



## GEOSatDB: global civil earth observation satellite semantic database

Ming Lin, Meng Jin, Juanzi Li & Yuqi Bai

To cite this article: Ming Lin, Meng Jin, Juanzi Li & Yuqi Bai (27 Mar 2024): GEOSatDB: global civil earth observation satellite semantic database, Big Earth Data, DOI: [10.1080/20964471.2024.2331992](https://doi.org/10.1080/20964471.2024.2331992)

To link to this article: <https://doi.org/10.1080/20964471.2024.2331992>



© 2024 The Author(s). Published by Taylor & Francis Group and Science Press on behalf of the International Society for Digital Earth, supported by the International Research Center of Big Data for Sustainable Development Goals, and CASEarth Strategic Priority Research Programme.



Published online: 27 Mar 2024.



Submit your article to this journal [↗](#)



View related articles [↗](#)



View Crossmark data [↗](#)

# GEOsSatDB: global civil earth observation satellite semantic database

Ming Lin<sup>a</sup>, Meng Jin<sup>a</sup>, Juanzi Li<sup>b</sup> and Yuqi Bai<sup>a,c</sup>

<sup>a</sup>Department of Earth System Science, Institute for Global Change Studies, Ministry of Education Ecological Field Station for East Asian Migratory Birds, Tsinghua University, Beijing, China; <sup>b</sup>Department of Computer Science and Technology, Tsinghua University, Beijing, China; <sup>c</sup>Tsinghua Urban Institute, Tsinghua University, Beijing, China

## ABSTRACT

Satellite remote sensing, characterized by extensive coverage, frequent revisits, and continuous monitoring, provides essential data support for addressing global challenges. Over the past six decades, thousands of Earth observation satellites and sensors have been deployed worldwide. These valuable Earth observation assets are contributed independently by various nations and organizations employing diverse methodologies. This poses a significant challenge in effectively discovering global Earth observation resources and realizing their full potential. In this paper, we describe the development of GEOsSatDB, the most complete semantic database of civil Earth observation satellites developed based on a unified ontology model. A similarity matching method is used to integrate satellite information and a prompt strategy is used to extract unstructured sensor information. The resulting semantic database contains 127,949 semantic statements for 2,340 remote sensing satellites and 1,021 observation sensors. The global Earth observation capabilities of 195 countries worldwide have been analyzed in detail, and a concrete use case along with an associated query demonstration is presented. This database provides significant value in effectively facilitating the semantic understanding and sharing of Earth observation resources.

## ARTICLE HISTORY

Received 21 November 2023  
Accepted 14 March 2024

## KEYWORDS

Earth observation; satellite; sensor; semantic representation; information extraction

## 1. Introduction

The widespread availability of coordinated and publicly accessible Earth observation (EO) data empowers decision-makers worldwide to comprehend global challenges and develop more effective policies (Annoni et al., 2023; Guo et al., 2022; Sudmanns et al., 2023). Space-based satellite remote sensing, which serves as the primary tool for EO, provides essential information about the Earth and its environment by measuring various

**CONTACT** Juanzi Li  [lijuanzi@tsinghua.edu.cn](mailto:lijuanzi@tsinghua.edu.cn)  Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China; Yuqi Bai  [yuqibai@tsinghua.edu.cn](mailto:yuqibai@tsinghua.edu.cn)  Department of Earth System Science, Tsinghua University, Beijing 100084, China

© 2024 The Author(s). Published by Taylor & Francis Group and Science Press on behalf of the International Society for Digital Earth, supported by the International Research Center of Big Data for Sustainable Development Goals.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

geophysical variables (Wulder et al., 2022; Zhao et al., 2021). This contributes significantly to our understanding of the fundamental Earth system and the impact of human activities.

Over the past few decades, many countries and organizations have markedly improved their regional and global EO capabilities by deploying a variety of advanced remote sensing satellites. The rapid growth of EO satellites and advances in on-board sensors have significantly enhanced remote sensing data quality by expanding spectral bands and increasing spatio-temporal resolutions (Bai & Jin, 2021). However, users face challenges in accessing available EO resources, which are often maintained independently by various nations, organizations, or companies (Boldrini et al., 2023; Roncella et al., 2023). As a result, a substantial portion of archived EO satellite resources remains underutilized (Ballari et al., 2023). Enhancing the discoverability of EO satellites and sensors can effectively utilize the vast amount of EO resources that continue to accumulate at a rapid pace, thereby better supporting data for global change research (Jin et al., 2022; Mazzetti et al., 2022). For example, thermal infrared (TIR) remote sensing, which detects thermal radiation emitted from the Earth's surface, plays a crucial role in monitoring land surface temperature (LST). The most widely used global EO data for this purpose include Landsat satellites (Chen et al., 2022; Gemtzi et al., 2021) and Moderate Resolution Imaging Spectroradiometer (MODIS) (Wang et al., 2022; Zhan & Liang, 2023). To determine which other EO satellites can provide similar TIR observing capabilities with comparable spatial resolution, a dedicated satellite and sensor database is highly anticipated. Such a database could maintain semantic descriptions of observing characteristics and capabilities for all satellites and sensors. It could also provide search capabilities to offer more comprehensive knowledge services.

In this paper, we present GEOSatDB, a semantic database of Earth observation satellites that contains a total of 127,949 semantic statements for 2,340 satellites, 1,021 sensors, and 2,331 wavebands. We have constructed a unified ontology model to formalize the semantic information of the satellites, sensors, and wavebands within this database. Given the significant variability of information across different satellite information databases, we present a fusion method for aggregating EO satellite information from them. A large language model was utilized with a specially designed prompt strategy to fulfill the information extraction from semi-structured web pages. A voting method with support from prior knowledge was used to derive the final information. To demonstrate the value of this database, a specific use case is presented, along with an analysis of Earth observation capabilities across 195 countries.

The remainder of this paper is structured as follows. [Section 2](#) provides an overview of existing EO satellite and sensor databases. [Section 3](#) details a unified ontology model, the methodology for integrating EO satellite information, and the approach for extracting and fusing EO sensor information. [Section 4](#) analyzes the results of the semantic database and the current status of EO capabilities in 195 countries. [Section 5](#) presents a specific query example and discusses the limitations of this study. [Section 6](#) offers a conclusion for the paper.

## 2. Related work

As summarized in [Table 1](#), there are several important information resources for EO satellites and sensors. The World Meteorological Organization (WMO) has developed the Observing Systems Capability Analysis and Review Tool (OSCAR) to assess EO

**Table 1.** Earth observation satellite and sensor databases.

Resource	Count*		Types	Services	License
	Satellite	Sensor			
WMO OSCAR	696	786	EO	A, F, W	OA
CEOS MIM Database	457	559	EO	F, W	OA
NASA GCMD	518	768	EO	A, F, W	OA
ITC Database	345	403	EO	W	OA
ESA eoPortal	657	-	EO	W	NC
UCS Satellite Database	6,718	-	All	F, W	OA
CelesTrak SATCAT	16,404	-	None	F, W	OA
GCAT	17,121	-	None	F	OA
UNOOSA OSOidx	16,568	-	None	W	OA
Nanosats Database	3,752	-	All	W	OA

Key: \*= Only successfully launched satellites and on-board sensors are counted.

- = No sensor information available.

Services: A = API; F = File export; W=Web page.

