



# Vocabulary and DMP in practice

Cristina Knapic

# General definitions

## Collection:

collection of objects gathered together according to strict selection criteria and intended for conservation for their intrinsic value or interest or for simple passion or personal pastime.

(publishing) series of literary works by a given author or on a given topic, theme or genre published by a publishing house

(fashion) the set of sartorial creations of a given designer or fashion house relating to a specific season

The collection is a group of resources that are related to each other in some identifiable way. The relationship might be through a topic, a place, a person etc.. The relation depends on the human logic and in astrophysics usually depends on field of investigation (cosmology, stellar physics, transients...), instrument characteristics and scientific purpose.

A collection is an aggregation of physical or digital items i.e. library or museum collections; library, museum or archive catalogues; digital archives; internet directories; internet subject gateways; collection of texts, images, sounds, datasets, software etc...

A collection may be made up of any number of items, from one to many.

A collection is a set of resources brought together for a particular audience or to serve a specific function.

# General definitions

A Collection is a growing organism: in particular in Astrophysics, all the items related to observations, simulations, ancillary information, calibrations, software for calibration and data analysis, visualization etc... might increase in number and type.

Three concepts:

- A collection is a single entity
- A collection is an entity that exists to serve a specific mission
- Several collections should be a unique part of a more larger collection (see lesson on Interoperability)

# General Definitions

Dataset:

A data set or dataset is a collection of data.

More commonly, a dataset constitutes a set of data structured in relational form, that corresponds to the content of a single database table, or to a single matrix of statistical data, in which each column of the table represents a particular variable, and each row corresponds to a specific member of the dataset in question. The size of the dataset is given by the number of members present, which form the rows, and by the number of variables of which it is composed, which form the columns.

**The term dataset can also be used more generically to indicate data in a set of closely connected items relating to a particular experiment or event.**

# General definitions

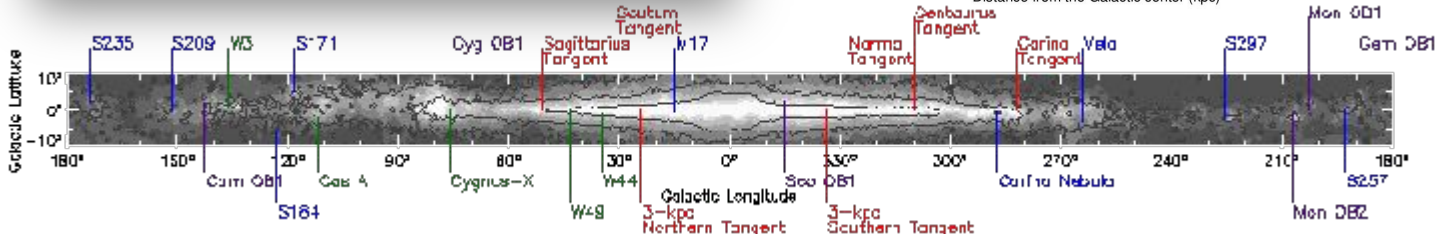
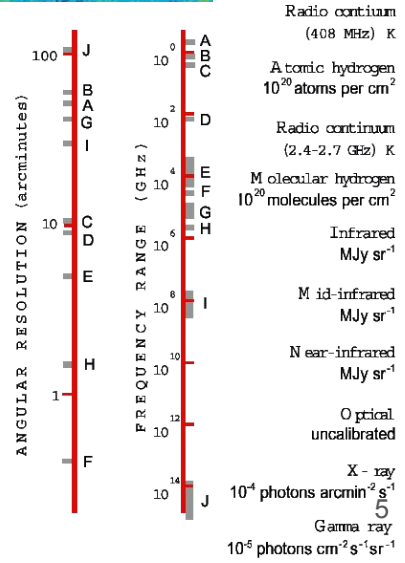
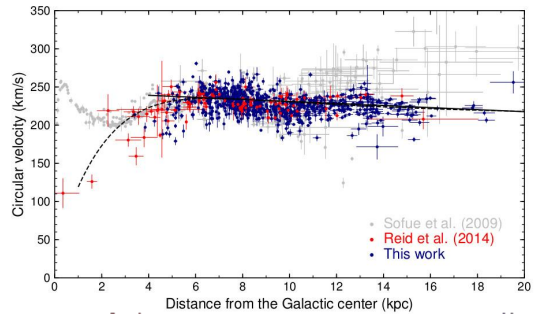
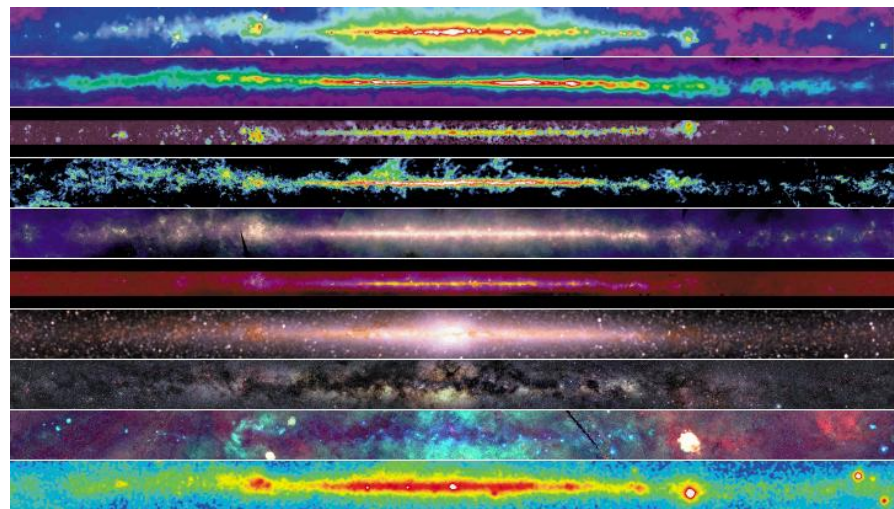
## Dataset

```

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> 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>
    
```



# General definitions

## What a dataset might look like

### Observations



24 Oct 2022

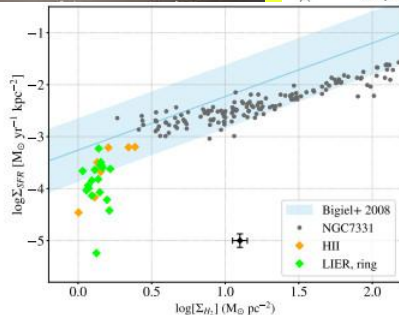


Figure 14. Relation between star formation rate surface density,  $\Sigma_{\text{SFR}}$ , and stellar mass surface density,  $\Sigma_{\text{star}}$ , for various galaxy environments with blue shading representing the  $3\sigma$  error on the slope. Data for NGC 7331 (Sutter & Padda 2022) are marked with grey dots. Median errorbars for the data from M104 (orange and green diamonds) are shown as a black cross in the lower part of the plot.

