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

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

[1]LIRA, Observatoire de Paris, Université PSL, Sorbonne Université, Université Paris Cité, CY Cergy Paris Université, CNRS, 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

1. Introduction

1.1. Context

Astrophysics brings together a wide range of dynamic communities: heliophysicists, planetary scientists, cosmologists, data scientists, instrument engineers, etc. These diverse experts collaborate across disciplines to tackle complex questions about the Universe. The astrophysics community has also been pioneering open-science with the early inception of the so-called *Virtual Observatory* [1, 2] in the early 2000's, which defines open and interoperable data and access standards for astronomy, that are maintained and developed by the IVOA [3]. Similar initiatives also exist for heliophysics (International Heliophysics Data Environment Alliance, IHDEA) [4, 5] and planetary sciences (International Planetary Data Alliance, IPDA) [6]. This interdisciplinary synergy is further amplified by the rapid development of computational tools, and globally shared data standards and open-access practices, such as those developed within the Research Data Alliance [7].

As data becomes increasingly interconnected, there is a growing need to ensure its interoperability. Specifically, the data discovery part of the science workflow is greatly facilitated when the data providers are using metadata with the same schema and terms from the same vocabularies. However, research institutions often do not use consistent naming conventions, particularly when referring to astronomy observation facilities (equivalent to the *platform* concept defined in OGC SensorML[1]), some refer to a telescope by its diameter and/or location, others by its nickname, and many variations exist that combine both. Additionally, historical labels (like "Mariner 11" for "Voyager 1") for the same facility can introduce further ambiguity, requiring a deeper understanding of the definitions and contexts around these entities.

OM 2025: The 20th International Workshop on Ontology Matching collocated with the 24th International Semantic Web Conference (ISWC 2025), November 2nd, 2025, Nara, Japan

liza.fretel@obspm.fr (L. Fretel); baptiste.cecconi@obspm.fr (B. Cecconi); laura.debisschop@obspm.fr (L. Debisschop)

https://github.com/Sazuna (L. Fretel)

0009-0001-4600-9954 (L. Fretel); 0000-0001-7915-5571 (B. Cecconi); 0000-0003-4688-6575 (L. Debisschop)

[1]Open Geospatial Consortium Sensor Model Language: https://docs.ogc.org/is/23-000/23-000.html

The main objective of this work is to propose a method to align multiple lists of astronomical observation facilities, standardize their aliases, and suggest a unique preferred label per physical entity. The source code of the functionalities described in this paper are available on GitHub [8]. The data can be re-downloaded by running *update.py*. A snapshot of the cached pages on 2025/09/04 is available on Zenodo [9].

1.2. Applications

The primary goal of this work is to enable smooth data discovery across data providers, in the scope of the several astronomy open data ecosystems (including the IVOA, IHDEA and IPDA). We thus need to propose, firstly, a vocabulary of commonly agreed terms, that shall be used to expose metadata in search interfaces; and secondly, a name resolver that suggests matching entities from the adopted vocabulary.

The implementation of the name resolver relies on an Elasticsearch API, powered by a JSON dictionary (example in Appendix G) that includes the preferred labels of observation facilities along with their aliases. The name resolver will have two interfaces: a *resolve* endpoint, which takes any input string and provides an ordered list of matching terms; and an *aliases* endpoint, which takes a term from the vocabulary and returns the list of known aliases. That second endpoint is also useful for expanding data discovery to databases that have not adopted the proposed vocabulary yet. In order to implement refined search and discovery, we also include a search by meronymy relation so that the data discovery is not restricted to the granularity choice of the data provider (for instance, a user searching for observations conducted by the *Voyager* space mission, shall also find the *Voyager 1* and *Voyager 2* spacecraft).

In parallel to this work, we develop OntoPortal-Astro[2] [10] to share astronomy ontologies while ensuring the FAIRness of their content. It aligns with the OntoPortal Alliance, a broader initiative to build portals for domain-specific ontologies, such as EarthPortal https://earthportal.eu/ and AgroPortal https://agroportal.lirmm.fr/. The ontology produced by our multi-source mapping will later be shared on OntoPortal-Astro, allowing the users to access statements from diverse resources and to track the statements' origin.

Finally, we also output an observation facilities list following the CSV format defined by the IVOA [11], so that it can be processed and listed on the IVOA vocabulary page[3]. This CSV file lists every physical entity and the meronymy relations between them.

