



Designing a data and computing infrastructure to support and disseminate research activities of CNR-IRPI

Paolo Allasia, Massimiliano Alvioli, Vinicio Balducci, Luca Ciabatta, Nunzio Luciano Fazio, Daniele Giordan, Giovanni Gullà, Diego Guenzi, Ivan Marchesini, Christian Massari, Mauro Rossi,

Alessandro Sarretta

CNR-IRPI

e-mail: alessandro.sarretta@irpi.cnr.it

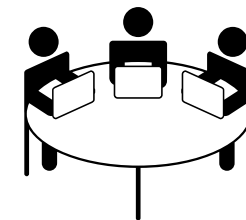
Obiettivo della presentazione

- Presentare il **tavolo di lavoro** sull'infrastruttura, delinearne motivazioni e obiettivi, descrivere i temi principali in discussione
- Descrivere le possibili/auspicabili **funzionalità e caratteristiche** dell'infrastruttura
- Presentare l'**attuale situazione** di dati e infrastruttura di calcolo **per sede**
- Delinare passi e prospettive di **implementazione e avanzamento**

Perché parliamo di un'infrastruttura di dati e calcolo per IRPI

- Aumentare la **conoscenza** (e visibilità) interna ed esterna e l'**accesso** alle molte e diverse risorse e dati prodotti e/o gestiti
- Fornire ai ricercatori **servizi** avanzati di elaborazione e calcolo
- Garantire un'**archiviazione** sicura e persistente di informazioni di valore per l'Istituto
- **Mettere a sistema** iniziative già esistenti nelle varie sedi
- **Stimoli/vincoli legislativi o amministrativi** (e.g. Direttiva Open data, Direttiva INSPIRE, Open Science, normativa AGID, CAD, ...)

Il tavolo di lavoro



Due riunioni preliminari

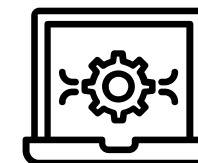
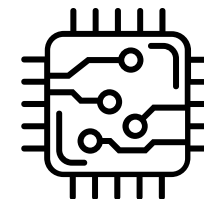
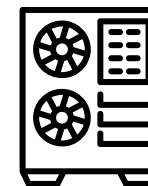
Rappresentanti di tutte le sedi:

- Bari
 - Nunzio Luciano Fazio
- Cosenza
 - Giovanni Gullà
- Padova
 - Alessandro Sarretta
- Perugia
 - Massimiliano Alvioli, Vinicio Balducci, Luca Ciabatta, Ivan Marchesini, Christian Massari, Mauro Rossi
- Torino
 - Paolo Allasia, Daniele Giordan, Diego Guenzi

round table by emilegraphics from the Noun Project

Di quale infrastruttura stiamo parlando

- Un'infrastruttura **informatica** (hardware, software, procedure) **distribuita e modulare per supportare attività di ricerca**
 - Concretamente: relativamente ai dati si comporrà di cataloghi di metadati, webGIS, DBMS e risorse di storage, servizi interoperabili; per quanto concerne il calcolo, si comporrà di risorse hardware/software dedicate al calcolo avanzato
 - Non stiamo parlando di strumentazione di laboratorio o di campo, o di server di posta elettronica, firewall



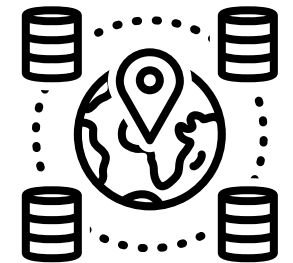
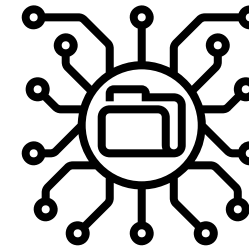
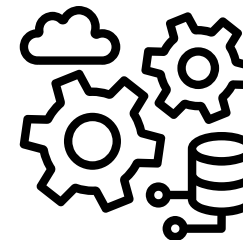
hardware by ProSymbols, Server by Ifki rianto, Software by remmachenasreddine from the Noun Project

Principali caratteristiche dell'infrastruttura

- Infrastruttura di **istituto**

Tre livelli principali

- Catalogo di **metadati**
- Organizzazione/condivisione di **dati**
- Risorse di **calcolo**



Metadata by M. Oki Orlando; Data access protection by Wichai Wi; geospatial by WEBTECHOPS LLP; data processing by DARAYANI from the Noun Project

Questioni aperte / criticità



- problematiche e criticità legate alla **disomogeneità** delle condizioni di partenza nelle diverse sedi e gruppi di ricerca in termini di competenze, approcci/propensioni e condivisione di dati e necessità di investimenti
- **costi**: la progettazione, creazione e gestione di qualunque tipo di infrastruttura richiede vari elementi da considerare: personale (già disponibile o da reclutare), risorse hardware, strategie di stabilizzazione e backup dei sistemi di alimentazione, risorse software, spazi interni e relativa strutturazione (e.g. climatizzazione)
 - ricerca di **finanziamenti** (e.g. Centro Nazionale HPC), progetti nazionali/internazionali per supportare creazione, mantenimento, sviluppo dell'infrastruttura

