DOI: 10.26314/GARR-Conf18-proceedings-04

La prima infrastruttura di data management per le nanoscienze

Rossella Aversa, Stefano Cozzini

CNR-IOM Istituto Officina dei Materiali

Abstract. In questo articolo presentiamo la prima infrastruttura di data management per le nanoscienze. L'infrastruttura, denominata IDRP (Information and Data management Repository Platform), è stata sviluppata dal CNR-IOM (Istituto di Officina dei Materiali) di Trieste all'interno di NFFA-EUROPE (Nanoscience Foundries & Fine Analysis). Questo progetto europeo Horizon 2020, coordinato dal CNR-IOM, coinvolge 20 partner ed ha lo scopo di fornire una infrastruttura di ricerca distribuita per la comunità di nanoscienze. La IDRP è nata con l'obiettivo di gestire ed archiviare in maniera FAIR la grande varietà di dati generati dalla strumentazione NFFA-EUROPE offrendo un accesso standardizzato.

Keywords. Data Repository, Metadati, Nanoscienze, SEM, Reti neurali

Introduzione

I dati scientifici, generati da esperimenti o da simulazioni numeriche, vengono prodotti ad un ritmo sempre maggiore. I concetti di riproducibilità, di interoperabilità, così come di open data, stanno assumendo un ruolo centrale nella ricerca. In effetti molti campi di ricerca dipendono quasi interamente dalla disponibilità e dalla condivisione di dati globali forniti tramite data repository aperti. I principi FAIR (Wilkinson et al. 2016) offrono linee guida per rendere i dati reperibili (Findable), accessibili (Accessible), interoperabili (Interoperable) e riutilizzabili (Reusable). Riconoscendo l'importanza della condivisione e del riutilizzo di dati scientificamente validi, il progetto NFFA-EUROPE (www.nffa.eu) ha implementato la prima piattaforma di data management, denominata IDRP, per la comunità di nanoscienze. L'obiettivo principale della IDRP è quello di automatizzare la raccolta di dati scientifici provenienti dai diversi strumenti presenti nelle strutture europee che partecipano al progetto, identificando i metadati corretti per consentire loro di essere ricercabili in conformità ai principi FAIR. Come primo esempio, ci siamo concentrati sui dati del microscopio a scansione elettronica (SEM). Questo è uno strumento estremamente versatile che viene utilizzato quotidianamente nelle nanoscienze e nelle nanotecnologie per esplorare la struttura dei materiali con risoluzione spaziale fino a 1 nanometro. Gli scienziati che lavorano con tecniche di microscopia elettronica sono particolarmente interessati a strumenti in grado di identificare e riconoscere automaticamente alcune caratteristiche specifiche all'interno delle immagini SEM. A tale scopo abbiamo cercato il modo di applicare algoritmi di classificazione di immagini nel campo delle nanoscienze.

Cagliari, 3-5 ottobre 2018 - ISBN: 978-88-905077-8-6

1. L'architettura

L'architettura del prototipo è mostrata in Fig.1. La IDRP è stata pensata come archivio di metadati relativi ai dati raccolti negli esperimenti fatti dagli utenti NFFA e salvati su storage/repository locali. Questo approccio permette di non interferire con la modalità di acquisizione dati presso le facility. Servizi di gestione dati come quello descritto in questo articolo permettono la registrazione automatica di tali metadati all'interno della IDRP.

L'infrastruttura è stata sviluppata sulla cloud OpenStack del CNR-IOM. I servizi sono configurati come macchine virtuali indipendenti l'una dall'altra. Le istanze più importanti sono il portale di accesso (portal.nffa.eu) e la IDRP per la gestione dei metadati (idrp.nffa.eu). Ci sono poi una istanza dedicata allo storage locale (datashare.iom.cnr.it) ed una per il servizio di classificazione delle immagini (sem-classifier.nffa.eu). Una volta effettuato il login nel portale di accesso, appare il link alla IDRP, dove si registrano e si gestiscono i metadati relativi ai dati che si trovano negli storage locali. Questi ultimi sono molto diversificati nei vari istituti; al CNR-IOM abbiamo adottato un sistema di data management basato su NextCloud. Abbiamo poi iniziato ad aggiungere dei servizi di analisi dati online, in aggiunta a quelli locali. Il primo ad essere stato sviluppato è un tool di classificazione di immagini SEM, che annota le immagini e aggiunge questa annotazione come metadato, oltre a quelli strumentali già presenti.