License: OA = open access with acknowledgement; NC = non-commercial use.

requirements, particularly within WMO application areas such as meteorology, hydrology, and climate studies (Balogh & Kurino, 2020; WMO, 2023). The Committee on Earth Observation Satellites (CEOS) maintains the Missions and Instruments Database to facilitate the exchange of information on civil EO satellites and instruments among CEOS member agencies (CEOS, 2023). Its primary goal is to improve the use of civil EO capabilities and to coordinate upcoming EO missions. The National Aeronautics and Space Administration (NASA) has established the Global Change Master Directory (GCMD) to help researchers, policymakers, and the general public locate and access data and services relevant to global change research (Parsons et al., 2022). The University of Twente's Faculty of Geo-Information Science and Earth Observation (ITC) has designed a satellite and sensor database, providing swift access to pertinent information about EO satellites and sensors (ITC, 2023). The European Space Agency (ESA) eoPortal serves as a reliable and precise gateway for accessing comprehensive information about EO satellite missions presented in web articles (ESA, 2023).

In addition to these specialized EO satellite resources, there are a number of valuable general satellite databases. The Union of Concerned Scientists (UCS) Satellite Database periodically publishes primary information about currently operational satellites in Excel format (UCS, 2023). CelesTrak Satellite Catalog (SATCAT) (Kelso, 2023), Outer Space Objects Index (OSOidx) (UNOOSA, 2023), and the General Catalog of Artificial Space Objects (GCAT) (McDowell, 2023) are three databases that focus on the safety of the Earth's orbital environment and include active payloads, rocket stages, and debris. As such, these databases encompass all successfully launched satellite resources, along with metadata primarily related to orbital parameters. In addition, the Nanosats Database collects detailed information on nanosatellites (Kulu, 2023).

These existing EO satellite and sensor databases have been built for different purposes and therefore exhibit significant variations in content and quantity. In addition, the richness of their metadata and the organization of information show considerable diversity. CelesTrak SATCAT, GCAT, and UNOOSA OSOidx represent general-purpose satellite databases. These databases encompass not only EO satellites but also those designed for other purposes, such as communication and navigation. They maintain information on satellite orbital parameters. However, there is no classification of satellite types in these



databases. The UCS satellite database is limited to satellites that are currently in orbit. The Nanosats database specializes exclusively in nanosatellites. Regarding EO databases, the ESA eoPortal primarily organizes resources in the form of unstructured webpages. NASA GCMD maintains only the names of the satellites as keywords, without additional information. The CEOS MIM database exclusively encompasses the satellites and sensors of its member agencies, featuring wavelength, spatial resolution, and swath width information primarily in the form of unstructured text. The ITC database initially includes models for wavelength, resolution, swath width, and revisit time, but the depiction of microwave information is still inadequate, and the total number of sensors is only 412. The WMO OSCAR database contains nearly 1,000 EO satellites and sensors, representing the most comprehensive database available. However, a majority of the information is maintained in semi-structured tables, which poses challenges for effective information retrieval and mining.

The purpose of this study is to propose a unified ontology model and develop a comprehensive methodology capable of in-depth analysis and integration of these databases and information sources to provide a more complete semantic representation of satellites, sensors and wavebands. The goal of this study is to provide comprehensive, structured semantic information about Earth observation satellites and sensors, and to facilitate advanced information retrieval and knowledge services for discoverability and reusability of EO resources.

### 3. Methodology

This section first presents the design of the ontology, including essential classes and relationships, and then provides a detailed description of the knowledge base creation process for EO satellites and their associated sensors. All information was obtained from open access sources listed in [Table 1](#).

#### 3.1. Ontology design

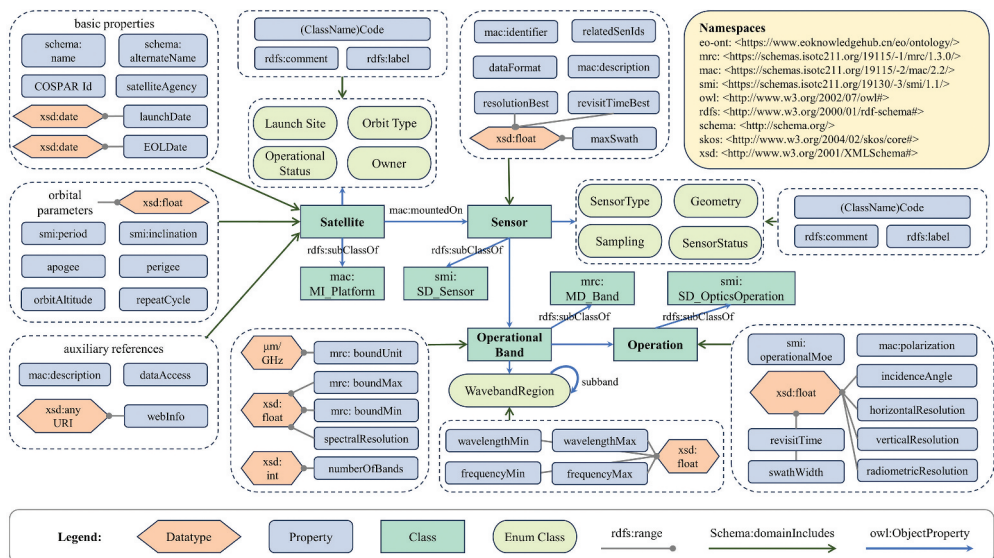
The ontology model establishes an understandable and common vocabulary to improve sharing and interoperability across diverse systems, thereby significantly increasing the reusability of domain knowledge. The primary goal of GEOSatDB is to provide a comprehensive database of Earth observation satellites by integrating different data sources. The GEOSatDB ontology has been developed based on the following principles:

- Incorporating valuable information from multiple data sources.
- Capturing essential observation parameters.
- Providing provenance information for GEOSatDB entities.

The construction of the ontology employs the RDF technology stack (Gandon et al., 2014; Prud'hommeaux et al., 2014) developed and maintained by the World Wide Web Consortium (W3C). The RDF data model is characterized by a variety of syntax notations and primarily consists of nodes and edges. Non-literal nodes are uniquely identified by Uniform Resource Identifiers (URIs), thereby enhancing resource association and integration. The methodology employed in building the GEOSatDB ontology integrates both

top-down and bottom-up approaches. Initially, the ontology is aligned with key international standards for Earth observation satellites and sensors, particularly those established by the International Organization for Standardization (ISO) Technical Committee 211 (ISO/TC211), a committee focused on geographic information. Adaptation and extension of the core concepts and metadata from the ISO 19115 (ISO, 2014b, 2019) and 19130 (ISO, 2014a, 2018, 2022) series were undertaken to ensure semantic interoperability. Subsequently, analysis and generalization of the data structures of existing resources were conducted, and application-specific properties were introduced within the proposed namespace. Enumerable properties are represented as classes to facilitate association and extension. To further advance semantic interoperability, generic metadata standards have been adopted. These include utilizing OWL (Hitzler et al., 2012) and RDFS (Brickley & Guha, 2014) for delineating relationships between classes and properties, employing SKOS (Miles & Bechhofer, 2009) for terminology definitions, applying XML Schema (Biron & Malhotra, 2004) for data type specification, and incorporating the publicly accessible vocabulary from Schema.org (Guha et al., 2016).

