The Evolution of GCMD Keywords — An instructive tale of data standards development and adoption

An ethnographic essay by

Mark A. Parsons
University of Alabama in Huntsville
https://orcid.org/0000-0002-7723-0950

Ruth Duerr
Ronin Institute
https://orcid.org/0000-0003-4808-4736

Øystein Godøy
Norwegian Meteorological Institute
https://orcid.org/0000-0001-6410-3488

Author Roles according to CRediT — the Contributor Roles Taxonomy:

Parsons: Conceptualization, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing
Duerr: Conceptualization, Writing – review & editing
Godøy: Conceptualization, Investigation, Writing – review & editing

Abstract

NASA established the Global Change Master Directory (GCMD) and supporting keywords in the early 1990s as part of implementing the GCMD through the Directory Interchange Format or DIF. The GCMD was developed to support the launch of the huge and enduring satellite-based Earth Observing System (EOS). The primary intent was to catalog EOS and related data, but the keywords have been implemented in many different systems and adopted in varying ways by many different organizations around the world. This essay provides an ethnographic examination of how the keywords have evolved and been managed and how they have been adopted over the last few decades. It illustrates how semantic approaches have evolved over time and provides insights on how standards and associated processes can be sustained and adaptable. Ongoing institutional commitment is essential, but so is transparency and technical flexibility. Understanding the different roles involved in standards creation, maintenance, and use of standards as well as the services that standards enable is also critical. It is apparent that semantic representations need to be mindful of different contexts and carefully define verbs as well nouns and categories. Understanding and representing relationships is central to interdisciplinary interoperability.

Introduction

Data discovery, interoperability, usability, and computation depend on standards — those detailed specifications and conventions that we take for granted when everything is working well. But standards are tricky. They come into force in different ways. Sometimes they emerge and establish quickly (e.g., http/html). More often they emerge through long periods of (open) consensus building either formally through standards organizations or informally through praxis (e.g ISO19115 and associated profiles). Sometimes standards emerge surreptitiously. A good idea for a particular purpose becomes broadly adopted for other purposes. (Russell 2014). The Global Change Master Directory (GCMD) set of keywords is one such example.

In this essay, we examine the almost accidental way that the GCMD keywords became a community standard throughout the course of several decades. We use this story of an enduring and influential "standard" to illustrate the complexities of how standards develop and how they are adopted.

GCMD Keywords are a hierarchical set of controlled Earth Science vocabularies that help ensure Earth science data, services, and variables are described in a consistent and comprehensive manner and allow for the precise searching of metadata and subsequent retrieval of data, services, and variables. Initiated over twenty years ago, GCMD Keywords are periodically analyzed for relevancy and will continue to be refined and expanded in response to user needs.

The categories of GCMD Keywords are as follows.

- Earth Science
- Earth Science Services
- Data Centers/Service Providers
- Projects
- Instruments/Sensors
- Platforms/Sources
- Locations
- Horizontal Data Resolution
- Vertical Data Resolution
- Temporal Data Resolution
- URL Content Types
- Granule Data Formats
- Measurement Names
- Chronostratigraphic Units

Earth Science keyword sets use this hierarchical structure:

Category > Topic > Term > Variable > Detailed Variable

To date, the only Category has been Earth Science. Topics represent disciplines and high level concepts (eg., Atmosphere). Terms and Variables define subject areas and parameters. Detailed Variables are uncontrolled values that can be added by users to more specifically describe data. [generally quoted from different parts of gcmd.nasa.gov]

The GCMD keywords are not a formal international standard in the sense of being endorsed by an organization like ISO or similar, but they are broadly adopted in different ways across many countries, agencies, and institutions. The keywords are a de facto standard interpreted and used differently in many different contexts. How did this de facto standardization come about? How consistent is the adoption? How satisfied are the users? What goals does the standard meet? What can we learn from all this? How might it influence our current practice? In particular, what can NASA learn from this as it develops and implements a new standards process across its entire Science Mission Directorate, which extends well beyond Earth science?

