DOI for Astronomical Data Centers: ESO

Hainaut, Bordelon, Grothkopf, Fourniol, Micol, Retzlaff, Sterzik, Stoehr [ESO]
Enke, Riebe [AIP]
DOI for Astronomical Data Centers: ESO

Hainaut, Bordelon, Grothkopf, Fourniol, Micol, Retzlaff, Sterzik, Stoehr [ESO]
Enke, Riebe [AIP]
What is ESO?

What do we do in terms of PIDs

How does that work with RDA’s recommendations
European Southern Observatory

- European Organisation for Astronomical Research in the Southern Hemisphere
European Southern Observatory

- European Organisation for Astronomical Research in the Southern Hemisphere

Who?

- Austria, Belgium, Brazil, the Czech Republic, Denmark, Finland, France, Germany, Italy, the Netherlands, Poland, Portugal, Spain, Sweden, Switzerland and the United Kingdom + Chile
European Southern Observatory

- European Organisation for Astronomical Research in the Southern Hemisphere

Why?

- provide state-of-the-art research facilities to astronomers
  - big telescopes…
- foster cutting-edge astronomical research collaborations
ESO?

What?
- La Silla
- Paranal
- Chajnantor
- Armazones
What?

- La Silla:
  - 3.6m + NTT
What?

- La Silla:
  - 3.6m + NTT

- Paranal:
  - VLT – 4x 8.2m
  - Surveys
  - VLTI
Chajnantor:
- APEX – sub/millimeter radio telescope (12m)
- ALMA – sub/millimeter radio interferometer (25km)
What?

Armazones:
- ELT – 39m, 2024
What?

- All the raw data (La Silla + Paranal + APEX + WFCAM)
  - > 20 years
  - > 20 instruments
What?

- All the raw data (La Silla + Paranal + APEX)
  - > 20 years
  - > 20 instruments

- A lot of processed data:
  - Internal data products: Standard processing into physical units
  - Advanced data products:
    - Data processing to address specific questions
    - Mosaic images, extracted + interpreted spectra…
    - Catalogues
Astronomical Data Centre?

What?

- All the raw data (La Silla + Paranal + APEX)
  - > 20 years
  - > 20 instruments

- A lot of processed data:
  - Internal data products: Standard processing into physical units
  - Advanced data products:
    - Data processing to address specific questions
    - Mosaic images, extracted + interpreted spectra…
    - Catalogues

Why?

- Preservation
- Distribution
- Archive-based science
How much?

- 40 million files
- 700 TB
- 30 billion metadata rows
How much?

- 40 million files *(each with unique, permanent identifier)*
- 700 TB
- 30 billion metadata rows
Astronomical Data Centre?

How much?
- 40 million files
- 700 TB
- 30 billion metadata rows

How?
- NGAS, Linux, RAID5/6
- 2d spinning copies
Astronomical Data Centre?

■ How much?
  ➢ 40 million files
  ➢ 700 TB
  ➢ 30 billion metadata

■ How?
  ➢ NGAS, Linux, RAID5
  ➢ 2d copies

■ Where?
  ➢ “Science Archive Facility” – ESO HQ
  ➢ Off-site copy at MP CDF, with off-off-site tape backup
**PIDS at ESO: DOI**

- DOI = PID for “digital objects”
  - DOI:10.18727/xxxxxx/yyyy/zzzz

- DOI registered at agency

- DOI resolves to a landing web page
  - Meta data about the digital object
  - Pointer to the actual digital object

- DOI is persistent
  - At least 10yrs

- ESO has an agreement with TIB / Leibniz
  - (sub-agency of DataCite for Germany)
  - DOIs minted free of charge
DOIs: WHY?

- **Track (by ESO):**
  - Track observation programmes used in publication:
    - Science output per programme
    - Science output per instrument

![Graph showing VLT instruments (1999-2015)]

![Graph showing ESO Publications 1996-2015](chart)
Track (by ESO):

- Track observation programmes used in publication:
  - Science output per programme
  - Science output per instrument

Cite:

- programme/dataset as citable entity

DOIs for Astronomical Data


DOIs: WHY?

- **Track (by ESO):**
  - Track observation programmes used in publication:
    - Science output per programme
    - Science output per instrument

- **Cite:**
  - programme/dataset as citable entity

- **Reproduce science result:**
  - Unique, reproducible dataset to reproduce a result
**DOIs: WHAT?**

- All raw data from an observing programme / run
  - DOI:10.18727/ESOprog/097.C-0123/B
  - Finite and well defined,
  - Obtained with a science goal
  - Continuity with previous system (citing prog.ID)
  - Metadata: information from the programme
    - Authors, telescope, instrument, programme abstract...
DOIs: WHAT?

