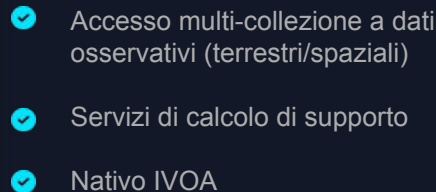




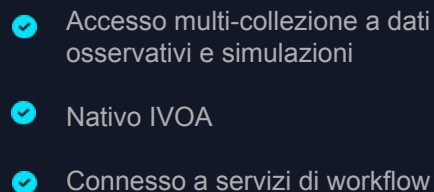
AI: come preparare gli archivi dell'era del Big Data per lo sfruttamento scientifico ad ampio spettro

Cristina Knapic

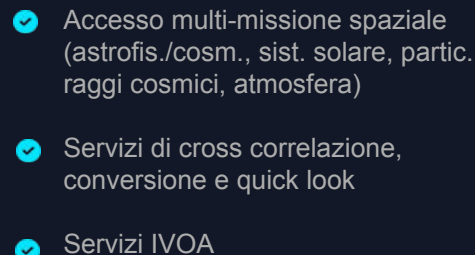
Canadian Astronomy Data Centre (CADC)



Italian Astronomical Archives (IA2)

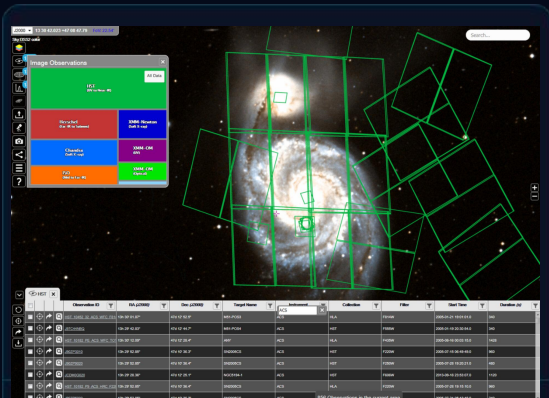


Space Science Data Center (SSDC)



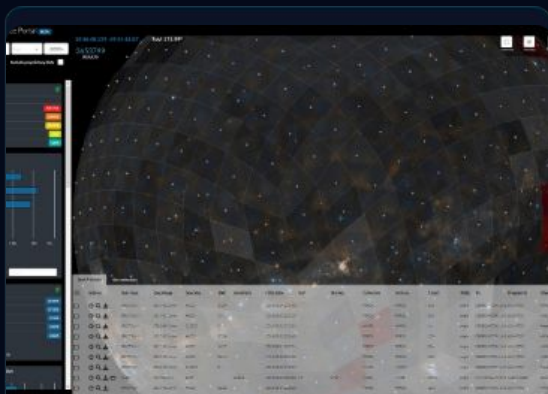
HUB ISTITUZIONALI: I PILASTRI GLOBALI

Le grandi agenzie spaziali hanno trasformato gli archivi da semplici repository a **ecosistemi interoperabili** basati sugli standard IVOA.



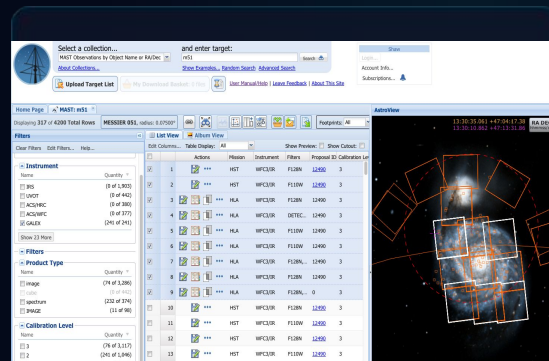
ESA Sky (ESAC)

- ✓ Accesso multi-missione "all-sky".
- ✓ Native IVOA: TAP, SIA, SSA.
- ✓ Ideale per generare cutouts multi-banda per training set IA.



ESO Science Archive

- ✓ Focus sui dati "Phase 3" (ridotti).
- ✓ Metadatezione ricca per spettroscopia.
- ✓ Elimina il bias del preprocessing strumentale per il Deep Learning.



MAST (NASA/STScI)

- ✓ Hub per JWST, Hubble e TESS.
- ✓ Supporto avanzato per **ASDF**.
- ✓ API integrate per integrazione con framework Python/Machine Learning.

IL FUTURO: SCIENCE PLATFORMS

Rubin Science Platform (RSP)

Rubin Science Platform

L'aumento esponenziale del volume dati (Petascale) impone il passaggio dal "Download" al paradigma **Code-to-Data**.

L'archivio di LSST non è un silo, ma un'infrastruttura di calcolo dove l'IA risiede accanto allo storage:

Portal

Discover data in the browser



Notebook Aspect: Ambienti Jupyter pronti per l'analisi massiva.

API-first design: Interfacce programmatiche IVOA

[Learn more about the portal.](#)

prioritarie rispetto alla UI.

SDSS Legacy: L'integrazione di SDSS in MAST come

benchmark per il training dei modelli.

Notebooks

Process and analyze LSST data with Jupyter notebooks in the cloud



[Learn more about notebooks.](#)

APIs

Learn how to programmatically access data with Virtual Observatory interfaces



Numeri e il perché delle SCIENCE PLATFORMS

Negli ultimi due decenni, l'astronomia è stata teatro della realizzazione di osservatori e survey multibanda, con strumenti sofisticati che acquisiscono un'enorme quantità di dati di eccezionale qualità.

