Dariusz Laskowski
*Globema Sp. z o.o.*

# Efektywne zbieranie informacji o infrastrukturze technicznej w terenie ze wsparciem sztucznej inteligencji

## Effective field data collection on technical infrastructure with support of artificial intelligence

### Wstęp

Skuteczne zebranie kompletnych i wartościowych danych z terenu jest niezmiennie, od lat, dużym problemem przedsiębiorstw i instytucji dysponujących majątkiem w terenie. Wąskie gardło stanowi styk człowieka i maszyny, tj. osób odpowiedzialnych za realizację czynności utrzymaniowych (monterów, serwisantów, obchodowych) oraz różnego rodzaju rozwiązań IT, które wprawdzie próbują oferować użytkownikowi przyjazny interfejs do wprowadzania danych, lecz nadal według klasycznego podejścia „przenoszenia kawałka biura w teren". Są to więc formularze do wprowadzania danych czy mapy z lokalizacją GPS naśladujące swoje analogowe odpowiedniki, które w trudnych warunkach terenowych często nie zdają egzaminu.

Ma to dwojakie konsekwencje: wiele firm, które próbowały wdrożyć procedury zbierania danych w terenie, poniosło porażkę spotykając się z frustracją i oporem pracowników terenowych, zaś firmy, które wdrożyły i z determinacją egzekwują te procedury, ponoszą duże koszty, które w skrajnym przypadku mogą przewyższyć zyski płynące z korzystania z pozyskanych danych.

Chcąc zaradzić temu problemowi i przełamać barierę interakcji człowiek – maszyna (Human – Machine Interaction, HMI) w profesjonalnych zastosowaniach technicznych, podjęliśmy się realizacji projektu badawczo-rozwojowego, który pozwoliłby nam to osiągnąć.

Celem projektu było stworzenie nowatorskiej platformy informatycznej o nazwie GlobIQ [wym. globajkju albo globik], wspierającej zbieranie informacji o obiektach majątku (głównie technicznego), rozproszonych w przestrzeni, w sposób naturalny – bazujący na komunikacji za pomocą głosu i obrazu – oraz konwersję tych informacji do postaci strukturalnej, umożliwiającej dalsze przetwarzanie w specjalistycznych aplikacjach biznesowych. A wszystko dzięki zaawansowanym metodom, opartym na sztucznej inteligencji, uzupełnionym o dane z sensorów urządzenia mobilnego, takich jak odbiornik GPS, kompas czy żyroskop.

*Zapraszamy do przyjrzenia się bliżej wybranym zagadnieniom, które mieliśmy okazję przebadać w projekcie, oraz rezultatom tych badań.*

### Interpretacja mowy

Mowa jest naturalnym sposobem komunikacji międzyludzkiej. Chcąc przekonać się, na ile efektywny jest to sposób w przypadku komunikacji na linii człowiek – maszyna, postanowiliśmy zbadać możliwości wykorzystania interpretacji mowy w trzech obszarach:

- wybudzania aplikacji,
- sterowania aplikacją,
- wprowadzania danych.

#### Wybudzanie aplikacji

Wybudzanie aplikacji ma na celu wprowadzenie uśpionej aplikacji w tryb aktywnego nasłuchu i interpretacji poleceń głosowych, np. sterowania aplikacją, wprowadzania i poprawiania danych, itp. Odbywa się za pomocą tzw. słowa wybudzającego (ang. wake word, hotword), którego mechanizm polega na ciągłej analizie dźwięku rejestrowanego przez mikrofon (telefonu komórkowego, słuchawki bluetooth, itp.) w poszukiwaniu wystąpienia zdefiniowanego słowa lub frazy aktywującej. Proces w całości realizowany jest na urządzeniu końcowym, bez konieczności wysyłania strumienia audio do zewnętrznych serwisów *speech-to-text*.

W toku prac nad mechanizmem *wake word* zbadaliśmy trzy dostępne w tamtym czasie narzędzia, realizujące funkcję wykrywania słowa wybudzającego, działające w systemie operacyjnym Android: PocketSphinx (open source), Porcupine (komercyjne) oraz Snowboy (komercyjne).

Przystępując do ich oceny, wzięliśmy pod uwagę następujące kryteria:

- skuteczność wykrywania słowa wybudzającego w sprzyjających warunkach,
- skuteczność wykrywania słowa wybudzającego w trudnych warunkach (szum, hałas),
- możliwość użycia słowa wybudzającego w języku polskim,
- możliwość douczania własnymi nagraniami audio,
- wrażliwość wyuczonego modelu na zmianę warunków otoczenia,
- koszt rozwiązania.

