LIRA, Observatoire de Paris, Université PSL, Sorbonne Université, Université Paris Cité, CY Cergy Paris Université, CNRS, 92190 Meudon, France

github repository: epn-vespa/FacilityList

publication: *Adaptive and Multi-Source Entity Matching for Name Standardization of Astronomical Observation Facilities*, Ontology Matching 2025

Towards a Standardized Vocabulary of Observation Facilities & Instruments

Liza FRETEL, Baptiste CECCONI

We built a pipeline to collect and unify astronomy observation facilities from multiple semantic sources, with source-specific cleaning and enrichment. Inter-source merging is done pairwise by resolving **identifier matches, generating candidate pairs, and filtering incompatible ones**. We compute a global similarity score combining **embeddings and string/numeric features**, then validate matches with an LLM. Finally, merged entities are consolidated and standardized, producing two vocabularies: OBSF and OBSI.

Multi-source entity collection

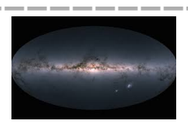
What are observation facilities ?

Telescopes

Type of facility that regroups space telescopes (on spacecraft), ground telescopes (independant or attached to an observatory), and telescope arrays (ex: NENUFAR radio telescope)

Instruments (sensors)

are located on platforms (mainly telescopes) to collect the observational data of a certain domain. They are not considered as "facilities" and will be released as a separate vocabulary.



Sky Surveys

Map or image of a region if the sky, created from data collected by certain missions.

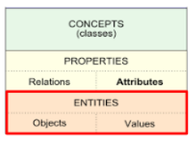


Figure: Ontology layers

Investigations

Observational projects that mostly imply launching spacecraft

Ground-based Observatories

are places or institutes that gathers astronomy researchers, telescopes, etc.

Spacecraft

Every facility sent to space.

Harmonizing metadata and values (units)

Collected sources so far

WIKIDATA · The International Astronomical Union Minor Planet Center · AAS PUBLISHING American Astronomical Society · IMCCE l'Observatoire de Paris | PSL · PDS Planetary Data System · N2YO.com · NSSDC National Space Science Data Center · NAIF The Navigation and Ancillary Information Facility

Harmonization & enrichment processes

With regular expressions, we extracted apertures from labels, converted all distances to meters. Then, we used **geolocator** to retrieve ground-based observatories address. Addresses are useful to enrich the observatories' semantic context.

Interoperability: metadata crosswalks to RDF

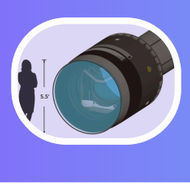
```
# SPASE metadata crosswalk
CROSSWALK = {
    "ResourceID": SKOS.notation,
    "type": RDF.type,
    "ResourceName": SKOS.prefLabel,
    "AlternateName": SKOS.altLabel,
    "Description": DCTERMS.description,
    "ObservatoryGroupID": DCTERMS.isPartOf,
    "StartDate": DCAT.startDate,
    "EndDate": 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}
```

Entity matching pipeline

Compatibility filters

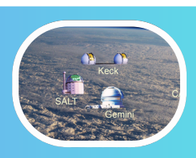
For each possible match (NxM), we eliminate incompatible pairs of entities by applying relevant filtering criteria. This largely reduces the amount candidate pairs.

Lense size mismatch



Type mismatch (ex ground VS space facilities)

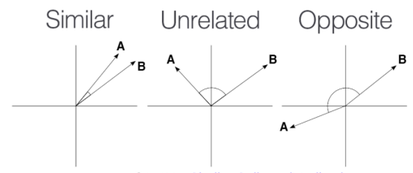
Launch year mismatch



Geodesic distance > n km
Country or state mismatch

For the remaining compatible candidate pairs, we match entities which have identical labels, then we compute additional scores and do a weighted sum; each score *i* having its own weight *i*.

$$score(p) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} s_i \times w_i$$



Source : Sindhu Selham (Medium)

Surface similarity

P(label 1 = acronym(label 2))

VEX | Venus Explorer
['v', ' ', 'e', ' '] 1st letters
[' ', 'e', ' ', 'x'] 2nd letters
[' ', ' ', ' ', ' '] stop words
[' ', ' ', ' ', ' '] other majs

$\times\ w_i$ +

BOOMERANG |Balloon Observations Of Millimetric Extragalactic Radiation ANd Geophysics
['b', ' ', 'o', ' ', ' ', ' ', ' ', 'm', ' ', 'e', ' ', 'r', ' ', ' ', ' ', ' ', ' ', 'g']
[' ', 'a', ' ', 'b', ' ', ' ', ' ', ' ', 'i', ' ', 'x', ' ', 'a', ' ', ' ', ' ', ' ', ' ']
[' ', ' ', ' ', ' ', ' ', ' ', 'f', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', 'd', ' ']
[' ', ' ', ' ', ' ', ' ', 'o', ' ', ' ', ' ', ' ', ' ', ' ', ' ', ' ', 'a', 'n', ' ', ' ']

