



MINISTERO DELL'AMBIENTE
E DELLA SICUREZZA ENERGETICA

Sistema Integrato di Monitoraggio (SIM)

Progetto Esecutivo

ALLEGATO _M-DARE_II



**Finanziato
dall'Unione europea**
NextGenerationEU

Storia del documento

Versione	Data	Autore	Autorizzato da	Descrizione delle modifiche
1.0	24/11/2023	RTI DXC	MASE	Rilascio prima versione

Sommario

1	Progetto M-DARE.....	4
1.1	Obiettivo del servizio applicativo.....	4
1.1.1	Introduzione.....	4
1.1.2	Scopo Generale.....	4
1.1.3	Esigenze e Requisiti Chiave.....	5
1.1.4	Tematiche e Obiettivi Correlati.....	6
1.1.4.1	Creazione e inserimento di asset digitale.....	6
1.1.4.2	Suite Operativa.....	8
1.1.5	Benefici Attesi.....	10
1.1.6	Vincoli e Limitazioni.....	10
1.1.7	Stakeholders Coinvolti.....	11
1.1.8	Conclusione e Riepilogo.....	11
1.2	Requisiti funzionali.....	12
1.2.1	Requisiti non Funzionali Correlati.....	15
1.3	Dati di input.....	15
1.3.1	Catalogo delle Fonti di Dati.....	15
1.4	Funzioni, Algoritmi e Modelli.....	17
1.5	Dati di output.....	19

1 Progetto M-DARE

1.1 Obiettivo del servizio applicativo

1.1.1 Introduzione

Il Centro Nazionale di Meteorologia e Climatologia Aerospaziale (CNMCA) di Pratica di Mare è custode di tutta la documentazione meteorologica prodotta dal Comparto Meteorologico dell'Aeronautica Militare a partire dalla sua costituzione, avvenuta più di un secolo fa, fino ai nostri giorni.

La documentazione connessa con la raccolta delle osservazioni riveste una grande importanza sia sotto l'aspetto meteorologico in sé sia sotto quello culturale.

Le informazioni meteorologiche relative a un determinato sito geografico sono infatti parte integrante del suo patrimonio culturale in quanto rappresentative delle condizioni ambientali che influenzano usi, abitudini e finanche modi di pensare.

Occorre esplicitare per grandi linee le caratteristiche dell'attività di rilevazione e misurazione dei fenomeni meteorologici.

Queste vengono condotte secondo rigorose metodologie indicate da importanti organizzazioni internazionali, l'Organizzazione Meteorologica Mondiale (OMM) e l'International Civil Aviation Organization (ICAO).

In un determinato istante si definisce informazione meteorologica, o messaggio, l'insieme dei valori misurati secondo un certo sistema di parametri, o codice, che dipende dalla finalità di utilizzo dell'informazione.

Ad esempio, nella Navigazione Aerea, ci si avvale del codice METAR, che contiene al suo interno rilevazioni della direzione e dell'intensità del vento in pista, della visibilità orizzontale, dell'entità della copertura nuvolosa e della sua organizzazione nel cielo e dell'altezza dal suolo della base delle nubi, della temperatura e dell'umidità.

Oltre al METAR vengono utilizzati anche diverse altre tipologie di parametrizzazione.

In base a quanto evidenziato ciascun sito di osservazione dell'A.M., fornisce quotidianamente l'insieme delle informazioni che definisce l'andamento del tempo in quel sito in un quel giorno.

Tutto ciò ha prodotto nel tempo una considerevole mole di dati e documentazione per lo più cartacea, poiché la trascrizione dei dati e delle informazioni sui modelli cartacei ha sempre avuto e, ancora oggi continua ad avere, anche se in maniera più limitata, una grande rilevanza.

1.1.2 Scopo Generale

Dagli anni '70 è iniziato il processo di digitalizzazione, ma il riempimento della base dati non ha riguardato tutte le tipologie di rilevazione ne ha coperto, per quelle inserite, tutto il periodo d'interesse.

Scopo del programma è di completare l'**Archivio meteo-climatico nazionale** conferendogli le caratteristiche di organicità, preservabilità e accessibilità, oltre alla necessaria conservazione del patrimonio documentale a oggi disponibile.

La digitalizzazione di serie storiche estratte da dati meteorologici rappresenta un'attività di interesse nazionale ed internazionale, ancor più motivato dall'accresciuta importanza a livello mondiale degli studi sui cambiamenti climatici.

