STAC, an open standard to describe and catalog geospatial data on the web

By Giorgio Basile

In the rapidly evolving world of geospatial data, efficiency and accessibility are key. As datasets grow exponentially in both size and complexity, so does the need for robust systems to catalog, manage, and distribute these.

Fig. 1 - The organizations supporting STAC. (Credits: STAC Website)

In this scenario, Spatio-Temporal Asset Catalog (STAC) has emerged as a widely used standard for describing and cataloging geospatial data so that it can be easily discovered and accessed.
It is centered around the concept of a spatiotemporal asset, defined as "*any file that represents information about the earth captured in a certain space and time*".
Files related to satellite imagery are typical spatiotemporal assets, but any other geospatial file usually has a spatial and temporal reference that qualifies it for easy description through STAC.

The specification is a joint effort of developers and organizations in the geospatial community, with a considerable impact on how providers have shared their datasets in the last few years.

The STAC specification

STAC provides a common structure for describing and cataloging spatiotemporal assets. The main goal of the specification is to standardize the semantics used to distribute geospatial datasets on the Web and allow providers to use pre-defined language and API instead of having to develop their own. This saves costs and allows seamless access for users across datasets and providers.
The benefits for all actors involved are straightforward: data providers have clear guidelines for publishing their data, developers have ready-made tools to deploy the software infrastructure to build data catalogs, and users have a unified language and API to query any geospatial dataset.
Beyond the excellent language and concepts it provides, STAC's power relies on the community behind it — some of the most prominent organizations and developers in geospatial tech — and the set of tools built around the specification to facilitate and streamline the development and deployment of catalogs.
Microsoft Planetary Computer, Copernicus Data Space Ecosystem and NASA Earthdata are just some of the most prominent examples in a growing list of organizations and products that have chosen STAC as their metadata language and API to catalog their geospatial assets.
Those hubs publish several petabyte-scale datasets like Sentinel, MODIS and Landsat, effectively facilitating user exploitation.
STAC is developed on GitHub, the most popular platform for developers to collaborate on open-source software projects.
It is based on the JSON and GeoJSON data formats and organized as a set of four semi-independent specifications

that outline the fundamental concepts—Catalog, Collection, Item—and the RESTful endpoints to search, filter and retrieve Items of interest.

Use cases

Typically, a provider publishes a Catalog that holds multiple data Collections, each usually associated with a dataset.
Within each Collection are multiple Items, defined as an atomic set of inseparable Assets - i.e., all the bands and metadata files part of a Sentinel-2 satellite image can be seen as Assets, all within the same Item. Metadata associated with Assets or the entire Item can be annotated and exposed through the STAC API, enabling search based on those properties, along with space and time filters.
This results in the ability to specify queries to retrieve, for example, all Sentinel-2 L2A images over Italy in June 2024 with less than 20% cloud coverage and more than 50% vegetation coverage.
While these concepts represent the core part of the standard, STAC is an extensible language agnostic regarding the usage of custom metadata keys and values. Nevertheless, there are often common domains and use cases where further standardization can benefit multiple actors.
That's where STAC Extensions come in, providing additional schemas related to particular domains, topics or scenarios, which typically require further metadata to be used at the Collection, Item or Asset levels. A growing list of extensions at different maturity levels addresses common vocabulary to describe raster, electro-optical, and projections attributes, down to more specialized schemas for aerial and SAR imagery, among many others.
It must be noted that STAC Catalogs can be static or dynamic. Static catalogs are made of JSON files served through any HTTP server or by an object storage bucket, which provides one by default. Those files can be produced along with and placed "next to" the geospatial assets of a dataset, resulting in an almost zero-infrastructure catalog that can be created by simply storing static files on, e.g., Amazon S3.
Dynamic catalogs, instead, require a dedicated database and a STAC API to be deployed. However, their dynamicity allows for sophisticated search queries and pairs well with the conditions of existing providers, who are very likely to already have a database that indexes their available assets.