P(label 1 = label 2)
Levenshtein similarity

		m	e	i	l	e	n	s	t	e	i	n
	0	1	2	3	4	5	6	7	8	9	10	11
l	1	1	2	3	3	4	5	6	7	8	9	10
e	2	2	1	2	3	3	4	5	6	7	8	9
v	3	3	2	2	3	4	4	5	6	7	8	9
e	4	4	3	3	3	3	4	5	6	6	7	8
n	5	5	4	4	4	4	3	4	5	6	7	7
s	6	6	5	5	5	5	4	3	4	5	6	7
h	7	7	6	6	6	6	5	4	4	5	6	7
t	8	8	7	7	7	7	6	5	4	5	6	7
e	9	9	8	8	8	7	7	6	5	4	5	6
i	10	10	9	8	9	8	8	7	6	5	4	5
n	11	11	10	9	9	9	8	8	7	6	5	4

$\times\ w_i$ +

$$LevensteinSimilarity = 1 - \frac{LevensteinDistance(|s_1|, |s_2|)}{\max(|s_1|, |s_2|)}$$

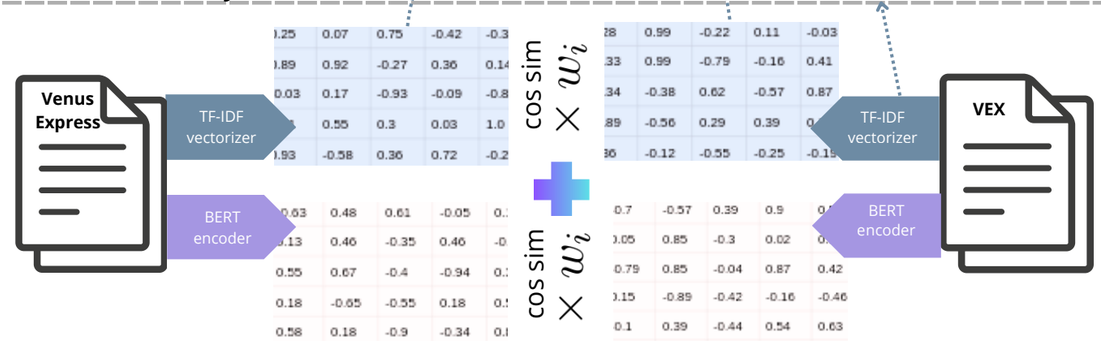
Semantic similarity

Cosine Similarity (cos sim)

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

$$tf\text{-}idf_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

N: total entity count
tf: term count in the entity's textual fields
df: term count in the ontology's textual fields



LLM and human validation of matches

```
obsf:239ca510-e322-49d2-b604-83159528e6b0 a sssom:Mapping ;
   rdfs:comment "entity1 (venus express) and entity2 (vex, venus-express) refer to the same spacecraft. \"venus express\" is the full
name, while \"vex\" is an abbreviation. the description of both entities aligns, indicating they represent the same esa mission focused on
venus."^^xsd:string ;
   obsf:acronym "1.0"^^xsd:float ;
   obsf:cosine_similarity "0.7893746033482804"^^xsd:float ;
   obsf:hybrid "0.9297915344494267"^^xsd:float ;
   obsf:levenshtein_similarity "1.0"^^xsd:float ;
   sssom:creator_label "gemma3:12b"^^xsd:string ;
   sssom:curation_rule "date",
      "distance",
      "identifier",
      "type" ;
   sssom:mapping_date "2025-12-08T19:23:57.839456"^^xsd:dateTimeStamp ;
   sssom:mapping_justification semapv:LexicalMatching ;
   sssom:mapping_set_id obsf:9f4c8fb5-b850-44d6-8116-00d0ace52d59 ;
   sssom:object_id spase:vex ;
   sssom:object_source obsf:spase_list ;
   sssom:predicate_id skos:exactMatch ;
   sssom:similarity_measure "hybrid"^^xsd:string ;
   sssom:similarity_score "0.9297915344494267"^^xsd:float ;
   sssom:subject_id pds:urn-esa-psa-context-investigation-mission.venus_express ;
   sssom:subject_source obsf:pds_list .
```

Predicted equivalence relations (skos:exactMatch)

Validation 101/118 of pds_list spase_list

gemma3:12b created this mapping on 2025-12-08T19:23:57.839456

entity1 (venus express) and entity2 (vex, venus-express) refer to the same spacecraft. "venus express" is the full name, while "vex" is an abbreviation. the description of both entities aligns, indicating they represent the same esa mission focused on venus.