Alcuni numeri:

- ✓ LBT : ~40 TB/anno → 500 TB
- ✓ Euclid : ~100 GB/giorno per 6 anni → 200 TB
- ✓ Pan-STARRS : >100 TB di dati
- ✓ JWST : ~70GB/giorno per 10 anni (e oltre) → 250 TB
- ✓ GAIA : ~1 PB in 5 anni
- ✓ Rubin/LSST : ~30 TB/notte per 10 anni → >100 PB
- ✓ CTA : ~40 PB/anno di cui 12 PB/anno archiviati → 100 PB
- ✓ SKA : ~ 700 Pbyte/anno → 35 EB
- ✓ VLT, E-ELT, ALMA



Numeri e il perchè dell'uso della AI

- Se raccogliamo dati a sufficienza con uno spazio dei parametri ad alta dimensionalità all'interno di un qualsiasi dominio di studio, allora avremo un modello statistico coerente per quel dominio.
- Ogni dato rappresenta una soluzione al problema di dominio, quindi in un regime Big Data abbiamo un ordine di $\sim 10^{20}$ soluzioni.

MAGICO!!!

Tutto ciò che vogliamo sapere su un dominio è specificato e codificato all'interno dei dati. Pertanto, l'obiettivo finale di qualsiasi scienza dei Big Data è trovare quelle codifiche, modelli e pezzi di conoscenza.



Data-driven Science = Scientific Knowledge Discovery in Databases

DAL MODEL-CENTRIC AL DATA-CENTRIC AI

L'evoluzione dell'AI in astrofisica richiede un cambio di paradigma:

Model-Centric: Focalizzato sull'ottimizzazione dell'architettura della rete neurale.

Data-Centric: Focalizzato sulla qualità, consistenza e struttura del dato (Andrew Ng, 2021).

Il **Data Manager** diventa il "progettista" del carburante per l'AI.

ML PROs & CONs

VANTAGGI

- ✔ Automazione dei processi: l'apprendimento automatico (ML) consente ai sistemi di prendere decisioni e adattarsi ai cambiamenti senza bisogno di un intervento umano diretto.
- ✔ Capacità predittiva: fornisce potenti strumenti per prevedere i risultati futuri analizzando i dati storici.
- ✔ Apprendimento continuo: i modelli possono essere aggiornati con nuovi dati, migliorandone le prestazioni nel tempo.
- ✔ Applicabilità trasversale: le tecniche di apprendimento automatico possono essere utilizzate in diversi settori, come medicina, agricoltura, marketing, industria... e astronomia!

SVANTAGGI

- ✔ Richiede grandi quantità di dati: le prestazioni del modello dipendono fortemente dalla disponibilità di dati pertinenti e rappresentativi.
- ✔ Rischio di overfitting: modelli eccessivamente complessi possono memorizzare i dati di addestramento e non riuscire a generalizzare bene su dati nuovi.
- ✔ Distorsione dei dati: se il set di dati di addestramento è distorto o sbilanciato, il modello può produrre risultati inaccurati o discriminatori.
- ✔ Interpretazione limitata: molti modelli complessi, come le reti neurali profonde, sono difficili da interpretare e spiegare.
- ✔ Costi computazionali: l'addestramento e l'utilizzo di modelli avanzati possono richiedere risorse hardware significative e tempi di elaborazione elevati.

DL PROs & CONs

VANTAGGI

- ✔ Estrazione automatica delle caratteristiche: questo è il vantaggio principale. Non è più necessario definire manualmente quali parametri fisici siano importanti (ad esempio, indici di colore o profili di Sersic). La rete "apprende" autonomamente le caratteristiche più discriminanti direttamente dai dati grezzi.
- ✔ Approssimazione universale: matematicamente, una rete neurale profonda (DNN) sufficientemente profonda può approssimare qualsiasi funzione continua. Questo la rende insuperabile nella modellazione di fenomeni fisici altamente non lineari e complessi.
- ✔ Scalabilità con i Big Data: a differenza dei modelli classici, le cui prestazioni si stabilizzano all'aumentare del volume dei dati, le DNN continuano a migliorare con la crescita del dataset. Questo è fondamentale per le indagini di nuova generazione come LSST (Vera Rubin) o Euclid.
- ✔ Integrazione multimodale: possono integrare senza soluzione di continuità fonti di dati eterogenee (ad esempio, immagini CCD + metadati tabellari + spettri) in un unico modello predittivo coerente.

SVANTAGGI

- ✔ Natura "a scatola nera" (interpretabilità): spesso è difficile decifrare perché una rete neurale abbia raggiunto una determinata conclusione. In un contesto scientifico in cui la causalità è fondamentale, questo rappresenta un ostacolo significativo. Tuttavia, tecniche come i valori SHAP contribuiscono a mitigare questo problema.
- ✔ Fame di dati: per funzionare in modo affidabile, le reti neurali profonde (DNN) richiedono migliaia (o milioni) di esempi etichettati. Se il campione di oggetti astronomici rari è ridotto, la rete potrebbe non riuscire a generalizzare efficacemente.
- ✔ Costo computazionale: l'addestramento richiede hardware dedicato (GPU/TPU) e un consumo energetico significativo. Non è sempre la soluzione più efficiente per compiti più semplici e a bassa dimensionalità.
- ✔ Overfitting e fragilità: con milioni di parametri, la rete può facilmente "memorizzare" il rumore presente nel set di addestramento anziché apprendere la fisica sottostante, con conseguente fallimento quando applicata a dati nuovi e mai visti prima.

Ma.....

Per far sì che la magia avvenga, abbiamo dei passi da compiere:

- ✓ caratterizzare i dataset e le collezioni;
- ✓ insegnare i pattern noti, assegnare correttamente i parametri di dominio;
- ✓ supervisionare le indagini su quanto non già noto;
- ✓ esplorare quanto non già noto senza supervisione;
- ✓ rilevare ed analizzare le anomalie;
- ✓ applicare su grandi moli di dati.