A rich ecosystem

As explained, one of the significant upsides of adopting an open standard like STAC is the community of developers and users gathering around it, resulting in valuable resources being created and shared among participants.
STAC Index is a website

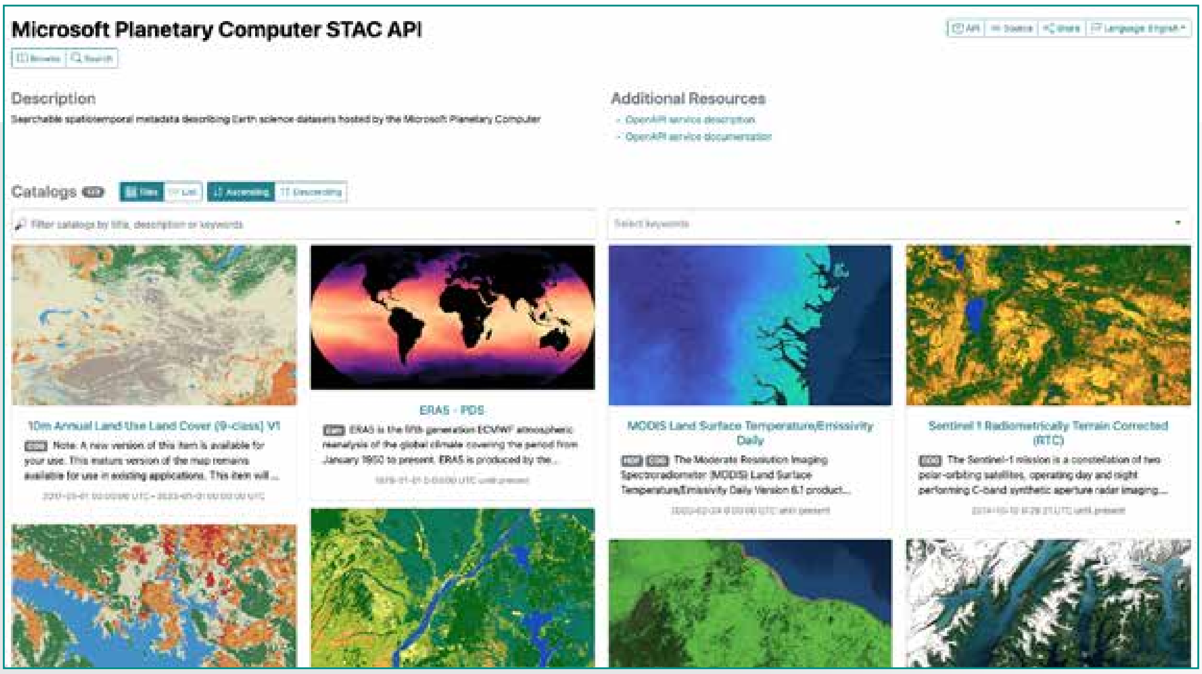

Fig. 2 - The Microsoft Planetary Computer is a reference implementation of STAC, leveraging several tools built by the community. Its catalog can be visualized through open-source software tools like STAC Browser.