### Paper

DRAFT VERSION OCTOBER 26, 2022  
Typeset using L<sup>A</sup>T<sub>E</sub>X two-column style in AASTeX.

#### A Molecular Gas Ring Hidden in the Sombrero Galaxy

JESSICA SUTTER<sup>1</sup> AND DARIO PADDA<sup>1</sup>

<sup>1</sup>SOFIA Science Center, USRA, NASA Ames Research Center, M.S. 222P-12 Moffett Field, CA 94035, USA

#### ABSTRACT

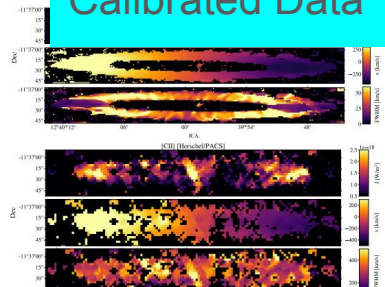
We present Herschel, ALMA, and MUSE observations of the molecular ring of Messier 104, also known as the Sombrero galaxy. These previously unpublished archival data shed new light on the content of the interstellar medium of M104. In particular, molecular hydrogen measured by CO emission and dust measured by far-infrared light are uniformly distributed along the ring. The ionized gas revealed by H $\alpha$  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 $\alpha$  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  $0.9 \times 10^6 M_{\odot}$ , a value intermediate between those of early type galaxies and the molecular ring of our galaxy.

INTRODUCTION  
Complex structures that can be shaped tors. Each galaxy has its own unique which can lead to the creation of spirals, or even a lack of structure all together to the distinct structural nature of observing perspective can have significant measurable properties of a galaxy. A k galaxy may appear devoid of optimization from our viewing angle, but a different picture if viewed face-on. Pieceations of a single galaxy across the extrum provides the context to discern structures and determine how viewing ally to distinguish galaxy character-

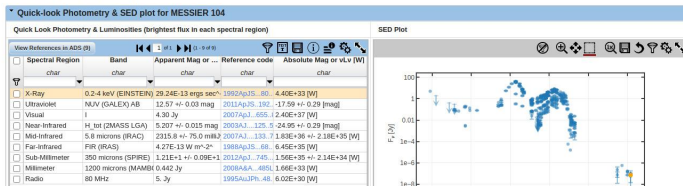
One particularly interesting galaxy for st unique histories and observing perspective certain diagnostics is the nearby galaxy 1 commonly referred to as the ‘Sombrero’ galaxy. general properties of this galaxy are summarized in Table 1. M 104 has been classified as a SAa type there has been some debate about the exact this galaxy (see e.g. Krause et al. 2006, an thesis). 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 ties similar to those of an elliptical galaxy (T 2006) and measurements of star-formation and dust temperatures match more closely type galaxies than similar nearby spirals. decomposition of the structures observed at that M 104 does not follow classical bulge/d and instead is better represented as an ellip with a large halo (Gadotti & Sánchez-Jans M 104 is of further interest due to the m of observational evidence indicating it hos

J. Jessica Sutter (moving to UCSD)  
Leon

### Calibrated Data



#### Photometry & SED for MESSIER 104



### Comparison

No.	Observed Passband	Photometry Measurement	Uncertainty	Units	Frequency	Flux Density	Upper limit...	Lower limit...	NED Uncertainty	NED Units
fit	char	double	char	char	double	double	double	double	char	-/yr
1	2-10 keV (Chandra)	-11.99 +/- 0.05		logerg/cm	1.45e+18	7.06e-8	3.25e-9	3.25e-9	+/-0.23E-09	yr
2	2-10 keV (Chandra)	8.02e-11 +/- 7.22E-13		erg/cm	1.45e+18	5.52e-8	4.98e-8	4.98e-8	+/-4.88E-08	yr
3	2-10 keV (MMM)	1.359e-12 +/- 11.556E-13		erg/cm	1.45e+18	9.37e-8	7.97e-8	7.97e-8	+/-7.97E-08	yr
4	2-10 keV (XMM)	1.39e-12		erg/cm	1.45e+18	9.59e-8				yr
5	0.3-8 keV (Chandra)	1.2e-12		erg/cm	1e+18	1.3e-7				yr
6	0.3-8 keV (Chandra)	6.987e-13 +/- 0.98E-14		erg/cm	1e+18	6.99e-8	9.8e-10	9.8e-10	+/-9.80E-10	yr
7	0.3-8 keV (Chandra)	1.60e-12		erg/cm	1e+18	1.60e-7				yr
8	0.3-8 keV (Chandra)	1.55e-12		erg/cm	1e+18	1.56e-7				yr
9	0.2-4 keV (XMM)	2.042e-12		ergs sec	5.25e+17	5.57e-7				yr
10	0.5-2 keV (Chandra)	2.25e-13 +/- 1.90E-13		erg/cm	3.02e+17	7.45e-8	6.29e-8	6.29e-8	+/-6.29E-08	yr
11	0.5-8 keV (XMM)	6.79e-13 +/- 6.46E-13		erg/cm	3.02e+17	2.24e-7	1.81e-7	1.81e-7	+/-1.81E-07	yr
12	1482A (RUE)	4.6e-15 +/- 0.27E-14		ergs cm	2.02e+15	3.37e-4	1.98e-4	1.98e-4	+/-1.98E-04	yr

Table 4. Bands used for SED Fitting

Filter	Wavelength	Beam	Pixel	$\sigma_{\text{cal}}$	Refs
	$\mu\text{m}$	arcsec	arcsec		
GALEX-FUV	0.152	4.2	1.5	0.05 mag	1
GALEX-NUV	0.227	5.3	1.5	0.03 mag	1
SDSS-g	0.354	1.4	0.4	2%	2

### Boundary Conditions

ZMASS_L	1.245	2.9	2.0	0.03 mag	3
ZMASS_H	1.662	2.8	2.0	0.03 mag	3
ZMASS_Ks	2.159	2.9	2.0	0.03 mag	3
IRAC_1	3.550	1.66	1.2	1.8%	4
IRAC_2	4.490	1.72	1.2	1.9%	4
IRAC_3	5.730	1.88	1.2	2.0%	4
IRAC_4	7.870	1.98	1.2	2.1%	4
WISE_3	11.56	6.5	2.75	4.5%	5
MIPS_24	23.70	4.9	2.5	4.0%	6
PACS_70	71.11	5.6	3.2	5%	7
---	---	---	---	3.2	5% 7
---	---	---	---	6.4	5% 7

SS are median seeing values. (7), (2) Padmanabhan et al. (4) Reach et al. (2005), (5) et al. (2007), (7) Balog et al.

# Open Archival Information System

The OAIS refers to the ISO OAIS Reference Model [1]. This reference model is defined by recommendation CCSDS 650.0-B-2 of the Consultative Committee for Space Data Systems. The CCSDS's purview is space agencies, but the OAIS model it developed has proved useful to other organizations and institutions with digital archiving needs. OAIS is widely accepted and utilized by various organizations and disciplines, both national and international, and was designed to ensure preservation. The OAIS standard, published in 2005, is considered the optimum standard to create and maintain a digital repository over a long period of time.

