IVOA 2025

The Building of a Vocabulary of Observation Facilities from Diverse Data Sources

November 2025

Liza FRETEL, Baptiste CECCONI, Laura DEBISSCHOP

LIRA Observatoire de Paris | PSL

Adaptive and Multi-Source Entity Matching for Name Standardization of Astronomical Observation Facilities

Liza Fretel[1,*,†], Baptiste Cecconi[1] and Laura Debisschop[1]

[1]Paris Observatory, Pl. Jules Janssen, 92190, Meudon, France

Abstract

This ongoing work focuses on the development of a methodology for generating a multi-source mapping of astronomical observation facilities. To compare two entities, we compute scores with adaptable criteria and Natural Language Processing (NLP) techniques (Bag-of-Words approaches, sequential approaches, and surface approaches) to map entities extracted from eight semantic artifacts, including Wikidata and astronomy-oriented resources. We utilize every property available, such as labels, definitions, descriptions, external identifiers, and more domain-specific properties, such as the observation wavebands, spacecraft launch dates, funding agencies, etc. Finally, we use a Large Language Model (LLM) to accept or reject a mapping suggestion and provide a justification, ensuring the plausibility and FAIRness of the validated synonym pairs. The resulting mapping is composed of multi-source synonym sets providing only one standardized label per entity. Those mappings will be used to feed our Name Resolver API and will be integrated into the International Virtual Observatory Alliance (IVOA) Vocabularies and the OntoPortal-Astro platform.

Keywords

Entity mapping strategy, Controlled Vocabularies, FAIR mapping, Astronomical observation facilities

LIRA Observatoire de Paris | PSL

Summary

1. Context
2. Update and map data
3. Activity Report (07/2025 - 11/2025)
4. Challenges and Perspectives

What are observation facilities ?

Objectives

Data sources

Metadata standardization

1. Context

Context
What are observation facilities ?

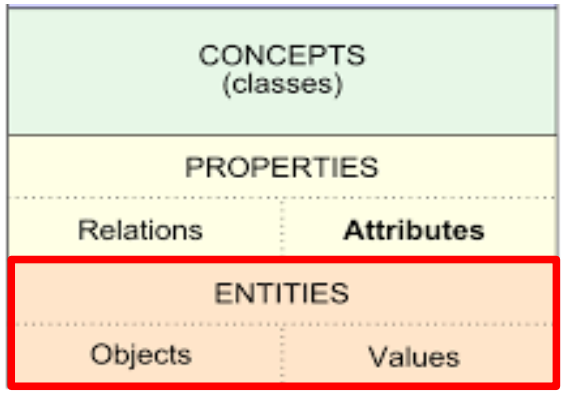
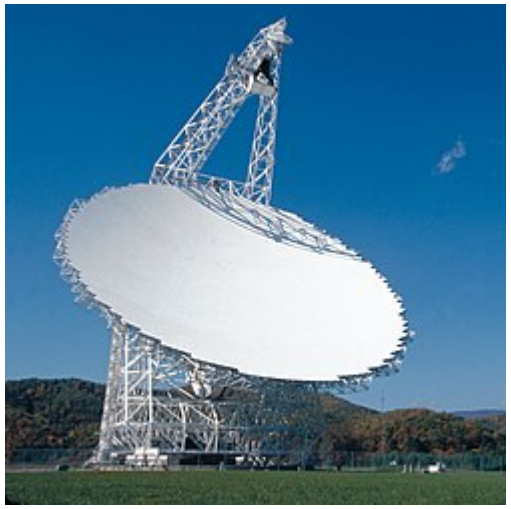


Figure: Ontology layers



Green Bank Telescope (USA)

Sphinx Observatory (Switzerland)

Astronomical observation facilities: entity level.

Observation Facility: Research infrastructure designed, constructed, and operated for the **purpose of acquiring astronomical data** that enables the **observation of celestial objects and phenomena** under controlled scientific conditions.

