MISSION INNOVATION ITALY

MINISTERO DELLA TRANSIZIONE ECOLOGICA

MISSION INNOVATION

accelerating the clean energy revolution

POA MATERIALI AVANZATI PER L'ENERGIA

PROGETTO IEMAP - Piattaforma Italiana Accelerata per i Materiali per l'Energia

Stato dell'arte nel settore per la Scienza dei Materiali

Claudio Ronchetti, Simone Giusepponi, Francesco Buonocore, Filippo Palombi, Roberto Grena, Beatrice Calosso, Massimo Celino

ENEA

Report IM21-24/3

STATO DELL'ARTE NELLA SCIENZA DEI MATERIALI

Claudio Ronchetti (ENEA), Simone Giusepponi (ENEA), Francesco Buonocore (ENEA), Filippo Palombi (ENEA), Roberto Grena (ENEA), Beatrice Calosso (ENEA), Massimo Celino (ENEA)

Maggio 2022

Report MISSION INNOVATION

Ministero della Transizione Ecologica - ENEA
Mission Innovation 2021-2024 - I annualità
Progetto: Piattaforma accelerata per i Materiali per l'Energia
Work package: IEMAP: Italian Energy Materials Acceleration Platform
Linea di attività: LA1.7
Responsabile del Progetto: Massimo Celino, ENEA
Responsabile della LA: Massimo Celino, ENEA

Indice

1 Sommario

Il presente rapporto descrive lo stato dell'arte nel settore della gestione dei dati nella scienza dei materiali per l'energia. Questa analisi è propedeutica alla pianificazione delle attività per lo sviluppo di una gestione dei dati per la piattaforma mission innovation IEMAP. Quindi dopo aver analizzato i database più utilizzati nel settore principalmente computazionali e analizzato le proposte per una gestione FAIR dei dati, si riporta lo studio realizzato per comprendere come i dati sono gestiti nei laboratori afferenti a IEMAP. Tale studio permette di trarre delle conclusioni e di proporre un format dei dati e una architettura per il database che li deve contenere.

2 Introduzione

La digitalizzazione della ricerca nel campo della scienza dei materiali implica la necessaria introduzione dei concetti principali che definiscono lo stato dell'arte nella progettazione del modello dati. Nei paragrafi seguenti vengono riportati e descritti i concetti e gli standard più diffusi.

2.1 Ontologie

Per ontologia si intende una rappresentazione del mondo reale che "traduce" il linguaggio naturale (ambiguo per sua natura) in un linguaggio formale (non ambiguo) utilizzabile per la realizzazione di sistemi informativi integrati.

Un'ontologia computazionale quindi – per un dominio di interesse – è una rappresentazione:

- **formale**, utilizza un linguaggio simbolico non ambiguo e processabile da elaboratori;
- **condivisa**, determinata dal consenso di una pluralità, il più ampia possibile, di soggetti competenti sulla materia rappresentata;
- **esplicita**, tutte le assunzioni sono rese in maniera esplicita.

I principali vantaggi della rappresentazione ontologica dei domini di interesse sono:

- *modellazione concettuale*, che non richiede di conoscere l'organizzazione fisica dei dati;
- *elevata espressività*, un concetto non è un elenco di attributi ma è la composizione ricorsiva di costrutti logici (intersezione, unione, complemento, disgiunzione o enumerazione di più concetti);
- *possibilità di uso di strumenti automatici di ragionamento*, si possono ottenere informazioni più complete ed esaustive ed effettuare ricerche più complesse ed efficienti;
- *condivisione di conoscenza e vocabolari comuni*, ottenendo un'interoperabilità a livello semantico tra uomo/uomo, uomo/macchina e macchina/macchina;
- separazione della conoscenza di dominio da quella operazionale.

Il linguaggio formale utilizzato nella modellazione delle ontologie è OWL – Web Ontology Language e rappresenta uno standard del W3C (World Wide Web Consortium). Le ontologie sono diffuse in diversi formati di serializzazione che includono RDF-XML, Turtle e Json-ld.

2.2 Principi FAIR

Nel 2014 sono stati elaborati un gruppo di principi fondamentali, denominati principi dei dati FAIR, per ottimizzare la riutilizzabilità dei dati della ricerca. Essi rappresentano un insieme di linee guida e migliori pratiche sviluppate per garantire che i dati, o qualsiasi oggetto digitale, siano **F**indable / Rintracciabili, **A**ccessible / Accessibili, **I**nteroperable / Interoperabili e **R**e-usable / Riutilizzabili:

- **Rintracciabili**: per poter rendere i dati riutilizzabili occorre che siano per prima cosa rintracciabili dagli esseri umani e dalle macchine. Il recupero automatico e affidabile di set di dati dipende dagli identificatori persistenti (PID) utilizzati, quali ad esempio DOI, Handle o URN, e dai metadati

descrittivi attribuiti ai dati, che devono essere registrati in "cataloghi" o in repository indicizzabili anche dalle macchine.

- **Accessibili**: i dati o almeno i loro metadati devono poter essere accessibili dagli esseri umani e dalle macchine anche attraverso sistemi di autenticazione e autorizzazione (non è necessario che i dati depositati siano open access) mediante l'uso di protocolli standard. I dati e i loro metadati devono essere depositati in archivi o repository che li rendano possibilmente persistenti nel tempo e rintracciabili in rete. Almeno i metadati dovrebbero rimanere sempre disponibili anche quando i dati non sono in open access.
- **Interoperabili**: i dati devono poter essere combinati e utilizzati insieme con altri dati o strumenti. Il formato dei dati deve pertanto essere aperto e interpretabile da vari strumenti, compresi altre basi di dati. Il concetto di interoperabilità si applica anche ai metadati. Ad esempio, i metadati dovrebbero utilizzare un linguaggio standardizzato e condiviso a livello internazionale dai diversi servizi di indicizzazione.
- **Riutilizzabili**: sia i metadati, sia i dati devono essere descritti e documentati nel migliore dei modi, a garanzia della loro qualità e perché possano essere replicati e/o combinati in contesti diversi. Il trattamento dei dati dovrebbe conformarsi agli standard o ai protocolli riconosciuti dalle comunità scientifiche di riferimento. Il riutilizzo dei metadati e dei dati dovrebbe essere dichiarato con una/o più licenze aperte chiare ed accessibili.

Varie iniziative internazionali hanno cercato e stanno cercando di definire strumenti e metriche per valutare il grado nel rispetto dei principi FAIR del proprio dato di ricerca. L'iniziativa europea, chiamata European Collaborative Data Infrastructure (EUDAT), ha definito una semplice lista di autovalutazione di controllo:

Findable / Rintracciabili

- È stato assegnato un identificatore persistente (es. DOI, Handle, URN) al dataset?
- Il dataset è stato descritto con metadati esaustivi, informativi e accurati?
- I metadati sono registrati in un catalogo online o in un data repository che sia indicizzato dai motori di ricerca?
- Fra i metadati è incluso anche l'identificatore persistente assegnato al dataset?

Accessible / Accessibili

- L'identificatore persistente associato al dataset risolve correttamente alla pagina dei matadati del dataset?
- Il protocollo di recupero dei dati e dei metadati rispetta un linguaggio standardizzato e riconosciuto come ad esempio quello delle pagine web (HTTP)?
- I metadati sono sempre pubblici, visibili e indicizzabili anche se i dati non sono in open access o non lo sono più?

Interoperable / Interoperabili

- I dati sono resi disponibili in formati aperti o almeno in formati documentati e diffusi?

- I metadati seguono schemi standard riconosciuti e condivisi?
- Sono stati utilizzati quanto più possibile vocabolari controllati tesauri o ontologie?
- Sono resi disponibili link o relazioni con altre risorse rilevanti per la comprensione dei dati come pubblicazioni o rapporti tecnici o applicazioni software?

Re-usable / Riutilizzabili

- I dati sono accurati, completi e descritti in modo che siano facilmente comprensibili e riproducibili?
- Al dataset è stata attribuita una licenza che ne specifica le possibilità di riutilizzo?
- Sono chiare dai metadati e dalla documentazione allegata le responsabilità scientifiche e finalità dei dati prodotti?
- I dati e i metadati rispettano gli standard e i protocolli di qualità del dominio di ricerca di riferimento?

2.3 Metadati

Il passaggio alla produzione documentaria digitale ha imposto l'adozione di specifici strumenti e la soddisfazione di requisiti necessari per governare la formazione, la gestione, la tenuta e la conservazione dei documenti. Fra di essi molto importante è l'utilizzo dei cosiddetti metadati, termine in uso nel linguaggio informatico per definire un insieme di informazioni sui dati. Essi sono spesso definiti anche come "dati sui dati". Il termine deriva dall'inglese metadata, che trae origine dal prefisso meta- (dalla preposizione greca metà "al di sopra") e dal plurale neutro latino data ossia "i dati".

I metadati sono quei dati che descrivono altri dati, in particolare in riferimento ai documenti digitali.

I metadati hanno vari utilizzi, possono sia costituire il documento informatico stesso (metadati di contenuto) sia descrivere un determinato documento per fare in modo che, una volta inserito in un sistema di archiviazione, esso possa essere facilmente recuperato.

I metadati più basilari sono il formato e il nome del file, le specifiche tecniche sulla versione del software e sul hardware, le date di creazione, di accesso e di ultima modifica, l'autore; quelli più complessi la descrizione, l'oggetto, i termini di rilascio, accesso e uso, ecc.

Tali elementi servono per attribuire al documento un'identità ben precisa. Si capisce quindi perché viene ritenuta cruciale la fase di formazione dell'archivio per la corretta tenuta e conservazione della documentazione digitale.

Il documento però una volta formato avrà un suo percorso di esistenza che lo porterà ad essere gestito da diversi sistemi e applicazioni con la contestuale produzione e associazione di ulteriori informazioni. Un altro aspetto da tenere in considerazione, dunque, è che i metadati non vengono attribuiti ai documenti e in generale agli oggetti digitali tutti nel medesimo tempo, ma tendono ad accumularsi nel corso della vita degli stessi per tracciarne l'utilizzo, ad esempio gli accessi, le modifiche, i trasferimenti, le copie, nonché le modalità della sua conservazione.

Si comprende quindi che i metadati sono dati che descrivono il contenuto, la struttura e il contesto dei documenti e la loro gestione nel tempo.

2.4 Il formato JSON

JSON (JavaScript Object Notation) è un semplice formato per lo scambio di dati. Per le persone è facile da leggere e scrivere, mentre per le macchine risulta facile da generare e analizzarne la sintassi. Si basa su un sottoinsieme del Linguaggio di Programmazione JavaScript, Standard ECMA-262 Terza Edizione - dicembre 1999.

JSON è un formato di testo completamente indipendente dal linguaggio di programmazione, ma utilizza convenzioni conosciute dai programmatori di linguaggi della famiglia del C, come C, C++, C#, Java, JavaScript, Perl, Python, e molti altri. Questa caratteristica fa di JSON un linguaggio ideale per lo scambio di dati.

2.5 Servizi API

Application Programming Interface (API) è un insieme di definizioni e protocolli per la creazione e l'integrazione di software applicativi.

Le API permettono ai prodotti o servizi software di comunicare con altri prodotti o servizi software senza che sia necessario sapere come vengano implementati, semplificando così lo sviluppo delle applicazioni software e consentendo un netto risparmi di tempo e denaro. Infatti, l'uso di questo tipo di protocollo definisce uno strato intermedio che permette il disaccoppiamento dalla tecnologia con cui sono stati implementati le diverse applicazioni o servizi.

I servizi offerti dalle API devono avere due caratteristiche principali, ovvero l'**interoperabilità** e la **componibilità**. L'interoperabilità è il grado in cui due o più sistemi riescono a cooperare e a scambiare informazioni in maniera più o meno completa e priva di errori, con affidabilità e con ottimizzazione delle risorse. La componibilità permette di ottenere un servizio, chiamato composto, attraverso la composizione di due o più servizi software, chiamati servizi componibili. A sua volta il servizio composto può dar luogo a nuovi servizi composti.