Il punto dolente è spesso la caratterizzazione di quanto noto: la disomogeneità e veridicità dei dati e della caratterizzazione delle collezioni, a partire dalla definizione dei metadati.

.... partiamo da delle definizioni.

I PILASTRI DELL'ARCHIVIO AI-READY



DMP

Piani di gestione dati orientati alla machine-actionability.



Data Models

Oltre il FITS: semantica e multi-dimensionalità.



IVOA Standards

Interoperabilità per l'accesso programmatico (API-first).



Archivi

Dalla conservazione passiva alla Science Platform.

Definizioni generali – DMP

Il Data Management Plan (DMP) è un documento fondamentale che descrive come i dati della ricerca verranno gestiti, organizzati, archiviati e condivisi durante l'intero ciclo di vita di un progetto. È un "living document" (documento dinamico) che cresce con il progetto e rappresenta spesso un requisito obbligatorio per enti finanziatori come *Horizon Europe*. Esso include:

- ✔ Descrizione dei dati: Quali dati verranno creati o riutilizzati.
- ✔ Standard e Metadati: Come verranno descritti e strutturati i dati.
- ✔ Principi FAIR: L'adesione a principi che rendono i dati *Findable* (reperibili), *Accessible* (accessibili), *Interoperable* (interoperabili) e *Reusable* (riutilizzabili).
- ✔ Etica e Legalità: Aspetti sulla privacy, proprietà intellettuale e sicurezza dei dati.
- ✔ Conservazione e Condivisione: Strategie per l'archiviazione a lungo termine e le modalità di condivisione pubblica.

Perché è importante:

- ✔ Obbligatorietà: Necessario per la maggior parte dei finanziamenti europei e nazionali.
- ✔ Efficienza: Previene la perdita di dati e ottimizza il lavoro di ricerca.
- ✔ Trasparenza: Favorisce l'Open Science (Scienza Aperta) e la trasparenza dei risultati.
- ✔ Definisce le finalità di utilizzo dei dati e indica le modalità di accesso programmatico.

I DMP devono essere sintetici, schematici e precisi, quantificando le risorse e specificando i repository scelti per la pubblicazione 12

Definizioni generali – Collezioni

Collezione:

☑ insieme di oggetti raccolti secondo rigorosi criteri di selezione e destinati alla conservazione per il loro valore intrinseco o interesse, o per semplice passione o passatempo personale.

(editoria) serie di opere letterarie di un determinato autore o su un determinato argomento, tema o genere pubblicata da una casa editrice

(moda) l'insieme delle creazioni sartoriali di un determinato stilista o casa di moda relative a una specifica stagione

☑ La collezione è un gruppo di risorse correlate tra loro in modo identificabile. La relazione può riguardare un argomento, un luogo, una persona ecc. La relazione dipende dalla logica umana e in astrofisica solitamente dipende dal campo di indagine (cosmologia, fisica stellare, transienti...), dalle caratteristiche dello strumento e dallo scopo scientifico.

☑ Una collezione è un'aggregazione di elementi fisici o digitali, ad esempio collezioni di biblioteche o musei; cataloghi di biblioteche, musei o archivi; archivi digitali; directory internet; gateway tematici internet; raccolte di testi, immagini, suoni, dataset, software ecc...

☑ Una collezione può essere composta da un numero qualsiasi di elementi, da uno a molti.

☑ Una collezione è un insieme di risorse riunite per un particolare pubblico o per assolvere a una funzione specifica.

| Definizioni generali - Collezioni

Una Collezione è un organismo in crescita: in particolare in Astrofisica, tutti gli elementi relativi a osservazioni, simulazioni, informazioni accessorie, calibrazioni, software per la calibrazione e l'analisi dei dati, visualizzazione ecc... possono aumentare in numero e tipologia.

Tre concetti:

- Una collezione è una singola entità
- Una collezione è un'entità che esiste per servire una missione specifica
- Diverse collezioni dovrebbero essere parte integrante di una collezione più ampia
(vedi concetto di Interoperabilità)

| Definizioni Generali - Dataset

Dataset:

Un data set o dataset è una collezione di dati.

Più comunemente, un dataset costituisce un insieme di dati strutturati in forma relazionale, che corrisponde al contenuto di una singola tabella di database, o a una singola matrice di dati statistici, in cui ogni colonna della tabella rappresenta una particolare variabile e ogni riga corrisponde a un membro specifico del dataset in questione. La dimensione del dataset è data dal numero di membri presenti, che formano le righe, e dal numero di variabili da cui è composto, che formano le colonne.

Il termine dataset può essere utilizzato anche in modo più generico per indicare dati in un insieme di elementi strettamente collegati relativi a un particolare esperimento o evento.