The resulting ontology is illustrated in Figure 1. There are four fundamental classes: Satellite, Sensor, Operational Band, and Operation. A remote sensing sensor, mounted on a satellite, collects data representing the state of the Earth's surface and atmosphere. An operational band represents the range of electromagnetic frequencies across which a sensor conducts observations. The modeling of the satellite class encompasses three principal facets: basic information, orbital information, and auxiliary information. Basic information refers to descriptive properties of a satellite, including its name, international designator, launch date, and end of life (EOL) date. Orbital information is critical in assessing a satellite's capability to perform an EO mission by specifying its orbital parameters. Auxiliary information provides references for various purposes, including



**Figure 1.** Ontology model of GEOSatDB. If no datatype is specified, properties default to xsd:string. Object properties are omitted when they start with a lowercase letter and match the class name. For example, in the tuple <OperationalBand, operation, Operation>, the term “operation” is excluded.

descriptions and data access. The sensor class is characterized by an emphasis on the observation technique and the attainment of optimal performance characteristics (e.g. highest resolution, maximum swath). The operational band class delineates the spectrum of bands available for a sensor. The operational mode and observing performance associated with an operational band are further delineated within the operation class.

### 3.2. Integrating EO satellite information

The first step is to select the objects classified as “payload” from the CSV file provided by the CelesTrak Satellite Catalog, which serves as the primary satellite database. This catalog primarily contains the satellite name, a unique international designator (also known as the COSPAR ID), and orbital parameter information. The CSV files from GCAT and UCS, along with information retrieved from the OSOidx and Nanosats databases, linked through the international designator, are utilized to populate the missing fields in our satellite database, encompassing satellite alternate name, orbit type, dry mass, end of life date, and applications.

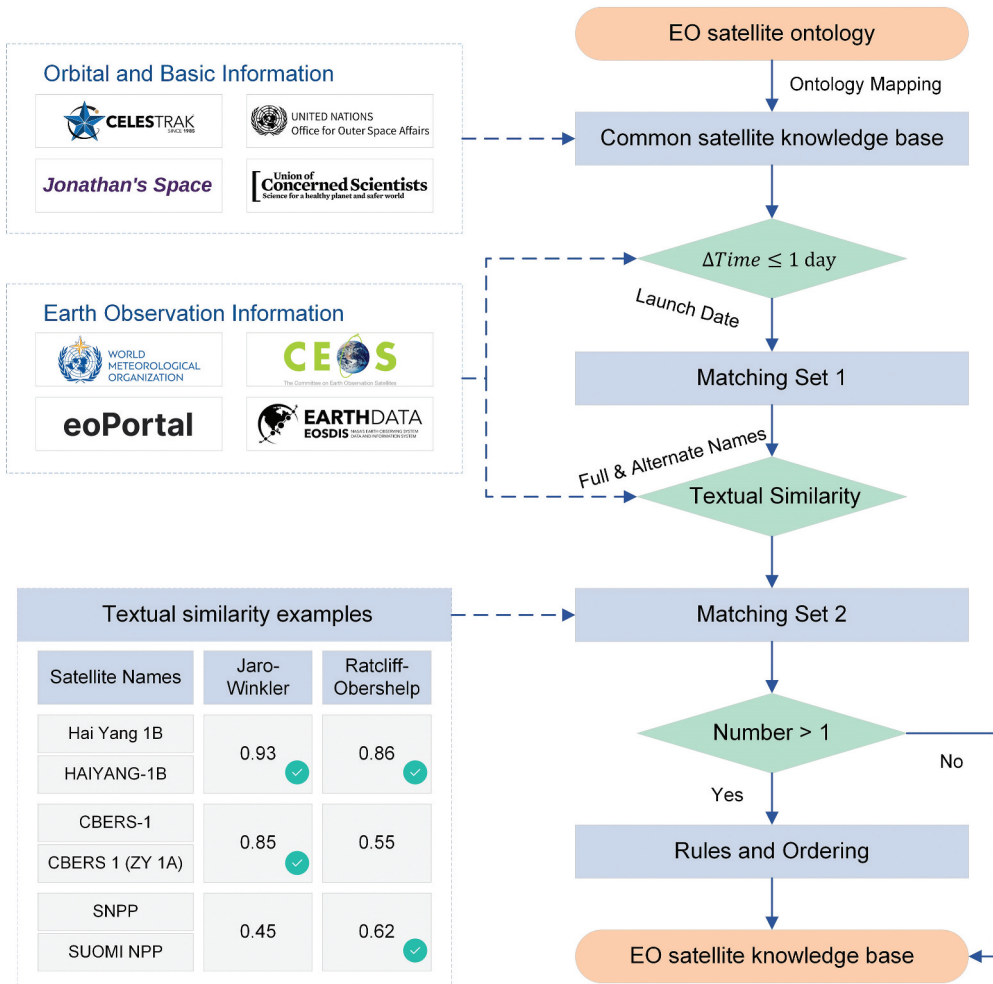
Although the OSCAR satellite database is a valuable resource for environmental satellite missions, it lacks the international designator required for direct mapping to our database. To overcome this limitation, we use a matching method that combines launch time and text similarity to establish the association, as shown in Figure 2. The OSCAR satellite database provides access to data from both the Excel file and the application programming interface (API) endpoint, but the information is not completely overlapping. We use the API endpoint data as the primary source and supplement it with the additional data provided in the Excel file. Satellite launch times in the OSCAR database are frequently provided independently by various organizations, resulting in the use of local time zones. To illustrate, consider the case of Feng-Yun 1A. OSCAR records its launch time as September 7, while GEOSatDB uses a UTC-based launch time of September 6. This discrepancy is due to the fact that Feng-Yun 1A was launched at 4 a.m. on September 7 in the UTC + 8 time zone. To account for such differences, we set a one-day launch time threshold.

We then use two similarity matching methods, Jaro-Winkler similarity (Cohen et al., 2003) and Ratcliff-Obershelp similarity (Kalbaliyev & Rustamov, 2021), to align OSCAR to our database using the satellite names and aliases. The Jaro-Winkler similarity measures the edit distance between two strings, with special emphasis on improving the similarity for strings with identical prefixes compared to the standard Jaro similarity. The formula for the Jaro similarity is

$$sim_j = \begin{cases} 0, & \text{if } m = 0 \\ \frac{1}{3} \left( \frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m} \right), & \text{if } m! = 0 \end{cases} \quad (1)$$

Where  $m$  is the number of character matches;  $t$  is half the number of characters that match but are out of order;  $|s_1|$  and  $|s_2|$  are the lengths of the two strings being compared. Thus, the Jaro-Winkler similarity is

$$sim_w = sim_j + 0.1 \times len \times (1 - sim_j) \quad (2)$$



**Figure 2.** Integration of Earth observation satellite information from multiple databases.

Where  $len$  is the number of characters that start the same in both strings (up to 4). The threshold for the Jaro-Winkler similarity is set to 0.6. If there are no matching satellites, the Ratcliff-Obershelp similarity is used to measure the sequence similarity between the two strings.

$$sim_{ro} = \frac{2 \times k_m}{|s_1| + |s_2|} \quad (3)$$

Where  $k_m$  represents the count of matching characters. A threshold of 0.6 is applied to the Ratcliff-Obershelp similarity. Comparable matching strategies are utilized for the CEOS, ITC, GCMD, and eoPortal satellite databases, all of which lack the international designator. These databases serve primarily to supplement EO-related information, such as repeat cycle, equatorial crossing time, and onboard observation sensors.

