danych Osobowych (RODO, *General Data Protection Regulation*, GDPR, [46]), infrastruktura informatyczna powinna oferować funkcjonalności umożliwiające użytkownikom dostęp do swoich danych, ich poprawianie lub usuwanie, ograniczenie lub sprzeciw wobec ich przetwarzania, a także przekazanie danych innemu administratorowi [8]. Aby spełnić te wymagania, w ramach systemu RAD-on stworzyliśmy usługę dostępu do danych obywatela. Dzięki niej użytkownicy mogą w jednym miejscu zweryfikować swoje dane, zgromadzone w wielu różnych systemach źródłowych.

Ze względu na zakres danych przetwarzanych w systemach źródłowych, usługa jest skierowana przede wszystkim do osób związanych ze szkolnictwem wyższym i nauką, w szczególności do studentów, nauczycieli akademickich i pracowników naukowych oraz ekspertów i recenzentów. Zalogowany użytkownik może za pośrednictwem swojego konta zweryfikować, czy jego dane są przetwarzane w systemach i bazach danych związanych ze szkolnictwem wyższym i nauką w Polsce. Do tych systemów należą między innymi: POL-on, Inventorum, Nauka Polska, System Wspomagania Wyboru Recenzentów, Polska Bibliografia Naukowa (systemy szerzej opisane są w rozdziale pierwszym 1.6). Następnie użytkownik może pobrać indywidualne raporty przedstawiające wszystkie jego dane przetwarzane w poszczególnych systemach. Raporty zawierają dane o przebiegu studiów i uzyskanych stypendiach, o projektach i publikacjach

naukowych, historię zatrudnienia w sektorze szkolnictwa wyższego i nauki, uzyskane stopnie i tytuły naukowe itp.

Podstawą usługi dostępu do danych obywatela są dwa rozwiązania:

- **hurtownia danych:** łączy, deduplikuje i agreguje dane o nauce i szkolnictwie wyższym ze wszystkich systemów dziedzinowych. Dane osobowe z różnych systemów transakcyjnych są przesyłane do hurtowni, a następnie integrowane z profilem osoby.
- **Moduł Centralnego Logowania (MCL):** zapewnia użytkownikom dostęp do usługi. MCL jest zintegrowany z Krajowym Węzłem Identyfikacji Elektronicznej[27] (KWIE, login.gov.pl).

Usługa pozwalająca na dostęp do danych obywatela używa danych zagregowanych w hurtowni, to znaczy korzysta z utworzonych profili osób. Serwis jest dostępny poprzez konto użytkownika. Aby wszelkie niepubliczne informacje dotyczące obywateli były odpowiednio zabezpieczone, dostęp do usługi wymaga uwierzytelnienia za pomocą systemu login.gov.pl. Dzięki integracji z KWIE, system MCL/RAD-on (występujący w roli dostawcy usług, DU) jest w stanie zidentyfikować użytkownika u wszystkich dostawców tożsamości zintegrowanych z login.gov.pl. Systemy identyfikacji podłączone do węzła muszą spełniać najwyższe standardy bezpieczeństwa. Potwierdzenie tożsamości jest możliwe przy użyciu wybranego przez użytkownika środka identyfikacji elektronicznej (np. profilu zaufanego, e-dowodu czy systemu bankowego). Celem takiego modelu dostępu do usługi jest udostępnianie danych osobowych wyłącznie osobie, której dotyczą.

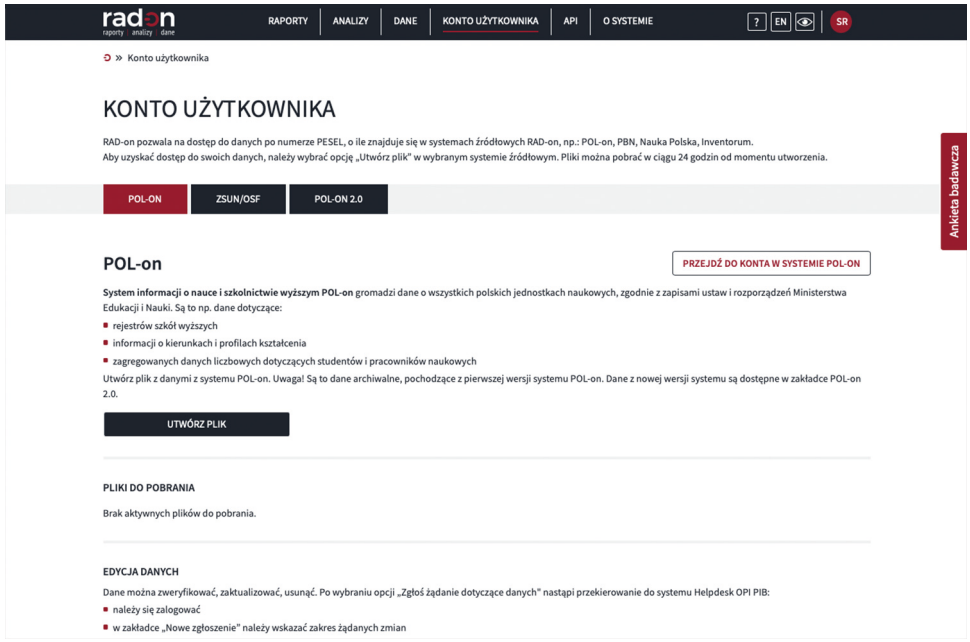
Proces potwierdzania tożsamości użytkownika działa na podstawie sfederowanego modelu tożsamości[28] login.gov.pl, który został przedstawiony na Ilustracji 2.5.

Ilustracja 2.5. Sfederowany model tożsamości

MCL/RAD-on kontaktuje się za pośrednictwem login.gov.pl z dostawcą tożsamości (DT) w celu uzyskania potwierdzenia tożsamości użytkownika, pragnącego skorzystać z usługi dostępu do danych obywatela. System DT uwalnia dane osobowe użytkownika po otrzymaniu żądania od systemu zewnętrznego (DU) i poprawnym zweryfikowaniu tożsamości użytkownika. Na podstawie danych osobowych uzyskanych od DT, po stronie hurtowni danych następuje weryfikacja, w jakich systemach źródłowych przetwarzane

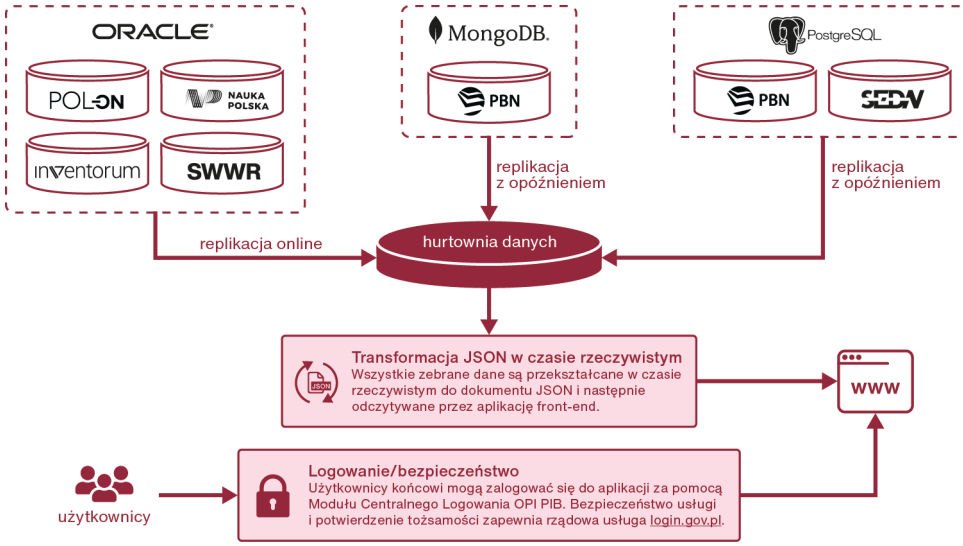
[27] gov.pl/web/login (dostęp: 02.03.2023)
[28] Podstawą sfederowanego (rozproszonego) modelu tożsamości są liczne środki identyfikacji elektronicznej, wydawane przez różne podmioty (publiczne i prywatne). Do publicznych środków identyfikacji elektronicznej można zaliczyć profil zaufany czy dowód osobisty z warstwą elektroniczną (tzw. e-dowód), natomiast do komercyjnych – bankowość elektroniczną.

są dane użytkownika. Informacja o systemach prezentowana jest użytkownikowi. Lista systemów może się różnić w zależności od tego, czy zalogowana osoba jest – na przykład – studentem, recenzentem czy autorem publikacji. Na Ilustracji 2.6 przedstawiono przykładowy widok przygotowany dla użytkownika, którego dane są zgromadzone w systemach POL-on i OSF.



Ilustracja 2.6. Moduł „Konto użytkownika" portalu RAD-on – widok po zalogowaniu

W niektórych przypadkach żaden system nie będzie dostępny, co oznacza, że w systemach źródłowych RAD-onu nie znaleziono danych tej osoby. Użytkownicy mogą zamawiać i pobierać indywidualne raporty zawierające ich dane z każdego systemu. Po zarejestrowaniu zlecenia utworzenia raportu z danymi z wybranego systemu, usługa pobiera dane z hurtowni i udostępnia użytkownikowi. Pliki można pobrać w ciągu 24 godzin w dwóch formatach: PDF (zalecany do uporządkowanej prezentacji danych), JSON (format do odczytu maszynowego zalecany do dalszego przetwarzania, na przykład do analizy lub przesyłania danych). Każde zlecenie utworzenia danych z systemu dotyczące konkretnej osoby oraz każde pobranie dokumentu jest rejestrowane w historii po stronie hurtowni danych. Całościowa, uproszczona architektura usługi przedstawiona została na Ilustracji 2.7.



Ilustracja 2.7. Uproszczona architektura usługi dostępu do danych obywatela

Po pobraniu danych obywatel może zweryfikować ich poprawność, kompletność oraz aktualność i w razie potrzeby zareagować na błędy. Usługa jest zintegrowana z systemem obsługi zgłoszeń i błędów ZSUN Helpdesk[29], za pośrednictwem którego użytkownik może zgłosić żądanie dotyczące swoich danych (np. korektę czy usunięcie, czyli tak zwane prawo do bycia zapomnianym). Każdy wniosek użytkownika jest analizowany i po pozytywnej weryfikacji dane są modyfikowane w odpowiednim systemie dziedzinowym. Należy jednak podkreślić, że nie wszystkie wnioski mogą zostać zrealizowane. Na przykład, w przypadku systemu POL-on ma zastosowanie przesłanka wykluczająca możliwość skorzystania z prawa do bycia zapomnianym i w konsekwencji usunięcia danych[30].

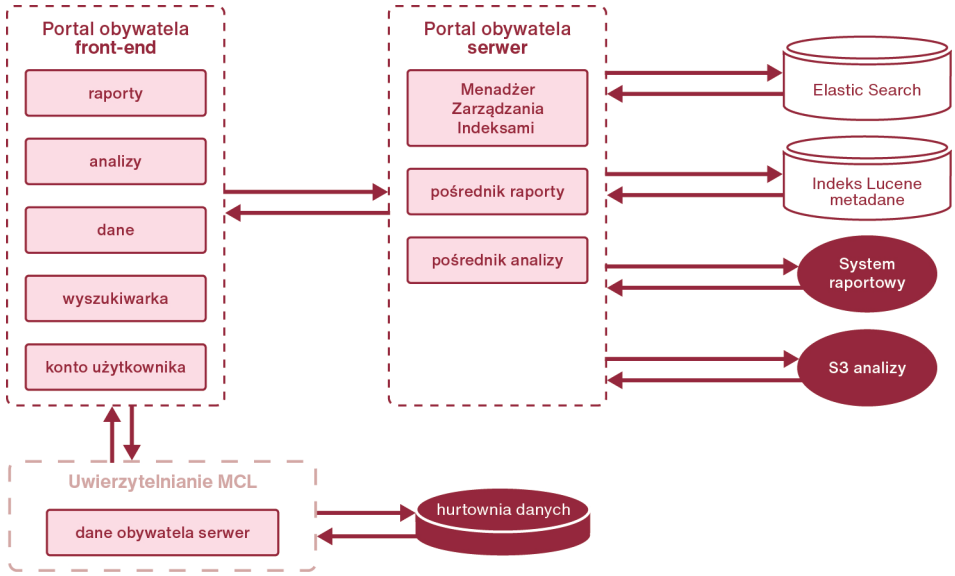
Dzięki zastosowanemu modelowi wymiany danych, zaktualizowane dane zasilają zintegrowane systemy oraz są publikowane na portalu RAD-on w ramach udostępnianych usług. Obywatele mają więc bezpośredni wpływ na poprawność danych dotyczących nauki i szkolnictwa wyższego w Polsce. Usługa, która pozwala użytkownikom na kontrolę poprawności swoich danych, pozwala nam dostarczać najbardziej aktualne, rzetelne i wiarygodne informacje o nauce i szkolnictwie wyższym. Jednocześnie umożliwia wypełnienie obowiązków wynikających z RODO oraz przyczynia się do dalszej cyfryzacji administracji publicznej.

[29] lil-helpdesk.opi.org.pl (dostęp: 02.03.2023)
[30] Przesłanka ta została określona w art. 17 ust. 3 lit. b) RODO, z której wynika, że uprawnienie realizujące „prawo do bycia zapomnianym" nie jest możliwe do zrealizowania, jeżeli przetwarzanie jest niezbędne do wywiązania się z prawnego obowiązku wymagającego przetwarzania na mocy prawa Unii lub prawa państwa członkowskiego, któremu podlega administrator, lub do wykonania zadania realizowanego w interesie publicznym lub w ramach sprawowania władzy publicznej powierzonej administratorowi.

2.2. Architektura portalu

Zarys architektury portalu obywatela przedstawia Ilustracja 2.8. Komponentami wchodzącymi w skład portalu są:

- **portal obywatela front-end:** aplikacja Angular JS[31];
- **portal obywatela serwer:** aplikacja JAVA Spring[32];
- **system raportowy:** aplikacja JAVA Spring;
- **S3 analizy:** serwer plików S3[33], na którym składowane są pliki z analizami;
- **dane obywatela serwer:** aplikacja JAVA Spring.



Ilustracja 2.8. Zarys architektury portalu obywatela

Użytkownikom udostępniana jest aplikacja internetowa (**portal obywatela front-end**) składająca się z pięciu modułów: 1) raporty; 2) analizy; 3) wyszukiwarka; 4) dane i 5) konto użytkownika. Moduły komunikują się z częścią serwerową za pośrednictwem API, który przedstawimy w ostatnim rozdziale 5.3.3. Omówimy teraz po kolei komunikację każdego z modułów z jego odpowiednikiem po stronie serwerowej.

[31] angularjs.org (dostęp: 02.03.2023)
[32] docs.spring.io (dostęp: 02.03.2023)
[33] min.io (dostęp: 02.03.2023)

Moduł raportowy, prezentujący interaktywne raporty, zasilany jest danymi pochodzącymi z **systemu raportowego**, opisanego w kolejnym rozdziale. Aplikacja internetowa komunikuje się z nim, wykorzystując proxy, jakim jest **portal obywatela serwer**. W ramach API udostępnione są zarówno metadane, jak i dane JSON raportów. Aplikacja webowa, po pobraniu metadanych, wyświetla użytkownikowi listę dostępnych raportów. Jeśli użytkownik zdecyduje się obejrzeć któryś z nich, pobierane są dane szczegółowe raportu. Na podstawie danych pobranych w formacie JSON generowane są wykresy. Interaktywną formę raportów zapewniają zestawy filtrów, które mogą być dowolnie zmieniane przez użytkownika. Obsługa filtrów również odbywa się za pośrednictwem API.

Serwer portalu obywatela jest również serwerem pośredniczącym dla modułu analiz. Podobnie jak w przypadku raportów, API udostępnia metadane o dostępnych analizach. Jeśli użytkownik postanowi pobrać plik zawierający jedną z dostępnych analiz, aplikacja webowa wysyła zapytanie do API i następuje pobieranie pliku. Wykorzystanie metadanych w modułach raportów i analiz umożliwia łatwe publikowanie nowych zasobów bez konieczności modyfikowania kodu aplikacji. Innymi słowy, treści w modułach raportów i analiz mogą być dynamicznie modyfikowane przez analityków bez konieczności wdrażania nowej wersji kodu.

Moduł danych zasilany jest rekordami zindeksowanymi na klastrze Elasticsearch (ES). Szczegóły techniczne zarządzania indeksami przez Menadżera Zarządzania Indeksami (MZI) zostały omówione w rozdziałach 5.3.2 i 5.3.3. Poza wyświetlaniem danych istnieje też możliwość pobrania ich w postaci plików w formatach PDF, CSV i XLSX. Za transformację rekordów ze źródłowego formatu JSON do formatów CSV i XLSX, odpowiedzialna jest część serwerowa portalu. Natomiast pliki PDF są generowane przez aplikację webową.

Wyszukiwarka umożliwia przeszukiwanie pełnotekstowe danych dostępnych w modułach raportów, analiz i danych. W przypadku raportów i analiz tworzony jest dodatkowy indeks Lucene [5], którym zarządza część serwerowa portalu obywatela. Dokumenty w indeksie tworzone są na podstawie metadanych o raportach i analizach. Informacje przedstawiane w module danych zindeksowane są na klastrze ES i zarządzane przez MZI (więcej w rozdziale 5.3.2). Zapytanie wykonane przez użytkownika za pośrednictwem API trafia do obu indeksów. Wyniki dla różnych modułów nie są łączone. Wyniki zgodne z zapytaniem klienta prezentowane są w odrębnych zakładkach.

Ostatnim z modułów jest konto użytkownika. Prezentuje on dane, których obejrzenie jest możliwe dopiero po zalogowaniu się. Do uwierzytelniania wykorzystywany jest MCL. Po zalogowaniu użytkownik może pobrać swoje dane w formatach PDF i JSON. Dzieje się to za pośrednictwem aplikacji **dane obywatela serwer**. Źródłem danych dla konta użytkownika jest hurtownia danych, z którą bezpośrednio komunikuje się aplikacja serwerowa. Aplikacja serwerowa zajmuje się również generowaniem plików PDF, które powstają ze źródłowych plików JSON.

ROZDZIAŁ 3

RAPORTY

dr Anna Knapińska
dr Aldona Tomczyńska
dr inż. Sławomir Dadas

3.1. Narzędzia BI a platformy analityczne

System raportowy, o którym wspomnieliśmy w poprzednim rozdziale, jest głównym narzędziem analitycznym dostępnym za pośrednictwem portalu RAD-on.

W tym rozdziale zademonstrujemy jego przydatność do przetwarzania danych. Pokażemy również przykłady interaktywnych raportów, prezentowanych na portalu i zbudowanych z wykorzystaniem systemu raportowego. W ramach wprowadzenia, rozpoczniemy jednak od zidentyfikowania różnic pomiędzy niezależnymi platformami analitycznymi a komercyjnymi narzędziami *business intelligence* (BI).

Wraz z gwałtownym przyrostem ilości przetwarzanych i gromadzonych danych cyfrowych rośnie również popularność systemów służących do ich analizy i wizualizacji. Są to kompleksowe rozwiązania, które łączą ze sobą różne technologie, we wszystkich fazach cyklu analizy danych, w celu zaspokojenia konkretnych potrzeb biznesowych. Ich działanie polega na przechowywaniu danych, zarządzaniu nimi, przygotowywaniu danych do analizy oraz przeprowadzaniu analiz w celu pozyskania cennych informacji.

Jednym z przykładów są aplikacje z obszaru BI, służące do analizy danych i dokonywania nowych, użytecznych spostrzeżeń pomocnych w zwiększaniu przewagi konkurencyjnej przedsiębiorstw. Termin *business intelligence* został wprowadzony przez Howarda Dresnera z Gartner Group w 1989 roku [40]. Czołowymi producentami oprogramowania w tej dziedzinie są obecnie Microsoft, Tableau i Qlik [44]. Głównymi odbiorcami aplikacji BI są duże przedsiębiorstwa, które korzystają z oferowanych funkcjonalności do analizy własnych biznesów. Większość użytkowników takiego narzędzia to zatem pracownicy korporacji.

Jeśli jednak rozpatrujemy znaczenie systemów analitycznych z perspektywy procesów podejmowania decyzji [63], to trzeba zwrócić uwagę, że jest to kluczowe działanie nie tylko w sektorze prywatnym, chociaż znacznie częściej wykorzystywane jest właśnie w nim [16]. Również sektor publiczny staje przed wyzwaniem przetwarzania, gromadzenia i analizy coraz większej ilości danych. Przeprowadzanie wyrafinowanych

działań analitycznych czyni procesy decyzyjne skuteczniejszymi, a to przyczynia się do zwiększania wydajności rozmaitych obszarów zarządzania publicznego i poszczególnych instytucji [66].

Nie inaczej jest w sektorze nauki i szkolnictwa wyższego, w którym uczelnie i inne instytucje badawcze rywalizują ze sobą o studentów, naukowców i fundusze na badania. Posiadanie i przetwarzanie szczegółowych informacji na ten temat ułatwia zyskiwanie przewagi konkurencyjnej, ale jednocześnie rodzi nowe wyzwania, związane chociażby z bezpieczeństwem danych oraz wzrostem wydatków. Nie ulega jednak wątpliwości, że zarówno dla pojedynczych instytucji naukowych, jak i podmiotów z obszaru polityki naukowej niezbędne jest intensywne wykorzystywanie danych do podejmowania decyzji i zwiększania skuteczności działań (np. [25, 49, 52]). Jest to jeden z wymiarów wciąż przyspieszającej transformacji cyfrowej, która staje się częścią kultury akademii [67]. Ważna jest przy tym nie tylko wiedza o tym, co się wydarzyło i co się aktualnie dzieje, ale też o tym, jakie są prognozy na przyszłość wraz z możliwymi scenariuszami i konsekwencjami potencjalnych decyzji [18].

Rozwiązania z zakresu analityki deskryptywnej, predykcyjnej i preskryptywnej zapewnia oprogramowanie BI, jednak z perspektywy instytucji publicznych najważniejszą jego wadą są zaporowe koszty licencji i pozalicencyjne [28]. Narzędzia te są zazwyczaj wyceniane dwojako, poprzez licencje na użytkownika i licencje na sprzęt (procesor). Podobnie wysokie i trudne do prognozowania są koszty rozwiązań BI w chmurze. Drugim poważnym problemem jest nieudostępnianie kodów źródłowych, bez czego niemożliwe staje się dostosowywanie usług do specyficznych potrzeb użytkowników końcowych.