Definizioni Generali

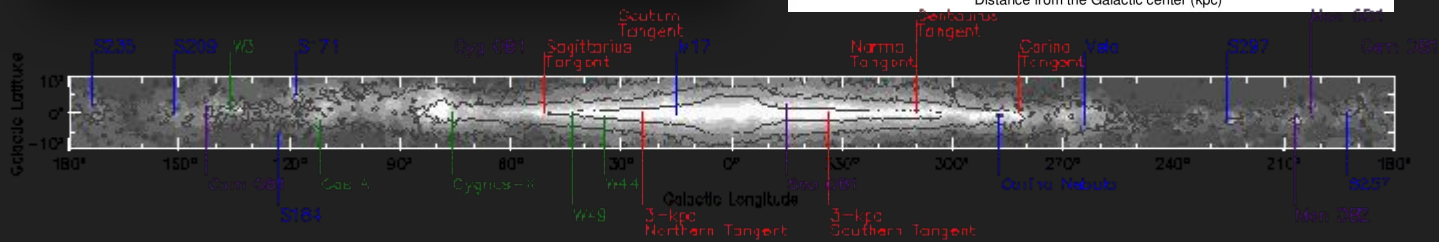
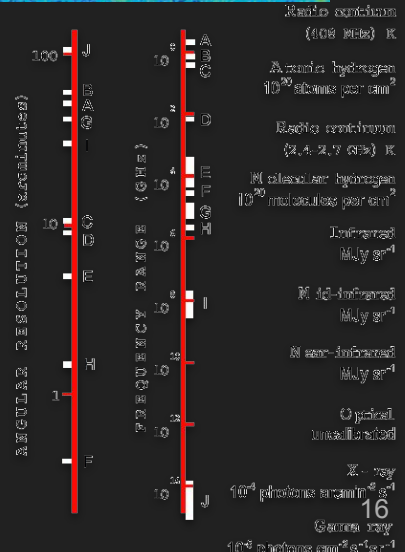
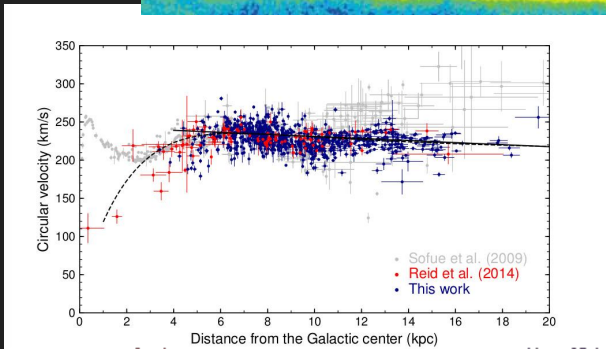
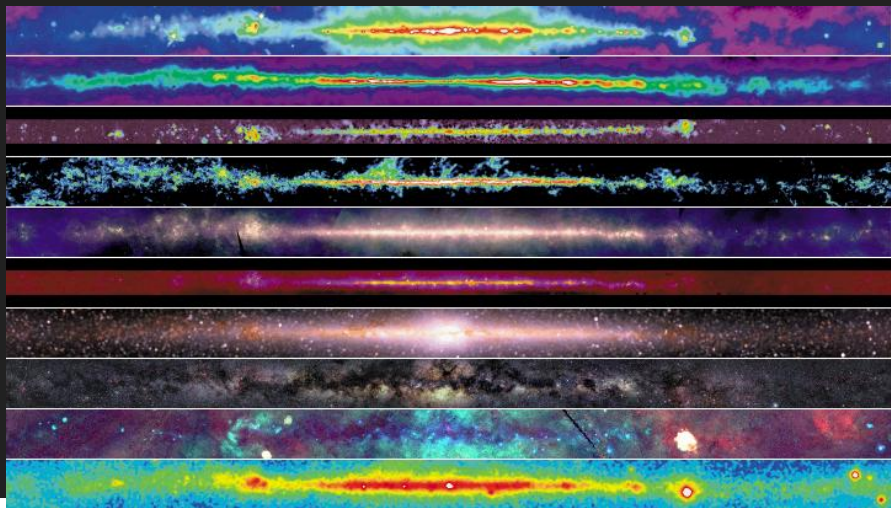
Dataset

```
vijinimallawaarachchi — mysql -u root -p — 81x36

[mysql> SELECT * FROM department;
+-----+-----+-----+
| dept_name | building | budget |
+-----+-----+-----+
| Biology   | Watson   | 90000.00 |
| Comp. Sci. | Taylor   | 100000.00 |
| Elec. Eng. | Taylor   | 85000.00 |
| Finance    | Painter   | 120000.00 |
| History    | Painter   | 50000.00 |
| Music      | Packard   | 80000.00 |
| Physics    | Watson    | 70000.00 |
+-----+-----+-----+
7 rows in set (0.00 sec)

[mysql> mysql> SELECT * FROM course;
+-----+-----+-----+-----+
| course_id | title | dept_name | credits |
+-----+-----+-----+-----+
| BIO-101   | Intro. to Biology | Biology | 4 |
| BIO-301   | Genetics | Biology | 4 |
| BIO-399   | Computational Biology | Biology | 3 |
| CS-101    | Intro. to Computer Science | Comp. Sci. | 4 |
| CS-190    | Game Design | Comp. Sci. | 4 |
| CS-315    | Robotics | Comp. Sci. | 3 |
| CS-319    | Image Processing | Comp. Sci. | 3 |
| CS-347    | Database System Concepts | Comp. Sci. | 3 |
| EE-181    | Intro. to Digital Systems | Elec. Eng. | 3 |
| FIN-201   | Investment Banking | Finance | 3 |
| HIS-351   | World History | History | 3 |
| MU-199    | Music Video Production | Music | 3 |
| PHY-101   | Physical Principles | Physics | 4 |
+-----+-----+-----+-----+
13 rows in set (0.00 sec)

mysql>
```



Definizioni generali - Dataset

Come potrebbe apparire un dataset

Osservazioni

Articolo

Dati Calibrati

Condizioni al Contorno

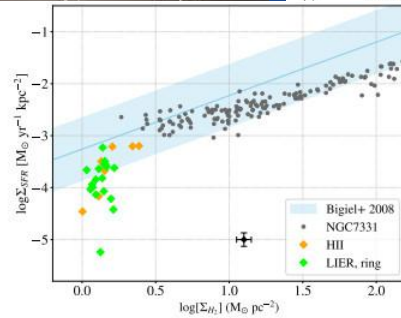


Figure 14: A scatter plot showing the relationship between the logarithm of the surface rate density ($\log \Sigma_{\text{core}} [\text{M}_{\odot} \text{yr}^{-1} \text{kpc}^{-2}]$) on the y-axis and the logarithm of the surface density ($\log \Sigma_{\text{H}_2} [\text{M}_{\odot} \text{pc}^{-2}]$) on the x-axis. The plot includes data points for various galaxies, with a blue shaded region representing the 3σ error on the slope. Data for NGC 7331 (orange dots), M104 (orange dots), and LIER ring (green diamonds) are marked with gray dots. Median error bars for the data from M104 (orange) and green diamonds are shown as a black cross in the lower part of the plot.