Tale attività di digitalizzazione è finalizzata a colmare i periodi di *gap* di disponibilità dell'archivio elettronico e ad allungare la serie storica dal 1925/30 al 1950 (data di inizio dell'archivio elettronico) delle sole stazioni che ancora oggi sono in esercizio.

L'attività si concretizza nella creazione di un duplicato digitale dei documenti esistenti, nell'estrazione da essi di dati mediante processi di riconoscimento dei caratteri riconoscimento dei caratteri attraverso tecnologie **Intelligent Character Recognition, Object Character Recognition, Tecniche di Intelligenza Artificiale** e nell'associare a ciascun documento i metadati relativi.

I vantaggi che possono scaturire da quest'attività possono essere sintetizzati come segue:

- Possibilità di accedere ai documenti in formato digitale, potendo individuare rapidamente la posizione del documento in archivio
- possibilità di associare a ogni documento i rispettivi metadati secondo gli standard internazionali vigenti;
- possibilità estrarre i dati contenuti nei documenti. In questo modo arrivare a colmare le lacune presenti nella base dati elettronica relazionale allungare le serie temporali
- Consideriamo in questo contesto oltre alla valenza meteorologica di alcuni documenti anche il loro valore storico (es le raccolte presso i territori dell'Eritrea, della Somalia, dell'Etiopia, della Libia durante il periodo coloniale).
- poter realizzare un "data lake" ad hoc con tutti i dati e le informazioni relative alle osservazioni meteorologiche raccolte dall'Aeronautica Militare ritenute necessarie per gli studi sui cambiamenti climatici condotti sia con strumenti convenzionali che con tecniche di Intelligenza Artificiale (IA) e di trattamento dei Big Data.

1.1.3 Esigenze e Requisiti Chiave

L'esigenza si inquadra negli ambiti della Data Rescue che dovrebbero essere intrapresi dai Servizi Meteorologici aventi nei propri archivi osservazioni e misurazioni storiche non ancora digitalizzate.

Tali attività hanno oggi assunto una notevolissima importanza in relazione alla possibilità che esse offrono di arricchire il patrimonio informativo grazie al quale con le nuove tecnologie è possibile affrontare con rinnovato vigore gli studi sui cambiamenti climatici.

In particolare, con l'avvento dei Big Data e/o dell'Intelligenza Artificiale, tale filone di studio ha assunto ancora maggior slancio.

1.1.4 Tematiche e Obiettivi Correlati

Tematica	Obiettivi
Creazione di asset digitale	Per ciascun documento deve essere realizzato un asset digitale costituito dalla sua immagine TIFF O PTIFF, da un PDF associato e dai relativi Metadati
Inserimento degli asset all'interno dell'infrastruttura digitale	L'inserimento degli asset all'interno del data Lake è finalizzato alle informazioni in essi contenute disponibili e facilmente reperibili.
Suite Operativa	Attraverso tecnologie Intelligent Character Recognition, Object Character Recognition, Tecniche di Intelligenza Artificiale , si mira ad estrarre dai documenti i dati necessari a colmare le lacune presenti, correggere errori e/o inconsistenze nei dati e allungare le serie storiche relativamente ai singoli siti di osservazione

1.1.4.1 Creazione e inserimento di asset digitale

Come rappresentato in sintesi, per ciascuno dei documenti in forma cartacea e su microfilm è richiesta la creazione di un asset digitale composto come di seguito descritto ovvero:

- dalle immagini digitalizzate in formato TIFF, per i documenti definiti di interesse storico il formato TIFF sarà sostituito dal formato TIFF piramidale (P-TIFF) con diverse risoluzioni di scansioni per mettere in evidenza i dettagli dei documenti;
- dall'omologo digitale in formato PDF/A;
- dai metadati ad esso relativi che verranno attribuiti in conformità agli standard riconosciuti (METS, PREMIS, Dublin Core etc.).
- I metadati saranno relativi da una parte alle caratteristiche esterne del documento con ciò intendendo tutte le proprietà che afferiscono alla sua provenienza, alla sua natura intrinseca, al contesto in cui esso è stato generato e alla sua posizione all'interno dell'archivio etc. e dall'altra alle caratteristiche interne del documento che possono essere desunte dal suo contenuto (tag).

