Report dal meeting degli Archivi Italiani

INAF Science Archives & the Big Data Challenge

17-19 June 2019
INAF
UTC timezone

Search...

- Overview
- Aim and Topics
- Important Date and Info
- Registration
- Call for Abstracts
- Timetable
- My Conference
  - My Contributions
- Participant List
- Location

The purpose of the workshop is to gather for discussion all main Italian actors involved in the use and management of astrophysical data, also within the interdisciplinary perspective of multimessenger. From an overview of the existing archives and their development, to the discussion of the Archive 2.0 concept for the Big Data, the different functionalities of archives will be presented. The use of modern era archives is no longer circumscribed to the search for scientific information, but it extends to providing the framework for the search, manipulation and analysis of data from telescopes, either terrestrial or satellite, of the new 2020 era.

SOC:

Cristina Knapic
23/10/2019

Goals

“The purpose of the workshop is to gather for discussion all main Italian actors involved in the use and management of astrophysical data, also within the interdisciplinary perspective of multimessenger. From an overview of the existing archives and their development, to the discussion of the Archive 2.0 concept for the Big Data, the different functionalities of archives will be presented. The use of modern era archives is no longer circumscribed to the search for scientific information, but it extends to providing the framework for the search, manipulation and analysis of data from telescopes, either terrestrial or satellite, of the new 2020 era.”

Partecipanti ed organizzazione

SOC:

- A. Antonelli
- A. Grado,
- C. Knapic (chair),
- E. Molinari,
- R. Morbidelli,
- M. Nanni,
- G. Polenta,
- R. Smareglia,
- A. Zanichelli.

LOC:

- C. Giorgieri
- R. Smareglia
- F. Tinarelli

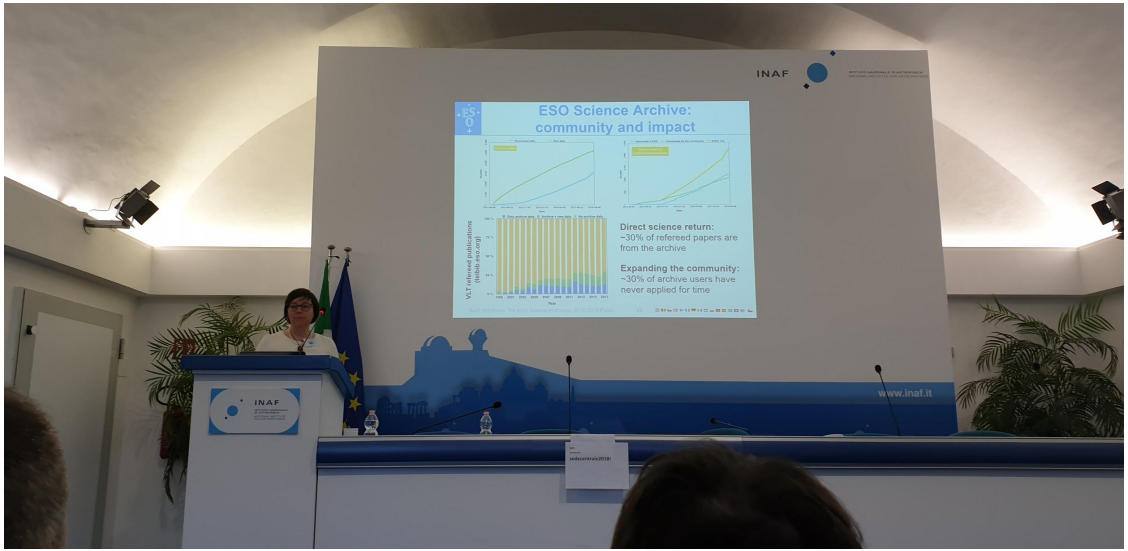
Invited Speakers:

- Severin Gaudet (CADC)
- Magda Arnaboldi (ESO)
- Stephan Schledstedt (CTAO)
- Pia Astone (INFN)
- Marco Molinaro (INAF)
- Rosie Bolton (SKAO) - R.Smareglia

Si ringrazia la DS per logistica e catering

Partecipanti ed organizzazione

Registrati: 80 (limite della capienza)
Presenti: circa 65 al giorno
Da remoto: circa 10
Not shown: circa 10