Le API, semplificando l'integrazione di nuovi componenti applicativi in un'architettura esistente, promuovono la collaborazione tra organizzazione e team IT. Per restare competitive e rispondere in modo agile ai costanti mutamenti dei mercati digitali, in cui nuovi concorrenti possono rivoluzionare un intero settore con una nuova app, le aziende devono adattarsi rapidamente e supportare lo sviluppo e il deployment di servizi innovativi. Lo sviluppo di applicazioni cloud native, basato sul collegamento di un'architettura applicativa di microservizi attraverso le API, consente di accelerare la velocità di sviluppo.

3 Implementazione

In questa sezione vengono riportati alcuni delle piattaforme che rappresentano lo stato dell'arte per la scienza dei materiali. I primi due, NOMAD e Materials Project, sono due piattaforme per la gestione e la fruizione di dati estratti da processi computazionali, mentre l'ultimo, Battery2030+, a partire da processi sperimentali.

3.1 NOMAD

3.1.1 Obiettivi

Novel Materials Discovery (NOMAD) crea, colleziona, archivia e ripulisce dati computazionali nel campo della scienza dei materiali, calcolati attraverso i più importanti codici in materia.

Inoltre, il NOMAD Laboratory sviluppa strumenti per l'estrazione di questi dati al fine di trovare strutture, correlazioni e nuove informazioni che non potrebbero essere scoperte studiando insiemi di dati più piccoli. Pertanto, NOMAD favorisce la ricerca e la scoperta di nuovi materiali.

E, soprattutto, NOMAD guida l'Open Science Movement nella scienza dei materiali, supportato dalla comunità globale rendendo tutti i dati liberamente accessibili e supportando al meglio i principi FAIR.

3.1.2 Metadati

NOMAD estrae ricchi metadati dai dati computazionali grezzi caricati dall'utente. Il processo di estrazione, mostrato in Figura 1, è composto essenzialmente da due macro task: caricamento e processamento/parsing dei dati.

Il caricamento dei dati da parte dell'utente è semplice; accedendo alla pagina web[1] è possibile effettuare direttamente il caricamento dei file computazionali, così come sono, in un formato zip che vengono immagazzinati all'interno del repository. I dati caricati vengono elaborati automaticamente e resi disponibili nei file grezzi caricati o nel relativo modulo di dati elaborati unificato. I parser[2] NOMAD convertono i file grezzi nel formato dati comune di NOMAD.

NOMAD supporta la maggior parte dei codici community e dei formati di file: abinit, aflow, amber, asap, asr, atk, band, bigdft, bopfox, castep, charmm, cp2k, cpmd, crystal, dftbplus, dlpoly, dmol3, eelsdb, elastic, elk, example , emozionante, fhiaims, fhivibes, fleur, fplo, gamess, gaussiano, gpaw, gromacs, gromos, gulp, lammps, libatoms, aragosta, molcas, mopac, mp, namd, nexus, nwchem, octopus, onetep, openkim, openmx, orca , fonopia, psi4, qball, qbox, quantumespresso, siesta, tinker, turbomole, vasp, wien2k, xtb, yambo.

NOMAD fornisce i dati in forma elaborata e normalizzata in un formato gerarchico e processabile dalla macchina. Questi dati elaborati, ovvero l'Archivio NOMAD, sono organizzati in sezioni nidificate di quantità con unità, tipi di dati, forme e descrizioni ben definiti. Queste definizioni sono chiamate NOMAD **Metainfo**, il cui schema è mostrato in Figura 2.

[1] Pagina web per il caricamento di dati in NOMAD: https://nomad-lab.eu/prod/v1/gui/user/uploads
[2] Guida all'uso dei parser di NOMAD: https://nomad-lab.eu/prod/rae/docs/client/parsers.html

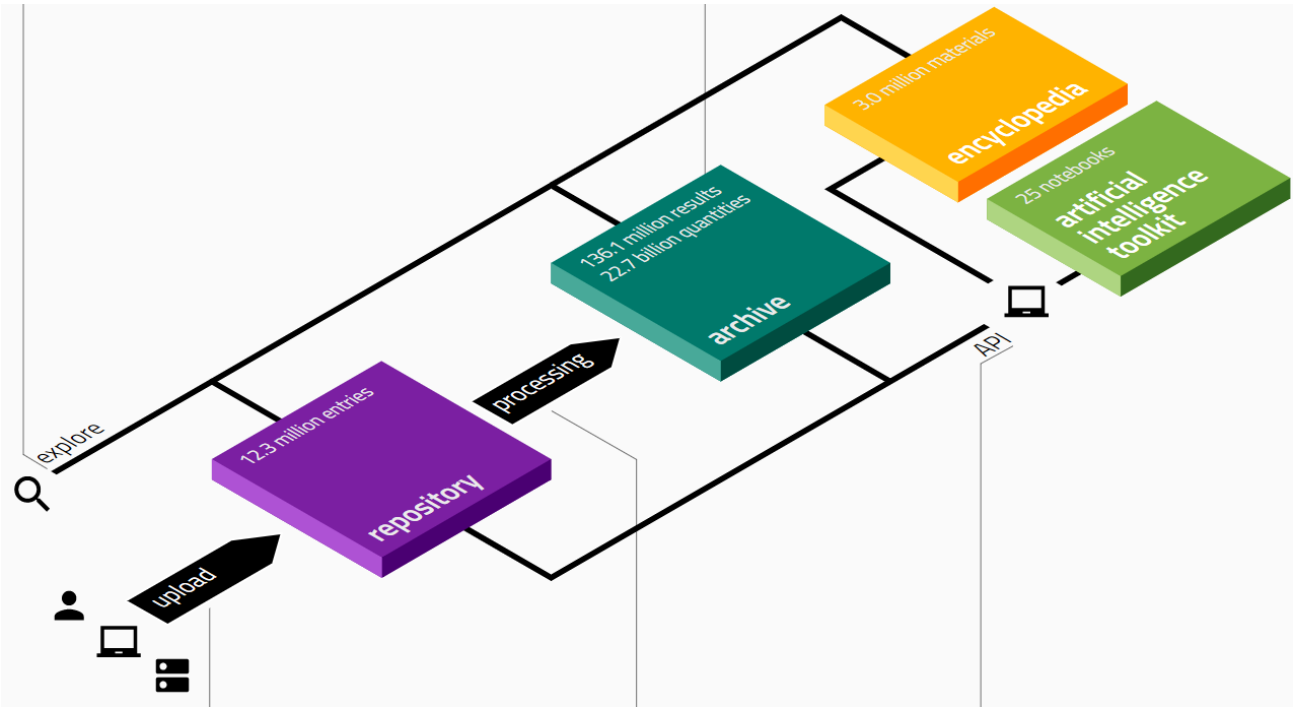

Figura 1. Rappresentazione grafica dei servizi offerti da NOMAD

NOMAD Metainfo consente di definire schemi per dati fisici indipendentemente dal formato di archiviazione utilizzato; di definire grandezze fisiche con tipi, forme complesse (vettori, matrici, ecc.), unità, collegamenti e descrizioni; di organizzare grandi quantità di queste quantità in gerarchie di contenimento di sezioni estensibili, riferimenti tra sezioni e categorie di quantità aggiuntive.

NOMAD utilizza le meta-informazioni per definire tutti i dati di archivio, i metadati del repository (e i dati dell'enciclopedia). Le meta-informazioni forniscono una comoda interfaccia Python per creare, manipolare e accedere ai dati mappandoli in vari formati di archiviazione, inclusi JSON, (HDF5), MongoDB ed Elastic Search.

In genere, le definizioni di (meta)dati vengono generate solo per un campo, un'applicazione o un codice predefinito e specifico. Al contrario, il NOMAD Metainfo considera tutte le informazioni pertinenti nei file di input e output dei codici supportati. Ciò garantisce una copertura completa di tutte le proprietà del materiale e della molecola, anche se alcune proprietà potrebbero non essere importanti quanto altre o mancano in alcuni file a supporto.

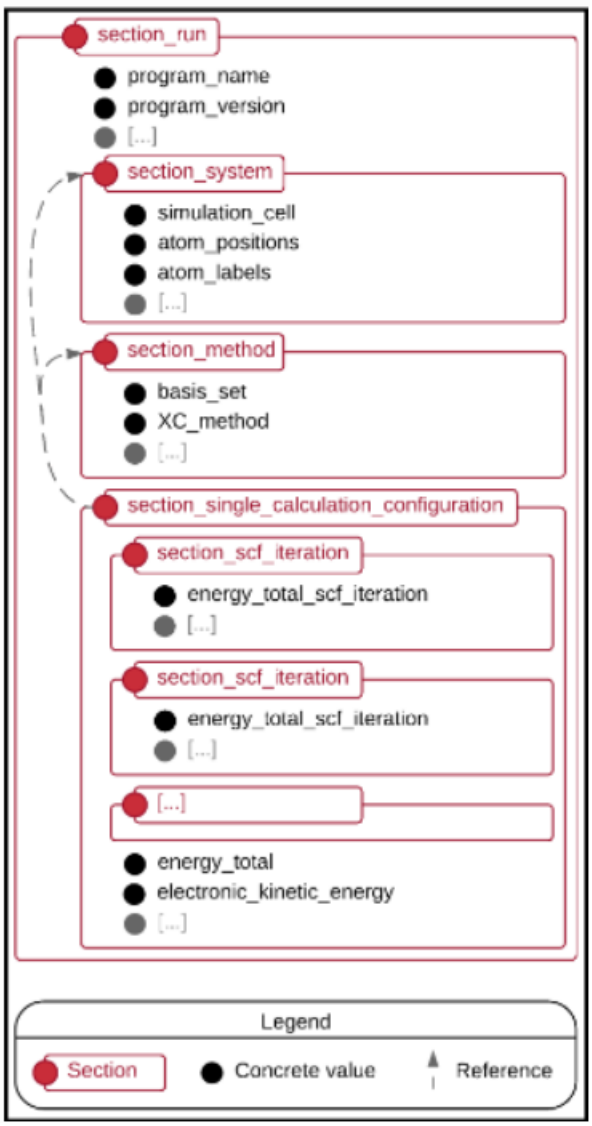

Figura 2. Schema di NOMAD Metainfo

3.1.3 Servizi

NOMAD offre tools allo stato dell'arte che ricercatori ed aziende possono utilizzare per sviluppare materiali migliori o completamente nuovi. I servizi offerti da NOMAD sono:

- **NOMAD Repository** è il repository contenente i file di input e output di oltre cento milioni di calcoli.
- **NOMAD Archive** offre una rappresentazione omogenea dei dati del repository processati, normalizzati e standardizzati.
- **NOMAD Encyclopedia** è uno strumento web per effettuare ricerche dettagliate di materiali nell'archivio dati NOMAD.
- **NOMAD Artificial Intelligence Toolkit** contiene strumenti di data mining applicati ai Big Data della scienza dei materiali per scoprire pattern o informazioni nuove. Tali strumenti utilizzano i più recenti approcci di machine learning e intelligenza artificiale.

I servizi offerti da NOMAD possono essere acceduti sia dall'utente umano tramite una interfaccia web, ma anche dall'utente macchina via API ReST. L'Application Programming Interface (API)[3] di NOMAD consente di accedere ai dati e alle funzioni NOMAD in modo automatico.

Esistono tuttavia altre API che possono accedere all'archivio dati di NOMAD e sono Open Data Integration for Materials Design (OPTIMATE)[4] e Data Catalog Vocabulary (DCAT)[5].

OPTIMADE mira a rendere interoperabili le basi di dati della scienza dei materiali attraverso la standardizzazione comune dei dati.

