Candid Core Framework: Enhancing Geospatial Metadata with Data Candor

Authors: Erin Trochim and Samapriya Roy

Remote sensing has a problem: we don't know where the forest is.

Despite our best efforts, the global abundance of geospatial data has created a paradox of choice. We've made better land use land cover data to identify trees, applied global definitions of forests, and estimated canopy heights. Yet, amid our pursuit of comprehensive and accurate datasets, we often overlook a critical question: Does our metadata genuinely equip users to understand and effectively leverage this information?

Relying on technical documentation and peer-reviewed papers assumes that users have the time and expertise to unpack inherent limitations and contradictions. But if you're exploring forests in the United States, can you confidently select a dataset that extends seamlessly into Canada? In our ambition to map the globe, have we lost sight of what our products are for and whether we'd use them in our neighborhoods?

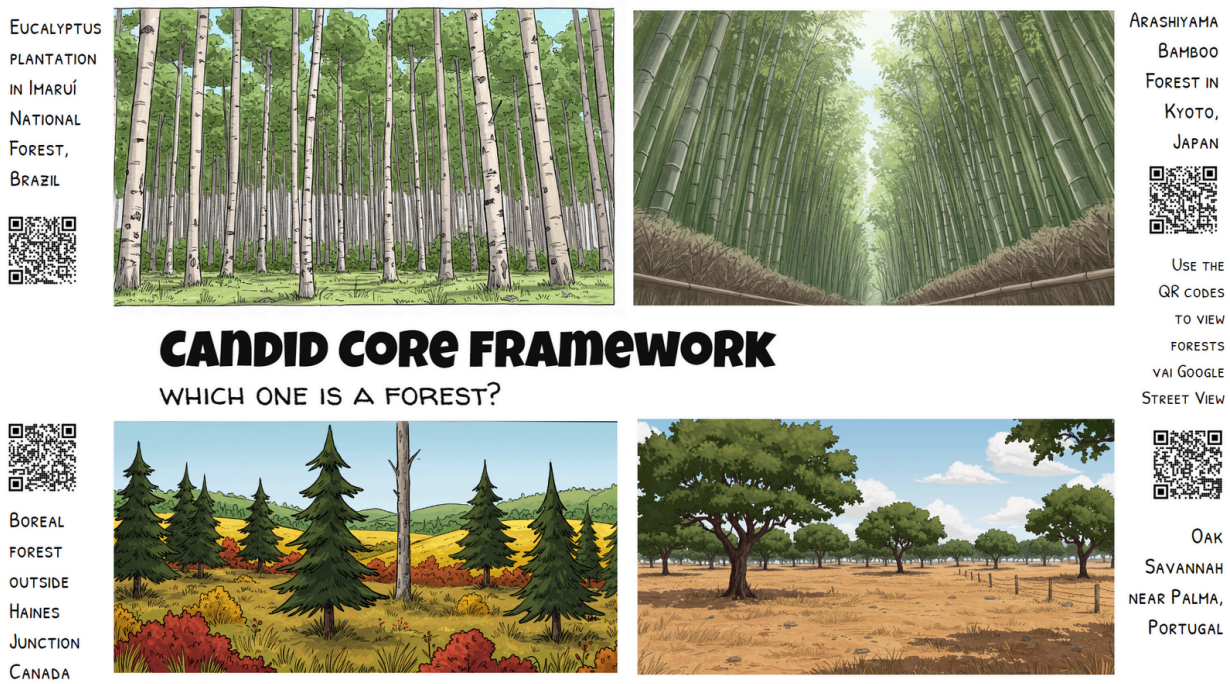

Figure 1: Forest or not? Demonstrating the inherent ambiguity of the term "forest," highlighting why datasets describing forests require transparent, context-rich metadata. QR codes (eucalyptus plantation, bamboo forest, boreal forest, and oak savannah) allow exploration of each location via Google Street View, emphasizing the importance of location-specific context. See related work here on A Comparison Shopper's Guide to Forest Cover Change Datasets.

Just as knowing only a tree's height and age reveals little about the broader ecological health of a forest, having extensive metadata often doesn't guarantee a meaningful understanding of data usability. In the quest to democratize data access, context matters more than ever. Are we ready to move beyond conventional metadata to embrace a more profound honesty—a radical candor in data stewardship? We believe the time is now. Building genuinely intelligent, user-focused, and trustworthy geospatial ecosystems requires putting metadata front and center, infused with transparency. Welcome to the age of **Data Candor**, where usability and openness are not afterthoughts but core pillars of our data infrastructure.