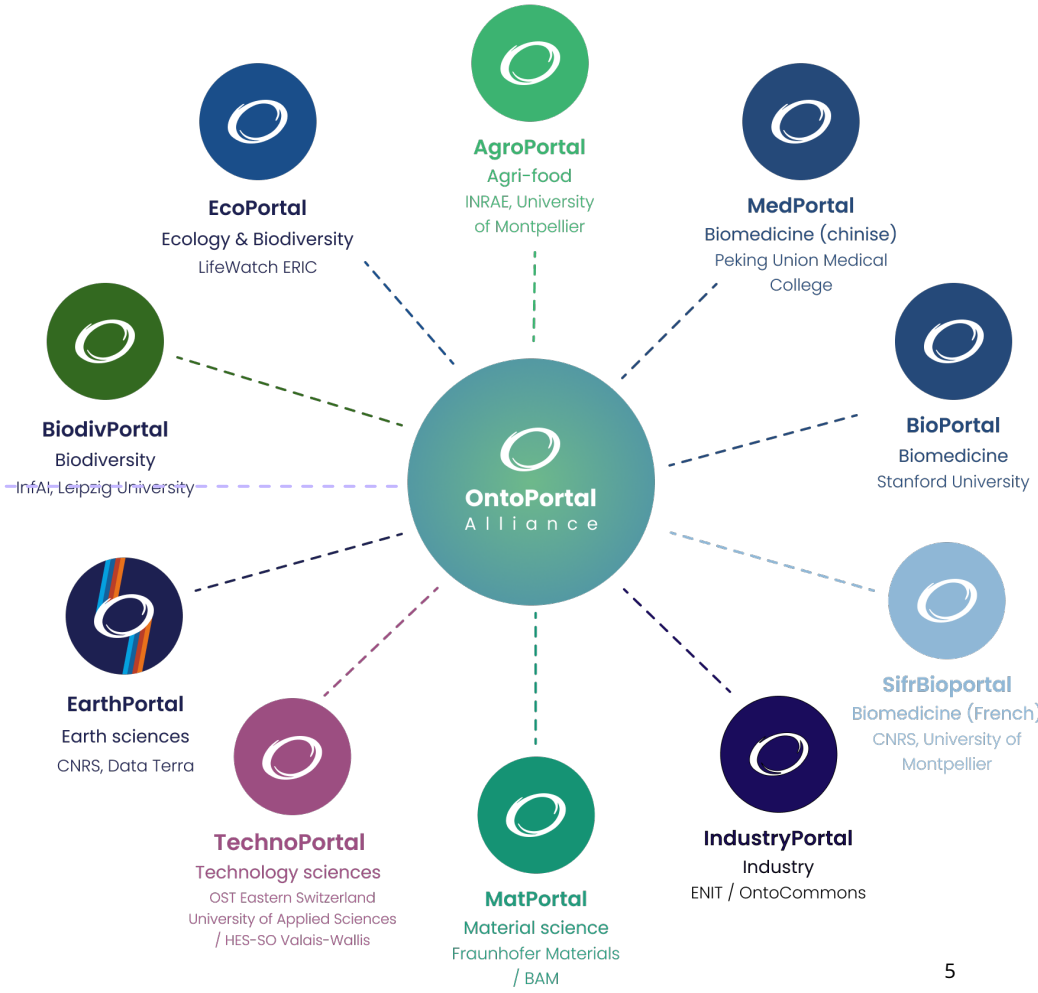
Hubble Space Telescope, HST

Context
Objectives

- Create and issue a cross-community vocabulary for the **IVOA** (International Virtual Observatory Alliance)

- Merge data and their metadata from semantic lists issued by different communities into one ontology and make it available on **OntoPortal-Astro**
  http://ontoportal-astro.eu/ontologies/OBSF *temporary private link

- Set a recommended label per observation facility to help researchers lookup and use the conventional labels (**Name Resolver**)



Context
Data sources

```
▼<Facility>
  <name>Siding Spring Observatory</name>
  <type>Observatory</type>
  <address>Australian National University Private Bag
  Coonabarabran, NSW</address>
  <country>Australia</country>
  <description>The Siding Spring Observatory, situated in New
  South Wales, Australia, is the site of the Anglo-Australian
  Telescope and numerous other facilities. It is known for its
  work in optical astronomy, particularly in the study of
  galaxies, star systems, and planetary science, and its critical
  role in various international research programs.</description>
</Facility>
```