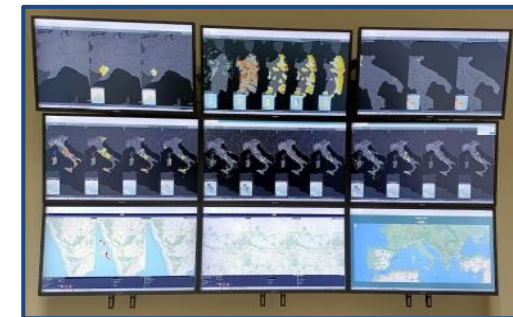
issue by SBTS from the Noun Project

Da dove partiamo: summary delle risorse per sede

Perugia

Sala CED dedicata con

- 4 armadi rack; sistemi di stabilizzazione (UPS); sistemi di backup alimentazione (gruppo elettrogeno diesel); sistemi di climatizzazione dedicati e ridondati; doppia connessione in fibra alla rete GARR (100 e 200 Gbit potenzialmente espandibile a 1 Gbit),
- 9 server adibiti alla virtualizzazione (5 ESXi VMware, 4 XCP-ng), sui quali girano attualmente 93 macchine virtuali utilizzate per gestione di sistemi di allerta e altri servizi verso l'eterno,
- 2 server con singolo sistema operativo che ospitano procedure di calcolo diversificate sviluppate in house (codici di calcolo paralleli e non, dedicati all'analisi numerica di dati geo-ambientali),
- 5 NAS e 2 SAN per la memorizzazione dei dati (con un totale di circa 220 TB di spazio effettivo per la memorizzazione dei dati).



La gestione del CED è quasi totalmente affidata ad un solo sistemista (che dedica larghissima parte del proprio tempo alla gestione del CED) coadiuvato in alcune operazioni da colleghi a Perugia e Torino

Dati: inventari di frana, database di piogge e dati ambientali, prodotti della ricerca, fototeche, dati satellitari, portate, contenuto d'acqua per un totale di alcuni TB di storage.

Torino

Principale **Infrastruttura IT**

- 2 workstation per l'elaborazione di modelli 3D e l'esecuzione di computazioni "CPU/GPU intensive"
- 6 NAS per la memorizzazione dei dati (con un totale di circa 150 TB di spazio effettivo per la memorizzazione dei dati)
- 5 server adibiti alla virtualizzazione, sui quali girano mediamente 30 macchine virtuali
- Sala server (in fase di ultimazione)

Principali **dati**

- Informazioni di diversa natura (rilievi drone/lidar, fotocamere/webcam, misure inclinometriche...) legate ai vari siti monitorati dal GMG, elencati alla pagina <http://gmg.irpi.cnr.it/index.php/it/dove-operiamo>
- Riprese aeree dal 1954 ad oggi, disponibili su <https://www.fototeca.to.irpi.cnr.it/lizmap/index.php/view/map/?repository=irpi&project=fototeca>
- Catasto delle frane di alta quota e altre informazioni legate al rapporto tra cambiamenti climatici e instabilità in contesto alpino, disponibili su <https://geoclimaalp.irpi.cnr.it>

Padova

Principale infrastruttura IT

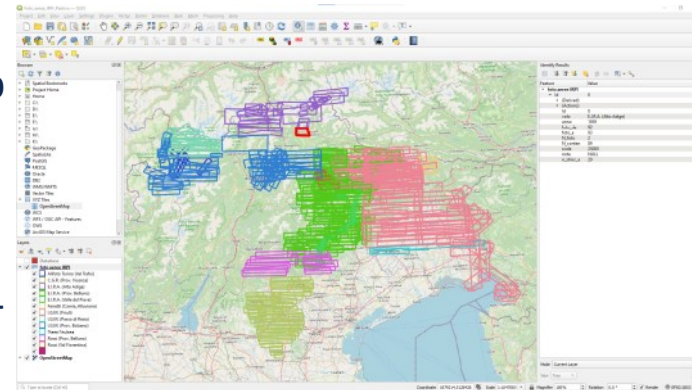
- 5 workstation per l'elaborazione InSAR, point cloud e modelli digitali ad alta risoluzione e modellazione numerica
- ~70 TB per lo storage dei dati
- 1 server per la gestione dei portali dei siti di monitoraggio

Principali dati

- Dataset di volumi e di monitoraggio di colate detritiche (pubblicati su repository open)
- Dati idrometeorologici e di monitoraggio di frane nel territorio Triveneto
- Riprese aeree dal 1954 (circa 25,000) parzialmente digitalizzate (circa 11,000) e consultabili in ambiente GIS
- Cartografia storica: Carta del Ducato di Venezia 1801-1805 - Von Zach, cartografia su eventi alluvionali
- Dati LiDAR aereo di alcuni bacini alpini

Desiderata

- Cloud storage di Istituto per back up di dataset di diversa natura
- Linee guida comuni per digitalizzazione e gestione delle foto aeree
- Individuazione di strumenti comuni per la condivisione e l'accesso collaborativo dei dati di istituto