The OAIS model can be applied to various archives, e.g., open access, closed, restricted, "dark", or proprietary.

The information being maintained has been deemed to need "long term preservation", even if the OAIS itself is not permanent. "Long term" is long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. "Long term" may extend indefinitely.

The archive defines the community and that definition is not fixed.

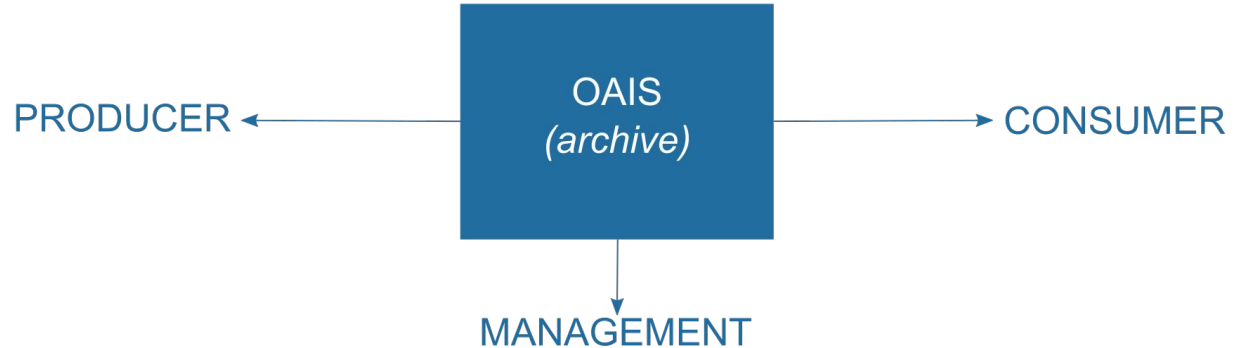
The "O" in OAIS represents the "open way the standard was developed", and does not represent "open access", or the usage of the term open in the Open Definition or Open Archives Initiative. The "I" in OAIS represents "information", meaning data that can be shared or exchanged.

[1]-<https://public.ccsds.org/pubs/650x0m2.pdf>

# Open Archival Information System

The OAIS environment involves the interaction of four entities:

- producers of information;
- consumers of information (or the designated community);
- management (person or group that sets policies for the content contained in the archive);
- archive.



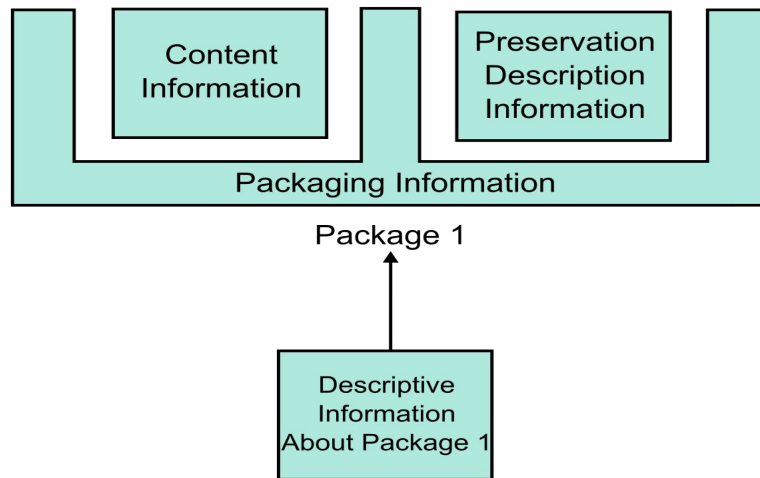


# Open Archival Information System

The OAIS model also defines an information model. Physical or digital items which contain information are known as data objects. Members of the Designated Community for an archive should be able to interpret and understand the information contained in a data object either because of their established knowledge base or with the assistance of supplementary "representation information" that is included with the data object.

An information package includes the following information objects:

- Content Information: this includes the data object and its representation information
- Preservation Description Information: contains information necessary to preserve its affiliated content information (such as information about the item's provenance, unique identifiers, a Checksum or other authentication data, etc.)
- Packaging Information: holds the components of the information package together
- Descriptive Information: metadata about the object which allows the object to be located at a later time using the archive's search or retrieval functions

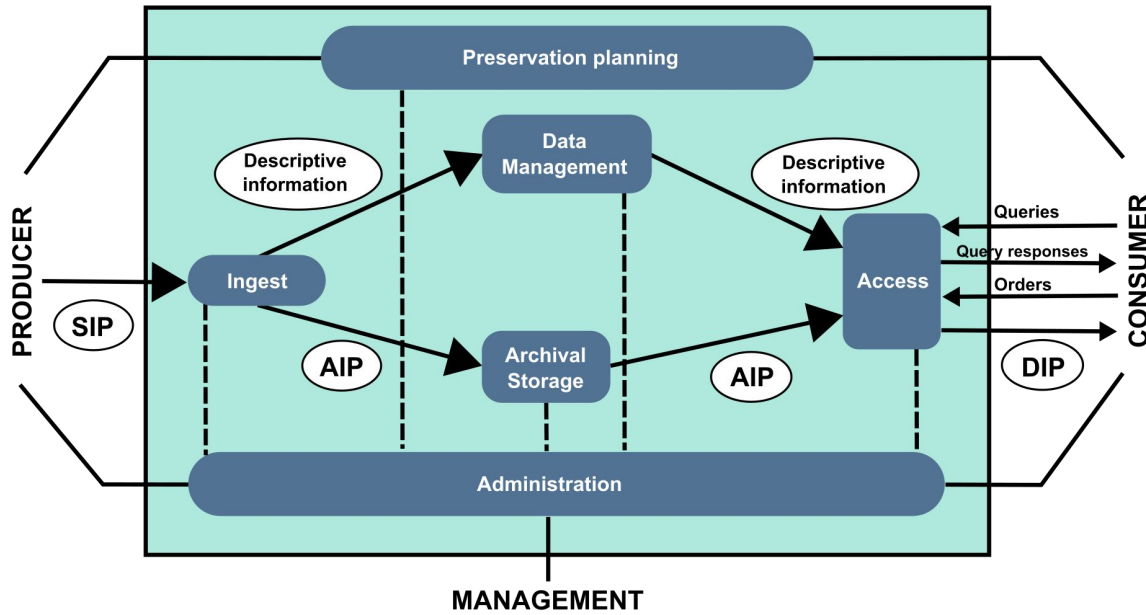


# Open Archival Information System

There are three types of information package in the OAIS reference model:

- Submission Information Package (SIP): which is the information sent from the producer to the archive
- Archival Information Package (AIP): which is the information stored by the archive
- Dissemination Information Package (DIP): which is the information sent to a user when requested

These three information packages may or may not be identical to each other.