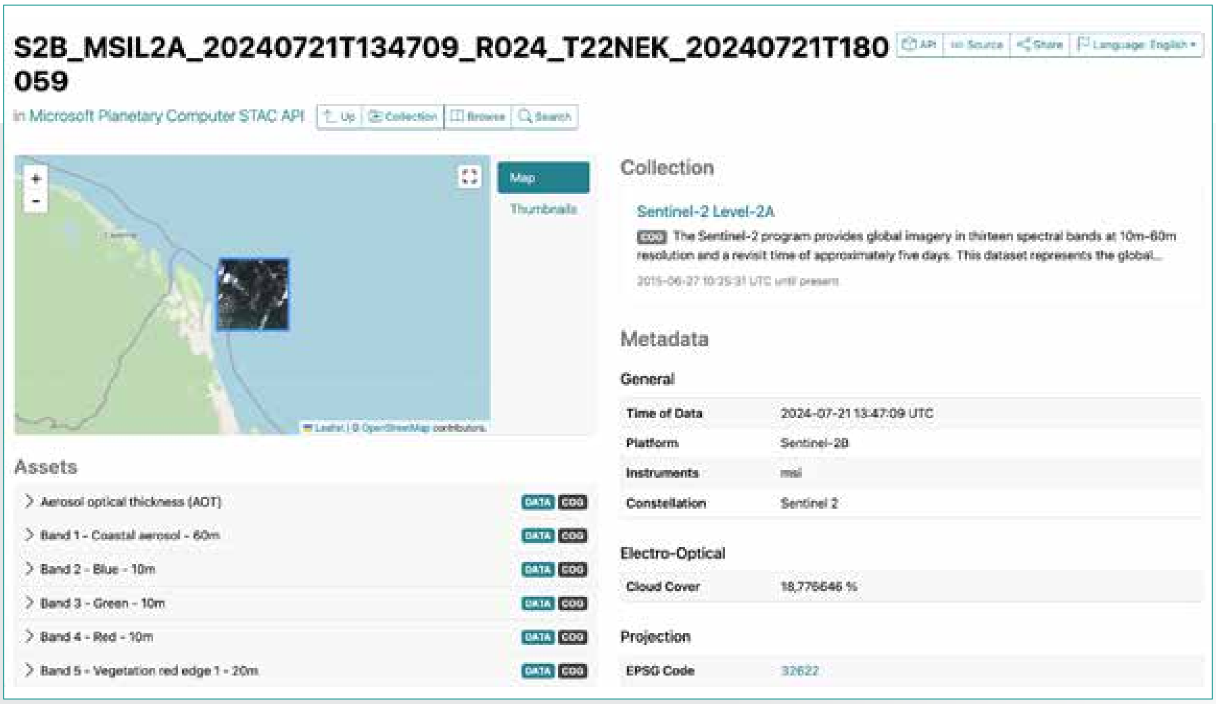

Fig. 3 - A STAC Item of the Sentinel-2 L2A collection offered by the Microsoft Planetary Computer. As bands are stored in separate files, each is referenced by a related Asset. Also, metadata is exposed through GeoJSON properties defined by dedicated STAC Extensions related to electro-optical and projection data, among others.

that collects resources to help newcomers build their catalogs or access data from any available ones. It also gathers several libraries developed in Python, Javascript and Rust, among others, that help manipulate STAC documents, store them in PostgreSQL or other databases, deploy STAC APIs through popular web frameworks, and query catalogs visually or programmatically. Moreover, STAC is supported by popular geospatial projects like GeoServer, QGIS, and OpenLayers, which enables interoperability with traditional frameworks for visualizing or analyzing geospatial data

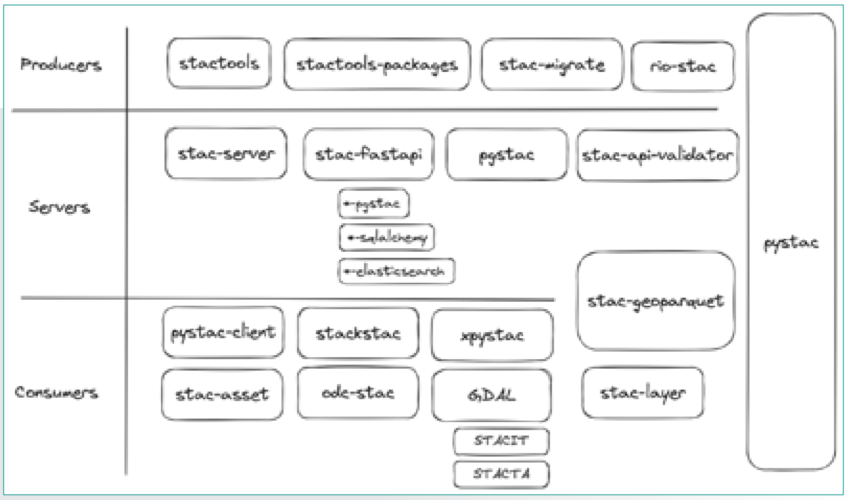

Fig. 4 - An overview of the tools provided by the STAC Community, divided into three categories. (Credits: stac-utils on GitHub).