NASA's **PDS** (Planetary Data Science)

```
<Spase xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns="http://www.sp
  <Version>2.2.0</Version>
  <Observatory>
    <ResourceID>spase://SMWG/Observatory/GOES/10</ResourceID>
    <ResourceHeader>
      <ResourceName>GOES 10</ResourceName>
      <AlternateName>1997-019A</AlternateName>
      <AlternateName>GOES-K</AlternateName>
      <ReleaseDate>2019-05-05T12:34:56Z</ReleaseDate>
      <Description>The Geostationary Operational Environmental Satellite GOES 10
```

SPASE (Space Physics Archive Search and Extract)

```
Code  Long.   cos      sin    Name
000   0.0000 0.62411 +0.77873 Greenwich
001   0.1542 0.62992 +0.77411 Crowborough
002   0.62   0.622   +0.781   Rayleigh
003   3.90   0.725   +0.687   Montpellier
004   1.4625 0.72520 +0.68627 Toulouse
005   2.231000.659891+0.748875Meudon
006   2.124170.751042+0.658129Fabra Observatory
007   2.336750.659470+0.749223Paris
```

IAU-MPC (International Astronomical Union Minor Planet Center)

LIRA Observatoire de Paris | PSL

Context
Data sources

Explorer 12

NSSDCA/COSPAR ID: 1961-020A

Description

Explorer 12 was a spin-stabilized, solar-cell-powered spacecraft instrumented to measure cosmic-ray particles, trapped particles, solar wind protons, and magnetospheric and interplanetary magnetic fields. It was the first of the S 3 series of spacecraft, which also included Explorers 14, 15, and 26.

NSSDC (National Space Science Data Center)

Pioneer 6 (Q11699464)

NASA space probe
Pioneer A

instance of	space probe
NSSDCA ID	1965-105A
NAIF ID	-6

Wikidata (query by subclass hierarchy)
+ Wikipedia's 1st paragraph (description to enhance semantic matches)

NAIF ID	NAME
-1	'GEOTAIL'
-3	'MOM'
-3	'MARS ORBITER MISSION'
-5	'AKATSUKI'
-5	'VCO'
-5	'PLC'
-5	'PLANET-C'
-6	'P6'
-6	'PIONEER-6'
-7	'P7'
-7	'PIONEER-7'
-8	'WIND'
-12	'VENUS ORBITER'

NASA's **NAIF** (Navigation and Ancillary Information Facility)

Metadata standardization

Developed Pipeline

Filtering Criteria

Surface and Semantic Scores

LLM Validation

2. Update and map data

(Partially introduced before the previous IVOA meeting)

LIRA Observatoire de Paris | PSL

Update and map data
Metadata standardization

- Properties standardisation (with crosswalks)
  - Conversion to SKOS, RDFS, OWL…
  - Use of IVOA's RDF vocabularies
    ex : wavebands
    https://www.ivoa.net/rdf/messenger/2020-08-26/messenger.html
  - OBSF custom namespace
    ex : obsf:waveband [property]
    obsf:launch_date [property]
    obsf:telescope [class]

```
# SPASE metadata crosswalk
CROSSWALK = {
    "ResourceID": SKOS.notation,
    "type": RDF.type,
    "ResourceName": SKOS.prefLabel,
    "AlternateName": SKOS.altLabel,
    "Description": DCTERMS.description,
    "ObservatoryGroupID": DCTERMS.isPartOf,
    "StartDate": DCAT.startDate,
    "EndDate": DCAT.endDate,
    "ObservatoryRegion": GEO.location,
    "Latitude": GEO.latitude,
    "Longitude": GEO.longitude,
    "Elevation": GEO.altitude,
    "Agency": OBSF.funding_agency,
    "PriorIDs": OBSF.prior_id,
    "PriorID": OBSF.prior_id}
```

Update and map data
Metadata standardization

Objectives:

- Have **comparable setups** for criteria filtering, that will help filter out « impossible » mappings
- Create one unique ontology with **homogeneous metadata** for publications

Inconvenients:

