IVOA 2025

Generating the Observation Facilities Vocabulary

via Multi-Source Data Alignment with
Word Embeddings and LLM Validation

L. FRETEL, B. CECCONI, L. DEBISSCHOP

FAIR-IMPACT
Expanding FAIR Solutions across EOSC

LIRA Observatoire de Paris | PSL

This work has received funding from the European Commission through the FAIR-IMPACT projet (grant agreement 101057344)

Context

- Observation facilities
  - Observatory, spacecraft, telescope…
- Lack of a reference database
  - Cross-community database
  - Naming convention
- Objectives
  - Named Entities alignment
  - Merge data from multiple sources into an ontology (http://ontoportal-astro.eu/ontologies/OBSF)

Summary

1. Facility lists
2. Data mapping & scoring
3. LLM validation
4. Output format & applications

1. Facility lists

Communities :

A : Celestial Astronomy (IVOA)
H : Heliophysics (IHDEA)
P : Planetary sciences (IPDA)
G : Geology (OGC)
O : Other, generic

List	A	H	P	G	O	Priority	Authoritative
AAS	x					+++	yes (A)
ADS	x	x	x			++	outdated?
Astroweb	x					+	
IAU-MPC	x		x			+++	yes (A,P)
IMCCE/Quaero	x	x	x			++	yes
IRAF	x					+	outdated?
IVOA obscore	x					++	
IVOA registry	x					++	
NAIF	x	x	x			+++	yes (P,H)
NASA/PDS		x	x	x		+++	yes (P)

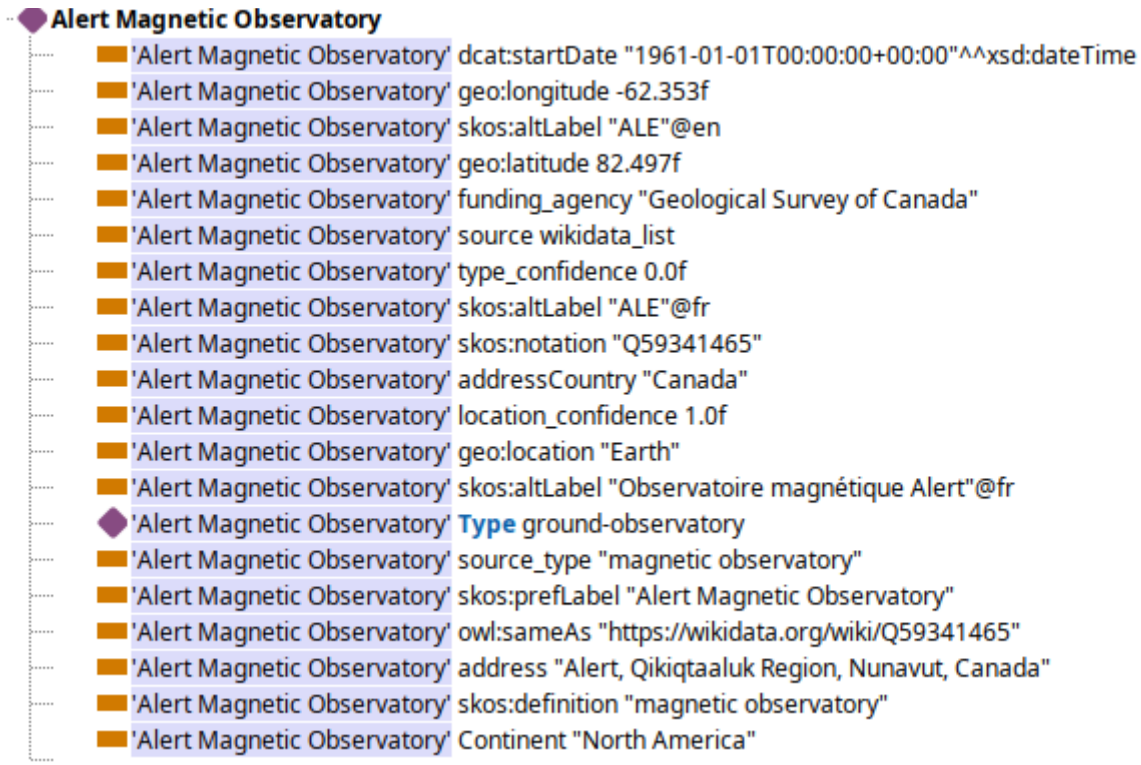
List	A	H	P	G	O	Priority	Authoritative
NED	x					++	yes (A)
NSSDC	x	x	x		x	++	yes
SANA	x	x	x		x	+	yes
SPASE		x	x			+++	yes (H)
VESPA		x	x			++	
WikiData	x	x	x	x	x	+++	
WISeREP	x					++	outdated?
WMO		x			x	+	yes
Xephem	x		x			+	

