Dati, pratiche e infrastrutture oltre i singoli progetti: una traiettoria dal PNRR al futuro dell'intelligenza artificiale

di Arnaud CÃ©ol

10-06-2026

Quando l'output non sono i modelli, ma i dati

Nel dibattito pubblico su intelligenza artificiale e supercalcolo, il ruolo dei dati è ormai ampiamente riconosciuto. Dati disponibili e facilmente utilizzabili costituiscono il presupposto essenziale per lo sviluppo di modelli, applicazioni e politiche basate sull'intelligenza artificiale. Accanto a questa consapevolezza, tuttavia, rimane spesso meno esplicitata una distinzione cruciale: quella tra i dati intesi come semplice risorsa tecnica e i dati concepiti come infrastruttura, cioè come componente strutturale di sistemi complessi e duraturi.

L'esperienza maturata all'interno del Centro Nazionale di Ricerca in HPC, Big Data e Quantum Computing offre un osservatorio privilegiato su questo passaggio. Nel corso del PNRR, il Centro Nazionale non si è limitato alla produzione di risultati scientifici o di dimostratori tecnologici, ma ha investito nella costruzione di infrastrutture e pratiche comuni progettate per estendere il proprio valore oltre la durata dei singoli progetti. Dataset strutturati e documentati, interoperabili e riutilizzabili nel tempo, sono stati integrati in archivi e servizi concepiti non come semplici depositi, ma come componenti strutturali a supporto di nuovi ecosistemi di ricerca e innovazione.

In questo contesto, i dati assumono una fisionomia più ampia. Non sono più soltanto input per modelli di intelligenza artificiale, ma diventano asset organizzati, governati e integrati all'interno di architetture tecnologiche e istituzionali complesse. Comprendere questa evoluzione è essenziale per interpretare correttamente sia l'esperienza del Centro Nazionale nel contesto del PNRR, sia lo sviluppo di iniziative successive come IT4LIA, e più in generale le sfide che attendono il sistema della ricerca e dell'innovazione nel passaggio verso architetture distribuite Edge-Cloud-HPC.

Il patrimonio: il Centro Nazionale PNRR e i dati come outcome

Il PNRR ha rappresentato per il sistema della ricerca italiana un'occasione senza precedenti, non soltanto per l'entità delle risorse mobilitate, ma soprattutto per l'impostazione strategica adottata. Con la nascita del Centro Nazionale di Ricerca in HPC, Big Data e Quantum Computing, l'obiettivo non era limitato al finanziamento di singoli progetti di eccellenza, ma orientato alla costruzione di capacità di sistema, intese come infrastrutture, competenze e pratiche in grado di produrre effetti durevoli oltre l'orizzonte temporale dei finanziamenti.

In questo quadro, i dati hanno assunto sin dall'inizio un ruolo centrale. Non semplicemente come materia prima delle attività di ricerca, ma come outcome strutturale del lavoro del Centro Nazionale e come componente portante dell'ecosistema nel suo complesso. Progetti e spoke di ogni dominio applicativo hanno contribuito a generare dati concepiti non per rimanere confinati in contesti locali, ma per confluire in un patrimonio più ampio, coordinato e progressivamente armonizzato.

Il modello Hub & Spoke ha reso possibile questa impostazione. La produzione dei dati è rimasta distribuita e aderente alle specificità disciplinari, ma è stata accompagnata da uno sforzo condiviso volto a definire regole, standard e pratiche comuni di gestione. In questo contesto, il valore non risiedeva prioritariamente nella quantità dei dati prodotti, quanto nella capacità collettiva di renderli utilizzabili nel tempo, anche da comunità e soggetti diversi da quelli che li avevano originariamente generati.

Questo approccio si è tradotto in attività operative in numerosi ambiti applicativi. I dati prodotti e