Most of these tools integrate seamlessly with the storage facilities of public cloud providers, like Amazon Web Services, Google Cloud Platform or Microsoft Azure. This is especially important, considering that most datasets are increasingly being released in object storage buckets on those platforms, leveraging modern data formats—like Cloud-Optimized GeoTIFF, Zarr or GeoParquet—that allow the user to subset only the pixels or features necessary for their task to be completed. All these are very powerful and beneficial circumstances for providers and users. There is no need to develop new software packages or APIs to support dataset releases or learn how to use them. Also, suppose a provider stops sharing its data for any reason. In that case, users can quickly point to other catalogs offering the same data with minimal effort, mainly

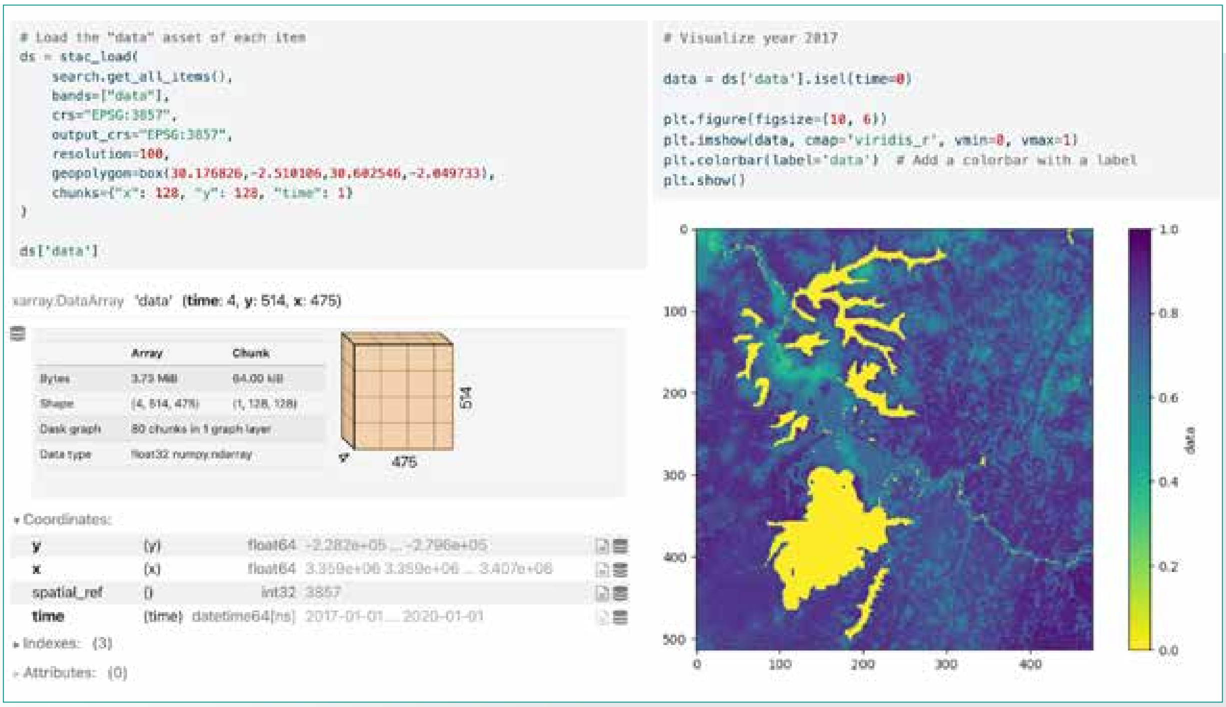

Fig. 5 - Python libraries like pystac-client and odc-stac allow users to find Items of interest and efficiently perform geospatial analysis. They interoperate with other popular data science libraries to simplify the work carried out by researchers and developers.

keeping the same workflow and processes in place.
Finally, a significant benefit brought by this ecosystem is the excellent education it provides regarding how to represent, share and access datasets on the web. It also lowers the barrier for newcomers to start using geospatial data and provides the necessary stability for developers to keep building more and better services and applications around it.