DCAT è un vocabolario RDF progettato per facilitare l'interoperabilità tra le banche di dati pubblicati sul Web.

3.2 Materials Project

3.2.1 Obiettivi

Materials Project, descritto in [JainOng2013], è una base di dati (open-access) contenente le proprietà calcolate di tutti i materiali conosciuti, mirando così a rimuovere le congetture della progettazione dei materiali in una varietà di applicazioni. Il set di dati computazionali permette alla ricerca sperimentale di essere indirizzata verso i composti più promettenti e ai ricercatori di essere in grado di estrarre trend scientifici per le proprietà dei materiali.

Materials Project usufruisce dei cluster di supercalcolo presenti nei laboratori degli Stati Uniti per elaborare calcoli, dati ed algoritmi ad alte prestazioni. I principali laboratori che vengono utilizzati sono il NERSC Scientific Computing Center e la Computational Research Division del Lawrence Berkeley National Laboratory, ma sono attivi anche con Leadership Computing Facility di Oak Ridge (OLCF), Leadership Computing Facility di Argonne (ALCF) e Super Computing di San Diego (SDSC).

Facendo uso dell'infrastruttura di calcolo dei diversi laboratori e di nuovi modelli predittivi sono riusciti ad identificare nuovi ossidi conduttivi trasparenti e materiali termoelettrici.

3.2.2 Metadati

A differenza di NOMAD, Materials Project genera internamente i dati estratti dai calcoli e per questo non vengono forniti i meccanismi di estrazione dei metadati. Tuttavia, il set di dati è accessibile ed esplorabile da chiunque sulla pagina web[6] e fornisce una serie di funzionalità su diversi portali:

- Explore Materials
- Explore Batteries,
- Crystal Toolkit
- Structure Predictor

[3] Documentazione delle API di NOMAD: https://nomad-lab.eu/prod/v1/api/v1/extensions/redoc
[4] https://www.optimade.org/
[5] https://www.w3.org/TR/vocab-dcat-2/
[6] https://materialsproject.org/

- Phase Diagram
- Pourbaix Diagram
- Calculate Reaction
- Thermo
- Compare Elements
- Nanoporous Explorer
- Explore Molecules
- RFB Dashboard
- XAS Matcher
- Interface Reactions

La sezione principale è sicuramente quella di esplorazione dei materiali applicando alcuni filtri (ad es. selezione degli elementi di interesse tramite una tavola periodica interattiva). La pagina restituirà una lista delle occorrenze a cui possono essere accedute nel dettaglio. Ciascun materiale riporta informazioni sulla struttura (lattice, siti, specie atomiche, volume), le proprietà calcolate (energia di formazione, energia finale, band gap, DoS, diffrazione a raggi X) e tutte le informazioni inerenti al calcolo (file di input e d i output, parametri). In aggiunta, il materiale riporta un identificativo univoco all'interno della piattaforma Materials Project per facilitare la ricerca e la condivisione e l'identificativo DOI con riferimento alla pubblicazione.

3.2.3 Servizi

Il progetto Materials rende disponibili i suoi dati e le sue analisi scientifiche attraverso l'API (Materials Application Programming Interface) e il pacchetto di analisi dei materiali Python Materials Genomics (pymatgen) open source. Mentre il front-end Web fornisce interfacce intuitive per la maggior parte degli utenti, l'API dei materiali e il pacchetto pymatgen forniscono agli utenti i mezzi per ottenere in modo efficiente grandi set di dati e sviluppare nuove analisi.

L'API dei materiali (MAPI)[7] è un'API per l'accesso ai dati di progetto sui materiali basata sui principi REpresentational State Transfer (REST). In un sistema RESTful, le informazioni sono organizzate in risorse che possono essere identificate in modo univoco tramite un identificatore di risorsa uniforme (URI).

Sebbene il MAPI sia progettato per essere indipendente dal codice che si utilizza e, quindi, possa essere plausibilmente utilizzato con qualsiasi linguaggio di programmazione che supporti le richieste http di base, nella libreria Python Materials Genomics (pymatgen) è stato implementato un comodo wrapper per facilitare i ricercatori nell'utilizzo di MAPI, chiamato MPRester.

3.3 Battery2030+

3.3.1 Obiettivi

[7] https://docs.materialsproject.org/downloading-data/using-the-api

BATTERY 2030+ è un'iniziativa di ricerca europea su larga-scala con lo scopo di creare una serie di strumenti atti a trasformare i processi di sviluppo e progettazione delle batterie in Europa (2020-2023).

L'iniziativa si compone di sette progetti (uno di coordinamento e 6 di ricerca):

- un'azione di coordinamento e supporto (CSA) coordinata da UU, Svezia;
- **BAT4EVER**, coordinato da VUB in Belgio;
- **BIG-MAP**, coordinato da DTU in Danimarca;
- **HIDDEN**, coordinato da VTT in Finlandia;
- **INSTABAT**, coordinato da CEA in Francia;
- **SENSIBAT**, coordinato da IKERLAN in Spagna;
- **SPARTACUS**, coordinato da Fraunhofer in Germania.

L'obiettivi dell'iniziativa BATTERY 2030+ sono:

- inventare batterie ad altissime prestazioni che siano sicure, convenienti e sostenibili, con una lunga durata.
- sviluppare nuovi prodotti chimici per batterie e concetti di batterie.
- fornire nuovi strumenti e tecnologie rivoluzionarie all'industria europea delle batterie lungo tutta la catena del valore, compresa la produzione e il riciclaggio.
- accelerare lo sviluppo delle batterie attraverso la realizzazione di nuovi strumenti nell'area della digitalizzazione.
- consentire una leadership europea a lungo termine in entrambi i mercati esistenti come i trasporti e lo stoccaggio stazionario e futuri settori emergenti come robotica, aerospaziale, dispositivi medici e Internet delle cose, ecc.

3.3.2 Metadati

In questa sezione, organizzata in uno o più capitoli, si descrive il lavoro svolto (teoria, metodologie sperimentali utilizzate, tecnologie sviluppate, ecc.) e s'illustrano i risultati ottenuti ed eventuali prodotti realizzati.

3.3.3 Servizi

In questa sezione, organizzata in uno o più capitoli, si descrive il lavoro svolto (teoria, metodologie sperimentali utilizzate, tecnologie sviluppate, ecc.) e s'illustrano i risultati ottenuti ed eventuali prodotti realizzati.

4 Descrizione delle attività svolte e risultati

La scienza dei materiali è una disciplina basata sulla chimica, sulla fisica e in parte sull'ingegneria, che tratta la progettazione, la produzione e l'uso di tutte le classi esistenti di materiali (tra cui i metalli, le ceramiche, i

semiconduttori, i polimeri e i biomateriali) e l'interazione dei materiali con l'ambiente, la salute, l'economia e l'industria.

I campi di applicazione di questa disciplina sono molteplici e, affinché venga definito un modello dati per la piattaforma IEMAP, è necessario comprendere lo scope e i limiti di dominio. Il progetto Mission Innovation si occupa delle aree di ricerca per la produzione di materiali per le batterie, gli elettrolizzatori e la perovskite. Questa risulta essere una base di partenza per la definizione dei dati gestiti della piattaforma. In aggiunta la piattaforma dovrà allocare informazioni ricavate sia da processi computazionali e sia da processi sperimentali. Come mostrato in precedenza, le principali piattaforme riconosciute a livello mondiale, come NOMAD, Materials Project e Battery2030+, si occupano di una fra le due tipologie di processo. La piattaforma IEMAP ha l'ambizione di aggregare le informazioni di entrambe le tipologie di processo.

Risulta necessario definire il dominio in funzione delle informazioni ricavate dai diversi partner di progetto. A tal proposito si è deciso di definire un sondaggio da condividere con i gruppi di lavoro, il quale viene descritto nel capitolo successivo.

4.1 Descrizione della survey

La survey ha l'obiettivo di definire le informazioni tecniche per i gruppi di lavoro, e quindi per i rispettivi partner, come il progetto di appartenenza, la tipologia di processo, i dati prodotti e la loro dimensione, gli strumenti o i programmi utilizzati, ecc. con finalità di raccordo tra i diversi gruppi, strumenti e codici, standard e protocolli utilizzati durante il processo di elaborazione.

Per la sua realizzazione è stato utilizzato lo strumento chiamato Google Forms, in quanto riporta una schermata user-friendly ed è facile da condividere.

La survey è composta da quattro sezioni che richiedono:

- L'identificazione del gruppo di lavoro
- Le informazioni sul processo computazionale (se previsto)
- Le informazioni sul processo sperimentale (se previsto)
- Il caricamento di file di esempio relativi all'attività e suggerimenti

La logica della survey è rappresentata dal diagramma di flusso seguente:

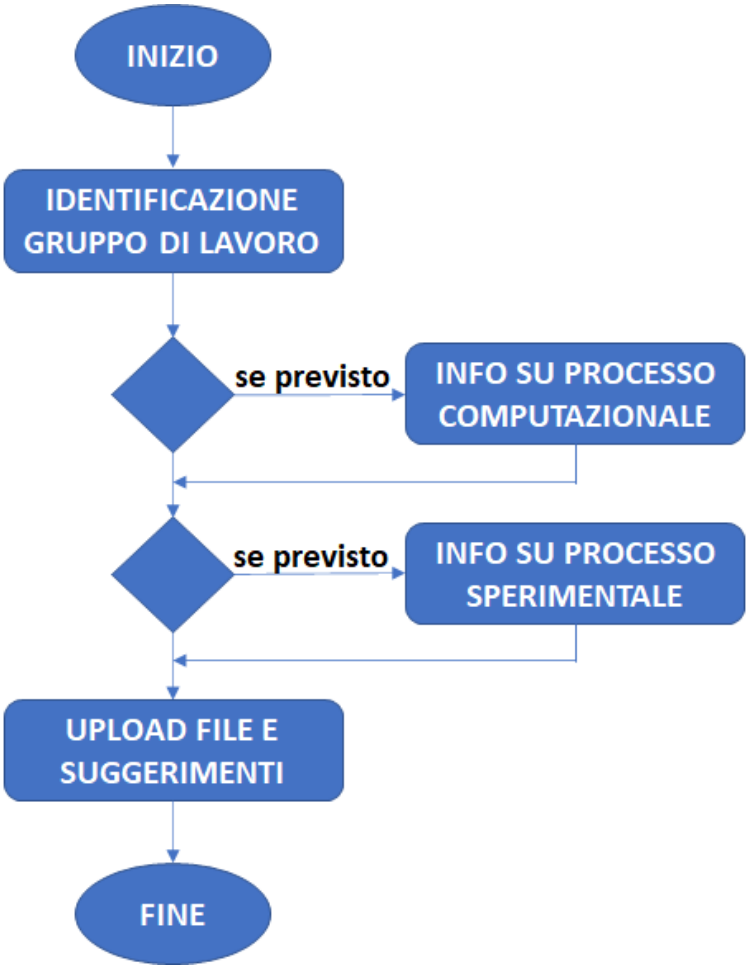

Figure 3. Diagramma di flusso della survey

Tutti i passaggi contenuti nel diagramma di flusso, riportato in

Figure 3, verranno descritti nel dettaglio nei capitoli successivi.

4.1.1 L'identificazione del gruppo di lavoro

La prima sezione della survey richiede all'utente di identificarsi riportando il proprio nome e cognome, l'affiliazione di appartenenza e una breve descrizione del caso d'uso. Tali informazioni sono necessarie a definire il contesto iniziale. In Figure 4 viene riportata la schermata appena descritta.