We cannot fully answer all these questions, but experience and observation may provide some wisdom or at least some rules of thumb. We approach this as long-time users and proponents of the keywords — as data stewards at two early NASA data centers and the Norwegian Meteorological Institute, as leaders in multiple international

communities working to establish data sharing and interoperability arrangements[1], and as active participants in and researchers of socio-technical systems.

We emphasize that this is a critique not a criticism. GCMD deserves immense credit for developing and especially maintaining a broadly adopted community standard. Our goal is to illustrate how the research community and their institutions came together to define and implement a particular standard. It is both a cautionary tale and a success story. And like all good stories, it provides some lessons and insight.

Early Days

NASA established the Global Change Master Directory (GCMD) and supporting keywords (see Box) in the early 1990s as part of implementing the GCMD through the Directory Interchange Format or DIF. The GCMD was developed to support the launch of the huge and enduring satellite-based Earth Observing System (EOS). The primary intent was to catalog all EOS *and related* data. That effort continues today through the NASA Common Metadata Repository (CMR), but the more instructional story around standards is the evolution of the keywords, their governance, and their use and adaptation beyond NASA as a de facto community standard. In many ways, the keywords became more important than the directory itself.

The science keywords, in particular, were quickly and widely adopted internationally in various Earth and climate science communities. Adoption continues today. An ongoing legacy of the effort is that subsets of the GCMD science keyword hierarchy frequently appear as the browse interface facets for many data repositories and portals. See, for example, a light implementation in Switzerland or this long-standing, rigorous adherence in Australia. The World Meteorological Association (WMO) included them in the WMO Core Metadata Profile and the associated WMO Information System. These are just a few examples. A more complete list of current keyword adopters can be found on the GCMD Keyword page including multiple US federal agencies and international organizations and many individual centers.

At first, the management and authority of the keywords was controlled by GCMD staff without explicit guidelines. Some keyword Categories like Project were easily amended with a simple request from the project. The science keywords, on the other hand, were more strictly controlled. GCMD staff would work with one or more organizations deemed to be an authority on the topic.

For example, in the late 1990's and early 2000s, the National Snow and Ice Data Center (NSIDC) supplemented and refined the definitions of multiple Variables within the Cryosphere Topic and developed the whole Frozen Ground Term and associated Variables. As the recognized snow and ice data archive for NASA, NSIDC was the naturally accepted authority on the Cryosphere Topic and had data that needed to be categorized accordingly. NSIDC worked

[1] Relevant international communities and organizations include: the International Polar Year, the Research Data Alliance, the Data Committee for the International Arctic Science Committee and Sustained Arctic Observing Networks, the World Meteorological Organization's Information System and Global Cryosphere Watch, the International Science Council's Committee on Data (CODATA) and World Data System, and the Earth Science Information Partners especially its Semantic Technologies Committee.

with the cryospheric research community to develop the terminology and definitions, drawing from the literature where appropriate, but it was an ad hoc and informal process. The basic approach from GCMD staff seemed to be: “If you have a bunch of metadata records to provide that don’t jive with current terminology and you represent a particular science community, then let’s talk.” This was a sensible and workable process at first. There were relatively few Earth science metadata standards and vocabularies at the time. Through the EOS program, NASA was beginning to produce unprecedented volumes, coverages, and varieties of Earth system science data. NASA understandably drove the conversation. While NASA was building an interchangeable directory, it was primarily building a system for NASA data.

In the early days, the GCMD organization, under the leadership of Lola Olsen, encouraged international organizations to contribute metadata to the directory. Many portals were developed by flagging certain records as part of a particular program or project. Venerable examples include the Committee on Earth Observing Satellites (CEOS) International Directory Network (IDN) and the Antarctic Master Directory. The directory was hosting metadata from organizations all around the world, but formal control of the metadata was held by GCMD. GCMD staff would even edit metadata submissions and sometimes add or remove science keywords. These edits were not necessarily communicated to the submitter of the metadata. The model was akin to creating a library catalog. It was only logical for an expert “librarian”, the GCMD staff, to control the content of the *Master Directory.*