# Open Archival Information System

There are six functional entities in an OAIS:

- Ingest function: receives information from producers and packages it for storage. It accepts a SIP, verifies it, creates an AIP from the SIP, and transfers the newly created AIP to archival storage;
- Archival Storage function: stores, maintains, and retrieves AIPs. It accepts AIPs submitted from the Ingest function, assigns them to long term storage, migrates AIPs as needed, checks for errors, and provides requested AIPs to the Access function;
- Data Management function: coordinates the Descriptive Information of the AIPs and the system information that supports the archive. It maintains the database that contains the archive's information by executing query requests and generating results, reports in support of other functions and updating the database;
- Administration function: manages the daily operations of the archive. This function attains submission agreements from information producers, performs system engineering, audits SIPs to ensure compliance with submission agreements, develops policies and standards. It handles customer service and acts as the interface between Management and the Designated Community in the OAIS environment;
- Preservation Planning function: supports all tasks to keep the archive material accessible and understandable over long terms even if the original computing system becomes obsolete, e.g. development of detailed preservation/ migration plans, technology watch, evaluation and risk analysis of content and recommendation of update and migration.
- Access function: This function includes the user interface that allows users to retrieve information from the archive. It generates a DIP from the relevant AIP and delivers it to the customer who has requested the information

# D-Identity, Users, Groups, Organizations

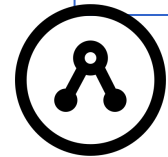
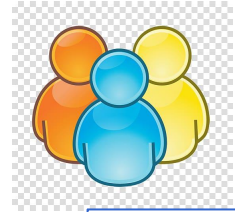
**Digital Identity:** is the set of digital resources uniquely associated with a natural person that identifies him, representing his will, during his digital activities;

**User :** person or device that uses data processing systems to obtain or process data and to exchange information;

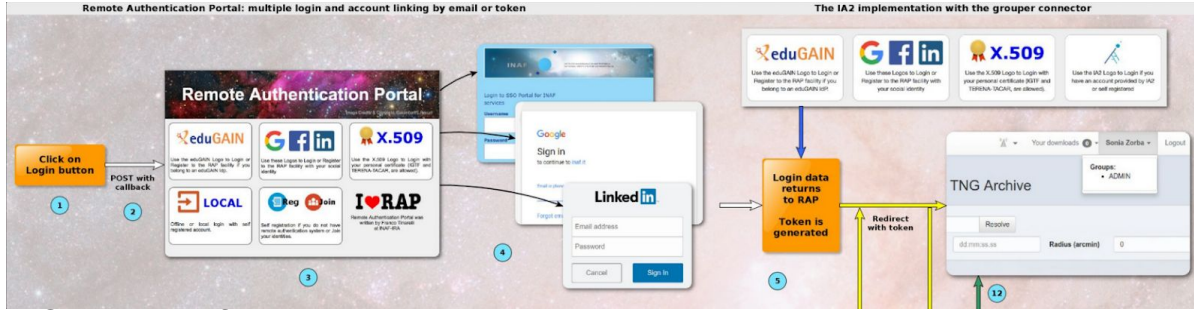
**Group:** a set of people who have similar interests, goals or concerns (i.e. transients, exoplanets, cosmology...);

**Group of groups:** several set of people who have similar interests and are part of a more extensive group (national observers at a telescope);

**Organization:** several set of groups that share an high level topic like a discipline or a scientific domain (i.e. INAF, SKA, CTA, CERN, INFN...);



# Remote Authentication Portal



Courtesy of F. Tinarelli



The same person,  
multiple identities,  
multiple user ID



One user,  
multiple identities,  
the same user ID



**ORCID**  
Connecting Research  
and Researchers

# Permissions and Roles

**Permissions:** define the actions that a user can perform in a community application (i.e. can write, read, delete, create, can execute digital items ,can see metadata, can retrieve data....);

**Roles:** provide a way for community administrators to group permissions and assign them to users or user groups (i.e. an administrator can do everything on an archive like store,delete, move, modify data; a PI user can retrieve only his/her own or public data **and** can execute routines on accessed data; a user can access only public data).

Roles definition as well as sharing necessities identification with the use of groups have particular relevance in data handling and in archive set up.

# Group Membership Service

The screenshot shows a web browser at the URL `sso.ia2.inaf.it/gms/#/`. The page title is "Group Membership Service". The breadcrumb navigation shows "ROOT / LBT / INAF", with "LBT" and "INAF" circled in red and purple respectively. The main content area has tabs for "Groups", "Members", and "Permissions", with "Groups" selected. A search bar is present above a list of groups. The first group in the list, "AO\_2021\_2022\_6", is circled in green. Below the list is a pagination control showing page 1 of 1, and a "Page size" dropdown set to 5. The total number of items is 687.

Archive (V-O)

Group of groups

Group

# Group Membership System

Group Membership Service

Help Search Cristina Knapic

Cristina Knapic (Google, LinkedIn, eduGAIN, IA2, Facebook): [Back](#)

Is member of

- LBT / INAF / test2\_cristina
- ROOT
- VOSpace / test1
- people / cristina.knapic

User info

User id: 2388

Identities (7):

Type	eduGAIN
Email	cristina.knapic@inaf.it
EPPN	cristina.knapic@inaf.it
Type	Google
Email	cristina.knapic@gmail.com
Type	LinkedIn
Email	cristina.knapic@linkedin.com
Type	eduGAIN
Email	cristina_k@inaf.it
EPPN	cristina.knapic@inaf.it
Type	Google
Email	cristina.knapic@gmail.com
Type	LinkedIn

Permissions	
Group	Permission
LBT / INAF / test2_cristina	MANAGE_MEMBERS
ROOT	ADMIN
VOSpace / test1	MANAGE_MEMBERS
people / cristina.knapic	VIEW_MEMBERS

Add permission

Search: knapic

Selected user: Cristina Knapic (Google, LinkedIn, eduGAIN, IA2, Facebook) [2388]

Permission:

Admin  Manage members  View members

[Cancel](#) [Add](#)

Group Membership Service

ROOT

Search group

group1

[Add collaborator](#)

Claudio Gheller (Google)

Add member

Search: knap

Selected user: Cristina Knapic (Google, LinkedIn, eduGAIN, IA2, Facebook) [2388]

[Cancel](#) [Add](#)



# Privacy policy

In European Public Research Institutes, all products of science research should be public since they are financed by government funds, in the perspective of applying the Open Science concepts (see E. Giglia talk [2]).

In INAF, for observational data, this is true for calibration data intake and data of the previous years, with one -or more in case of long observing programs- year of embargo in order to guarantee a reserved time of exclusive access to the PI and his/her group of collaborators for publishing paper/s.

Open Science foreseen the embargo period but strongly suggest the shortest duration as possible.