1.3. Related works

The previous version of our algorithm [12] used to set Wikidata as a *pivot* ontology, trying to map every other ontology to it, but our attempts were unsuccessful, due to the non-exhaustiveness of Wikidata. That is why we introduced the mapping strategy, that permits to link resources in any order and on any criterion. The necessity of a multi-strategy approach to map plural and heterogeneous entity collections was highlighted in the original RiMOM paper [13].

In our work, we employ an LLM to validate or invalidate a candidate pair, which is composed by two semantically close entities. This method was exploited by LLM-Align [14], a framework that embeds entities and selects the k-nearest target entities before asking an LLM to decide which target entity matches the source entity.

2. Data updating

2.1. Data sources, processing and updates

Currently, we only processed eight out of 19 identified vocabularies [15]: Wikidata [16], AAS (American Astronomy Society), PDS (Planetary Data System), IAU-MPC[4] (The International Astronomical Union

[2] https://ontoportal-astro.eu
[3] https://ivoa.net/rdf
[4] IAU-MPC entities are referred to with an identifier in Wikidata (Minor Planet Center ID)

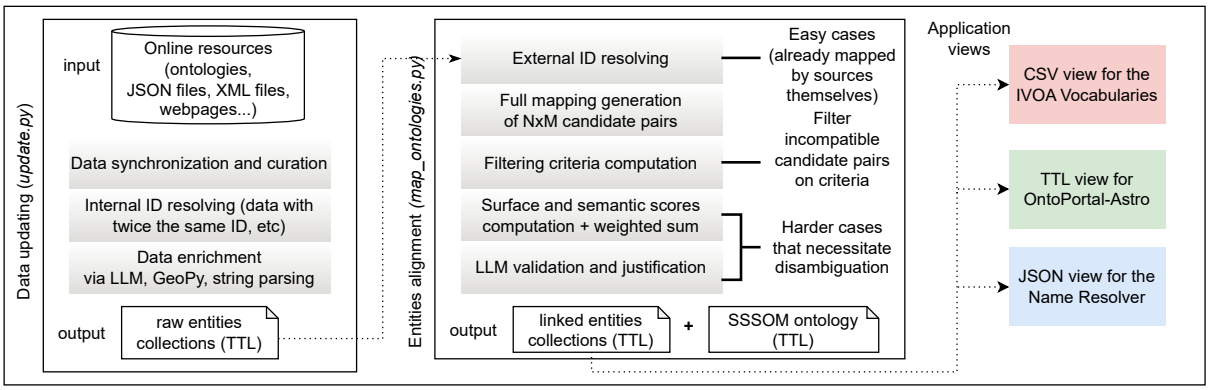

Figure 1: Data processing pipeline. During data updating (**left**), described in section 2, we collect observation facilities' records and save them in turtle files. Those files are used as inputs of the entity alignment steps (**right**), described in sections 3, 4 and 5. This outputs a linked ontology containing all entities from each list with their matching relations (*SKOS:exactMatch*), along with an associated SSSOM ontology. The application views can be generated from the linked ontology. Their purposes are explained in the subsection 1.2.

Minor Planet Center), NAIF[5] (The Navigation and Ancillary Information Facility) , NSSDC[6] (NASA Space Science Data Coordinated Archive), SPASE (Space Physics Archive Search and Extract) and IMCCE (Celestial Mechanics and Ephemerides Calculus Institute).The extracted observation facilities fit into five classes: *Telescope, Observatory, Spacecraft, Airborne platform*, and *Investigation.* Those classes are convenient for the ontology mapping task, but do not embody the complexity of the taxonomy of observation facilities. We might consider adding other classes when needed.

2.1.1. Data synchronization and curation

In the previous version of the data format (detailed in [12]), data were not typed and only labels and identifiers were mapped with a threshold on the Levenshtein similarity [17]. However, in doing so, the labels did not provide enough information to be properly disambiguated. To counter this, we extracted every field that might add a semantic or numerical value to the entities by enriching them with a class if possible, a description and/or definition, an observation waveband, a launch date for spacecraft, a funding agency, a latitude and longitude for ground-based facilities, etc. A sample of the extracted data is available in Appendix B.

One of the nice-to-have features mentioned in [12] is the automated update feature. We developed a *version manager* that detects any change on an entity whenever *update.py* is called with the *–no_cache* argument. It updates its *DCTERMS:Modified* value and sets a *Deprecated* flag to entities that no longer exist in their original source.

After data were synchronized, we fix the identified errors in the data (data curation step).