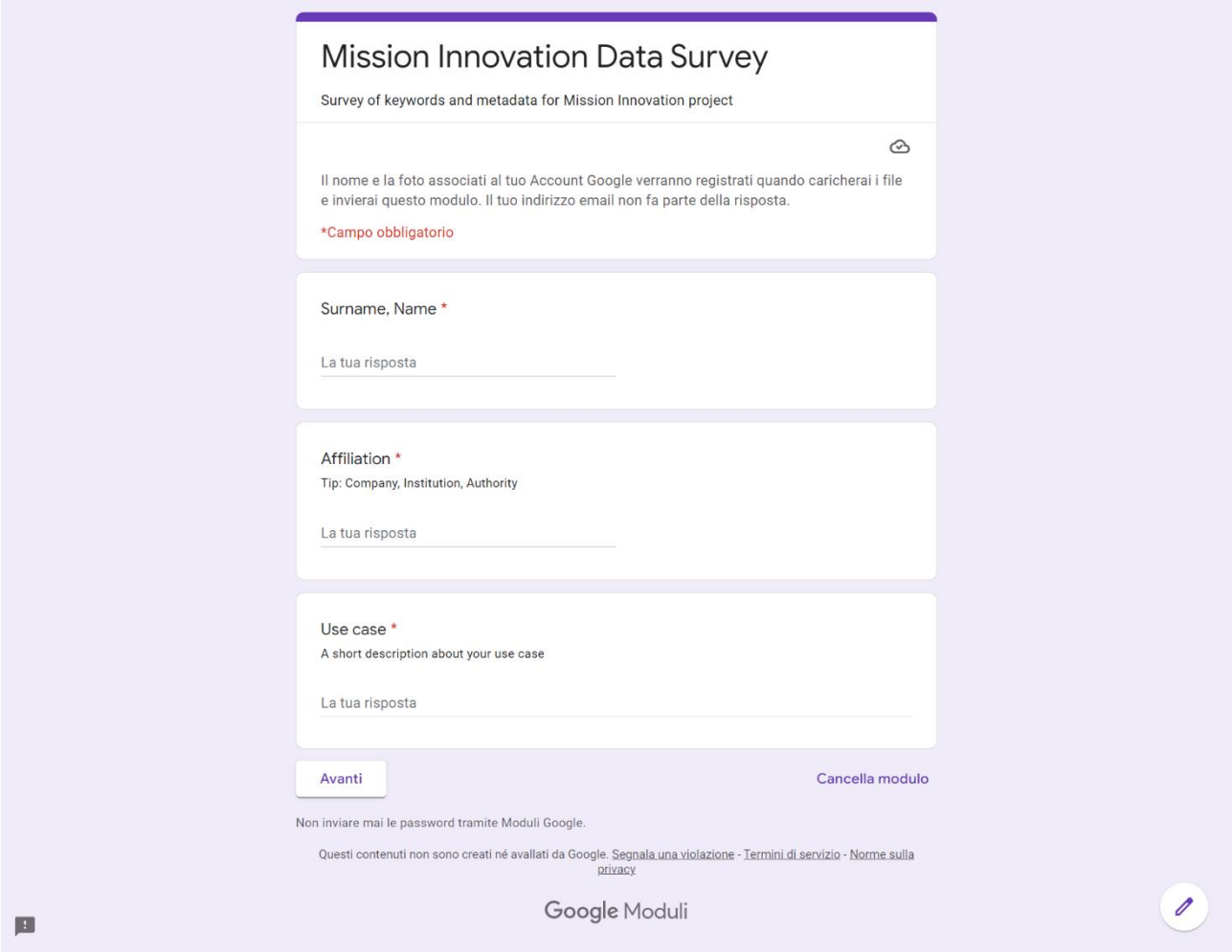

Figure 4. Modulo di identificazione del gruppo di lavoro

4.1.2 Richiesta informazioni su processo computazionale

La seconda sezione della survey richiede all'utente le informazioni inerenti al processo computazionale (se previsto). Tali informazioni vengono mostrate in Figure 5 e riguardano il proprio processo di elaborazione, detto workflow, il software utilizzato per l'elaborazione e informazioni tecniche sul formato, la dimensione e la frequenza di produzione del file di output.

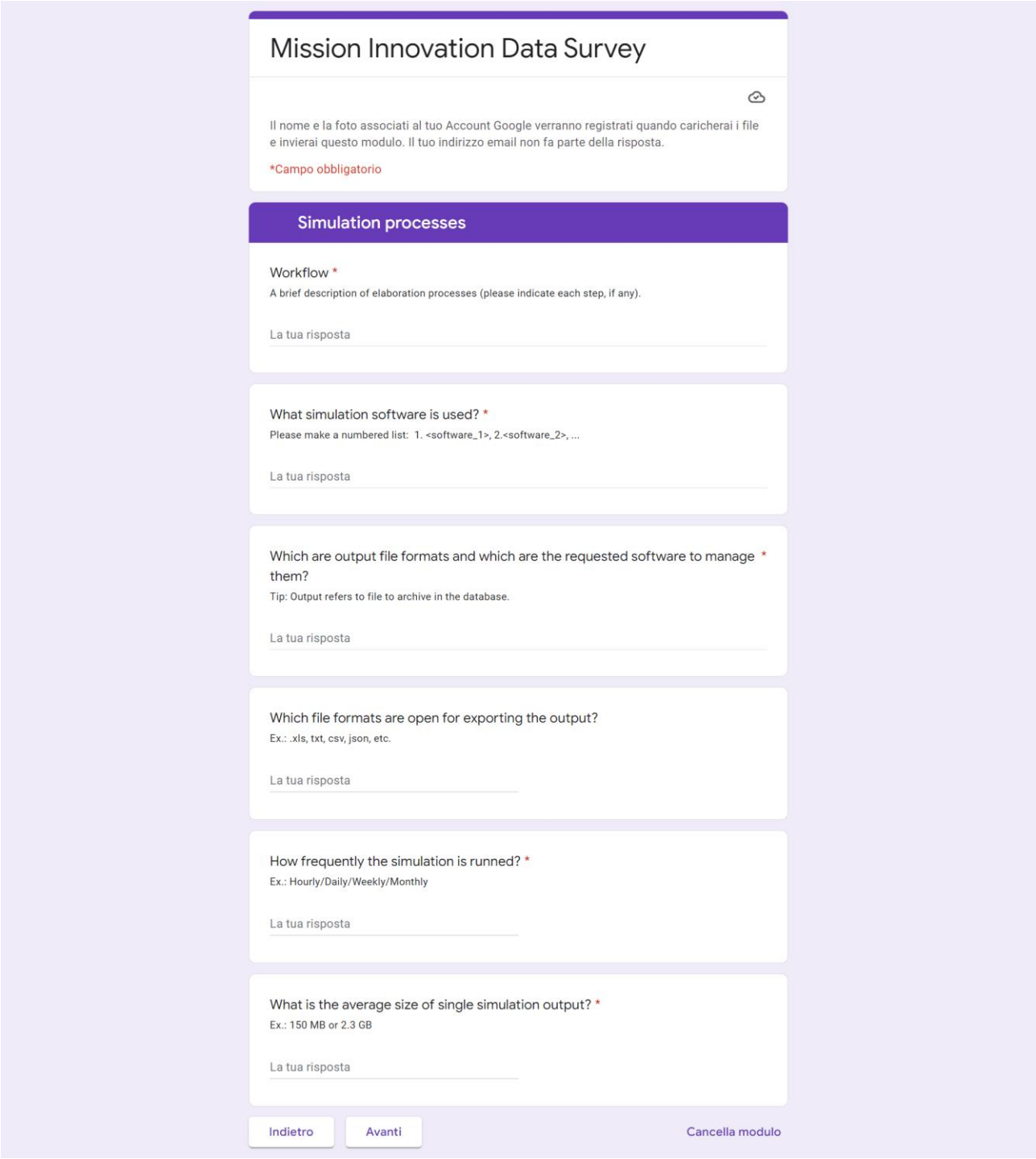

Figure 5. Modulo di informazioni per il processo computazionale

4.1.3 Richiesta informazioni su processo sperimentale

La terza sezione della survey richiede all'utente le informazioni inerenti al processo sperimentale (se previsto) e si compone di quattro parti, più una di raccordo. Quest'ultima viene mostrata in

Figure 6 e richiede informazioni tecniche riguardo la frequenza e la dimensione in memoria dell'esperimento.

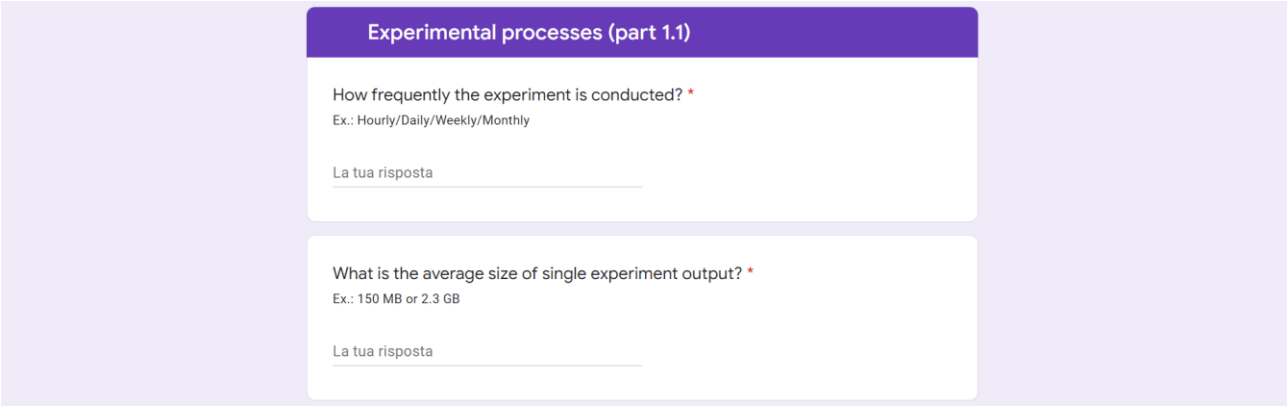

Figure 6. Modulo di informazioni per il processo sperimentale (parte I)

Negli step successivi, se presenti, si richiedono informazioni sui dati grezzi generati dallo strumento (

Figure 7), sul formato di immagini e dati tabellari (rispettivamente in

Figure 8 e Figure 10) e su possibili protocolli o standard utilizzati per generare l'output (

Figure 9).