### 3.3. Extraction and fusion of EO sensor information

Unfortunately, there is a lack of databases that allow advanced retrieval of sensor information. For example, the OSCAR sensor database provides simple filtering options based only on sensor type and spectral domain. This limitation is due to the fact that much of the valuable information is presented in an unstructured text format. Recently, there has been a growing interest in unsupervised information extraction using large language models (LLMs). For example, GPT-3 has demonstrated potential as an effective extractor of clinical information through zero and few-shot prompts (Agrawal et al., 2022). Employing few-shot prompting with GPT-3 achieves performance nearing the state-of-the-art (SOTA) in relation extraction tasks (Wadhwa et al., 2023). ChatIE (Wei et al., 2023), a two-step framework relying on LLMs, achieves remarkable performance on zero-shot information extraction tasks. In this context, we propose the utilization of a structured prompt strategy to guide large language models in extracting information including resolution, revisit time, swath width, spectral range, polarization, incidence angle. We use the GPT-4 language model for this purpose, details of which are shown in Figure 3. Initially, we retrieve the unprocessed content of each sensor in OSCAR using WMO APIs and web crawlers. We then perform the following preprocessing steps:

- Numerical information, such as signals and wave numbers, is filtered based on predefined rules. Complex unit representations pose a challenge to the language model, resulting in potential extraction inaccuracies that can affect the accuracy of other numerical entities.
- Raw content is consolidated into key-value pairs for use as model input. Unnecessary line breaks and indentation are removed. This reduction significantly shortens the model input, allowing for more input text and cost savings.

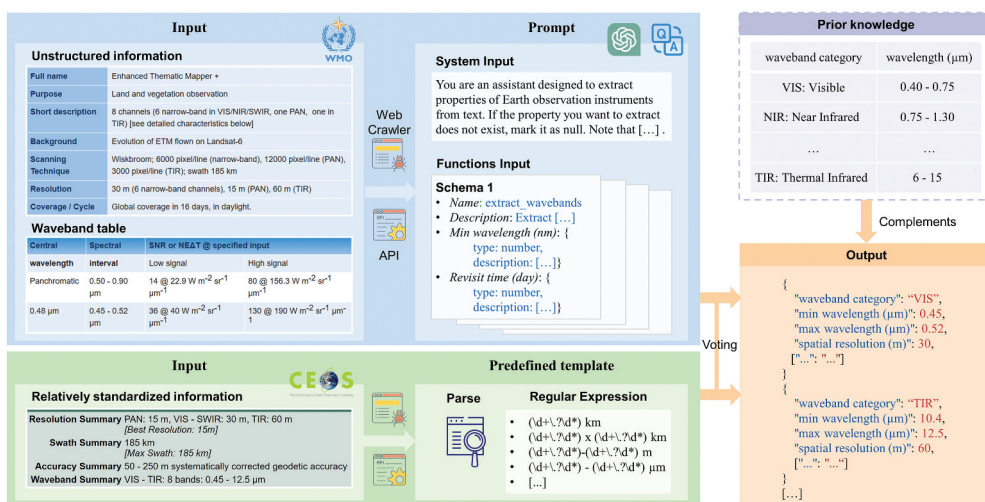


Figure 3. Extracting EO sensor information from web pages using large language models and regular expressions.

The prompt string consists of two components: a system message and a function call. The system message provides instructions for guiding the model according to the following principles:

- Explicitly state that this task pertains to information extraction.
- Exercise caution by not extracting uncertain data and marking it as null.
- Pay special attention to unit differences and conversions.

The function call is used to define the JSON object for output from the model, following the JSON schema and specifying the type and description of each entity, including the unit for numeric types.

In scenarios involving resources such as the CEOS sensor database, the data, although in unstructured text form, adheres to a more consistent and specific format. The Excel files were obtained from the CEOS sensor database and the web pages for each sensor were downloaded, each containing a significant amount of valuable unstructured information. Information was extracted using regular expression templates that primarily included numeric fields such as spatial resolution and swath width. The resulting content from various sources was then merged through a voting process to produce a singular outcome. Initially, all extracted results were aggregated and the entry with the highest number of votes was selected as the output. In the absence of a clear majority, results derived from regular extraction methods were prioritized. Prior knowledge, including wavelength ranges corresponding to waveband categories, was employed for refinement. Finally, the automatically extracted information is progressively corrected and supplemented by domain experts and users to achieve optimal results.

## 4. Results

### 4.1. Overview of GEOSatDB

GEOSatDB employs a unified ontology model consisting of 4 core classes, 9 enumeration classes and 61 associated properties. In total, it contains 127,949 semantic statements. In particular, it completes the description of 2,340 satellites, 1,021 sensors and 2,331 wavebands. [Figure 4](#) illustrates the comparison between GEOSatDB and four prominent existing EO satellite databases. Evidently, GEOSatDB exhibits a distinct advantage in both the quantity of EO satellites and the richness of information. Both the OSCAR and CEOS databases fail to match GEOSatDB in about 6% of the satellite count, considering a combination of launch time and text similarity. An analysis of these shortcomings shows that they are mainly due to OSCAR and CEOS misclassifying failed launches as inactive or inaccurately recording the launch time. The GCMD and ITC databases have a lower correlation with GEOSatDB, mainly because the launch times they provide are only precise at the year level.

### 4.2. EO capability analysis

The trend in the number of EO satellites launched over the years is depicted in [Figure 5\(a\)](#). Over the past two decades, a significant increase in the number of EO satellites has been



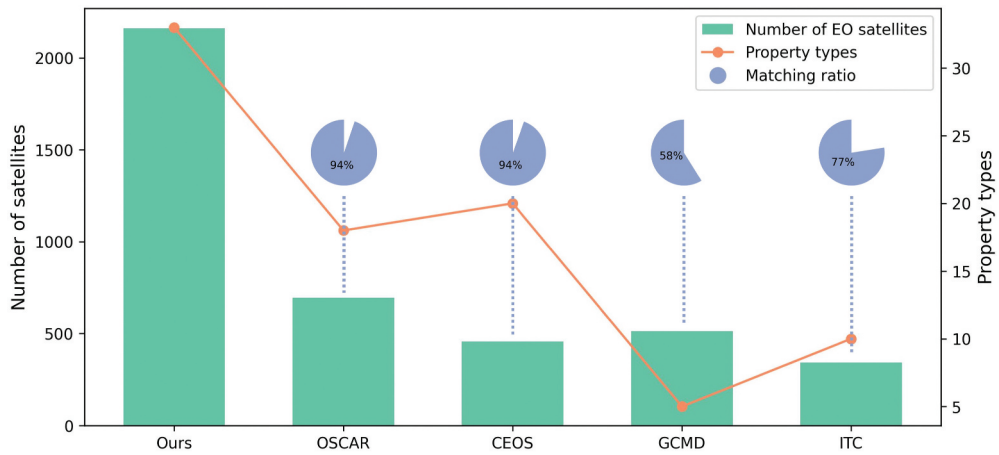


Figure 4. Comparative analysis of our EO satellite database and existing resources.