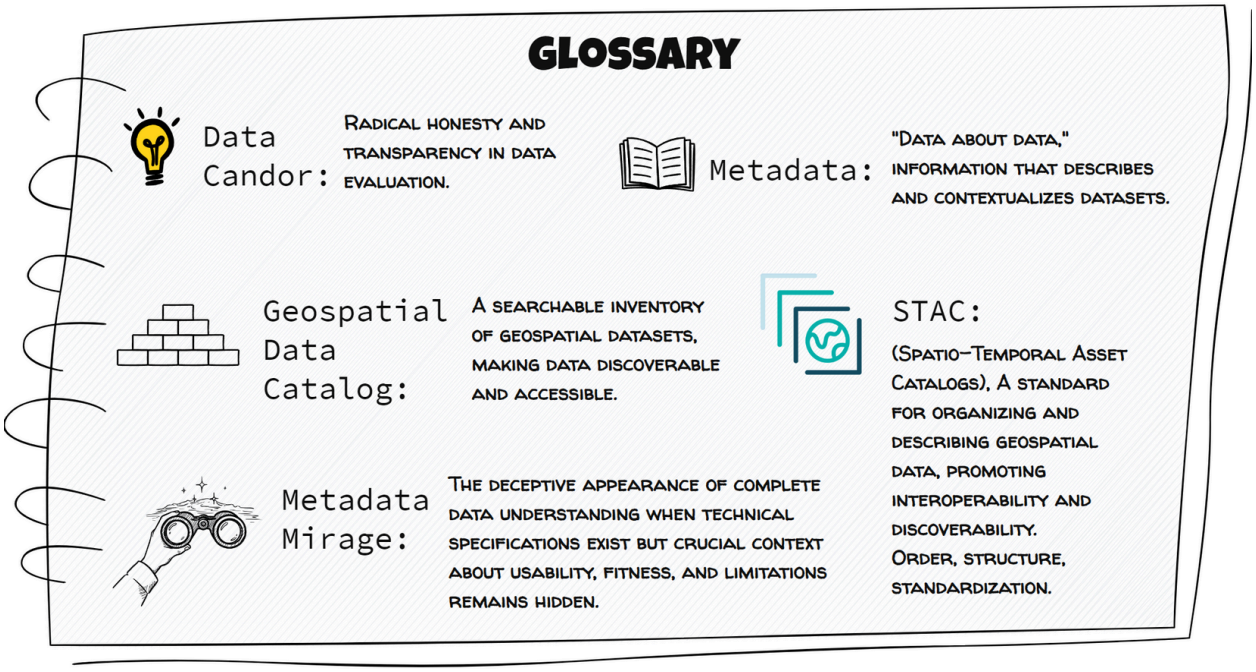

Figure 2: Glossary of key terms in Candid Core Framework

TL;DR

- Geospatial data is abundant, but finding the correct data for a specific task remains challenging. Current metadata (even STAC) often lacks crucial information about a dataset's limitations and appropriate uses.
- We call this the **Metadata Mirage** – the difference between finding data and truly understanding its fitness for purpose.
- The **Candid Core Framework** suggests a new layer of principles to existing metadata, providing honest, user-focused information about data quality, limitations, and recommended uses.

- By prioritizing human-curated evaluative information alongside technical specifications, we aim to build more trustworthy and usable geospatial data ecosystems.

WHERE ARE THE FORESTS

At first glance, asking "Where are the forests?" might seem simple, but this seemingly straightforward question quickly reveals layers of complexity. Definitions of forests vary significantly between countries, institutions, and even individual researchers, shaped by cultural, ecological, and economic perspectives. For some, a forest is identified simply by canopy cover percentage; for others, it's determined by specific ecosystem characteristics or land-use categories. Consequently, the datasets representing forests often embody diverse assumptions, challenging direct comparisons or combinations.

This ambiguity extends beyond forests to wetlands, coastlines, water bodies, and numerous other geospatial categories facing similar definitional complexities. When datasets are created with varying standards or purposes, the core question shifts from mere availability to a more subtle, essential consideration: How do we determine their practical value? The heart of this dilemma lies in recognizing that value is not inherently tied to just data quantity, ease of access, or following specific low-level technical standards but instead profoundly embedded in usability, clarity of context, and transparent documentation.

INTRODUCING THE CANDID CORE FRAMEWORK

Building on the challenges identified in geospatial data discovery and application, the **Candid Core Framework** offers a structured yet adaptable solution that embodies the principle of **Data Candor**—enabling users to evaluate datasets for their specific needs confidently. This framework empowers users to confidently and systematically assess geospatial datasets to accurately determine their fitness for specific applications. Rather than replacing established metadata standards like STAC, the Candid Core Framework complements them by adding an essential, curated layer of evaluative information, clearly communicating practical usability and limitations.