Web Services as an Inflection Point

As the scale of climate science, data, and collaboration grew, this centralized and closed process was no longer appropriate to the increasingly interdisciplinary problem at hand. Historically, computing systems have vacillated between centralized and distributed models. Perhaps because of the success of Google over Yahoo, the already distributed Web was moving away from centralized or authoritative directories to more distributed, scale-free networks. By the mid to late aughts, web services were seen as standards to connect different scientific resources across the web.[2] The greatest value of the GCMD was increasingly the broad acceptance of its science keywords and not the centralized catalog function. The GCMD organization seemed to be slow to recognize this. In many ways, the keywords had become a standard but there was no API or easy way to adapt to new versions. Furthermore, metadata in the GCMD may not be in sync with that held at the local repository, which may in turn be shared with other systems on the Web. Metadata authority and interoperability began to drift.

The experience of the International Polar Year (IPY) 2007-8 illustrates the issue. IPY was a US$1.2 billion investment in polar research involving 50,000 participants from 60 nations in an intense, coordinated polar observation and research program (Carlson 2010). IPY had a very forward looking data policy generally requiring open and interoperable data sharing. While there were no funds to develop a formal data service, an international committee worked to coordinate federated discovery and interoperability across dozens of multi-disciplinary data centers around the world. The Antarctic community had been using the GCMD for years, and the formal Standing Committee on Antarctic Data Management was embarking on an effort to

[2] This includes “abandoned” standards like SOAP/WSTL, formal standards like those from OGC, de facto standards like OpenDAP, and the general REST-like services now generally presented as APIs.

improve their metadata in the GCMD and provide more direct links to data.[3] The GCMD director offered, and it seemed logical for the less-coordinated Arctic community, to simply join that effort and create an IPY master directory.

After the first meeting of the IPY data committee and an associated community workshop, a more nuanced approach was recommended. The Arctic brought in much more scientific diversity than the Antarctic including much more terrestrial ecology, medical and social sciences, and the explicit inclusion of Indigenous knowledge. Therefore, the idea was to have multiple catalogs interconnected through open web protocols in a “union catalog”. Each catalog could be tailored to its own community while also sharing basic metadata with others. This was an early conception of graph-based, federated search systems implemented today through schema.org and related web technologies (e.g. Google dataset search). The group explicitly recommended “using appropriate harvesting technology and working closely with existing metadata portals, notably the Global Change Master Directory” (Parsons et al. 2006, p.7). Unfortunately, this didn’t really happen.

Many repositories were willing to adopt the DIF as a basic discovery-level metadata format, but they were frustrated that the GCMD was not set up to automatically provide or harvest metadata records (that did eventually happen but long after IPY). Much more frustrating were the vocabularies — the science keywords. And GCMD had strict keyword requirements. They strongly encouraged if not required the inclusion of at least one keyword down to the Variable level for a DIF to be accepted in the GCMD. Many of these diverse new communities already had their own terminologies or found the science keywords inappropriate, and it was unclear how to modify them. Ecology in particular already had established vocabularies of their own. Moreover, everything was captured under the broad category of “Earth Science” which did not apply for medicine, most social science, and much environmental science. The scope of the GCMD was expanding, but that was not immediately recognized. See more discussion in Parsons et al. (2011).

This discordance is to be expected. Developing shared terminology is difficult and time consuming, and it is not just a problem for GCMD staff. The real problem was that at the time, the GCMD organization and most of the community failed to realize that the existing keywords had become a critical *service* that needed to be maintained and community-driven to ensure and sustain adoption. Even geoscience data centers became frustrated. The largest geoscience data center in Europe, PANGEA, was and still is unwilling to submit their hundreds of records to the GCMD, largely because of disagreements on the required keywords and granularity. Some early keyword adopters began to abandon them or at least stop keeping up with the latest versions. A recent update at the time was not fully backwards compatible and therefore out of alignment with many local systems. So some system managers chose not to upgrade to the new version, especially because there was no web service or API for the new terms and definitions. Few would want to go on record with this, but it was a common topic in the informal conversations that go on around the formal meetings or on particular community email lists[4].