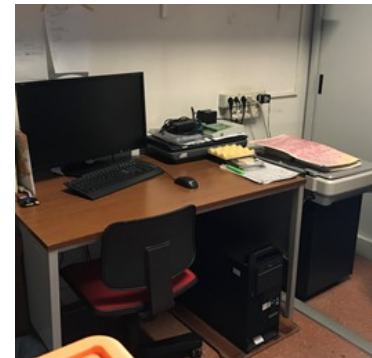
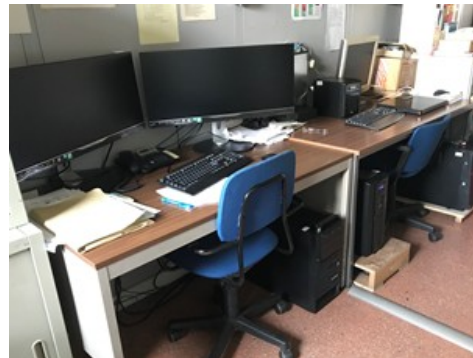
Cosenza

Censimento in itinere....

Connessione 300MBps

- 3 workstation per elaborazioni grafiche e calcolo numerico
- Server per gestione sottorete e produzione VM
- Nas Storage
- UPS dedicato
- Sistema di climatizzazione dedicato e centralizzato

- Sala CED (in via di ristrutturazione):
- Server rack con UPS integrato
- UPS aggiuntivi con stabilizzatori di tensione
- Connessione alla rete GARR con FW a 1Gbps
- Sistema di climatizzazione dedicato



Bari

Principale infrastruttura IT

- 8 workstation per modellazione numerica e gestione dati multipiattaforma
- 2 workstation per fotointerpretazione digitale
- ~20 TB per lo storage dei dati
- 1 server per la gestione delle risorse in rete della sede

Principali dati

- Database geomorfologico completo e aggiornato della Daunia (progetto GeoPuglia)
- Database delle caratteristiche meccaniche, litologiche, stratigrafiche e geofisiche acquisite nell'ambito del progetto Microzonazione Sismica Puglia
- Database cavità antropiche e sinkhole
- Database caratteristiche meccaniche calcareniti pugliesi
- GeoDatabase acquiferi del mediterraneo (gruppo di idrogeologia)
- Database dati climatici Italia Meridionale
- Database foto aeree Puglia

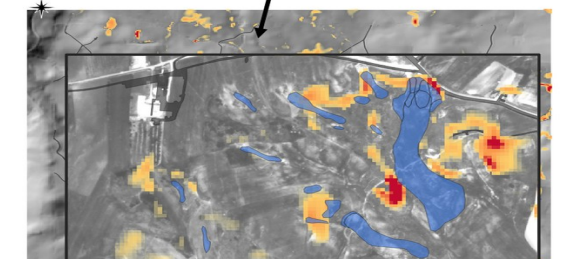
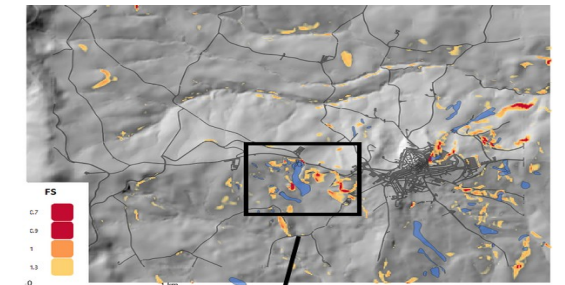
Prospettive future

- Piattaforma di calcolo numerico integrata con un database condiviso e interattivo per l'analisi numerica (FEM, LEM, ...) real time, multitemporale dei dissesti geo-idrologici anche attraverso tool già presenti in letteratura scientifica.
- Standardizzazione dei formati di condivisione del dato.

Esempio:

Dati necessari:

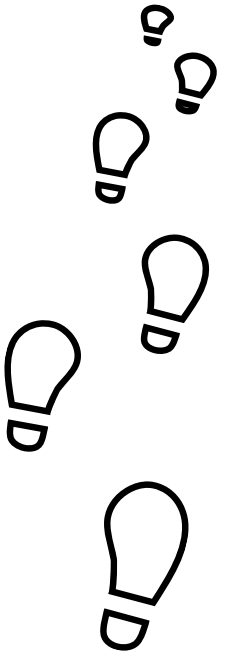
- DEM
- Dati di umidità pre-evento
- Dati pluviometrici
- Parametri fisico-meccanici dei terreni di copertura area di studio
- Spessore terreni di copertura



Ugenti et al. (2021, in prep.)

Prossimi passi, azioni da attivare

- **Completamento** del tavolo di lavoro con identificazione di momenti di discussione interni all'IRPI
- Consolidamento del **censimento** per sede
- Analisi delle **priorità e ipotesi/tempi/costi** di implementazione
- Definizione/scrittura di una **roadmap**
- Identificazione di un infrastruttura dati e calcolo **ottimale per l'IRPI** in relazione alla sua utilizzabilità, alla sua **sostenibilità** economica e manutenibilità (e.g. necessarie risorse umane ed economiche dedicate), e alla sue specificità legate alle attività dell'IRPI



footsteps by Ataur Rahman from the Noun Project