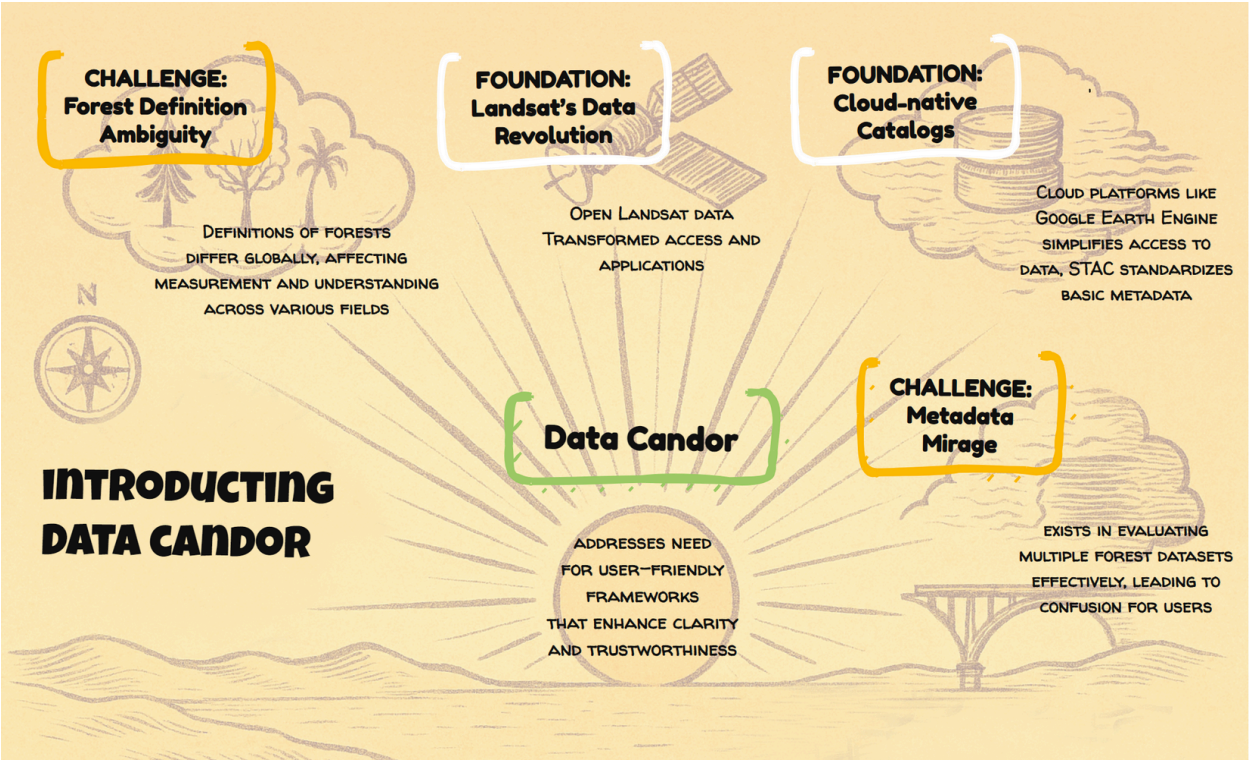

Figure 3: The Candid Core Framework's conceptual foundation. This visualization illustrates how Data Candor (the rising sun) illuminates and connects the four key areas in geospatial data management: Forest Definition Ambiguity, Landsat's Data Revolution, Cloud-native Catalogs, and the Metadata Mirage. As Data Candor rises, it brings clarity to these interconnected challenges through a user-centered approach to metadata transparency.

The Candid Core Framework is organized around key evaluation categories, where each dimension adds nuance for understanding a dataset's honest strengths, candid limitations, and appropriate and responsible applications. These categories, detailed in the sections that follow, include aspects designed to promote Data Candor, such as:

- **Dataset Description** - *Honest Context*: Going beyond basic technical specifications to capture the nuances of data collection, processing, and intended purpose with complete transparency and candor.
- **Suggested Uses** - *Candid Applications*: Providing curated examples of how and by whom the dataset has been or could be effectively utilized, drawing upon community knowledge and expert insights to offer honest appraisals of application suitability.
- **Application Limitations** - *Transparent Caveats*: Explicitly and candidly outlining known limitations, potential biases, and inappropriate use cases to promote responsible data application and effortless data understanding.

- **Review & Trustworthy Provenance Information:** Providing full transparency into the data's review process and sources, enhancing user trust and understanding of data quality through honest accounting of data origins.
- **Community Vibes** - *Authentic Perspective*: Capturing qualitative and contextual information that may not be easily formalized is crucial for understanding the dataset's broader relevance, impact, and even its unique "personality" with honesty and authenticity.

While some aspects of the Candid Core Framework conceptually overlap with emerging STAC extensions—particularly those addressing data quality, provenance, or application-specific details—this overlap is deliberate and complementary. In the spirit of Data Candor, the Framework provides a curated perspective, leveraging STAC's extensibility to surface evaluative insights transparently and consistently.

The Role of LLMs and Human Curation

At the heart of the Candid Core Framework is that human oversight remains essential for building trustworthy, reliable knowledge ecosystems. While Large Language Models (LLMs) can support category standardization or initial content generation, the Framework is intentionally designed around human curation, prioritizing expert review and community-driven input. Human expertise, transparent community contributions, and continuous refinement ensure the provided evaluative information's accuracy, relevance, and trustworthiness. In alignment with Data Candor, our goal is not exhaustive, automatically generated metadata that overwhelms users—but concise, meaningful insights crucial for quickly assessing a dataset's practical utility and promoting responsible, informed use.

IN THE BEGINNING: EVOLUTION OF GEOSPATIAL DATA ACCESS

Early Challenges: The Download Era

Cloud-native geospatial data catalogs, co-locating data with compute, represented a pivotal innovation within the geospatial community—and Landsat data served as a crucial demonstration of this potential. Landsat offered a vast repository of high-value remote-sensing imagery. By 2008, it had moved to a new business model where data was freely available, a shift that unlocked immense long-term benefits. What if the need to download data was removed, and the archive of all data could be accessed for global applications looking at temporal variability? This approach paved the way for the utility of cloud-native geospatial data catalogs.