Dato il particolare contenuto dei documenti da digitalizzare e al fine di rendere il più efficiente possibile il processo, il progetto che si intende realizzare prevede i seguenti passaggi:

- Preparazione e Normalizzazione dei Documenti:
 - Preparazione dei documenti cartacei prima della scansione.
 - Normalizzazione dei documenti per una gestione elettronica efficiente.
- Scansione Documentale:
 - Utilizzo di scanner professionali per creare un database di immagini digitali (es. scanner per microfilm, scanner planetari).
 - Scansione in diversi formati (b/n, scala di grigio, colori) e risoluzioni.
 - Controllo qualità: verifica, pulizia, eliminazione fogli bianchi, rotazione, e rescan se necessario.
- Classificazione Avanzata:

- Classificazione documenti tramite machine learning e intelligenza artificiale.
- Uso di Tecnologie di data recognition (OCR, ICR, optical mark recognition,)
- Analisi e definizione dei task di classificazione.
- Creazione File Indice:
 - Aggiunta di metadati per una ricerca rapida e accurata dei documenti.
 - Indicizzazione manuale e automatica tramite OCR.
 - Integrazione con database esterni per arricchire i metadati.
- Creazione del Data Lake con le Immagini e i Metadati:
 - Salvataggio dei dati su db
- Accesso e Condivisione
 - Accessibilità immediata delle informazioni attraverso interfacce utente intuitive e fruibili su tutti i dispositivi.

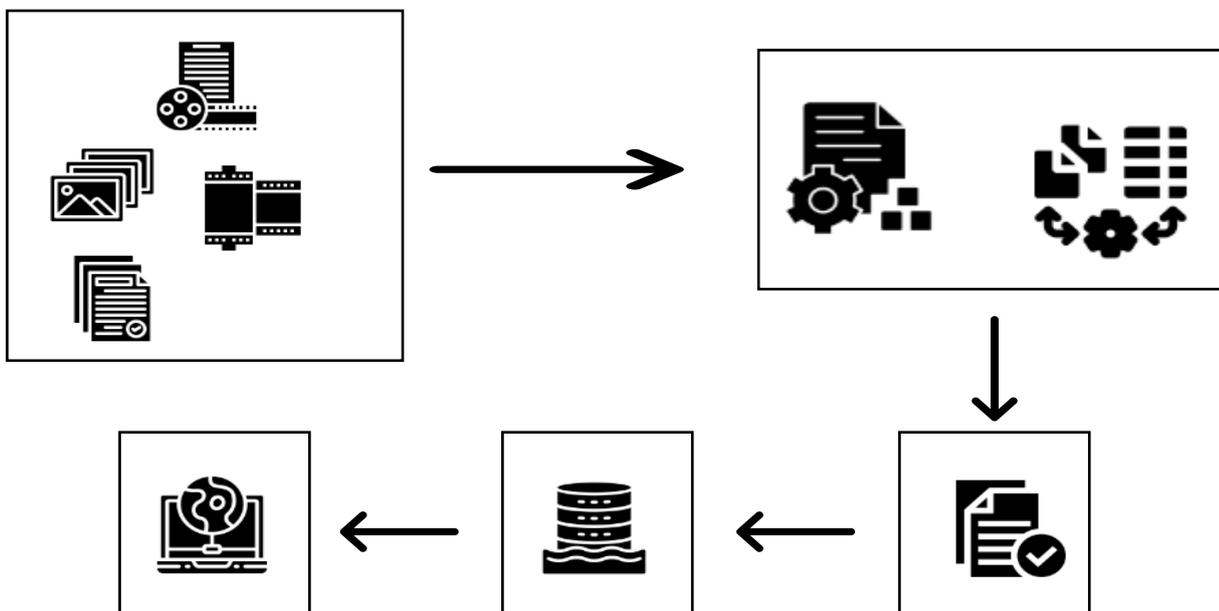


Figura 1 - Processo di Dematerializzazione

L'implementazione del progetto porta alla disponibilità di informazioni che ad oggi non sono facilmente fruibili da molti e che portano all'arricchimento del patrimonio informativo e ad un significativo miglioramento nell'efficienza operativa e nella gestione delle informazioni.

Una parte fondamentale del processo di digitalizzazione è rappresentato dalla pianificazione accurata del **set di metadati** che sono associati alle immagini digitali. Questi devono consentire le relazioni gerarchiche tra i documenti. Pertanto, verrà effettuata una progettazione di descrizione archivistica avanzata nel pieno rispetto degli standard archivistici (come ISAD (G); ISAAR (CPF); EAC-CPF) e seguendo il principio di una descrizione multilivello, adeguata alle specificità dei complessi documentari. I metadati devono includere, al minimo, informazioni che rappresentano l'unità da documentare e includono:

- Identificativo univoco;
- Tipo o oggetto del documento;
- Data del documento originale;
- Collocazione fisica dell'originale;
- Informazioni di natura gestionale;
- Data di creazione/validazione;

Per ogni file: nome, formato, impronta, dimensioni, data di creazione e codice operatore di scansione. Le informazioni che saranno estratte dal processo di catalogazione saranno integrate con i dati preesistenti dall'Aeronautica Militare.