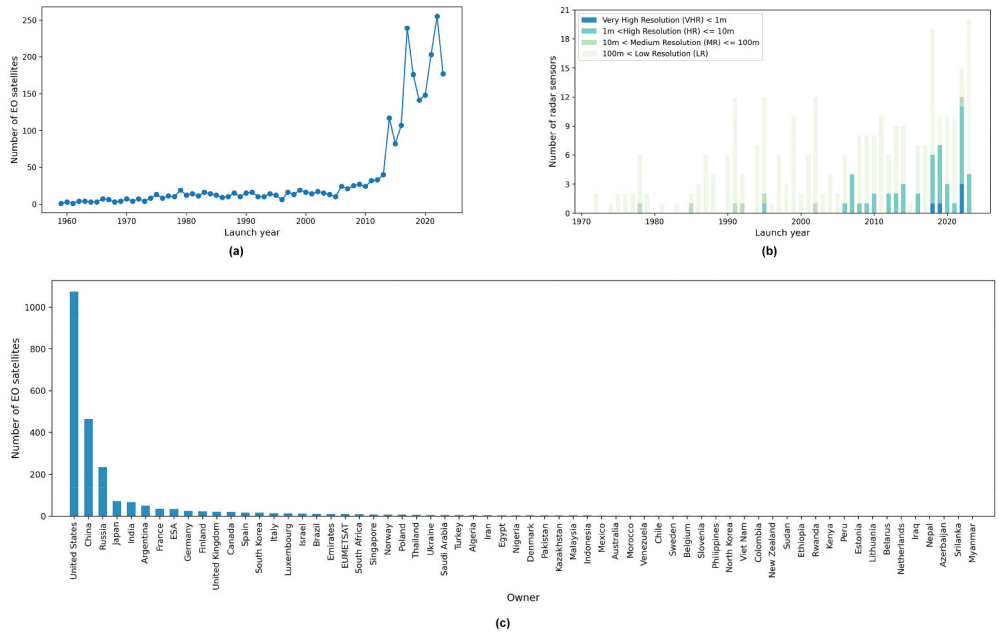


Figure 5. (a) Number of EO satellites launched annually. (b) Number of radar sensors of different resolutions launched annually. (c) Number of EO satellites operated by different countries or organizations. Data from international organizations such as ESA are listed only under themselves and are not duplicated in each member country.

observed, driven by continuous technological advances. Especially since 2014, numerous low-cost small satellite constellations have emerged, as exemplified by the Planet’s Dove satellites. Additionally, Figure 5(b) demonstrates the evolution of the spatial resolution of civil radar sensors. The early low-resolution SeaSat and Nimbus series of satellites were primarily deployed to verify the feasibility of meteorological and oceanographic



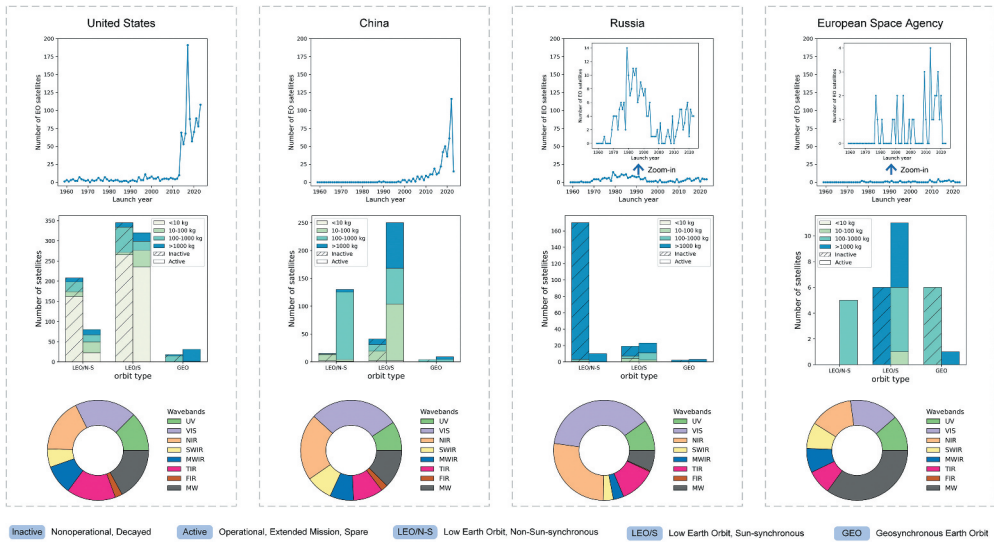
observations from space. Subsequently, satellites such as the European Remote Sensing Satellite (ERS), equipped with a comprehensive suite of radar sensors, have provided continuous and stable observations. Over the past decade, significant advances in radar technology have enabled improved image resolution and diversified modes of operation. The C-band synthetic aperture radar (SAR) onboard Sentinel-1A, which was launched in 2014, has facilitated high-resolution measurements in certain modes. In recent years, commercial satellites such as COSMO-SkyMed Second Generation (CSG) and Capella Space have transitioned to sub-meter SAR observations. Looking forward, the focus on higher resolution will also underscore the harmonized development of data quality.

Figure 5(c) shows the number of EO satellites operated by different countries and organizations. It is evident that many countries in the southern hemisphere tend to have limited satellite Earth observation capabilities. Furthermore, only a few African countries possess the capacity to launch and sustain their own EO satellites. This implies that these nations are predominantly dependent on global EO satellites operated by other countries for observing their national space. This study recommends that spacefaring nations should strengthen their collaboration with the Group on Earth Observations (GEO), a leading intergovernmental organization dedicated to Earth observation. GEO envisions a future in which decisions and actions for the benefit of humanity are informed by well-coordinated, in-depth, and continuous Earth observation. In their engagement with GEO, spacefaring nations ought to adopt a more open and inclusive approach in sharing their global satellite observation data and products. Furthermore, it is advised that they provide technical support for the development and maintenance of EO satellites to developing countries, particularly focusing on African nations, guided by the principles of cooperation and mutual benefit.

We conducted a detailed analysis of the three countries with the highest number of EO satellites – the United States, China, and Russia – along with ESA, a prominent international organization in space exploration, as depicted in Figure 6. The United States holds a significant lead with a total of 1,037 EO satellites, characterized by a balanced ratio of retired to in-orbit satellites and an equitable distribution across various spectral bands. This capability enables the United States to maintain continuous observation of specific targets over extended periods. In contrast, China's EO satellites consist primarily of newly launched operational satellites in recent years, suggesting significant growth potential in the coordinated development of small EO satellites. Russia has launched fewer new EO satellites, and the majority of these have been decommissioned and no longer participate in observation activities. ESA has consistently launched and maintained high quality satellites throughout its 50-year history, providing stable and enduring global services.