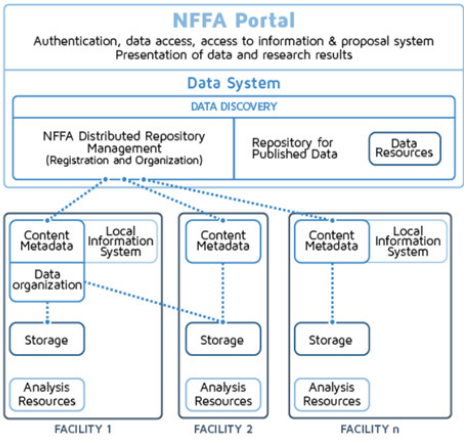


Fig. 1
Architettura della infrastruttura. Dall'alto verso il basso: attraverso il portale NFFA si accede alla IDRP, dove vengono registrati i metadati relativi ai dati prodotti alle varie facility, che vengono salvati negli storage locali.

2. Classificazione di immagini SEM

La classificazione, il riconoscimento di immagini, ed altri algoritmi basati su deep learning, ampiamente applicati in molte aree, non sono ancora sfruttati appieno in nanoscienze. Queste tecniche possono essere uno strumento potente, in particolare quando all'analisi scientifica è associata la gestione dei dati. Una rete neurale per classificare le immagini SEM offre molti vantaggi ai ricercatori in nanoscienze: le annotazioni automatiche, che evitano la necessità di classificare manualmente ciascuna immagine registrata; un database che consenta agli scienziati di trovare una categoria specifica di immagini SEM; la possibilità di adattare la rete neurale per svolgere compiti specifici relativi all'elaborazione dell'immagine, ad esempio quantificare la frazione di nanotubi coerentemente allineati

nelle immagini SEM (Modarres et al. 2017).

2.1 Creazione del dataset

Un set di 18577 immagini SEM è stato manualmente annotato e convalidato da un gruppo di scienziati, i quali hanno convenuto su una suddivisione in 10 categorie, per formare il primo dataset di immagini SEM classificate, disponibile pubblicamente (Aversa et al. 2018). Le categorie sono state stabilite tenendo conto delle caratteristiche visive delle immagini piuttosto che di classi astratte relative a nozioni specifiche. Un'immagine rappresentativa per ciascuna delle categorie scelte è mostrata in Fig.2.

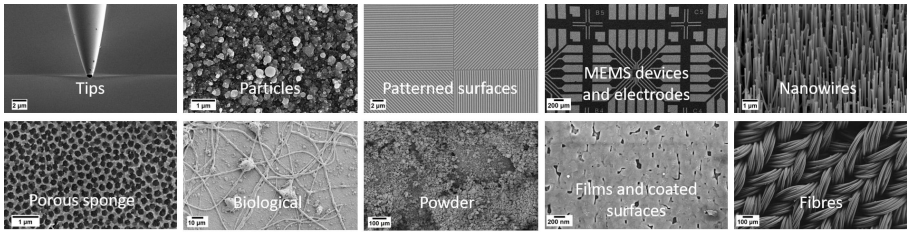


Fig. 2
Categorie del dataset SEM

2.2 Training della rete neurale

Il dataset annotato è stato usato per eseguire il training delle più note reti neurali convoluzionali (Aversa et al. 2019); l'architettura adottata dopo aver confrontato i risultati di accuratezza è Inception-v3 (Szegedy et al. 2015).

