

L'Italian Arctic Data Center (IADC)

L'INFRASTRUTTURA INFORMATICA DISTRIBUITA PER LA GESTIONE E CONDIVISIONE
DEI DATI ARTICI E DEL PRA

L'INFRASTRUTTURA COME "SISTEMA DI SISTEMI"

L'architettura dell'infrastruttura informatica si basa sul concetto di "Sistema di Sistemi" (System-of-Systems - SoS): un insieme di sistemi interconnessi tra di loro per mezzo di servizi di mediazione e adattamento, che sono comunemente chiamati "brokering services" (Nativi et al. 2012), implementati da una componente software centrale –di seguito chiamato il "common node" (Nativi et al. 2015).

Quest'architettura distribuita, consente di mettere in comune le risorse e le capacità dei sistemi esistenti (e di quelli che verranno in un futuro prossimo) per dar vita ad un sistema più complesso che offre funzionalità e prestazioni che non sono la semplice somma dei sistemi esistenti.

L'approccio architetturale proposto (di tipo "bottom-up") garantisce l'autonomia gestionale e le diversità tecnologiche (interfacce standard) che caratterizzano i singoli sistemi informativi/dati, e al tempo stesso la piena interoperabilità delle loro risorse.

L'*interoperabilità* consente a banche dati, servizi Web e sistemi informativi di scambiare contenuti attraverso interfacce standard di comunicazione condivisi, senza che i gestori debbano affrontare particolari costi/oneri relativi alla conversione dei dati e delle informazioni (IEEE 1990).

Infatti, i "brokering service" forniti da un nodo centrale, "common node", mediano tra le differenti interfacce standard di comunicazione implementate dai sistemi dati e informativi esistenti, che quindi non devono modificare né la loro tecnologia né gli standard comunicativi adottati.

Nel mentre il "common node" implementerà i "brokering services" per il SoS, le altre componenti del sistema verranno sviluppate in modo da realizzare e implementare una infrastruttura del tipo **SaaS (Software as a Service)** per l'archiviazione, analisi e visualizzazione dei dati. L'architettura verrà sviluppata in modo coordinato e condiviso e tale da poter erogare servizi di vario livello di complessità, e che siano accessibili attraverso differenti interfacce: Web API (machine-to-machine), Desktop API or tool/software, Web GUI, Desktop GUI, VRE (Virtual research environments).

LE FUNZIONI

Italian Arctic Data Center (IADC)

Il sistema digitale distribuito per la gestione dei dati raccolti dalla comunità nazionale in Artico costituirà l'Italian Arctic Data Center (IADC). IADC è implementato e gestito dal CNR, che ha il compito di implementare il PRA, e dalle altre Istituzioni chiamate a predisporre il Programma di Ricerche in Artico: ENEA, INGV, OGS.

Il sistema è finalizzato a:

- Garantire la **stabile disponibilità** dei dati acquisiti in Artico, e in particolare dei dati presso le stazioni osservative e dei progetti supportati dal PRA, rendendoli sempre fruibili per la ricerca e la divulgazione scientifica;
- Agevolare la **diffusione dei dati** prodotti in artico verso la comunità scientifica nazionale e internazionale;
- Garantire la **qualità e la tracciabilità dei dati** raccolti, con particolare attenzione ai dati acquisiti a Dirigibile Italia e Thule e nell'ambito di progetti finanziati dal PRA.
- Accrescere la **fruibilità e il valore aggiunto** per gli utenti, attraverso lo sviluppo di data products e servizi a livello base (QA/QC, downloading, visualizzazione) e alto (confronto, integrazione/combinazione, processamento).

Le funzioni principali d'interoperabilità di IADC

IADC è costruito utilizzando i sistemi già esistenti presso i "partner" del PRA. Esso deve essere in grado di:

- Costituire un punto d'accesso comune per la scoperta, l'accesso e l'utilizzo dei dati e dei servizi.
- Integrare dati/informazioni e strumenti per la loro elaborazione indipendentemente dal sistema che li gestisce –e che farà parte dell'infrastruttura distribuita;
 - *integrare piattaforme e strumenti software eterogenei;*
 - *supportare l'evoluzione tecnologica nel settore dell'Information Technology (IT);*
- Essere flessibile così da includere i nuovi sistemi dati/informativi che vorranno contribuire in futuro al PRA e alla ricerca artica nazionale;
 - *integrare nuove capacità;*
- Garantire l'interoperabilità sia con i programmi e le infrastrutture Nazionali che con quelli Internazionali;
 - *utilizzare standard "open" sia internazionali che de-facto in una Comunità scientifica;*
- Garantire l'autonomia dei sistemi dati/informativi che contribuiscono a IADC.
- Adottare i principi dell'architettura di SoS –con particolare riferimento ai principi di FAIR, e alle "best practices" definite in ambito GEOSS e in ambito POLAR DATA FORUM.