Name	Affiliation
Eleonora Alei	UNIPD/INAF-OAPD
Magda Arnaboldi	ESO
Pia Astone	INFN, sezione di Roma
Ugo Becciani	INAF - OACT
Daniela Bettoni	Istituto Nazionale di Astrofisica (INAF)
Andrea Bignamini	Istituto Nazionale di Astrofisica (INAF)
Andrea Botteon	IRA-INAF
Sandra Burkutean	IRA-INAF, Italian ARC
Deborah Busonero	Istituto Nazionale di Astrofisica (INAF)
Alberto Buzzoni	INAF - OAS Bologna
Guido Cupani	INAF–OATs
Michele Fabrizio	INAF
Fabiana Faustini	SSDC-ASI and INAF-OAR
Stefano Gallozzi	INAF-Osservatorio Astronomico di Roma
Vincenzo Galluzzi	INAF - Osservatorio Astronomico di Trieste
Séverin Gaudet	NRC/CADC
Stavro L. Ivanovski	Istituto Nazionale di Astrofisica (INAF)
Cristina Knapic	Istituto Nazionale di Astrofisica (INAF)
Mario G. Lattanzi	INAF - Osservatorio Astrofisico di Torino
Saverio Lombardi	INAF-OAR and ASI-SSDC
Silvia Marinoni	Istituto Nazionale di Astrofisica (INAF) & SSDC
Marcella Massardi	INAF-IRA / Italian ARC
Marco Molinaro	Istituto Nazionale di Astrofisica (INAF)
Giuseppe Murante	INAF - OATs
Ilaria Musella	INAF
Mauro Nanni	Istituto Nazionale di Astrofisica (INAF)
Luciano Nicastro	INAF-OAS
Diego Paris	INAF - OAR
Nicolo' Parmiggiani	Istituto Nazionale di Astrofisica (INAF)
Fabio Pasian	INAF - OATs
Matteo Perri	INAF-OAR and SSDC-ASI
Carlotta Pittori	INAF-OAR & ASI-SSDC
Marzia Rivi	INAF - IRA
Stefan Schlenstedt	CTAO
Marco Scodeggio	INAF IASF Milano
Simone Silvestro	INAF OACN
Riccardo Smareglia	Istituto Nazionale di Astrofisica (INAF)
Franco Tinarelli	INAF - IRA
Fabio Vitello	
Sheng Yang	Observatory of Padova, INAF
Alessandra Zanichelli	Istituto Nazionale di Astrofisica (INAF)
Angelo Zinzi	SSDC - ASI

Organizzazione delle sessioni

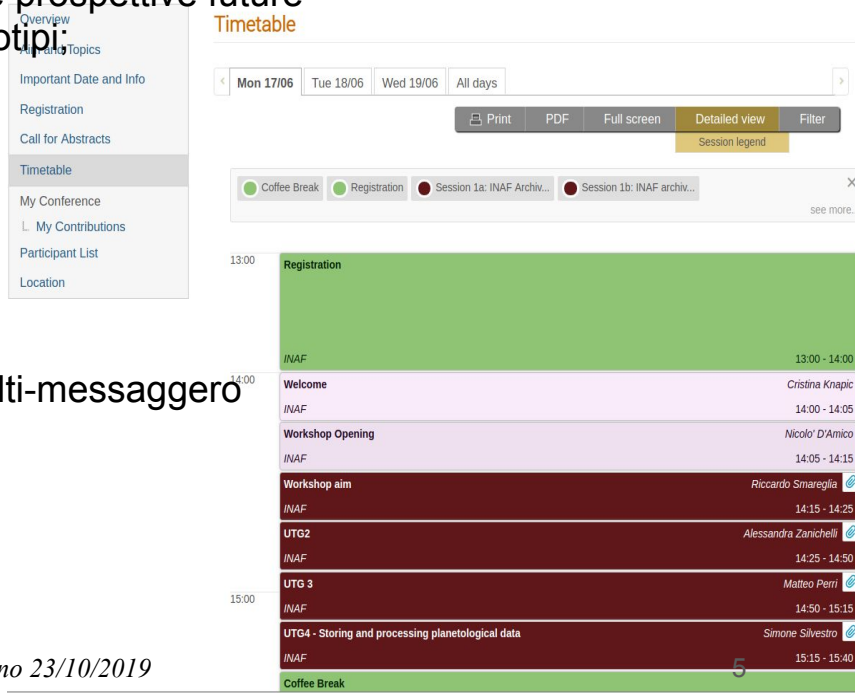
Sessione 1: Status degli archivi in Italia
- A. Report dalle UTG 2,3,4;
- B. Report dalle maggiori facility di archivio esistenti in Italia.

Sessione 2: Dai dati alla scienza e ritorno: lo stato dell'arte e le prospettive future
- A. Esempi di Facility internazionali allo stato dell'arte e prototipi;
- B. Nuove frontiere degli archivi

Sessione 3: data processing and pipelines
- A. Processing and data;
- B. Workflow management systems and pipelines

Sessione 4: Nuove frontiere dell'astronomia
- A. Contesto internazionale ed Europeo per l'astronomia multi-messaggero
- B. Interoperabilità.