LIRA Observatoire de Paris | PSL

1. Facility lists

- Reference lists : more complete, less control
  - Provide synonyms, translations & more information to help the mapping. Used as reference only.
    ex : Wikidata, NSSDC.
- Authoritative lists : domain-specific, more reliable, not as complete
  - released by scientific institutions (archives, data alliances).
    ex : AAS, IAU-MPC.

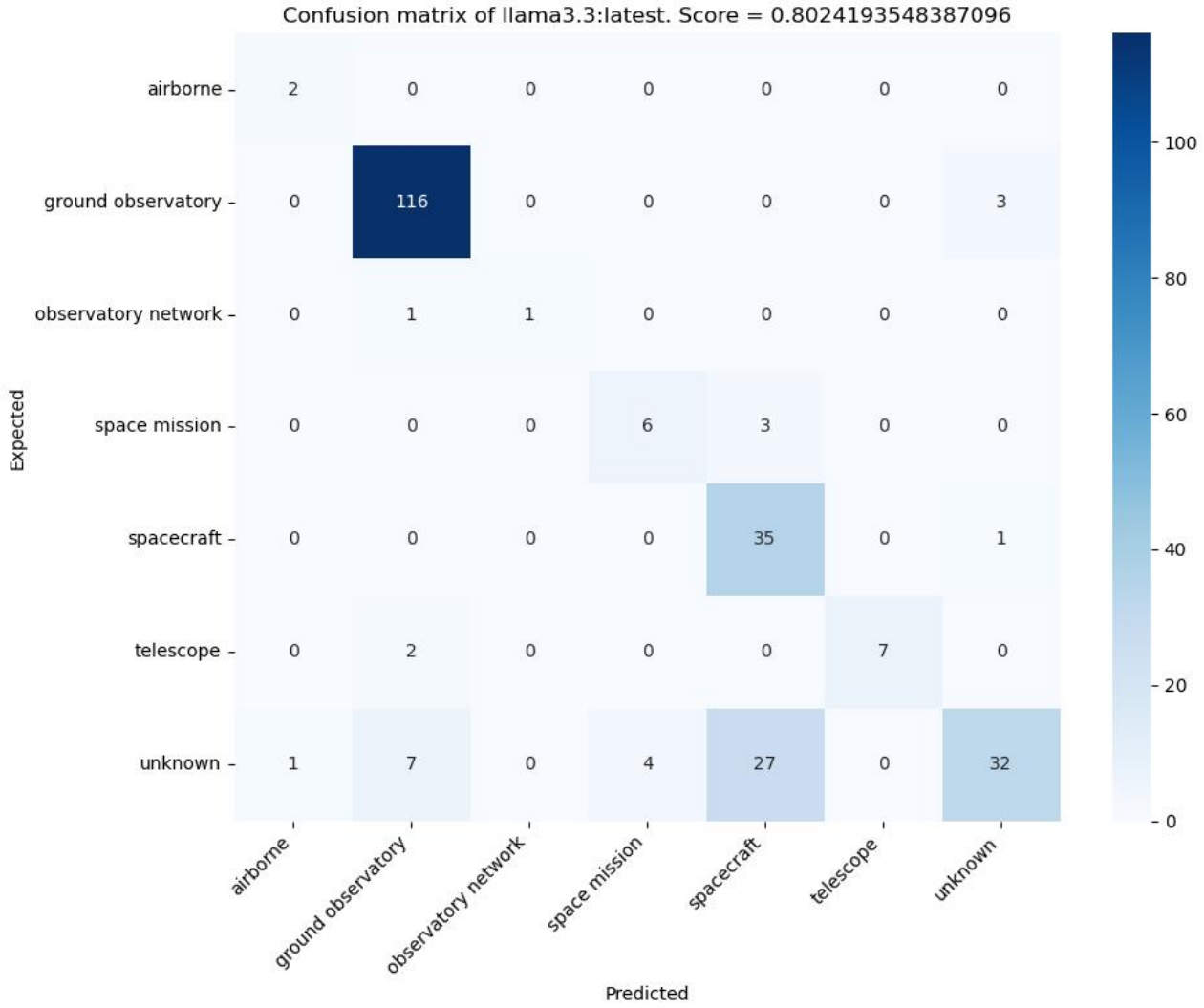
1. Facility lists



Extracted data from Wikidata for Alert Magnetic Observatory (Canada)