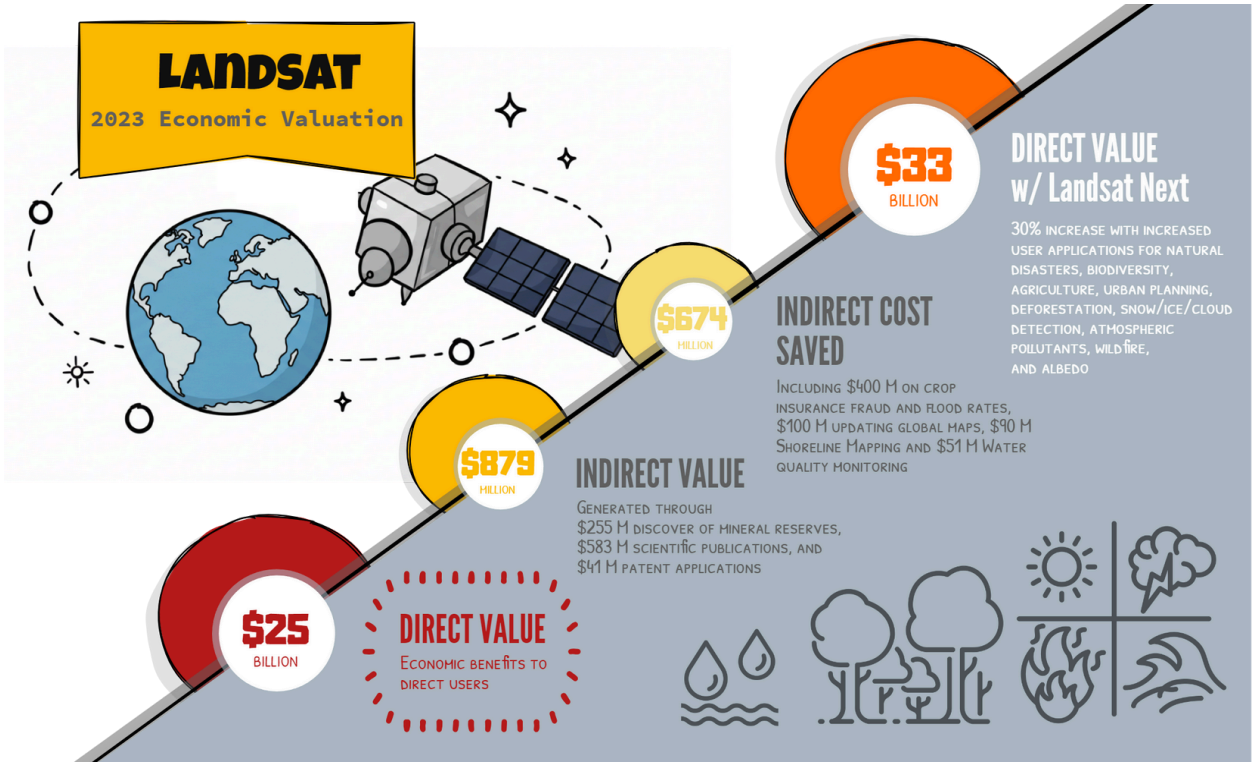

Figure 4: Infographic is derived from the "Economic Valuation of Landsat and Landsat Next 2023" report. It indicates that the direct economic value of Landsat in 2023 was $25.6 billion.

Geospatial data, traditionally represented in raster and vector formats since the 1960s, faced significant access hurdles in its early digital era. While online directories like GLIS (1993) and EarthExplorer (2000) provided remote access to Landsat and other earth observation data, the "download model" prevailed. Raster data remained underutilized throughout the 1990s and early 2000s due to slow download speeds and limited local computing power, and data was often distributed using physical media like CDs. This technological constraint kept data processing local, making Geographic Information Systems dependent on users downloading data to their systems. Those with access to faster networks and more memory had a clear advantage in their analysis. However, this era of data scarcity and localized processing began to recede in the 2010s with the advent of cloud-based geospatial platforms, setting the stage for a fundamental shift—from simply accessing data to practically evaluating and leveraging its value.

The Cloud Revolution: Co-location and Catalogs

Google Earth Engine marked a turning point in geospatial data usability, widely recognized for pioneering one of the first cloud-native geospatial data catalogs. The Earth Engine data catalog documents each dataset using the following schema: (1) a dataset *Description*; (2) a description of image *Bands* including names, units, min/max; (3) the *Image* or *Table* properties

including data type and description; (4) *Terms of Use* which include the license information; and (5) attribution information listed under *Citations*. In 2018, it introduced per-dataset code snippets for immediate visualization, which can be found under each data description.