W ten sposób dostęp do danych podlega poważnym ograniczeniom, a to stoi w sprzeczności z polityką otwartego dostępu (*open access*), jedną z ważniejszych reguł w państwach należących do Organizacji Współpracy Gospodarczej i Rozwoju (OECD) oraz Unii Europejskiej. W Polsce ustawa z dnia 11 sierpnia 2021 roku o otwartych danych i ponownym wykorzystywaniu informacji sektora publicznego określa zasady otwartości danych, zasady i tryb udostępniania i przekazywania informacji sektora publicznego w celu ponownego wykorzystywania oraz podmioty, które udostępniają lub przekazują te informacje. Szczególne znaczenie ma zmiana systemowa polegająca na udostępnianiu danych i wiedzy naukowej, co poprawia jakość prowadzonych badań i sprawia, że mogą one lepiej odpowiadać na potrzeby społeczne. To dlatego w programach ramowych Horyzont 2020 i Horyzont Europa, największych w historii UE przedsięwzięciach w zakresie badań naukowych i innowacji, od beneficjentów wymaga się takiego zarządzania danymi badawczymi, aby dane były możliwe do znalezienia, dostępne, interoperacyjne i nadawały się do ponownego wykorzystania (*findable, accessible, interoperable, reusable* – FAIR) [53] (więcej w rozdziale 1.1). Inicjatywą wartą odnotowania jest również Europejska Chmura dla Otwartej Nauki (European Open Science Cloud, EOSC), na którą składają się również zaawansowane narzędzia i zasoby do składowania, udostępniania i przetwarzania danych [9].

Koszty systemów komercyjnych z jednej strony, a nacisk na otwartość publicznych danych z drugiej doprowadziły do tego, że w sektorze publicznym tworzone są otwarte platformy analityczne. W obszarze nauki i szkolnictwa wyższego platformy stanowią

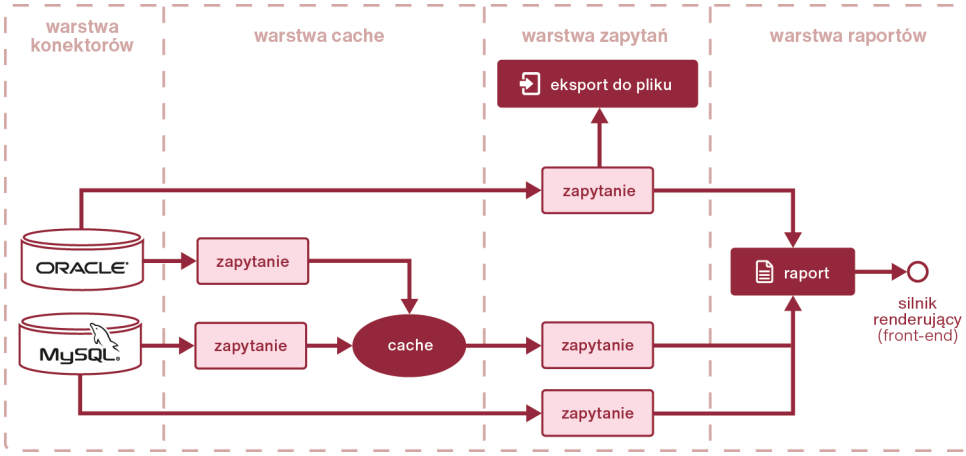
wsparcie dla twórców polityki naukowej oraz samych badaczy, nie tylko dzięki sprawniejszemu gromadzeniu danych i statystyk oraz przyspieszaniu i ułatwianiu wyciągania wniosków i podejmowania decyzji opartych na tych informacjach. Równie istotne jest zapewnianie standardów interoperacyjności do celów badań porównawczych (takich jak ocena osiągnięć naukowych) na poziomie lokalnym, krajowym i międzynarodowym. Bazy danych i pulpity nawigacyjne (*dashboards*) prezentujące statystyki publiczne na temat badań naukowych, innowacji i szkolnictwa wyższego dostarczane są na przykład przez Europejski Urząd Statystyczny (Eurostat) i OECD, o czym pisaliśmy w rozdziale pierwszym. W Polsce źródłem otwartych informacji o szkolnictwie wyższym i nauce jest omawiany w niniejszej monografii system RAD-on [36, 41].

Platforma analityczna OPI PIB działa bez kosztów licencyjnych i na dowolnej stronie internetowej, dlatego może stanowić alternatywę dla komercyjnego oprogramowania BI. Oprócz analiz i danych, nieodłączną częścią systemu RAD-on jest dashboard w postaci raportów zawierających interaktywne tabele i wykresy, które mogą służyć jako narzędzie kreowania polityki opartej na dowodach i wpisują się w ideę rządu jako otwartej platformy dostarczającej społeczeństwu wiarygodne dane [7]. Raporty te są stale rozwijane przy użyciu specjalnego silnika, który omówimy poniżej.

3.2. Zaplecze platformy analitycznej RAD-onu

Kreator raportów (*reports engine*[34]) jest serwisem wspierającym generowanie stron HTML lub plików wyjściowych zawierających graficzną reprezentację danych, które są edytowane przez analityków danych w celu stworzenia określonych raportów, a następnie wyświetlane na publicznych lub prywatnych stronach internetowych. Dane prezentowane w pojedynczym raporcie mogą pochodzić z wielu zapytań odnoszących się do różnych relacyjnych baz danych, a także plików płaskich – CSV i XLSX. Aby możliwe było zapanowanie nad strukturą tych zależności, architektura platformy analitycznej została podzielona na cztery warstwy abstrakcji widoczne na poniższym schemacie: konektorów, cache, zapytań i raportów (por. Ilustracja 3.1).

[34] Kreator raportów został również wykorzystany do stworzenia GENDERACTION Data Dashboard (genderaction-data-dashboard.opi.org.pl (dostęp: 02.03.2023)) w ramach projektu GENDer equality in the ERA Community To Innovate policy implementatiON. Coordination and Support Action, finansowanego w programie ramowym Horyzont 2020. Dashboard prezentuje statystyki dotyczące równości płci w Europie.



Ilustracja 3.1. Warstwy kreatora raportów

3.2.1. Warstwy architektury systemu

Każda warstwa odpowiada za zarządzanie jednym rodzajem obiektów. Odpowiedzialność poszczególnych warstw jest następująca:

1. **Warstwa konektorów:** umożliwia definiowanie i zarządzanie połączeniami do zewnętrznych źródeł danych. W przypadku relacyjnej bazy danych, konektorem jest pula połączeń do dowolnej bazy, na przykład Oracle lub MySQL.
2. **Warstwa cache:** pozwala na szybszy dostęp do danych. Platforma analityczna umożliwia wykonywanie raportów i zapytań bezpośrednio na bazach źródłowych, jednak dzięki posiadaniu lokalnej bazy H2, która może przechowywać kopie danych pochodzących z zewnętrznych źródeł, nie jest wymagane ciągłe odpytywanie serwerów źródłowych. Technicznie proces ten przeprowadza się tak, że w momencie otrzymania żądania utworzenia cache, system tworzy w bazie danych tabelę o tymczasowej losowej nazwie (np. TMP_36847234764). Następnie dane zwrócone przez zapytanie źródłowe są kopiowane do tej tabeli. Ze względu na fakt, że wynik zapytania nie zawiera żadnych informacji o metadanych tabel źródłowych, cache nie będzie posiadał żadnych kluczy głównych, obcych ani indeksów. W momencie gdy wszystkie dane zostaną prawidłowo skopiowane, nazwa tabeli jest zmieniana na właściwy identyfikator cache. W przypadku cache z odświeżaniem cyklicznym ten proces jest powtarzany. Dzięki temu podczas odświeżenia danych cache jest niedostępny tylko przez krótki moment usunięcia starej wersji tabeli i zmiany nazwy tabeli tymczasowej na docelową. Zanim cache odświeży się w pełni, możliwe jest korzystanie z jego poprzedniej wersji. Jeżeli odświeżenie z jakiegoś powodu się nie powiedzie, aktywna pozostanie wersja cache z ostatniego prawidłowego odświeżenia. Warto podkreślić, że z punktu widzenia systemu raportowego warstwa cache jest kolejnym typem konektora o identyfikatorze „cache". Definiowanie zapytań do niego odbywa się w taki sam sposób, jak w przypadku dowolnego innego konektora bazodanowego.

3. **Warstwa zapytań:** umożliwia definiowanie i wykonywanie zapytań. Obiekt zapytania składa się z odwołania do obiektu konektora, treści zapytania (SQL) oraz opcjonalnego zestawu nazwanych parametrów. Jeżeli zapytanie jest sparametryzowane, to wartości zdefiniowanych parametrów powinny być przekazywane podczas wykonywania zapytania. Dzięki temu można tworzyć filtry wykorzystywane w unikatowych raportach. Żądanie wykonania zapytania powinno również zawierać jedną lub więcej definicji tak zwanych wyjść *(*output), które instruują system, jak przetworzyć wynik zapytania. Najczęściej spotykane wyjścia obejmują zwrócenie danych w odpowiedzi HTTP, zapisanie danych do pliku oraz zwrócenie statystyk zapytania. Wyjścia pozwalają na bezpośrednią interakcję z warstwą zapytań. System wykorzystuje również wewnętrzne przekazywanie danych do kolejnej warstwy, czyli do raportu korzystającego z wyników zapytania.
4. **Warstwa raportów:** zawiera definicje raportów. Specyfikacja raportu jest obiektem złożonym, dla którego definiowany jest zbiór składający się z listy zapytań, które zostaną wykorzystane w raporcie, listy parametrów raportu, która pokrywa się z parametrami wcześniej zadeklarowanych zapytań oraz listy sekcji raportu. W procesie generowania raportu obiekty specyfikacji są przekształcane w obiekty wyników, zawierające dane do zaprezentowania użytkownikowi końcowemu. Kluczowy jest tutaj silnik renderujący (*front-end*), który przekształca obiekty wynikowe w gotowe strony prezentujące użytkownikom końcowym raporty z danymi.

Zazwyczaj w procesie przygotowania raportu, analityk danych jest odpowiedzialny za dostarczenie zapytania SQL lub pliku XLSX zawierającego odpowiednie dane. Na podstawie tego tworzony jest właściwy cache. Kolejne zapytanie z buforowanych danych przekształca prostą tabelę danych w dynamiczny raport z filtrami.

3.2.2. Warstwa raportów

Raport jest obiektem, który integruje elementy zdefiniowane w poprzednich warstwach aplikacji. Między tymi obiektami a obiektami zdefiniowanymi wewnątrz raportu, takimi jak parametry czy sekcje, powstają współzależności. Aby zrozumieć działanie warstwy raportów, konieczne jest zrozumienie tych współzależności i ich implikacji; zmiana jednego elementu raportu często wymaga zmiany innych elementów, które od niego zależą. Specyfikacja raportu składa się z następujących części:

1. **Lista zapytań:** pobierana ze zdefiniowanego źródła danych. Umieszcza się tutaj wszystkie zapytania pochodzące z warstwy zapytań, z których korzysta raport. Wynik uruchomienia pojedynczego zapytania może być wykorzystywany przez różne elementy raportu jednocześnie – na przykład służy do prezentacji wykresów i tabel oraz do uzupełnienia słowników dla parametrów raportu. Jeżeli kilka elementów korzysta z tego samego zapytania, to zostanie ono uruchomione tylko raz.
2. **Lista parametrów:** suma wszystkich parametrów zadeklarowanych zapytań. Na przykład jeśli raport zawiera dwa zapytania Q1 i Q2, gdzie zapytanie Q1 posiada parametry A, B i C, a zapytanie Q2 posiada parametry B, C i D, raport powinien zawierać definicję co najmniej czterech parametrów: A, B, C i D. Zestaw danych wymaganych do

uzupełnienia dla parametrów raportu jest szerszy niż w przypadku samych zapytań – należy między innymi uzupełnić etykietę parametru, wartość domyślną pokazującą się w momencie pierwszego uruchomienia raportu czy listę wartości słownikowych, które parametr może przyjąć.

3. **Lista sekcji:** elementy raportu widoczne dla użytkownika końcowego. Każda sekcja ma własną specyfikację (opis, w jaki sposób ma być prezentowana oraz z jakich danych ma korzystać) oraz reprezentację wynikową (wygenerowany zestaw informacji, na podstawie którego można zaprezentować sekcję użytkownikowi). Na przykład specyfikacja sekcji wykresu może zawierać informacje, że wykres posiada dwie serie, a dane do nich powinny być pobrane odpowiednio z zapytania A i zapytania B. W reprezentacji wynikowej zwrócona zostanie lista serii z już wypełnionymi danymi, ale bez informacji o zapytaniach, z których pochodzą te dane, jako nieistotnych dla użytkownika końcowego. Istnieje wiele rodzajów sekcji, zostaną one szczegółowo omówione w kolejnej części rozdziału.

Aby stworzyć gotowy do publikacji raport, w tej warstwie dodawane są rozmaite funkcjonalności, takie jak filtry, wizualizacje danych czy teksty. Poniżej znajduje się opis sekcji, które można wykorzystać do prezentacji danych.

3.2.3. Rodzaje sekcji

O strukturze każdego raportu przeglądanego przez użytkowników końcowych decydują analitycy danych. Ponieważ różne typy sekcji pełnią różne funkcje, to każda sekcja zawiera indywidualny zestaw pól, które kontrolują jej zachowanie i sposób prezentacji. Kreator raportów systemu RAD-on zawiera następujące sekcje:

1. **Sekcja tekstów (TEXT):** służy do prezentacji panelu z opisem tekstowym (np. wstęp). Sekcja akceptuje kod HTML, zatem możliwe jest dowolne formatowanie tekstu.
2. **Sekcja filtrów (FILTERS):** służy do interakcji z raportem poprzez filtrowanie jego parametrów. Specyfikacja sekcji zawiera informacje o tym, jakie typy elementów powinny być wyświetlane (np. wielokrotny wybór, tekst, filtrowanie selektywne, autouzupełnianie) oraz z jakimi parametrami raportu są połączone.
3. **Sekcja wykresów (CHART):** służy do prezentacji różnych wizualizacji (wykresy – kolumnowe, powierzchniowe, punktowe, kołowe, radarowe; mapy). Zawierają one dane z zapytań zadeklarowanych w liście zapytań. Wizualizacja może posiadać jedną lub więcej serii danych, a każda seria może pochodzić z innego zapytania. Seria danych jest definiowana jako lista punktów typu (x, y), gdzie x jest etykietą określonego punktu, a y – wartością punktu. Niektóre wykresy wspierają dodatkowy wymiar danych (x, y, z) i w takiej sytuacji wartość z może być prezentowana na przykład w postaci rozmiaru bąbla na wykresie bąbelkowym. Specyficznym rodzajem wykresów są mapy, zawierające wartości liczbowe indeksowane nazwami lub kodami obszarów geograficznych (państwa, województwa, powiaty).
4. **Sekcja tabeli (TABLE i CHART_TABLE):** służy do prezentacji danych pochodzących z zapytania w formie tabelarycznej. Dla TABLE wyświetlane są dane identyczne z tymi, które zostały zwrócone przez zapytanie, natomiast dla CHART_TABLE dane są

pobierane bezpośrednio z wykresu. W drugim przypadku kolumnami w tabeli będzie lista etykiet odpowiedniego wykresu i wartości poszczególnych serii danych; opcjonalnie wyświetlić można dodatkową kolumnę sumującą wartości wszystkich serii.

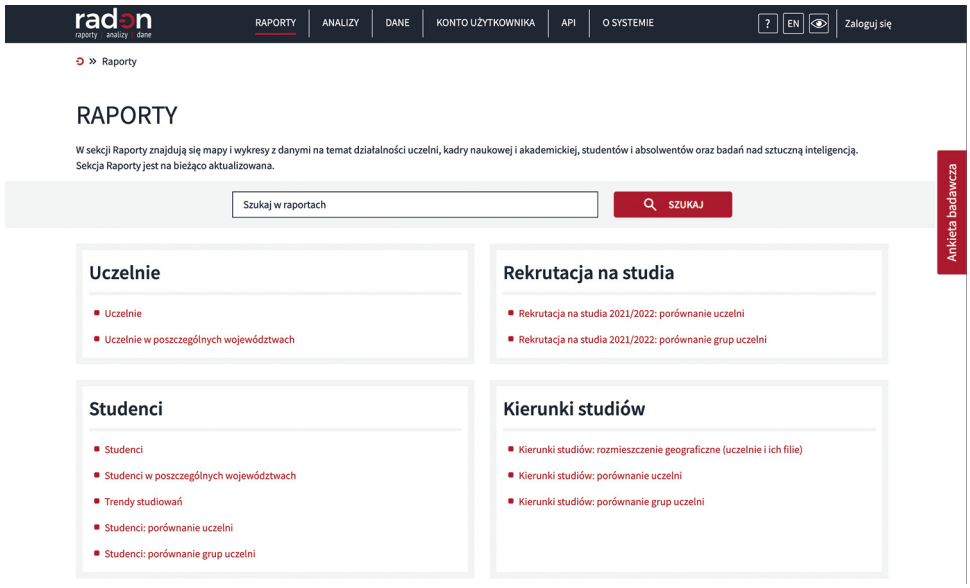
5. **Sekcja podsumowania (SUMMARY):** służy do wyświetlania wartości liczbowych syntetyzujących raport, pochodzących z odpowiednich kolumn w zapytaniach zadeklarowanych w liście zapytań. Na kaflach podsumowujących mogą znajdować się na przykład wzrosty lub spadki określonych wartości w czasie, zatem dobrą praktyką jest przypisywanie im zapytań zwracających pojedynczy rekord ze zagregowaną wartością.

Podsumowując, proces tworzenia raportów jest dość prosty. Do przygotowania zbiorów danych służą analitykowi zapytania SQL lub nawet proste pliki CSV. Najważniejszą częścią jest parametryzacja zapytań, dzięki którym użytkownicy końcowi raportu mogą wchodzić w interakcję z danymi. Wykorzystując warstwy i sekcje, analityk może skonstruować wiele różnych raportów i w efekcie wdrożyć na stronę internetową gotowy do publikacji dashboard. Jego przykład, czyli raporty z systemu RAD-on przedstawimy na kolejnych stronach.

3.3. Raporty w RAD-onie jako narzędzie rozpowszechniania informacji i podejmowania decyzji

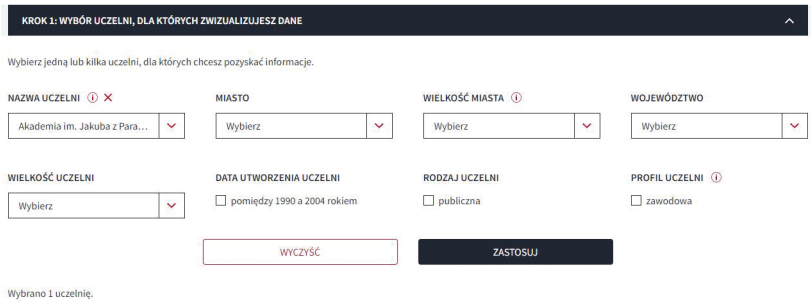
Dashboard można zdefiniować jako „wizualne przedstawienie najważniejszych informacji potrzebnych do osiągnięcia pewnych celów; skonsolidowane i rozmieszczone tak, aby możliwy był monitoring danych na pierwszy rzut oka" ([19], s. 26) lub „zestaw wizualnych zasobów, które umożliwiają odbiorcom zrozumienie wyświetlanych danych i/lub wglądu w nie" [59]. Wśród najważniejszych jego zalet wymienia się między innymi zapewnienie szybkiego wglądu w zróżnicowane dane, z jednoczesnym szczegółowym badaniem zjawisk (*drill into detail*), możliwość analizowania trendów w czasie, wreszcie – ułatwianie podejmowania decyzji na poziomie instytucjonalnym i strategicznym [35].

W kontekście powyższych walorów opiszemy teraz dwa przykładowe raporty zaprojektowane przy zastosowaniu opisanego wyżej kreatora i zaimplementowane w systemie RAD-on: „Studenci: porównanie uczelni" i „Studenci: porównanie grup uczelni", znajdujące się w folderze „Studenci" (por. Ilustracja 3.2). Raporty przedstawione zostaną z perspektywy użytkowników, ale pokażemy również, jakie czynności musi wykonać analityk, aby osiągnąć efekt końcowy w postaci dynamicznego raportu.



Ilustracja 3.2. Moduł „Raporty" portalu RAD-on

Założeniem obu raportów jest umożliwienie porównania bądź poszczególnych uczelni między sobą, bądź grup uczelni, w odniesieniu do charakterystyk studentów. Użytkownik dokonuje wyboru interesujących informacji w kilku krokach. Najpierw definiowane są uczelnie, dla których zwizualizowane zostaną dane. Dostępne są następujące filtry: 1) nazwa uczelni; 2) miasto będące siedzibą uczelni; 3) województwo; 4) wielkość miasta mierzona liczbą mieszkańców; 5) wielkość uczelni mierzona liczbą studentów; 6) data utworzenia uczelni (do 1939 roku; w latach 1940–1989; w latach 1990–2004; od 2005 roku); 7) rodzaj uczelni (publiczna; niepubliczna, kościelna) oraz 8) profil uczelni (akademicka; zawodowa). Filtry są od siebie zależne, co oznacza, że wybierając określoną uczelnię i wciskając przycisk „zastosuj", użytkownik dostaje jednocześnie odfiltrowane informacje na jej temat. Na przykład, gdy interesuje go Akademia imienia Jakuba z Paradyża (por. Ilustracja 3.3), dowie się, że jest to zawodowa uczelnia publiczna, utworzona w latach 1990–2004, w której kształci się powyżej 1 tys. do 5 tys. studentów, mająca siedzibę w Gorzowie Wielkopolskim, czyli mieście w województwie lubuskim liczącym powyżej 100 tys. do 250 tys. mieszkańców. Analogicznie, jeśli w filtrze „Rodzaj uczelni" zaznaczy tylko wyższe szkoły niepubliczne, to w odpowiedzi otrzyma wszystkie informacje na ich temat.



Ilustracja 3.3. Przykład filtrów zależnych

W przypadku pozyskiwania danych dla Akademii im. Jakuba z Paradyża, jeśli użytkownik przefiltruje dane wyłącznie dla studentów pierwszego roku, to otrzyma osiem wizualizacji w postaci wykresów. Pokazują one ogólną liczbę studentów, a następnie odpowiednio udziały studentów w podziałach według płci, miejsca zamieszkania, pochodzenia, wieku oraz trybu, poziomu i profilu studiów. W przypadku raportu do porównywania grup uczelni podstawowa różnica polega na tym, że wartości wskaźników są wyliczane wspólnie dla całej grupy; wyliczana jest również średnia dla grup uczelni. Na wykresach z udziałami po najechaniu myszką na poszczególne kolumny wykresów ukażą się odpowiadające udziałom wartości liczbowe. Ponadto, klikając na wpisy legendy można drążyć w danych, czyli wyłączać lub włączać podgląd określonych danych na wykresie, poszerzając lub zawężając zakres wyświetlanych informacji.