2.3 Classificatore online

Il risultato finale è un sito pubblico (sem-classifier.nffa.eu), la cui pagina iniziale è mostrata in Fig.3, dove l'utente carica una o più immagini SEM, la rete neurale fa inferenza e mostra il risultato più probabile. Cliccando su ciascuna immagine si possono vedere i dettagli dell'inferenza; si può anche decidere di selezionare una diversa categoria o di inserirne una nuova oltre alle 10 presenti al momento, che non sono rappresentative di tutto ciò che viene visto con il SEM, ma sono basate sulle immagini a nostra disposizione.

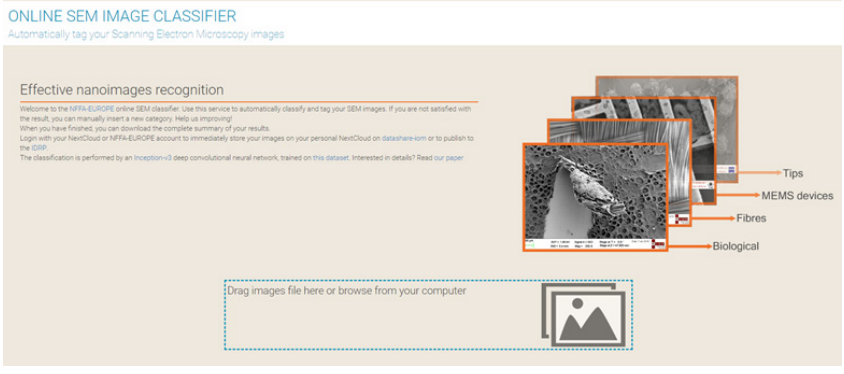


Fig. 3
Pagina iniziale del sito di classificazione di immagini SEM

4. Conclusioni

Abbiamo sviluppato un'infrastruttura per gestire i dati di nanoscienze, ed un sito di analisi online per le immagini SEM, dal quale l'utente può attivare il caricamento nello storage locale. Tutti i metadati vengono automaticamente estratti dalle immagini e possono essere registrati sulla IDRP, dove diventano ricercabili e quindi riutilizzabili dalla comunità scientifica. In futuro verranno aggiunte nuove reti neurali e verrà esteso il dataset, sia in termini di numero di immagini che di categorie.

Riferimenti bibliografici

Aversa R., Modarres M.H., Cozzini S., Ciancio R., Chiusole A. (2018), The first annotated set of scanning electron microscopy images for nanoscience, Scientific Data 5 (180172)

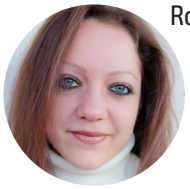
Aversa R., De Nobili C., Cozzini S., (2019) Deep learning for Nanoscience SEM image recognition, in prep.

Modarres M.H., Aversa R., Cozzini S., Ciancio R., Leto A., Brandino G.P. (2017), Neural Network for Nanoscience Scanning Electron Microscope Image Recognition, Scientific Reports 7 (13282)

Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. (2015), Rethinking the Inception Architecture for Computer Vision, arXiv:1512.00567v3

Wilkinson M.D., et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data 3 (160018)

Autori



Rossella Aversa - aversa@iom.cnr.it

Rossella Aversa si è laureata nel 2011 in Astrofisica e Cosmologia all'università di Bologna e nel 2015 ha ottenuto il dottorato in Astrofisica alla SISSA di Trieste. Ha poi frequentato il Master in HPC a Trieste, diplomandosi nel 2016. Da allora è assegnista di ricerca al CNR-IOM di Trieste e lavora per il progetto NFFA-EUROPE.

Stefano Cozzini - cozzini@iom.cnr.it

Stefano Cozzini è tecnologo presso il CNR-IOM con oltre 20 anni di esperienza nelle gestione di infrastrutture IT per HPC e data management. Ha gestito diversi progetti di ricerca a livello internazionale e sta coordinando le attività di data management dei progetti NFFA-EUROPE ed EUSMI. È inoltre coinvolto nel master in HPC e nella laurea magistrale in "Data Science and Scientific Computing", entrambi promossi da diversi enti di Ricerca a Trieste, tra cui CNR-IOM.