---

[3] An ongoing issue for GCMD and other catalogs is that many records only point to high-level pages and not the actual data in question.

[4] See for example, the 2006 discussion on the CF-metadata list: http://mailman.cgd.ucar.edu/pipermail/cf-metadata/2006/011072.html

The bigger issue was that data volumes and diversity were growing more rapidly than ever, and other vocabularies were emerging in climate and related disciplines. People collecting the data and designing the measurement instruments were grappling with the complex issues of their science and paid little heed to standard vocabularies. Their focus was communication with their peers not facilitating standard computerized communication. But even the data professionals who were trying to develop and implement the formal vocabularies struggled to coordinate (Carbotte et al. 2007, esp. the presentation by R. Lowry). For example, use of the Climate Forecast (CF) extensions (which include vocabularies) for the netCDF file format had become standard in much of the climate modelling and meteorological communities. CF and GCMD are not competing standards — CF metadata enable data use while GCMD metadata enable discovery — but their vocabularies were developed independently. Mapping between them became a challenge, especially since these and other vocabularies were continuing to change (Bermudez, et al. 2005).

So just when IPY investigators were generating most of their data, the initial, tentative agreement on use of GCMD Variables was fading. All the IPY data managers could hope to do was agree on Terms. Agreement at the Term level would have been useful, but few thought so at the time. It seemed superficial, obvious, and irrelevant to real science problems. That was naive. Any level of semantic agreement is a worthy achievement that can lead to greater agreement or greater understanding of context. More critically, few realized or accepted that semantics was moving to a more open and *linked* world. The connection, the link, was more important than the map or schema.

Current Practice

It has since become apparent that while keywords (typically nouns and categories) are important, their context may be even more important. What is the relationship between the keywords? We must define the verbs and predicates as well as the nouns. That is tricky. We shouldn't define a specific relationship until we have a clear context to do so — a principle of late semantic binding. Avoid making assumptions and do not specify the verb until you need to. And when you do, record it and its context (e.g., namespace, relation to projects, institutions, etc.). Strict hierarchies make this impossible even when you put the same Term under different Topics as in the GCMD.

Formal efforts in this direction began in earnest in 2005, when the late Rob Raskin from the NASA Jet Propulsion Labs, developed a semantic, linked-data representation of the GCMD keywords in the "Semantic Web for Earth and Environmental Terminology (SWEET)" (Raskin and Pan, 2005). This was an early attempt to convert the hierarchical taxonomy of the science keywords into a set of formal ontologies. Rob's team at JPL continued to evolve SWEET until Rob's untimely death in 2012. After somewhat of a hiatus, the Semantic Technologies Committee of the Earth Science Information Partners (ESIP) has taken over the maintenance and development of the ontologies and they are available through the ESIP Common Ontology Repository and GitHub under a CC0 waiver.

Today, SWEET no longer has any direct relationship to the GCMD science keywords. There are not full ontologies for all the Topics in the GCMD hierarchy, but the topics that are addressed are actively maintained by the relevant community and are increasingly adopted and interconnected to other ontologies and services. For example, the polar community has an active group working within ESIP and internationally to address semantic interoperability and metadata sharing through multiple protocols. GCMD staff participate in these activities, but the GCMD science keywords do not hold the sway they once did, even in the Antarctic community. But SWEET has also not been broadly embraced by Earth science data centers as a standard ontology or vocabulary service. Current metadata federation mechanisms, notably schema.org, work best with scalable, graph-based semantics where vocabularies are represented as classes with properties rather than in strict hierarchies (Jones et al. 2021). Yet this is not necessarily apparent to researchers and data providers, especially since these federation approaches do not have vocabularies inherently included like the DIF includes the GCMD keywords.