Czynności analityczne polegały przede wszystkim na zdefiniowaniu w kreatorze wszystkich niezbędnych sekcji tekstowych, czyli paneli pozwalających wyodrębnić widoki z określonym typem informacji (na Ilustracji 3.3 jest to pasek tematyczny opisujący krok pierwszy, ale raporty zawierają również panele wstępu, filtrów i objaśnień metodologicznych). Określono również, które panele mają być rozwinięte (*expanded*) w widoku głównym oraz czy powinny być rozwijalne (*expandable*). W drugiej kolejności ustanowiono wszystkie potrzebne filtry i dodano podpowiedzi (*tooltips*) do niektórych z nich (na Ilustracji 3.3 widać tooltip przy filtrze „profil uczelni"; opisuje on cechy uczelni akademickich i zawodowych, które przynależą do tego filtra). Trzeci krok analityczny polegał na utworzeniu wizualizacji i obejmował między innymi wybór zapytania, z którego czerpane są dane do wykresu, wybór typu wizualizacji (*horizontal*) oraz wybór sposobu agregacji danych (suma, średnia). Analityk określił również, czy na wykresach należy pokazywać wartości (*hide values*) i kategorie w porządku odwróconym (*reverse order*), a ponadto nadał tytuły zakładkom i wykresom, zdefiniował, co umieścić na osi X i Y, wpisał nazwy etykiet osi (*label*) oraz minimalne (*min*) i maksymalne (*max*) wartości przedstawiane na osi, określił, czy wartości mają być skumulowane (*stacked*) oraz czy osi wykresu powinny być przewijane (*scrollbar*), a także dobrał odpowiednie kolory kolumn.

Sekcja raportów RAD-onu jest stale wzbogacana, obecnie zawiera zbiór raportów o uczelniach, rekrutacji na studia, studentach, trendach w studiowaniu, absolwentach, doktorantach i nauczycielach akademickich. Wszystkie raporty udostępniane są w wer-

sji polskiej i angielskiej. Dane będące podstawą wizualizacji mogą być pobierane w formatach XLSX lub CSV, a same wykresy ściągać można jako obrazki w formacie PNG. Ponadto, sukcesywnie do poszczególnych raportów dodawane są komentarze eksperckie, zawierające skróconą interpretację danych, a także odniesienia do zewnętrznych dokumentów i analiz dotyczących określonych tematów. Komponenty raportów stanowią zatem wsparcie przy tworzeniu własnych, spersonalizowanych zestawień czy prezentacji, mogących być podstawą podejmowania decyzji na różnych poziomach.

3.3.1. Wyzwania w tworzeniu raportów

Nie negując niewątpliwych zalet pulpitów nawigacyjnych, trzeba zauważyć również, że ich projektowanie i obsługa jest skomplikowanym procesem. Jeśli nie przezwyciężymy pojawiających się ryzyk, stracimy zaufanie użytkowników. Dashboardy oceniane jako niewiarygodne stają się nieprzydatne. Na przykładach konkretnych raportów RAD-onu przedstawimy teraz najistotniejsze z wyzwań [35], a także sposoby radzenia sobie z nimi.

Podstawową zaletą systemu RAD-on jest wysoka jakość danych. W większości przypadków podstawą raportów są dane wprowadzane przez uczelnie do systemu POL-on na potrzeby sprawozdawczości Głównego Urzędu Statystycznego (GUS), są więc częścią statystyki publicznej w Polsce. Uznaje się, że jakość danych i stosowanych wskaźników wywiera kluczowy wpływ na proces podejmowania przez użytkowników decyzji o tym, z jakiego narzędzia korzystać [12, 23, 50]. Fakt, że właścicielem większości danych jest Ministerstwo Edukacji i Nauki, a ich administratorem OPI PIB, neutralizuje ryzyko rozproszonej odpowiedzialności za jakość przekazywanych informacji [35].

Wysoka jakość danych wyjściowych jest ważna, jednak na kolejnym etapie pojawia się wyzwanie w postaci czyszczenia danych i przygotowania ich do analizy [14]. W przypadku raportów RAD-onu zajmuje się tym kilkuosobowy Zespół Data Science w Laboratorium Baz Danych i Systemów Analityki Biznesowej, w skład którego wchodzą analitycy, statystycy, programiści oraz naukowcy prowadzący badania dotyczące obszaru nauki i szkolnictwa wyższego. Ponadto, wsparcia w dbałości o jakość przekazywanych danych udziela również Zespół Raportowania i Analiz z tego samego laboratorium. Połączenie wiedzy technicznej z ekspercką wiedzą o sektorze nauki i szkolnictwa wyższego przynosi efekt w postaci dobrze przygotowanych zapytań do baz danych, a następnie wiarygodnych raportów.

Osobną sprawą pozostaje kwestia dostarczenia użytkownikom tak zaprojektowanych raportów, by mogli je bezproblemowo obsługiwać użytkownicy z różnych grup, między innymi decydenci w obszarze polityki naukowej, zarządzający uczelniami i innymi instytucjami naukowymi, dziennikarze, przyszli studenci etc. (*user-centered design*; [37]). Kwestia, czy można tworzyć dashboardy dla wszystkich (*one-size-fits-all*; [54]), jest otwarta, natomiast reguła „dostosowuj, personalizuj, adaptuj" [59] wciąż obowiązuje i oznacza kompleksowe podejście. W przypadku raportów RAD-onu oznacza to po pierwsze objaśnienie używanych wskaźników i terminów (np. różnice między uczel-

niami akademickimi i zawodowymi są oczywiste dla urzędników ministerstwa, ale nie muszą być dla pozostałych odbiorców). Po drugie, mimo że raporty są konsultowane przez specjalistów UX, problemem pozostaje brak wytycznych w projektowaniu pulpitów z danymi z sektora publicznego [1]. Na przykład wprowadzone przez nas filtry zależne (Ilustracja 3.3) mogą być nieintuicyjne dla osób przyzwyczajonych do częstego odwiedzania stron sklepów internetowych czy systemów rezerwacji hoteli, jednak w naszej opinii są dodatkowym źródłem informacji i rezygnacja z nich byłaby błędem.

Również problem z niewłaściwą interpretacją informacji [35] można przezwyciężyć, wprowadzając dodatkowe podpowiedzi dla użytkowników. Zespół Data Science opracowuje szczegółowe objaśnienia metodologiczne ułatwiające zrozumienie raportów (np. przekazywana jest informacja o tym, że studenci uczący się na kilku kierunkach liczeni są kilkukrotnie), wraz z odniesieniami do zewnętrznych źródeł (np. strony pomocy systemu POL-on). Wgląd w dane ułatwiają także sukcesywnie dodawane komentarze eksperckie oraz – tam gdzie jest to wskazane – przedstawianie informacji w różnej formie (np. wizualizacje i tabele). Poza tym na bieżąco analizuje się i uwzględnia sugestie użytkowników, które pojawiają się już po wdrożeniu raportów.

Kolejne ryzyko wiąże się z aktualizacją dashboardów [4], w tym ograniczonymi zasobami kadrowymi i czasowymi potrzebnymi do kontrolowania tego procesu. Samo przygotowanie raportu do wdrożenia produkcyjnego, zwłaszcza gdy ma on odmienną formę od raportów już istniejących, jest długotrwałym procesem analitycznym, czasem liczonym w miesiącach. Następnie należy sprawnie poprawiać ewentualne błędy, a także wprowadzać aktualizacje, bez których użytkownicy uznają dashboard za mało wiarygodny i nieprzydatny. Raporty w RAD-onie aktualizowane są jedynie raz w roku, przedstawiając stan na 31 grudnia (po sprawozdaniu się uczelni do GUS oraz weryfikacji tego procesu przez resort nauki) i chociaż cały proces jest do pewnego stopnia zautomatyzowany, to wyzwaniem jest dostosowywanie się do zmieniającego się porządku prawnego (np. gdy w wyniku nowego rozporządzenia ministra nauki do klasyfikacji dziedzin i dyscyplin dodane są nowe dyscypliny, filtr analitycy muszą „przepisać" istniejące wcześniej dyscypliny). Czasem w wyniku zachodzących zmian rodzą się nowe pytania, które mogą wymagać dodatkowych danych, a to ponownie zwiększa ilość i czas pracy.

Konserwacji i aktualizacji wymaga także silnik – kreator raportów. We współpracy z zespołem deweloperów podejmowane są działania w celu usprawniania jego funkcjonalności tak, aby z jednej strony ułatwiać pracę analitykom a z drugiej – aby dostarczać lepszych informacji użytkownikom. Przykładem drugiego ulepszenia jest dodanie mapy powiatów, która w większym stopniu niż mapa województw pozwala uchwycić specyfikę szkolnictwa wyższego w Polsce – zamiast powszechnie znanej informacji, że najwięcej uczelni jest w województwie mazowieckiem, można pokazać uczelnie w poszczególnych powiatach, z uwzględnieniem ich filii, a to daje bardziej zniuansowany obraz sektora. Należy przy tym zdawać sobie sprawę z przyzwyczajania się użytkowników do wyglądu interfejsów i niechęci do zbyt często wprowadzanych zmian funkcjonalności [65].

Ostatnie, kluczowe ryzyko polega na ograniczonym wykorzystaniu dashboardów przez użytkowników, dla których zostały zaprojektowane [35]. Raporty RAD-onu są szeroko promowane przez Ośrodek Przetwarzania Informacji – Państwowy Instytut Badaw-

czy w środkach masowego przekazu, mediach społecznościowych oraz podczas konferencji naukowych i branżowych. Najważniejsze zadanie Zespołu Data Science polega jednak na tym, by dostarczać użytkownikom raporty dobrze zaprojektowane, opisane i łatwe w obsłudze.

W niniejszym rozdziale przedstawiliśmy alternatywę dla standardowych aplikacji *business intelligence*, które mogą być budowane i wykorzystywane przez poszczególne organizacje. Ich znaczenie cały czas rośnie, jednak w pewnych przypadkach, opisanych poniżej, platformy analityczne budowane przez zespoły IT mogą być korzystniejszym wyborem.

Po pierwsze, platformy analityczne są lepsze w sytuacjach, gdy potencjalna liczba użytkowników końcowych jest duża i trudna do przewidzenia, na przykład gdy planowane jest opublikowanie dashboardu z otwartymi danymi na publicznej stronie internetowej. Przy nieograniczonej liczbie odbiorców licencjonowane komercyjne narzędzia BI są znacznie droższe.

Po drugie, platformy analityczne sprawdzają się, gdy konieczne jest dostosowanie się do rozwijających się potrzeb użytkowników końcowych. Każdy rodzaj oprogramowania BI zawiera unikatowy zestaw funkcji i zazwyczaj ta lista jest zamknięta. Platformy analityczne są elastyczne i rozszerzalne, a to pozwala łatwo dostosować się do zachodzących zmian.

Po trzecie, platformy analityczne są znacznie bardziej użyteczne, gdy kluczowy jest element wizualny dashboardu, na przykład, gdy musi on odpowiadać wyglądowi nadrzędnej strony internetowej. Większość narzędzi BI oferuje ograniczone układy graficzne, natomiast platformy analityczne łatwiej jest zintegrować z istniejącymi projektami.

Wszystkie powyższe sytuacje znajdują zastosowanie dla raportów systemu RAD-on omawianych w tym rozdziale.

ROZDZIAŁ 4

HURTOWNIA I MODEL WYMIANY DANYCH

Łukasz Błaszczyk
Sylwia Ostrowska
Emil Podwysocki

4.1. Model wymiany danych

Jednym z głównych celów projektu było zwiększenie dostępu do danych z obszaru nauki i szkolnictwa wyższego, zgodnie z ideą otwartych danych rządowych (*open government data*, OGD). W sytuacji gdy dane z sektora są rozproszone (tzn. istnieje wiele odrębnych systemów, baz i mikroserwisów, przetwarzających dane w różnych celach), realne zwiększenie dostępu do danych wymaga nie tylko udostępnienia obywatelom większej liczby danych z oficjalnych źródeł, ale przede wszystkim udostępnienia ich w jednym miejscu. Zgodnie z założeniami FAIR (więcej w rozdziale 1.1), dane powinny być łatwe do znalezienia, dostępne, interoperacyjne i wielokrotnego użytku (*findable, accessible, interoperable, reusable*), a zatem konieczna jest także ich integracja i ujednolicenie. W przypadku systemu RAD-on, warstwę integracyjną (obok hurtowni danych) stanowi model wymiany danych (MWD) – centralny element systemu, którego zaprojektowanie i wdrożenie było kluczowe dla powodzenia projektu. W tej części przedstawimy proces wyboru właściwej technologii dla modelu wymiany danych, a także jego założenia i architekturę, uzupełnione o przykłady praktycznych zastosowań.

4.1.1. Technologia

Model wymiany danych odpowiada za integrację systemów źródłowych, a także za udostępnianie danych na potrzeby serwisów maszynowych (np. publiczne API RAD-onu zapewniające dostęp do danych z systemów źródłowych), usług dostępnych na portalu RAD-on (np. publiczne zestawienia danych) oraz innych systemów korzystających z danych. MWD uczestniczy pośrednio w realizacji wszystkich usług systemu, dlatego wybór właściwej technologii był kluczową decyzją techniczną na początkowym etapie projektu. W celu doboru odpowiedniego rozwiązania, przeprowadziliśmy analizę trzech technologii, które potencjalnie mogły zostać wykorzystane do realizacji MWD:

- Mule Enterprise Service Bus (ESB)[35];
- Spring Integration[36];
- Apache Kafka[37].

Każda z technologii została zbadana pod kątem cech takich, jak: stabilność, skalowalność, ograniczenia programistyczne, łatwość rozwoju, łatwość wdrażania, dokumentacja, asynchroniczność, integracja z bazami danych, integracja z różnymi web serwisami, integracja z Apache Hadoop[38], monitoring (narzędzia graficzne), globalne logowanie i monitorowanie zapytań oraz monitorowanie dostępności serwisów. Analiza została poparta implementacją praktycznych rozwiązań, co pozwoliło na ocenę możliwości zastosowania w tym konkretnym projekcie. W wyniku przeprowadzonej analizy zdecydowaliśmy się na odejście od kanonicznego modelu danych (w rozumieniu architektury opartej na usługach – *Service-Oriented Architecture*, SOA), w kierunku komunikacji z użyciem systemu kolejkowego Apache Kafka.

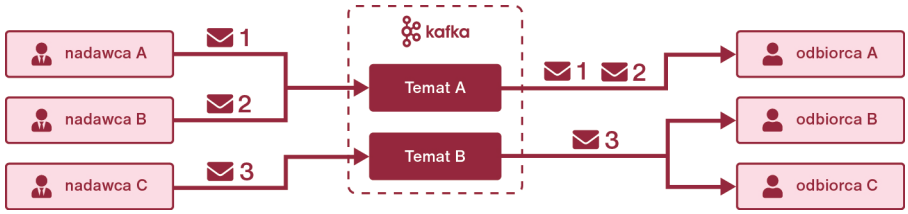
Apache Kafka działa w modelu publicznej subskrypcji (*public subscribe*), który pośredniczy w wymianie komunikatów między aplikacjami (systemami) i umożliwia ich propagację w czasie rzeczywistym. System został pierwotnie zaprojektowany dla serwisu LinkedIn[39], aby służyć jako scentralizowana platforma potokowania zdarzeń do zadań integracji danych online [58]. Aktualnie technologia Apache Kafka jest dostępna jako otwarte oprogramowanie. Ponieważ może działać na dowolnej liczbie serwerów, niezależnie od ich lokalizacji, cechuje ją również odporność na awarie. Integracja oparta na Apache Kafka zapewnia wydajną i bezpieczną komunikację między modułami i systemami. Posiada wbudowane mechanizmy zapewniające replikację i ochronę wiadomości przed utratą w środowisku produkcyjnym (por. [26, 33, 64]). W kontekście systemu RAD-on niewątpliwą zaletą Apache Kafka jest również jego skalowalność i imponujący katalog praktycznych zastosowań.

4.1.2. Założenia modelu wymiany danych

Model wymiany danych został zbudowany z wykorzystaniem otwartego oprogramowania Apache Kafka. Aplikacja ta umożliwia publikowanie i subskrybowanie danych poprzez kolejkę uporządkowanych wiadomości. Aplikacja odbiera rekordy danych, a następnie przechowuje je w taki sposób, aby każdy zainteresowany ich pobraniem system miał do nich dostęp. Tworzy kolejkę wiadomości, porządkując odebrane rekordy według kolejności wysłania ich z systemu źródłowego. Każdy rekord danych, nazywany wiadomością (*message*) zawiera klucz, wartość oraz znacznik czasu, które umożliwiają systemom pobierającym dane zorientować się, czy wiadomość wymaga odczytania

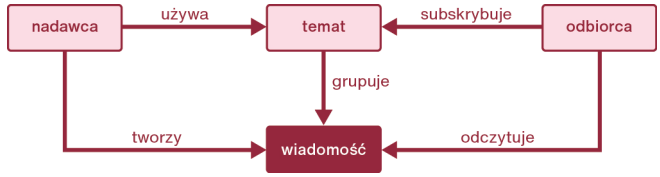
[35] mulesoft.com/resources/esb/what-mule-esb (dostęp: 02.03.2023)
[36] spring.io/projects/spring-integration (dostęp: 02.03.2023)
[37] kafka.apache.org (dostęp: 02.03.2023)
[38] hadoop.apache.org (dostęp: 02.03.2023)
[39] about.linkedin.com/pl-pl (dostęp: 02.03.2023)

i przetworzenia. Pojedyncza wiadomość wysłana przez system źródłowy jest porcją danych (tablicą bajtów), która ma znaczenie dla odbiorców. Odbiorcy mogą ją przechować lub przetworzyć w celu realizacji własnych zadań. Apache Kafka pośredniczy w wymianie wiadomości pomiędzy aplikacjami (systemami) i umożliwia ich propagację w czasie rzeczywistym. Na Ilustracji 4.1 został przedstawiony schemat działania rozwiązania opartego na Apache Kafka wraz z wyszczególnieniem komponentów uczestniczących w wymianie wiadomości.



Ilustracja 4.1. Ogólny schemat działania modelu wymiany danych opierającego się na Apache Kafka

Apache Kafka łączy dwa modele wymiany danych pomiędzy systemami rozproszonymi. Po pierwsze, jest to podejście, w którym dane udostępniane są w postaci uporządkowanej kolejki wiadomości, umożliwiającej odczyt każdego rekordu przez dowolną liczbę odbiorców. Po drugie, Kafka integruje model komunikacji oparty na kolejkach z modelem subskrypcyjnym. Model subskrypcyjny zakłada, że każdy system zainteresowany otrzymywaniem danych z systemu źródłowego może subskrybować wiadomości udostępniane przez Apache Kafka. Subskrypcja umożliwia przekazywanie danych tylko tym systemom, które z nich korzystają lub je przetwarzają. Dodatkowo Apache Kafka umożliwia przesyłanie tej samej wiadomości do wielu grup odbiorców (*consumer groups*). Apache Kafka zapewnia mechanizmy umożliwiające przetwarzanie tej samej wiadomości przez wielu odbiorców równolegle. Główne terminy używane w kontekście aplikacji tej technologii zostały przedstawione na Ilustracji 4.2.

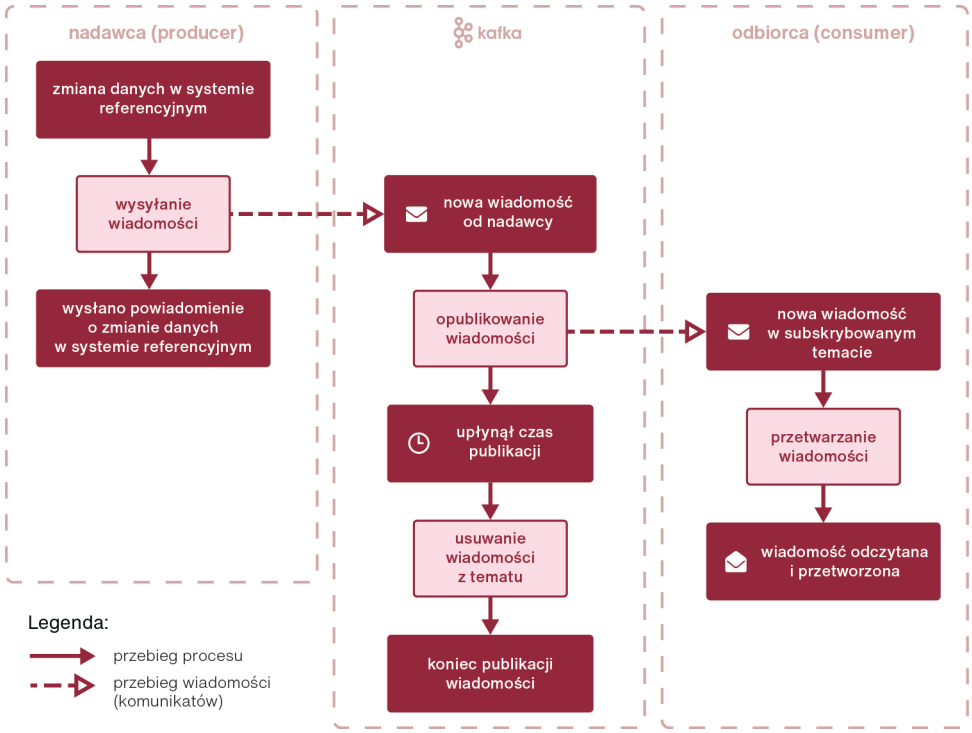


Ilustracja 4.2. Podstawowe klasy obiektów występujące w modelu wymiany danych opierającym się na Apache Kafka

Aplikacja, która odczytuje komunikaty, nazywana jest odbiorcą lub konsumentem (*consumer*). Aplikacja, która tworzy wiadomości, określana jest mianem producenta lub nadawcy (*producer*). Ważnym założeniem architektury opartej na Apache Kafka jest to, że komunikaty nie mają z góry zdefiniowanego odbiorcy, który byłby zainteresowany wiadomością. Główną odpowiedzialnością nadawcy jest sklasyfikowanie wiadomości poprzez przypisanie jej do odpowiedniego tematu, a następnie wysłanie do Apache

Kafka. Nadawca i odbiorca powiązani są ze sobą poprzez temat konwersacji (*topic*). Temat grupuje wiadomości, które z punktu widzenia odbiorców są ze sobą powiązane lub dotyczą podobnych zagadnień dziedzinowych. Z jednego tematu może korzystać wielu odbiorców, co jest jedną z zalet integracji poprzez Apache Kafka. Jeżeli w ramach integracji systemów zostanie utworzony temat, każdy zainteresowany system może dokonać subskrypcji i odbierać komunikaty.

Architektura wykorzystująca Apache Kafka opiera się na centralnym punkcie komunikacji, gdzie publikowane są wiadomości dostępne dla wszystkich zainteresowanych systemów. Rozwiązanie takie zapobiega wzajemnemu odpytywaniu systemów oraz umożliwia systemom dynamiczne reagowanie na zmiany mające miejsce w systemie referencyjnym. Wiadomości mogą zawierać rekordy danych, który umożliwiają aktualizację danych w systemach zależnych bez konieczności cyklicznego pobierania przez nie pełnego zestawu danych i żmudnego analizowania różnic. Niewątpliwą zaletą Apache Kafka jest skalowalność; jest to rozwiązanie sprawdzone w dużych systemach rozproszonych, dostępnych jako otwarte oprogramowanie. Technologia jest odporna na awarie, ponieważ może być uruchamiana na dowolnej liczbie serwerów znajdujących się w różnych lokalizacjach. Integracja oparta na Apache Kafka zapewnia mechanizm wydajnej i bezpiecznej komunikacji pomiędzy modułami lub systemami. Ponadto, wbudowane mechanizmy zapewniają replikację i zabezpieczenie wiadomości przed ich utratą w środowisku produkcyjnym. Na Ilustracji 4.3 pokazano schemat przepływu danych pomiędzy uczestnikami komunikacji z wykorzystaniem Apache Kafka.