Sessione 5: Tecnologia e Science Gateways
- A. Tools per i science gateways;
- B. Tecnologie e strategie per il computing.



Le discussioni

Grande spazio è stato lasciato volutamente per le discussioni. I temi sono stati:

- Le nuove prospettive e frontiere dell'accesso ai dati
  - gli archivi, le interfacce, l'approccio, le ricerche e la facilità di utilizzo
    - ridurre il numero di filtri di ricerca per minimizzare errori nella ricerca o viceversa?
    - nel multi-wavelength o nel multi-messenger usare approcci custom per ciascun caso o un approccio generico?
  - l'utilizzo dei dati di archivio e il monitoring delle esigenze della comunità
    - gli archivi stanno rispondendo alle esigenze della comunità ora?
    - e alle esigenze di domani (vedi SKA o LSST o CTA o Euclid)?
  - i dati science ready: le policy di accesso: cosa esiste al momento;
  - quali tools ha senso agganciare agli archivi (cut-out, mosaics, …);
  - riusabilità dei dati: gli science goals non sono l'unica informazione contenuta nei dati, si potrebbe estrarre molto altro;
  - aumentare la connessione tra le comunità lavorando su vari fronti: archivi, tools, strumentazione, standards e interoperabilità;

continua …….

Le discussioni

- I sistemi di gestione del workflow e le pipelines
  - le pipeline di riduzione sono specifiche per strumento e spesso non sono pubbliche.
    - Come possiamo permettere l'utilizzo di queste preziose risorse da parte della comunità?
    - come gestire la data quality nel caso di dati ri ridotti dagli utenti?
    - quali utenti sono abilitati a gestirle?
    - Contributo sull'esito del workshop spettroscopico
  - le funzionalità dei sistemi di workflow;
    - pipelines private;
    - workflow interattivo;
    - ultima versione dello stack software;
  - gestione dei data products
    - policy dei dati processati e proprietà intellettuale;
    - pubblicazione dei dati processati (DOI?)

continua…..

Le discussioni

- Future solutions:
  - Visione della Direzione Scientifica:
    - “INAF astronomical community coming late with respect to other communities (f.i. genomics) wrt big data, HPC, archiving. What we need is mentality, effort and infrastructure” -> discussione con gli invited sulle facilities internazionali
    - “New figures in INAF are needed, who are in the middle between science and technology..” -> discussione sulle figure a supporto e di supporto dei centri dati/computing;
  - Cosa si fa in Europa: EOSC -> not for free! E i dati devono essere curati!!
  - Preservazione: DOI e Open Science, il futuro delle nostre pubblicazioni. Abbiamo il servizio DOI ma come possiamo massimizzare l’utilizzo e il problema della proprietà intellettuale delle pubblicazioni su rivista o meno;
  - Importanza dell’interoperabilità -> esistono degli standard e delle best practices, usiamole!
    - il concetto di data curation: si estende dalla conservazione dell’accessibilità al dato e al fatto che i formati divengono desueti al completamento delle informazioni corollarie (dalla data quality alla provenance..)
    - meta descrittori e WCS sono concetti trascurati perchè non si ha la visione del riuso;
  - Modelli, formati e servizi: una catena frammentata che deve diventare un tutt’uno!

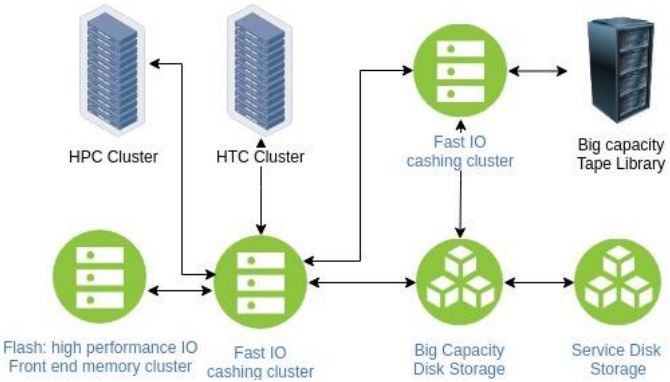
continua…..