Na podstawie przeprowadzonych eksperymentów zadecydowaliśmy, że wykorzystamy mechanizm oferowany przez PocketSphinxa. Po wyborze narzędzia przyszła kolej na przygotowanie własnego słowa wybudzającego. Nasza intuicja, uzupełniona eksperymentami, podpowiedziała nam, że najlepiej sprawdzi się słowo wybudzające:

- trzysylabowe lub dłuższe,
- zakończone samogłoską,
- zawierające spółgłoski dźwięczne (b, d, g, w, z, ź, ż, l, ł, r, m, n, j, dz, dź, dż).

Rozważaliśmy kilka opcji, ale ostatecznie wybór padł na słowo „globiku", które spełnia większość wymienionych założeń. Na stronach projektu PocketSphinx dostępnych jest wiele modeli ogólnego przeznaczenia dla wybranych języków. Aby wykorzystać PocketSphinxa do wykrywania hotworda, zaleca się, aby uniwersalny model dotrenować nagraniami audio, co też uczyniliśmy.

Najlepsze wyniki osiągaliśmy po dotrenowaniu uniwersalnych modeli en-us-ptm oraz de-ptm. Trudno jest jednoznacznie wskazać, jaka jest odpowiednia liczba nagrań douczających. Pojedynczy eksperyment (92 nagrania douczające od dziewięciu osób, 92+76 nagrań testowych „globiku", 24 min nagrań „bez globiku") pokazał, że dotrenowane modele, w subiektywnej ocenie, są już dość dobre.

### *Trudności związane z wybudzaniem aplikacji*

Do typowych problemów, związanych z wykorzystaniem słowa wybudzającego, należy zaliczyć możliwość przypadkowego wybudzenia aplikacji innym słowem (co u nas zdarzało się bardzo rzadko) czy brak wybudzenia, mimo użycia właściwego słowa. Takie sytuacje miały miejsce częściej, co może być irytujące dla użytkownika. Oprócz tego trzeba liczyć się ze zwiększonym zużyciem energii przez urządzenie. Z wymienionych powodów nie polecamy korzystać z wybudzania głosowego w przypadku, gdy użytkownik korzysta z aplikacji częściowo w sposób tradycyjny, trzymając urządzenie w dłoni, ponieważ można je zastąpić przyciskiem uruchamiającym aktywny nasłuch na żądanie. Natomiast z pewnością może mieć zastosowanie w sytuacjach, w których użytkownik ma zajęte obie ręce i chciałby obsługiwać aplikację wyłącznie za pomocą głosu, np. podczas prac na wysokości.

## Sterowanie aplikacją i wprowadzanie danych

Wprowadzanie danych przy pomocy małej klawiatury ekranowej urządzenia mobilnego nie jest zbyt wygodne, szczególnie dla osób o większych dłoniach, stąd pomysł, by wykorzystać do tego celu mowę. W poszukiwaniu efektywnego sposobu komunikacji głosowej pomiędzy użytkownikiem a urządzeniem mobilnym, przetestowaliśmy dwie opisane poniżej koncepcje.

### *Asystent głosowy*

W tej koncepcji komunikacja przebiega w opisany poniżej sposób.

1. Użytkownik wybudza asystenta za pomocą słowa wybudzającego, a następnie wydaje polecenie uruchamiające wybraną funkcję aplikacji, np. wprowadzania informacji o nowym obiekcie.
2. Asystent zadaje użytkownikowi pytania, uruchamiając za każdym razem mikrofon i oczekując na odpowiedź.
3. Pytania zadawane są jedno po drugim do momentu zebrania wszystkich danych i użytkownik nie ma możliwości ingerowania w ich kolejność, poza użyciem słowa „cofnij", aby powrócić do poprzedniego pytania.
4. Dialog z asystentem wyświetlany jest na ekranie urządzenia w formie czatu, gdzie widoczne są zarówno pytania zadane przez asystenta, jak i odpowiedzi udzielone przez użytkownika.