Ilustracja 4.3. Ogólny schemat przepływu danych w modelu wymiany danych opierającym się na Apache Kafka

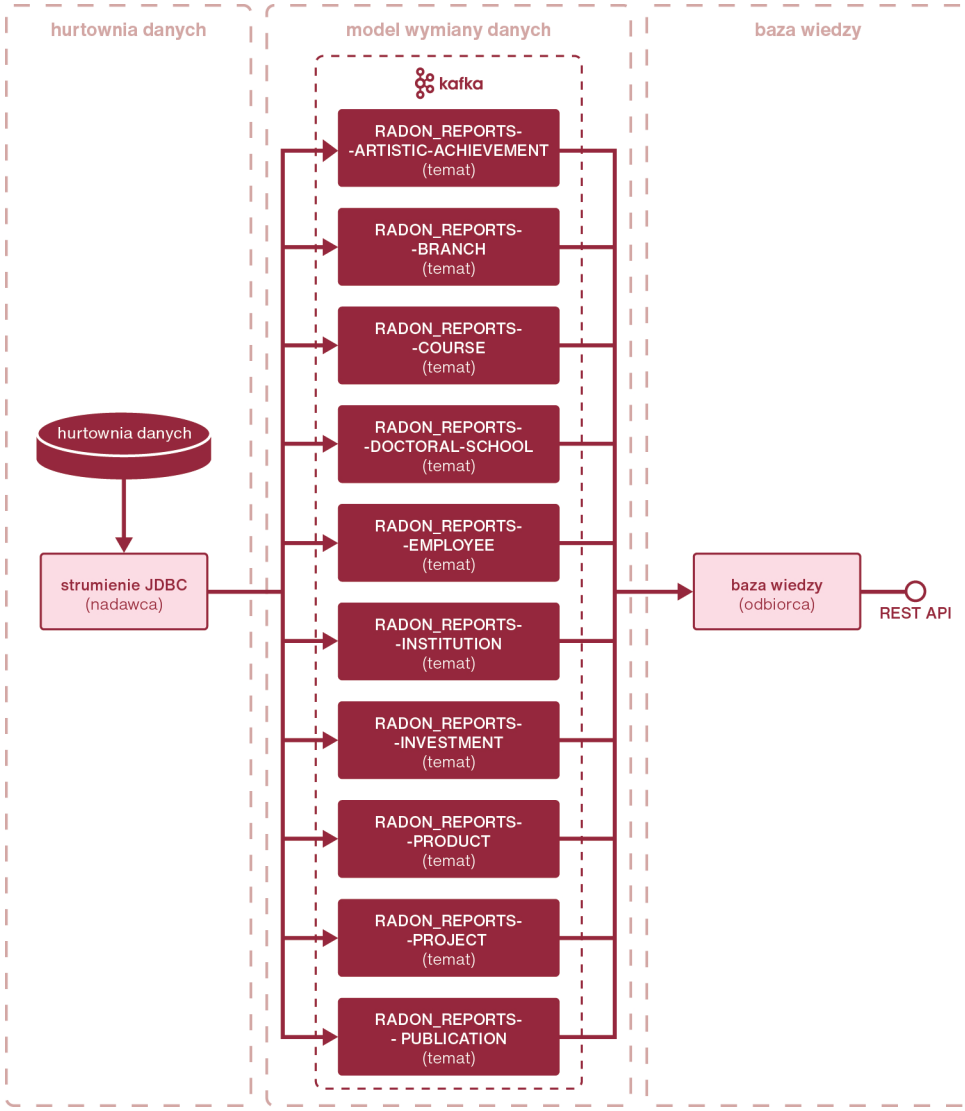
4.1.3. Zastosowanie modelu wymiany danych

Jedną z przewag modelu wymiany danych opartego o technologię Apache Kafka jest szeroki zakres możliwości jego praktycznego zastosowania. Dzięki wdrożeniu rozwiązania o opisanej powyżej architekturze i założeniach, mamy możliwość obsługi wielu procesów z różnych obszarów. Aby zobrazować wszechstronność MWD, w tej części omówimy trzy różne przykłady jego wykorzystania:

1. Udostępnienie danych publicznych (na portalu RAD-on oraz za pomocą otwartego API).
2. Udostępnienie danych systemowi dziedzinowemu.
3. Wykorzystanie MWD w procesie obsługi komunikatów administracyjnych.

ZASTOSOWANIE MWD DO UDOSTĘPNIANIA PUBLICZNYCH ZASOBÓW NAUKI I SZKOLNICTWA WYŻSZEGO

Model wymiany danych odgrywa kluczową rolę w procesie udostępniania zasobów nauki i szkolnictwa wyższego na portalu RAD-on. Użytkownicy portalu mogą korzystać z różnych usług i form udostępniania danych, w zależności od swoich potrzeb i poziomu umiejętności analitycznych. Portal umożliwia między innymi przeglądanie zestawień danych z rozbudowaną sekcją filtrów, możliwością pobrania plików CSV/XLSX oraz wyszukiwanie pełnotekstowe. Dane są również dostępne do pobrania maszynowego za pomocą publicznych usług API (architektura portalu oraz zakres usług opisane są w rozdziale 2). Zarówno wyszukiwarka, jak i powiązane z nią usługi sieciowe zostały zbudowane na podstawie modelu wymiany danych. Dane źródłowe, które podlegają publicznemu udostępnieniu, są agregowane przez hurtownię danych, a następnie przekształcane na potrzeby modelu wymiany danych. MWD udostępnia dane poszczególnym usługom portalu. Szczegółowy opis tego procesu oraz architektura rozwiązania zostaną szerzej omówione w rozdziale 5. Ilustracja 4.4 przedstawia poglądowy schemat przepływu danych z systemów dziedzinowych do hurtowni danych i do bazy wiedzy portalu RAD-on.

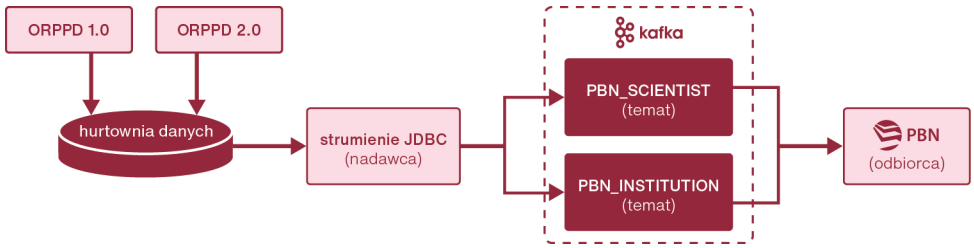


Ilustracja 4.4. Schemat przepływu danych pomiędzy hurtownią danych a bazą wiedzy portalu RAD-on z wykorzystaniem modelu wymiany danych

ZASTOSOWANIE MWD DO REJESTRACJI INFORMACJI O OSIĄGNIĘCIACH NAUKOWYCH W SYSTEMIE PBN

Polska Bibliografia Naukowa[40] (PBN) to portal Ministerstwa Edukacji i Nauki (MEiN) gromadzący informacje o publikacjach polskich naukowców, dorobku publikacyjnym instytucji naukowych oraz o polskich i zagranicznych czasopismach. Stanowi on część Zintegrowanego Systemu Informacji o Nauce i Szkolnictwie Wyższym POL-on[41]. Dorobek publikacyjny rejestrowany w systemie PBN jest powiązany z instytucjami szkolnictwa wyższego i nauki (m.in. uczelniami i instytucjami naukowymi) oraz ich pracownikami. Dane instytucji i pracowników gromadzone są w odrębnych modułach systemu POL-on.

W celu zapewnienia systemowi PBN dostępu do aktualnych i poprawnych danych w procesie rejestracji informacji o publikacjach naukowców i dorobku naukowym instytucji, został wdrożony mechanizm udostępniania danych z zastosowaniem modelu wymiany danych. Zaletą tego rozwiązania jest brak konieczności odpytywania o dane poszczególnych mikroserwisów POL-onu. Dane źródłowe dotyczące pracowników i instytucji są agregowane przez hurtownię danych. Następnie hurtownia, za pomocą rozwiązania o nazwie JDBC Streams (więcej w rozdziale 5.3.1), przekształca zagregowane dane na potrzeby modelu wymiany danych. Przekształcone i opublikowane dane są odczytywane i wykorzystywane przez PBN, w celu powiązania publikacji z pracownikami i instytucjami zarejestrowanymi w systemie POL-on. Schemat przepływu danych z systemów dziedzinowych do hurtowni danych i systemu PBN został przedstawiony na Ilustracji 4.5.



Ilustracja 4.5. Schemat przepływu wiadomości z wykorzystaniem modelu wymiany danych pomiędzy systemami dziedzinowymi a systemem PBN 2.0

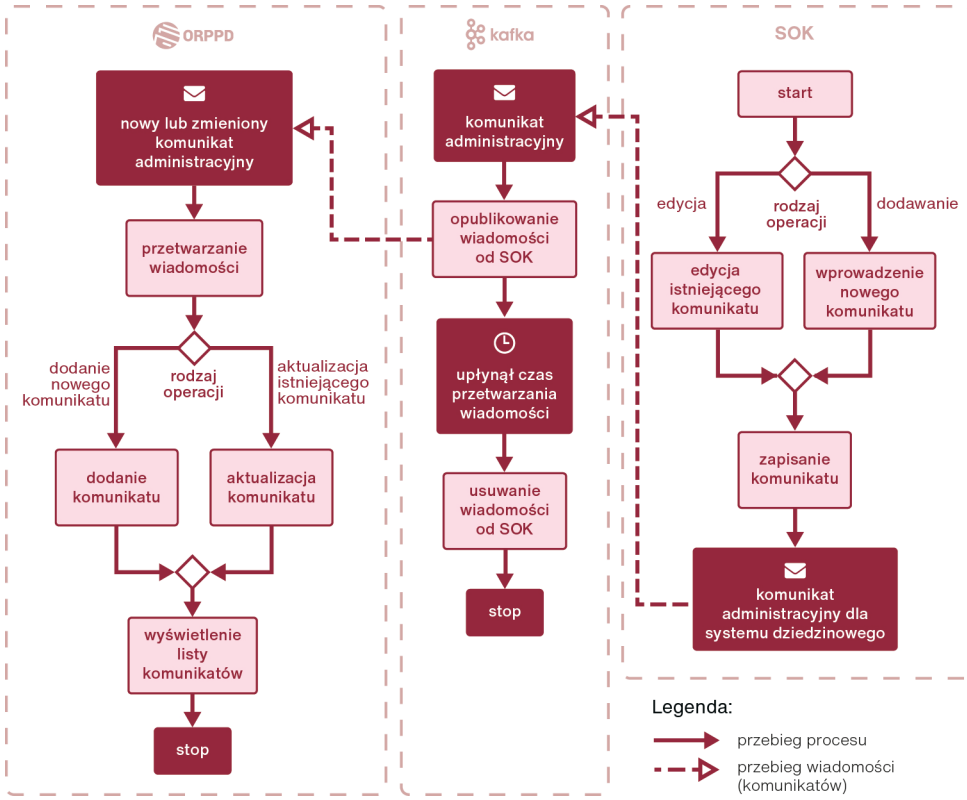
ZASTOSOWANIE MWD DO WYMIANY KOMUNIKATÓW ADMINISTRACYJNYCH

System obsługi komunikatów (SOK) jest aplikacją, która umożliwia wysyłanie powiadomień administracyjnych do systemów zintegrowanych w ramach RAD-onu. Komunikat administracyjny to ważna wiadomość dla użytkowników systemu dziedzinowego, która powinna być wyświetlona w widocznym miejscu, tak aby każdy użytkownik mógł

[40] pbn.nauka.gov.pl (dostęp: 02.03.2023)
[41] polon2.opi.org.pl/siec-polon (dostęp: 02.03.2023)

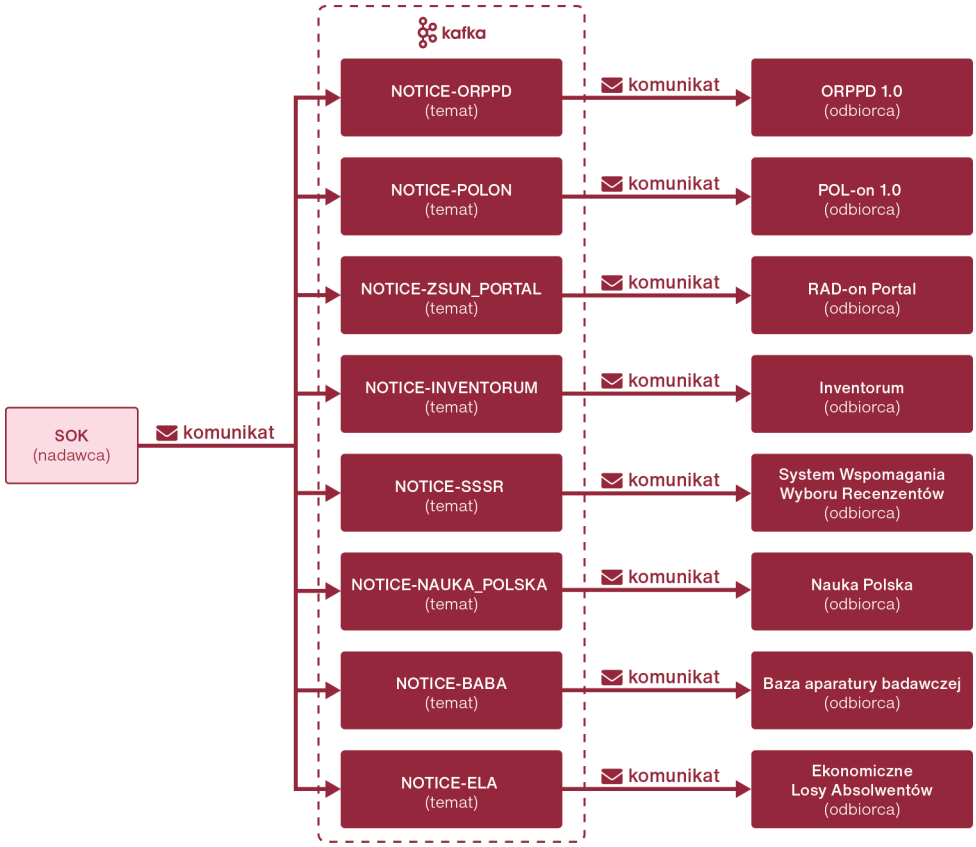
się z nią zapoznać. W szczególności są to komunikaty o planowanych wdrożeniach i przestojach konserwacyjnych systemów. SOK udostępnia interfejs graficzny umożliwiający użytkownikom tworzenie oraz modyfikowanie komunikatów administracyjnych. Posiada on własną, niezależną bazę danych, w której przechowywane są wszystkie opublikowane komunikaty administracyjne. SOK został zintegrowany z modelem wymiany danych, który jest odpowiedzialny za propagowanie informacji o zmianach mających miejsce w systemie. Jego ważną funkcją jest klasyfikowanie komunikatów według ich wagi (ostrzeżenia, informacje, błędy) wraz z określeniem ram czasowych publikacji komunikatu w systemach dziedzinowych. Oprócz treści, komunikat administracyjny zawiera identyfikator komunikatu, identyfikator systemu, którego dotyczy oraz informację, czy określony komunikat jest nadal aktywny. Wiadomość dostępna w Apache Kafka ma na celu poinformowanie o dodaniu lub modyfikacji komunikatu administracyjnego z poziomu interfejsu graficznego.

Proces integracji systemów z SOK przewiduje dwa scenariusze działania przepływu komunikatów administracyjnych. Pierwszy scenariusz wykorzystywany jest w przypadku, gdy system zintegrowany z SOK jest uruchomiony i działa bez zakłóceń. W tym scenariuszu komunikaty dodawane lub modyfikowane przez użytkowników SOK są automatycznie przypisywane do odpowiedniego tematu w Apache Kafka i publikowane. Proces tworzenia, zarządzania i współdzielenia komunikatów został przedstawiony na Ilustracji 4.6



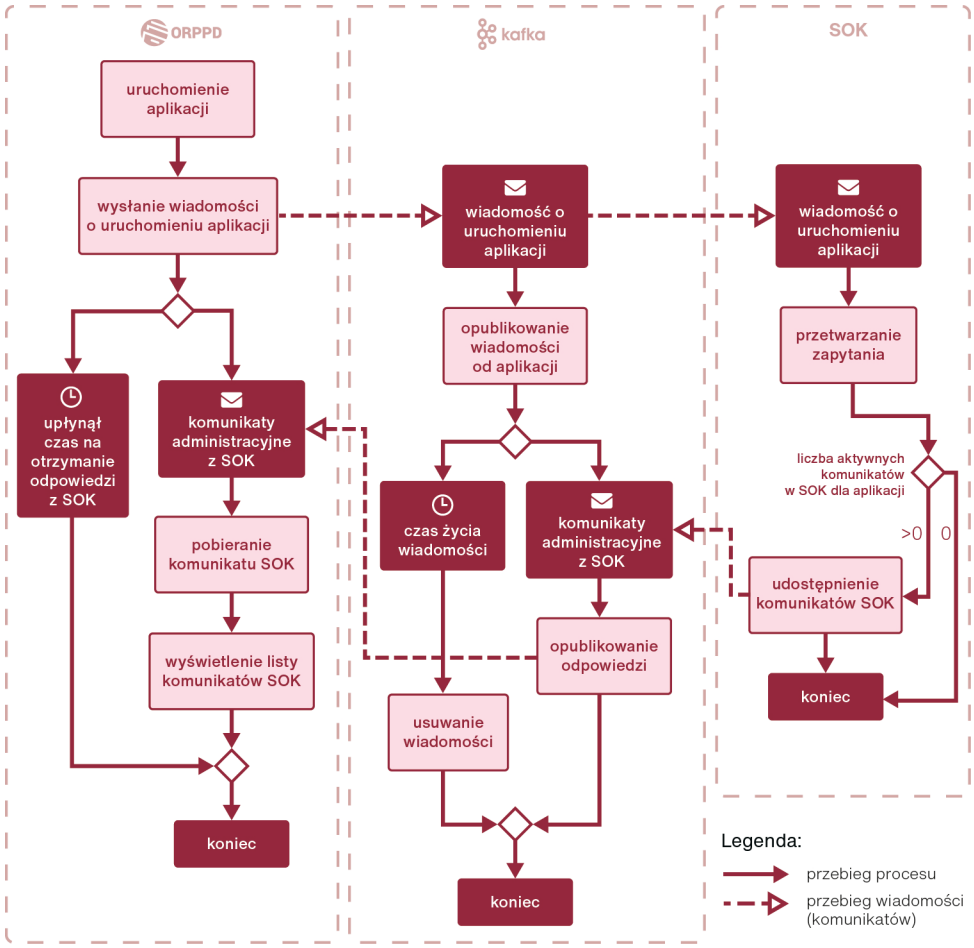
Ilustracja 4.6. Proces tworzenia komunikatów administracyjnych z wykorzystaniem modelu wymiany danych na przykładzie ORPPD

Systemy subskrybujące temat, w ramach którego pojawiła się nowa lub zmodyfikowana wiadomość, pobierają opublikowany komunikat, a następnie wyświetlają w miejscu do tego przeznaczonym. Utworzono szereg tematów odpowiedzialnych za komunikaty administracyjne dla poszczególnych systemów dziedzinowych. Tematy udostępniające komunikaty administracyjne poszczególnym systemom dziedzinowym zostały przedstawione na Ilustracji 4.7.



Ilustracja 4.7. Integracja systemów dziedzinowych z Systemem Obsługi Komunikatów (SOK) z wykorzystaniem modelu wymiany danych

Drugi scenariusz integracji SOK z systemami dziedzinowymi obsługuje sytuację, gdy system dziedzinowy został zamknięty i ponownie uruchomiony. Proces pobierania komunikatów administracyjnych po ponownym uruchomieniu systemu dziedzinowego został przedstawiony na Ilustracji 4.8.



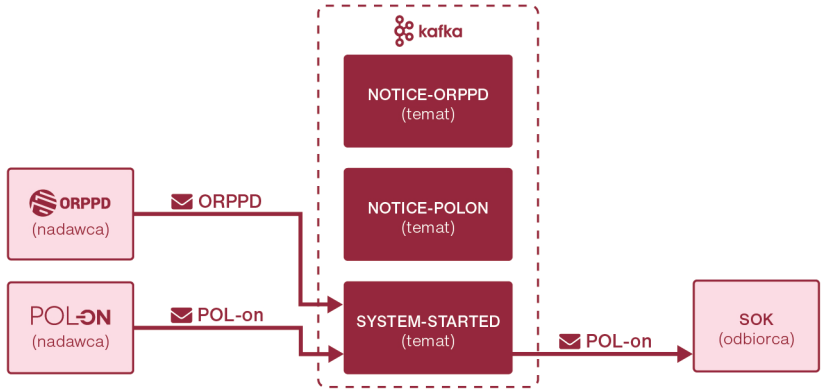
Ilustracja 4.8. Proces pobierania komunikatów administracyjnych po ponownym uruchomieniu aplikacji dziedzinowej z wykorzystaniem modelu wymiany danych na przykładzie ORPPD

Systemy dziedzinowe nie przechowują trwale zapisanych komunikatów z SOK, ponieważ pobierają je w czasie rzeczywistym i przechowują w pamięci podręcznej. W przypadku zamknięcia i ponownego uruchomienia aplikacji konieczne jest pobranie wszystkich komunikatów, tak by system dziedzinowy mógł odtworzyć swój stan sprzed zamknięcia.

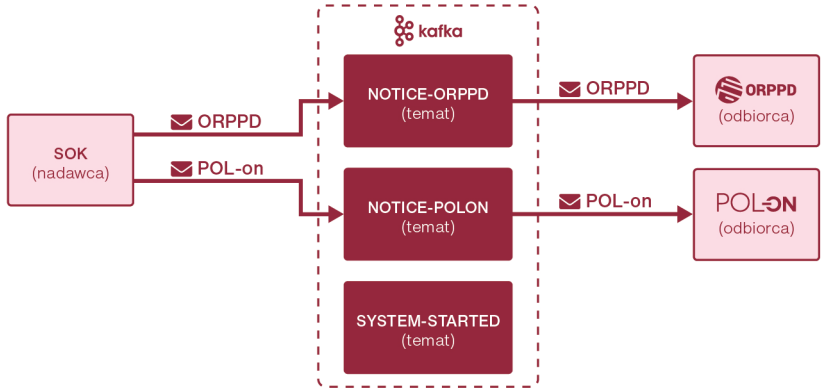
Wiadomości publikowane za pomocą Apache Kafka mają zdefiniowany czas życia w systemie kolejek. Po upłynięciu zdefiniowanego w konfiguracji serwera czasu publikacji, udostępniane przez Kafkę wiadomości są usuwane z systemu kolejek. Wiadomości, które zostały usunięte z systemu kolejek, mogą nadal pozostawać w pamięci podręcznej aplikacji dziedzinowej. Po zamknięciu aplikacji dziedzinowej wszystkie komunikaty zostają usunięte z pamięci podręcznej. Ponownie uruchomiona aplikacja dziedzinowa wymaga kompletnej listy komunikatów utworzonych w SOK. W celu przywrócenia sta-

nu sprzed zamknięcia systemu dziedzinowego, został zaimplementowany mechanizm umożliwiający poinformowanie SOK o fakcie ponownego uruchomienia aplikacji. Uruchamiany system wysyła do Apache Kafka wiadomość zawierającą informację, jaki system dziedzinowy został zrestartowany. SOK odczytuje wiadomość i wysyła do Apache Kafka pełną listę komunikatów administracyjnych. W opisanym powyżej scenariuszu każdy z systemów zintegrowanych poprzez Apache Kafka występuje zarówno w roli odbiorcy, jak i nadawcy. Etapy przywracania listy komunikatów w systemie dziedzinowym oraz komponenty biorące udział w komunikacji zostały przedstawione na Ilustracji 4.9 na przykładzie systemu ORPPD 1.0 i POL-on 1.0.