The GCMD keywords may have lost some influence, but they remain essential in certain communities. GCMD has modernized and still plays an important role in maintaining a broadly used and useful terminology. Clear and formal governance guidance was published in 2016 under the auspices and control of NASA's Earth Science Data and Information System (ESDIS) Standards Office. Roles are defined and review experts are listed. Importantly, there is also a reasonably active community forum and a "fast track" process for minor revisions. Periodic comprehensive reviews are conducted, and a new version 10.0 that better aligns with NOAA was released in March 2021. There is a helpful viewer and RESTful API to access the keywords, and the keywords are downloadable in RDF, JSON, XML, and CSV formats. Moreover, one can request past versions as well as the current version. Note, all the provided formats maintain the hierarchical representation rather than present a fully linked graph or ontology. For example, they have multiple identifiers for the same Term depending on which Topic it is listed under. This can confuse machines and automated workflows which expect one identifier for the same definition.

Today, it seems that the GCMD organization is more focused on a tighter mission. The directory function has been taken over by the CMR, which works hard to improve the quality and consistency of its NASA metadata and thereby provides greater services (Bugbee et al. 2021). The international presence and identity is the IDN (gcmd.nasa.gov redirects to the IDN). The IDN provides greater search and access functionality than it used to if the nodes provide sufficient metadata. With the notable exception of the Antarctic data, IDN nodes are primarily satellite remote-sensing data centers where a hierarchical, engineering-based model works well. Perhaps the flexibility and contextual ambiguity of a linked-data approach is more fraught but it is something NASA will need to increasingly explore in an open interdisciplinary world (cf. Parsons and Fox, 2013). Meanwhile, after roughly 30 years, the GCMD keywords continue to underpin the description of thousands of data sets and are integrated into hundreds of data systems. They are more flexible, inclusive, and better governed than ever. It is perhaps one of NASA's most valuable but least recognized informatics achievements.

It is futile but fun to speculate what the world of polar semantics would be like if current GCMD and related services were available 15 years ago. What could have been a great moment of early ontological alignment instead became one of ontological drift. This is *not the fault* of the GCMD or anyone. This is *an exampl*e of how influential institutions must maintain controlled

sustainability while also remaining flexible and adaptable to changing technology and community needs. Easier said than done.

Looking Forward

Many lessons can be learned from a close examination of the GCMD keyword history. Much is about balancing different interests and concerns. Many of the lessons are at a conceptual level, but these conceptual lessons can guide everyday decisions around system strategy and design. We offer some initial thoughts:

- Both institutional commitment and community engagement are essential for a healthy standard. Yet the objectives of the “institution” and the “community” may not always align.
- Sustained institutions like GCMD, i.e. NASA, and ongoing commitment are critical to data access and interoperability, but these institutions need to remain agile in their methods and even in their mission or audience. The bureaucracy necessary to maintain an institution is often in tension with the need or desire to change. This tension won’t go away, but recognizing it can help foster solutions.[5]
- Maintaining a balance of centralized control and distributed adoption/adaption is an ongoing effort. Over time, the focus of computing vacillates between the centralized and the distributed. Power dynamics are inherent in standardization. Sometimes NASA can lead, even dictate. Sometimes NASA must follow, or at least accommodate, other approaches.
- Public, transparent maintenance of a standard may be more important than its development. People need to know what’s going on and how they can have a voice.
- The incentives must be clear on why someone should adhere to a standard. There is much goodwill but unless there is tangible benefit, providers will only endure so much pain to stay current and compliant. Any level of semantic agreement should be considered a win.
- Transparent decision mechanisms with active and valued community engagement is critical even if “consensus” is not necessary.
- Services may be more important as products, and they are harder to maintain.
- Service providers need to pay attention to the scope and objectives of their service. It may have changed without you realizing it. Much like building a more just society, we need to pay attention to the impact, not just the intent, of our actions.
- It is necessary to define roles and authorities in maintaining a standard and recognize these may need to change. Correspondingly, it is important for data managers and providers to pay attention to current activity and developments. This can be overwhelming, but it is necessary. It happens at multiple levels including the details of defining particular terms and their relationships, the discussion of where they are used and to what else they relate, the incorporation into other standards, and the formal codification by organizations and governments.

[5] Friction is an inevitable and necessary element of collaboration and consensus. Without friction with the ground, a spinning wheel goes nowhere. Anna Lowenhaupt Tsing (2005) discusses this concept in compelling detail in an international case study around the preservation of tropical rainforests.