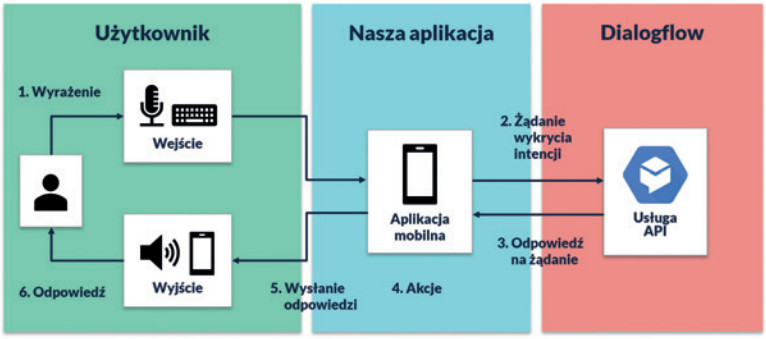

Przykład dialogu z asystentem

W celu skrócenia i przyspieszenia komunikacji, doświadczony użytkownik może wydawać polecenia wraz z informacjami, które ich dotyczą, np.:

Dodaj złącze Mickiewicza 8 o numerze 12345
polecenie nazwa złącza numer złącza

Dzięki temu asystent nie zadaje pytań o podane informacje, ograniczając się jedynie do zebrania pozostałych danych.

Do stworzenia asystenta wykorzystaliśmy narzędzie Dialogflow firmy Google. Podany schemat obrazuje interakcje zachodzące między użytkownikiem, aplikacją i usługą Dialogflow.



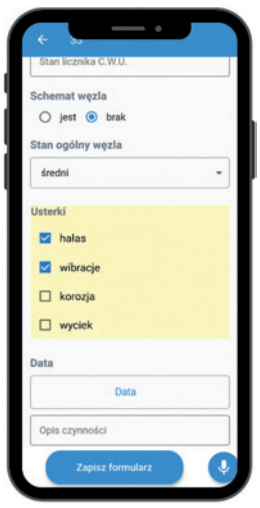

Asystent głosowy okazał się dobrym rozwiązaniem jako element sterowania aplikacją oraz w przypadku, gdy trzeba poprowadzić użytkownika „za rękę" przez jakiś proces, poprosić o wykonanie zdjęcia czy innej czynności. Sprawdzi się szczególnie tam, gdzie ilość pozyskiwanych informacji jest stosunkowo niewielka. Nie możemy go natomiast polecić jako efektywnego narzędzia osobom, które na co dzień wprowadzają masowo, w powtarzalny sposób, duże ilości danych. Tutaj asystent okazał się po prostu zbyt wolny.

*Hurtowe dyktowanie danych*

W przypadku zbierania informacji o obiektach majątku rozproszonych w terenie często mamy do czynienia z sytuacją, w której pracownik zbiera szereg informacji dotyczących odwiedzonego obiektu, notując je na papierowym lub elektronicznym formularzu. A gdyby tak, zamiast odręcznego zapisywania, mógł te informacje po prostu swobodnie podyktować? W tym celu opracowaliśmy, alternatywny dla asystenta głosowego, sposób komunikacji użytkownika z aplikacją, a mianowicie hurtowe dyktowanie danych, z analizą wypowiedzi i wyświetlaniem wyników w czasie rzeczywistym.

**Jak to działa?**

- Na ekranie urządzenia wyświetlany jest formularz z polami do wypełnienia.
- Użytkownik włącza nasłuch, a następnie dyktuje informacje, np.: **temperatura** *dwadzieścia (stopni)*, **ciśnienie zasilania** *jeden przecinek cztery (bara)*, **schemat węzła** *brak*, **stan węzła** *średni*, **usterki** *hałas i wibracje,* itd.
- Wypowiedź użytkownika analizowana jest w czasie rzeczywistym.
- Po wypowiedzeniu nazwy pola, aplikacja podświetla na ekranie odpowiednie miejsce, dzięki czemu użytkownik ma wizualne potwierdzenie swojej wypowiedzi już podczas mówienia.
- Wypowiadane dane wpisywane są do formularza i wyświetlane na bieżąco, w trakcie mówienia.
- Nasłuch kończy się po kilku sekundach ciszy lub po dotknięciu ikony mikrofonu.




Formularz
z podświetlonym polem
„Usterki"

Dzięki wprowadzeniu tzw. aliasów dla nazw pól, użytkownik może wypowiadać nazwę danego pola na wiele sposobów, niekoniecznie dokładnie w takiej postaci, w jakiej widnieje ona na formularzu. Alternatywnie może również dyktować same wartości, bez nazw pól, zgodnie z kolejnością ich występowania na formularzu.

Do realizacji wyżej opisanego sposobu komunikacji wykorzystaliśmy usługę SpeechToText firmy Google, zamieniającą mowę na tekst, w połączeniu z napisanym przez nas analizatorem tekstu.