Moreover privacy policy can interest different part of the archive:

- **partial or full metadata description:**
  - only selected users can see full metadata body (admin or PI);
  - privileged users can see a subset of metadata (PI collaborators);
  - public users can see a basic subset of metadata (user not part of the collaboration);
- **partial or full data sets:**
  - only collaboration users can access and download scientific data;
  - public user can download only calibration data;
- **partial or full collection:**
  - users involved in an V-Organization can access only partially the collection both in metadata and data access;
  - collection administrator can access full collection;
  - system admin can do everything.

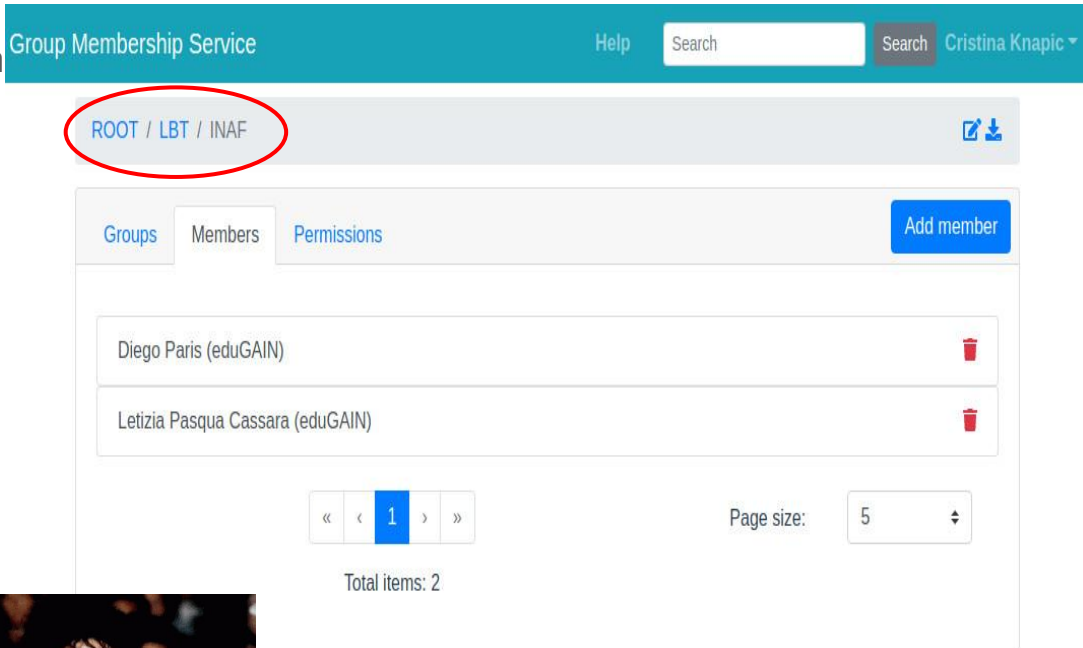
Open Science states full metadata body should be public

# Privacy policy

Moreover, the privacy policy interests to different number of users, groups, groups of groups and (v-)organizations with different member roles. As previously mentioned, some types of permissions define a role, and different roles can be applied to aggregation of users.

Example:

- public user can see only metadata but can't download data;
- members of a group of researchers can see and download data;
- PI of observations can see, download data and assign privileges to other users;
- administrator of the archive/group of groups can see and manage all permissions;
- system administrator can do everything including delete the archive!!



The screenshot shows the 'Group Membership Service' interface. At the top, there is a teal header with 'Group Membership Service', 'Help', a search bar, and a user profile 'Cristina Knapic'. Below the header, the breadcrumb 'ROOT / LBT / INAF' is circled in red. The main content area has tabs for 'Groups', 'Members', and 'Permissions', with 'Members' selected. An 'Add member' button is in the top right. The member list contains two entries: 'Diego Paris (eduGAIN)' and 'Letizia Pasqua Cassara (eduGAIN)', each with a red trash icon. A pagination control shows '1' in a blue box, and a 'Page size: 5' dropdown. Below the list, it says 'Total items: 2'. At the bottom right, there is a logo for 'IA2' and the text 'Powered by IA2'.



# Data

## Definition:

- originally Data is plural for “datum”, a Latin word;
- a “datum” is a single factual, a single entity, a single point of matter;
- Datums are most often called “data points”;
- Data represent a collection of data points;
- Data contains the scientific content, the research topic (i.e. a image, a spectrum, a sensor output) and often are described by other information, called metadata

# MetaData

Metadata (or metainformation) is "data that provides information about other data", but not the content of the data itself, such as the text of a message or the image itself. There are many distinct types of metadata, including:

- **Descriptive metadata** – the descriptive information about a resource. It is used for discovery and identification. It includes elements such as title, abstract, author, and keywords.
- **Structural metadata** – metadata about containers of data and indicates how compound objects are put together, for example, how pages are ordered to form chapters. It describes the types, versions, relationships, and other characteristics of digital materials.
- **Administrative metadata** – the information to help manage a resource, like resource type, permissions, and when and how it was created.
- **Reference metadata** – the information about the contents and quality of statistical data.
- **Statistical metadata** – also called process data, may describe processes that collect, process, or produce statistical data.
- **Legal metadata** – provides information about the creator, copyright holder, and public licensing, if provided.

Metadata is not strictly bound to one of these categories, as it can describe a piece of data in many other ways.



# MetaData

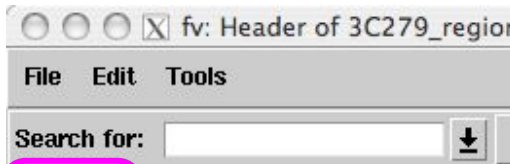
- Metadata: this is some data describing some other data.
- Example:
  - The bibliographical reference describing a book.  

*The metadata*
*The data*
  - Key takeaway: data without metadata can be worthless  
 -> What would you do with a pile of 10,000 books without any indication on their title, authors, or date of publication?
  - The difference between data and metadata is not always relevant  
 -> In the alumni network dataset, what is data and what is metadata?

This textual data is in digital form (because it is stored in bits on a computer, not by hand writing on a piece of paper)	(as opposed to analog).
The tweet is textual (as opposed to numerical. In programming, text can also be called a <code>String</code> )	this is the <b>type</b> (or format) of the data
The tweet appears plain text "plain text" is one sort of format for text. Others formats are <code>JSON</code> , <code>XML</code> or <code>CSV</code> !	this is the <b>format</b> of the data
The text of the tweet is encoded in UTF-8	this is the <b>encoding</b> of the data
The tweet is part of a list of tweets I collected	this is the <b>data structure</b>
The tweet is stored in a Word file on my laptop	this is the <b>format</b> of the data

Notice the ambiguity in the terminology!