utilizzati comprendono dataset meteoclimatici e ambientali, incluse osservazioni satellitari e serie temporali ad alta risoluzione; dati per il monitoraggio del territorio e dei rischi naturali; dati sanitari e biomedicali, soggetti a vincoli normativi stringenti; dati industriali e ingegneristici generati in collaborazione con grandi imprese; fino agli output di simulazioni numeriche, modelli HPC, digital twin e applicazioni di machine learning. La gestione di questa varietà è stata resa possibile dalla complementarità delle competenze dei diversi nodi e datacenter coinvolti, ciascuno con ruoli distinti in relazione a specifici domini, volumi di dati e tipologie di processamento.
Da queste esperienze è emersa con chiarezza l'eterogeneità strutturale dei dati e dei processi che li generano. I dataset del Centro Nazionale differiscono per tipologia, formato, scala e grado di maturità: coesistono dati grezzi e dati fortemente elaborati, risorse di pochi megabyte e archivi dell'ordine del petabyte, dataset statici e flussi quasi in tempo reale. A questa varietà corrispondono pipeline di trattamento complesse, che comprendono ingestione multisorgente, simulazioni ad alte prestazioni, integrazione di modelli HPC, data fusion e tecniche avanzate di analisi e apprendimento automatico.
In tale contesto, la gestione dei dati ha riguardato non solo gli output finali, ma anche gli input, le trasformazioni intermedie e le condizioni di utilizzo dei modelli. Aspetti quali tracciabilità, versioning e riproducibilità sono diventati centrali non per ridurre la complessità, ma per renderla governabile, mantenendo coerenza tra discipline scientifiche, comunità di utenti e bisogni operativi profondamente diversi.
È qui che si colloca una delle lezioni più rilevanti del PNRR: le infrastrutture dati non emergono per accumulazione spontanea, ma richiedono progettazione, investimenti continuativi e modelli di governance espliciti. Archivi federati, cataloghi, pipeline di ingestione, servizi di metadatazione e strumenti di tracciabilità non hanno costituito semplici sottoprodotti tecnici, ma risultati strategici del lavoro del Centro Nazionale.
Parte integrante di questo percorso è stata la gestione dei diversi regimi di accesso ai dati. Il Centro Nazionale ha operato fin dall'inizio in un contesto intrinsecamente eterogeneo: accanto a dati pensati per essere aperti e riutilizzabili, sono presenti dati sensibili, in particolare in ambito sanitario, e dati di origine privata o industriale, utilizzabili esclusivamente entro perimetri progettuali autorizzati. La capacità costruita non è stata quella di imporre un modello unico di apertura, ma di rendere queste differenze esplicite, trasparenti e governabili.
Anche l'accesso limitato a dati aziendali non ha rappresentato un vincolo, bensì un'opportunità. Il lavoro su dati reali, non sempre disponibili in contesti puramente accademici, ha consentito ai progetti del Centro Nazionale di confrontarsi con condizioni operative concrete e di sviluppare soluzioni più robuste. È anche in questa capacità di bilanciare apertura, protezione e responsabilità che i dati del Centro Nazionale assumono pienamente il ruolo di infrastruttura, non solo per i contenuti informativi che veicolano, ma per l'insieme di regole, pratiche e relazioni fiduciarie che ne rendono possibile l'utilizzo nel tempo.
Guardando a posteriori, uno dei risultati più rilevanti del periodo del PNRR non risiede in singoli esiti scientifici o tecnologici, ma nell'affermazione dei dati come infrastruttura, attraverso pratiche, competenze e modalità di gestione che ne rendono possibile l'uso affidabile su larga scala. Questo patrimonio di metodi ed esperienze non esaurisce gli sviluppi successivi, ma costituisce uno dei contributi che alimentano scenari più ampi e articolati.
Nel contesto di IT4LIA, il lavoro sui dati si colloca all'interno di un quadro europeo in evoluzione, orientato a renderli progressivamente utilizzabili da ricercatori, pubblica amministrazioni e industria nei workflow dell'intelligenza artificiale. La continuità che si osserva non è quindi istituzionale, ma tematica: riguarda il ruolo centrale dei dati e delle loro modalità di gestione nel passaggio dal PNRR al presente e agli sviluppi futuri dell'ecosistema nazionale ed europeo dell'AI.

Il presente: IT4LIA e la maturità dei dati

Se il Centro Nazionale PNRR ha rappresentato la fase di costruzione delle infrastrutture dati e delle pratiche comuni di gestione, IT4LIA ne segna l'ingresso in una fase pienamente operativa. È il momento in cui le capacità sviluppate vengono applicate e consolidate su casi d'uso reali, all'interno di un ecosistema europeo competitivo e regolato, e in cui emerge con chiarezza un concetto chiave: quello di AI Data Readiness.

Nel contesto di IT4LIA, non è più sufficiente che i dati siano semplicemente disponibili o correttamente gestiti. Per poter sostenere in modo affidabile lo sviluppo e l'impiego di sistemi di intelligenza artificiale, i dati devono essere pronti all'uso lungo l'intero ciclo di vita dei modelli e dei servizi AI. La readiness riguarda innanzitutto la comprensibilità del dato, garantita da metadata coerenti e da informazioni di contesto che ne chiariscano origine, significato e limiti di utilizzo. Include poi la tracciabilità, intesa come capacità di ricostruire le trasformazioni subite dal dato prima di alimentare un modello, evitando opacità che renderebbero difficile l'interpretazione dei risultati. Un ulteriore elemento è la coerenza semantica, che consente di integrare dati provenienti da contesti eterogenei senza ambiguità, riducendo i tempi di avvio delle attività e favorendo il riuso.