### **4.3. An information retrieval and mining use case**

To demonstrate the information mining potential of GEOSatDB, we employed the LST monitoring scenario outlined in the introduction as a concrete example. Figure 7 illustrates satellite resources that provide enhanced thermal infrared resolution compared to the broadly utilized Thermal Infrared Sensor (TIRS) on board Landsat-8 and Landsat-9. On the China-Brazil Earth Resources Satellite-4 (CBERS-4), the TIR sensor possesses a single TIR band with 80 m spatial resolution. The TIR camera on the HJ-2A/B satellite features nearly the same band



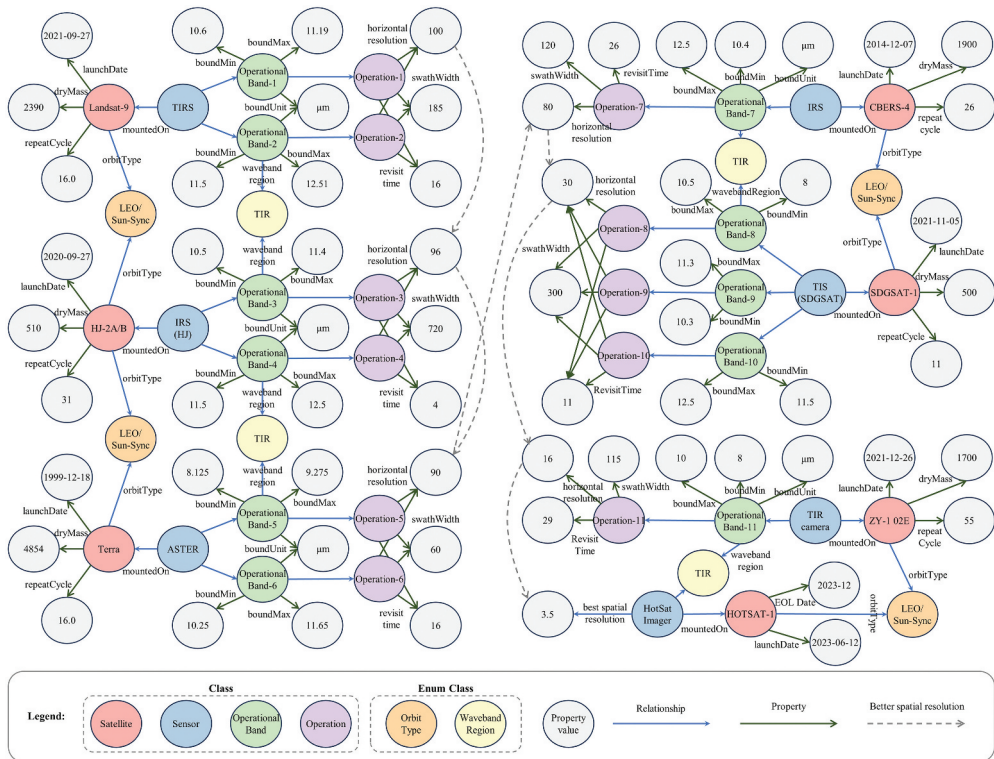
**Figure 6.** Analysis of the EO capabilities of the United States, China, Russia, and ESA. Line charts show the number of EO satellites launched over the years. Bar charts illustrate the operational status and mass of satellites in various orbit types. Pie charts demonstrate the distribution of wavebands among the sensors aboard these satellites.

segmentation and spatial resolution as the Landsat TIRS, coupled with a larger swath width and faster revisit time. The SDGSAT-1 satellite is equipped with a TIR sensor featuring a novel 3-band design that combines a large swath width of 300 km with a high spatial resolution of 30 m. The TIR band of the ZY-1-02E satellite operates within a special wide channel of 8–10  $\mu\text{m}$ , establishing it as a leader in the field with a high spatial resolution of 16 m. In addition, the recent deployment of the HOTSAT-1 commercial satellite has provided a highly competitive TIR resolution of 3.5 m. However, an anomaly that occurred in December 2023 resulted in the cessation of data production.

## 5. Further discussions

### 5.1. Query example


GEOSatDB is published and distributed in Turtle format and supports the SPARQL query language. As a W3C standard, SPARQL ensures consistency and interoperability in querying RDF graphs. This standardization is crucial for developers and organizations engaged in working with web data. Figure 8 demonstrates the SPARQL query statement for the specific case depicted in Figure 7, comprising a main query and a subquery. The subquery is initially utilized to obtain the thermal infrared spatial resolution of Landsat-9 TIRS. Subsequently, the main query, based on a filter expression, aims to identify all sensors with a thermal infrared spatial resolution higher than that of Landsat-9 TIRS. In addition, GEOSatDB has been integrated into our Earth Observation Knowledge Hub, enhancing service accessibility for users unfamiliar with SPARQL, as illustrated on the right side of Figure 8.




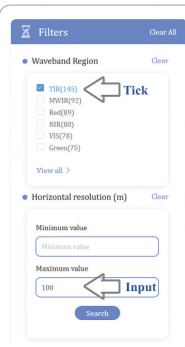
**Figure 7.** Satellite remote sensing resources that provide better spatial resolution than Landsat-8 in the thermal infrared spectrum.


```

1 PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
2 PREFIX eo-ont: <https://www.eoknowledgehub.cn/eo/ontology/>
3 PREFIX schema: <http://schema.org/>
4 PREFIX mac: <https://schemas.iscte211.org/19115/-2/mac/2.2/>
5 PREFIX eor: <https://www.eoknowledgehub.cn/eo/resource/>
6 SELECT DISTINCT ?satellite ?sensor ?band_tir_res WHERE {
7   # Spatial resolution of Landsat-9 TIRS
8   {
9     SELECT ?tirs_band_res WHERE {
10      ?Landsat_9 rdf:type eo-ont:Satellite;
11      schema:name ?name;
12      schema:alternateName ?alternateName.
13      FILTER((CASE(?name)) = "Landsat-9"@en ||
14             ((CASE(?alternateName)) = "Landsat-9"@en))
15      ?tirs_mac:mountedOn ?Landsat_9;
16      eo-ont:operationalBand ?tirs_band.
17      ?tirs_band eo-ont:wavebandRegion eor:waveband_region.TIR ;
18      eo-ont:operation ?tirs_band_op.
19      ?tirs_band_op eo-ont:horizontalResolution ?tirs_band_res.
20    }
21   }
22   # Spatial resolution for all thermal infrared
23   ?satellite rdf:type eo-ont:Satellite;
24   schema:name ?satellite_name.
25   ?sensor mac:mountedOn ?satellite;
26   schema:name ?sensor_name;
27   eo-ont:operationalBand ?band.
28   ?band eo-ont:wavebandRegion eor:waveband_region.TIR ;
29   eo-ont:operation ?band_op.
30   ?band_op eo-ont:horizontalResolution ?band_tir_res.
31 }
32 FILTER(?band_tir_res <= ?tirs_band_res)
33 }
                
```









**Figure 8.** SPARQL query and online search to identify satellites and sensors with higher spatial resolution than Landsat-9 in the thermal infrared band.

## 5.2. Limitations

The GEOSatDB database is currently limited in the inclusion of EO sensors and lacks satellite constellation information. The construction of the ontology integrates a top-down approach, adhering to international standards, and a bottom-up approach, informed by empirical data sources. As a result, certain properties currently have no or minimal values, notably repeat cycle, data access, data format, and radiometric resolution. Nevertheless, these properties are critical for characterizing EO capabilities and data use, so they are included in the ontology to enable future extensions. In future iterations, we plan to augment the database by automatically extracting sensor and constellation details from official satellite websites. Furthermore, building on this initial release, there is potential to fill data gaps through collaborations with GEO and satellite operators. Regarding unsupervised information extraction, the generalized GPT model has shown impressive precision but suffers from a lack of recall. The development of a domain-specific model may prove to be an effective way to improve extraction performance.