1. Facility lists

- Data augmentation for missing attributes :
  - Geopy => geolocalisation of ground entities
    - Latitude, longitude, address, city, country, continent
    - *0 ≤ location_confidence ≤ 1*
  - LLM => typing of entities with non-explicit types
    - type : telescope, spacecraft...
    - *type_confidence = 0*

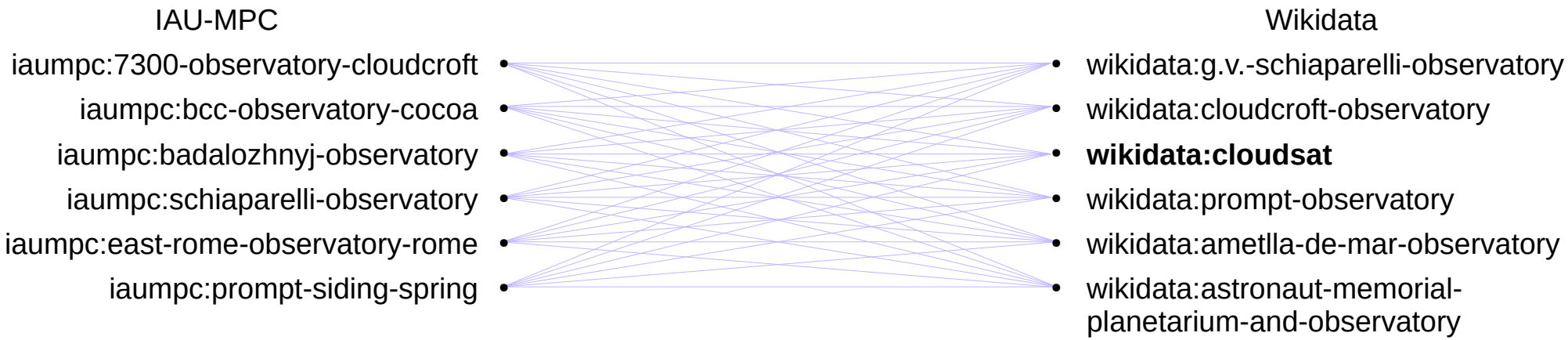


Automatic typing by LLM

2. Data mapping

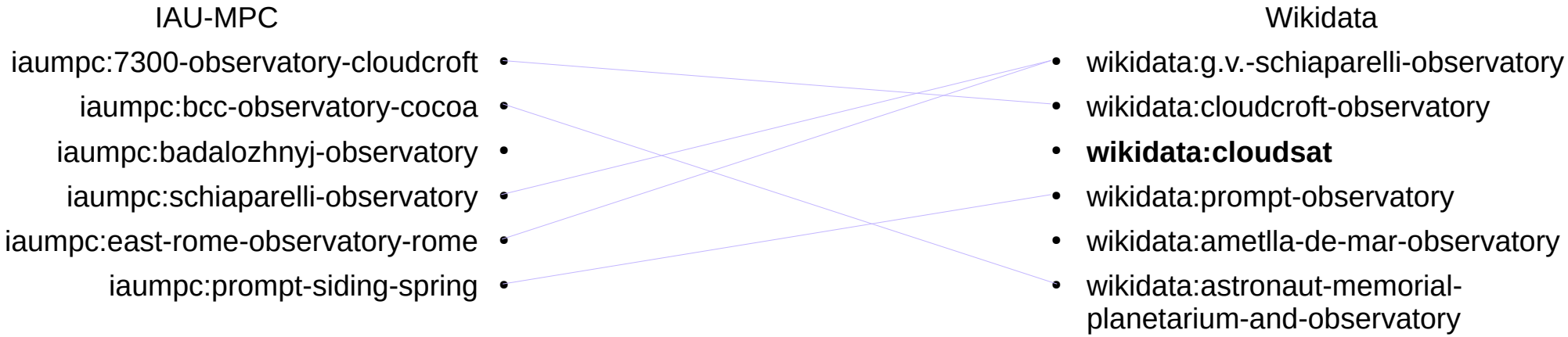
Step 1 : generate a full mapping between two lists