In termini operativi, questo cambiamento incide direttamente sui diversi tipi di utenti dell'ecosistema. Per i ricercatori, comporta una riduzione dei tempi di avvio dei progetti, grazie alla disponibilità di dati già comprensibili e contestualizzati. Per le imprese, significa poter utilizzare i dati all'interno di pipeline affidabili e ripetibili, compatibili con contesti regolati. Per gli attori pubblici, infine, implica la possibilità di basare decisioni e servizi su risultati dell'intelligenza artificiale più trasparenti, tracciabili e quindi più facilmente valutabili e adottabili.

Questa impostazione completa e rende pienamente operative le attività avviate durante il PNRR attraverso l'evoluzione del Data Management Plan, che in IT4LIA assume un ruolo esplicitamente funzionale. Il DMP guida in modo concreto le modalità con cui i dati vengono ingeriti, monitorati, trasformati e resi disponibili nei diversi ambienti dell'AI Factory, contribuendo a renderli non solo corretti dal punto di vista formale, ma anche stabili, riutilizzabili e mantenibili nel tempo.

Un ambito particolarmente rappresentativo di questo approccio è la data ingestion, concepita come un processo continuo e monitorabile. I dati vengono identificati, descritti e arricchiti con metadata, per poi essere preparati all'utilizzo all'interno di workflow di intelligenza artificiale. Attraverso questo percorso, evolvono progressivamente in data products, risorse scopribili e utilizzabili da comunità diverse senza la necessità di ricostruire ogni volta il contesto originario.

Accanto allo sviluppo dei servizi, ICSC contribuisce anche direttamente sul piano dei dati. Quando possibile, i dataset prodotti o curati nell'ambito del Centro Nazionale vengono utilizzati per completare e arricchire il catalogo di IT4LIA, assicurando continuità con il patrimonio costruito durante il PNRR. In questo modo, l'esperienza maturata all'interno di ICSC continua a generare valore all'interno dell'AI Factory.

Un ulteriore elemento distintivo di IT4LIA è rappresentato dallo sviluppo dei Data Labs, spazi operativi in cui dati, servizi e competenze convergono per supportare l'uso dell'intelligenza artificiale in contesti applicativi concreti. In prospettiva, questi ambienti si inseriscono nel quadro dei EU AI Data Labs come strumenti per favorire la sperimentazione, l'integrazione e il riuso dei dati all'interno di ecosistemi tematici. IT4LIA opera così come nodo di una rete europea di laboratori dedicati ai dati per l'intelligenza artificiale, contribuendo alla costruzione di capacità comuni.

In prospettiva, man mano che i servizi dati si stabilizzano, la loro presenza tende a diventare meno visibile, non perché meno rilevante, ma perché maggiormente integrata nei processi operativi. Quando discovery, ingestion, preparazione, gestione e orchestrazione sono allineate all'interno di un'unica architettura operativa, l'attenzione può spostarsi dall'infrastruttura all'innovazione e all'impatto. È proprio questa maturità operativa, fondata sulla AI Data Readiness, a configurare IT4LIA come passaggio chiave tra l'eredità del PNRR e gli scenari futuri dell'intelligenza artificiale.

Il futuro: Edge-Cloud-HPC, ECHO-TWIN ed etica dei dati

Se IT4LIA rappresenta il presente della maturità dei dati, il futuro che si delinea è caratterizzato da una trasformazione più profonda: il passaggio verso un continuum Edge-Cloud-HPC. In questo scenario, i dati non sono più generati prevalentemente in contesti centralizzati e stabili, ma nascono in prossimità delle fonti fisiche che descrivono, come sensori, infrastrutture, città e territori. Il dato diventa così parte integrante dei sistemi osservati e deve poter essere utilizzato con tempi di latenza ridotti per supportare analisi, simulazioni e processi decisionali.

Questa evoluzione incide in modo diretto sulle modalità di gestione dei dati. Non sempre è possibile, né opportuno, trasferire grandi volumi informativi verso centri di calcolo remoti. In numerosi contesti, i dati devono essere selezionati, filtrati, arricchiti e aggregati lungo il percorso che li conduce all'utilizzo. Il valore informativo non risiede più esclusivamente nella conservazione del dato, ma nella capacità di governarne il ciclo di vita, definendo in modo consapevole dove, quando e in quale forma esso viene trasformato e impiegato.