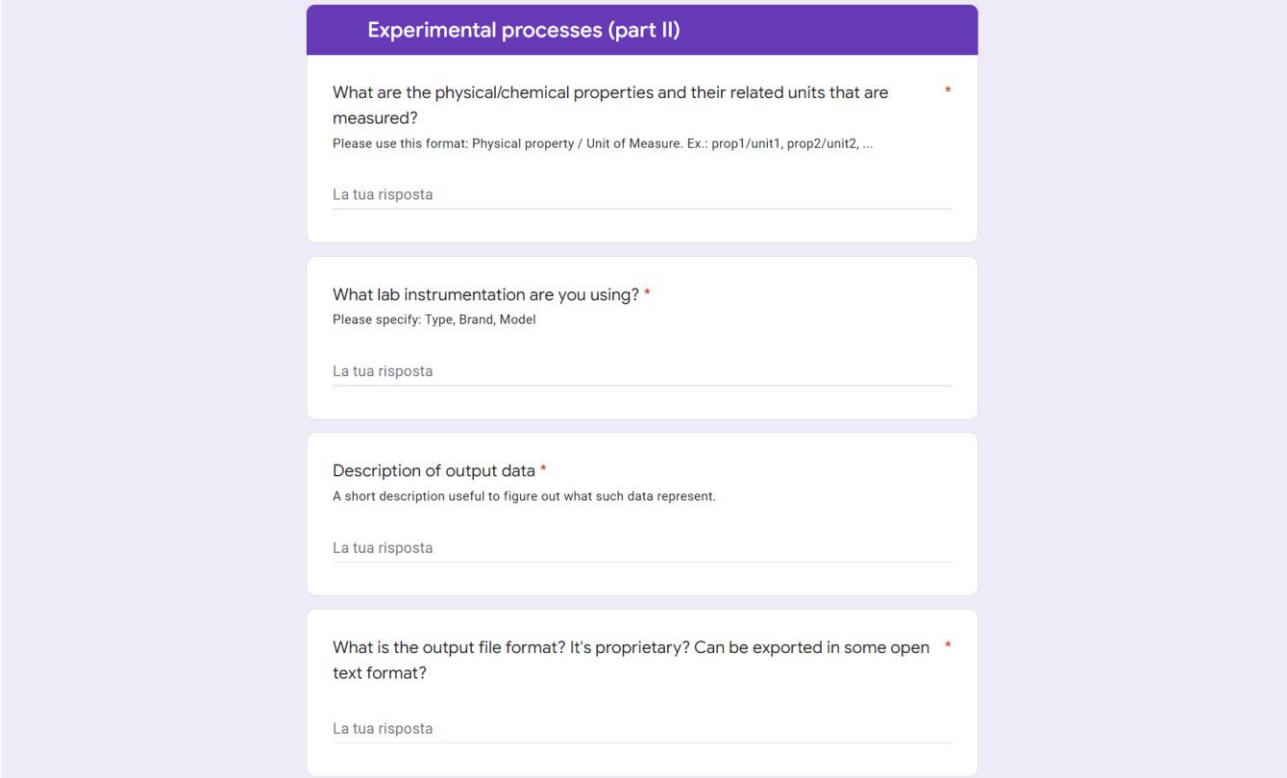

Figure 7. Modulo di informazioni per il processo sperimentale (parte II)

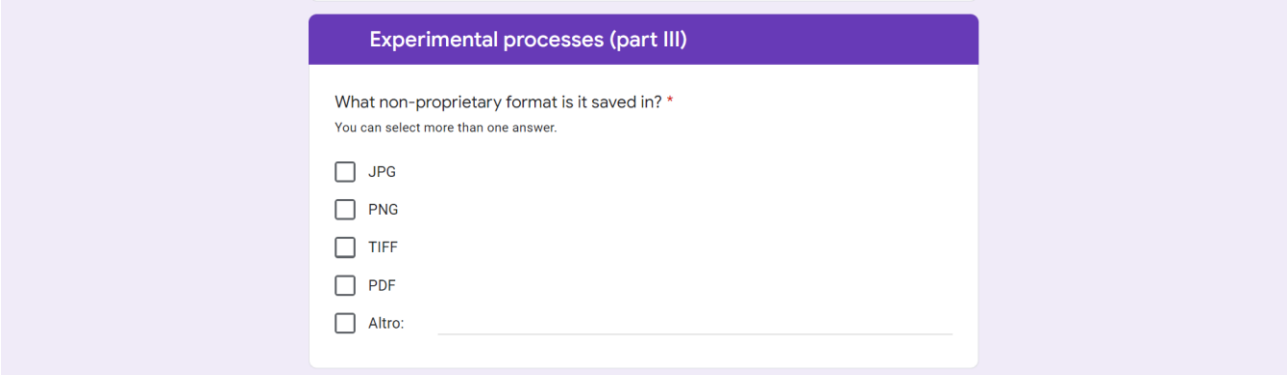

Figure 8. Modulo di informazioni per il processo sperimentale (parte III)

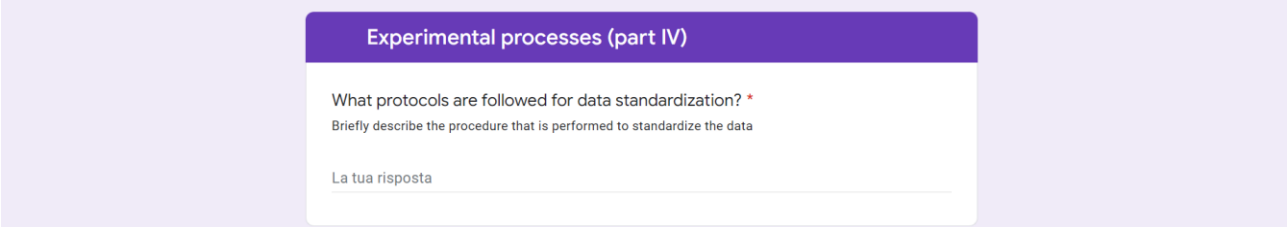

Figure 9. Modulo di informazioni per il processo sperimentale (parte IV)

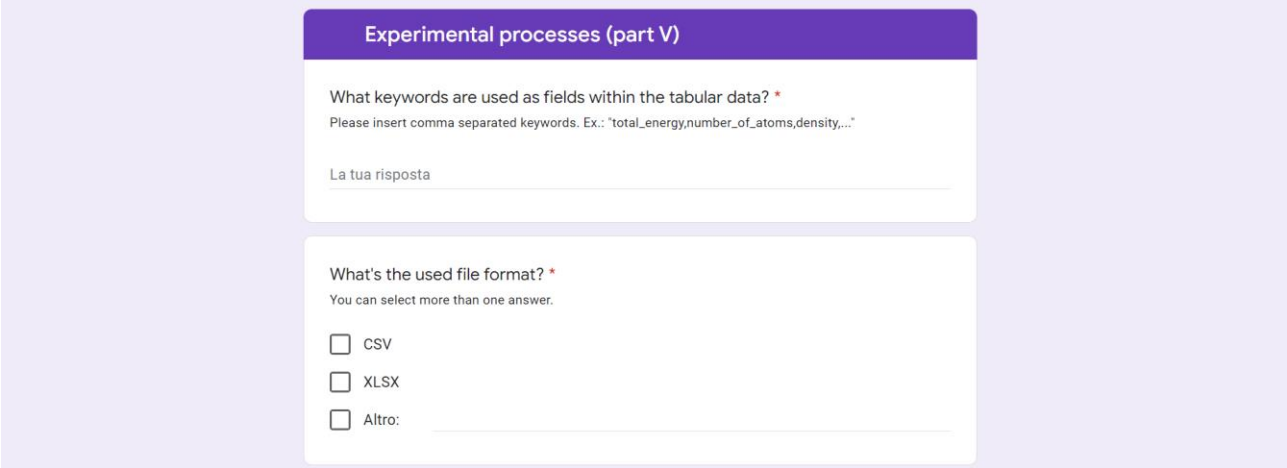

Figure 10. Modulo di informazioni per il processo sperimentale (parte V)

4.1.4 Conclusioni

La quarta ed ultima sezione della survey richiede all'utente il caricamento di file di dati da considerare come esempio e suggerimenti (

Figure 11).

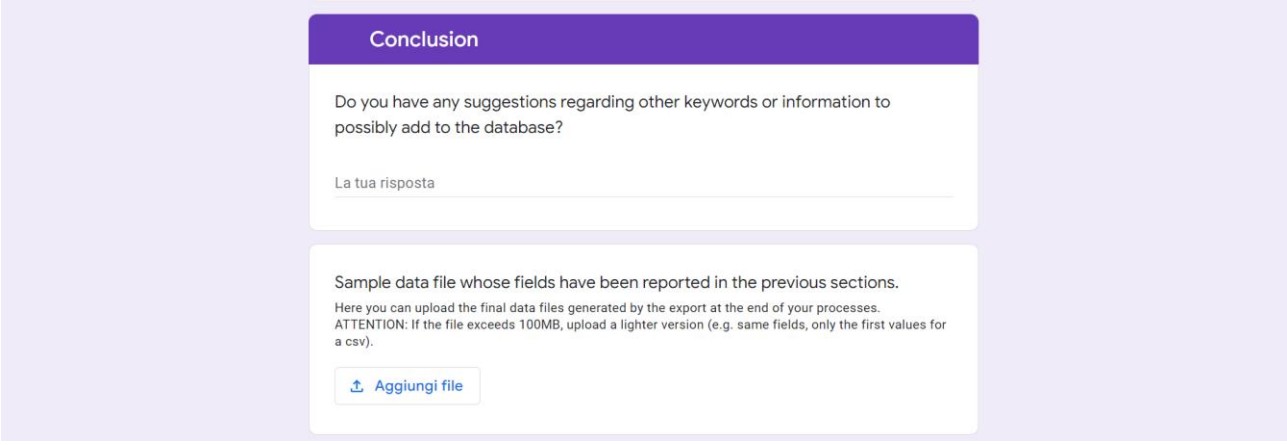

Figure 11. Modulo conclusivo

4.2 Risultati della survey

La survey ha permesso di definire gli aspetti generali e tecnici derivanti dalle diverse applicazioni gestite dai gruppi di lavoro.

Nello specifico, le persone che hanno partecipato sono elencate di seguito ordinati per nome:

- Abagnale Giovanni (RSE),
- Arciniegas Milena (IIT),
- Buonocore Francesco (ENEA),
- Celino Massimo (ENEA),
- Ferrario Alberto (CNR-ICMATE),
- Filippone Francesco (CNR-ISM),
- Fontana Danilo (ENEA),
- Giovanni Battista Appetecchi (ENEA),
- Liotta Leonarda Francesca (CNR-ISMN),
- Lisi Nicola (ENEA),
- Mercaldo Lucia Vittoria (ENEA),
- Montanino Maria (ENEA),
- Nicola Briguglio (CNR-ITAE),
- Protopapa Maria Lucia (ENEA),
- Siracusano Stefania (CNR-ITAE),
- Zani Lorenzo (CNR-ICCOM)

Il numero di gruppi di lavoro che hanno partecipato alla survey per ciascun ente è riportato in Figure 12. Il Consiglio Nazionale di Ricerca (CNR) è costituito da diversi istituti di ricerca dislocati geograficamente sul territorio italiano e sono l'Istituto di Struttura della Materia (ISM), l'Istituto di Chimica della Materia Condensata e di Tecnologie per l'Energia (ICMATE), l'Istituto di Tecnologie Avanzate per l'Energia (ITAE), l'Istituto di Chimica dei Composti Organometallici (ICCOM) e l'Istituto per lo studio dei materiali nanostrutturati (ISMN).

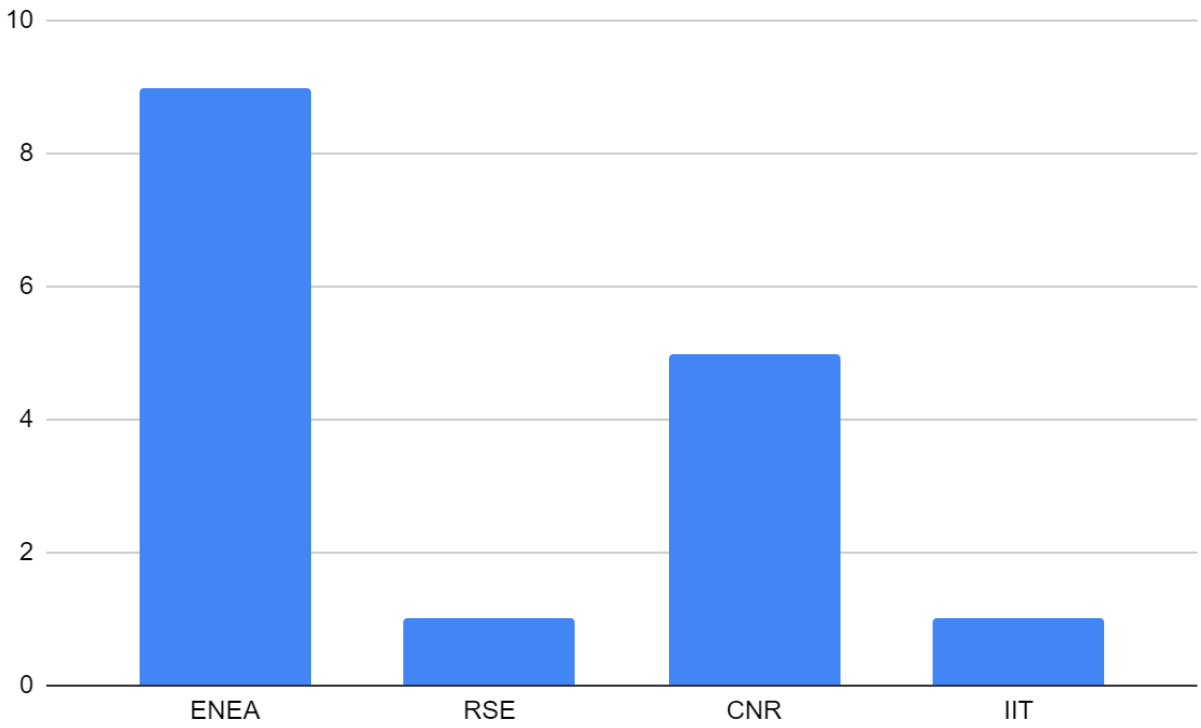

Figure 12. Numero di partecipanti per ente

Dalla survey si è dedotto che i gruppi lavoreranno principalmente su processi sperimentali, come mostrato in **Errore. L'origine riferimento non è stata trovata.**, e solo il gruppo del partner 'Ricerca sul Sistema Energetico' (RSE) è coinvolto in entrambi i processi.