N × M ***Candidate Pairs***



2. Data mapping

Step 2 : compute **discriminant criteria*** & eliminate *n* incompatible pairs : N × M – *n_incompatible* ***Candidate Pairs***



*Discriminant criteria :

Geodesic distance > n km ;

≠ continent or country ;

≠ class (spacecraft, observatory, telescope...);

≠ launch date, start date or end date

2. Data mapping

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

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

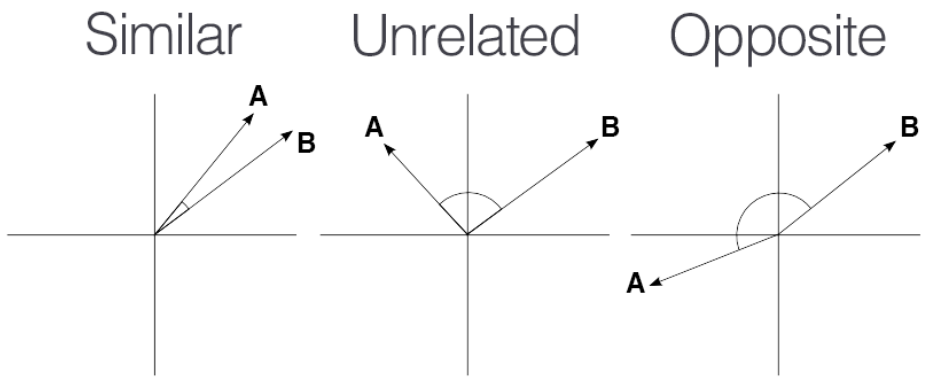
Source : Ted Mei (Medium)

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

Scores	*Effective fields*	*Multilingual support*
TF-IDF cosine similarity(*x, y*)	labels, description, definition	No
Sentence Transformer cosine similarity(*x, y*)	””	Yes
LLM embeddings 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

Global score
(weighted sum of the semantic scores)

Cosine Similarity



Source : Sindhu Selham (Medium)

LIRA Observatoire de Paris | PSL

2. Data mapping

Disambiguation algorithm

List 1 / List 2	list1: Pioneer XI	list1: HUYGENS PROBE	list1: Vega 2	list1: Voyager II
list2: GOES-16	0.73	0.27	0.51	0.26
list2: P-11	0.88	0.43	0.31	0.75
list2: VG-2	0.50	0.48	0.89	0.79
list2: VG-1	0.40	0.60	0.83	0.74

2. Data mapping

Disambiguation algorithm

List 1 / List 2	list1: Pioneer XI	list1: HUYGENS PROBE	list1: Vega 2	list1: Voyager II
list2: GOES-16	0.73	0.27	0.51	0.26
list2: P-11	0.88	0.43	0.31	0.75
list2: VG-2	0.50	0.48	0.89	0.79
list2: VG-1	0.40	0.60	**0.83**	**0.74**