L'Architettura e i Nodi funzionali del Sistema

L'infrastruttura digitale da realizzare è SoS che prevede l'attivazione di vari componenti (di seguito chiamati nodi funzionali del sistema) distribuite per la gestione e la condivisione dei dati e delle informazioni. Ogni nodo è costituito da un sistema dati e/o informativo esistente –o in fase di realizzazione presso una delle organizzazioni che contribuiscono al PRA e alla infrastruttura IADC.

I nodi funzionali sono interconnessi e federati tra loro (vedi Fig.1) per mezzo di interfacce e modelli di dati/metadati standard –internazionali o largamente supportate dalle comunità scientifiche.

Facendo riferimento alla Fig. 1 sono individuati per l'architettura IADC di base da un punto di vista "funzionale" 2 tipi di nodi così denominati:

- Nodo Centrale Comune (Common Node);
- Nodo di I livello;

Tale struttura differisce da quella del NADC sia perché tende a non considerare la possibilità di dover gestire nodi di secondo livello, sia perché prevede che il sistema che presso il CNR gestisce le funzionalità del "common node" abbia anche la capacità di nodo I livello. Lo sviluppare l'architettura senza considerare i nodi di II livello, ha come conseguenza che i nodi di I livello debbano possedere maggiori capacità di acquisizione e gestione dei dati rispetto agli equivalenti nodi del NADC, e porre più attenzione allo sviluppo in comune di quei servizi necessari ad aumentare la fruizione e il valore aggiunto per l'utente. I nodi di I livello implementeranno una infrastruttura del tipo SaaS (Software as a Service) facendo leva quando possibile/necessario sul paradigma del cloud computing.

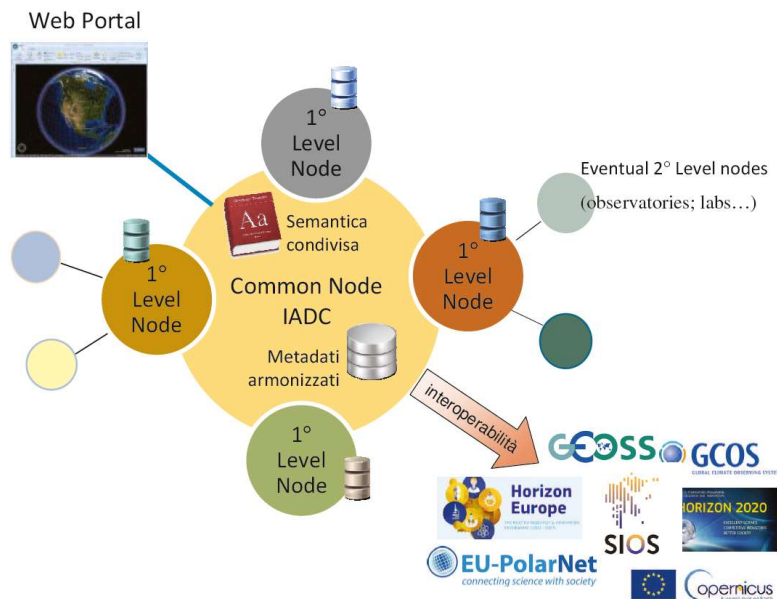


Figura 1. Schema dei nodi funzionali dell'Italian Arctic Data Center (IADC)

Funzionalità del Nodo Centrale Comune (Common Node)

Il Nodo Centrale Comune deve (vedi Fig. 2):