2.1.2. Internal ID resolving

There are several constraints regarding the URIs management. Since the resulting data will be mainly exploited in the IVOA ecosystem, we have to conform to their specifications, such as the use of the facilities' labels to generate human-readable URIs for historical reasons. This is an issue for sources that have non-unique labels, which we solved by adding keywords or by re-using the source's identifiers as URIs, like in IAU-MPC or PDS, in which an investigation and a spacecraft are distinct records but share the same label. Moreover, we defined a namespace per source (see Appendix A for namespaces used in this document) to prevent entities from two resources but with the same label to end up with the same URI. In addition, the NAIF list contains non-unique identifiers for the same or different entities, which required domain experts to be resolved.

[5]NAIF entities are referred to with an identifier in Wikidata (NAIF ID).
[6]NSSDC entities are referred to with an identifier in Wikidata (COSPAR ID and NSSDCA ID)

2.1.3. Data enrichment

Finally, the data enrichment step consists of using tools to fill missing attributes. As most of the entities' classes can be deduced from their metadata (Wikidata), their URL (PDS) or the web page layout (NAIF), some of the lists (SPASE) do not explicitly classify entities. We prompted an LLM to classify non-explicitly typed resources into our five categories (six with the *Unknown* class for entities that do not belong to any category, like space military facilities). Unfortunately, our best attempt with the model LLama3.1-8B [18] only reached 80.20% accuracy on the classification task, and the results were too sensitive to the prompt. Because we did not want to introduce biases at the early stages of the mapping process, we decided to consider those entities as untyped, therefore we will try to pair them with every entity instead of creating one sub-mapping per type (e.g. a mapping between IAU-MPC and Wikidata's observatories).

Then, we used Geopy (version 2.4.1) to retrieve the ground-based entities' location information such as a detailed address, latitude and longitude from the entities' metadata. The address can benefit the semantic scores because some facilities share their name with their street, city or country.

String parsing is used for example on PDS spacecraft's aliases to extract their NSSDC identifiers and infer their launch years, as well as on AAS facilities to extract their apertures (size of the primary mirror or the input lens of a telescope) and some aliases. An example of data enrichment can be found in C.

3. Mapping strategy

The mapping strategy is a configuration file that allows the engineer to program a mapping path (see an example in Appendix E) on a certain type and on certain scores only, proceeding by pairs of lists. In this way, we iteratively increase the size of the synonym sets, which are made of validated synonym pairs, enhancing the subsequent mapping steps.

The mapping strategy splits the entities alignment problem into three levels of difficulty. Those levels are displayed in the right part of Figure 1). When being confronted to two sets of entities, we start with the most obvious alignments — external identifiers (external ID resolving step), linking entities that are already refer to each other. After that, we generate a full mapping with the remaining single entities (second step). In this step, $N \times M$ candidate pair objects are generated, with N and M varying between 561 (PDS) and 10.000 (IMCCE) entities. We then apply discriminant criteria (third step), reducing the complexity depending on the criteria and the lists' features, and we end up with the most complex mappings, which require the computation of surface and semantic scores, followed by LLM validation to disambiguate them (the last two steps).

4. Discriminant criteria and scores

For each line of the mapping strategy, the algorithm will apply a set of filtering criteria, surface and semantic scores on the candidate pairs.

4.1. Filtering criteria

Filtering criteria are elimination rules that will accept or disqualify candidate pairs in prior to any further computation. This has two purposes: reduce the mapping's complexity by removing incompatible pairs; and prevent incoherent decisions during the subsequent steps. It includes a label match (accepting criterion)[7], a mismatch in location, date, class, aperture size, or identifier (rejecting criteria). For instance, the "date" filter compares and disqualifies two spacecraft with different launch years. Ground-based facilities like telescopes or observatories often have an associated latitude and longitude, allowing a geodesic distance computation. We set the maximum distance between two entities to 4 km to account for rounding tolerance; beyond that, they are considered distinct.

[7]The label match works well on spacecraft and investigations, as their names do not vary much.

4.2. Similarity scores

Once this preliminary step is done, only compatible pairs are left. To disambiguate, we apply some linear similarity scores, that take different aspects of the entities into account.

4.2.1. Surface scores

The Levenshtein distance [17] is an edition distance between two strings. By applying the following formula, we obtain a similarity score between 0 and 1 for strings $|s_1|$ and $|s_2|$:

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

For example, it outputs a high score between "observatory" and "observatoire" (French), therefore it is able to detect slight translation variations or typos. For the digits match score, we extract all numbers from the entity's strings with a regular expression, apply truncation and rounding to match numbers from both entities and compute a matching ratio. The acronym probability score computes the probability of a label to be the acronym of another label.