- Collectively, we need to continue to move the emphasis from hierarchy to graph and from noun to verb. Understanding and representing relationships is central to interdisciplinary interoperability.

More lessons could be learned from closer socio-technical examination of this and other journeys through data standards development, acceptance, maintenance, and adaptation. This examination must involve and interrogate the practitioners — those people who create, manage, and use the data in question — and also those who codify and formally define the standards. This need for collaboration across different perspectives is not a new insight, but the GCMD story illustrates how there can be an ill-perceived disconnect between different users and the developers of a given standard. Perhaps this can inform the GCMD organization as they continue to expand and evolve the keywords in response to user and programmatic needs. Other standards efforts, notably NASA's emergent Science Mission Directorate-wide process, would also benefit from similar critical examinations of semantics and pragmatics in addition to robust systems-engineering approaches.

Acknowledgements

This work was partially supported by NASA Grant NNM11AA01A as part of the Interagency Implementation and Advanced Concepts Team (IMPACT) program. Huge thanks to Kaylin Bugbee (NASA MSFC) for valuable comments that improved the clarity and accuracy of this essay. Opinions and errors are solely the responsibility of the authors.

Many of the ideas presented in this essay were strongly influenced by the ideas and methods of Prof. Peter Arthur Fox. Peter would have likely been a contributor to this essay, and he would have made it a clearer and more insightful if he had not tragically died this spring. We miss him.

Bibliography

Bermudez, L., Graybeal, J., Isenor, A., Lowry, R., & Wright, D. 2005. Construction of Marine Vocabularies in the Marine Metadata Interoperability Project. *Proceedings of OCEANS 2005 MTS/IEEE.* https://doi.org/10.1109/OCEANS.2005.1640159

Bugbee, K., le Roux, J., Sisco, A., Kaulfus, A,. Staton, P., Woods, C., Dixon, V., Lynnes, C .and Ramachandran, R. 2021. Improving Discovery and Use of NASA's Earth Observation Data Through Metadata Quality Assessments. *Data Science Journal*. https://doi.org/10.5334/dsj-2021-017

Carbotte, S., K. Lehnert, S. Tsuboi, W. Weinrebe, and Workshop Participants. 2007. Building a Global Network for Studies of Earth Processes: Report of the International Data Exchange Workshop. May 9-11, 2007, Kiel, Germany, 44 pp. http://www.nsf-margins.org/Datawkshp07 and (Accessed 2021-05-24)

Carlson, D. J. 2010 Why do we have a 4th IPY? In: Barr S and Lo¨decke C (eds) The History of the International Polar Years. Berlin: Springer-Verlag.

Jones, M., Richard, S., Vieglais, D., Shepherd, A., Duerr, R., Fils, D., and McGibbney, L. J..2021. Science-on-Schema.org v1.2.0 (Version 1.2). Zenodo. http://doi.org/10.5281/zenodo.4477164

Parsons, M. A., et. al. 2006. International Polar Year Data Management Workshop, 3-4 March 2006, Cambridge, UK. *Glaciological Data Series, GD-33.* https://nsidc.org/sites/nsidc.org/files/files/Glaciological_Data_33.pdf (Accessed 2021-04-29)

Parsons, M. A., Godøy, Ø., LeDrew, E., de Bruin, T. F., Danis, B., Tomlinson, S. et al. 2011. A conceptual framework for managing very diverse data for complex interdisciplinary science. *Journal of Information Science.* https://doi.org/10.1177/0165551511412705.

Parsons, M. A. and Fox, P. A. 2013. Is data publication the right metaphor? *Data Science Journal.* http://doi.org/10.2481/dsj.WDS-042

Raskin, R. G., & Pan, M. J. 2005. Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). *Computers & Geosciences*. https://doi.org/10.1016/j.cageo.2004.12.004

Russell, A. L. 2014. *Open Standards and the Digital Age* (Cambridge Studies in the Emergence of Global Enterprise). Cambridge University Press. Kindle Edition.

Tsing, A. L. 2005. *Friction: An Ethnography of Global Connection.* Princeton University Press.