Risultati

DRAFT VERSION OCTOBER 26, 2022
Typeset using L^AT_EX two-column style in AASTeX

A Molecular Gas Ring Hidden in the Sombrero Galaxy

JESSICA SUTTER¹ AND DARIO PADD²

¹SOPHIA Science Center, USRA, NASA Ames Research Center, M.S. N220-12 Moffett Field, CA 94035, USA

ABSTRACT

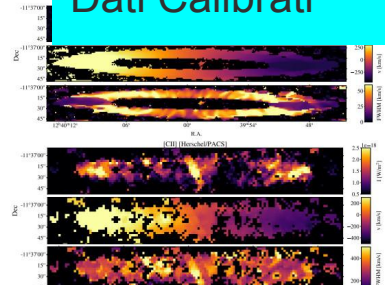
We present Herschel, ALMA, and MUSE observations of the molecular ring of Messier 104, also known as the Sombrero galaxy. These previously unpublished archival data shed new light on the content of the interstellar medium of M104. In particular, molecular hydrogen measured by CO emission and dust measured by far-infrared light are uniformly distributed along the ring. The ionized gas revealed by H α and [CII] emission is distributed in knots along the ring. Despite being classified as an SAA galaxy, M104 displays features typical of early-type galaxies. We therefore compared its [CII] and dust emission to a sample of early-type galaxies observed with *Herschel* and SOFIA. The [CII]/FIR ratio of M104 is much lower than that of typical star-forming galaxies and is instead much more similar to that of early-type galaxies. By classifying regions using optical emission line diagnostics we also find that regions classified as HII lie closer to star-forming galaxies in the [CII]/FIR diagram than those classified as low-ionization emission regions. The good match between [CII] and H α emission in conjunction with the lack of correlation between CO emission and star formation suggest that there is very limited active star formation along the ring and that most of the [CII] emission is from ionized atomic gas rather than molecular gas. From the total intensity of the CO line we estimate a hydrogen mass of $0.9 \times 10^8 \text{ M}_{\odot}$, a value intermediate between those of early type galaxies and the molecular ring of our galaxy.

INTRODUCTION

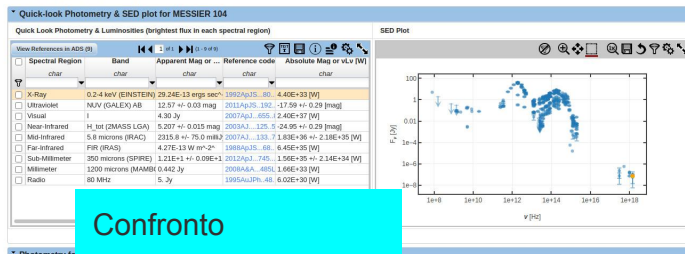
Galaxies are complex structures that can be shaped by torques. Each galaxy has its own unique shape which can lead to the creation of spirals, or even a lack of structure at all, leading to the distinct structural nature of galaxies. The distinct structural nature of galaxies has been some debate about the physical properties of a galaxy. A galaxy may appear devoid of structure from one viewing angle, but a different picture if viewed face-on. The physical properties of a galaxy can be determined by the context to distinguish galaxy characteristics and determine how viewing angle affects the appearance of a galaxy.

J. Jessica Sutter (moving to UCSD)
Lcom

One particularly interesting galaxy for its unique histories and observing perspective is the nearby galaxy M104, commonly referred to as the 'Sombrero' galaxy. General properties of this galaxy are summarized in Table 1. M104 has been classified as an SAA type galaxy (see e.g. Krause et al. 2006, an elliptical). While it appears as an edge-on galaxy with a thick ring of dust, visible as a dark HST image shown in Figure 1, the bulge is similar to those of an elliptical galaxy (Taddei et al. 2006) and measurements of star-formation and dust temperatures match more closely type galaxies than similar nearby spirals. The decomposition of the structures observed at M104 does not follow classical bulge/disk and instead is better represented as an elliptical galaxy with a large halo (Gadotti & Sánchez-Janssen 2014) of further interest due to the mass of observational evidence indicating it has



Photometry & SED for MESSIER 104



Open Archival Information System

L'OAIS si riferisce al modello di riferimento ISO OAIS. Questo modello di riferimento è definito dalla raccomandazione CCSDS 650.0-B-2 del Consultative Committee for Space Data Systems. L'ambito del CCSDS sono le agenzie spaziali, ma il modello OAIS da esso sviluppato si è dimostrato utile per altre organizzazioni e istituzioni con esigenze di archiviazione digitale. L'OAIS è ampiamente accettato e utilizzato da varie organizzazioni e discipline, sia nazionali che internazionali, ed è stato progettato per garantire la conservazione. Lo standard OAIS, pubblicato nel 2005, è considerato lo standard ottimale per creare e mantenere un deposito digitale per un lungo periodo di tempo.

Il modello OAIS può essere applicato a vari archivi, ad esempio, ad accesso aperto, chiusi, limitati, "nascosti" o proprietari.