4.2.2. Semantic scores

The TF-IDF (term-frequency inverse document frequency) is used in many OM frameworks [19].

$$\textit{tf-idf}_{i,j} = \textit{tf}_{i,j} \times \log\left(\frac{N}{\textit{df}_i}\right)$$

In the context of OM, the term frequency *tf* is how often a token appears in an entity's textual fields; N is the amount of entities that constitute the corpus; *df*, the document frequency, is how many times the token appears in the ontology's textual fields. It emphasizes the importance of rarer tokens such as proper nouns, while lowering the impact of recurring tokens like "mission", but it does not embed the synonyms and sentences' meaning. To train a TF-IDF encoder, we simulate a reference corpus by extracting all of the textual fields (labels, descriptions and definitions) and filter out English, Spanish and French stop words. After that, we can encode the source and the target entities and compute a cosine similarity between their vectors (A and B):

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

To encode entities, we also implemented the sentence transformer's cosine similarity. We chose the all-Mini-LM-L6-v2 model. It was pre-trained on a general and multilingual corpus. After entities are vectorized and saved on the disk, we compare them by computing their cosine similarity.

Some LLMs provide an encoding function that generates embeddings for a given text, allowing entities comparison via the cosine similarity. We tested different LLMs of different sizes: LLama3.1-8B [18] (8 billion parameters), DeepSeek-V3 [20] (671 billion parameters, version `deepseek-v3:671b-q4_K_M`) and Astrollama [21] (7 billion parameters, fine-tuned on 300.000 arXiv articles). Despise being fine-tuned on astronomy data, it did not outperform DeepSeek and its embeddings were too space-greedy.

4.2.3. Global score

By combining those scores via a weighted sum, we obtain a global score on each candidate pair:

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

where p is a candidate pair, s_i is a score value (for instance the TF-IDF's cosine similarity score) and w_i is its score's weight, a constant that we fix beforehand (for instance we set 0.5 for the Levenshtein similarity). Unlike in a traditional vector space, which is usually tied to a single unique similarity score, this encompasses both the surface and the semantic aspect of the data.

5. Iterative validation by LLM prompting

For each pair, starting with the one with the highest global score, we prompt an LLM to accept or reject the pair and justify its decision [22]. In the prompt, both entities are represented by a string containing all of their textual features. We also give the instruction to consider a narrower distinct from its broader entity. After rejecting a certain amount of pairs in a row, we interrupt the process and output an auxiliary SSSOM ontology [23], which keeps track of the mappings' decision time, LLM justifications, which scores were the decisive ones, etc., for each positive match. Two SSSOM match examples can be found in Appendix F.

To evaluate the LLM validation, we used the AAS and PDS facility lists, that share similar features (aperture, naming convention, etc). We annotated a list of 30 pairs of potential matches with compatible features with a "same|distinct" label. This experiment was made using DeepSeek-V3 [20] (671 billion parameters, version `deepseek-v3:671b-q4_K_M`) as the candidate pairs' validator. Out of 30 candidate pairs, none were wrongly reviewed: 19 were true positives and 11 true negatives.

6. Conclusion and perspectives

In this work, we have introduced the mapping strategy methodology to perform data alignment of astronomy observation facilities. Our contribution mostly consists of the development of an adaptable mapping strategy combining filtering criteria, surface and semantic scores and the iterative aspect of the mapping, enhancing the semantic similarity capabilities by adding new information as we discover new synonyms.

As we focused on a limited amount of resources, in the future, we would like to align more vocabularies together. Furthermore, we are currently annotating a dataset of candidate pairs with a "same|distinct" annotation with our European collaborators for the disambiguation task on observation facilities. It will help giving insights about each scores' relevance to different mappings as well as evaluating different LLMs on the validation task. Currently, we use a generalist LLM to validate the candidate pairs and justify its decision; but in the future, we hope to fine-tune a Small Language Model or any relevant architecture by using this annotated dataset, that could run quicker and outperform DeepSeek due to its training on the specific task and data. Lastly, we are going to explore the use of an agentic MCP (Model Context Protocol) server to allow the validation LLM to search for further information online about an entity and solicit a human expertise when it is unsure about a candidate pair. This could benefit the quality of the resulting mapping, countering the lack of information of some resources.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly and Chat-GPT in order to grammar and spell check, and improve the text readability. After using the tool, the authors reviewed and edited the content as needed to take full responsibility for the publication's content.