VENUS EXPRESS (N/A)
▸ **Details**
exactMatch
VEX (N/A)
▾ **Details**

alt_label	Venus-Express
code	spase://CNES/Observatory/CDPP-AMDA/VEX
description	ESA Mission. Venus Express is a satellite optimised for studying the atmosphere of Venus, from the surface right up to the ionosphere. It arrived at Venus in April 2006 and continued operating for more than eight years
exact_match	urn-esa-psa-context-investigation-mission.venus_express
label	VEX
location	Venus, Heliosphere.Inner
source	spase_list
start_date	Wed, 09 Nov 2005 03:33:00 GMT
type	space-facility
type_confidence	1
url	https://github.com/hpde/hpde.io/tree/master/CNES/Observatory/CDPP-AMDA/VEX.json

◉ Same ○ Distinct ○ Broad ○ Narrow
Selected relation: exactMatch [Keep previous relation (reset)]
Justification (not required)
[Validate]

Terminology generation pipeline

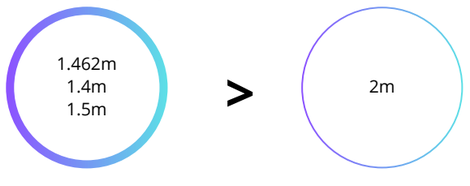
After all equivalence links were predicted by the entity matching pipeline, we need to merge the equivalent entities" information.

Majority vote & rounding

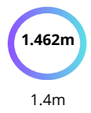
Aperture values according to 4 different sources:

- 1.462m
- 1.4m
- 1.5m
- 2m

1. Group by rounding and truncation and keep the majority



2. Keep the most precise value



Standardization of labels and definitions

1. Standardized format of main labels and terms (URI): rules + LLM

Generate a term by extending every acronym when possible (ASPOC →Active Spatial Potential Control

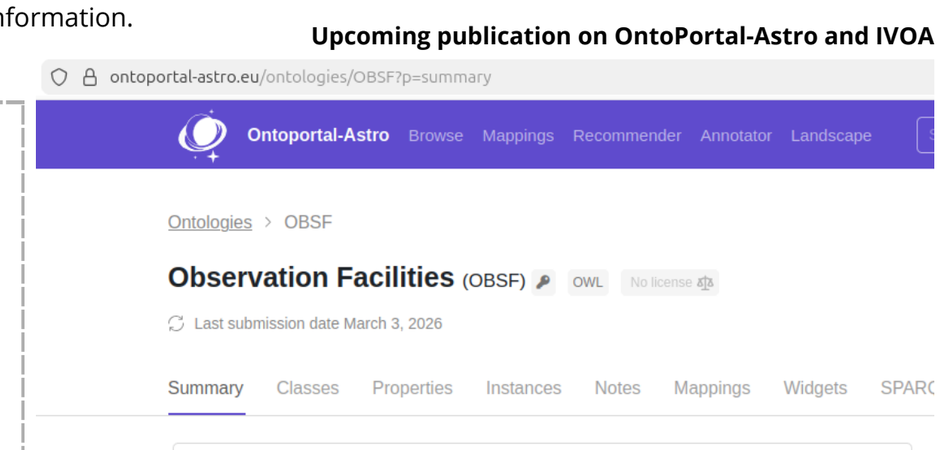
(aperture m)	Entity name	Entity type	(at Broader entity)	(, Country)
1.2m	Mercator	Telescope	at Observatorio del Roque de Los Muchachos	, Spain
2.6m	VLT Survey	Telescope	at Paranal Observatory	, Chile

Instrument name	on Host name
Plasma Electron and Current Experiment	on Cluster-1
Sun Earth Connection Coronal and Heliospheric Investigation	on STEREO A

2. LLM generation of short summary

"A 1.2-meter telescope at Observatorio del Roque de Los Muchachos on La Palma, operated by Katholieke Universiteit Leuven and the University of Geneva, focusing on exoplanet research."
generated by Mistral-Large

Upcoming publication on OntoPortal-Astro and IVOA



This work and the OntoPortal-Astro portal have been supported by FAIR-IMPACT, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101057344, as well as OPAL (Ontology Portal for Astronomy Linked-data) cascading grant (number 49-437) of the Open Science Clusters' Action for Research and Society (OSCARS) European project under grant agreement No. 101129751. The OntoPortal-Astro server is hosted by PADC (Paris Astronomical Data Center). OntoPortal-Astro is based on OntoPortal, and is jointly developed with the AgroPortal team at LIRMM and INRAE (in Montpellier, France).