Le informazioni salvate necessitano di "conservazione a lungo termine", anche se l'OAIS stesso non è permanente. "Lungo termine" è un periodo sufficientemente lungo da essere influenzato dai cambiamenti tecnologici, compreso l'aggiornamento per nuovi supporti e formati di dati, o da una comunità di utenti che cambia. Il "lungo termine" può estendersi indefinitamente.

L'archivio definisce la comunità e tale definizione non è fissa.

La "O" in OAIS rappresenta il "modo aperto in cui lo standard è stato sviluppato" e non rappresenta l' "accesso aperto" o l'uso del termine aperto nella Open Definition o nella Open Archives Initiative. La "I" in OAIS sta per "informazione", ovvero dati che possono essere condivisi o scambiati.

Open Archival Information System

L'ambiente OAIS prevede l'interazione di quattro entità:

- produttori di informazioni;
- consumatori di informazioni (o la comunità designata);
- management (persona o gruppo che stabilisce le politiche per il contenuto dell'archivio);
- archivio.

Utenti e IA

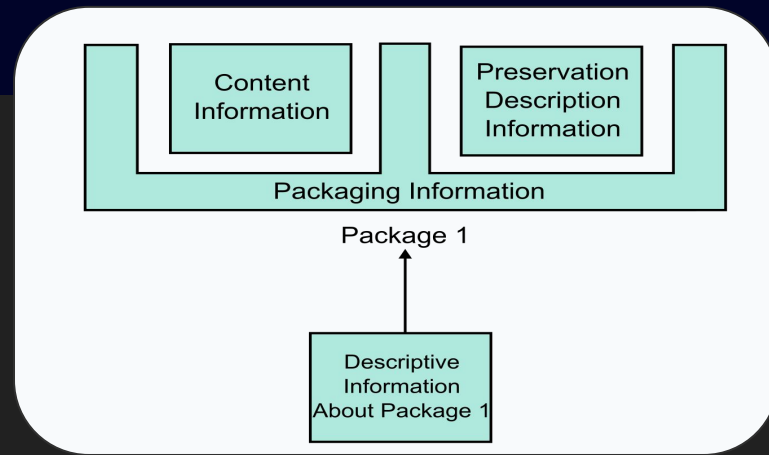


Open Archival Information System

Il modello OAIS definisce anche un modello di informazione. Gli elementi fisici o digitali che contengono informazioni sono noti come oggetti dati. I membri della Comunità Designata per un archivio dovrebbero essere in grado di interpretare e comprendere le informazioni contenute in un oggetto dati sia grazie alla loro base di conoscenze consolidata sia con l'assistenza di "informazioni di rappresentazione" supplementari incluse con l'oggetto dati.

Un pacchetto informativo comprende i seguenti oggetti informativi:

- **Informazioni sul Contenuto:** includono l'oggetto dati e le sue informazioni di rappresentazione
- **Informazioni sulla Descrizione della Conservazione:** contengono le informazioni necessarie per conservare le informazioni sul contenuto (come informazioni sulla provenienza dell'oggetto, identificatori unici, un Checksum o altri dati di autenticazione, ecc.)
- **Informazioni sul Pacchetto:** tengono insieme i componenti del pacchetto informativo
- **Informazioni Descrittive:** metadati sull'oggetto che consentono di localizzarlo in un momento successivo utilizzando le funzioni di ricerca o recupero dell'archivio

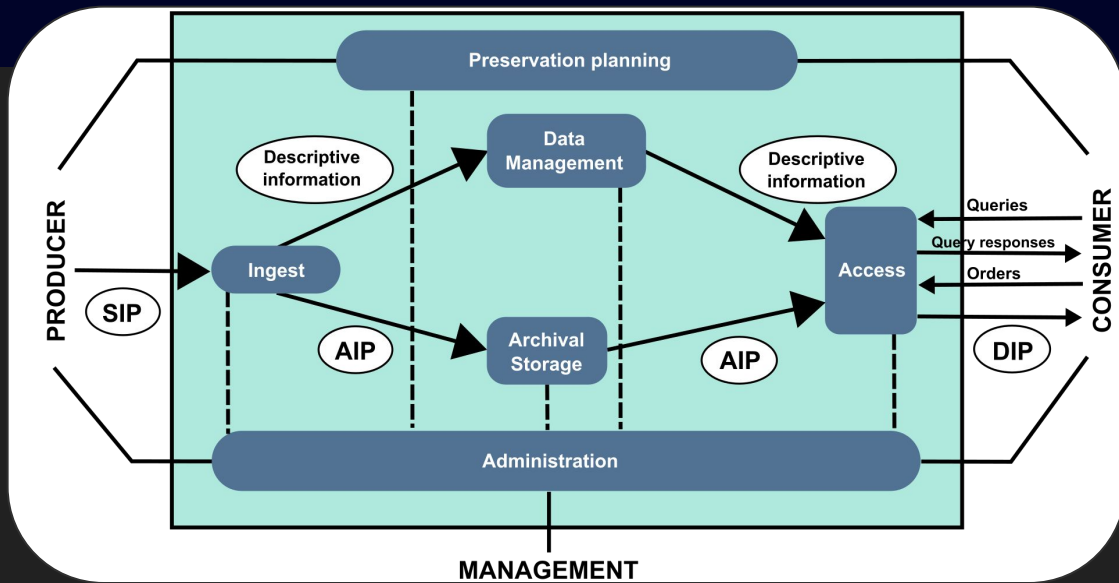


Open Archival Information System

Esistono tre tipi di pacchetti informativi nel modello di riferimento OAIS:

- ✓ Submission Information Package (SIP): l'informazione inviata dal produttore all'archivio
- ✓ Archival Information Package (AIP): l'informazione conservata dall'archivio
- ✓ Dissemination Information Package (DIP): l'informazione inviata a un utente su richiesta

Questi tre pacchetti informativi possono essere o meno identici tra loro.