Acknowledgments

This work has been supported by: the Europlanet 2020 Research Infrastructure (EPN2020-RI) and Europlanet 2024 Research Infrastructure (EPN-2024-RI) projects, which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654208 and 871149, respectively; the FAIR-IMPACT project, which received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement no 101057344; and OPAL cascading grant from the the OSCARS project, which received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement no 101129751. The authors also acknowledge support from CNRS and Observatoire de Paris, and especially Stéphane Aicardi and Philippe Hamy, from the Direction Informatique de l'Observatoire (DIO). They also thank Mireille Louys, Emmannuelle Perret and Sébastien Derrière, from CDS (Centre de Données Astronique de Strasbourg, France); and Markus Demleitner (University of Heidelberg, Germany).

References

[1] F. Genova, D. Egret, O. Bienaymé, F. Bonnarel, P. Dubois, P. Fernique, G. Jasniewicz, S. Lesteven, R. Monier, F. Ochsenbein, M. Wenger, The CDS information hub. On-line services and links at the Centre de Données astronomiques de Strasbourg, Astron. Astrophys. Suppl. 143 (2000) 1–7. doi:10.1051/aas:2000333.

[2] C. Arviset, M. Allen, A. Aloisi, B. Berriman, C. Boisson, B. Cecconi, D. Ciardi, J. Evans, G. Fabbiano, F. Genova, T. Jenness, B. Mann, T. McGlynn, W. OMullane, D. Schade, F. Stoehr, A. Zacchi, The VO: A powerful tool for global astronomy, 2018. doi:10.48550/arXiv.1803.07490.

[3] C. Arviset, S. Gaudet, IVOA TCG, The IVOA Architecture, in: P. Ballester, D. Egret, N. P. F. Lorente (Eds.), Astronomical Data Analysis Software and Systems XXI, volume 461 of *Astronomical Society of the Pacific Conference Series*, 2012, p. 259.

[4] D. A. Roberts, J. Thieman, V. Génot, T. King, M. Gangloff, C. Perry, C. Wiegand, D. De Zeeuw, S. F. Fung, B. Cecconi, S. Hess, The SPASE Data Model: A Metadata Standard for Registering, Finding, Accessing, and Using Heliophysics Data Obtained From Observations and Modeling, Space Weather 16 (2018) 1899–1911. doi:10.1029/2018SW002038.

[5] S. F. Fung, A. Masson, L. F. Bargatze, T. King, R. Ringuette, R. M. Candey, C. Wiegand, L. K. Jian, D. De Zeeuw, K. Muglach, R. M. McGranaghan, D. Aaron Roberts, B. Cecconi, N. André, V. Génot, J. Vandegriff, M. A. Reiss, SPASE metadata as a building block of a heliophysics science-enabling framework, Advances in Space Research 72 (2023) 5707–5752. doi:10.1016/j.asr.2023.09.066.

[6] S. Slavney, R. Beebe, D. Crichton, S. Hughes, J. Zender, The International Planetary Data Alliance, in: 38th Annual Lunar and Planetary Science Conference, Lunar and Planetary Science Conference, 2007, p. 1336.

[7] R. Showstack, Initiative to establish Research Data Alliance moves forward, EOS Transactions 93 (2012) 354–354. doi:10.1029/2012EO370002.

[8] L. Fretel, L. Debisschop, B. Cecconi, M. Louys, E. Perret, T. Al-Ubaidi, epn-vespa/facilitylist, 2025. URL: https://doi.org/10.5281/zenodo.17199128. doi:10.5281/zenodo.17199128.

[9] L. Fretel, B. Cecconi, L. Debisschop, Cache folder of FacilitiesList, 2025. URL: https://doi.org/10.5281/zenodo.17078681. doi:10.5281/zenodo.17078681.

[10] B. Cecconi, L. Debisschop, S. Derrière, M. Louys, C. Corre, N. Grau, C. Jonquet, OntoPortal-Astro, a Semantic Artefact Catalogue for Astronomy, Astronomy and Computing 53 (2025) 100991. doi:10.1016/j.ascom.2025.100991. arXiv:2504.12897.

[11] M. Demleitner, N. Gray, M. Taylor, Vocabularies in the VO Version 2.1, IVOA Recommendation 06 February 2023, 2023. doi:10.5479/ADS/bib/2023ivoa.spec.0206D.

[12] B. Cecconi, L. Debisschop, M. Louys, E. Perret, M. Demleitner, Using Wikidata for an Observation Facility Vocabulary, IVOA Note, Version 1.0, 2023. URL: https://www.ivoa.net/documents/ObsFacilityWikidata/20231115, semantics Working Group, published 2023-11-15. Editor: Baptiste Cecconi.