1. Interconnettere tutti i Nodi di I livello implementando gli adattatori (*adapters*) necessari per interfacciarsi con loro.
2. Eseguire un "harvest" regolare dei metadati pubblicati dai Nodi di I livello, armonizzandoli secondo uno schema comune e implementando i mediatori (*mediators*) necessari per mappare gli schemi dei metadati pubblicati dai Nodi di I livello. Tra i metadati veicolati dal Common Node devono anche esserci quelli utili per l'accesso ai dati che risiedono presso i Nodi di I livello.
3. Consentire l'accesso remoto ai dati condivisi dai Nodi di I livello, per mezzo di un reindirizzamento sul Web. Questo servizio può fornire funzionalità aggiuntive –ove possibile se richiesto dall'utenza.
4. Mantenere un vocabolario comune per una semantica condivisa per consentire l'armonizzazione dei metadati pubblicati dai Nodi di I livello e facilitare la scoperta delle risorse di IADC.
5. Supportare e pubblicare il portale Web dell'Infrastruttura per consentire la ricerca, l'accesso e l'utilizzo delle risorse condivise dai Nodi di I livello (dati e servizi). Il target sono gli utenti finali.

6. Supportare la scoperta e l'accesso di strumenti software e prodotti condivisi dalla Comunità Artica Nazionale. Il target sono gli utenti intermedi.
7. Garantire l'interoperabilità verso i principali programmi e le iniziative Europee e internazionali nel settore della ricerca polare.
8. Garantire un adeguato livello di servizio (Quality-of-Service: QoS).

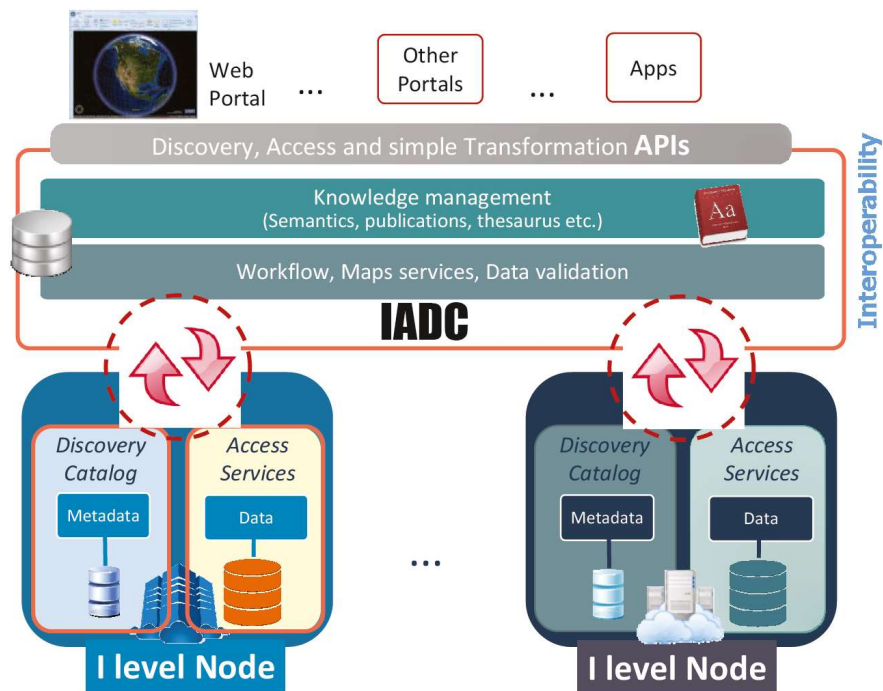


Figura 2. Interfacce d'interoperabilità tra il Common Node e i Nodi di I Livello

Nodi di I livello

I nodi di primo livello sono sistemi eterogenei che rappresentano Enti e/o Organizzazioni diverse. Queste contribuiscono a IADC mettendo a disposizione dati, informazioni e conoscenze garantendone la loro cura e preservazione. Proprio al fine di garantire la loro sostenibilità ed evoluzione nel tempo, la loro autonomia e le loro diversità tecnologiche devono essere garantite dall'infrastruttura complessiva.

I nodi di I livello devono implementare le seguenti funzionalità (vedi Fig. 2):

1. Pubblicare su Internet uno o più servizi di discovery delle risorse (dati, informazioni, servizi, tools, etc.) condivise;
2. Pubblicare su Internet uno o più servizi di accesso alle risorse (dati, informazioni, servizi, tools, etc.) condivise;
3. Condividere una Data Policy comune;
4. Garantire il livello di qualità richiesto (i.e. completezza dei metadati, formato/i standard dei dati, etc.) alle risorse condivise. Tale livello sarà fissato in modo condiviso in sede di governance di IADC.
5. Garantire l'aggiornamento, la manutenzione e la preservazione dei servizi e delle risorse condivise all'interno di IADC.