Etap I. Wysyłanie wiadomości o ponownym uruchomieniu systemów dziedzinowych



Etap II. Wysyłanie wiadomości z kompletną listą komunikatów administracyjnych dla systemów dziedzinowych



Ilustracja 4.9. Odczytanie komunikatów z SOK z użyciem modelu wymiany danych po ponownym uruchomieniu systemu dziedzinowego na przykładzie ORPPD i POL-on

SOK subskrybuje w Apache Kafka temat o nazwie SYSTEM-STARTED, który umożliwia wszystkim systemom zgłaszanie faktu ponownego uruchomienia. SOK odbiera wia-

domość, a następnie publikuje kompletną listę komunikatów administracyjnych w ramach odpowiedniego tematu.

4.2. Hurtownia danych

W rozwijanym w ramach RAD-onu ekosystemie IT brakowało komponentu integrującego i porządkującego dane wytwarzane przez poszczególne systemy. Efektem braku centralnego źródła danych była konieczność raportowania bezpośrednio z systemów działających na środowiskach produkcyjnych, a procesy integracji i wymiany danych pomiędzy systemami realizowane były w sposób nieustandaryzowany, bardzo często generując nieuzasadnione obciążenie systemów działających produkcyjnie. Realizacja złożonych analiz, wymagająca przekrojowego przeanalizowania danych z wielu systemów, była bardzo kosztowna, ponieważ każdorazowo wymagana była ręczna integracja, czyszczenie i porządkowanie danych manualnie pobranych z systemów dziedzinowych. Utrzymujący się ciągły wzrost kosztów przygotowania raportów cyklicznych, przekrojowych analiz na żądanie użytkownika czy odpowiedzi na zapytania w ramach dostępu do informacji publicznej doprowadziły do podjęcia decyzji o konieczności wdrożenia centralnej hurtowni danych wraz z systemem klasy *business intelligence* (BI). W ramach prac nad systemem RAD-on zaprojektowaliśmy oraz wdrożyliśmy hurtownię danych. Stała się ona naturalnym źródłem kompletnych oraz wiarygodnych danych prezentowanych na portalu RAD-on (usługi i dane publikowane na portalu opisaliśmy w rozdziale 2).

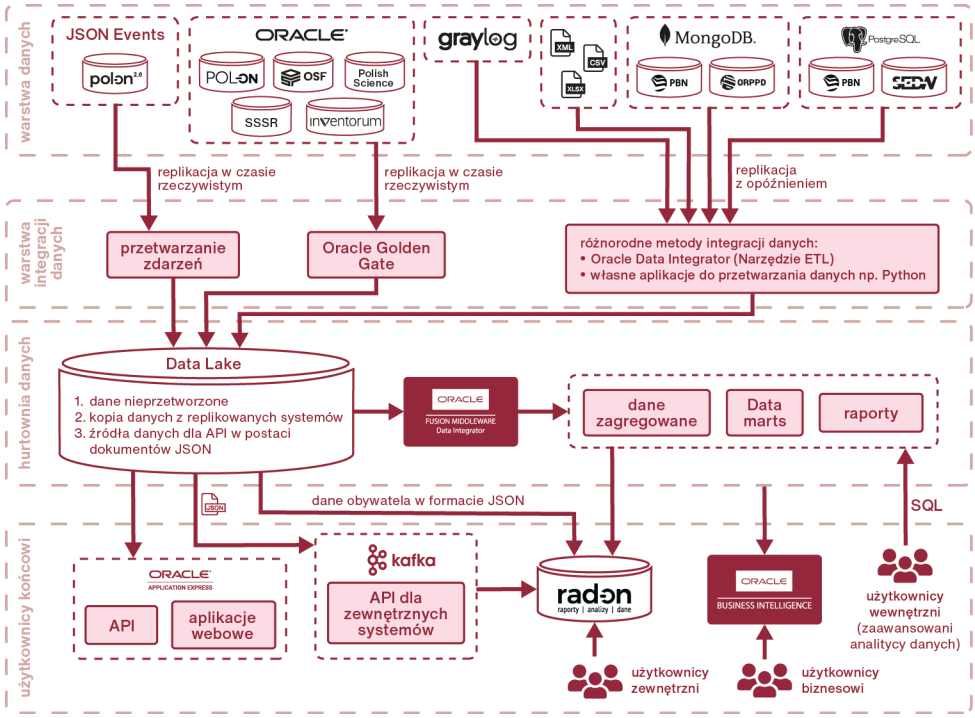
4.2.1. Architektura hurtowni danych

Rozwijane przez OPI PIB systemy informatyczne wykorzystują różnorodne technologie przechowywania danych. Dominującą technologią składowania danych jest baza relacyjna, bardzo często wykorzystywana do przechowywania dokumentów typu JSON lub XML. W hurtowni danych zostały zintegrowane następujące źródła danych:

- relacyjne bazy danych Oracle;
- relacyjne bazy danych PostgreSQL;
- relacyjne bazy danych MySQL;
- relacyjne bazy danych MSSQL;
- obiektowe bazy danych MongoDB;
- broker komunikatów Apache Kafka;
- dokumenty XML;
- REST API;
- statyczne pliki XLXS, CSV, TXT itp.

Wymienione powyżej źródła danych zostały pokazane na Ilustracji 4.10 (warstwa źródła danych), która w uproszczony sposób obrazuje architekturę hurtowni danych oraz przepływu informacji pomiędzy systemami dziedzinowymi, hurtownią danych i – finalnie – systemem RAD-on. Z powodu dominującej pozycji technologii Oracle (wykorzystują

ją największe systemy rozwijane w OPI PIB, czyli POL-on i OSF – więcej o systemach w rozdziale 1.6), jako silnik bazy danych dla hurtowni wybraliśmy bazę właśnie tej firmy.



Ilustracja 4.10. Uproszczona architektura hurtowni danych

Realizacja procesu integracji wielu nieheterogenicznych źródeł danych wymaga przygotowania rozbudowanego procesu przetwarzania danych (*Extract, Transform, Load*, ETL). W tym celu skorzystaliśmy z różnych narzędzi i metod integracji danych. Dane z baz Oracle replikowane są do hurtowni danych w czasie rzeczywistym, z wykorzystaniem specjalnego narzędzia Oracle Golden Gate[42]. Wykorzystanie tego produktu umożliwiło nam efektywne przenoszenie zmian w danych, dokonywanych w systemach produkcyjnych bez ich obciążania. Opóźnienie w replikacji danych nie przekracza jednej sekundy. W przypadku pozostałych źródeł danych do integracji danych wykorzystujemy Oracle Data Integrator (ODI)[43] lub autorskie rozwiązania napisane w języku Python[44], które następnie uruchamiane są z wykorzystaniem ODI. W tym przypadku opóźnienie w replikacji danych zależy od wymagań biznesowych. Dla krytycznych procesów biznesowych dane mogą być replikowane nawet co kilka minut, dla pozostałych

[42] oracle.com/pl/integration/goldengate (dostęp: 02.03.2023)
[43] oracle.com/pl/middleware/technologies/data-integrator.html (dostęp: 02.03.2023)
[44] python.org (dostęp: 02.03.2023)

natomiast do kilku razy na dzień. Widać tu wyraźne odejście od najważniejszego do niedawna trendu odświeżania hurtowni danych raz dziennie w godzinach nocnych. W naszym przypadku procesy krytyczne, wymagające najbardziej aktualnych danych, zostały odseparowane od standardowych lub cyklicznych procesów raportowych.

Produktem końcowym procesu ETL są oczyszczone, usystematyzowane i pogrupowane tematycznie zbiory danych (*Data Mart*, DM). Struktury danych tworzące DM mogą zawierać zintegrowane dane z wielu systemów lub obszarów. Przykładem modelu danych łączącego dane z wielu systemów jest zbiór danych na temat instytucji, który integruje dane z trzech systemów: POL-on, Nauka Polska i Inventorum (szerzej opisane w rozdziale 1.6). Przygotowany w ramach procesu przetwarzania danych „złoty rekord" [15] jest następnie w całości wykorzystywany przez system RAD-on i stanowi referencyjną informację dla pozostałych systemów OPI PIB, jak również dla systemów zewnętrznych, na przykład systemów uczelni, które chcą się integrować maszynowo z systemami domeny MEiN.

Hurtownia danych została zaprojektowana z wykorzystaniem dwóch metodologii: 1) klasycznego modelu gwiazdy, stanowiącego źródło danych do raportów implementowanych w systemie BI oraz 2) Data Vault, wykorzystywanego ze względu na częste zmiany w strukturach danych oraz bardzo elastyczne rozwiązanie w pracy z dokumentami (np. w formacie JSON).

W ramach projektu zaimplementowaliśmy dwadzieścia obszarów danych, które służą do obsługi procesów w ramach systemu RAD-on, Apache Kafka oraz BI. Na początku 2020 roku do architektury hurtowni danych dołączył jeszcze jeden komponent: platforma niskokodowa (*low-code; Rapid Application Development*, RAD) Oracle APEX. Wykorzystanie tego rozwiązania w znaczący sposób uprościło integrację systemów OPI PIB z hurtownią danych, dzięki zastosowaniu szybkich i prostych w implementacji usług REST API. Zalety technologii RAD znacząco rozwinęły możliwość wykorzystania hurtowni danych w komunikacji z użytkownikami. Powstały między innymi przyjazne interfejsy umożliwiające samodzielne zasilanie hurtowni własnymi danymi, bez konieczności angażowania zespołu programistów.

4.2.2. Wdrożenie narzędzia klasy BI

W celu umożliwienia łatwej i przyjaznej komunikacji użytkowników końcowych z hurtownią danych, wdrożyliśmy narzędzie Business Intelligence Oracle Analytics Server (OAS, dawniej Oracle Business Intelligence EE OBIEE). Zgodnie z założeniami projektu głównymi odbiorcami danych, udostępnianych w postaci interaktywnych pulpitów informacyjnych mieli być specjaliści MEiN. Obowiązujący wcześniej proces zakładał ręczne generowanie raportów w trybie „na żądanie" przez analityków danych. Nie uwzględniał on hurtowni danych czy mechanizmów automatyzujących proces raportowy. Zaletą takiego procesu było indywidualne podejście do każdej analizy. Największą wadę stanowił koszt utrzymania procesu, w którym każde zamówienie raportu było analizowane i przygotowywane niezależnie od poprzednich. Drugą wadą był długi czas oczekiwa-

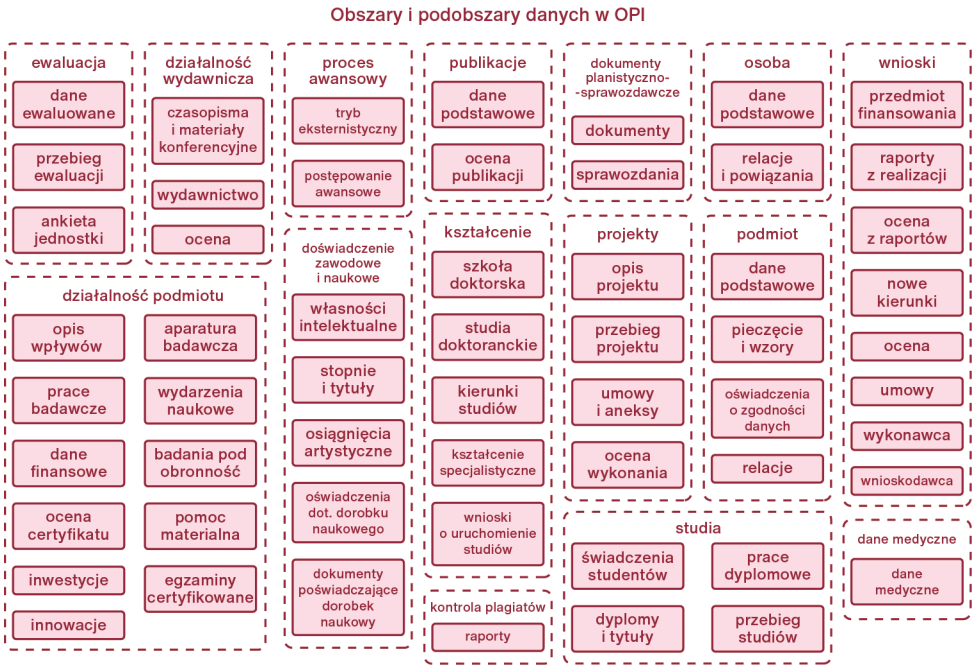
nia na indywidualną analizę. Wdrożenie narzędzia BI umożliwiło zautomatyzowanie cyklicznych procesów raportowych. Po przeanalizowaniu zamówień na raporty stworzyliśmy dedykowane pulpity informacyjne, dostarczające odpowiedzi na najczęściej zadawane pytania, umożliwiające użytkownikom wykonanie samodzielnej analizy danych, z uwzględnieniem różnych przekrojów. Najciekawsze statystyki dotyczące wykorzystania BI i hurtowni danych przedstawiają się następująco:

- **24** tematyczne pulpity informacyjne;
- **107** unikatowych użytkowników;
- **10** unikatowych użytkowników dziennie;
- ponad **61 tysięcy** raportów wygenerowanych w 2022 roku;
- ponad **255 milionów** wierszy w pobranych w 2022 roku raportach;
- **osiem** instytucji publicznych korzystających z BI;
- **82** unikatowe procesy ETL zasilające hurtownie danych;
- **2,2 terabajta** danych składowanych w hurtowni danych;
- **191** udostępnionych interfejsów REST API.

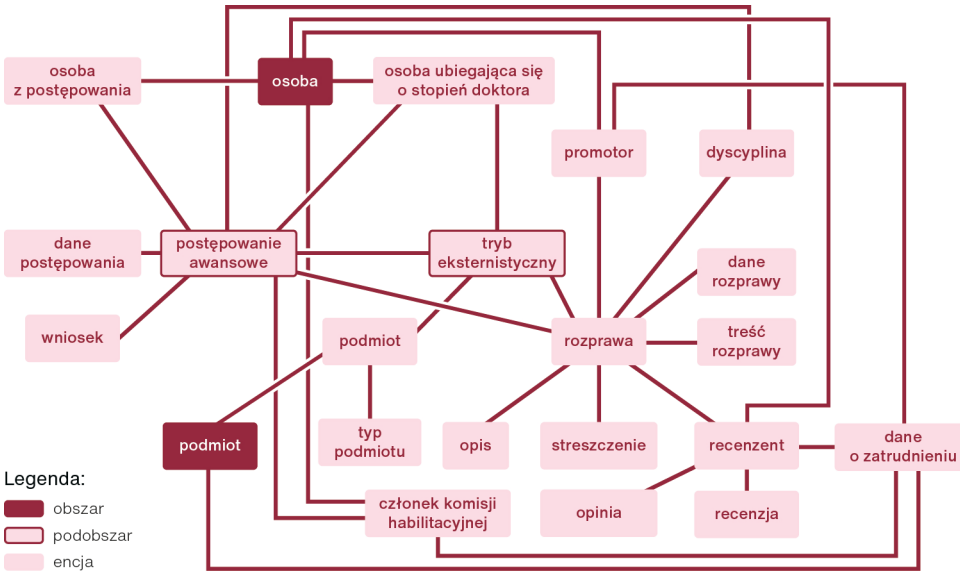
4.2.3. Zarządzenie danymi

Posiadanie dużej ilości danych nie zawsze jest powiązane z umiejętnością wykorzystania potencjału informacji. OPI PIB przetwarza w wytwarzanych w instytucie systemach bardzo duże ilości różnorodnych danych. Z racji specyfiki wytwarzanych systemów, wiedza dziedzinowa o danych gromadzona jest w zespołach wytwarzających systemy i zespołach lub osobach zajmujących się tworzeniem raportów analitycznych. Wysoki poziom rozproszenia informacji w systemach nie ułatwia analizy danych i wyciągania wniosków, a wydobycie przekrojowych informacji wymaga zaangażowania wielu osób. W związku z tym w 2021 roku podjęliśmy decyzję o uruchomieniu programu zarządzania danymi (*Data Governance*, DG). Mając świadomość zawiłości związanych z zarządzaniem danymi i wyzwaniami wiążącymi się z wdrażaniem DG w organizacji [2], za główne cele uznaliśmy zwiększenie w OPI PIB świadomości potencjału danych, którymi zarządza organizacja, oraz wyznaczenie osób odpowiedzialnych za poszczególne obszary danych (*data owners* – właściciele danych). Po ponadrocznej pracy udało nam się przeprowadzić ewidencję posiadanych zbiorów danych oraz powołać macierzową strukturę organizacyjną odpowiedzialną za powodzenie programu. Zainicjowaliśmy również proces tworzenia korporacyjnego modelu danych. Na obecnym etapie prac zidentyfikowaliśmy 15 dużych biznesowych obszarów danych oraz 58 mniejszych (Ilustracja 4.11), którymi zarządza 12 właścicieli danych. Podstawą do implementacji platformy RAD-on są dane, dlatego kluczowym elementem projektu było zapewnienie odpowiedniej jakości, niezawodności i wysokiej dostępności danych. Dzięki wdrożeniu polityk zarządzania danymi, platforma RAD-on stała się efektywnym i wiarygodnym dostawcą publicznych informacji z sektora nauki i szkolnictwa wyższego w Polsce.

Na Ilustracji 4.12 przedstawiony został biznesowy model danych dla obszaru dotyczącego procesu awansowego.



Ilustracja 4.11. Biznesowe obszary danych

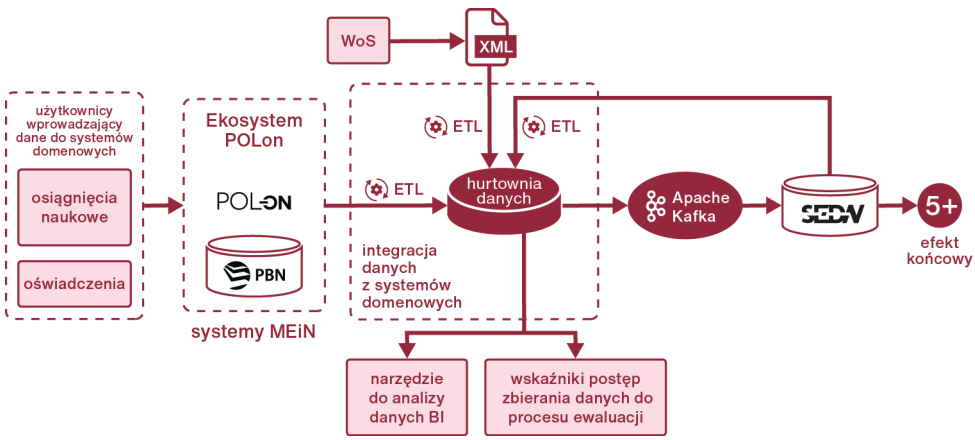


Ilustracja 4.12. Biznesowy model danych dla obszaru „Proces awansowy”

Prace związane z wdrażaniem elementów programu DG w OPI PIB pozwoliły zrozumieć, jak ważne dla organizacji jest właściwe zarządzanie danymi. Wykształcenie w organizacji świadomości i rozumienia posiadanych danych ułatwia codzienną pracę, poprawia bezpieczeństwo i umożliwia dostarczanie lepszych produktów w krótszym czasie.

4.2.4. Rola hurtowni danych w architekturze systemów informatycznych OPI PIB

W obecnej architekturze systemów informatycznych rozwijanych przez instytut hurtownia bardzo często pełni rolę integratora danych. Z założenia wszystkie systemy informatyczne powinny zostać zintegrowane z hurtownią, a ta powinna przetrzymywać kopie danych ze zintegrowanych systemów. Dlatego gdy określony system wymaga dostarczenia wiarygodnych, przetransformowanych i oczyszczonych danych, to naturalnym ich dostawcą będzie właśnie hurtownia. W OPI PIB takie podejście znacząco skróciło czas implementacji funkcjonalności w systemach biznesowych, ponieważ programiści skupiali się na oprogramowaniu określonej funkcjonalności, nie musząc myśleć o integrowaniu danych z wielu źródeł. Jednym z przykładów wykorzystania hurtowni danych jako integratora był proces ewaluacji działalności naukowej instytucji (opisany szerzej w rozdziale 1.3). W procesie ewaluacji hurtownia danych zbierała, integrowała, czyściła i transformowała do docelowego formatu dane z systemów POL-on, PBN oraz Web of Science (WoS), które następnie zasilały System Ewaluacji Dorobku Naukowego (SEDN). Cały proces zobrazowany jest na Ilustracji 4.13.



Ilustracja 4.13. Wykorzystanie hurtowni danych w procesie ewaluacji jakości działalności naukowej podmiotów systemu nauki i szkolnictwa wyższego w Polsce

ROZDZIAŁ 5

API

Marcin Białas
dr inż. Marcin Mirończuk

5.1. Geneza API

Interfejs programowania aplikacji (*Application Programming Interface*, API) to zestaw ścisłych reguł i konwencji, które określają sposób, w jaki programy mogą komunikować się ze sobą, jakie mogą wymieniać informacje wraz z określeniem ścisłej ich struktury, oraz jak mają wykonywać wspólne zadania [30]. Interfejsy programowania aplikacji są używane w wielu różnych dziedzinach, w tym w informatyce, telekomunikacji i elektronice, a ich celem jest ułatwienie tworzenia programów i umożliwienie im współpracy. Dodatkowo API może korzystać z komponentów graficznego interfejsu użytkownika (*Graphical User Interface*, GUI) [34]. Dobre API ułatwia budowę oprogramowania, sprowadzając ją do łączenia przez programistę bloków elementów w ustalonej konwencji.

Jednym z celów projektu było udostępnienie zestandaryzowanych danych źródłowych z zasobów nauki i szkolnictwa wyższego za pomocą szeregu API, dostarczających usług użytkownikom końcowym – zarówno ludziom, jak i maszynom z organizacji sektora publicznego i prywatnego.