## 6. Conclusions

This paper presents GEOSatDB, a specialized semantic database that provides extensive and semantically enriched information on both active and retired EO satellites and their onboard sensors. Its primary goal is to improve the discoverability of EO resources, thereby assisting researchers in accessing newly available EO data and satellite operators in assessing current EO capabilities for more effective coordination of future EO missions. In summary, the main contributions of this paper are as follows:

- GEOSatDB serves as the most extensive knowledge base for Earth observation satellites and sensors, covering 4 core classes, 9 enumeration classes, 61 properties, 2,340 satellites, 1,021 sensors, and a total of 127,949 semantic statements. It has been developed by integrating data from diverse sources and employs semantic representation.
- We propose a matching method that combines launch time and text similarity to address the challenge of linking and integrating diverse satellite databases. Additionally, we propose the use of a structured prompt strategy to guide LLMs in extracting sensor information from unstructured text.
- Our research reveals a significant North-South divide in satellite Earth observation capabilities, with the majority of African nations unable to launch or maintain their EO satellites. We advocate for spacefaring nations to strengthen their collaboration with GEO, both to expand and share global EO resources and to support developing countries in building their own EO satellites through technical assistance.

Our future research goals include the following: (1) Develop a specialized, large EO-specific language model to enhance the extraction of EO entities and relationships. (2) Augment GEOSatDB by retrieving information from EO satellite and sensor detail pages, including platforms such as eoPortal and official websites. (3) Create a comprehensive and interconnected EO knowledge graph, encompassing Earth science variables, EO satellites, scientific datasets, scholarly literature, and other pertinent components.

## Acknowledgements

We gratefully acknowledge the publicly available satellite and sensor resources, including WMO OSCAR, CEOS MIM Database, NASA GCMD, ITC Database, ESA eoPortal, UCS Satellite Database, CelesTrak SATCAT, GCAT, Nanosats Database, and UNOOSA OSOidx. We are also grateful to the editors and three anonymous reviewers whose constructive comments significantly improved the quality of the manuscript.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Major Program of the National Natural Science Foundation of China [42090015].

## Notes on contributors



**Ming Lin** received his bachelor's degree from the College of Surveying and Geo-Informatics at Tongji University in 2020. He is currently a Ph.D. candidate at the Department of Earth System Science, Tsinghua University. His current research focuses on knowledge service pathways linking satellite remote sensing observations to sustainable development goals.



**Meng Jin** received her Ph.D. degree from the Department of Earth System Science at Tsinghua University in 2023. Her current research focuses on spatio-temporal big data analysis.



**Juanzi Li** received her Ph.D. degree from the Department of Computer Science and Technology at Tsinghua University in 2000, and has been working here since she finished her research work as a post-doctor in the Department of Electronic Engineering at Tsinghua University in 2001. She is currently a professor in the Department of Computer Science and Technology at Tsinghua University. Her research areas are semantic web and semantic web services, and text and social network mining. Currently, her research focuses on studying key technologies in semantic content management, and applying them in the domains of news, social networks, and web services.



**Yuqi Bai** received his Ph.D. degree in cartography and GIS from the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing, in 2003. He is a professor of Earth and space information science in the Department of Earth System Science at Tsinghua University. His research interests include cyber-infrastructure studies and applications in environmental health assessments. He serves as the Co-chair of the Group on Earth Observations (GEO)'s GEOSS Infrastructure Development Task Team (GIDTT), a member of the GEO Executive Committee (ExCom), a member of the GEO Programme Board (PB), and the Convenor of ISO/TC 211/ WG7 Information Community and AHG 11 Climate Change.

## Data availability statement

GEOSatDB was developed using widely adopted W3C technologies and standards, facilitating ease of extension and reuse. It is distributed under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license, permitting free use for non-commercial purposes with proper acknowledgement. The data are available for download in the standard RDF Turtle format at <https://doi.org/10.57760/sciencedb.11805>. GEOSatDB undergoes quarterly updates, involving the addition of new satellites and sensors, revisions based on expert feedback, and the implementation of additional enhancements. In addition, a user-friendly portal (<http://www.geosatdb.cn>) is being developed to ensure easy access to GEOSatDB.

## References

- Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022, December). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, ACL (pp. 1998–2022). <https://doi.org/10.18653/v1/2022.emnlp-main.130>
- Annoni, A., Nativi, S., Çöltekin, A., Desha, C., Eremchenko, E., Gevaert, C. M., Giuliani, G., Chen, M., Perez-Mora, L., Strobl, J., & Tumamos, S. (2023). Digital Earth: Yesterday, today, and tomorrow. *International Journal of Digital Earth*, 16(1), 1022–1072. <https://doi.org/10.1080/17538947.2023.2187467>
- Bai, Y., & Jin, M. (2021). Constant-level spatio-temporal integrated search algorithm for repeating sun-synchronous orbit satellite images. *International Journal of Digital Earth*, 14(8), 943–958. <https://doi.org/10.1080/17538947.2021.1907463>
- Ballari, D., Vilches-Blázquez, L. M., Orellana-Samaniego, M. L., Salgado-Castillo, F., Ochoa-Sánchez, A. E., Graw, V., Turini, N., & Bendix, J. (2023). Satellite earth observation for essential climate variables supporting sustainable development goals: A review on applications. *Remote Sensing*, 15(11), 2716. <https://doi.org/10.3390/rs15112716>
- Balogh, W., & Kurino, T. (2020). The World Meteorological Organization and space-based observations for weather, climate, water and related environmental services. In S. Ferretti (Ed.), *Space capacity building in the XXI century* (pp. 223–232). Springer International Publishing. [https://doi.org/10.1007/978-3-030-21938-3\\_20](https://doi.org/10.1007/978-3-030-21938-3_20)
- Biron, P. V., & Malhotra, A. (2004). *XML Schema Part 2: Datatypes Second Edition* (W3C Recommendation). <https://www.w3.org/TR/xmlschema-2/>
- Boldrini, E., Nativi, S., Hradec, J., Santoro, M., Mazzetti, P., & Craglia, M. (2023). GEOSS Platform data content and use. *International Journal of Digital Earth*, 16(1), 715–740. <https://doi.org/10.1080/17538947.2023.2174193>
- Brickley, D., & Guha, R. V. (2014). *RDF schema 1.1*. W3C Recommendation. <https://www.w3.org/TR/rdf-schema/>
- CEOS. (2023). *CEOS Missions, Instruments, Measurements and Datasets Database*. Retrieved September 1, 2023, From <https://database.eohandbook.com/index.aspx>