Le discussioni

- Science Gateways:
  - Non solo experiment devoted ma sistemi in grado di permettere un reale riutilizzo del contenuto scientifico dei dati acquisiti e fluidi a sufficienza da garantire un incremento graduale delle risorse a disposizione. Lo science gateway è visto come un catalizzatore di risorse (dati, strumenti e applicazioni) a supporto della comprensione scientifica del caso.
    - strumenti: computazione, storage, supporto…
    - applicazioni: analisi, visualizzazione, calibrazione, cooperazione, servizi di archivio, interoperabilità…
    - dati: ricerca, accesso, produzione, condivisione…
  - Aggancio della parte archivi con le necessità di calcolo;
  - Il calcolo ai dati: in prospettiva di grandi infrastrutture osservative internazionali, è ragionevole pensare di ottimizzare i tempi scala di processamento avanzato presso i data center spostando la computazione verso i dati. Questo implica un cambiamento di prospettiva per chi sviluppa gli science gateway e soprattutto di chi li utilizzerà.
    - USER SPACE:
      - non più archivio da cui scaricare in locale dati, ma ambiente integrato con la possibilità di fare calcolo, applicare algoritmi, trovare tools per l'analisi;
    - SUPPORTO: la complessità di alcuni progetti implica la necessità di avere supporto sia scientifico che informatico per la corretta riduzione ed analisi dei dati.

Le discussioni

- Archivi e computazione (HW):
  - Archivi:
    - multi-tiered;
    - distribuiti;
    - business logic coordinata e interoperabile;
  - Computazione:
    - nuove tecnologie (esempio di LOFAR): quanto sono facilmente utilizzabili?
    - la mentalità vs cambio approccio;
    - i PoC con i provider cloud esterni come sono andati?
    - è comunque indispensabile il supporto scientifico tecnico;



Considerazioni preliminari

a. I dati provengono da diversi data providers (telescopi, surveys, simulazioni) e rilasciati da vari siti/istituzioni (info da UTG) in situazione di eterogeneità;
b. Quali infrastrutture di archivi astronomici ci sono:
    i. Infrastruttura INAF -> IA2 ;
    ii. ASI DC attualmente dedicato a missioni spaziali anche con personale INAF;
    iii. varie infrastrutture specifiche per progetto;
c. Un sacco di contributi da molti campi e progetti diversi, con necessità e casi d'uso completamente diversi;
d. E' emersa l'esistenza di molti strumenti per la data handling / visualization / analysis;
e. Abbiamo visto due differenti prospettive di Archivi Internazionali (ESO , CADC) che usano entrambi standards ben definiti e sono dotati di sistemi che garantiscono l'interoperabilità;
f. Pipelines: non sempre disponibili ma.. è nato il gruppo di Coordinamento Spettroscopia!
g. Direttore Scientifico da due informazioni importanti per il futuro:
    i. Ci vorrà un cambiamento di mentalità, uno sforzo congiunto e le infrastrutture;
    ii. Serviranno nuove figure in INAF a metà tra scienza e tecnologia.

Considerazioni preliminari

h. Non è fondamentale avere tutte le risorse in house ma riuscire a fare sistema
i. “Se è ESO a richiedere i dati (phase2), tutti sull’attenti……”;
j. Sarebbe auspicabile la presenza di rappresentanti della comunità degli utenti per feedbacks;
k. Uso degli Standards: FAIR, IVOA, RDA, common data models in modo da non fare impazzire l’utente con formati e nomenclature ;
l. Machine learning e data mining è strategico per la data exploitation!
m. I casi più demanding prevedono la compresenza di archivi e calcolo, possibilmente usando degli science gateways.

Vantaggi

Abbiamo cominciato a parlarne!

- Scambio di idee e diffusione delle nuove tendenze;
- serve per confrontarsi con realtà internazionali;
- serve per avere feedback dalla comunità;
- le challanges tecnologiche ci spronano a aprire gli orizzonti anche per le realtà meno intensive;
- supporto alla comunità;
- creazione di gruppi di lavoro su tasks specifici;
- coordinamento a livello ICT;
- sistemi per la condivisione: dai collaborative agli sharing tools

Criticità

- Ricostrustruzione e riusabilità usando i dati ancillari (telemetria,..)
- pipelines private/ non facilmente usabili
- standard non adottati comunemente
- scarsa circolazione delle informazioni sui nuovi sviluppi.. si reinventa la ruota
- scarsa interazione tra i gruppi, tra i progetti e i teams

la comunità non vede ancora il vantaggio di fare sinergia… mancanza di chiara comunicazione e comunione di intenti.

ROADMAP

- chiarire cosa si intende per coordinarsi (gruppo tecnologico);
- chiarire i motivi per cui conviene coordinarsi (sia tecno che science);
- trovare il consenso, con il supporto della DS, su:
  - interazione nazionale su attività comuni;
  - definizione di strategie comuni, forte cooperazione e coordinazione;
  - evidenziare i contributi individuali all'interno di un panorama nazionale;
- capire quali sono i servizi di supporto utili e condivisibili;
  - delegare il più possibile il lavoro all'utente;
  - fornire supporto all'utilizzo dei servizi
- cambiare l'approccio dell'utenza: convincere la comunità dei benefici degli science gateways;
  - cambio della percezione del concetto di archivio: data+computing+supporto

THANK YOU!