Advanced Data Product package

- DOI:10.18727/ESOdata/0123546*
- Data releases from Survey, from large programme…
- Each release is
  - Finite, well defined,
  - Well documented
  - Heavily used by the community

* Unique, semi-sequential number
DOIs: WHAT?

- Arbitrary data set – user generated
  - DOI:10.18727/ESOdata/00123547
  - List of unique data files identifiers
    - raw science, calibration, processed or combination
    - Upload the list of IDs to ESO → ESO generates the DOI
  - Metadata: author, abstract with purpose of the dataset
  - Unique dataset
    - Author can refer to his dataset, citation
    - Allows others to reproduce the exact same investigation
    - Allows ESO to track
      - Programmes
      - Instrument
      - Type of data (raw, processed, advanced…).
…14 recommendations on how to adapt a data source for providing identifiable subsets for the long term, elaborated by the RDA WG on Dynamic Data Citation.

These recommendations are mostly for databases. How do they apply to our data?
Recommendations

- **R1 - Data Versioning:** Apply versioning to ensure earlier states of data sets can be retrieved.
  **OK:** Each of our asset (be it raw or a data product) has a unique identifier based on its creation time. Changes are always incremental.

- **R2 – Timestamping:** Ensure that operations on data are timestamped.
  **OK:** Each of our asset (be it raw or a data product) has a unique identifier based on its creation time. Always incremental.

  \[\text{EMMI}.2013-01-04T23:42:12.123\]
Recommendations

- **R3 - Query Store Facilities:** Provide means for storing queries and the associated metadata in order to re-execute them in the future.
  
  **N/A:** Only the result of the queries resulting in a data download are preserved (and stored with a query ID). The actual query or selection process leading to that dataset is not preserved.

- **R4 - Query Uniqueness:** Re-write the query to a normalised form so that identical queries can be detected.
  
  **NOT supported:** only the result of the query is stored, not the query itself.
Recommendations

■ **R5 - Stable Sorting:** Ensure that the sorting of the records in the data set is unambiguous and reproducible.

**OK:** for the dataset discussed: a list of frames is unambiguous and reproducible.

■ **R6 - Result Set Verification:** Compute fixity information (also referred to as checksum or hash key)

**OK:** The individual frames have a check sum. The list of frames constituting the result of a query does not, but is simple to compare (eg, diff).
Recommendations

- **R7 - Query Timestamping**: Assign a timestamp to the query based on the last update to the database (for privacy)
  
  **OK NO**: The archive queries are time stamped. DB is ~continuously updated; privacy not an issue(?)

- **R8 - Assigning Query PID**: Assign PID to query if either the query or the result is new. Otherwise, return the existing PID of the earlier query.
  
  **OK NO**: PID is issued, but not check for uniqueness. Not an issue in our case?
Recommendations

- **R9 - Store the Query:** Store query and metadata (e.g. PID, original and normalised query, query and result set checksum, timestamp, superset PID, dataset description, and other) in the query store.

  **OK:** Query itself is not relevant; result list is stored with metadata.

- **R10 - Automated Citation Text Generation:**
  Generate citation texts in the format prevalent in the designated community. Include the PID into the citation text snippet.

  **OK:** this is the purpose of the DOI.

  Landing page includes a citable snippet in BiTeX.
R11 - Landing Pages: Make the PIDs resolve to a human readable landing page that provides the data and metadata. 

**OK:** this is a requirement of the DOI.

R12 - Machine Actionability: Provide an API / machine actionable landing page to access metadata and data via query reexecution. 

**To be implemented:** simple via the programmatic interface of the Archive Services.
Recommendations

- **R13 - Technology Migration**: When data is migrated to a new representation, migrate also the queries and associated fixity information. OK: lists are resilient; migration is core mission of the Science Archive Facility.

- **R14 - Migration Verification**: Verify successful data and query migration, ensuring that queries can be re-executed correctly. OK: core mission of the Science Archive Facility
Final thoughts

We generate DOIs for
- Noteworthy datasets
- User-assembled dataset
to make the dataset
- Trackable by us
- Citable in the literature
- Reproducable by other researchers
Final thoughts

- We generate DOIs for
  - Noteworthy datasets
  - User-assembled dataset

- to make the dataset
  - Trackable by us
  - Citable in the literature
  - Reproducable by other researchers

- Do you have comments / recommendations?
  → ohainaut@eso.org