- Chen, X., Peng, J., Song, Z., Zheng, Y., & Zhang, B. (2022). Monitoring persistent coal fire using Landsat time series data from 1986 to 2020. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–16. <https://doi.org/10.1109/TGRS.2022.3142350>
- Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). A comparison of string distance metrics for name-matching tasks. In *Proceedings of the 2003 International Conference on Information Integration on the Web*, Acapulco, Mexico. <https://www.cs.cmu.edu/~wcohen/postscript/ijcai-ws-2003.pdf>
- ESA. (2023). *eoPortal*. Retrieved September 1, 2023, From <https://www.eoportal.org/>
- Gandon, F., INRIA, Schreiber, G., & Amsterdam, V. U. (2014). *RDF 1.1 XML syntax*. W3C Recommendation. <https://www.w3.org/TR/rdf-syntax-grammar/>
- Gemitzi, A., Dalampakis, P., & Falalakis, G. (2021). Detecting geothermal anomalies using Landsat 8 thermal infrared remotely sensed data. *International Journal of Applied Earth Observation and Geoinformation*, 96, 102283. <https://doi.org/10.1016/j.jag.2020.102283>
- Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.Org: Evolution of structured data on the web. *Communications of the ACM*, 59(2), 44–51. <https://doi.org/10.1145/2844544>
- Guo, H., Liang, D., Sun, Z., Chen, F., Wang, X., Li, J., Zhu, L., Bian, J., Wei, Y., Huang, L., Chen, Y., Peng, D., Li, X., Lu, S., Liu, J., & Shirazi, Z. (2022). Measuring and evaluating SDG indicators with big earth data. *Science Bulletin*, 67(17), 1792–1801. <https://doi.org/10.1016/j.scib.2022.07.015>
- Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P. F., & Rudolph, S. (2012). *OWL 2 web ontology language: Primer* (2nd ed.). W3C Recommendation. <https://www.w3.org/TR/2012/REC-owl2-primer-20121211/>
- ISO. (2014a). *Geographic Information — Imagery Sensor Models for Geopositioning — Part 2: SAR, InSAR, Lidar and Sonar (ISO/TS 19130-2: 2014)*. <https://www.iso.org/standard/56113.html>
- ISO. (2014b). *Geographic Information — Metadata — Part 1: Fundamentals (ISO 19115-1:2014)*. <https://www.iso.org/standard/53798.html>
- ISO. (2018). *Geographic Information — Imagery Sensor Models for Geopositioning — Part 1: Fundamentals (ISO 19130-1)*. <https://www.iso.org/standard/66847.html>
- ISO. (2019). *Geographic Information — Metadata — Part 2: Extensions for Acquisition and Processing (ISO 19115-2:2019)*. <https://www.iso.org/standard/67039.html>
- ISO. (2022). *Geographic Information — Imagery Sensor Models for Geopositioning — Part 3: Implementation Schema (ISO/TS 19130-3:2022)*. <https://www.iso.org/standard/74074.html>
- ITC. (2023). *ITC Satellite and Sensor Database*. Retrieved 2023-09-01, From <https://www.itc.nl/research/research-facilities/labs-resources/satellite-sensor-database/>
- Jin, M., Lin, M., Liu, Y., & Bai, Y. (2022). An earth observation potential evaluation model and its application to SDG indicators. *International Journal of Digital Earth*, 15(1), 1187–1203. <https://doi.org/10.1080/17538947.2022.2095447>
- Kalbaliyev, E., & Rustamov, S. (2021, June). Text similarity detection using machine learning algorithms with character-based similarity measures. In *Digital Interaction and Machine Intelligence*. [https://doi.org/10.1007/978-3-030-74728-2\\_2](https://doi.org/10.1007/978-3-030-74728-2_2)
- Kelso, T. S. (2023). *CelesTrak Online Satellite Catalog*. Retrieved September 1, 2023, From <https://celestrak.org/>
- Kulu, E. (2023). *Nanosats Database*. Retrieved September 1, 2023, From <https://www.nanosats.eu/>
- Mazzetti, P., Nativi, S., Santoro, M., Giuliani, G., Rodila, D., Folino, A., Caruso, S., Aracri, G., & Lehmann, A. (2022). Knowledge formalization for earth science informed decision-making: The GEOEssential knowledge base. *Environmental Science & Policy*, 131, 93–104. <https://doi.org/10.1016/j.envsci.2021.12.023>
- McDowell, J. (2023). *General Catalog of Artificial Space Objects*. Retrieved September 1, 2023, From <https://www.planet4589.org/index.html>
- Miles, A., & Bechhofer, S. (2009). *SKOS simple knowledge organization system*. W3C Recommendation. <https://www.w3.org/2004/02/skos/>
- Parsons, M. A., Duerr, R., & Godøy, Ø. (2022). The evolution of a geoscience standard: An instructive tale of science keyword development and adoption. *Geoscience Frontiers*, 14(5), 101400. <https://doi.org/10.1016/j.gsf.2022.101400>



- Prud'hommeaux, E., Carothers, G., & Machina, L. (2014). *RDF 1.1 Turtle: Terse RDF triple language*. W3C Recommendation. <https://www.w3.org/TR/2014/REC-turtle-20140225/>
- Roncella, R., Zhang, L., Boldrini, E., Santoro, M., Mazzetti, P., & Nativi, S. (2023). Publishing China satellite data on the GEOSS Platform. *Big Earth Data*, 7(2), 398–412. <https://doi.org/10.1080/20964471.2022.2107420>
- Sudmanns, M., Augustin, H., Killough, B., Giuliani, G., Tiede, D., Leith, A., Yuan, F., & Lewis, A. (2023). Think global, cube local: An earth observation data Cube's contribution to the digital earth vision. *Big Earth Data*, 7(3), 831–859. <https://doi.org/10.1080/20964471.2022.2099236>
- UCS. (2023, January 01). *UCS Satellite Database*. Retrieved September 1, 2023, From <https://www.ucsusa.org/resources/satellite-database>
- UNOOSA. (2023). *Outer Space Objects Index*. Retrieved September 1, 2023, From [https://www.unoosa.org/oosa/osoindex/index.jsp?lf\\_id=](https://www.unoosa.org/oosa/osoindex/index.jsp?lf_id=)
- Wadhwa, S., Amir, S., & Wallace, B. (2023, July). Revisiting Relation Extraction in the era of Large Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, Toronto, Canada, ACL (pp. 15566–15589). <https://doi.org/10.18653/v1/2023.acl-long.868>
- Wang, Y.-R., Hessen, D. O., Samset, B. H., & Stordal, F. (2022). Evaluating global and regional land warming trends in the past decades with both MODIS and ERA5-land land surface temperature data. *Remote Sensing of Environment*, 280, 113181. <https://doi.org/10.1016/j.rse.2022.113181>
- Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., Jiang, Y., & Han, W. (2023). Zero-Shot Information Extraction via Chatting with ChatGPT. *arXiv, abs/2302.10205*. <https://doi.org/10.48550/arXiv.2302.10205>
- WMO. (2023). *Observing Systems Capability Analysis and Review Tool*. Retrieved September 1, 2023, From <https://space.oscar.wmo.int/spacecapabilities>
- Wulder, M. A., Roy, D. P., Radeloff, V. C., Loveland, T. R., Anderson, M. C., Johnson, D. M., Healey, S., Zhu, Z., Scambos, T. A., Pahlevan, N., Hansen, M., Gorelick, N., Crawford, C. J., Masek, J. G., Hermosilla, T., White, J. C., Belward, A. S., Schaaf, C., Woodcock, C. E. . . . Cook, B. D. (2022). Fifty years of Landsat science and impacts. *Remote Sensing of Environment*, 280, 113195. <https://doi.org/10.1016/j.rse.2022.113195>
- Zhan, C., & Liang, S. (2023). Generation of global 1-km daily top-of-atmosphere outgoing longwave radiation product from 2000 to 2021 using machine learning. *International Journal of Digital Earth*, 16(1), 2002–2012. <https://doi.org/10.1080/17538947.2023.2220611>
- Zhao, T., Cosh, M. H., Roy, A., Mu, X., Qiu, Y., & Shi, J. (2021). Remote sensing experiments for earth system science. *International Journal of Digital Earth*, 14(10), 1237–1242. <https://doi.org/10.1080/17538947.2021.1977473>