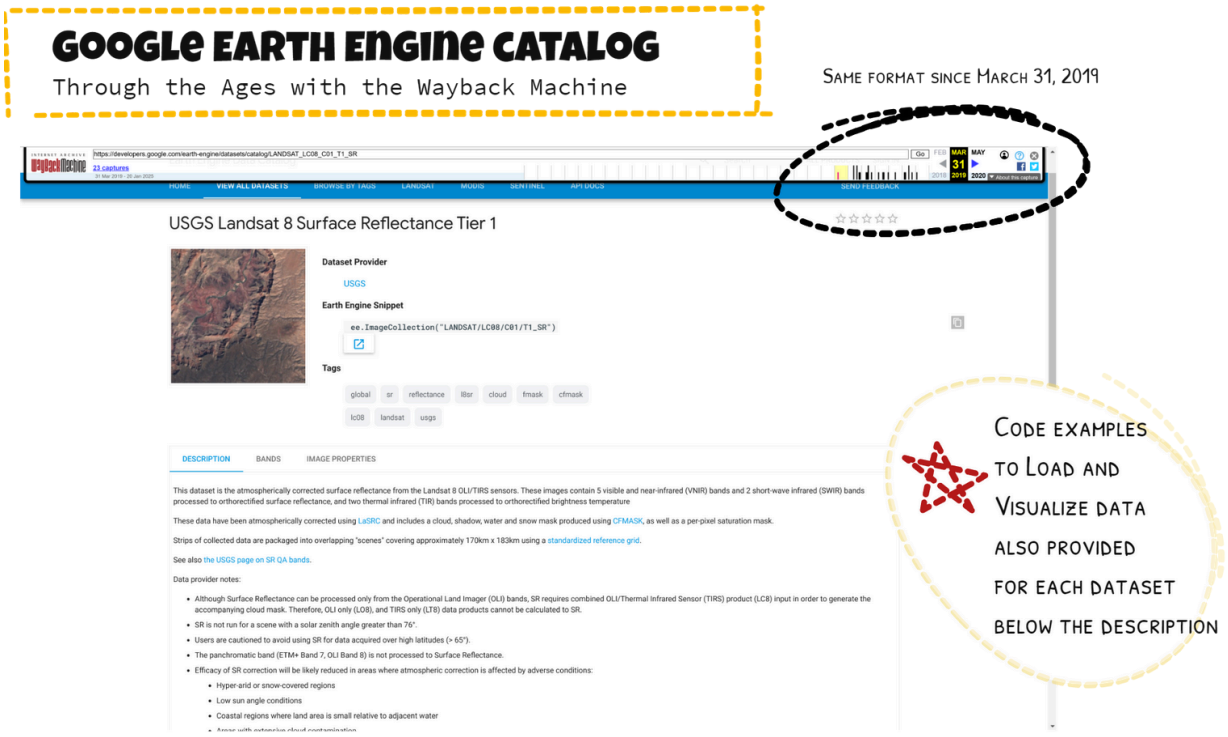

Figure 5: A screenshot from the Wayback Machine (March 31, 2019) showing a deprecated entry in the original Google Earth Engine Data Catalog for the USGS Landsat 8 Surface Reflectance Tier 1 dataset. This illustrates the catalog's format, which has remained stable until now.

The **Awesome Google Earth Engine Datasets - Community Catalog** emerged in 2021 and solidified into its current form in 2022, building on Earth Engine's proven catalog structure. By offering detailed descriptions, visualizations, and code snippets, this community-driven approach allowed users to discover datasets and actively refine and contextualize them. Like the original Google Earth Engine Data Catalog, which uses its own Issue Tracker, the Community Catalog uses Github to have users suggest or contribute datasets, including issues with datasets and community examples with the data. However, even with this level of detailed documentation and community curation, users still needed deeper guidance to understand each dataset's practical suitability for their specific needs.

Recognizing this challenge, the **SpatioTemporal Asset Catalog (STAC)** surfaced as a standardized framework for describing and organizing geospatial data. It provides a consistent way to describe geospatial data, regardless of the source or format. Enabling programmatic

access to geospatial data automating data retrieval and processing, STAC also helps to create searchable catalogs and APIs, making it easier for users to find the needed data. Interoperability is promoted between geospatial systems and tools, leading to a more connected and efficient ecosystem. STAC is an open standard with a growing community of users and developers, promoting collaboration and innovation through developing new extensions. Yet, despite these improvements, standardization alone didn't fully resolve users' uncertainty regarding a dataset's real-world applicability.