A steady adoption

In 2022, Planet Labs gathered statistics on about 500 million geospatial assets published through STAC. Those only take into account public catalogs referenced by the STAC Index, but a growing number of organizations are using STAC in their private infrastructure to power their Data teams and streamline their geospatial workloads.

The Microsoft Planetary Computer (https://planetarycomputer.microsoft.com) is one of the most important implementations of STAC, both for the relevance of the company behind it and for the choice to develop a core project for its sustainability-related efforts using open standards and technologies and publish all the software repositories on GitHub.
NASA Earthdata (https://www.earthdata.nasa.gov), the space agency's go-to geospatial data hub, provides an STAC endpoint in its Common Metadata Repository (CMR), allowing users to query data gathered from several Distributed Active Archive Centers (DAACs).
The Copernicus Data Space Ecosystem (https://dataspace.copernicus.eu/) is a recent effort by the European Union to provide a centralized hub for all data produced by its missions. STAC is among the supported protocols to access the available data.
Google Earth Engine (https://earthengine.google.com/) provides STAC metadata about all its datasets, although only Collection-level metadata is provided, given the very nature of Google's platform, which abstracts away the concept of geospatial assets or files.
Element84, a company that focuses on advancing geospatial technologies and promoting their adoption, curates EarthSearch (https://www.element84.com/earth-search) a STAC endpoint that catalogs assets related to datasets available in the AWS Open Data Registry.
These are just a few examples, in a growing list of companies like Planet Labs, UP42, VIDA.place and many others that have embraced STAC as their cataloging language and ecosystem of tools, fostering

interoperability within the overall geospatial community.

Conclusions

SpatioTemporal Asset Catalog (STAC) substantially improves geospatial data management by providing a standardized and flexible framework for cataloging and accessing spatiotemporal assets. This innovation enhances data discovery, integration, and interoperability, benefiting applications ranging from environmental monitoring to urban planning.

The adoption of STAC has streamlined the integration and interoperability of diverse datasets, enhancing users' ability to discover and utilize geospatial data efficiently. By adhering to a common specification, STAC facilitates the sharing of data across different platforms and organizations, reducing redundancy and promoting collaboration. Its open and extensible nature allows it to evolve with emerging needs, supporting a growing volume of diverse geospatial data.

In summary, STAC offers a robust solution for managing spatiotemporal data, fostering better decision-making and a deeper understanding of our world through improved geospatial data accessibility and collaboration.

REFERENCES

- STAC Website - https://stacspec.org
- STAC Specification - https://github.com/radiantearth/stac-spec
- STAC Index - https://stacindex.org
- stac-utils on GitHub - https://github.com/stac-utils
- Microsoft Planetary Computer - https://planetarycomputer.microsoft.com/
- State of STAC, by Planet Labs - https://developers.planet.com/blog/2022/Aug/31/state-of-stac/
- Building the VIDA Data Catalog with STAC and Cloud-Native Geospatial formats, by the Engineering team at VIDA.place, appeared on Medium - https://tinyurl.com/vida-catalog

KEYWORDS

Geospatial; data engineering; cloud native; Stac

ABSTRACT

The STAC is a recent geospatial standard that allows to describe and catalog geospatial assets. It is part of a broader innovation effort called Cloud-Native Geospatial, providing modern standards and tools to efficiently access raster and vector data in the cloud.

AUTHOR

Giorgio Basile

Geospatial Data Engineer
VIDA Engineering


INTERGEO®
2024
SEPT. 24 – 26
STUTTGART
INSPIRATION
FOR A SMARTER
WORLD
WWW.INTERGEO.DE
GET YOUR
TICKET NOW!
VOUCHER CODE: IG24-GEOMEDIA
EXPO
CONFERENCE
STAGE
NETWORKING
DVW
DVW e.V. – Gesellschaft für Geodäsie, Geoinformation und Landmanagement
SPONSORS:
AUTODESK
esri
HEXAGON
Trimble