6. Partecipare e contribuire allo sviluppo di servizi che abbiano lo scopo di aumentare la fruibilità delle risorse (dati in particolare) da parte degli utenti finali, e la possibilità di ottenere nuova informazioni e valore aggiunto dagli stessi.

LO SVILUPPO

Il Common Node

Il nodo centrale comune è sviluppato dal CNR che ne cura la manutenzione evolutiva e il potenziamento in termini di hw e sw.

Verranno sviluppati e configurati i moduli software necessari per realizzare il SoS, in linea con gli attuali scenari e programmi Europei ed Internazionali (vedi open access, open science, brokering approach, etc.).

Verrà definita una Quality-of-Service (QoS) dei servizi offerti dal Common Node verso l'esterno –sia per gli Utenti intermedi che finali. Tale QoS sarà definita in piena sinergia e condivisione con i Nodi di I livello, così da assicurare che le QoS per le due tipologie di nodi siano ben integrate.

Verranno implementate APIs di livello alto (leggi semplici) per garantire ai suoi Utenti intermedi (es. sviluppatori di applicazioni e/o servizi) l'interoperabilità. Tenendo conto delle funzioni che IADC si prefigge (cfr. pag. 6), utilizzatori intermedi privilegiati saranno i gestori dei Nodi di Primo livello.

Verrà sviluppato e configurato il Web Data Portal (Portale generale IADC).

Verrà garantita la gestione e la manutenzione evolutiva del Common Node.

Il livello delle funzioni e dei servizi di IADC sarà garantito dalla sua governance, sviluppata in modo adeguato alla dimensione del sistema ed utenti (cfr. più avanti). Nel caso risultasse necessario, dei "Interoperability Agreement" ad hoc saranno stipulati dal Common Node con alcuni dei Nodi di I livello.

Verrà preparato e fornito (on-line) il materiale descrittivo circa le funzionalità del Common Node. Laddove necessario si provvederà a organizzare dedicate sezioni di formazione per sviluppatori e gestori dei nodi di I livello.

Nella fase operativa dell'Infrastruttura, il CNR, anche avvalendosi di ISP, individuerà un Amministratore per IADC e il portale dati.

Lo sviluppo del Portale prevede un approccio in tre fasi:

- (a) Il rilascio di una **prima infrastruttura prototipale** fornendo le funzionalità di base quali discovery, accesso, e visualizzazione. In questa fase verrà testato il funzionamento dell'infrastruttura e la fruibilità degli archivi.
- (b) Il rilascio di una **versione consolidata dell'infrastruttura**, realizzando funzionalità avanzate – quali la discovery semantica, il raffinamento della discovery e trasformazioni per aumentare l'usabilità delle risorse.
- (c) Il rilascio dell'**infrastruttura operativa finale**, con l'interconnessione di tutti i Nodi di I Livello previsti e la piena operatività di tutte le funzioni avanzate.

Nodi di I livello

Ogni Organizzazione/Ente individuato per realizzare un Nodo di I Livello provvederà a disegnare, sviluppare e configurare i moduli software che costituiranno la piattaforma digitale per realizzare il Sistema Informativo che si deve interfacciare con il Common Node.

Provvederà, altresì, allo sviluppo delle procedure e tools (es. Metadata editor, procedure di metadattazione automatica, etc.) per garantire la corretta metadattazione dei futuri dati acquisiti dall'Ente che gestisce il Nodo.

Verrà, inoltre, effettuata una ricognizione dei dataset esistenti da pubblicare e condividere attraverso il Nodo di I Livello.

L'Organizzazione/Ente garantisce la qualità dei prodotti pubblicati e condivisi tramite il Common Node, nonché la gestione e la manutenzione evolutiva del Nodo.

In una logica di SoS, il *Nodo I Livello* si interfaccia con il *Common Node* al fine di condurre insieme i test necessari per l'integrazione e l'interoperabilità.