Analizator składa się z czterech części.

1. **Source** – dostarcza kolejne słowa ze źródła.
2. **Parser** – czyta kolejne słowa za pomocą źródła i dopasowując je do nazw kategorii i pól w formularzu generuje tokeny zawierające informacje o polu formularza, jego wartości oraz ewentualnej jednostce.
3. **StringValuesConverter** – dopasowuje wartości typu tekstowego (String) do typów wymaganych przez pola formularza.
4. **FormValuesSetter** – ustawia odpowiednie źródło, czyta strumień tokenów parsera i ustawia odpowiednie wartości pól na formularzu.

Po przeanalizowaniu wypowiedzi aplikacja przewija ekran i podświetla pole wskazane w ostatnim tokenie przekazanym przez parser, czyli przechodzi do pola, którego nazwa została wypowiedziana jako ostatnia. Następnie użytkownik wypowiada wartość, którą aplikacja wpisuje do podświetlonego pola. Podświetlenie pola lub wpisanie wartości następuje po każdym cząstkowym rozpoznaniu mowy przez usługę SpeechToText, tak aby użytkownik miał wrażenie pracy w czasie rzeczywistym.

W przeciwieństwie do asystenta głosowego, ten sposób pracy z aplikacją sprawdzi się również tam, gdzie zbierane są bardzo duże ilości danych.

## Interpretacja obrazu

Dzięki upowszechnieniu się urządzeń mobilnych, zawierających wbudowany aparat fotograficzny, wykonywanie zdjęć stało się niemal standardowym elementem wizyt terenowych na obiektach majątku. Zdjęcia są, co prawda, cennym źródłem informacji, ale dane, które się na nich znajdują, nie mogą być wykorzystane od razu przez firmowy system czy przetworzone dalej. Zdjęcia wymagają bowiem pobrania z urządzenia mobilnego, uporządkowania, spisania potrzebnych danych i wprowadzenia do firmowej bazy. To zajmuje cenny czas i może być źródłem błędów.

I tutaj z pomocą przychodzi rozpoznawanie obrazu. Dzięki wykorzystaniu sztucznej inteligencji i aparatu wbudowanego w urządzenie mobilne, specjalistyczna aplikacja jest w stanie automatycznie rozpoznawać obiekty oraz ich cechy, a następnie uzupełnić odpowiednimi danymi bazę danych. Rola użytkownika sprowadza się do wykonania zdjęcia, a następnie weryfikacji i akceptacji wyświetlonych wyników rozpoznania.

### Klasyfikacja obrazu

Klasyfikacja obrazu to podstawowe zadanie z obszaru rozpoznawania obrazu, którego celem jest zrozumienie i zaklasyfikowanie całego obrazu (wykonanego zdjęcia) do jednej z ustalonych kategorii. Mogą to być np. różne rodzaje zwierząt: psy, konie, krowy czy samochodów: osobowy, dostawczy, ciężarowy. W naszym przypadku jedną kategorię stanowiła kombinacja dwóch cech energetycznego złącza kablowego:

1) rodzaj złącza – wolno stojące, wnękowe,
2) rodzaj obudowy – betonowa, metalowa, z tworzywa,

co w efekcie dało nam sześć kategorii (2 rodzaje złącza x 3 rodzaje obudowy).

**Jak to działa w praktyce?**

- Użytkownik wykonuje zdjęcie złącza.
- Wyuczony model klasyfikacyjny automatycznie rozpoznaje, do której kategorii należy wykonane zdjęcie i prezentuje wynik rozpoznania użytkownikowi.
- Użytkownik zapoznaje się z wynikiem rozpoznania i akceptuje go lub wprowadza korektę, jeżeli zachodzi taka potrzeba.
- Aplikacja zapisuje zdjęcie oraz wynik w postaci strukturalnej, gotowej do dalszego wykorzystania.


Ekran z wynikiem rozpoznania obrazu

Do nauki modelu wykorzystaliśmy chmurowy produkt do klasyfikacji zdjęć – Google Cloud AutoML Vision (obecnie zastąpiony przez platformę Vertex AI) – uzyskując skuteczność przekraczającą 95%. AutoML Vision pozwolił na realizację całego procesu uczenia, począwszy od utworzenia zbioru do nauki, poprzez trening modelu, jego ocenę, aż po eksport modelu do wdrożenia na urządzeniu mobilnym. Znacząca część prac leżała po stronie przygotowania zbioru do nauki, tj. zebrania i sklasyfikowania zdjęć, które powinny spełniać m.in. następujące warunki: przedstawiać złącza pod różnymi kątami, w różnych rozdzielczościach i na różnym tle, przy rekomendowanej liczbie około 1000 zdjęć per kategoria (w naszym przypadku było to kilkaset zdjęć per kategoria).