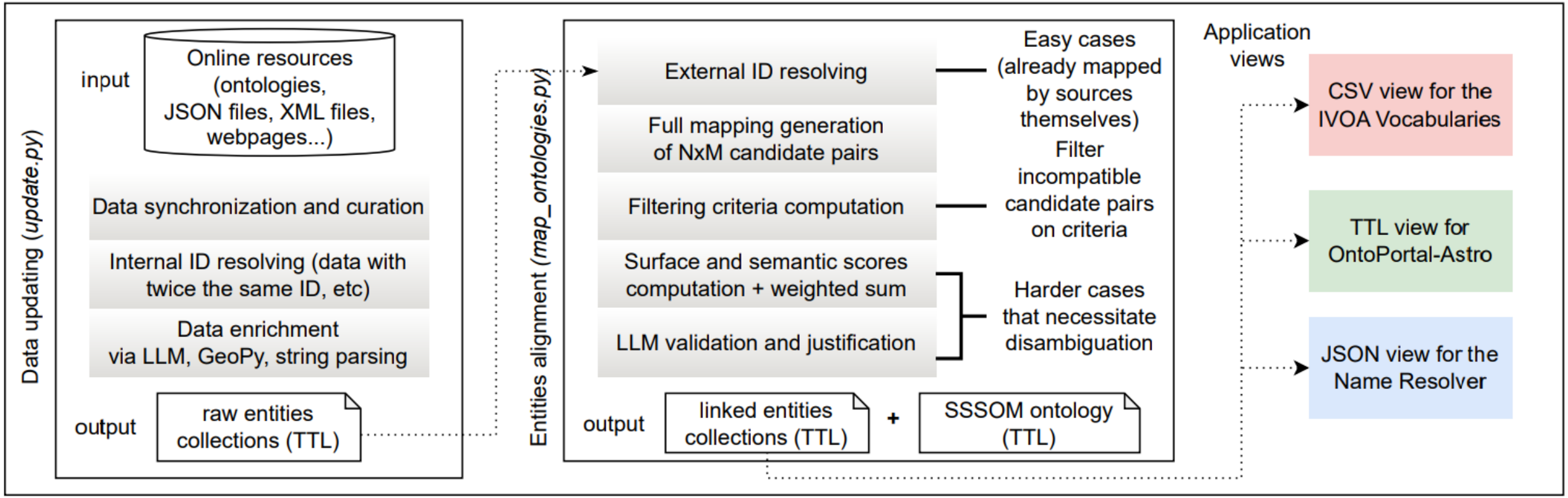
- Need to manually make **crosswalks between each resource's fields**, common ontological properties, IVOA properties or our custom properties.

Advantages:

- Towards a more **controlled vocabulary**

=> could be automatized (Property Matching) if the schemas were more complex

Update and map data
Developed Pipeline



Update and map data
Filtering Criteria

Geodesic distance > 10km for ground facilities, location (country, continent) mismatch
Challenge : incorrect latitude & longitude in data → manual fixes

Different aperture (telescope lense's dimension)
Challenge : different units & rounding → inches conversions to meters, rounding

Different launch date
Challenge : years only in certain sources → rounded to January 1st

Unmatching type
Challenge : entities typed differently (SPASE) or untyped (AAS) → manual fixes, **hierarchical type comparisons**

Unmatching external identifier (COSPAR-ID, NAIF-ID, NSSDCA-ID…)

...

Update and map data
Surface and Semantic Scores

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF
Term x within document y

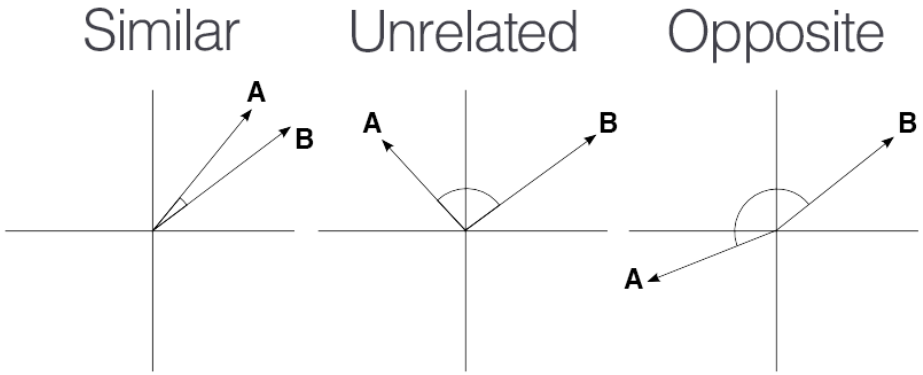
$tf_{x,y}$ = frequency of x in y
df_x = number of documents containing x
N = total number of documents