Definizioni generali: Metadati

I metadati (o metainformazioni) sono "dati che forniscono informazioni su altri dati", ma non il contenuto dei dati stessi, come il testo di un messaggio o l'immagine stessa. Esistono molti tipi distinti di metadati, tra cui:

- **Metadati descrittivi** – le informazioni descrittive su una risorsa. Vengono utilizzati per la scoperta e l'identificazione. Includono elementi come titolo, abstract, autore e parole chiave.
- **Metadati strutturali** – metadati sui contenitori di dati che indicano come gli oggetti composti sono messi insieme, ad esempio, come le pagine sono ordinate per formare capitoli. Descrivono tipi, versioni, relazioni e altre caratteristiche dei materiali digitali.
- **Metadati amministrativi** – le informazioni per aiutare a gestire una risorsa, come il tipo di risorsa, i permessi e quando e come è stata creata.
- **Metadati di riferimento** – le informazioni sui contenuti e sulla qualità dei dati statistici.
- **Metadati statistici** – chiamati anche dati di processo, possono descrivere processi che raccolgono, elaborano o producono dati statistici.
- **Metadati legali** – forniscono informazioni sul creatore, sul detentore del copyright e sulla licenza pubblica, se fornita.

I metadati non sono strettamente legati a una di queste categorie, poiché possono descrivere un dato in molti altri modi.



Metadata

fv: Header of 3C279_region

File Edit Tools

Search for:

```
TFORM21 = 'E' / data format
TTYPE22 = DIFRSP4 / Diffuse response component
TFORM22 = E / data format
CHECKSUM= UAI0a39NU9GNa99N' / HDU checksum updated 2010-08-06T15:24:21
DATASUM= 3158868525' / data unit checksum updated 2010-08-06T15:24:21
TELESCOP= GLAST / name of telescope generating data
INSTRUME= LAT / name of instrument generating data
EQUINOX= 2000. / equinox for ra and dec
RADECSYS= FK5 / world coord. system for this file (FK5 or FK4)
DATE= 2010-08-06T15:22:36.0000' / file creation date
DATE-OBS= 2008-08-04T15:43:36.0000' / start date and time
DATE-END= 2009-08-04T15:59:58.0000' / end date and time
OBSERVER= Peter / GLAST/LAT PI
ORIGIN= LISOC / name of organization making data
EXTNAME= EVENTS / name of this binary table
HDUCLASS= OGIP / format conforms to OGIP standard
HDUCLAS1= EVENTS / extension contains events
HDUCLAS2= ALL / extension contains all events
TSTART= 239557417. / mission time of the start of observation
TSTOP= 255398400. / mission time of the end of observation
MJDREFI= 51910. / Integer part of MJD correction
MJDREFF= 7.428703703703703D-4 / Fractional part of MJD correction
TIMEUNIT= s / units for the time related to TSTART and TSTOP
TIMEZERO= 0. / clock correction
```

File Edit Tools

Help

Index

Extension

Type

Dimension

View

☐ 0

Primary

Image

0

Header

Image

Table

☐ 1

EVENTS

Binary

22 cols X 1102 rows

Header

Hist

Plot

All

Select

☐ 2

GTI

Binary

2 cols X 1 rows

Header

Hist

Plot

All

Select

KEY

VALUE

COMMENT

Columns. Each represents an attribute of the data.

Header: these are the names of the attributes.

Rows, or lines.
Each represents a data point

A23410						
1	A	B	C	D	E	F
2	id	civitate	particule	first name	name	maiden name
1	10997	M		William	Pruitt	
2	10998	F		Marian	Oconnor	
3	10999	M		Sammie	Robertson	
4	22529	M		Efren	Smith	
5	22528	M		Nigel	Simon	
6	22527	M		Bruce	Bowers	
7	22526	M		Chester	Hicks	
8	22525	M		Bernardo	Lott	
9	22524	F		Elisabeth	Nash	
10	22523	M		Kristopher	Stanton	
11	10990	M		Dennis	Sparks	
12	22522	M		Sean	Ewing	
13	10991	M		Cedrick	Hoffman	

A value.
(can be empty).

DataSet e Identificatori di Oggetti Digitali

L'identificatore di oggetti digitali (acronimo DOI) è uno standard che consente l'identificazione duratura e univoca di oggetti di qualsiasi tipo all'interno di una rete digitale e l'associazione ad essi dei relativi dati di riferimento - i metadati -, secondo uno schema strutturato ed estensibile.

Il DOI si differenzia dai comuni indicatori Internet, come gli URL, nell'identificare direttamente un oggetto, come entità di prima classe, e non semplicemente attraverso qualche attributo, come il luogo in cui l'oggetto si trova.

Il DOI si distingue anche dagli identificatori legati agli standard bibliografici (ISBN, ISSN, ISRC, ecc.), in quanto può essere immediatamente attivato online e utilizzato per lo sviluppo di servizi specifici quali motori di ricerca, certificazioni di autenticità, ecc.

Un identificatore DOI può essere registrato su oggetti di qualsiasi forma materiale (digitale o fisica) o su entità astratte (come opere testuali) quando vi è una necessità funzionale di distinguerli da altri oggetti.

Un oggetto può essere arbitrariamente identificato a qualsiasi livello di granularità. Ciò significa che, ad esempio, un DOI può essere registrato sul titolo di una rivista, sul suo singolo numero, sul singolo articolo di un dato numero, sulla singola tabella di un dato articolo.