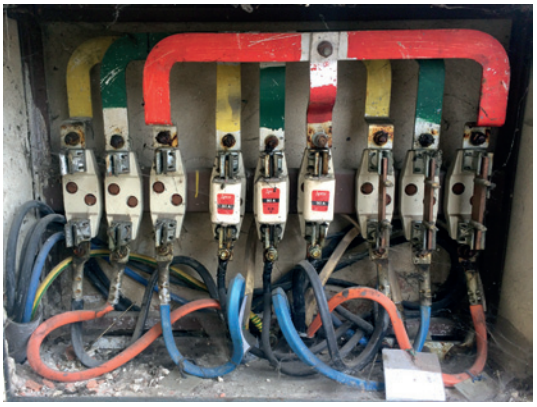
### Detekcja obiektów i OCR

Podczas zbierania informacji na temat obiektów majątku technicznego w terenie zazwyczaj zachodzi konieczność ich szczegółowego opisu. Po pierwsze, pracownik terenowy stwierdza sam fakt istnienia bądź braku obiektu, po drugie określa jego rodzaj, a następnie zbiera szczegółowe dane na jego temat, jak np. typ katalogowy. Dzięki wykorzystaniu rozpoznawania obrazu, a dokładniej detekcji obiektów i OCR, szczegółowe dane o obiektach mogą być zbierane w sposób automatyczny. Jak to działa, pokażemy na przykładzie złącza kablowego niskiego napięcia.

Zdjęcie przedstawia wnętrze złącza kablowego niskiego napięcia.

Zadaniem pracownika terenowego jest zebranie informacji m.in. na temat wyposażenia znajdującego się we wnętrzu złącza:

- rodzaju łącznika: bezpiecznik, zwora, brak łącznika,
- wartości amperażu bezpiecznika,
- rozmieszczenia łączników w tzw. polach złącza.



Wnętrze złącza kablowego niskiego napięcia

Z punktu widzenia użytkownika, zasada działania jest analogiczna do klasyfikacji obrazu: wykonanie zdjęcia, rozpoznanie obrazu, prezentacja wyniku użytkownikowi, akceptacja lub wprowadzenie korekty, zapis zdjęcia i wyniku w postaci strukturalnej. Różnica tkwi w metodach rozpoznawania obrazu, wykorzystanych do realizacji tego zadania, zatem przyjrzyjmy się im bliżej.

Wydobywanie informacji ze zdjęcia przebiega w następujący sposób:

1. Detekcja obiektów ⟶ 2. Detekcja napisów ⟶ 3. Postprocessing

Na zdjęciu złącza, wykonanym przez użytkownika, w pierwszym kroku dokonywana jest detekcja, czyli wykrywanie obiektów. Wytrenowany model analizuje zdjęcie w poszukiwaniu obecności określonych obiektów, tj. bezpiecznika, zwory lub pustej podstawy bezpiecznikowej, a każdy znaleziony obiekt oznacza prostokątną ramką oraz etykietą. Do detekcji obiektów wykorzystaliśmy model YOLO v3.

W drugim kroku następuje wykrywanie napisów, znajdujących się na bezpiecznikach, w celu określenia wartości amperażu. Aby proces detekcji przebiegał szybciej, z obszaru całego zdjęcia do dalszej obróbki wybierane są tylko fragmenty obrazu, znajdujące się wewnątrz ramek oznaczających bezpieczniki, wykrytych w kroku pierwszym. Dalej następuje wykrywanie tekstu za pomocą modelu EAST, a następnie rozpoznawanie liter i cyfr z wykorzystaniem gotowego modelu z biblioteki Tesseract.

W ostatnim kroku ma miejsce oczyszczenie rozpoznanego tekstu z niepotrzebnych napisów oraz przypisanie bezpiecznikowi finalnej wartości amperażu, zgodnej z listą dopuszczalnych wartości.