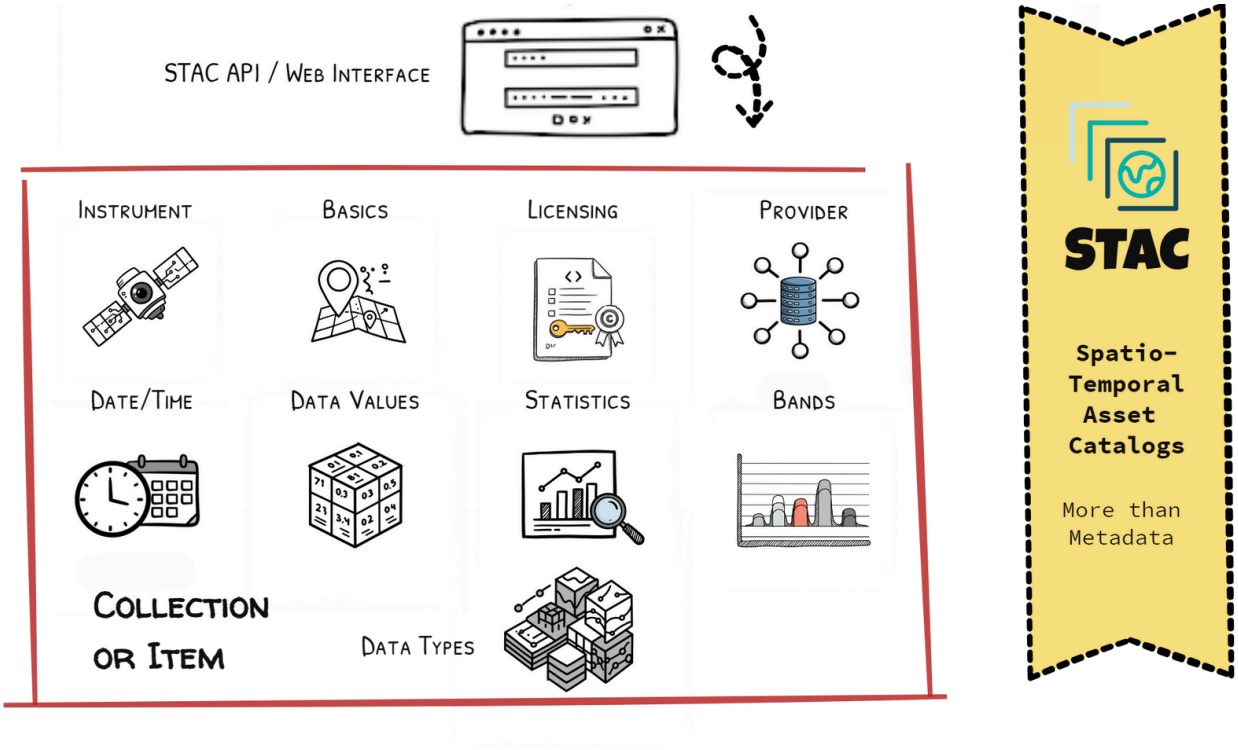

Figure 6: Common elements of STAC accessed through API / web interface. All pieces of information generally need to be filled out.

Google Earth Engine now employs STAC standards for dataset metadata, significantly enhancing dataset discoverability and cross-platform integration. But for many users, this standardization has only partly solved the deeper problem. While STAC made datasets more straightforward to find and technically compatible across platforms, it has not fundamentally enhanced users' ability to judge whether a dataset meets their specific analytical needs quickly. Much of STAC's metadata was already accessible through earlier documentation, leaving a critical challenge unsolved: **how can users swiftly, accurately, and confidently assess a dataset's practical value?**

The Remaining Gap: Beyond Discoverability

Building upon the foundation of cloud-native geospatial data catalogs and standardized metadata like STAC, the geospatial community faces a pivotal question: Are we genuinely candid about data usability and evaluability, especially for broader, more diverse user groups? Are we transparent about inherent dataset complexities and limitations? Addressing these questions directly, an "AGU Retreat" was convened in Washington, D.C., preceding the 2024 American Geophysical Union Fall Meeting. More than a metadata workshop, it was a call for greater transparency and user-centricity. Representatives from major Earth Engine catalog initiatives, a researcher specializing in data usability, and a product manager dedicated to user experience converged around a shared vision: embedding **Data Candor** directly into the core of geospatial metadata.

THE AGU RETREAT: A CALL FOR DATA CANDOR

The retreat's central goal was identifying what type of fundamental honesty was missing from current geospatial metadata practices. Starting with the existing STAC framework—which effectively describes basic geospatial data parameters—the group explored the core conceptual components needed for a metadata framework embracing **Data Candor** in evaluating geospatial datasets, moving beyond mere technical description to transparent and user-centered insights. A key pain point fueled this discussion: after significant effort to build accessible data catalogs, the critical bottleneck wasn't discoverability anymore but clarity around usability and evaluability. The new challenge centered data discovery on users' ability to use metadata to quickly and candidly assess a dataset's suitability – and limitations – for their specific applications. In short, effective data use now depends less on locating datasets and more on swiftly understanding their real-world value through candid evaluation.

Figure 7: The goals and key elements discussed during the AGU Retreat in December 2024 illustrate the focus on evaluability, data properties, and user attention.

Traditional metadata approaches are crucial for data documentation and interoperability. Still, they often prioritize strict ontologies and predefined properties – essential for the data's technical integrity but not always for user understanding. The retreat participants, embracing the spirit of **Data Candor**, questioned whether these rigid structures were sufficiently candid to address the nuanced challenges of data evaluation from a user's perspective. They also explored new tools. Could LLMs, guided by Data Candor principles, assist in standardizing and harmonizing evaluation categories, moving towards user-focused curation rather than exhaustive and sometimes overwhelming documentation? Could LLMs help scale the candid expert guidance experienced users provide informally – the honest insights into data strengths and weaknesses – with newcomers navigating the complex geospatial landscape? Ultimately, participants envisioned metadata as an honest, practical guide, not a technical barrier.

The fundamental challenge identified at the retreat—the **Metadata Mirage**—remains the issue limiting the full potential of geospatial data to connect to applications. This connection gap needs a bridge.