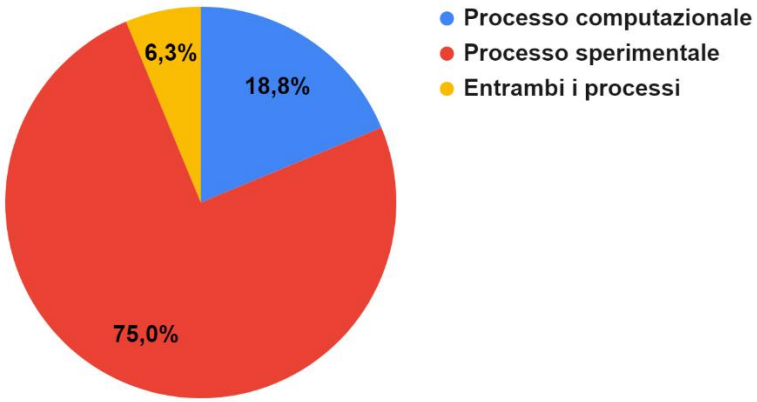

Figure 13. Numero di gruppi di lavoro coinvolti sui processi computazionali, sperimentali o entrambi

I gruppi coinvolti nei processi computazionali utilizzeranno codici custom scritti con il linguaggio di programmazione C++ oppure codici prodotti da terzi e riconusciuti a livello mondiale con licenza a pagamento o open-source, come Quantum Espresso, AiiDA, Yambo, Orca, CP2K, Gibbs2 e Critic2. I risultati generati come output da questi software hanno una dimensione dell'ordine di qualche MB, ad esclusione di

dati definiti durante la fase di elaborazione (GB), e saranno prodotti con frequenza per lo più giornaliera in formati aperti (txt, json, xls, hdf5, yaml) o binari.

Table 1. Lista degli strumenti di laboratorio

ENTE	Partecipante	Tipologia	Brand	Modello
CNR	Ferrario Alberto	Impedance Analyzer	Gamry	Reference 600
	Zani Lorenzo	UV-Vis spectrometer	Shimadzu	UV-2600
		Spectrofluorimeter	JASCO	FP-8300
		A source meter unit	Agilent	B2910
	Liotta Leonarda Francesca	X-ray powder Diffractometer	Bruker-Siemens	D5000
		Micromeritics	Autochem	2910
		Micromeritics	ASAP	2020
		Thermical Analyzer	Mettler	TGA/DSC1 STAR
		Potentiostatical frequency-response analyzer	Autolab	PGSTAT30
ENEA	Mercaldo Lucia Vittoria	Spectrophotometer	Perkin Elmer	lambda 900
	Lisi Nicola	VG Escalab MKII		
		Oscilloscopes Tektronix		
		Sourcemeter Keithley		
		Plasma optical emission spectrometer Ocean Optics		
		Optical Microscopy		
		Electron Microscopy		
		X ray Diffractograms		
	Montanino Maria	VISCOSIMETER		
		PROBES CONDUCTIMETER		
		SCANNING ELECRTON MICROSCOPY		

ENTE	Partecipante	Tipologia	Brand	Modello
		GRAVURE PRINTER		
		BALANCE		
		DENSIMETER		
		PH-METER		
		THERMOMETER		
	Giovanni Battista Appetecchi	Hood		
		lab-glassware		
		mixer		
		rotary-evaporator		
		vacuum even		
		glove-bis		
	Protopapa Maria Lucia	Raman spectrometer LabRam HR Evolution (HORIBA)		
		AFM (HORIBA),		
		FTIR IS50 (ThermoFisher) provided with ATR module		
		XRD (Empyrean Panalytical)		
		SEM/EDX		
		Dynamic Light Scattering (Malvern)		
IIT	Arciniegas Milena	Spectrometer		
RSE	Abagnale Giovanni			Aixtron G4

I gruppi coinvolti nei processi sperimentali fruirà delle macchine strumentali di laboratorio riportate in Table 1. I file generati dalla strumentazione e dalle analisi svolte probabilmente saranno prodotte con frequenza giornaliera o mensile, riportando una dimensione che si aggira tra le decine di MB e un GB. I dati prodotti vengono principalmente memorizzati nei formati proprietari del produttore che potranno essere riportati anche in formati open-source (ad es. csv, txt). Più della metà dei gruppi riporta anche dati in immagini nei formati non proprietari, come JPG, PNG, TIFF e PDF.

I dati riporta informazioni su parametri impostati nel processo sperimentale e le proprietà ricavate da esso. Tali informazioni sono riportate in Table 2.

Table 2. Dati dei processi sperimentali

Property name	Property unit	Alias	Formula	Description
coeff.esp.termica	$°C^{-1}$	CLTE/CTE	$\alpha = \Delta L / (L0 * \Delta T)$	Coefficient of Linear Thermal Expansion
cost.reticolare	Angstrom	lattice	(a, b, c), (α, β, γ)	
Young Module	dyne/cm2	Young's modulus		
band gap	eV	Bandgap		
Temperature	°C			
Pressure	mbar			
Flow	ml			
Wavelength	nm			
Transmittance	%			
Reflectance	%			
Surface composition	%			
chemical state	%			
Workfunction	eV			
Resistivity	ohm			
plasma characteristics	(I , V , W)			
WEIGHT	g			
DENSITY	g/cm3			
PRINTING FORCE	N			
PRINTING SPEED	m/min			
CORONA TREATMENT	W			
VISCOSITY	mPas			
LAYER CONDUCTIOVITY	S/cm2			
Purity	wt.%			
Ionic conductivity	mS cm-1			
Thermal stability	°C			
electrochemical stability	V			
photoluminescence quantum yield	%			
Frequency	Hz			
Current	Ampere			
Chemical bonds/Intensity of ligh as a function of Raman shift	Raman analysis			
Chemical bonds/intensity of ligh as a function of light frequency	FTIR analysis			
Semi-quantitative recognition of crystallographic phases and crystallite size/Intensity as a function of the diffracted angle	X-Ray diffraction			

Property name	Property unit	Alias	Formula	Description
Surface morphology/ Height profile as a function of the position	AFM			
Particle size dimension distribution/ Number of particles as a function of particle size	Dinamic Light Scatytering			

4.3 Analisi dei requisiti

In questa sezione vengono chiariti i requisiti proponendo i casi d'uso (Use Cases, UC), le domande sulle competenze (Competency Questions, CQ) e restrizioni aggiuntive.

I casi d'uso, che sono stati identificati tramite uno studio della letteratura e discussi tra gli esperti del dominio e gli ingegneri con esperienza pregressa basata sull'uso di database della scienza dei materiali, sono elencati di seguito:

- UC1: La piattaforma IEMAP sarà utilizzata per rappresentare la conoscenza nella scienza dei materiali di base come la fisica dello stato solido e la teoria della materia condensata nel campo delle batterie, degli elettrolizzatori e del fotovoltaico
- UC2: La piattaforma IEMAP sarà utilizzata per memorizzare processi computazionali e sperimentali eterogenei sui materiali

A partire dai risultati della survey, lo scope e i limiti della piattaforma sono stati definiti e tradotti in domande di competenze che il sistema dovrà poter rispondere (riportate in Table 3). Le domande sulle competenze della piattaforma si basano sulla discussione con gli esperti del dominio e contengono le domande che gli esperti vorrebbero porre al database.

Table 3. Domande sulle competenze della piattaforma IEMAP

ID	Processo computazionale	Processo sperimentale
CQ1	Quali sono le proprietà calcolate e i relativi valori prodotti da un calcolo sui materiali?	Quali sono le proprietà analizzate e i loro valori prodotti da un esperimento sui materiali?
CQ2	Quali sono le strutture di input e output di un calcolo dei materiali?	Quali sono le composizioni di un esperimento sui materiali?
CQ3	Qual è il tipo di gruppo spaziale di una struttura?	
CQ4	Qual è il tipo di reticolo di una struttura?	
CQ5	Qual è la formula chimica di una struttura?	Qual è la formula chimica di una composizione?
CQ6	Per una serie di calcoli sui materiali, quali sono le composizioni dei materiali con un intervallo specifico di una proprietà calcolata (ad esempio band gap)?	Per una serie di esperimenti sui materiali, quali sono le composizioni dei materiali con un intervallo specifico di una proprietà analizzata (ad esempio la conduttività)?

ID	Processo computazionale	Processo sperimentale
CQ7	Per un materiale specifico e un dato intervallo di una proprietà calcolata (ad esempio, band gap), qual è il tipo di reticolo della struttura?	
CQ8	Per un materiale specifico e un tipo reticolare previsto di struttura di output, quali sono i valori delle proprietà calcolate dei calcoli?	
CQ9	Qual è il metodo di calcolo utilizzato nel calcolo dei materiali?	Qual è la tecnica sperimentale utilizzato in un esperimento sui materiali?
CQ10	Qual è il valore per un parametro specifico (ad es. energia di taglio) del metodo utilizzato per il calcolo?	Qual è il valore per un parametro specifico (ad es. temperatura) del metodo utilizzato per l'esperimento?
CQ11	Quale software ha prodotto il risultato di un calcolo?	Quale strumento ha prodotto il risultato di un esperimento?
CQ12	Chi sono gli autori del calcolo?	Chi sono gli autori dell'esperimento?
CQ13	Con quale software o codice viene eseguito il calcolo?	
CQ14	Quando sono stati pubblicati i dati di calcolo nel database?	Quando sono stati pubblicati i dati dell'esperimento nel database?
CQ15	Ci sono pubblicazioni relative ai dati?	Ci sono pubblicazioni relative ai dati?

Inoltre, proponiamo una lista di restrizioni aggiuntive che aiutano nella definizione dei concetti.

- Una proprietà del materiale può riferirsi ad una struttura
- Un processo computazionale ha esattamente un metodo di calcolo
- Un processo sperimentale ha esattamente una tecnica sperimentale
- Una struttura è composta da un certo numero di elementi
- Un processo computazionale viene eseguito da alcuni software o codici
- Un processo sperimentale viene analizzato da alcuni strumenti o macchinari
- Una struttura e una proprietà possono essere pubblicate da banche dati o pubblicazioni (riferimenti).

4.4 Sviluppo di un modello logico

Un modello di dati, in poche parole, è un insieme di specifiche e diagrammi di dati per spiegare i requisiti dei dati e i relativi progetti. Un modello logico di dati serve a definire come un sistema deve essere implementato, indipendentemente dal sistema di gestione di database utilizzato. Gli architetti di dati e gli analisti aziendali sono di solito i creatori di un modello logico di dati. L'obiettivo della creazione di un modello logico di dati è sviluppare una mappa altamente tecnica delle regole e delle strutture di dati alla base.