L'integrazione prevede la creazione di:

- algoritmi per la corretta catalogazione dei dati sia a livello temporale (datazione) sia a livello spaziale (area di riferimento) corredandoli altresì dell'informazioni riguardanti la stazione di rilevamento.
- algoritmi che vadano a sanare eventuali inesattezze dei dati preesistenti.

Al termine delle procedure di digitalizzazione si ottiene una base informativa contenente oggetti in diversi formati (es: PDF, JPEG, TIFF, PTIFF, ...) con caratteristiche di omogeneità, integrità e completezza.

1.1.4.2 Suite Operativa

Le tecniche di estrazione delle caratteristiche interne del documento, che hanno una natura dinamica e possono crescere nel tempo, si configurano con l'Object Character Recognition (O.C.R.), l'Intelligent Character Recognition (I.C.R.) e l'Intelligenza Artificiale.

La suite operativa con la quale verrà svolto il servizio di estrazione di dati e informazioni, indicizzazione e metadattazione interna ed esterna si compone delle seguenti componenti fondamentali:

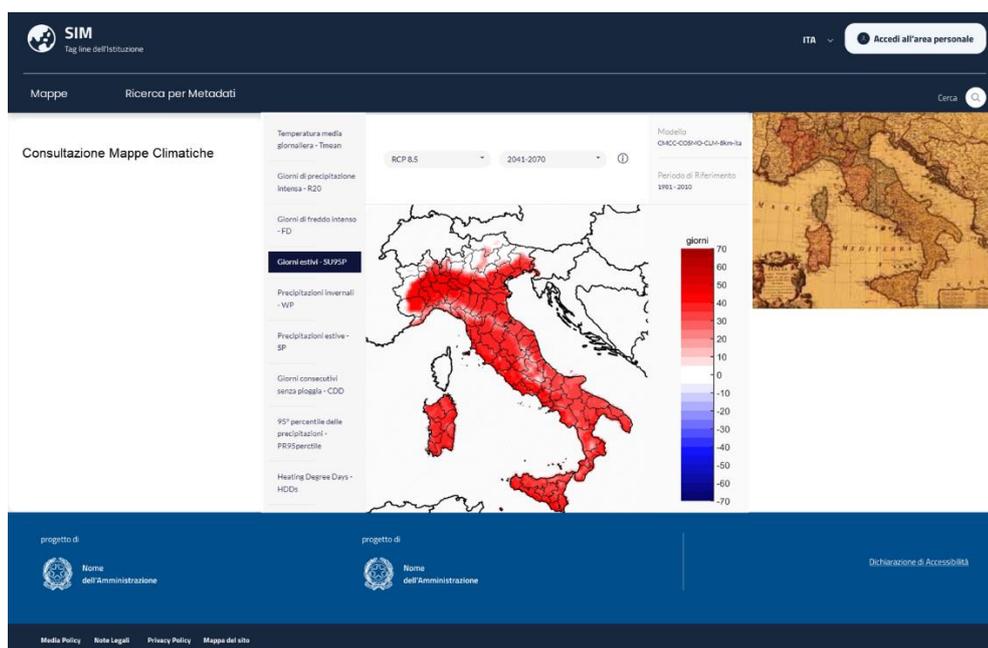
- workflow manager, opera come un **gestore di workflow completo**, offrendo un approccio integrato per la gestione delle funzionalità legate alla produzione, trasformazione e distribuzione di contenuti ed è progettato per eseguire in maniera sequenziale un'ampia gamma di attività di elaborazione su diversi tipi di contenuti. All'interno di questa soluzione, le attività preimpostate si estendono dalla ricerca e archiviazione alla metadattazione e all'esportazione dei contenuti ed, inoltre, possibile eseguire delle personalizzazioni, come la conversione del formato di un contenuto, la gestione di richieste di archiviazione in diversi siti. Il sistema è in grado di gestire una varietà di componenti degli asset documentali, inclusi contenuti video, audio, foto, pdf, oltre ai metadati tecnici e editoriali.
- motore OCR e ICR: Il modulo dedicato alla metadattazione interna ed esterna integra un motore OCR e ICR che verrà selezionato (es. Cloud Vision API di Google Cloud) che consente di convertire il testo presente nelle immagini in formato digitale, sia esso manoscritto o stampato, offrendo così una soluzione avanzata per la metadattazione dei documenti.
- A valle dell'applicazione delle tecniche di OCR e ICR, qualora l'affidabilità dei risultati dovesse attestarsi al di sotto di una soglia prestabilita, tali tecniche verranno riapplicate unitamente a

tecniche di Intelligenza Artificiale per migliorare l'estrazione di dati e informazioni dal contenuto del documento e per la metadazione interna