Geospatial data is inherently complex, starting from its fundamental diversity in representation—ranging from scalar fields to vector geometries and multidimensional tensors—and extending to various sensing methods, such as multispectral imaging, radar, lidar,

and more. Additionally, datasets may derive from direct physical measurements, sophisticated modeling processes, or combinations thereof. Such complexity confronts users with a bewildering range of choices when selecting suitable datasets or analytical approaches, regardless of whether their focus is academic research or practical problem-solving.

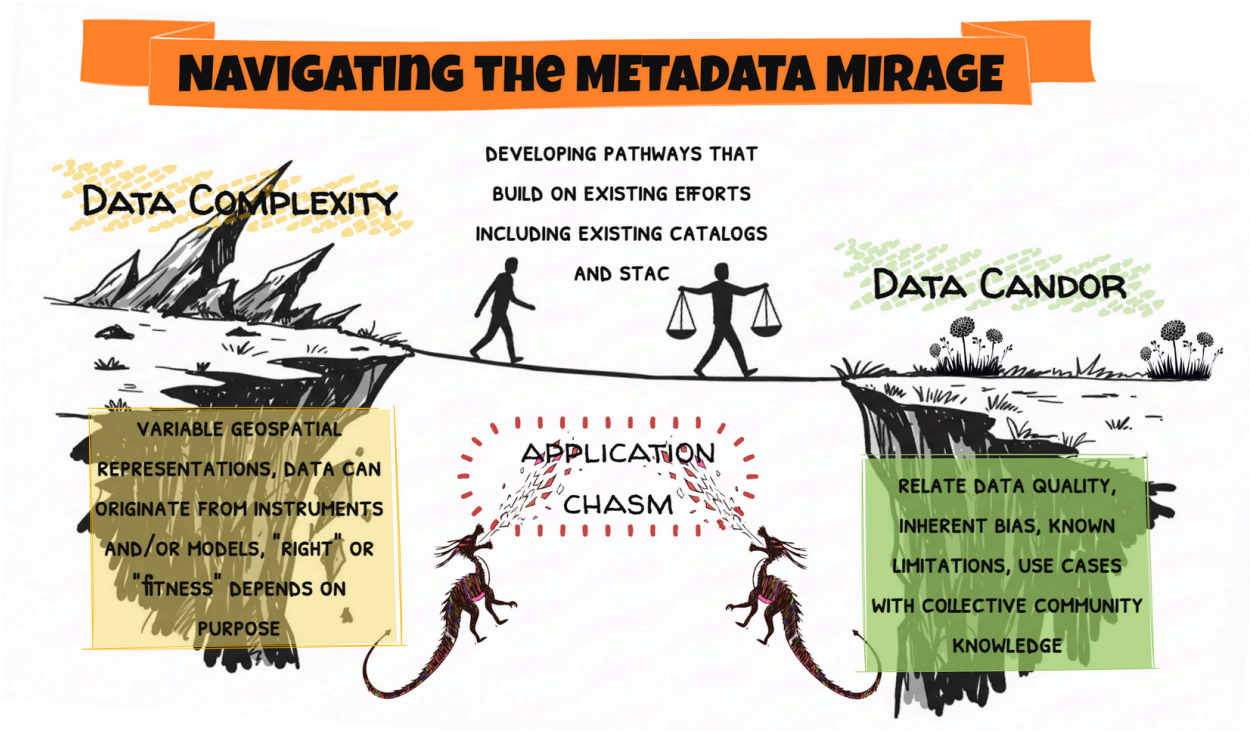

Figure 8: Illustrates the characteristics of data complexity that contribute to the Metadata Mirage, highlighting the need for Data Candor to navigate the chasm between data availability and practical application.

To illustrate, consider returning to our forest example. If we ask, *"What forest datasets are available for Canada?"* a newly released LLM-powered EE catalog search returns at least five different Canada-specific datasets on forest inventory, forest harvest disturbance, forest land cover, wildfire disturbance, and forest age. All these datasets leverage Landsat imagery, yet each captures a different facet of forest characteristics. At this critical juncture, how can a user confidently evaluate dataset appropriateness beyond merely relying on assumptions or superficial descriptions about what each product delivers?