# Metadata



```
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= 2009-08-05T15:59:58.0000' / 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 this table
TIMEZERO= 0. / clock correction
```

Index	Extension	Type	Dimension	View
<input type="checkbox"/> 0	Primary	Image	0	Header Image Table
<input type="checkbox"/> 1	EVENTS	Binary	22 cols X 1102 rows	Header Hist Plot All Select
<input type="checkbox"/> 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

**A value.** (can be empty).

id	civilitate	particule	first name	name	maiden name	year of birth
10997	M		William	Pruitt		unknown
10998	F		Marian	Oconnor		unknown
10999	M		Sammie	Robertson		unknown
22529	M		Efren	Smith		1970
22528	M		Nigel	Simon		unknown
22527	M		Bruce	Bowers		unknown
22526	M		Chester	Hicks		1987
22525	M		Bernardo	Lott		unknown
22524	F		Elisabeth	Nash		unknown
22523	M		Kristopher	Stanton		unknown
10990	M		Dennis	Sparks		1989
22522	M		Sean	Ewing		1950
10991	M		Cedrick	Hoffman		1983

# DataSet and Digital Object Identifiers

The digital object identifier (acronym DOI) is a standard that allows the lasting and unique identification of objects of any type within a digital network and the association with them of the relevant reference data - the metadata -, according to a structured and extensible scheme.

The DOI differs from common Internet indicators, such as URLs, in identifying an object directly, as a first-class entity, and not simply through some attribute, such as the place where the object is located.

The DOI is also distinguished from identifiers such as those linked to bibliographic standards (ISBN, ISSN, ISRC, etc.), as it can be immediately activated online and used for the development of specific services such as search engines, certifications of authenticity, etc.

A DOI identifier can be recorded on objects of any material form (digital or physical) or on abstract entities (such as textual works) when there is a functional need to distinguish them from other objects.

An object can be arbitrarily identified at any level of granularity. This means that, for example, a DOI can be registered on the title of a journal, on its single issue, on the single article of a given issue, on the single table of a given article

The problem of granularity is not still codified, but follows the rational sense of opportunity to be consumed by the reference community.

# DataSet and Digital Object Identifiers

In INAF there are two ways to create a Unique Identifier for data:

- DOI request
  - <https://www.ict.inaf.it/index.php/ict-inaf/doi>;
  - there is no limit in space;
  - the data are checked before DOI registration, so useful suggestions can be given on data structure and granularity;
- Open Access handle for datasets request
  - <https://openaccess-info.inaf.it/oa-in-inaf>
  - in case of software there is the possibility to attach user manuals or documentations;
  - in the case of databases it is possible to attach user manuals or other supporting material;
  - for what concern Vizier, a read.me file can be uploaded with the link and indication of using Vizier;
  - for other systems, add e read.me file with the link to dataset.



# General Definitions

**Storage:** physical or digital space where items are collected. Digital storage can be redundant for preservation purposes.

**Repository:** digital storage that offer a way to store, manage and retrieve information in a organized manner. Each user organizes data by her/him selves. In INAF institutional repositories are:

- OwnCloud (<https://owncloud.ia2.inaf.it/index.php/login>)
- VOSpace (<http://vospace.ia2.inaf.it/ui/>)
- Google Drive (<https://drive.google.com>)

**Archive:** repository managed in a way the search capabilities are well enhanced, there is a reach description of content, users are able to retrieve data after filtering them via computer applications. In INAF there are several scientific archives:

- Astrophysical Observatory Archives;
- Scientific high level products Archives;
- Simulated Data Archives;
- Survey Archives;
- Catalogues
- ....

# Archives, Repositories and storage space

As defined in OAIS, an archive is a complex infrastructure, growing and involving four main actors:

- data providers;
- user community;
- management;
- archive itself.

Data providers, as the term suggests, are the entities that produce and procure data (scientific items, content and their description). Moreover Data Providers suggest the important relationship between single datum (file, tuple), component and subcomponents relevant for successive data handling, calibration, analysis, publication and sharing policies.

Data Provider role is crucial and fundamental for a good understanding of scientific products. In DP duties are:

- definition of the Data Model (see A. Bignamini session);
- definition of the whole services:
  - storage space required;
  - type of access (web, ssh, both, email request..)
  - preservation mechanisms: depending on the data (custom or standard) a policy to prevent the obsolescence has to be foreseen;
  - sharing policies (who can access the collection? With which roles?);
  - privacy policy: definition of the level of privacy in time and related to showed meta/data information;
  - number of accesses foreseen on daily bases;
  - typical dimension of one dataset and how many dataset are reasonably downloaded at time (how many files do compose each observation?);
  - format conversion;
  - responsibilities on data integrity;
  - backup policies;
  - services to solve inconsistencies issues;

# Preservation in INAF

INAF - Italian Astronomical Archives (IA2) facility manage all archives and internal data repositories:

- Open Access (<https://openaccess-info.inaf.it>);
- OwnCloud (<https://owncloud.ia2.inaf.it>);
- VOSpace (<http://vospace.ia2.inaf.it>);

Moreover, IA2 hosts temporary and permanent repositories and archives in on-line (1.6PB) and cold storage system (2PB).

The IA2 cold storage system is used to store and preserve checked data and to provide a permanent storage for DOIs. The cold storage system is a tape library IBM TS4500 equipped with a IBM Spectrum Archive control system (under upgrade to Spectrum Protect for backup purposes).

The main features of this system are:

- long persistent storage;
- low power consumption for non frequently used data;
- **big data storage**;
- unique filesystem for each V-Organization;
- cheapest solution;
- write once, read more;

Drawback:

- difficulties on data update, remove or reorganize;



# Data Management Plan

Why develop a data plan?

There are many benefits to managing and sharing your data:

- you can find and understand your data when you need to use it
- there is continuity if project staff leave or new researchers join
- you can avoid unnecessary duplication e.g. re-collecting or re-working data
- the data underlying publications are maintained, allowing for validation of results
- data sharing leads to more collaboration and advances research
- your research is more visible and has greater impact
- other researchers can cite your data so you gain credit

Funders expect data plans to outline how data will be created, managed, shared and preserved, justifying any restrictions that need to be applied. The plans are an opportunity to demonstrate your awareness of good practice and reassure funders that your proposal is in line with their data policy.

# DMP and metadata

While data curators, and increasingly researchers, know that good metadata is key for research data access and re-use, figuring out precisely what metadata to capture and how to capture it is a complex task. Fortunately, many academic disciplines have supported initiatives to formalise the metadata specifications the community deems to be required for data re-use.

In particular tomorrow there will held two specific presentations on Data Models and Interoperability standards, so to have a strong base and references on already developed work on Astrophysic (and more).

Let's proceed: what are the FAQ about DMP?