- un motore di ricerca per l'indicizzazione dei contenuti degli asset che permette la ricerca di grandi quantità di dati e delle informazioni in tempo reale e di effettuare analisi su di essi. È fondamentale sia per l'individuazione dei contenuti quando un utente richiede l'accesso ad un asset documentale o a parte di esso, sia per l'esecuzione di studi o analisi sugli asset documentali in elaborazioni avanzate. Per questo è richiesto che esso sia in grado di eseguire oltre alla fase di ricerca anche la fase di analisi in maniera integrata. Per quanto concerne le ricerche di testo completo il motore di ricerca richiesto deve possedere un sistema automatizzato integrato di analisi del linguaggio e il suo database deve essere in grado di analizzare porzioni di testo in qualsiasi lingua e formato, indicizzando in modo efficiente termini chiave e sinonimi. È scalabile e garantisce affidabilità e velocità, soprattutto nella ricerca di testo completo e nell'analisi delle metriche. La sua configurazione è altamente personalizzabile per adattarsi a diversi casi d'uso.
- un'applicazione di visualizzazione dei dati che lavorando in abbinamento con il motore di ricerca consente agli utenti di esplorare, analizzare e visualizzare i dati in modo intuitivo e interattivo. Esso deve supportare una vasta gamma di tipi di dati, tra cui dati strutturati e non strutturati, e si richiede che esso disponga di una vasta gamma di strumenti di visualizzazione dei dati. Agli utenti dovrà essere consentito di creare dashboard personalizzate per visualizzare i dati in modo chiaro, diretto e conciso inserendo grafici, tabelle, mappe e altre visualizzazioni per realizzare rappresentazioni efficaci offrendo allo scopo una vasta gamma di opzioni di personalizzazione consentendo di modificare colori, layout e le impostazioni di visualizzazione dei dati in base alle esigenze. Grazie alla sinergia con il motore di ricerca il visualizzatore deve poter mettere in evidenza all'interno dei documenti brani di testo completo e suggerimenti sui dati e/o estrarre informazioni significative dai dati. Il processo di visualizzazione dei dati nel "data lake" integrato nella piattaforma SIM, contenente tutte le informazioni e i dati relativi alle osservazioni meteorologiche raccolte dall'Aeronautica Militare per gli studi sui cambiamenti climatici, può essere descritto come segue:

1. Accesso al Data Lake: Gli utenti autorizzati possono accedere al "data lake" attraverso un appropriato sistema o interfaccia.
2. Identificazione dei Dati: Gli utenti possono identificare e selezionare specifici set di dati correlati alle osservazioni meteorologiche di interesse.
3. Applicazione di Filtri e Query: Utilizzando funzioni di ricerca avanzate, gli utenti possono applicare filtri e query per raffinare i dati in base a parametri specifici come la data, la località o altri criteri pertinenti.
4. Visualizzazione dei Dati: I dati selezionati vengono visualizzati in formati comprensibili e utilizzabili, che possono includere grafici, mappe o altre rappresentazioni visive. Questo passo consente agli utenti di interpretare rapidamente le informazioni e identificare pattern o tendenza.

5. **Analisi Interattiva:** Gli utenti possono condurre analisi interattive esplorando ulteriormente i dati, soffermandosi su dettagli specifici e regolando la visualizzazione in base alle loro esigenze.
6. **Rapporti e Condivisione:** Gli utenti possono generare rapporti basati sui dati visualizzati e condividere le informazioni con altri stakeholder o ricercatori coinvolti negli studi sui cambiamenti climatici.



1.1.5 Benefici Attesi

Il servizio in esame consentirà di avere a disposizione un sistema di dati integrati completi e consultabili, da cui estrarre informazioni meteorologiche storiche utili per l'analisi e lo studio in materia climatica.