Il problema della granularità non è ancora codificato, ma segue il senso razionale dell'opportunità di essere consumato dalla comunità di riferimento.

SEMANTICA: UCDS COME TRADUTTORI

L'IA non "capisce" i nomi delle colonne. Gli **Unified Content Descriptors (UCD)** sono essenziali:

phot.mag;em.opt.V indica univocamente una magnitudine ottica.

Permette il **Transfer Learning** tra cataloghi di missioni diverse.

Abilita l'integrazione automatica di ontologie semantiche.

Vantaggio IA

Riduzione del preprocessing manuale dell'80% attraverso il mapping automatico delle features fisiche.

L'IA COME "POWER USER" DELL'IVOA

Gli standard IVOA non servono solo agli astronomi, ma agli agenti computazionali:

TAP (Table Access Protocol): Query ADQL per estrarre training sets massivi.

SIAv2 / SSA: Accesso programmatico a immagini e spettri.

VOSpace: Archiviazione remota ad accesso federato.

DISCOVERY AUTOMATICA

Machine-Actionable Interfaces

Un sistema di AI deve poter scoprire i dati senza intervento umano tramite il **Registry IVOA**.

L'uso di interfacce VO-standard assicura che il dataset di addestramento sia scalabile da 10 a 10.000 sorgenti con la stessa riga di codice.

CREARE IL "GROUND TRUTH"

La preelaborazione dei dati è il primo passo in qualsiasi pipeline di analisi dei dati o di apprendimento automatico. Consiste nella pulizia, trasformazione e organizzazione dei dati grezzi per garantirne l'accuratezza, la coerenza e la predisposizione alla modellazione. Ha un impatto significativo sulla creazione dei modelli, ad esempio:

- Dati puliti e ben strutturati consentono ai modelli di apprendere pattern significativi anziché rumore.
- Dati elaborati correttamente prevengono input fuorvianti, portando a previsioni più affidabili.
- Dati organizzati semplificano la creazione di input utili per il modello, migliorandone le prestazioni.
- Dati organizzati supportano una migliore analisi esplorativa dei dati (EDA), rendendo pattern e tendenze più interpretabili.

SCIENCE PLATFORMS: CODE-TO-DATA

Nell'era del Big Data, scaricare i dati è impossibile. L'archivio diventa una piattaforma di calcolo:

- Notebook Jupyter pre-configurati con framework ML (TensorFlow, PyTorch).
- Accesso diretto a bucket S3/Cloud storage.
- **Parallelismo massivo:** GPU attigue allo storage dei dati.

Definizioni Generali - Archivi

- ✔ **Storage:** spazio fisico o digitale in cui vengono raccolti gli elementi. Lo storage digitale può essere ridondante per scopi di conservazione.
- ✔ **Repository:** storage digitale che offre un modo per archiviare, gestire e recuperare informazioni in modo organizzato. Ogni utente organizza i dati da solo/a.
- ✔ **Archivio:** repository gestito in modo tale che le capacità di ricerca siano notevolmente potenziate, sia presente una ricca descrizione dei contenuti e gli utenti siano in grado di recuperare i dati dopo averli filtrati tramite applicazioni informatiche. In INAF esistono diversi archivi scientifici:
 - Archivi degli Osservatori Astrofisici;
 - Archivi di prodotti scientifici di alto livello;
 - Archivi di dati simulati;
 - Archivi di Survey;
 - Cataloghi
- ✔ **Science Gateways / Platforms:** Insieme di servizi offerti per l'accesso, la riduzione, l'esplorazione, l'analisi e la condivisione dei risultati. Possono prevedere l'interazione con AI (chatbot o LLM customizzati per ambito di ricerca). Un esempio è EVA (ESA Virtual Assistant)....

EVA

L'ESA (Agenzia Spaziale Europea) ha sviluppato e implementato un sistema di chatbot chiamato EVA (ESA Virtual Assistant). EVA è una piattaforma basata sull'intelligenza artificiale progettata per facilitare l'interazione con i vari portali scientifici dell'ESA tramite linguaggio naturale (testo o voce).

- ESASky: È integrato nell'interfaccia di esplorazione astronomica ESASky. Permette di cercare oggetti celesti, navigare nella mappa del cielo e accedere rapidamente alle informazioni sulle missioni senza usare il mouse.
- ESA Cosmos: Supporta gli utenti nel portale principale delle missioni scientifiche (Cosmos), aiutando a trovare documentazione, manuali e FAQ in tempo reale.
- Supporto Tecnico (SOCCI): Funziona come un ingegnere di helpdesk virtuale per fornire risposte istantanee a problemi tecnici comuni.
- Osservazione della Terra: L'ESA ha recentemente lanciato o sta sviluppando assistenti digitali specifici (come PhilEO o prototipi di "Digital Assistant") per interrogare archivi complessi come quelli di Sentinel-1 e 2 usando domande semplici.

CONCLUSIONI: L'ORIZZONTE DELL'EXASCALE

~600 PB

SKA PHASE 1 / ANNO

~20 TB

RUBIN LSST / NOTTE

Il Futuro è Data-Centric

-  **ML & Deep Learning:** Non sono più solo strumenti, ma i "filtri cognitivi" indispensabili per gestire facility che operano oltre la scala umana.
-  **Metadattazione e DMP:** Senza una curatela semantica profonda (IVOA), l'oceano di dati dell'Exascale rimarrebbe un rumore inaccessibile.
-  **Science Platforms:** L'archivio diventa il laboratorio virtuale dove la conoscenza viene estratta *in-situ*, trasformando il dato in scoperta scientifica di ampio spettro.

Thanks for attention!

Questions?