Nel continuum Edge-Cloud-HPC, la gestione dei dati assume quindi una valenza architetturale e di responsabilità. La maggiore prossimità del dato ai contesti operativi rende infatti più diretto il legame tra informazione e decisione, accrescendo l'importanza della qualità, della trasparenza e della ricostruibilità dei processi che portano da un'informazione grezza a un risultato utilizzabile.

È in questo quadro che iniziative come ECHO-TWIN assumono un ruolo specifico. Più che come un singolo progetto, ECHO-TWIN può essere interpretato come un laboratorio sperimentale in cui vengono sviluppate e testate architetture dei dati destinate a supportare digital twin distribuiti, confrontati con flussi informativi eterogenei, dinamici e in continuo aggiornamento. I gemelli digitali sviluppati nell'ambito dell'iniziativa richiedono infatti l'integrazione coerente di dati provenienti da sorgenti diverse, mantenendo significato e consistenza lungo l'intero processo di utilizzo.

In ECHO-TWIN, i dati non sono trattati come oggetti statici, ma come flussi informativi strutturati, che attraversano fasi successive di ingestione, validazione, arricchimento semantico e integrazione. Per gestire tale complessità, il progetto opera su data space federati, in cui metadata, ontologie e modelli condivisi consentono di mantenere interoperabilità anche in presenza di sorgenti eterogenee e aggiornamenti continui. Attraverso questo approccio, ECHO-TWIN affronta in modo operativo i temi della qualità, dell'affidabilità e della ricostruibilità dell'informazione nei digital twin. Il dato assume così il ruolo di infrastruttura a pieno titolo, che deve essere progettata, monitorata e governata con lo stesso livello di attenzione riservato agli altri componenti del sistema.

Man mano che i dati si avvicinano alle persone e ai territori emergono con maggiore evidenza le dimensioni etiche e sociali della loro gestione. Nel continuum Edge-Cloud-HPC, le decisioni basate sui dati non sono più astratte o distanti, ma possono incidere direttamente su ambiti quali mobilità, ambiente, salute e sicurezza. Ciò rende insufficiente una visione esclusivamente tecnica della governance dei dati: privacy, protezione delle informazioni sensibili, mitigazione dei bias e trasparenza diventano requisiti strutturali, da integrare fin dalle fasi di progettazione delle infrastrutture e dei servizi.

Le esperienze maturate nel Centro Nazionale e consolidate attraverso ECHO-TWIN evidenziano l'importanza di strumenti di governance dedicati, quali linee guida condivise anche sociali ed etiche, in grado di affiancare lo sviluppo tecnologico con una riflessione continua sugli impatti sociali. In questo contesto, l'etica non si configura come un vincolo esterno, ma come una condizione di sostenibilità per l'utilizzo dei dati in sistemi complessi e distribuiti.

Conclusione: dai dati come mezzo ai dati come capacità

Guardando al percorso nel suo insieme, dal PNRR alle iniziative più recenti, emerge come uno dei risultati più duraturi non risieda in singoli esiti scientifici o tecnologici, ma nella progressiva

affermazione dei dati come infrastruttura. Un'infrastruttura fatta di pratiche, regole e strumenti che rendono possibile l'uso affidabile dei dati in contesti complessi, distribuiti e ad alto impatto.
Questo percorso si inserisce in modo naturale nel quadro europeo dell'European Open Science Cloud (EOSC), di cui ICSC coordina il Nodo Nazionale italiano. Il lavoro svolto dal Centro Nazionale sulla gestione dei dati, sui regimi di accesso e sulla loro interoperabilità non rappresenta quindi un'esperienza isolata, ma contribuisce alla costruzione di un'infrastruttura federata europea, pensata per rendere i dati della ricerca scopribili, accessibili e riutilizzabili nel rispetto dei contesti disciplinari, tecnologici e normativi.
In questa prospettiva, l'esperienza maturata durante il PNRR prepara il terreno a una continuità che va oltre i singoli programmi di finanziamento. I dati non sono soltanto un output del Centro Nazionale, ma diventano un ponte strutturale tra l'ecosistema italiano della ricerca e quello europeo, accompagnando il passaggio dalla produzione dei dati alla loro integrazione in spazi condivisi e operativi. È in questo ruolo di connessione e garanzia che si rafforza la funzione di ICSC come attore di coordinamento, capace di assicurare coerenza, sostenibilità e affidabilità all'uso dei dati nel presente e nel futuro dell'intelligenza artificiale e del supercalcolo in Europa.