Sarà di supporto per effettuare le analisi e gli studi relativi ai cambiamenti climatici

1.1.6 Vincoli e Limitazioni

Tra la limitazioni/punti critici possiamo evidenziare quelli riportati di seguito che vengono richieste come attività a corredo della digitalizzazione, dovendosi ritenere il progetto come un 'chiavi in mano':

- Il servizio di messa in sicurezza richiede un'attività di ritiro della documentazione, trasporto in sicurezza, sanificazione, spolveratura, archiviazione e custodia della stessa.
- Inoltre, il materiale cartaceo che non potrà essere digitalizzato verrà messo in sicurezza e salvaguardato rispetto alla possibilità di ulteriore deterioramento. Esso verrà conservato per

essere sottoposto a eventuali future tecniche di trattamento che consentano di estrarre da esso dati e/o informazioni.

La salvaguardia del materiale non digitalizzabile non è nello scope del progetto M-DARE.

1.1.7 Stakeholders Coinvolti

CNMCA – Centro Nazionale di Meteorologia e Climatologia (Pratica di Mare) e le Stazioni Meteo Periferiche, in qualità di fornitori del materiale cartaceo/microfilmato da digitalizzare e inserire all'interno del Data Lake.

Tramite un'interfaccia grafica lo stesso personale specializzato dell'Aeronautica Militare potrà effettuare le analisi mirate agli studi meteorologici d'interesse

1.1.8 Conclusione e Riepilogo

Le finalità principali del progetto M-DARE possono inquadrarsi in prima istanza nella realizzazione di una base dati completa e integrata in materia climatica, a partire dal materiale cartaceo e microfilmato attualmente disponibile, che conseguentemente, poiché sostituito nelle consultazioni dalla relativa versione digitale, verrà salvaguardato da usura o danni.

In seconda istanza, attraverso l'asset digitale realizzato, essendo esso integrato, completo e attendibile, si consentirà agli utenti specializzati di effettuare analisi rapide e interattive finalizzate agli studi climatologici particolarmente strategici in fase di cambiamenti climatici.

1.2 Requisiti funzionali

Elenco dei Requisiti Funzionali

id applicativo_id_rf	descrizione_rf	progettazione_rf	implementazione_rf
VMDARE_RF001	Accesso all'applicativo	Realizzazione di una interfaccia utente di utilizzo che sarà accessibile solo tramite autenticazione. La profilazione utente dovrà consentire di attribuire le giuste utilities e le corrette grant agli utilizzatori	L'autenticazione deve seguire gli standard supportati del SIM, gestiti dai servizi forniti dal sottosistema "PROFILAZIONE"
VMDARE_RF002	Accesso multidevice	Realizzazione di un sistema adeguatamente fruibile da diversi dispositivi	L'applicativo deve essere esercibile attraverso differenti tipologie di dispositivo e devono essere supportati i sistemi Android o IOS
VMDARE_RF003	Sincronizzazione delle Operazioni	Realizzazione di una funzione che consenta la sincronizzazione delle operazioni di INPUT/OUTPUT	L'applicativo deve usufruire dei metodi che consentendo la corretta sincronizzazione su Cloud, in modo da poter interrompere liberamente il lavoro e riuscire a riprenderlo anche da un'altra postazione
VMDARE_RF004	Interfaccia personalizzata	Realizzazione di un cruscotto personalizzato per utente	L'applicativo deve essere tale da consentire all'utente una volta acceduto di poter disporre delle funzionalità e del cono dati specifici per la sua utenza
VMDARE_RF005	Consultazione	La componente realizzata consentirà agli utenti di effettuare ricerche all'interno della banca dati anche utilizzando le tecniche ICR e OCR. In questo modo è possibile selezionare e identificare specifici set di dati	Verrà realizzato un motore di ricerca che consentirà all'utente di effettuare la ricerca dei dati e delle informazioni e di effettuare analisi su di essi

id applicativo_id_rf	descrizione_rf	progettazione_rf	implementazione_rf
VMDARE_RF006	Navigazione dell'interfaccia attraverso l'utilizzo di filtri e query	La componente realizzata consente di effettuare ricerche avanzate, applicare filtri e query per raffinare i dati in base a parametri specifici come la data, la località o altri criteri pertinenti.	
VMDARE_RF007	Visualizzazione e informazioni	I dati selezionati vengono visualizzati in formati comprensibili e utilizzabili, che possono includere grafici, mappe o altre rappresentazioni visive. Questo passo consente agli utenti di interpretare rapidamente le informazioni e identificare pattern o tendenze.	
VMDARE_RF008	Analisi interattiva	Gli utenti possono condurre analisi interattive esplorando ulteriormente i dati, soffermandosi su dettagli specifici e regolando la	



MINISTERO DELL'AMBIENTE
E DELLA SICUREZZA ENERGETICA

Sistema Integrato di Monitoraggio (SIM) –
Progetto esecutivo



Finanziato
dall'Unione europea
NextGenerationEU

id_applicativo_id_rf	descrizione_rf	progettazione_rf	implementazione_rf
		visualizzazione in base alle loro esigenze.	
VMDARE_RF009	Rapporti e Condivisioni	Gli utenti possono generare rapporti basati sui dati visualizzati e condividere le informazioni con altri stakeholder o ricercatori coinvolti negli studi sui cambiamenti climatici.	