2. Data mapping

Disambiguation algorithm

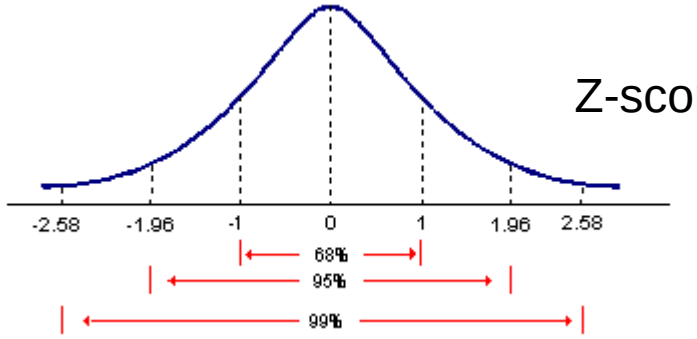
List 1 / List 2	list1: Pioneer XI	list1: HUYGENS PROBE	list1: Vega 2	list1: Voyager II
list2: GOES-16	0.73	0.27	0.51	0.26
list2: P-11	0.88	0.43	0.31	0.75
list2: VG-2	0.50	0.48	0.89	0.79
list2: VG-1	0.40	0.60	**0.83**	**0.74**

2. Data mapping

List 1 / List 2	list1: Pioneer XI	list1: HUYGENS PROBE	list1: Vega 2	list1: Voyager II
list2: GOES-16	0.73	0.27	0.51	0.26
list2: P-11	0.88	0.43	0.31	0.75
list2: VG-2	0.50	0.48	0.89	0.79
list2: VG-1	0.40	0.60	**0.83**	**0.74**

- Standardize columns' and rows' values
  - prevent an excessively high column or row (elsewise the disambiguation algorithm always proposes the same row or column's pairs, *e.g. « OBO », Oxford Belarus Observatory*)
- Validation at each Candidate Pair :
  - No validation (select the highest score each time)
  - Human validation (too long)
  - LLM validation through prompting : *« Are those two entities the same or distinct ? »*
- Threshold (stopping condition) :
  - Z-score : stop after 2.5 % of the best Candidate Pairs were reviewed
  - N times « distinct » in a row (N is proportional to data size)