Source : Ted Mei (Medium)

Considering entities *x* and *y*, we can compute relevant **semantics scores** between *x, y :*

Scores	*Effective fields*	*Multilingual support*
− TF-IDF cosine similarity(*x, y*)	labels, description, definition	No
− Transformer cosine similarity(*x, y*)	””	Yes
− P(*x* acronym of *y*)	labels	only same alphabet
− Levenshtein similarity (*x, y*) = $1 - \frac{\text{Levenshtein distance}(x, y)}{\max(\lvert x\rvert, \lvert y\rvert)}$	””	only same alphabet

Cosine Similarity



Source : Sindhu Selham (Medium)

Update and map data
LLM Validation

Deepseek (400B Parameters, Open Source & available on Ollama for local run)

{"**answer**": "distinct",
"**justification**": "the two entities are distinct because they refer to different spacecraft with separate missions and launch dates. **the nasa 2001 mars odyssey orbiter** was launched in 2001 and focuses on observing neutrinos, particles, and gravitational waves across various wavebands. on the other hand, **the mars reconnaissance orbiter** (mro) was launched in 2005 with a primary mission to search for water on mars and support future missions. they are part of different nasa programs and have distinct objectives."}

```
   obs:justification "both entities refer to the same telescope, \"the nrao 300-foot green bank telescope,\" with consistent attributes such as
location (green bank, west virginia), continent (north america), and waveband (radio). the additional names provided in entity2 (\"gbo:300ft\",
\"radaiteileascóp 300 troigh\") are aliases for the same telescope. therefore, they represent the same entity despite slight differences in how
the information is presented." ;
   obs:provenance "deepseek-v3:latest validation" ;
```

Update and map data
Track of mapping decisions

```
obsf:ff0a8988-4821-41bc-9adc-4c08e4ea6c6a a sssom:Mapping ;
    sssom:mapping_date "2025-11-12T11:26:40.187974"^^xsd:dateTimeStamp ;
    sssom:mapping_set_id obsf:ba4b1c10-193f-4dc1-82e2-89d171967308 ;
    sssom:match_string "2000-017A"^^xsd:string ;
    sssom:object_id imcce:image ;
    sssom:object_match_field skos:altLabel ;
    sssom:object_source obsf:imcce_list ;
    sssom:predicate_id skos:exactMatch ;
    sssom:similarity_measure semapv:StringEquality ;
    sssom:subject_id nssdc:image ;
    sssom:subject_match_field obsf:NSSDCA_ID ;
    sssom:subject_source obsf:nssdc_list .
```

Objective: FAIR mappings

Namespaces:

- SSSOM (track mapping decisions, author of mappings, justification strings by human or LLM validator)
- SEMAPV (reference pre-processing algorithms, surface and semantic scores applied)