1.2.1 Requisiti non Funzionali Correlati

id_applicativo_id_rf	descrizione_rf	progettazione_rf
VMDARE_RNF001	Sicurezza	Il sistema deve essere protetto da accessi non autorizzati
VMDARE_RNF002	Performance	Il sistema deve essere in grado di gestire il numero richiesto di utenti senza subire degrado prestazionale
VMDARE_RNF003	Scalabilità	Il sistema deve essere in grado di aumentare o diminuire in base alle esigenze
VMDARE_RNF004	Disponibilità	Il sistema deve essere disponibile e accessibile
VMDARE_RNF005	Manutenzione	Il sistema deve essere di facile manutenzione e aggiornamento
VMDARE_RNF006	Portabilità	Il sistema deve garantire il funzionamento su piattaforme diverse con modifiche minime
VMDARE_RNF008	Affidabilità	Il sistema deve essere affidabile e soddisfare i requisiti dell'utente
VMDARE_RNF009	Usabilità	Il sistema deve essere facile da usare e da capire
VMDARE_RF010	Digitalizzazione Documentale	Sarà realizzato un prodotto software ed una procedura esecutiva che consentirà la digitalizzazione della documentazione della Committente.
VMDARE_RF011	Creazione e inserimento Asset digitale	Una volta proceduto alla digitalizzazione del materiale cartaceo/microfilmato, lo stesso verrà inserito in base dati
VMDARE_RF012	Gestione Workflow operativa	La componente realizzata dovrà consentire la gestione l'elaborazione e la distribuzione informativa, nei diversi formati, verso i vari richiedenti
VMDARE_RF013	Omogeneizzazione Informativa	
VMDARE_RF014	Backup	
VMDARE_RF015	Integrazione	

1.3 Dati di input

1.3.1 Catalogo delle Fonti di Dati

Id	Nome Sorgente Dati	Proprietà dei Dati (owner)	Modalità di Accesso	Uso del Dato
VMDARE_DI001	Documentazione cartacea e microfilmata Meteorologica	CNMCA	Consultazione manuale	Manuten- tivo Docume- ntale



MINISTERO DELL'AMBIENTE
E DELLA SICUREZZA ENERGETICA

*Sistema Integrato di Monitoraggio (SIM) –
Progetto esecutivo*



**Finanziato
dall'Unione europea**
NextGenerationEU

Id	Nome Sorgente Dati	Proprietà dei Dati (owner)	Modalità di Accesso	Uso del Dato
VMDARE_ DI002	Asset digitale residente all'interno del Data Lake	CNMCA	Consultazione Analitica	Analitico

1.4 Funzioni, Algoritmi e Modelli

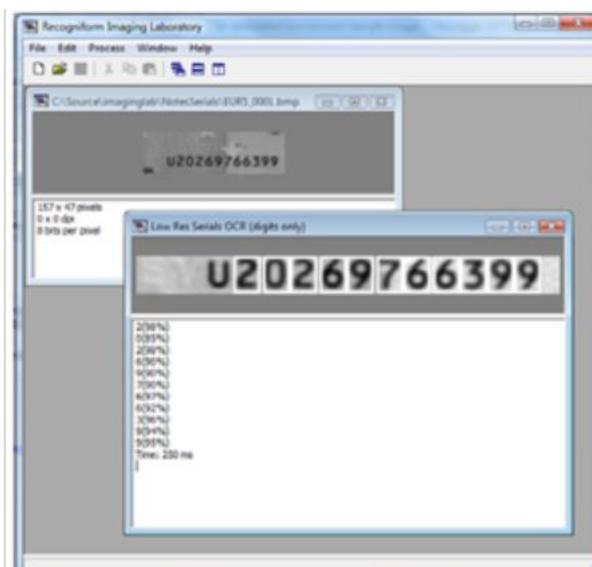
Le funzioni principali utilizzate all'interno dell'applicativo M-Dare sono riconducibili al processo di metadattazione interna ed esterna, del materiale cartaceo/microfilmato. In particolare, parliamo delle già citate tecnologie OCR (Optical Character Recognition) e ICR (Intelligent Character Recognition).