[13] J. Li, J. Tang, Y. Li, Q. Luo, RiMOM: A Dynamic Multistrategy Ontology Alignment Framework, IEEE Trans. Knowl. Data Eng. 21 (2009) 1218–1232. doi:10.1109/TKDE.2008.202.

[14] X. Chen, T. Lu, Z. Wang, LLM-Align: Utilizing Large Language Models for Entity Alignment in Knowledge Graphs, 2024. URL: https://arxiv.org/abs/2412.04690. arXiv:2412.04690.

[15] L. Fretel, B. Cecconi, L. Debisschop, Generating the Observation Facilities Vocabulary, 2025. doi:10.5281/zenodo.15862784.

[16] D. Vrandečić, M. Krötzsch, Wikidata: A Free Collaborative Knowledgebase, Commun. ACM 57 (2014) 78–85. doi:10.1145/2629489.

[17] V. I. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, Soviet Physics Doklady 10 (1966) 707.

[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, 2023. arXiv:2302.13971.

[19] M. Gulić, I. Magdalenić, B. Vrdoljak, Ontology matching using tf/idf measure with synonym recognition, in: T. Skersys, R. Butleris, R. Butkiene (Eds.), Information and Software Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 22–33.

[20] DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Wang, J. Chen, J. Chen, J. Yuan, J. Qiu, J. Li, J. Song, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Wang, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Wang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Zhang, R. Pan, R. Wang, R. Xu, R. Zhang, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ye, S. Ma, S. Wang,

S. Zhou, S. Yu, S. Zhou, S. Pan, T. Wang, T. Yun, T. Pei, T. Sun, W. L. Xiao, W. Zeng, W. Zhao, W. An, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Zhang, X. Chen, X. Nie, X. Sun, X. Wang, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yu, X. Song, X. Shan, X. Zhou, X. Yang, X. Li, X. Su, X. Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Y. Zhang, Y. Xu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Yu, Y. Zheng, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Tang, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Wu, Y. Ou, Y. Zhu, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Zha, Y. Xiong, Y. Ma, Y. Yan, Y. Luo, Y. You, Y. Liu, Y. Zhou, Z. F. Wu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Huang, Z. Zhang, Z. Xie, Z. Zhang, Z. Hao, Z. Gou, Z. Ma, Z. Yan, Z. Shao, Z. Xu, Z. Wu, Z. Zhang, Z. Li, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Gao, Z. Pan, DeepSeek-V3 Technical Report, 2025. arXiv:2412.19437.

[21] T. D. Nguyen, Y.-S. Ting, I. Ciucă, C. O'Neill, Z.-C. Sun, M. Jabłońska, S. Kruk, E. Perkowski, J. Miller, J. Li, J. Peek, K. Iyer, T. Różański, P. Khetarpal, S. Zaman, D. Brodrick, S. J. R. Méndez, T. Bui, A. Goodman, A. Accomazzi, J. Naiman, J. Cranney, K. Schawinski, UniverseTBD, AstroLLaMA: Towards Specialized Foundation Models in Astronomy, 2023. arXiv:2309.06126.

[22] R. Peeters, A. Steiner, C. Bizer, Entity Matching using Large Language Models, 2024. arXiv:2310.11244.

[23] N. Matentzoglu, J. P. Balhoff, S. M. Bello, C. Bizon, M. Brush, T. J. Callahan, C. G. Chute, W. D. Duncan, C. T. Evelo, D. Gabriel, J. Graybeal, A. Gray, B. M. Gyori, M. Haendel, H. Harmse, N. L. Harris, I. Harrow, H. B. Hegde, A. L. Hoyt, C. T. Hoyt, D. Jiao, E. Jiménez-Ruiz, S. Jupp, H. Kim, S. Koehler, T. Liener, Q. Long, J. Malone, J. A. McLaughlin, J. A. McMurry, S. Moxon, M. C. Munoz-Torres, D. Osumi-Sutherland, J. A. Overton, B. Peters, T. Putman, N. Queralt-Rosinach, K. Shefchek, H. Solbrig, A. Thessen, T. Tudorache, N. Vasilevsky, A. H. Wagner, C. J. Mungall, A Simple Standard for Sharing Ontological Mappings (SSSOM), Database 2022 (2022) baac035. doi:10.1093/database/baac035.