Nad strukturą danych, komunikacją i wymianą informacji pomiędzy poszczególnymi elementami systemu RAD-on, jak i użytkownikami a wykonanym systemem, w postaci aplikacji internetowej, czuwa utworzone REST (*Representational State Transfer*) API. Utworzone REST API umożliwia realizację założonych w systemie RAD-on funkcji poszczególnych serwisów i modułów, jak i komunikowanie ich ze sobą w razie potrzeby. REST API jest rodzajem interfejsu programowania aplikacji, który jest oparty na architekturze klient–serwer i przestrzeni nazw w postaci ujednoliconych identyfikatorów zasobów (*Uniform Resource Identifier*, URI). REST API używa protokołu HTTP do komunikacji między klientem a serwerem i jest przeznaczone do udostępniania zasobów w sposób, który pozwala na ich odczytanie, modyfikację, dodawanie lub usuwanie za pomocą standardowych metod HTTP, takich jak GET, POST, PUT i DELETE. Ponadto API jest często używane w aplikacjach internetowych i mobilnych, ponieważ pozwala na łatwą integrację różnych systemów i usług.

Obecnie system RAD-on udostępnia dane za pośrednictwem REST API oraz API w postaci interfejsu graficznego. Udostępnia on standardowe kategorie danych, które stano-

wią podstawę jego usług i modułów. Każda kategoria odpowiada jednemu z zasobów związanych ze szkolnictwem wyższym i nauką (np. instytucje, publikacje, pracownicy uczelni). Większość tych danych jest stale modyfikowana – podmioty nauki i szkolnictwa wyższego mogą zmieniać nazwy, pojawiają się nowe publikacje, pracownicy zmieniają zatrudnienie. Dodatkowo zmiana struktury danych oraz samych danych jest konieczna zawsze wtedy, gdy pojawiają się nowe wymagania dotyczące udostępnianych informacji lub nowe modele biznesowe w systemach źródłowych. Możliwe jest też pojawienie się nowych kategorii danych. Wszystkie te scenariusze muszą zostać uwzględnione w procesie integracji i udostępniania danych, co wymaga od systemu spełnienia wielu wymagań. W celu zarządzania tak dynamicznie zmieniającymi się danymi, została zaprojektowana i zaimplementowana nowatorska strategia, architektura i system, dzięki którym możliwe jest szybkie i sprawne reagowanie na zmieniające się potrzeby i wymagania interesariuszy.

Omawiane w bieżącym rozdziale monografii API powiązane jest szczególnie z usługami udostępniania maszynowego zasobów szkolnictwa wyższego i nauki oraz udostępniania metadanych. Opinie na temat zapotrzebowania zaprojektowanych usług zostały ustalone w trakcie badania potrzeb opisanego w rozdziale pierwszym 1.4.

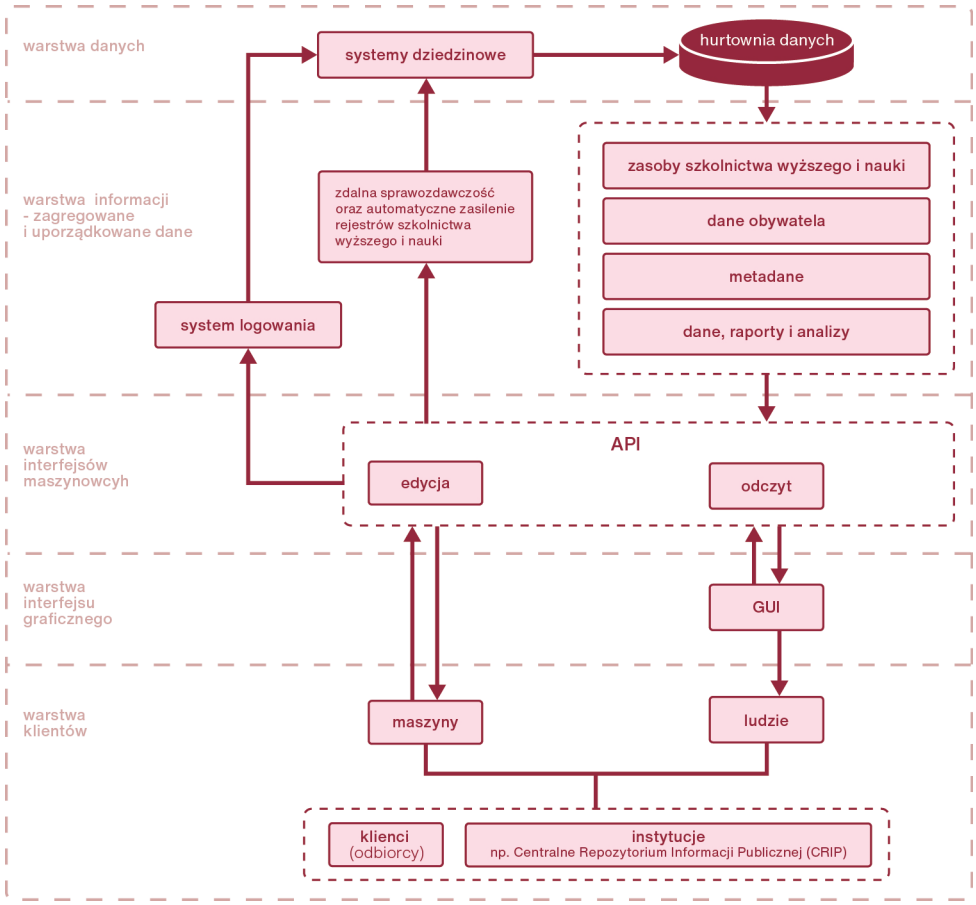
Usługa udostępniania maszynowego zasobów szkolnictwa wyższego i nauki została dostosowana do potrzeb co najmniej kilku grup odbiorców instytucjonalnych, w tym uczelni i instytucji naukowych, agencji finansujących naukę (NCN, NCBR, NAWA) oraz Polskiej Komisji Akredytacyjnej. Z punktu widzenia administracji uczelni i instytucji naukowych usługa zaspokaja potrzeby takie jak automatyczne i maszynowe pobieranie różnego rodzaju zestawień z zakresu nauki i szkolnictwa wyższego.

W przypadku usługi udostępniania metadanych, której celem było zwiększenie widoczności i dostępności zasobów polskiej nauki, planowane rozwiązanie miało być użyteczne dla szerokiej grupy odbiorców, w tym indywidualnych naukowców i badaczy, przedstawicieli instytucji rządowych odpowiedzialnych za politykę naukową kraju, instytucji badawczo-naukowych, dziennikarzy oraz obywateli zainteresowanych dostępem do danych o nauce i szkolnictwie wyższym. Potrzeby tych interesariuszy zostały zaspokojne głównie dzięki możliwości wysyłania metadanych o udostępnionych danych do Centralnego Repozytorium Informacji Publicznej (CRIP) i aktualizacji metadanych w razie zmian. Dzięki integracji z CRIP zwiększona została widoczność zasobów dotyczących nauki w Polsce.

W trzech kolejnych podrozdziałach opisujemy architekturę zaproponowanego rozwiązania z uwzględnieniem kluczowych elementów systemu i przepływu danych. Opis architektury został podzielony ze względu na jego szczegółowość. W pierwszej kolejności zostały przedstawione ogólne założenia i kluczowe komponenty architektury systemu RAD-on (podrozdział 5.2). Następnie szczegółowo opisano techniczne aspekty zrealizowanej strategii (podrozdział 5.3) wraz z użytymi technologiami (podrozdział 5.4).

5.2. Architektura systemu w ujęciu ogólnym

Architektura utworzonego systemu składa się z interfejsów odpowiadających za bezpośrednią komunikację zarówno pomiędzy systemami wewnętrznymi i zewnętrznymi, jak i z użytkownikami systemu. Poglądowa architektura systemu została przedstawiona na Ilustracji 5.1.



Ilustracja 5.1. Ogólna architektura utworzonego systemu

Stworzony system, pokazany na Ilustracji 5.1, składa się z następujących warstw:

1. **Warstwa danych:** zajmuje się integracją zasobów informacyjnych systemów dziedzinowych i udostępnianiem zestandaryzowanych danych źródłowych dla usług przez model wymiany danych.
2. **Warstwa informacji:** przechowuje dane poddane na przykład procesowi porządkowania i agregacji w formie raportów i analiz.

3. **Warstwa interfejsów maszynowych:** służy do wprowadzania i edycji danych (jest to inaczej warstwa serwisów do zasilania rejestrów informacyjnych).
4. **Warstwa interfejsów maszynowych do odczytu danych** (serwisy dostępu do danych, analiz i raportów): udostępnia standardowe API, dzięki któremu można pobierać ujednoliconą informację z warstwy informacyjnej.
5. **Warstwa interfejsu graficznego** (portal obywatelski) w postaci portalu internetowego: umożliwia użytkownikom wygodny dostęp do informacji z zakresu szkolnictwa wyższego i nauki.
6. **Warstwa użytkowników:** składa się z ludzi (zarówno użytkowników instytucjonalnych, jak i indywidualnych) oraz programów (maszyn) stworzonych przez użytkowników lub na zamówienie firm i instytucji.

Z architekturą systemu związane jest także pojęcie modelu wymiany danych, który został dokładnie opisany w rozdziale 4. W celu przypomnienia, model wymiany danych stanowi abstrakcyjny element, który łączy wszystkie wymienione powyżej warstwy. Ma on na celu integrację systemów w dziedzinie nauki i szkolnictwa wyższego oraz umożliwienie współpracy z innymi modelami, na przykład modelami informacyjnymi na poziomie krajowym. Model wymiany danych łączy systemy wewnętrzne i zewnętrzne oraz udostępnia dane z systemów źródłowych za pośrednictwem serwisów maszynowych i portalu.

Na **warstwę danych** składają się następujące komponenty:

- dziedzinowe (źródłowe) systemy informacyjne: różnego rodzaju systemy informacyjne i ich rejestry danych, utrzymywane przez OPI PIB lub podmioty zewnętrzne;
- hurtownia danych: przetwarza dane źródłowe w użyteczne informacje, takie jak analizy, statystyki, raporty statyczne i dynamiczne, poprzez procesy ładowania, transformacji i ekstrakcji danych oraz narzędzia analityki biznesowej. Informacje te są udostępniane jako raporty przez bazę wiedzy w portalu oraz przez narzędzia *business intelligence* (BI) dla zaawansowanych użytkowników.

Warstwa informacji składa się z następujących komponentów:

- baza wiedzy (raporty, analizy, dane): indeks semantyczny do wyszukiwania. Udostępniona jest wyszukiwarka pełnotekstowa i semantyczna, umożliwiająca przeszukanie zasobów wszystkich systemów dziedzinowych i połączenie semantyczne danych;
- zasoby szkolnictwa wyższego i nauki: zunifikowane rejestry danych i informacji zgromadzone w systemach dziedzinowych odpowiadających za realizację procesów biznesowych w obrębie nauki i szkolnictwa wyższego;
- dane obywatela: rejestr zawierający zunifikowane informacje pochodzące z różnych systemów dziedzinowych, w tym dane osobowe i powiązane informacje;
- metadane: rejestr zawierający informacje o udostępnionych usługach;
- logi zdarzeń: rejestr składujący wszelkie zdarzenia w postaci zapisu i odczytu danych poprzez udostępnione API.

Warstwa interfejsów maszynowych do wprowadzania i edycji danych zawiera serwis do realizacji zadań związanych ze zdalną sprawozdawczością i automatycznym zasilaniem rejestrów szkolnictwa wyższego i nauki za pomocą systemów zewnętrznych, które mają dostęp do maszynowych serwisów oraz umożliwiają realizację udostępniania i edycję danych.

Warstwa interfejsów maszynowych do odczytu danych składa się z następujących komponentów:

- serwis udostępniający zasoby bazy wiedzy za pomocą API;
- serwis udostępniający zasoby szkolnictwa wyższego: zestaw serwisów sieciowych, które umożliwiają pobieranie danych przez zewnętrzne systemy. Serwisy te są zintegrowane z systemami źródłowymi danych za pomocą modelu wymiany danych, co umożliwia separację zewnętrznych systemów od źródeł danych i ujednolica model komunikacji;
- serwis dostępu do danych obywatela: użytkownik ma dostęp do swoich danych za pośrednictwem usługi dostępu do danych obywatela dostępnej za pośrednictwem portalu. Portal wyświetla wszystkie dane dotyczące określonego użytkownika, ponieważ model wymiany danych sprawdza zasoby wszystkich systemów źródłowych i udostępnia kompletny zestaw danych;
- serwis udostępniający metadane: model wymiany danych posiada informacje na temat danych i usług udostępnianych przez siebie. Użytkownik może dowiedzieć się, jakie usługi i dane są dostępne poprzez portal. Także zewnętrzne systemy mogą zdobyć tę samą wiedzę za pośrednictwem serwisów maszynowych.

Warstwa interfejsu graficznego składa się z następujących komponentów:

- raporty, analizy, dane i wyszukiwarka: elementy umieszczone w portalu internetowym, umożliwiające dostęp do informacji zgromadzonych w modułach: baza wiedzy, dane obywatela, metadane, zasoby szkolnictwa wyższego. Portal łączy się z warstwą informacyjną za pomocą standardowych API. Wyszukiwarka umożliwia przeszukiwanie danych w języku naturalnym oraz przeszukanie wszystkich systemów dziedzinowych połączonych semantycznie w bazie wiedzy;
- dane obywatela: element pozwalający użytkownikowi na sprawdzenie swoich danych zgromadzonych w systemach dziedzinowych, a także dający możliwość ich edycji, na przykład zgłoszenia potrzeby poprawy danych.

Warstwa użytkowników składa się z następujących komponentów:

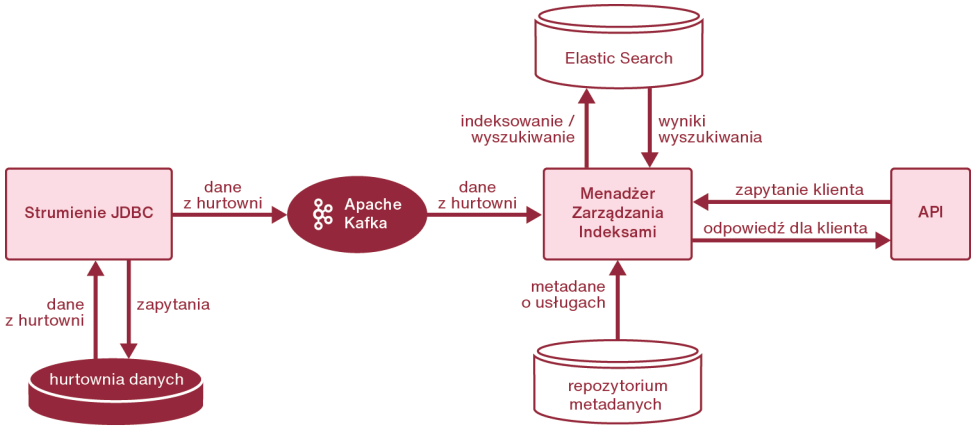
- systemy zewnętrzne: dowolne systemy informatyczne, które mogą wymieniać dane z systemem za pomocą serwisów dostępu do danych. Te systemy mogą pobierać i przekazywać dane z i do systemów źródłowych;
- systemy współpracujące: systemy będące w rzeczywistości systemami zewnętrznymi, które współpracują ściśle z systemem, wymieniając dane lub usługi. Kluczowe systemy współpracujące to CRIP i Krajowy Węzeł Identyfikacji Elektronicznej (KWIE);

- użytkownicy: osoby, które mogą być zlokalizowane w dowolnym miejscu i które mogą korzystać z systemu poprzez przeglądarkę internetową na dowolnym urządzeniu elektronicznym.

5.3. Architektura systemu w ujęciu szczegółowym

W tej części skupimy się na opisaniu tych elementów systemu, które biorą udział w procesie udostępniania danych. Zanim dane trafią do użytkownika końcowego, przechodzą przez wiele istotnych komponentów. Lista poglądowa komponentów znajduje się poniżej, natomiast zarys architektury i przepływu danych przedstawia Ilustracja 5.2.

- **hurtownia danych:** baza danych Oracle;
- **Strumienie JDBC:** aplikacja JAVA Spring obsługująca strumieniowe przesyłanie danych między hurtownią a Apache Kafka;
- **Apache Kafka [21]:** platforma do przetwarzania strumieni danych;
- **Menadżer Zarządzania Indeksami (MZI) i API:** aplikacja JAVA Spring;
- **Elasticsearch (ES) [24]:** silnik wyszukiwania pełnotekstowego;
- **repozytorium metadanych:** repozytorium GIT służące do przetrzymywania metadanych o usługach.



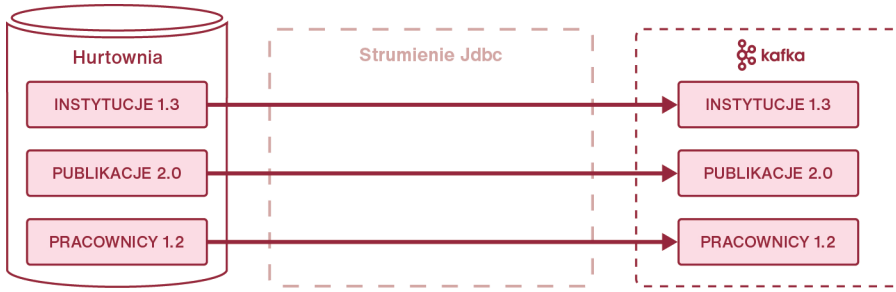
Ilustracja 5.2. Zarys architektury systemu z uwzględnieniem przepływu danych

W celu intuicyjnego przedstawienia działania systemu prześledzimy drogę, jaką pokonują dane od hurtowni do użytkownika końcowego. Pierwszym komponentem biorącym udział w transporcie danych jest aplikacja **Strumienie JDBC**. Jest ona pomostem między **Apache Kafka** a hurtownią danych. Jej podstawowa funkcjonalność polega na cyklicznym, wcześniej zaplanowanym wykonywaniu zapytań do hurtowni oraz transporcie danych do Apache Kafka. Dodatkowo pozwala ona zarządzać kolejkami komunikatów na Apache Kafka. Kolejki komunikatów będziemy dalej nazywać tematami. Gdy nowa paczka danych trafi do Apache Kafka, jest ona odczytywana przez **Menadżera Zarzą-**

dzania Indeksami (MZI). MZI jest centralnym komponentem systemu, pełniącym kilka kluczowych funkcji. Jedną z nich jest wspomniane już pobieranie danych z Apache Kafka. Po pobraniu danych MZI, opierając się na wcześniej zdefiniowanych instrukcjach, przetwarza dane do postaci dokumentów zgodnych ze specyfikacją **Elasticsearch (ES)**. Instrukcje określające strukturę dokumentów ES zależą od typu indeksowanych danych, a dokładniej od tego, jakie zapytania mogą być najbardziej interesujące dla użytkowników RAD-onu. Instrukcje dla różnych typów danych zgromadzone są w repozytorium metadanych, z którego są pobierane przez MZI. Przetworzone dokumenty są następnie indeksowane na klastrze ES. Kolejne zadanie MZI polega na obsłudze zapytań klienta przychodzących z warstwy **API**. Zapytania użytkowników trafiają do MZI, który następnie przetwarza je do postaci zgodnej ze specyfikacją ES i inicjuje wyszukiwanie. W dalszej kolejności wyniki wyszukiwania są przekazywane do klientów. Zasady korzystania z API są udostępnione klientom w postaci interaktywnej dokumentacji **Swagger**[45]. Cała dokumentacja, podobnie jak instrukcje indeksowania, składowana jest w **repozytorium metadanych**, z którym komunikuje się MZI. W kolejnych podrozdziałach skupimy się na omówieniu poszczególnych elementów systemu.

5.3.1. Zarządzanie strumieniami

Strumienie JDBC są komponentem systemu służącym do wysyłania danych z baz relacyjnych do Apache Kafka. Podstawową funkcjonalnością aplikacji jest możliwość definiowania strumieni. Strumień definiujemy jako uruchamiane cyklicznie zapytanie do bazy danych, którego wyniki są wysyłane do tematu Apache Kafka o nazwie identycznej jak nazwa strumienia. Każdy strumień obsługuje jedną wersje encji, a nadana mu nazwa jest połączeniem tych dwóch informacji. Przykładowo, w przypadku encji „instytucje w wersji 1.3" nazwa strumienia i temat będą mieć postać INSTYTUCJE-1.3. Wykorzystanie takiego nazewnictwa pozwala w logiczny sposób powiązać różne elementy systemu, o czym była już mowa w podrozdziale 5.3.2. Schemat transportu danych przedstawia Ilustracja 5.3. Widać na nim trzy różne tabele w hurtowni zawierające dane dla encji „instytucje", „publikacje" i „pracownicy".



Ilustracja 5.3. Transport danych między hurtownią danych a Apache Kafka

[45] swagger.io (dostęp: 02.03.2023)

Przekazywane dane nie są przechowywane przez aplikację, gdyż jej główną funkcją jest transport danych. W zależności od tego, w jaki sposób napisane jest zapytanie, strumień może pobierać całą zawartość tabeli lub tylko nowe rekordy. W ramach strumienia definiowana jest również częstotliwość, z jaką aplikacja ma odpytywać bazę danych. Wybranie odpowiedniej strategii zależy od dynamiki zmian, jakim podlegają dane i jest optymalizowany indywidualnie dla każdej encji. Dla danych, które podlegają częstym modyfikacjom, interwał między odpytaniami może wynosić kilka sekund czy minut. Z kolei dane podlegające rzadkim modyfikacjom mogą być odświeżane na przykład raz na dobę. Podczas definiowania częstotliwości istotne jest również wzięcie pod uwagę wydajności bazy danych – zbyt duża ilość zapytań wykonanych w tym samym czasie mogłaby doprowadzić do niewydolności systemu lub nawet awarii. Dzięki możliwości definiowania dla każdego strumienia indywidualnej strategii pobierania danych, system działa stabilnie i jednocześnie dostarcza użytkownikowi możliwe najświeższe dane.

Z poziomu panelu administratora możliwe jest monitorowanie aktualnego stanu aplikacji. Każdy strumień opisywany jest przez dwa odrębne stany: stan funkcjonalny i stan związany z ostatnią wykonaną akcją. Stan funkcjonalny określa aktualny stan działania strumienia. Możliwe wartości tego stanu to:

- *OCZEKIWANIE:* strumień oczekuje na kolejne uruchomienie;
- *PRZETWARZANIE:* strumień jest w trakcie pobierania danych z bazy i przesyłania na Apache Kafka;
- *ZATRZYMANY:* strumień został zatrzymany z powodu zbyt dużej ilości błędów.

Z kolei stan związany z ostatnią wykonaną akcją może przyjąć następujące wartości:

- *STABILNY:* ostatnie zapytanie wykonało się poprawnie;
- *NIESTABILNY:* ostatnie zapytanie wykonało się, ale przetwarzane rekordy zawierały błędy;
- *BŁĄD:* zapytanie zostało przerwane z powodu poważnego błędu lub dużej ilości błędnych rekordów;
- *WYCZYSZCZONY:* stan strumienia został zresetowany przez administratora.