La modellazione logica dei dati appartiene al modello entità-relazione, costruito usando un diagramma di relazione fra le entità (noto come ERD), una tecnica di modellazione standard usata come strumento di comunicazione dai modellatori di dati di tutto il mondo.

Sulla base dei concetti e delle restrizioni discussi nel capitolo precedente, sono stati definiti due modelli logici inerenti ai dati memorizzati e gestiti dalla piattaforma IEMAP, chiamati Core Data e Log Data. Il primo modello

rappresenta le informazioni inerenti ai processi computazionali e sperimentali, mentre il secondo ha l'obiettivo di gestire i file caricati ed accoppiarli con i dati estratti dei vari processi.

Prima di procedere con i modelli dati è giusto accennare al processamento dell'informazione sulla piattaforma IEMAP.

Uno dei casi d'uso della piattaforma IEMAP è quello di permettere il caricamento dei dati e dei file annessi. Il dato viene archiviato sulla base di dati (MongoDB) e racchiude le informazioni principali e si lega ad un singolo processo computazionale o sperimentale. All'interno della base di dati vengono memorizzate solo le informazioni oggetto di query (interrogazione). Il dato deve contenere anche i riferimenti ai file di processo e, come indicato nella survey, possono essere di varia natura e formato (file testuali, immagini o binari). I file e il loro contenuto non sono oggetto di query; quindi, non vengono incapsulati all'interno del dato.

4.4.1 Core Data

I dati fondamentali inerenti ai processi computazionali e sperimentali sono stati riportati nel modello chiamato Core Data, che si basa sull'articolo [Andersen2021]. Il modello logico del Core Data, mostrato in Figure 14, è costituito da quattro moduli principali, tre sono specifici del dominio, *Material*, *Process* e *Property*, mentre uno è aggiuntivo ed è relativo alla provenienza delle informazioni, *Provenance*.

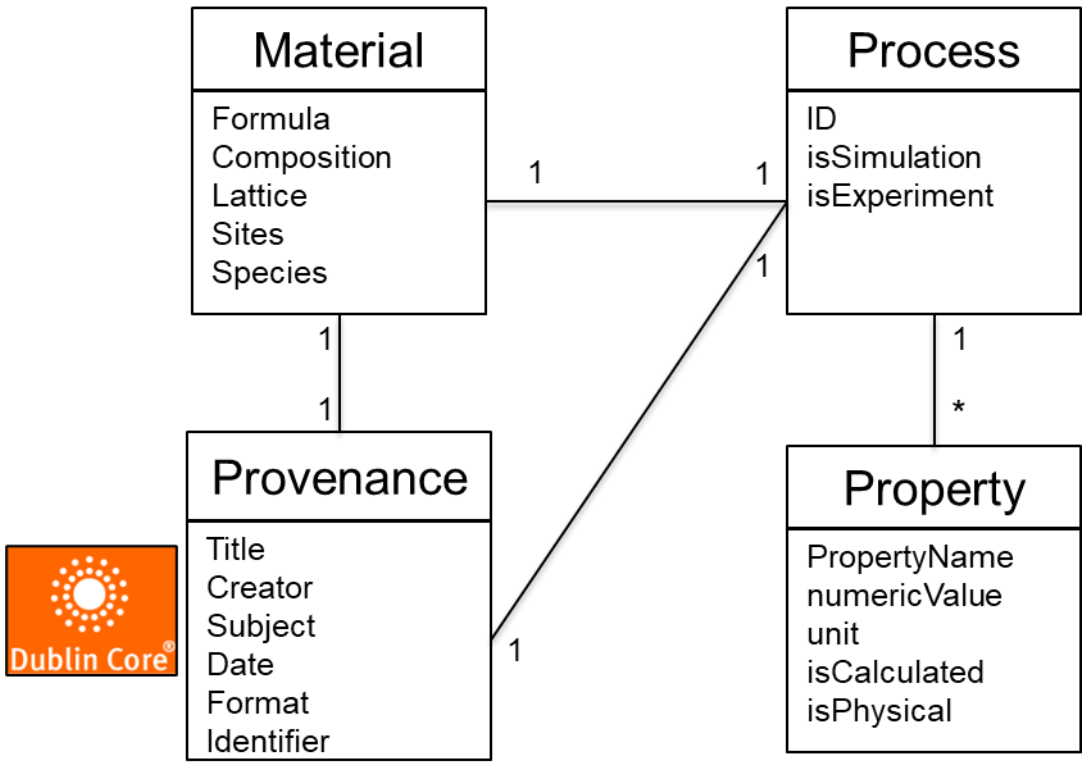

Figure 14. Modello logico relativo al Core Data

Il modulo ***Material***, come mostrato in Figure 15, rappresenta le informazioni strutturali dei materiali. Ogni struttura ha esattamente una composizione che definisce quali elementi chimici compongono la struttura e la percentuale degli elementi all'interno della struttura. La composizione ha diverse rappresentazioni di

formule chimiche. In aggiunta, i processi computazionali definiscono l'occupazione della struttura che mette in relazione i siti (S*ites*) con le specie atomiche (*Species*) che le occupano.

Il modulo ***Process***, come mostrato in

Figure 16, rappresenta le informazioni inerenti ai processi. Il modulo è specializzato nei concetti disgiunti di *Calculation* o *Esperiment*, rispettivamente classificati dal metodo di calcolo o dalla tecnica sperimentale e dal concetto di *Agent*, che identifica il software o lo strumento utilizzato durante l'esecuzione o l'analisi di processo. In aggiunta, ogni processo ha alcuni parametri.

Il modulo ***Property*** rappresenta le informazioni relative alle proprietà calcolate e non del materiale.

Il modulo ***Provenance*** rappresenta le informazioni di provenienza del dato del materiale e del processo. Il modulo fa riferimento allo standard *Dublin Core*.

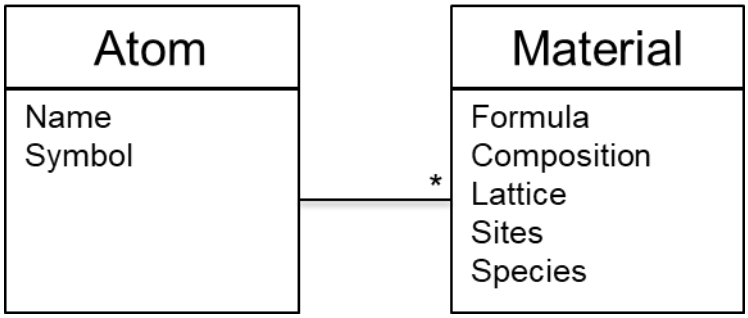

Figure 15. Modulo logico di 'Material'

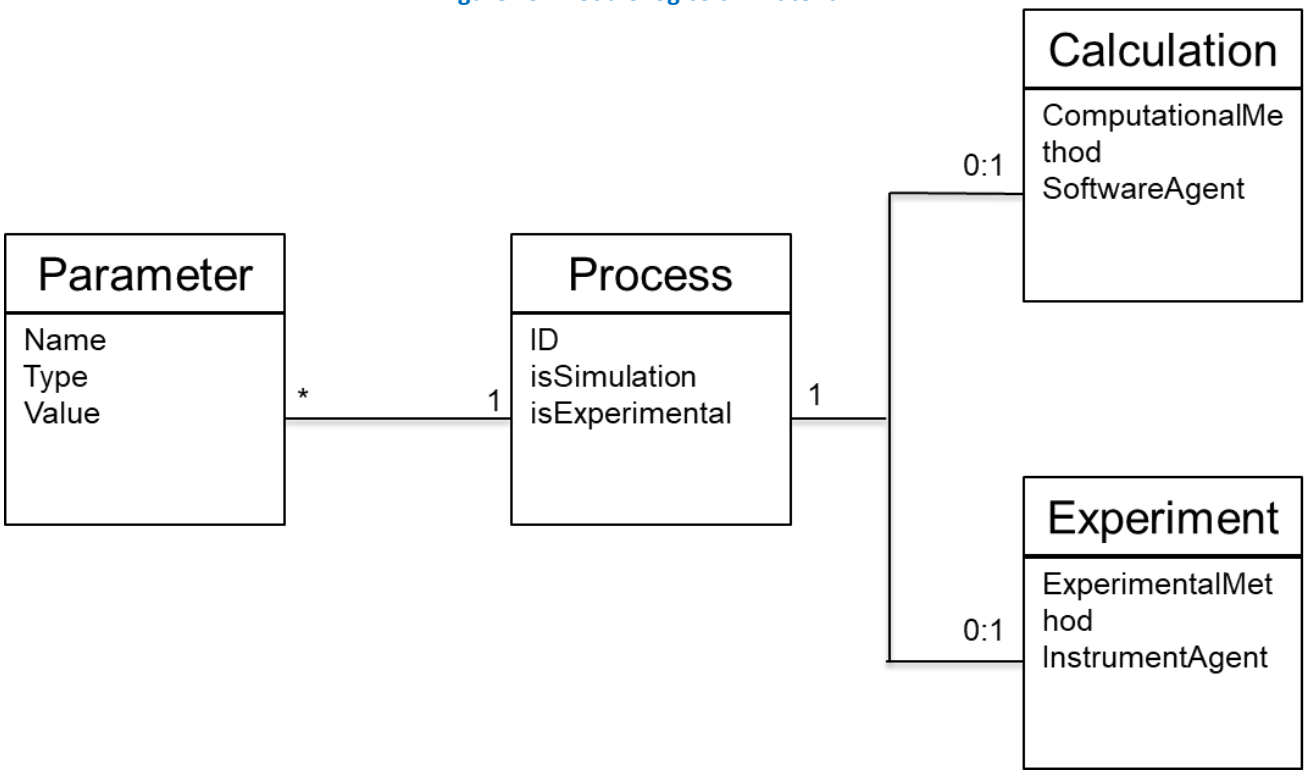

Figure 16. Modulo logico di 'Process'

4.4.2 Log Data

Un registro delle transazioni è un file (cioè un log) delle comunicazioni tra un sistema e gli utenti di quel sistema, o un metodo di raccolta dati che acquisisce automaticamente il tipo, il contenuto o l'ora delle transazioni effettuate da una persona da un terminale con tale sistema.

Il modello logico dei registri, chiamato Logs, rappresenta le informazioni legate al caricamento dei dati al fine di monitorare le transazioni di ogni singolo dato. Attraverso questo modello, è possibile definire la provenienza del dato, ovvero chi ha caricato il dato, quando è stato caricato e a quale progetto si lega.

Il modulo, mostrato in Figure 17, rappresenta le informazioni di registro generali dei dati caricati. I concetti *Upload* e *File* rappresentano il caricamento dei file indicati dall'utente, mentre *User* e *Project* rappresentano rispettivamente le informazioni principali sull'utente e sul progetto.