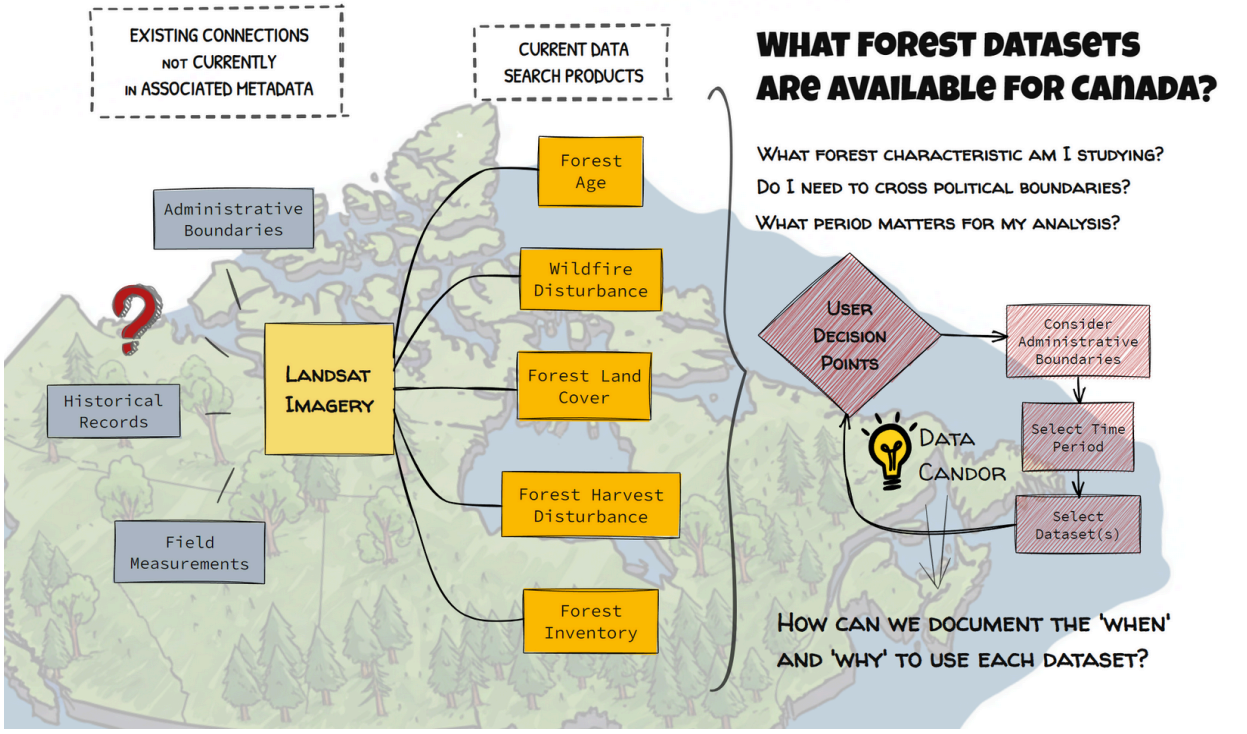

Figure 9: Forest Dataset Selection Framework for Canada. This visualization illustrates how multiple specialized datasets derive from common Landsat imagery, requiring additional contextual information (gray boxes) not typically included in technical metadata. Data Candor principles bridge the gap between data availability and effective use by documenting when and why to select specific datasets for particular applications.

Further muddling this issue is the inherent trade-off between spatial and temporal resolutions: datasets with finely detailed spatial coverage might lack the temporal frequency needed to capture dynamic phenomena adequately, and vice versa. These intertwined characteristics emphasize why "fitness for purpose" heavily depends on the user's specific application context and domain expertise—and why transparent, candid metadata guidance (Data Candor) is indispensable.

Dispelling this **Metadata Mirage** requires fundamentally rethinking geospatial metadata beyond purely technical specifications. Effective metadata must transparently communicate evaluative insights into crucial factors such as articulated data quality, explicitly acknowledged biases, candid limitations, recommended use cases informed by real-world insights, and collective community knowledge shared transparently. This essential evaluative information remains fragmented, partially known to data producers, hidden within user trial-and-error experiences, or implicitly held by domain specialists. This fragmentation creates a critical chokepoint, constraining efficient and scalable data evaluation and, ultimately, limiting the practical utility of available datasets. Recognizing this critical gap motivated the structured, scalable approach embodied by the **Candid Core Framework**.

CONCLUSION: BUILDING TRUSTWORTHY KNOWLEDGE ECOSYSTEMS

The **Candid Core Framework** represents a fundamental shift in approaching geospatial metadata – while STAC offered incremental improvements, we need something closer to Landsat's shift to free data access. This fundamental business model change unleashed over $25 billion in economic value by expanding who could use Earth Observation data and what they would do with it. **Data Candor** has similar revolution potential.

Even incremental implementations of the concept can change our impact. We don't need to abandon existing infrastructure; our challenge is more on supporting better "community-curated data resources in the Earth sciences, highly valuable but systematically underfunded" that form the "critical foundations of most scientific disciplines." The **Metadata Mirage** persists because, like these vulnerable long-tail databases, the contextual knowledge about dataset applications remains fragmented and undervalued despite its enormous potential worth. We need coordination and support from private and public institutions to recognize both the societal and economic benefits that twin supporting these trustworthy knowledge ecosystems.

The scientific impact of **Data Candor** could rival or exceed that of Landsat's open data policy. By embracing radical transparency as a core principle, we won't just incrementally improve metadata—we'll fundamentally transform how Earth observation data creates value, moving beyond merely connecting users to data and beginning to connect them to actionable knowledge.

When we move beyond asking 'Where are the forests?' to understanding forest context, we unlock new value. Similarly, our focus must shift from questioning if **Data Candor** is necessary to identify which organizations will drive this revolution and help fulfill its promise for science and society.

Thanks to Simon Ilyushchenko, Renee Johnston, and David Schottlander of Google for their valuable insights and contributions. This work was supported by a Google Research academic research award, with support from Matthew Abraham and Sameera Ponda. AI was used to edit and revise text and assist in generating figures. This work can be found at https://doi.org/10.5281/zenodo.15227664