# 1. Data Types, Formats, Standards and Capture Methods

## Questions

- What data outputs will your research generate?
  - outline volume, type, content, quality and format of the final dataset
    - useful to capture your data in (or convert it to) community-accepted data formats;
    - Open or non-proprietary formats are preferable
- Outline the metadata, documentation or other supporting material that should accompany the data for it to be interpreted correctly
  - allow your data to be understood and discovered by others using a rich set of metadata and contextual details;
- What standards and methodologies will be utilised for data collection and management?
  - standards defined on disciplines base;
  - help in optimize existing standard;
  - refer to RDA for exhaustive list of multidisciplinary scientific standards;
- State the relationship to other data available in public repositories e.g.
  - existing data sources that will be used by the research project
  - gaps between available data and that required for the research
  - the added value that new data would provide in relation to existing data

## 2. Ethics and Intellectual Property

### Questions:

- Demonstrate that you have sought advice on and addressed all copyright and rights management issues that apply to the resource;
  - note that INAF data is property of the institute and not of the PIs;
- Make explicit mention of consent, confidentiality, anonymisation and other ethical considerations, where appropriate
  - astrophysical data do not need anonymization or other ethical considerations;
  - GDPR is not necessary in astrophysical data like is for example in other disciplines;
- Are any restrictions on data sharing required – for example to safeguard research participants or to gain appropriate intellectual property protection?
  - embargo for PI data confidentiality is foreseen and last one year;
  - data sharing have to be formalized in order to create the appropriate groups in GMS;
  - restrictions on data (and also metadata) should be defined beforehand.

# 3. Access, Data Sharing and Reuse

- What are the further intended and/or foreseeable research uses for the completed dataset(s)?
  - brief description of the scientific goal and research purposes;
  - plan for data reuse;
- How you will make the resource accessible to the potential audience(s) identified.
  - Where will you make the data available?
    - web portal;
    - private repository;
    - .....
  - How will other researchers be able to access the data?
    - authorization and permissions - > roles!
  - Will a data sharing agreement be required? -> licencing
  - What is the timescale for public release of the data? -> publication and reuse
  - Using existing infrastructure avoid to reinvent the wheel (standards, formats, hardware..)
- State any expected difficulties in data sharing, along with causes and possible measures to overcome these difficulties.
- How will data sharing provide opportunities for coordination or collaboration?
  - sharing between collaborators;
  - sharing with community members;
  - sharing with all: citizen science, open science...



## 4. Short-Term Storage and Data Management

- Describe the planned quality assurance and back-up procedures (security/storage)
  - data integrity checking;
  - data recovery;
  - Define data management support: outline what provision is available to you within your institution and any additional skills or resources that you need to secure;
- Specify the responsibilities for data management and curation within research teams at all participating institutions
  - Be clear about who will be responsible for different tasks
  - Strong file-naming conventions and versioning applications may be of use to keep track of the development process, particularly when several people are working together;
    - prefer the use of personal access method to facility instead of group account, to be able to reconstruct issues;
  - Apply appropriate levels of data management:
    - encrypting data or using secure online storage;
    - check data integrity after data transfer;
    - check the timescale for data transfer related to the net bandwidth used;
    - foreseen the right space required after elaboration;

# 5. Deposit and Long-Term Preservation

- Identify which of the data sets produced are considered to be of long-term value:
  - Select data of long-term value since not all data are scientifically valuable in particular in observations (i.e. bad pointing, achievement of target precision, errors.);
  - Safeguard the data behind the graph: tables and graphs in papers are of interest for preservation;
- Outline the plans for preparing and documenting data for preservation and sharing:
  - Assure that your data will remain accessible;
  - assure datacenter is trustable;
- Explain your archiving/preservation plan to ensure the long-term value of key datasets

## 6. Resourcing

- What resources will you require to deliver your plan?
  - Outline and justify costs:
    - outsource services;
    - pay for data management support;
    - link resources with roles and responsibilities;
- Outline additional hardware, software and technical expertise, support and training that is likely to be required and how it will be acquired
  - Don't underestimate the human effort required;
  - Use as much as possible already existing infrastructure and expertise, it will reduce costs.

# DMP final considerations:

Data plans are an integral part of grant applications – not an afterthought

Data plans are enhanced through collaboration

Data plans are living documents – they will change

The short list:

[https://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP\\_Checklist\\_2013.pdf](https://www.dcc.ac.uk/sites/default/files/documents/resource/DMP/DMP_Checklist_2013.pdf)

# An Example

Try by yourself to generate a DMP with a web tool originally developed by Digital Curation Centre :

<https://www.dcc.ac.uk/dmponline>

A descriptive manual:

<https://dmponline.dcc.ac.uk/help>



Notice: Your password was changed successfully. You are now signed in.

#### Info:

As part of our routine maintenance, we have upgraded our SSO login to enhance security.

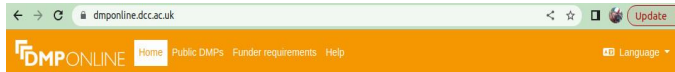
- If your account was not linked to your institutional credentials, please log in as normal.
- If your account was linked to your institutional credentials, you will now need to re-link your account please log in using your DMPonline email and password.
- Next, go to **Edit profile** > scroll down to the point **Institutional credentials**, and select the **Link your**
- After re-linking your account, you need to refresh your browser to complete the process. Remember

## My Dashboard

The table below lists the plans that you have created, and that have been shared with you by others. You can or remove these plans at any time.

Project Title	Template	Edited	Role	Test	Visible
<a href="#">IA2 Management Plan</a>	DCC Template	08-02-2022	Owner	<input checked="" type="checkbox"/>	N/A

Create plan



#### Info:

As part of our routine maintenance, we have upgraded our SSO login to enhance security.

- If your account was not linked to your institutional credentials, please log in as normal.
- If your account was linked to your institutional credentials, you will now need to re-link your account as part of this upgrade. To do this, please log in using your DMPonline email and password.
- Next, go to **Edit profile** > scroll down to the point **Institutional credentials**, and select the **Link your institutional credentials** option.
- After re-linking your account, you need to refresh your browser to complete the process. Remember to save your updated settings.

## Plan to make data work for you

Data Management Plans that meet institutional funder requirements.



DMPonline helps you to create, review, and share data management plans that meet institutional and funder requirements. It is provided by the Digital Curation Centre (DCC).

Sign in Create account

\* Email

\* Password

[Forgot password?](#)

Remember email

Sign in

- or -


Sign in with your institutional credentials



# Call for space request

## Call #3 - Request for Computing e-infra INAF

cristina.knopic@inaf.it [Cambia account](#)

 Bozza salvata

Il nome e la foto associati al tuo Account Google verranno registrati quando caricherai i file e invierai questo modulo. Solo l'indirizzo email che inserisci fa parte della risposta.

### Collaborators

Please specify row by row the list of the Collaborators ( name and e-mail )

List of collaborators

Francesca, Marco, Andrea, Serena

[Indietro](#)

[Avanti](#)

Pagina 2 di 8

[Cancella modulo](#)

Non inviare mai le password tramite Moduli Google.

Questo modulo è stato creato all'interno di INAF Istituto Nazionale di Astrofisica. [Segnala abuso](#)

Google Moduli

## Call #3 - Request for Computing e-infra INAF

cristina.knopic@inaf.it [Cambia account](#)

 Bozza salvata

Il nome e la foto associati al tuo Account Google verranno registrati quando caricherai i file e invierai questo modulo. Solo l'indirizzo email che inserisci fa parte della risposta.