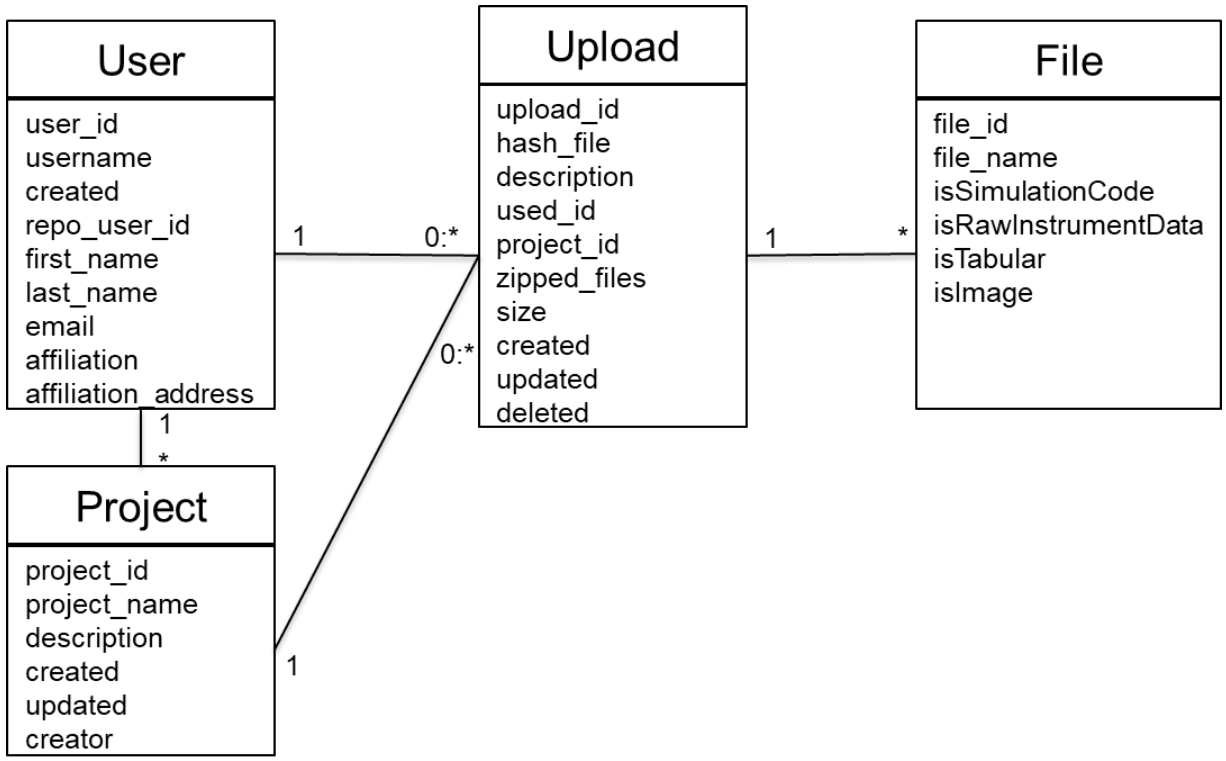

Figure 17. Modello logico relativo al 'Log Data'

4.5 Sviluppo di un modello fisico

Il modello fisico di dati riguarda il modo in cui il sistema sarà implementato e i fattori del sistema di gestione dei database specifico. Questo modello è solitamente creato dagli sviluppatori. L'idea è più che altro definire come il database reale sarà usato o implementato per scopi di business. In generale, la modellazione logica dei dati è un'attività di tipo "analisi dei requisiti", mentre la modellazione fisica dei dati è considerata un'attività di progettazione.

Un modello logico di dati serve come base per un modello fisico di dati, incorporando requisiti di business e raccogliendo metadati. La modellazione logica può essere effettuata usando tecniche standard e notazioni

di modellazione dei dati. La modellazione dei dati è un'attività orientata all'organizzazione della semantica dei dati, alla loro descrizione e ad affrontarne i limiti di coerenza.

Le tipologie di dati che devono essere memorizzati sono due:

- Dati oggetto di query e di audit
- File di elaborazione e di analisi

I dati oggetto di query e di audit vengono archiviati nella base di dati, mentre i file relativi all'elaborazioni dei processi computazionali o all'analisi dei processi sperimentali vengono memorizzati su un repository.

4.5.1 Dati oggetto di query e di audit

Nonostante il modello logico (Capitolo 4.4) venga riportato come diagramma Entità-Relazioni, il modello dati fisico si deve legare al tipo di base di dati, ovvero MongoDB; un potente sistema di database open source e gratuito non relazionale, che invece di memorizzare i dati in righe e colonne, adotta un design orientato ai documenti che rappresenta i dati in vari documenti e collezioni simili a JSON. Questi documenti contengono una serie di coppie di valori o chiavi di diversi tipi, come documenti annidati e array. Le coppie chiave/valore possono essere strutturate diversamente da un documento all'altro.

MongoDB offre maggiore sicurezza, affidabilità ed efficienza oltre alla flessibilità di modificare la struttura o lo schema dei dati. In cambio si ottengono requisiti superiori in termini di velocità e archiviazione.

Il modello dati fisico dei dati oggetto di query e di audit viene definito come un documento in formato BSON strutturato dalle coppie chiave/valore, i cui campi chiave sono mostrati in Table 4.

Table 4. Schema dati del modello fisico

Nome	Genitore	Tipo	Indice	Descrizione
_id	root	Object		Identificativo del documento BSON legato ad un singolo processo (computazionale o sperimentale)
createdAt	root	Date		Data di creazione del documento nel formato
updatedAt	root	Date		Ultima data di aggiornamento del documento
userEmail	root	String		Email dell'utente che ha caricato i dati
Affiliation	root	String		Affiliazione dell'utente che ha caricato i dati
projectName	root	String		Nome del progetto a cui il documento si lega
projectWP	root	String		Work Package di progetto
projectDescription	root	String		Descrizione del progetto
projectLabel	root	String	Yes	Etichetta di progetto

Nome	Genitore	Tipo	Indice	Descrizione
Process	root	Object		Entità che definisce i dati inerenti al processo
iemapID	process	String		
isExperiment	process	Boolean	Yes	True se il processo è sperimentale; altrimenti False
isSimulation	process	Boolean	Yes	True se il processo è computazionale; altrimenti False
Parameters	process	Array		Lista dei parametri di processo
Calculation	process	Object		Incapsula le informazioni del processo computazionale
Method	process.calculation	String		Metodo computazionale utilizzato per il processo
Agent	process.calculation	Object		Software o codice utilizzato per il processo
Name	process.calculation.agent	String	Yes	Nome del software o del codice
Version	process.calculation.agent	String		Versione del software o del codice
Experiment	process	Object		Incapsula le informazioni del processo sperimentale
Method	process.experiment	String		Tecnica sperimentale utilizzata per il processo
Agent	process.experiment	Object		Strumentazione utilizzata prt il processo
Name	process.experiment.agent	String	Yes	Nome dello strumento
Version	process.experiment.agent	String		Versione dello strumento (ad esempio la versione del firmware)
Material	process	Object		Informazioni sul materiale elaborato o analizzato
Formula	process.material	String	Yes	Formula
Elements	process.material	Array		
chemicalComposition	process.material	Array		
Input	process.material	Object		Informazioni sul materiale ricevuto in input dal processo computazionale
Lattice	process.material.input	Object		Informazioni sul reticolo del materiale di input definito dalle due terne

Nome	Genitore	Tipo	Indice	Descrizione
				(a, b, c) e (alpha, beta, gamma)
A	process.material.input.lattice	String		
B	process.material.input.lattice	String		
C	process.material.input.lattice	String		
Alpha	process.material.input.lattice	String		
Beta	process.material.input.lattice	String		
Gamma	process.material.input.lattice	String		
Sites	process.material.input	Array		Lista dei siti dove sono posizionati gli atomi
Species	process.material.input	Array		Lista delle specie atomiche in relazione all'ordine riportato nei siti
Output	process.material	Object		Informazioni sul materiale generato come output dal processo computazionale
Lattice	process.material.output	Object		Informazioni sul reticolo del materiale di output definito dalle due terne (a, b, c) e (alpha, beta, gamma)
A	process.material.output.lattice	String		
B	process.material.output.lattice	String		
C	process.material.output.lattice	String		
Alpha	process.material.output.lattice	String		
Beta	process.material.output.lattice	String		
Gamma	process.material.output.lattice	String		
Sites	process.material.output	Array		Lista dei siti dove sono posizionati gli atomi
Species	process.material.output	Array		Lista delle specie atomiche in relazione all'ordine riportato nei siti
Properties	process	Array		Lista delle proprietà calcolate o non dal processo
Files	root	Array		Lista dei percorsi dei file caricati dall'utente legati al processo
_v	root	String		Versione dello schema dati

Lo schema dati non è da considerarsi definitivo in quanto possono essere previste variazioni, sia aggiunte che modifiche. Per tale motivo è stato introdotto il campo di versioning, chiamato *_v*, che permette di identificare lo schema dei dati utilizzato per ogni dato.

4.5.2 File di elaborazione e di analisi relativi ai processi computazionali e sperimentali

I file derivanti dai processi computazionali e sperimentali possono essere caricati sulla piattaforma IEMAP. I file caricati vengono riportati tramite riferimenti nei documenti BSON nella base dati, memorizzati nel repository nell'infrastruttura ENEA, chiamata *General Parallel File System* (GPFS), e rinominati attraverso una funzione hash.

La funzione hash è uno dei meccanismi di sicurezza basilari del moderno mondo tecnologico, un sistema crittografico in grado di garantire la veridicità (e quindi l'affidabilità) di un messaggio, un sito Web o anche di un file scaricato dalla Rete. La funzione di hashing permette il caricamento dei file garantendo l'unicità degli hash evitando le cosiddette "collisioni", ovvero che il file venga sovrascritto.

5 Conclusioni

Questa sezione sintetizza i risultati complessivi del lavoro e riporta eventuali raccomandazioni per possibili ulteriori sviluppi della ricerca.

6 Riferimenti bibliografici

1. [Andersen2021] Andersen et al, "OPTIMADE, an API for exchanging materials data", Sci. Data 8, 217 (2021) 10.1038/s41597-021-00974-z
2. [JainOng2013] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, "The Materials Project: A materials genome approach to accelerating materials innovation", APL Materials, 2013, 1(1), 011002.doi:10.1063/1.4812323

7 Abbreviazioni ed acronimi

In questa sezione sono state riportate le abbreviazioni e gli acronimi che sono presenti nel documento.

Table 5. Lista delle abbreviazioni e degli acronimi

ID	Acronimo o abbreviazione	Nome	Descrizione
1	ALCF	Argonne Leadership Computing Facility	
2	API	Application Programming Interface	
3	BSON	Binary JSON	
4	CNR	Centro Nazionale delle Ricerche	

ID	Acronimo o abbreviazione	Nome	Descrizione
5	CQ	Competency Questions	
6	CSV	Comma-separated Values	
7	DOI	Digital Object Identifier System	
8	ECMA	European Computer Manufacturers Association	
9	ENEA	Agenzia nazionale per le nuove tecnologie, l'energia e lo sviluppo economico sostenibile	
10	ERD	Entity–Relationship Diagram	
11	EUDAT	European Collaborative Data Infrastructure	
12	FAIR	Findability, Accessibility, Interoperability and Reuse	
13	GB	Giga Byte	
14	GPFS	General Parallel File System	
15	HTTP	Hypertext Transfer Protocol	
16	ICCOM	Istituto di Chimica dei Composti Organometallici	
17	ICMATE	Istituto di Chimica della Materia Condensata e di Tecnologie per l'Energia	
18	ICT	Information and Communication Tecnologies	
19	IEMAP	Italian Energy Materials Acceleration Platform	
20	IIT		
21	ISM	Istituto di Struttura della Materia	
22	ISMN	Istituto per lo Studio dei Materiali Nanostrutturati	
23	IT	Information Technology	

ID	Acronimo o abbreviazione	Nome	Descrizione
24	ITAE	Istituto di Tecnologie Avanzate per l'Energia	
25	JPG	Joint Photographic Experts Group	
26	JSON	JavaScript Object Notation	
27	JSON-LD	JavaScript Object Notation – Linked Data	
28	MB	Mega Byte	
29	NOMAD	Novel Materials Discovery	
30	OLCF	Oak Ridge Leadership Computing Facility	
31	OWL	Ontology Web Language	
32	PDF	Portable Document Format	
33	PID	Persistent Identificator	
34	PNG	Portable Network Graphics	
35	RDF	Resource Description Framework	
36	RSE	Ricerca sul Sistema Energetico	
37	SDSC	San Diego Supercomputer Center	
38	TIFF	Tagged Image File Format	
39	TXT	Text	
40	UC	Use Cases	
41	URN	Uniform Resource Name	
42	W3C	World Wide Web Consortium	
43	XML	EXtensible Markup Language	