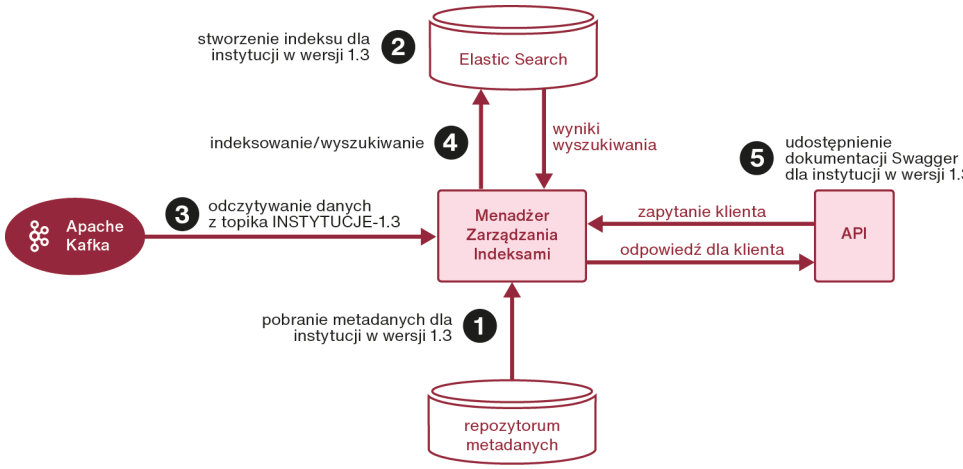
Prezentacja aktualnych stanów strumieni umożliwia skutecznie monitorowanie przesyłu danych i szybkie reagowanie na zaistniałe błędy.

Dane transportowane przez aplikację mają format JSON i składowane są jako odrębne rekordy w ramach tematu Apache Kafka. Aplikacja Strumienie JDBC w trakcie przesyłania danych może również weryfikować zgodność przesyłanych rekordów z wcześniej zdefiniowanym schematem JSON dla określonej encji. Pozwala to wyeliminować błędne dane z dalszego przetwarzania. W ramach strumienia, w zależności od potrzeb, można zdefiniować liczbę dopuszczalnych błędów, której przekroczenie spowoduje wstrzymanie dalszych zapytań i przejście strumienia w stan *BŁĄD*.

5.3.2. Menadżer Zarządzania Indeksami (MZI)

Zarządzanie różnymi wersjami kategorii danych, które dalej będziemy nazywać encjami, umożliwia Menadżer Zarządzania Indeksami (MZI). Wersjonowanie encji jest konieczne i wynika wprost z ciągłych modyfikacji, którym podlegają dane. Dla każdej wersji encji tworzona jest nowa dokumentacja API, jak również mogą być tworzone nowe reguły indeksowania, jeśli wymagają tego nowe filtry w API. Przy każdej zmianie struktury danych lub schematu zapytań wprowadzana jest nowa wersja metadanych, zawierająca aktualną dokumentacje Swagger i, jeśli to konieczne, nowe instrukcje indeksowania. Metadane encji składowane są w repozytorium z systemem kontroli wersji GIT. Dodatkowo, gdy zmiany pociągają za sobą modyfikacje schematu zapytań API, konieczne są też niewielkie zmiany w samym kodzie aplikacyjnym obsługującym zapytania klienta. MZI udostępnia interfejs użytkownika dla administratorów, który pozwala dodać lub usunąć encję, jak również zmienić jej wersję na inną dostępną w repozytorium.

Prześledźmy dodawanie nowej encji na przykładzie instytucji. Załóżmy, że wprowadzamy nową wersję instytucji o numerze 1.3. Każdej wersji encji odpowiada temat na Apache Kafka, który jest wcześniej zasilony danymi, w tym przykładzie temat będzie się nazywał INSTYTUCJE-1.3. Aby rozważana encja była dostępna z poziomu interfejsu MZI, jej metadane muszą się znajdować w repozytorium. Proces dodawania encji dla rozważanego przykładu zobrazowany jest na Ilustracji 5.4.



Ilustracja 5.4. Proces dodawania encji

Operacje wykonywane przez MZI w związku z udostępnianiem nowej wersji encji zostały oznaczone czarnym kolorem i ponumerowane. Omówimy je teraz zgodnie z przyjętą numeracją. **Operacja pierwsza** polega na pobraniu metadanych z repozytorium w momencie uruchomienia interfejsu użytkownika MZI. Zostają wtedy pobrane informacje o wszystkich dostępnych encjach, wśród których znajdzie się również ta z omawianego przykładu. Wśród metadanych znajdziemy zarówno instrukcje indeksowania, jak

również aktualną dokumentację Swagger. W kolejnym kroku, zilustrowanym jako **operacja druga**, MZI tworzy nowy indeks ES, do którego będą trafiały dokumenty zawierające dane instytucji w wersji 1.3. Po stworzeniu indeksu, MZI subskrybuje się do tematu INSTYTUCJE-1.3 **(operacja trzecia)** i zaczyna się strumieniowe pobieranie danych. Na podstawie instrukcji zawartych w metadanych następuje indeksowanie na klastrze ES rekordów pobranych z Apache Kafka **(operacja czwarta)**. Cały cykl kończy **operacja piąta**, kiedy to MZI zaczyna udostępniać nową dokumentację klientom. Wszystkie wspomniane operacje są wykonywane w bardzo krótkim czasie, zaledwie po kilku sekundach od udostępnienia nowej wersji klienci mogą już pobierać nowe dane.

Po zasubskrybowaniu się do tematów na Apache Kafka, MZI na bieżąco pobiera rekordy przez cały czas działania aplikacji. Pozwala to efektywnie dokonywać modyfikacji zindeksowanych dokumentów i utrzymywać aktualny stan danych. Każdy rekord na Apache Kafka ma unikalny identyfikator (UUID), stały dla określonego obiektu biznesowego. W ramach pojedynczego tematu może pojawić się wiele rekordów o tym samym UUID. Przykładem może być instytucja, której nazwa zmieniała się kilka razy. W tej sytuacji, po każdej zmianie nazwy, zostanie dodany nowy rekord. W trakcie procesu indeksowania UUID jest wykorzystywany jako identyfikator dokumentu w indeksie ES. Jeśli dokument o danym identyfikatorze nie istnieje w indeksie, to dodawany jest nowy dokument. W przeciwnym razie istniejący dokument jest zastępowany nowym. Pozwala to na bezobsługowe utrzymywanie aktualnych danych w indeksie. Zawsze zindeksowany jest ostatni rekord o określonym UUID. Szczególnym przypadkiem jest usunięcie dokumentu. Wówczas na Apache Kafka trafia rekord zawierający dodatkowe pole informujące o konieczności usunięcia elementu z indeksu. Na wypadek występowania takiego pola, MZI przeszukuje każdy rekord. Gdy pole się pojawi, z indeksu jest usuwany odpowiedni dokument serwisu internetowego RAD-on.

Zarządzanie przez MZI przepływem danych odbywa się za pomocą obiektów Topic Reader (TR). TR wiąże ze sobą kolekcje encji i odpowiada za skomunikowanie ze sobą wszystkich elementów systemu. TR może znajdować się w jednym z czterech stanów:

- *UTWORZONY:* TR został utworzony, ale nie czyta jeszcze danych z Kafki;
- *ZATRZYMANY:* czytanie danych z Kafki przez TR zostało wstrzymane;
- *URUCHOMIONY:* TR aktywnie czyta dane z Kafki;
- *ZAWIESZONY:* Na skutek poważnego błędu czytanie danych zostało zawieszone, wkrótce TR podejmie kolejną próbę odczytu.

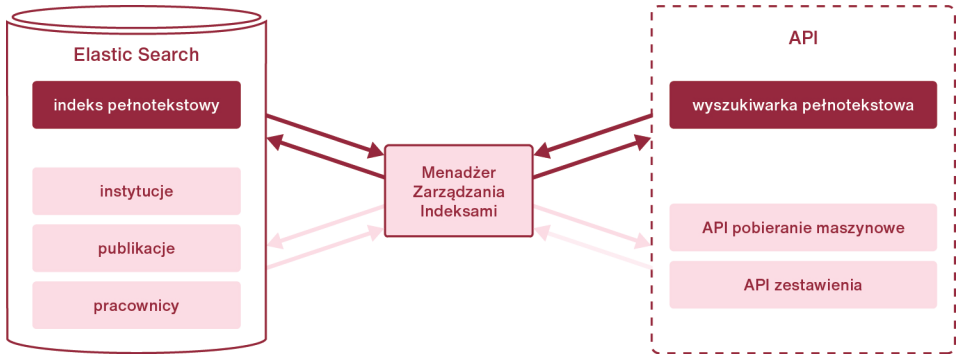
Podczas tworzenia nowego TR wybierana jest kolekcja encji, którymi będzie on zarządzał, a dla każdej z wybranych encji tworzony jest odrębny indeks na klastrze ES. Aby rozpocząć pobieranie danych z Apache Kafka, wykonywana jest akcja *START*, która powoduje subskrypcje TR do tematów z tożsamą wersją encji i równoległe pobieranie danych dla każdej z nich. Dane następnie są indeksowane na klastrze ES, a TR przechodzi w stan *URUCHOMIONY*. Jeśli pobrana z Apache Kafka kolekcja rekordów została bezbłędnie zindeksowania na klastrze ES, TR pobierze kolejną kolekcję nieprzeczytanych jeszcze rekordów i przystąpi do ich indeksowania. W przypadku błędu na którymś z etapów przetwarzania danych, TR przechodzi w stan *ZAWIESZONY* i po odczekaniu trzech minut następuje kolejna próba pobrania tych samych danych. Zapre-

zentowany mechanizm zapobiega wyciekowi danych, na przykład w sytuacji chwilowej awarii klastra ES bądź innego błędu, który na jakimś etapie przetwarzania uniemożliwiłby prawidłowe zindeksowanie pobranych rekordów. Każdy błąd, który doprowadził do restartu TR, jest utrwalany w bazie danych. Co godzinę MZI sprawdza tabelę zawierającą informacje o błędach. W przypadku pojawienia się nowych rekordów wysyła wiadomość e-mail informującą o awarii. Dzięki tej informacji w stosunkowo krótkim czasie można podjąć działania mające na celu naprawę problemu. Do zatrzymania przepływu danych służy akcja *STOP*. Wykonanie jej powoduje przejście TR w stan *ZATRZYMANY* i przerwanie procesu indeksowania serwisu internetowego RAD-on. Gdy TR jest w stanie *ZATRZYMANY*, możliwe jest modyfikowanie kolekcji encji, którymi zarządza, takie jak dodawanie nowych encji i usuwanie istniejących. Jeśli zachodzi potrzeba zmiany wersji encji, wykonuje się akcję *SKASUJ*, która usuwa wybraną kategorię danych, a wraz z nią jej indeks na klastrze ES, a następnie dodaje się ją jeszcze raz (akcja *DODAJ*), tym razem w nowej, pożądanej wersji. Utrzymywanie odrębnych indeksów dla encji pozwala w szybki sposób dokonywać zmian wersji. Utrzymywanie jednego wspólnego indeksu dla wszystkich encji wymagałoby zindeksowania wszystkich danych ponownie, co zajmowałoby bardzo dużo czasu i niepotrzebnie obciążało klaster ES. Przyjęte rozwiązanie pozwala zmodyfikować wersje pojedynczej encji bez konieczności przebudowywania pozostałych, zarządzanych przez TR.

Jeśli encje wchodzące w skład TR są również dostępne w wyszukiwarce pełnotekstowej serwisu internetowego RAD-on, budowany jest jeden dodatkowy indeks o uproszczonej strukturze, wykorzystywany specjalnie do tego celu. Dokumenty w tym indeksie zawierają tylko jedno pole tekstowe, umożliwiające wyszukiwanie. Pole to może łączyć różne informacje tekstowe wybrane indywidualnie dla każdej encji, zgodnie z potrzebami użytkowników. Taki indeks tworzony jest równolegle z pozostałymi, o których była mowa wyżej i zawiera wszystkie kategorie danych, które są udostępniane w wyszukiwarce serwisu internetowego RAD-on. Stworzenie indeksu pełnotekstowego pozwala pozycjonować wyniki wyszukiwania na podstawie zadanej przez użytkownika frazy, w obrębie wszystkich kategorii. W sytuacji, gdy określona encja jest usuwana z TR, dokumenty należące do tej encji są również usuwane z indeksu pełnotekstowego i przestają być dostępne w wynikach wyszukiwania.

5.3.3. Obsługa zapytań z API

W tym podrozdziale opiszemy proces komunikacji między API a silnikiem ES. System udostępnia trzy różne rodzaje API oparte o silnik wyszukiwania ES. Dwa z nich są wykorzystywane przez portal obywatelski. Trzecie API służy do bezpośredniego pobierania maszynowego danych. Przepływem zapytań za pośrednictwem aktywnego TR steruje MZI, co przedstawiono na Ilustracji 5.5



Ilustracja 5.5. Przepływ zapytań między API a silnikiem Elasticsearch

W zilustrowanym przykładzie widzimy klaster ES, na którym znajdują się trzy odrębne indeksy dla każdej obsługiwanej przez TR encji („instytucje", „publikacje", „pracownicy") i jeden indeks pełnotekstowy, który łączy wszystkie rekordy. Zapytania przychodzące z wyszukiwarki pełnotekstowej są przekierowywane do indeksu pełnotekstowego, co pozwala zwrócić najlepiej pasujące wyniki dla zadanej frazy spośród wszystkich kategorii. Wprowadzenie jednego indeksu pełnotekstowego pozwala wykorzystać pozycjonowanie wyników zgodne z algorytmem zaimplementowanym w ramach ES. Alternatywne rozwiązanie wymagałoby pobierania wyników ze wszystkich indeksów odrębnych i zaimplementowanie dodatkowego algorytmu, który by te wyniki pozycjonował. Zastosowanie takiego rozwiązania znacznie skomplikowałoby obsługę zapytań i zwiększyło czas oczekiwania na odpowiedź.

Indeksy odrębne są wykorzystywane zarówno przez użytkowników pobierających dane maszynowo, jak i przez zestawienia prezentowane na portalu obywatela. Każde zestawienie prezentuje dane jednej z encji i jest zasilane danymi z jednego z indeksów odrębnych. Względnie niewielkie wielkości indeksów skracają czas zapytań, przez co zapewniają dużą responsywność systemu. Wykorzystanie tego samego indeksu zarówno dla przetwarzania maszynowego, jak i zestawień dostępnych w portalach zapewnia spójność danych.

Poza danymi znajdującymi się w zestawieniach, RAD-on udostępnia również dane zintegrowane, które dostępne są jedynie za pośrednictwem API. Dla tych danych tworzony jest odrębny TR, w obrębie którego nie ma łączącego encje indeksu pełnotekstowego. W rezultacie, w ramach aplikacji funkcjonują zawsze dwa aktywne TR, z których pierwszy zajmuje się obsługą zapytań danych dostępnych w zestawieniach, a drugi udostępnia dane zintegrowane.

5.3.4. System logowania zdarzeń

System logowania zdarzeń służy do zbierania wszystkich żądań i odpowiedzi API. Dzięki temu możliwe jest logowanie ruchu i zdarzeń na utworzonych serwisach API. Ten

system logowania jest ważnym składnikiem architektury API, ponieważ pozwala zbierać ilościowe informacje o tym, jak serwisy API są wykorzystywane i jakie jest ich obciążenie.

Architektura logowania zdarzeń składa się z następujących kluczowych elementów:

- **artefakt:** dostarcza ujednoliconego interfejsu opisującego, jakie atrybuty i ich wartości są logowane. Określa on schemat logu z API oraz co ma się w nim znajdować. Aktualnie logowane są informacje takie, jak IP klienta oraz dane geolokalizacyjne, na przykład miejscowość, z której pochodzi zapytanie do API. Zbierane są także informacje dotyczące czasu przetwarzania zapytania, nazwy serwisu i wywołanej metody, a także odpowiedzi z serwisów API, wraz z ewentualnymi wyjątkami;
- **serwisy API:** implementują interfejs dostarczony przez artefakt logowania. Uzupełniają one określony schemat danych odpowiednimi danymi, czyli przypisują odpowiednie wartości atrybutom dostarczonym przez artefakt;
- **Apache Kafka:** transportuje logi z rozproszonego środowiska różnych serwisów API do wyznaczonych odbiorców;
- **Graylog**[46]**:** główny odbiorca logów z Apache Kafka, który przechowuje je w pełnotekstowym indeksie. Dzięki temu możliwe jest szybkie wstępne analizowanie informacji przepływających z systemów dziedzinowych API oraz tworzenie roboczych analiz i zestawień dotyczących obciążenia serwisów czy pojawiających się błędów;
- **hurtownia danych:** pobiera dane z aplikacji Graylog i, w zależności od potrzeb i wymogów biznesowych, dokonuje ich odpowiedniej analizy i agregacji w podziale na zadane okresy czasowe dla poszczególnych usług;
- **końcowy serwis RAD-on GUI (backend):** odpowiada za pobieranie przetworzonych i ujednoliconych informacji z hurtowni danych oraz prezentację ich w graficznym interfejsie użytkownika (RAD-on GUI).

5.3.5. System prezentacji danych

System prezentacji danych składa się z katalogu dostępnych usług API wraz z dokumentacją techniczną i instrukcją obsługi usług.

Dokumentacja jest tworzona za pomocą biblioteki **Swagger**, która umożliwia wygodne projektowanie i dokumentowanie API. Katalog dostępnych API znajduje się w portalu RAD-on. Katalog podzielony jest na dwa główne podkatalogi zawierające usługi udostępniania i edycji danych[47]. Aktualnie[48] dostępne są 34 usługi, dzięki którym użytkownik może pobrać interesujące go dane za pomocą publicznie dostępnych API. W przypadku katalogu edycji danych, dostępne są usługi dające dostęp do danych pochodzących z systemów dziedzinowych, takich jak POL-on oraz PBN, za pomocą których autoryzowani użytkownicy mogą dokonywać zmian w rejestrach danych.

[46] graylog.org (dostęp: 02.03.2023)
[47] radon.nauka.gov.pl/api/katalog-udostepniania-danych (dostęp: 02.03.2023)
[48] Stan na 21 grudnia 2022.

Takie moduły systemu RAD-on, jak: raporty[49], analizy[50] i dane[51] również działają, opierając się na usługach i pozyskanych przez nie danych. Moduły te zostały dokładnie opisane w rozdziale 2.

System RAD-on udostępnia również codziennie aktualizowane statystyki użycia wybranych API[52]. Statystyki te przedstawiają między innymi ilość udostępnionych w ramach systemu danych, liczbę pobrań poszczególnych usług oraz liczbę pobrań dokumentów zawierających informacje z sektora publicznego.

Warto dodać, że informacje o danych publikowanych w systemie RAD-on są zamieszczane w CRIP, czyli na portalu dane.gov.pl, o którym była już mowa w rozdziale 1.

5.4. Technologie

Opisywany system został wykonany z wykorzystaniem dziesięciu głównych technologii: Java Enterprise Edition, JDK8, Spring, Hibernate, hurtownia Oracle, ES, Apache Kafka, serwer aplikacji Tomcat, Graylog i Swagger. Ponadto, zmiany w wykorzystaniu i dostępie do nowych technologii są na bieżąco śledzone. Aktualizacje dokonywane są, gdy pojawiają się zarówno potrzeby techniczne (narzędzia wytwarzania oprogramowania), jak i biznesowe (nowe przypadki użycia, nowe funkcjonalności).

Elementy budujące API, w szczególności serwisy REST API, są skalowalne, dzięki ich konteneryzacji przy wykorzystaniu środowiska Kubernetes. Kubernetes (technologia zwana również K8s) to system, z otwartym kodem źródłowym, który służy do zarządzania chmurą obliczeniową i pozwala na automatyzację instalacji, skalowania i optymalizacji działania aplikacji w środowisku wielu maszyn wirtualnych lub fizycznych[53]. Jest to jedno z najpopularniejszych narzędzi do zarządzania aplikacjami w chmurze obliczeniowej. Kubernetes jest wykorzystywany przez wiele dużych firm technologicznych, takich jak Nokia, Spotify czy Yahoo[54]. Umożliwia on programistom skupienie się na tworzeniu aplikacji, a nie na zarządzaniu infrastrukturą, co pozwala na szybsze wdrażanie aplikacji i serwisów oraz elastyczne reagowanie na zmieniające się warunki.

[49] radon.nauka.gov.pl/raporty (dostęp: 02.03.2023)
[50] radon.nauka.gov.pl/analizy (dostęp: 02.03.2023)
[51] radon.nauka.gov.pl/dane (dostęp: 02.03.2023)
[52] radon.nauka.gov.pl/o-systemie/statystyki (dostęp: 02.03.2023)
[53] kubernetes.io (dostęp: 02.03.2023)
[54] kubernetes.io/case-studies (dostęp: 02.03.2023)

ZAKOŃCZENIE

Zgodnie ze stanem naszej wiedzy RAD-on jest obecnie największym pod względem zakresu zbieranych i udostępnianych danych narodowym systemem informatycznym z obszaru nauki i szkolnictwa wyższego w Europie. Jest to platforma analityczna prezentująca oficjalne statystyki, wpisująca się w ideę otwartych danych rządowych, danych FAIR oraz trend kreowania procesów politycznych i zarządczych na podstawie danych. RAD-on stanowi uzupełnienie ekosystemu IT w obszarze nauki i szkolnictwa wyższego, jaki tworzony jest w Polsce przez Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy od ponad dekady. Może służyć jako inspiracja do tworzenia podobnych systemów w innych państwach lub regionach.

Osiągnięcie poszczególnych celów projektu RAD-on było możliwe dzięki wprowadzanym stopniowo procesom zarządzania danymi i budowie rozległej infrastruktury informatycznej.

Integracja danych o szkolnictwie wyższym i nauce w Polsce pochodzących z różnych autonomicznych baz danych dokonana została poprzez wprowadzenie modelu wymiany danych.

Otwarty dostęp do najbardziej aktualnych i wiarygodnych danych dotyczących B+R, nauki i szkolnictwa wyższego zapewniony został przez stworzenie portalu obywatelskiego. Za jego pośrednictwem obywatele mają dostęp do zestawień danych zawierających rozbudowaną sekcję filtrów, funkcjonalność pobierania danych w formacie CSV/XLSX oraz wyszukiwarką pełnotekstową, umożliwiającą szybkie dotarcie do informacji. Osoby związane z sektorem nauki, w tym decydenci polityczni, mogą przeglądać gotowe raporty zawierające tabele i wykresy. Z myślą o dziennikarzach i badaczach za pośrednictwem portalu RAD-on publikowane są kompleksowe analizy pozwalające na dogłębną interpretację danych. Użytkownicy portalu mogą także w jednym miejscu zweryfikować swoje dane osobowe pochodzące z wielu systemów, a także zgłosić konieczność ich korekty. Potwierdzenie tożsamości przy użyciu usługi uwierzytelniającej login.gov.pl zapewnia odbiorcom usługi bezpieczeństwo.