OCR (Optical Character Recognition)

A prescindere da quale tipologia di sistema OCR utilizzare (monofont, multfont o omnifont) il lavoro che un tale sistema svolge è suddivisibile in tre fasi:

- segmentazione,
- estrazione delle caratteristiche
- classificazione.

La prima fase, la segmentazione, consiste nell'identificare i caratteri da leggere: qualora l'area da leggere sia ampia o sia costituita da un'intera pagina, può essere necessaria anche una fase di analisi del layout, che consiste nell'individuare colonne, paragrafi, linee di testo e parole. La seconda fase è rappresentata dall'estrazione delle caratteristiche di ogni singolo carattere segmentato. Esistono diversi metodi per estrarre queste caratteristiche, alcuni basati su procedimenti statistici (es: istogrammi delle frequenze di pixel), altri su procedimento di tipo geometrico (es: curvatura e direzionalità delle linee). In ogni caso l'idea di base è che queste caratteristiche debbano essere presenti in ogni carattere dello stesso tipo, debbano essere sufficienti per identificarlo e che debbano tollerare rumore e distorsioni introdotte al processo di scansione. La terza fase è quella della classificazione: le caratteristiche estratte sono analizzate in modo da risalire al carattere partendo dalla sua forma, usando una conoscenza a priori basata sui prototipi del set dei caratteri da riconoscere.



Il risultato di questo processo è quindi quello di attribuire a ciascun carattere il codice ASCII (o UNICODE) che lo rappresenta e, eventualmente, una confidenza di lettura che indica quanto il sistema è certo di aver classificato correttamente il dato.

Per fare in modo da minimizzare il più possibile gli errori dovuti all'OCR in alcuni sistemi di lettura ottica è possibile fare in modo di limitare il set di caratteri che ci si aspetta di trovare: in tal modo, ad esempio, se si devono leggere dati costituiti solo da numeri, l'OCR non confonderà degli zero con delle lettere "O" oppure dei "2" con delle lettere "Z", e così via. Sebbene la tecnologia OCR sia molto progredita, talvolta i risultati ottenuti possono essere sotto le aspettative a causa della scarsa qualità di stampa o della scarsa qualità di scansione dei documenti processati: i maggiori problemi di riconoscimento, infatti, si incontrano laddove più caratteri risultino attaccati tra loro, oppure laddove un singolo carattere sia spezzettato in più parti.

ICR (Intelligent Character Recognition)

L'ICR (Intelligent Character Recognition), utilizzato per il riconoscimento del testo scritto a mano in stampatello, può considerarsi un'evoluzione ed una differenziazione dell'OCR. Nel momento in cui i sistemi OCR sono stati estesi al riconoscimento di dati scritti a mano in stampatello, si è fatto ricorso soprattutto alla tecnologia delle reti neurali: per sottolineare che la tale nuova tecnologia era basata sull'intelligenza artificiale è stato allora coniato il nuovo acronimo, sostituendo alla "O" di Optical, la "I" di Intelligent. Tuttavia, la maggior parte dei sistemi ICR è in grado di leggere solo parzialmente il testo stampato o dattiloscritto, quindi qualora si abbia necessità di estrarre sia dati stampati che manoscritti è necessario utilizzare sistemi di lettura ottica che dispongano anche di OCR.

I sistemi ICR si possono distinguere in due categorie: constrained ed unconstrained. I primi sono quelli che riescono a riconoscere testo scritto a mano in stampatello in cui i caratteri sono totalmente separati gli uni dagli altri, mentre i secondi sono quelli che tollerano che alcuni caratteri possano toccarsi da loro.

1.5 Dati di output

Id	Nome Sorgente Dati	Proprietà dei Dati (owner)	Modalità di Accesso	Uso del Dato
VMDARE_DO001	Documentazione Meteorologica digitale	CNMCA	Consultazione manuale	Manut entivo Docu menta le
VMDARE_DO002	Dashboard	CNMCA	Navigazione orientata all'analisi dei dati/serie storiche	Analiti co
VMDARE_DO002	Reporting	CNMCA	Produzione della reportistica esplicativa da condividere, riportante i risultati della fase di analisi	Consul tativo

Di seguito è riportato un esempio di dashboard navigabile, funzionale all'attività di analisi.