A. Table of namespaces used in this paper

Sematic Artefact name	Namespace	URI
American Astronomical Society	AAS	https://voparis-ns.obspm.fr/rdf/obsfacilities/aas#
Dublin Core Metadata Initiative (DCMI)	DCTERMS	http://purl.org/dc/terms
WGS84 Geo Positioning	GEO1	http://www.w3.org/2003/01/geo/wgs84_pos#
Observation Facilities (our NS)	OBSF	https://voparis-ns.obspm.fr/rdf/obsfacilities#
Planetary Data System	PDS	https://voparis-ns.obspm.fr/rdf/obsfacilities/pds#
Resource Description Framework Schema	RDFS	http://www.w3.org/2000/01/rdf-schema#
Schema	SCHEMA	https://schema.org/
Semantic Mapping Vocabulary	SEMAPV	https://w3id.org/semapv/vocab/
Simple Knowledge Organization System	SKOS	http://www.w3.org/2004/02/skos/core#
A Simple Standard for Sharing Ontology Mappings	SSSOM	https://w3id.org/sssom/

Note The namespaces of the https://voparis-ns.obspm.fr/rdf/ domain are temporary ones, and are subject to change in further versions.

B. Extracted data

After collecting data from multiple semantic artifacts, we standardize them into a unique ontology.

```
aas:european-southern-observatory-1.52m-telescope-at-la-silla-observatory
        a obsf:observatory ;
    dcterms:isPartOf aas:la-silla-observatory ;
    geo1:latitude "-29.2552104"^^xsd:float ;
    geo1:location "South America" ;
    geo1:longitude "-70.739507"^^xsd:float ;
    skos:exactMatch pds:1.52-m-spectrographic-cassegrain-coude-reflector ;
    skos:notation "ESO:1.52m" ;
    skos:prefLabel "European Southern Observatory 1.52m Telescope at La Silla
        Observatory" ;
    obsf:aperture "1.52m" ;
    obsf:waveband wb:infrared,
```

```
        wb:optical .


pds:1.52-m-spectrographic-cassegrain-coude-reflector a obsf:telescope ;
   dcterms:description "The 1.52-m spectrographic Cassegrain/Coude reflector is a 1.52 m
    telescope located at -29.255028, 289.267975 at the European Southern Observatory.
    Operational 07/1968+" ;
    dcterms:isPartOf pds:european-southern-observatory,
        pds:european-southern-observatory-la-silla ;
    geo1:latitude "-29.255028"^^xsd:float ;
    geo1:location "Earth" ;
    geo1:longitude "289.267975"^^xsd:float ;
    skos:altLabel "urn:nasa:pds:context:telescope:eso.1m52" ;
    skos:notation "urn:nasa:pds:context:telescope:eso-la_silla.1m52",
        "urn:nasa:pds:context:telescope:eso.1m52" ;
    skos:prefLabel "1.52-m spectrographic Cassegrain/Coude reflector" ;
   schema:url "https://pds.nasa.gov/data/pds4/context-pds4/telescope/eso-la_silla.1m52_1.1.xml" ;
    obsf:altitude "2347" ;
    obsf:aperture "1.52m" ;
    obsf:coordinate_source "Astronomical" ;
```

C. Data enrichment by string parsing

```
aas:nasa-0.85m-spitzer-space-telescope a obsf:telescope ;
    geo1:location "Space" ;
   skos:altLabel "NASA 0.85m Spitzer Space Telescope (SST formerly Space Infrared Telescope Facility
        "NASA 0.85m Spitzer Space Telescope (SST)",
        "SIRTIF",
        "SST",
        "Space Infrared Telescope Facility",
        "Space Infrared Telescope Facility (SIRTIF)" ;
    skos:notation "Spitzer" ;
    skos:prefLabel "NASA 0.85m Spitzer Space Telescope" ;
    obsf:aperture "0.85m" ;
    obsf:location_confidence "0.5"^^xsd:float ;
    obsf:source obsf:aas_list ;
    obsf:type_confidence "1"^^xsd:float ;
    obsf:waveband wb:infrared .
```

The alternate labels were extracted from the full label "NASA 0.85m Spitzer Space Telescope (SST formerly Space Infrared Telescope Facility or SIRTIF) Satellite Mission" using string parsing, as well as the telescope aperture (0.85m) which relies on a regular expression.