I nodi di I livello realizzeranno ed implementeranno una infrastruttura del tipo SaaS (Software as a Service) rivolta ad assicurare l'archiviazione dei dati, così come la piena fruibilità in una logica open e produzione di nuova informazione (added-value) attraverso servizi di vario livello (visualizzazione, confronto, comparazione, elaborazioni di più dataset e data products). L'adozione di tecnologie cloud computing verrà laddove possibile privilegiata. Così come la possibilità di accedere e utilizzare diverse interfacce, in particolare VRE.

Nella fase operativa dell'infrastruttura, ogni Organizzazione/Ente che amministra un Nodo di I Livello provvederà a nominare un Amministratore del Nodo che si interfacerà e collaborerà con il personale CNR che ha il compito di mantenere ed avanzare il Common Node.

Costituzione dei Nodi di I Livello

La classificazione dei nodi di I livello può essere fatta per area tematica o per appartenenza ad Organizzazione/Ente. Ognuna delle due opzioni presenta vantaggi e svantaggi come sinteticamente riportato di seguito:

Aggregazione per aree tematiche:

- *vantaggi:* il numero dei nodi è "definito" a priori perché coincidente con le aree tematiche stabilite.
- *svantaggi:* necessita l'aggregazione di strutture e soggetti diversi afferenti ad una stessa area tematica.

Aggregazione per Ente/Organizzazione

- *vantaggi:* ogni Organizzazione/Ente si organizza sulla base della propria infrastruttura digitale e tematica.
- *svantaggi:* il numero dei nodi non può essere stabilito a priori. Maggiore complessità per armonizzare le possibili sovrapposizioni.

Laddove necessario un Nodo di I Livello stabilisce un “interoperability agreement” con il Common Node che prevede aspetti Organizzativi (inclusa la Data Policy) e Tecnologici. Il CNR avrà il compito di preparare e far circolare una prima versione di tale “interoperability agreement”. La governance di IADC avrà il compito di rivedere e raffinare tale draft adottando una versione finale.

In una prima fase vengono individuati i seguenti quattro nodi di I livello: CNR, INGV, OGS, ENEA. Il sistema potrà in seguito allargarsi aggiungendo altri nodi, ogni qualvolta Organizzazioni/Enti fortemente coinvolti nella ricerca polare potranno farne richiesta e dimostrare di poter garantire il Quality of System (QoS) che è stato definito in ambito IADC.

In questo senso lo sviluppo futuro dell'architettura di IADC nel medio periodo si distingue fortemente dall'analogo sviluppo del NADC. In questo non è al momento previsto un ampliamento del numero dei nodi di I livello (attualmente 5), mentre esiste una geometria variabile di Nodi di II livello. Ciò è legato alla profonda differenza nel panorama della ricerca PNRA e della ricerca Artica.

RIFERIMENTI

European Commission, 2016, “Open Science (Open Access)”, URL:

<https://ec.europa.eu/programmes/horizon2020/en/h2020-section/open-science-open-access>

FAIR principles - <https://www.go-fair.org/fair-principles/>

INSPIRE, 2013, “Guidelines for the encoding of spatial data” available at:

http://inspire.ec.europa.eu/documents/Data_Specifications/D2.7_v3.3rc3.pdf

Nativi, S., Craglia, M., Pearlman, J., 2012, The Brokering Approach for Multidisciplinary Interoperability: A Position Paper. International Journal of Spatial Data Infrastructures Research, Vol.7, 1-15. Available at: <http://ijmdir.irc.ec.europa.eu/index.php/ijmdir/article/view/281/319>

Nativi, S., Mazzetti, P., Santoro, M., Papeschi, F., Craglia, M., Ochiai, O., 2015, Big Data challenges in building the Global Earth Observation System of Systems, Environmental Modelling & Software, Volume 68, June 2015, Pages 1-26.

Ottavianelli Giuseppe et al., 2015, “Update on the Copernicus Coordinated data Quality Control (CQC) activities”, ESA Copernicus Space Component Data Access Workshop, Frascati (Rome). Available at:

http://congrexprojects.com/custom/15M34/Ppt%20Copernicus%20only%20PDF/01%20Day-1/Session%201/05_CSCDA_Workshop20150420_CQC_v06.pdf

Rosanzki, N., and Woods, E., 2012, “Software Systems Architecture 2nd Edition”, Addison-Wesley, Sep 2012, 678 pages.

Wilkinson, Mark, D., et. al., 2016, The FAIR Guiding Principles for scientific data management and stewardship, Sci Data 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>