\* Indica una domanda obbligatoria

### Project - Scientific Justification

Please report here the Scientific Justification of the project by detailing

Project Title \*

Test storage request

Scientific description (extended, max 6000 characters) \*

I need to store and organize my observation and simulated data of galactic transients, and ...

[Indietro](#)

[Avanti](#)

Pagina 3 di 8

[Cancella modulo](#)

Non inviare mai le password tramite Moduli Google.

# Call for space request

## Call #3 - Request for Computing e-infra INAF

cristina.knopic@inaf.it [Cambia account](#)

 Bozza salvata

Il nome e la foto associati al tuo Account Google verranno registrati quando caricherai i file e invierai questo modulo. Solo l'indirizzo email che inserisci fa parte della risposta.

\* Indica una domanda obbligatoria

### Project - Platform details

Please select here the platform requested for the project

#### Computing e-infra Request \*

- Computing (Pleiadi)
- Long term Storage (IA2)
- Computing (Cineca)

[Indietro](#)

[Avanti](#)

Pagina 4 di 8

[Cancella modulo](#)

Non inviare mai le password tramite Moduli Google.

Questo modulo è stato creato all'interno di INAF Istituto Nazionale di Astrofisica. [Segnala abuso](#)

Google Moduli

Requested long-term storage space and technical justification \*

All data are fundamental for creating a catalog of galactic transients, I need a place where to preserve them for future uses and when users can easily retrieve specific objects....

Type and format of data and description of the structure of the data collection \*

My collection will be composed by three type of data (images, spectra and time series) that will be grouped together based on object or target. Also calibration data will accompany the datasets and documentation will describe specific procedures I used to calibrate and analyze data. The Data Model can be found ([link](#)) and the reference publication standard are [SIAP](#), [SSAP](#), [TAP](#). Each dataset (group of different files aggregated in a tar file) can reach 5 GB.

Expected access frequency \*

I foresee about 3 accesses to data daily, with an expected download of about [45GB](#) per day (3 datasets per access)

Access policy \*

Only a group of collaborators of mine can access the data until the paper will be accepted, than everyone will be able to retrieve data.


Do you plan to public your data? \*

If your data are for private access only, answer No. Otherwise, provide any useful information to make your data public, according to the FAIR principles, and how you plan to do it.

I plan to publish data using [VOspace](#) or I plan to build an archive using [TAP](#) services.

# Call for space request

## Call #3 - Request for Computing e-infra INAF

cristina.knopic@inaf.it [Cambia account](#)  Bozza salvata

Il nome e la foto associati al tuo Account Google verranno registrati quando caricherai i file e invierai questo modulo. Solo l'indirizzo email che inserisci fa parte della risposta.


*\* Indica una domanda obbligatoria*

### Timing

Please insert the expected start date and end date of the project


**Start date \***

Data




**End date \***

Data



Una copia delle risposte verrà inviata via email all'indirizzo fornito.

[Indietro](#) [Invia](#)  Pagina 8 di 8 [Cancella modulo](#)

Non inviare mai le password tramite Moduli Google.



# Long term storage

<https://www.ia2.inaf.it/index.php/ia2-services/data-sharing-preservation/long-term-preservation>

At this page you can find useful information to request data preservation:

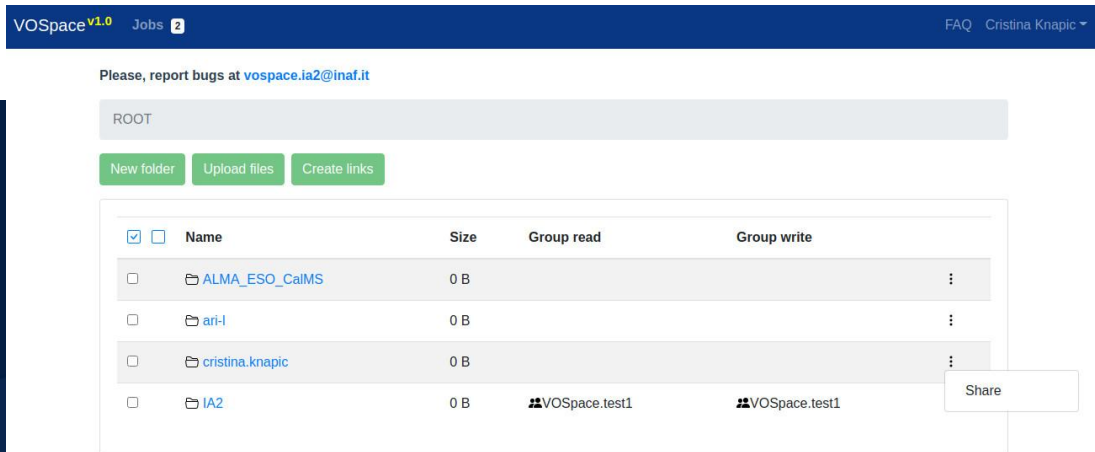
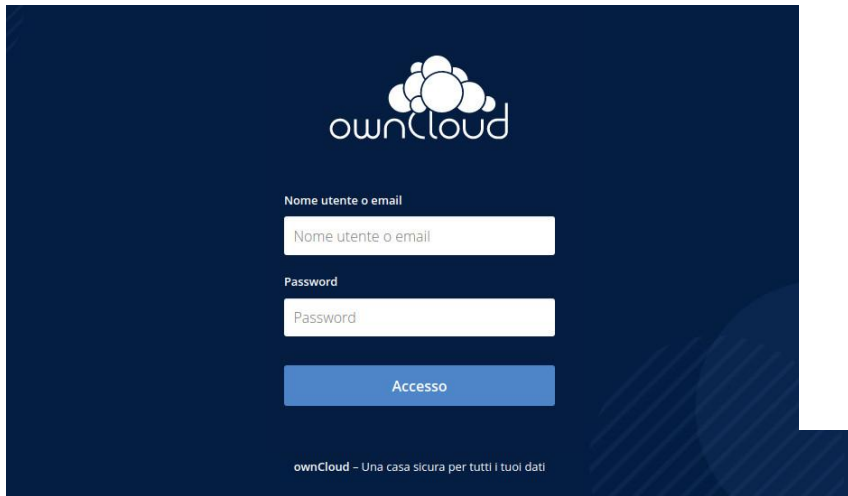
- how the tape library works;
- how to request an account;
- how to write a data management plan (in brief!);
- how to ingest data;
- how to retrieve data;

# Data sharing

More details on data sharing and preservation can be found at

<https://www.ia2.inaf.it/index.php/ia2-services/data-sharing-preservation>

<https://owncloud.ia2.inaf.it/index.php/login>



<http://vospace.ia2.inaf.it/ui/>

Thanks for attention!  
Questions?