LIRA Observatoire de Paris | PSL

Data curation & post-processing

Human validation of mappings

Automated updates support

+ SSSOM Ontology for FAIRness

+ Code optimization

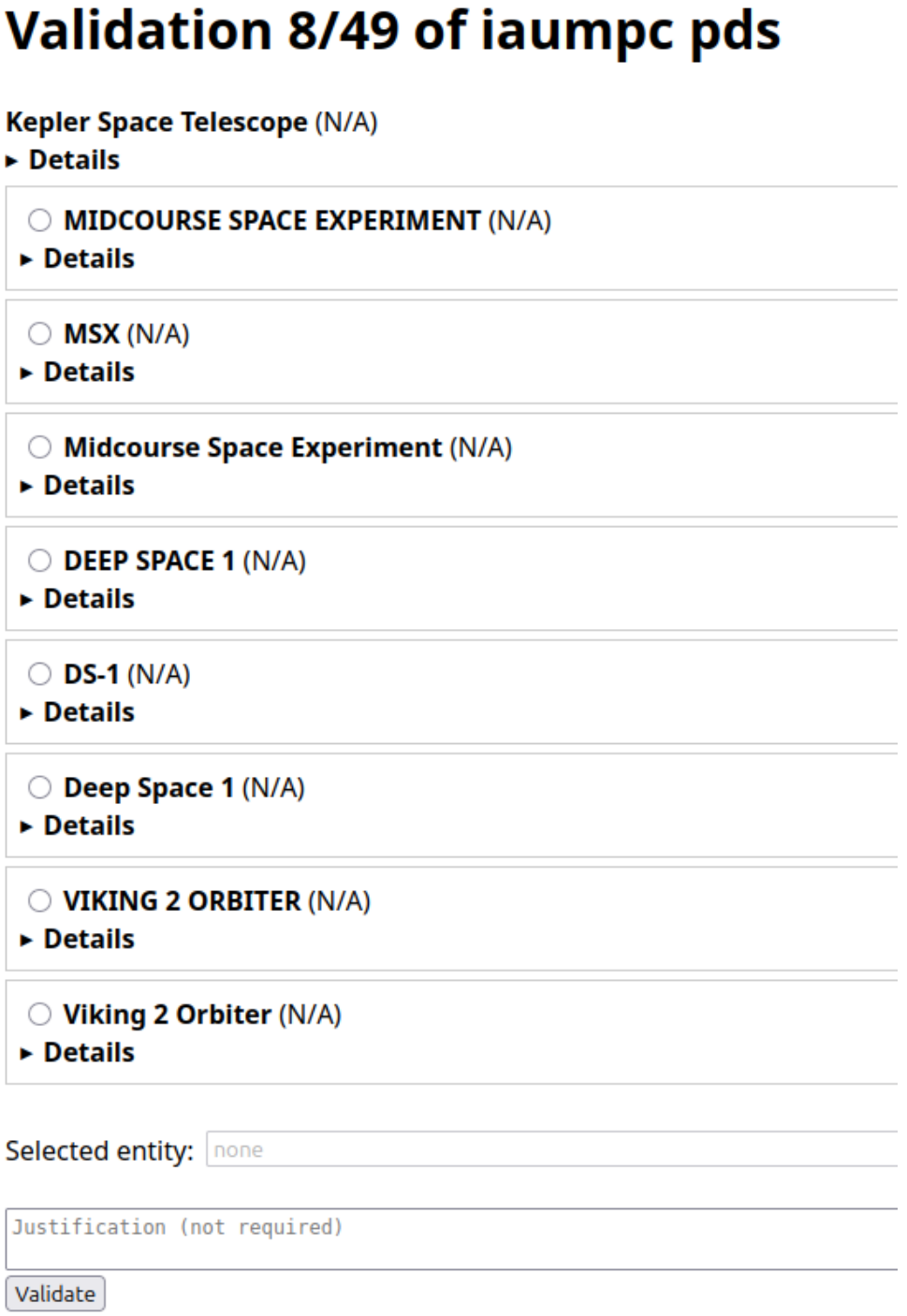
3. Activity report

Activity report (2025-07 to 2025-11)
Data curation & post-processing

- Detection and curation of PDS longitude errors, and fix Wikidata, AAS, SPASE errors in data (manually)
- Location comparison enabled for entities with no latitude and longitude : continent and country name standardization
- Add filtering criteria: lense aperture, external identifier mismatch
- Extraction of Surveys, Instruments types
- Create types Space facility, Ground facility to be compatible with the SPASE metadata schema

Activity report (2025-07 to 2025-11)
Human Validation of Mappings

- For small updates, human validation is prefered over LLM validation.



Activity report (2025-07 to 2025-11)
Automated Updates Support

Objective:

Make one update of the vocabularies every month

Map the newly added or modified entities

Progress:

Support for mapping checkpoints (execution stop)

Compare entities changes and update the last modified dates

Mapping of instruments and surveys

Create an annotated dataset of validated mappings

LLM validation

Automatic updates of vocabularies

4. Challenges and perspectives

Challenges and Perspectives

Mapping of instruments and surveys

- Instruments are part of platforms.
- The platforms must be mapped first => then the instruments will be easier to map.