D. Scores available for mapping strategies

Score name	Type	Description
Label match	Accept	Entities match if any or their aliases are equal.
Identifier	Reject	Eliminate candidate pairs when one of their identifiers mismatch (NAIF ID), COSPAR ID, NSSDCA ID).
Distance limit	Reject	Eliminate ground-based facilities if their geodesic distance > 4km.
Date mismatch	Reject	Eliminate launch, start and/or end date mismatch.
Aperture mismatch	Reject	Eliminate facilities with a different lens aperture.
Acronym probability	Surface	Probability of a label to be the acronym of another label.
Levenshtein similarity	Surface	Edit distance between labels. Can detect typos or small naming divergences.
Digits Match	Surface	Digits match ratio between all fields.
TF-IDF	Semantic	Encode each word of the entity's textual fields and compute a cosine similarity. BoW approach, not multilingual, not synonym-aware.
Sentence transformer	Semantic	Encode semantic fields' sentences and compute a cosine similarity. Multilingual, synonym-aware.
LLM embeddings	Semantic	Cosine similarity on embeddings produced by an LLM. Multilingual, synonym-aware.

E. Mapping strategy configuration file

This strategy does not use neural network scores (sentence transformer or LLM embeddings) in order to run faster.

```
iaumpc, wikidata[spacecraft]: label_match, identifier, levenshtein, tfidf, digit
iaumpc, wikidata[all,-spacecraft]: label_match, identifier, distance, type, levenshtein, tfidf
spase, nssdc: type, label_match, identifier, date, levenshtein, tfidf, digit
spase, iaumpc: type, label_match, identifier, distance, date, levenshtein, tfidf, digit
pds, wikidata: label_match, distance, date, levenshtein, tfidf, digit
pds, aas: distance, type, date, aperture, label_match, levenshtein, tfidf, digit
imcce, naif[spacecraft]: label_match, date, levenshtein, tfidf, digit
```

F. SSSOM ontology

If the candidate pair is accepted by a label match, it does not need to be reviewed by an LLM. We save a simple Mapping entity in the SSSOM ontology:

```
obsf:b7cee265-8355-4d4c-a05d-dc22cd63592c a sssom:Mapping ;
    obsf:label_match "1"^^xsd:float ;
    sssom:mapping_date "2025-07-23T11:12:18.890248"^^xsd:dateTimeStamp ;
    sssom:mapping_tool "FacilityList/merge.py" ;
    sssom:object_id aas:observatorio-del-teide> ;
    sssom:predicate_id skos:exactMatch ;
    sssom:similarity_measure "label_match" ;
    sssom:similarity_score "1"^^xsd:float ;
    sssom:subject_id pds:observatorio-del-teide> .
```

If the validation is done by an LLM, we save the scores details and the reviewing metadata in the SSSOM Ontology:

```
obsf:10d6f11e-4552-4343-a986-295e206543ed a sssom:Mapping ;
    rdfs:comment "both entities refer to a 1.52m telescope located at la silla
    observatory in chile, operated by the european southern observatory (eso).
    they share identical attributes such as aperture size (1.52m), location
    (south america, chile), and waveband capabilities (infrared, optical).
    the slight differences in namingc onventions (\"european southern
```

```
    observatory 1.52m telescope at la silla observatory\" vs. \"1.52-m
    spectrographic cassegrain/coude reflector\") do not indicate distinct
    entities but rather different descriptive labels for the same telescope.
    therefore, they are the same entity." ;
    obsf:levenshtein_similarity "0.4444444444444444"^^xsd:float ;
    obsf:weighted_sum "0.3277075222694277"^^xsd:float ;
    obsf:tfidf_cosine_similarity "0.2926864456169227"^^xsd:float ;
    sssom:justification semapv:LexicalMatching ;
    sssom:mapping_date "2025-07-23T20:08:51.577381"^^xsd:dateTimeStamp ;
    sssom:mapping_tool "FacilityList/merge.py" ;
    sssom:object_id aas:european-southern-observatory-1.52m-telescope-
                        at-la-silla-observatory> ;
    sssom:predicate_id skos:exactMatch ;
    sssom:reviewer_label "deepseek-v3:671b-q4_K_M" ;
    sssom:similarity_measure "weighted_sum" ;
    sssom:similarity_score "0.3277075222694277"^^xsd:float ;
    sssom:subject_id pds:1.52-m-spectrographic-cassegrain-coude-reflector> .
```

G. Sample of the generated JSON for the name resolver

```
{...
  "3.6-m-equatorial-cassegrain-coude-reflector": [
    "3.6-m equatorial Cassegrain/Coude reflector",
    "urn:nasa:pds:context:telescope:eso.3m6"
  ],
  "isee-magnetometer-nain-station": [
    "ISEE Magnetometer Nain station",
    "spase://IUGONET/Observatory/ISEE/Induction/NAI"
  ],
  "cosmos-1221": [
    "1980-090A",
    "12058",
    "COSMOS 1221"
  ],
  ...}
```