Z-score



Source : analystnotes

3. LLM Validation

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

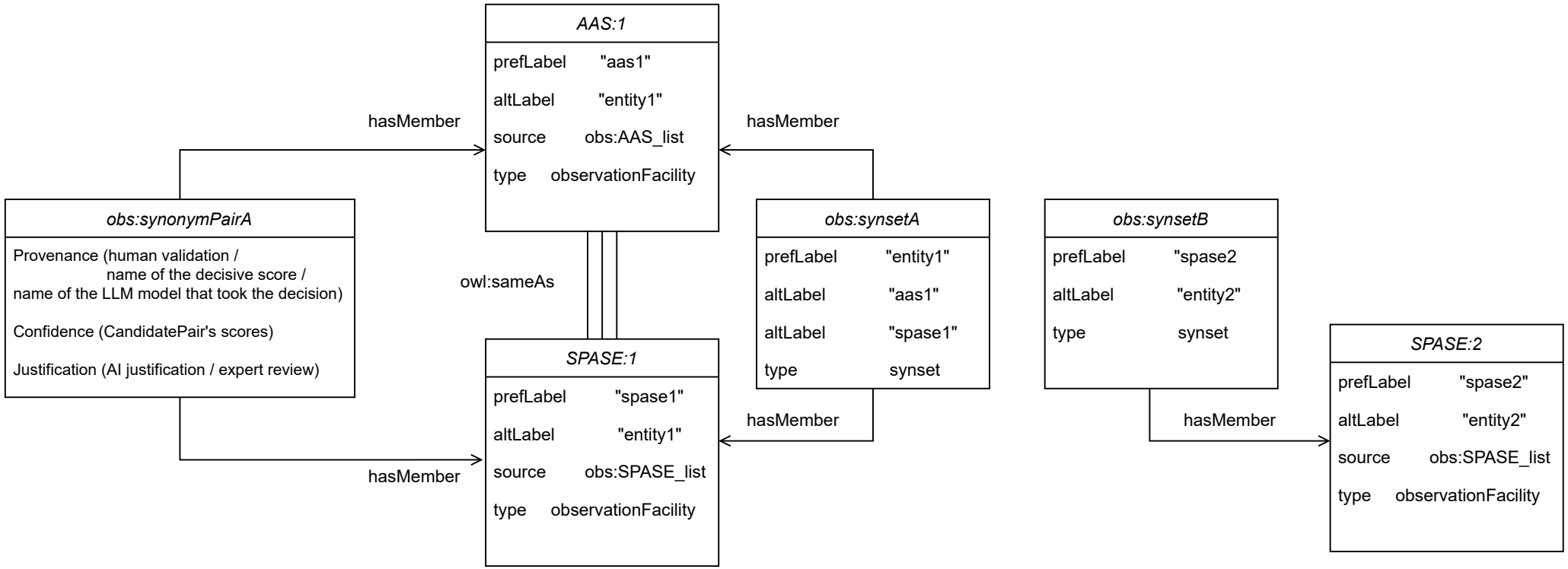
Negative match

{"**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."}

Positive match

```
obs:25182b58-3a43-4379-8b49-f92eb8c0429f a obs:SynonymPair ;
    obs:date "-1"^^xsd:float ;
    obs:distance "-1"^^xsd:float ;
    obs:fuzzy_levenshtein "1.0"^^xsd:float ;
    obs:global "0.7473960092740229"^^xsd:float ;
    obs:hasMember aas:the-nrao-300-foot-green-bank-telescope,
        wikidata:300-foot-radio-telescope ;
    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" ;
    obs:sentence_cosine_similarity "0.8889001607894897"^^xsd:float ;
    obs:tfidf_cosine_similarity "0.5301106605407628"^^xsd:float .
```

4. Output format & Applications



4. Output format & Applications

List of distinct observation facilities, following IVOA specification :
https://www.ivoa.net/documents/Vocabularies/20230206/REC-Vocabularies-2.1.html#tth_sEcA.1

CSV :

Term ; level ; label ; description ; more_relations

japan-aerospace-exploration-a...	1	Japan Aerospace eXploration Agency Hinotori Sol...	
japan-aerospace-exploration-a...	1	Japan Aerospace eXploration Agency Tenma X-ray	
jodrell-bank-centre-for-astroph...	1	Jodrell Bank Centre for Astrophysics 76.2m Lovell Tel.	skos:broader(jodrell-bank-observatory)
jodrell-bank-observatory	1	Jodrell Bank Observatory	skos:sameAs("https://wikidata.org/wiki/Q1569783")
johns-hopkins-university-0...	1	Johns Hopkins University 0.9m Hopkins Ultraviolet	
kamioka-gravitational-wave-d...	1	Kamioka Gravitational wave detector, Large-scale...	skos:sameAs("https://wikidata.org/wiki/Q725081")
kao	1	KAO	skos:sameAs("https://wikidata.org/wiki/Q4329973")
karl-schwarzschild-observatory	1	Karl Schwarzschild Observatory	skos:sameAs("https://wikidata.org/wiki/Q646078")
kashima-space-research-cen...	1	Kashima Space Research Center 34m Radio Telesco...	
kgj-vla	1	KGJ VLA	skos:sameAs("https://wikidata.org/wiki/Q461382")
kiepenheuer-institut-fur-sonn...	1	Kiepenheuer-Institut fur Sonnenphysik 0.70m Vac...	skos:broader(observatorio-del-teide)
kilodegree-extremely-little-tel...	1	Kilodegree Extremely Little Telescope at Winer O...	
kirkland-airforce-base	1	Kirkland airforce base	
kitt-peak-bok-observatory	1	Kitt Peak-Bok observatory	skos:sameAs("https://wikidata.org/wiki/Q116162881") skos:broader(...
korea-astronomy-and-space-sc...	1	Korea Astronomy and Space Science Institute 0.6...	skos:broader(soao)

4. Output format & Applications

Name Resolver


```
"david-dunlap-observatory": [
  "David Dunlap Observatory",
  "David Dunlap Observatory (DDO)",
  "Q531487",
  "779",
  "DDO"
],
"manastash-ridge-observatory": [
  "Manastash Ridge Observatory",
  "Manastash Ridge Observatory (MRO)",
  "Q6747042",
  "664",
  "MRO"
```


Query : `"DDO"`

Answer : `"david-dunlap-observatory"`

Conventional label : `"David Dunlap Observatory"`

Aliases : `"Q531487", "779", "DDO", "David Dunlap Observatory (DDO)"`

Thank you

LIRA Observatoire de Paris | PSL