Portal z pewnością jest najważniejszym, choć nie jedynym, efektem projektu. Ograniczenie biurokracji, między innymi poprzez odciążenie użytkowników w obszarze zasilania baz danymi, możliwe było dzięki budowie API, a także poprzez zwiększenie możliwości maszynowego przetwarzania i ponownego wykorzystania istniejących danych w innowacyjnych aplikacjach i usługach.

Szczególnym rozwiązaniem jest oferowana przez RAD-on alternatywa dla standardowych aplikacji *business intelligence* (BI), która wspiera procesy decyzyjne poprzez

dostarczenie właściwych narzędzi informatycznych umożliwiających analizę i interpretację udostępnianych danych. Te narzędzia służą wizualizacji i pobieraniu danych już przetworzonych, opracowanych przez ekspertów i zawierających odpowiedni komentarz.

System jest nadal rozwijany, co oznacza, że zakres udostępnianych w nim danych będzie się systematycznie zwiększał, a użyteczność poszczególnych usług będzie rosła. Planowane prace obejmują m.in. tworzenie raportów i analiz naukometrycznych, współpracę z podmiotami finansującymi naukę w zakresie udostępniania danych o przyznawanych w Polsce grantach na badania, promocję wykorzystania API w systemach wewnętrznych uczelni oraz przez badaczy sektora nauki, a także dalszy rozwój narzędzi analitycznych służących wizualizacji danych.

Najważniejszym wyzwaniem, przed którym stoimy jest sprawne dostosowywanie usług do zmieniającego się otoczenia prawnego. W trakcie prac nad projektem system szkolnictwa wyższego i nauki uległ znaczącej transformacji. Spodziewamy się dalszych zmian w zakresie gromadzonych danych i zasad ich przetwarzania.

Inną grupą wyzwań są te związane z warstwą UX platformy analitycznej. Na podstawie ankiety satysfakcji użytkowników szacujemy, że jeden na pięciu użytkowników systemu ma problem z odnalezieniem interesujących go danych. W większości przypadków jest to związane z faktem, że nie wszystkie dane możemy udostępnić. W części jest to natomiast kwestia pracochłonności procesu otwierania już przetworzonych danych. Najlepiej oceniana jest warstwa estetyczna systemu (64 proc. pozytywnych opinii i 29 proc. neutralnych), a także forma prezentacji danych i użyteczność usług, najgorzej zaś – łatwość dotarcia do informacji (50 proc. pozytywnych opinii, 28 proc. neutralnych oraz 22 proc. negatywnych). Użytkownicy zwracają uwagę na brak dostępności niektórych funkcji w wersji mobilnej systemu, chwalą natomiast stale poszerzaną ofertę raportów.

System RAD-on ma w zamierzeniu służyć każdemu, niezależnie od poziomu kompetencji cyfrowych i wiedzy merytorycznej o systemie szkolnictwa wyższego i nauki. Mierzymy się zatem z problemem, jak zapewnić równowagę między precyzją w opisie danych a ich zrozumieniem przez obywateli i dziennikarzy. Mimo, że projektowane usługi są konsultowane przez specjalistów UX, problemem pozostaje brak wytycznych w projektowaniu pulpitów z danymi z sektora publicznego. Należy także zdawać sobie sprawę z przyzwyczajania się użytkowników do wyglądu interfejsów i niechęci do zbyt często wprowadzanych zmian funkcjonalności. Uważamy, że jest to jeden z obszarów wymagających dalszych badań.

Mamy nadzieję, że podzielenie się doświadczeniami w projektowaniu architektury systemu RAD-on, a także omówienie korzyści i najważniejszych wyzwań związanych z tworzeniem systemu o takiej skali i celach będą przydatne dla wszystkich, którzy chcą projektować i wykorzystywać platformy analityczne jako narzędzia wspierające podejmowanie decyzji w sektorze nauki i szkolnictwa wyższego.

SPIS ILUSTRACJI

1.1. Systemy zintegrowane w ramach RAD-onu ... 23
2.1. Strona główna portalu RAD-on ... 28
2.2. Moduł „Dane" portalu RAD-on ... 29
2.3. Moduł „Analizy" portalu RAD-on ... 32
2.4. Wyniki wyszukiwania na portalu RAD-on ... 33
2.5. Sfederowany model tożsamości ... 34
2.6. Moduł „Konto użytkownika" portalu RAD-on – widok po zalogowaniu ... 35
2.7. Uproszczona architektura usługi dostępu do danych obywatela ... 36
2.8. Zarys architektury portalu obywatela ... 37
3.1. Warstwy kreatora raportów ... 42
3.2. Moduł „Raporty" portalu RAD-on ... 46
3.3. Przykład filtrów zależnych ... 47
4.1. Ogólny schemat działania modelu wymiany danych opierającego się na Apache Kafka ... 53
4.2. Podstawowe klasy obiektów występujące w modelu wymiany danych opierającym się na Apache Kafka ... 53
4.3. Ogólny schemat przepływu danych w modelu wymiany danych opierającym się na Apache Kafka ... 55
4.4. Schemat przepływu danych pomiędzy hurtownią danych a bazą wiedzy portalu RAD-on z wykorzystaniem modelu wymiany danych ... 57
4.5. Schemat przepływu wiadomości z wykorzystaniem modelu wymiany danych pomiędzy systemami dziedzinowymi a systemem PBN 2.0 ... 58
4.6. Proces tworzenia komunikatów administracyjnych z wykorzystaniem modelu wymiany danych na przykładzie ORPPD ... 60
4.7. Integracja systemów dziedzinowych z Systemem Obsługi Komunikatów (SOK) z wykorzystaniem modelu wymiany danych ... 61
4.8. Proces pobierania komunikatów administracyjnych po ponownym uruchomieniu aplikacji dziedzinowej z wykorzystaniem modelu wymiany danych na przykładzie ORPPD ... 62
4.9. Odczytanie komunikatów z SOK z użyciem modelu wymiany danych po ponownym uruchomieniu systemu dziedzinowego na przykładzie ORPPD i POL-on ... 63

4.10. Uproszczona architektura hurtowni danych 65
4.11. Biznesowe obszary danych 68
4.12. Biznesowy model danych dla obszaru „Proces awansowy” 68
4.13. Wykorzystanie hurtowni danych w procesie ewaluacji jakości działalności naukowej podmiotów systemu nauki i szkolnictwa wyższego w Polsce 69
5.1. Ogólna architektura utworzonego systemu 73
5.2. Zarys architektury systemu z uwzględnieniem przepływu danych 76
5.3. Transport danych między hurtownią danych a Apache Kafka 77
5.4. Proces dodawania encji 79
5.5. Przepływ zapytań między API a silnikiem Elasticsearch 82

BIBLIOGRAFIA

[1] A. Abduldaem, A. Gravell, *Principles for the design and development of dashboards: Literature review*, *Proceedings of INTCESS 2019 6th International Conference on Education and Social Science*, 2019, s. 1307–1316, ocerints.org/intcess19_e-publication/papers-412.pdf (dostęp: 02.03.2023).

[2] I. Alhassan, D. Sammon, M. Daly, *Critical success factors for data governance: A theory building approach*, *Information Systems Management*, 2019, 36(2), s. 98–110, DOI: 10.1080/10580530.2019.1589670.

[3] A. Andhavarapu, *Learning Elasticsearch*, Packt Publishing, 2017, ISBN 978-1787129917.

[4] B. Ballou, D. L. Heitger, L. Donnell, *Creating effective dashboards: How companies can improve executive decision making and board oversight*, *Strategic Finance*, 2010, 91(9), s. 27–33, go.gale.com/ps/i.do?id=GALE%7CA221186517&issn=1524833X&p=AONE&it=r&v=2.1&linkaccess=abs (dostęp: 02.03.2023).

[5] A. Białecki, R. Muir, G. Ingersoll, *Apache Lucene 4*, *Proceedings of the SIGIR 2012 Workshop on Open Source Information Retrieval*, 2012, s. 17–24.

[6] A. Bochenek red., *Systemy informatyczne wspierające naukę i szkolnictwo wyższe. OSF: Obsługa Strumieni Finansowania*, Ośrodek Przetwarzania Informacji – Państwowy Instytut Badawczy, 2021, opi.org.pl/wp-content/uploads/2021/12/OSF-ebook.pdf (dostęp: 02.03.2023).

[7] A. Brown, J. Fishenden, M. Thompson, W. Venters, *Appraising the impact and role of platform models and Government as a Platform (GaaP) in UK government public service reform: Towards a Platform Assessment Framework (PAF)*, *Government Information Quarterly*, 2017, 34(2), s. 167–182, DOI: 10.1016/j.giq.2017.03.003.

[8] W. Buchanan, A. Gobeo, C. Fowler, *GDPR and Cyber Security for Business Information Systems*, River Publishers, 2022, ISBN 978-1003338253, DOI: 10.1201/9781003338253.

[9] P. Budroni, J. Claude-Burgelman, M. Schouppe, *Architectures of knowledge: The European open science cloud*, *ABI Technik*, 2019, 39(2), s. 130–141, DOI: 10.1515/abitech-2019-2006.

[10] S. Burke, R. MacIntyre, G. Stone, *Library data labs: Using an agile approach to develop library analytics in UK higher education*, *Information and Learning Science*, 2018, 119(1/2), s. 5–15, DOI: 10.1108/ILS-05-2017-0035.

[11] E. Cebrián, J. Domenech, *Is Google Trends a quality data source?*, *Applied Economics Letters*, 2023, 30(6), s. 811–815, DOI: 10.1080/13504851.2021.2023088.

[12] I. Chengalur-Smith, D. Ballou, H. Pazer, *The impact of data quality information on decision making: an exploratory analysis*, *IEEE Transactions on Knowledge and Data Engineering*, 1999, 11(6), s. 853–864, DOI: 10.1109/69.824597.

[13] M. Christopher, *The Agile Supply Chain: Competing in Volatile Markets*, *Industrial Marketing Management*, 2000, 29(1), s. 37–44, DOI: 10.1016/S0019-8501(99)00110-8.

[14] X. Chu, I. F. Ilyas, S. Krishnan, J. Wang, *Data Cleaning: Overview and Emerging Challenges*, *Proceedings of the 2016 International Conference on Management of Data*, SIGMOD '16, New York, NY, USA, 2016, Association for Computing Machinery, s. 2201–2206, DOI: 10.1145/2882903.2912574.

[15] D. Deng, W. Tao, Z. Abedjan, A. K. Elmagarmid, I. F. Ilyas, S. Madden, M. Ouzzani, M. Stonebraker, N. Tang, *Entity consolidation: The golden record problem*, *ArXiv e-prints*, 2017, 1709.10436v2, DOI: 10.48550/arXiv.1709.10436.

[16] N. Denwattana, A. Saengsai, *A framework of Thailand higher education dashboard system*, *2016 International Computer Science and Engineering Conference (ICSEC)*, 2016, s. 1–6, DOI: 10.1109/ICSEC.2016.7859883.

[17] U. Dombrowski, T. Mielke, C. Engel, *Knowledge Management in Lean Production Systems*, *Procedia CIRP*, 2012, 3, s. 436–441, ISSN 2212-8271, DOI: 10.1016/j.procir.2012.07.075.

[18] C. El Morr, H. Ali-Hassan, *Descriptive, Predictive, and Prescriptive Analytics*, *Analytics in Healthcare: A Practical Introduction*, rozdz. 3, s. 31–55, ISBN 978-3030045067, DOI: 10.1007/978-3-030-04506-7_3.

[19] S. Few, *Information Dashboard Design: The Effective Visual Communication of Data*, O'Reilly, Sebastopol, CA, 2006.

[20] Y. Gao, M. Janssen, C. Zhang, *Understanding the evolution of open government data research: Towards open data sustainability and smartness*, *International Review of Administrative Sciences*, 2021, DOI: 10.1177/00208523211009955.

[21] N. Garg, *Apache Kafka*, Packt Publishing, 2013, ISBN 978-1782167938.

[22] M. Gibbons, C. Limoges, H. Nowotny, S. Schwartzman, P. Scott, M. Trow, *The New Production of Knowledge: The Dynamics of Science and Research in Contemporary Societies*, SAGE Publications Ltd, London, 2010, DOI: 10.4135/9781446221853.

[23] R. Gitzel, S. Turring, S. Maczey, *A data quality dashboard for reliability data*, *2015 IEEE 17th Conference on Business Informatics*, tom 1, 2015, s. 90–97, DOI: 10.1109/CBI.2015.24.

[24] C. Gormley, Z. Tong, *Elasticsearch: The definitive guide*, O'Reilly, wyd. 1, 2015, ISBN 978-1449358549.

[25] I. Guitart, J. Conesa, *Adoption of Business Strategies to Provide Analytical Systems for Teachers in the Context of Universities*, *International Journal of Emerging Technologies in Learning (iJET)*, 2016, 11(7), s. 34–40, DOI: 10.3991/ijet.v11i07.5887.

[26] B. R. Hiraman, C. Viresh M., K. Abhijeet C., *A Study of Apache Kafka in Big Data Stream Processing*, *2018 International Conference on Information, Communication, Engineering and Technology (ICICET)*, 2018, s. 1–3.

[27] *How to make your data FAIR*, openaire.eu/how-to-make-your-data-fair (dostęp: 02.03.2023).

[28] P. Howard, *Total cost of ownership for business intelligence*, crozdesk.com/software-research/total-cost-of-ownership-for-business-intelligence (dostęp: 02.03.2023).

[29] P. Huston, V. L. Edge, E. Bernier, *Reaping the benefits of Open Data in public health*, *Canada Communicable Disease Report*, 2019, 45(10), s. 252–256, DOI: 10.14745/ccdr.v45i10a01.

[30] D. Ince, *A Dictionary of the Internet*, Oxford University Press, Oxford, wyd. 4, 2019, DOI: 10.1093/acref/9780191884276.001.0001.

[31] J. Jabłecka, B. Lepori, *Between historical heritage and policy learning: the reform of public research funding systems in Poland, 1989–2007*, *Science and Public Policy*, 2009, 36(9), s. 697–708, ISSN 0302-3427, DOI: 10.3152/030234209X475263.

[32] E. Kulczycki, *Assessing publications through a bibliometric indicator: The case of comprehensive evaluation of scientific units in Poland*, *Research Evaluation*, 2017, 26(1), s. 41–52, ISSN 0958-2029, DOI: 10.1093/reseval/rvw023.

[33] P. Le Noac'h, A. Costan, L. Bougé, *A performance evaluation of Apache Kafka in support of big data streaming applications*, *2017 IEEE International Conference on Big Data*, 2017, s. 4803–4806, DOI: 10.1109/BigData.2017.8258548.

[34] W. L. Martinez, *Graphical user interfaces*, *WIREs Computational Statistics*, 2011, 3(2), s. 119–133, DOI: 10.1002/wics.150.

[35] R. Matheus, M. Janssen, D. Maheshwari, *Data science empowering the public: Data-driven dashboards for transparent and accountable decision-making in smart cities*, *Government Information Quarterly*, 2020, 37(3), s. 101.284, ISSN 0740-624X, DOI: 10.1016/j.giq.2018.01.006.

[36] M. Michajłowicz, M. Niemczyk, J. Protasiewicz, K. Mroczkowska, *POL-on: The Information system of science and higher education in Poland*, *European Journal of Higher Education IT 2018-1*, 2018, ISSN 2519-1764, eunis.org/download/2018/EUNIS_2018_paper_70.pdf (dostęp: 02.03.2023).

[37] D. A. Norman, S. W. Draper, red., *User Centered System Design: New Perspectives on Human-computer Interaction*, Lawrence Erlbaum Associates, wyd. 1, 1986.

[38] H. Nowotny, P. Scott, M. Gibbons, *Re-Thinking Science: Knowledge and the Public in an Age of Uncertainty*, Wiley, 2001, ISBN 978-0745626086.

[39] OECD, *Open, Useful and Re-usable data (OURdata) Index: 2019*, *OECD Public Governance Policy Papers*, 2020, 01, DOI: 10.1787/45f6de2d-en.

[40] D. Power, *A Brief History of Decision Support Systems*, wersja 4.0, 2007, DSSresources.COM/history/dsshistory.html (dostęp: 02.03.2023).

[41] J. Protasiewicz, E. Podwysocki, S. Ostrowska, A. Tomczyńska, *Open access to data on higher education and science: A case study of the RAD-on platform in Poland*, S. Bolis, J.-F. Desnos, L. Merakos, R. Vogl, red., *Proceedings of the European University Information Systems Conference 2021*, tom 78 w *EPiC Series in Computing*, EasyChair, 2021, s. 9–21, DOI: 10.29007/gz8q.

[42] J. Protasiewicz, S. Rosiak, I. Kucharska, E. Podwysocki, M. Niemczyk, M. Michajłowicz, *RAD-on: An integrated System of Services for Science – Online Elections for the Council of Scientific Excellence in Poland*, *European Journal of Higher Education IT 2019-1*, 2019, eunis.org/download/2019/EUNIS_2019_paper_20.pdf (dostęp: 02.03.2023).

[43] A. Ramesh Babu, Y. Singh, *Determinants of research productivity*, *Scientometrics*, 1998, 43, s. 309–329, DOI: 10.1007/BF02457402.

[44] J. Richardson, R. Sallam, K. Schlegel, A. Kronz, J. Sun, *Magic quadrant for analytics and business intelligence platforms*, 2020, Gartner ID G00386610.

[45] G. Rieder, J. Simon, *Datatrust: Or, the political quest for numerical evidence and the epistemologies of Big Data*, *Big Data & Society*, 2016, 3(1), DOI: 10.1177/2053951716649398.

[46] *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*, Dz. Urz. UE. L Nr 119, 4.5.2016.

[47] *Rozporządzenie Ministra Nauki i Szkolnictwa Wyższego z dnia 6 marca 2019 r. w sprawie danych przetwarzanych w Zintegrowanym Systemie Informacji o Szkolnictwie Wyższym i Nauce POL-on*, Dz.U. 2019 poz. 496.

[48] E. Ruijer, F. Détienne, M. Baker, J. Groff, A. J. Meijer, *The Politics of Open Government Data: Understanding Organizational Responses to Pressure for More Transparency*, *The American Review of Public Administration*, 2020, 50(3), s. 260–274, DOI: 10.1177/0275074019888065.

[49] B. Scholtz, A. Calitz, R. Haupt, *A business intelligence framework for sustainability information management in higher education*, *International Journal of Sustainability in Higher Education*, 2018, 19(2), s. 266–290, ISSN 1467-6370, DOI: 10.1108/IJSHE-06-2016-0118.

[50] B. A. Schwendimann, M. J. Rodríguez-Triana, A. Vozniuk, L. P. Prieto, M. S. Boroujeni, A. Holzer, D. Gillet, P. Dillenbourg, *Perceiving Learning at a Glance: A Systematic Literature Review of Learning Dashboard Research*, *IEEE Transactions on Learning Technologies*, 2017, 10(1), s. 30–41, DOI: 10.1109/TLT.2016.2599522.

[51] W. Solesbury, *The Ascendancy of Evidence*, *Planning Theory & Practice*, 2002, 3(1), s. 90–96, DOI: 10.1080/14649350220117834.

[52] A. Sorour, A. S. Atkins, C. F. Stanier, F. D. Alharbi, *The Role of Business Intelligence and Analytics in Higher Education Quality: A Proposed Architecture*, *2019 International Conference on Advances in the Emerging Computing Technologies (AECT)*, 2020, s. 1–6, DOI: 10.1109/AECT47998.2020.9194157.

[53] D. Spichtinger, J. Siren, *1. The Development of Research Data Management Policies in Horizon 2020*, J. B. Thestrup, F. Kruse, red., *Research Data Management – A European Perspective*, s. 11–24, ISBN 978-3110365634, DOI: 10.1515/9783110365634-002.

[54] S. D. Teasley, *Student Facing Dashboards: One Size Fits All?*, *Technology, Knowledge and Learning*, 2017, 22(3), s. 377–384, ISSN 2211-1670, DOI: 10.1007/s10758-017-9314-3.

[55] B. V. Thummadi, O. Shiv, K. Lyytinen, *Enacted Routines in Agile and Waterfall Processes*, *2011 Agile Conference*, 2011, s. 67–76, DOI: 10.1109/AGILE.2011.29.

[56] *Ustawa z dnia 3 marca 2018 r. – Prawo o szkolnictwie wyższym i nauce*, Dz.U. 2022 poz. 574 z późn. zm.

[57] A. F. van Veenstra, B. Kotterink, *Data-Driven Policy Making: The Policy Lab Approach*, P. Parycek, Y. Charalabidis, A. V. Chugunov, P. Panagiotopoulos, T. A. Pardo, O. Sæbø, E. Tambouris, red., *Electronic Participation*, Lecture Notes in Computer Science, Cham, 2017, Springer International Publishing, s. 100–111, DOI: 10.1007/978-3-319-64322-9_9.

[58] D. Vohra, *Apache Kafka*, *Practical Hadoop Ecosystem: A Definitive Guide to Hadoop-Related Frameworks and Tools*, Berkeley, CA, 2016, Apress, s. 339–347, DOI: 10.1007/978-1-4842-2199-0_9.

[59] A. Vázquez-Ingelmo, F. J. Garcia-Peñalvo, R. Therón, *Information Dashboards and Tailoring Capabilities – A systematic Literature Review*, *IEEE Access*, 2019, 7, s. 109.673–109.688, DOI: 10.1109/ACCESS.2019.2933472.

[60] V. Weerakkody, Z. Irani, K. Kapoor, U. Sivarajah, Y. K. Dwivedi, *Open data and its usability: An empirical view from the Citizen's perspective*, *Information Systems Frontiers*, 2017, 19(2), s. 285–300, DOI: 10.1007/s10796-016-9679-1.

[61] *What is open data*, data.europa.eu/trening/what-open-data (dostęp: 02.03.2023).

[62] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, *The FAIR Guiding Principles for scientific data management and stewardship*, *Scientific Data*, 2016, 3, s. 160.018, DOI: 10.1038/sdata.2016.18.

[63] B. Williamson, *The hidden architecture of higher education: building a big data infrastructure for the 'smarter university'*, *International Journal of Educational Technology in Higher Education*, 2018, 15(1), s. 12, DOI: 10.1186/s41239-018-0094-1.

[64] H. Wu, Z. Shang, K. Wolter, *Performance prediction for the Apache Kafka messaging system*, *2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, 2019, s. 154–161, DOI: 10.1109/HPCC/SmartCity/DSS.2019.00036.

[65] M. Zavyiboroda, *Why user interface changes annoy people and how to handle it*, speckyboy.com/annoying-ui-changes (dostęp: 02.03.2023).

[66] N. H. Zulkifli Abai, J. Yahaya, A. Deraman, A. R. Hamdan, Z. Mansor, Y. Yah Jusoh, *Integrating Business Intelligence and Analytics in Managing Public Sector Performance: An Empirical Study*, *International Journal on Advanced Science, Engineering and Information Technology*, 2019, 9(1), s. 172–180, DOI: 10.18517/ijaseit.9.1.6694.

[67] L. Šereš, V. Pavlićević, P. Tumbas, *Digital transformation of higher education: Competing on analytics*, *INTED2018 Proceedings*, 12th International Technology, Education and Development Conference, IATED, 2018, s. 9491–9497, DOI: 10.21125/inted.2018.2348.