Skuteczność detekcji obiektów i napisów w trakcie testów dochodziła do 98%, a rozpoznawania liter i cyfr z wykorzystaniem biblioteki Tesseract do 53%. Wynik ten był dla nas niezadowalający i wynikał m.in. z faktu, że biblioteka Tesseract przystosowana jest do rozpoznawania słów i tekstów, a nie pojedynczych znaków. W toku prac rozwojowych

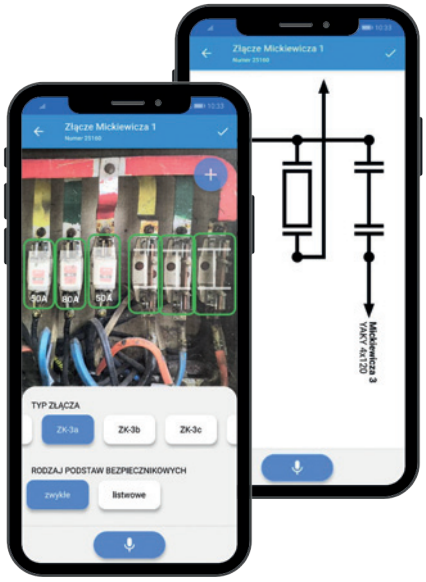
zdecydowaliśmy się skorzystać z modułu OCR dostępnego w pakiecie Google ML Kit, dzięki czemu udało się znacznie skrócić czas oraz zwiększyć skuteczność rozpoznawania do poziomu 80%.

Wynikiem całego procesu są automatycznie zebrane informacje na temat rozmieszczenia łączników w złączu, ich rodzaju oraz amperażu, gotowe do wysłania do firmowej bazy danych, a wszystko dzięki jednemu wykonanemu zdjęciu.

## Podsumowanie

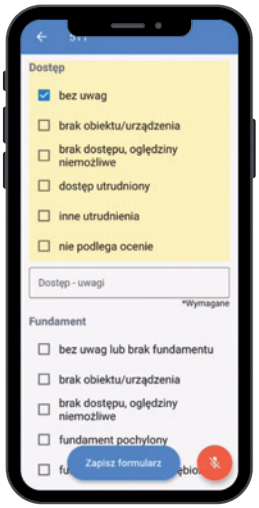
Nasze badania nad różnymi sposobami komunikacji, w celu przełamania bariery interakcji człowiek – maszyna w profesjonalnych zastosowaniach technicznych, pokazały, że sztuczna inteligencja jest w stanie w znacznym stopniu zniwelować tę barierę, ułatwiając i przyspieszając pozyskiwanie informacji. Należy jednak pamiętać, aby określone techniki wykorzystywać tam, gdzie realnie ułatwiają pracę. Jeśli bardziej wydajne jest dyktowanie informacji, warto użyć rozpoznawania mowy. Tam, gdzie łatwiej wykonać zdjęcie i wyczytać z niego pożądane informacje, lepiej zastosować rozpoznawanie obrazu. Warto również dać użytkownikowi możliwość wyboru preferowanej przez niego techniki interakcji z aplikacją i np. nie zmuszać do dyktowania danych, jeśli woli wprowadzić je ręcznie.

W efekcie prac wykonanych w projekcie powstała platforma do inteligentnego wsparcia procesu zbierania danych majątkowych, która dała początek dwóm specjalistycznym aplikacjom mobilnym: Topologia nN i Mobile Data Collector.




**Topologia nN** to łatwa w użyciu, sterowana głosem aplikacja mobilna, która wykorzystuje automatyczne pozyskiwanie danych technicznych ze zdjęć oraz buduje topologię sieci elektroenergetycznej w terenie.

Użycie mechanizmów sztucznej inteligencji, wraz z odpowiednio zaprojektowanym interfejsem użytkownika umożliwiło uproszczenie wprowadzania danych przez pracownika terenowego.




**Mobile Data Collector** to aplikacja mobilna, służąca do pozyskiwania i aktualizacji danych o majątku w trybie operacyjnym, w szczególności w zakresie przeglądów i konserwacji obiektów i urządzeń, przeznaczona dla różnych branż.

Aplikacja ułatwia i przyspiesza proces wprowadzania informacji technicznych w terenie na urządzeniu mobilnym dzięki wykorzystaniu rozpoznawania głosu, obrazu i czujników urządzenia. Posiada bogate możliwości konfiguracji, co upraszcza dostosowywanie jej do aktualnych potrzeb użytkownika.

*Platforma **GlobIQ** powstała w wyniku projektu B+R o nazwie **GlobIQ**, dofinansowanego w ramach Regionalnego Programu Operacyjnego Województwa Mazowieckiego na lata 2014-2020*

Fundusze Europejskie Program Regionalny
Rzeczpospolita Polska
Mazowsze. serce Polski
Unia Europejska Europejski Fundusz Rozwoju Regionalnego