Create an annotated dataset of validated mappings

Objective :

- Evaluate an automated mapping
- Few-shot training of validators

Annotation	Entity1	Entity2
o	Centre National de la Recherche Scientifique 2m Bernard Lyot Telescope at Observatoire Midi-Pyrenees	Bernard Lyot 2.0m https://pds.nasa.gov/data/pds4/context-pds4/telescope/pic_du_midi.bernardlyot_2m0_1.0.xml
x	El Observatorio Astronomico Nacional/Sierra San Pedro Martir	1.5 m https://pds.nasa.gov/data/pds4/context-pds4/telescope/ensenada.1m5_1.0.xml
o	University of California 3m C. Donald Shane Telescope at Lick Observatory	3.05-m Shane reflector https://pds.nasa.gov/data/pds4/context-pds4/telescope/lick.shane3m05_1.0.xml
o	National Research Foundation of South Africa 1.9m Radcliffe Telescope at South African Astronomical Obser	Radcliffe 1.88m telescope https://pds.nasa.gov/data/pds4/context-pds4/telescope/saao.radcliffe_1m88_1.0.xml
x	Warner and Swasey Observatory	1.83-m Perkins Warner & Swasey reflector https://pds.nasa.gov/data/pds4/context-pds4/telescope/lowell.perkins_
o	Astronomical Institute of Academy of Sciences of Czech Republic 0.65m Telescope at Ondrejov Observatory	0.65 m https://pds.nasa.gov/data/pds4/context-pds4/telescope/ondrejov.0m65_1.0.xml
o	Carnegie Institution for Science 2.5m Irenee du Pont Telescope at Las Campanas Observatory	Irenee du Pont 2.5m Telescope https://pds.nasa.gov/data/pds4/context-pds4/telescope/las_campanas.ireneedupont_2
o	University of Arizona/Smithsonian Institution 6.5m MMT Telescope at Fred Lawrence Whipple Observatory	6.5-m Single Mirror https://pds.nasa.gov/data/pds4/context-pds4/telescope/mmt.single_mirror6m5_1.1.xml
x	Asteroid Terrestrial-impact Last Alert System	Double Asteroid Redirection Test Spacecraft https://pds.nasa.gov/data/pds4/context-pds4/instrument_host/spacecraft.
o	California Institute of Technology 1.2m Samuel Oschin Telescope at Palomar Observatory	1.24/1.83-m Oschin Schmidt photographic equat. telescope https://pds.nasa.gov/data/pds4/context-pds4/telescope/pa
o	Carnegie Institution for Science 6.5m Walter Baade Telescope at Las Campanas Observatory	Walter Baade Telescope at Las Campanas Observatory https://pds.nasa.gov/data/pds4/context-pds4/telescope/las_ca

Challenges and Perspectives
LLM validation

LLM validation is a binary problem: same | distinct

- Time and energy consuming
- Can only consider one potential match at a time
  => intelligent candidates selection with scores and thresholds

- Not 100 % precise
- Output format is imprevisible text
  => MCP implementation to increase precision and control format

- Confusion between relation types (is part of VS exact match)
  => more annotation categories (same | distinct | is part of | has part)

Challenges and Perspectives
LLM validation

MCP (Model Context Protocols): enhance LLMs with functions that they may call when necessary

=> Could MCPs improve the mappings' validation quality of LLMs ?

MCP tools ideas:

- Search engine through scientific papers
- Provide access to other associated vocabularies (ex: space agencies synonym lists, etc)
- Possibility to delegate to human for later review when it is unsure about a mapping
- Validation function for output format controllability (answer, justification)

Challenges and Perspectives
Automatic updates of vocabularies

- Objective : issue a new version of vocabularies once a month

TODO to achieve this:

- Compute metrics to let the user determine whether to perform manual or AI mapping (depending on how many new entities are to be mapped)

- Create a method to only map entities modified after a certain date
  - if last modified date > previous update: if the label or an identifier has changed, might need to be mapped again
  - if creation date > previous update: try to map this entity

- Implement more identified facility lists over time (data extractor + metadata crosswalk)

- Let an LLM generate short descriptions of mapped entities

Thank you

LIRA Observatoire de Paris | PSL