Blog post: https://www.esipfed.org/esip-interviews/making-data-matter-with-ken-casey

Blog title: Making Data Matter with Ken Casey

Interviewee: Kenneth S. Casey, PhD, National Center for Environmental Information (NCEI) at National Oceanic and Atmospheric Administration (NOAA)
Interviewer: Arika Virapongse
Date of interview: July 19, 2018

Blog highlight: "In this day and age, trust is more important than it's ever been."

Arika: Could you tell me about how you got started working in the field of data and informatics, particularly in regards to Earth Science?

Ken: I started as an undergraduate at the University of Miami (UM). I was a physics and marine science major. After my freshman year (1988), I landed a job at UM's marine school--Rosenstiel School of Marine and Atmospheric Science (RSMAS). It was my start in environmental sciences, as well as data management. That was a great experience because I got to go to sea and collect observations on marine chemistry--dropping sampling bottles all the way to the bottom of the ocean and collecting freon concentrations.

I got much more heavily involved in informatics when I was a graduate student (University of Rhode Island, received PhD in 1997). Peter Cornillon, who is a long-time ESIP (Earth Science Information Partners) member, was my graduate advisor. I started working with satellite sea surface temperature datasets. About mid-way through my graduate career was the dawn of the world wide web. We were working on the development of sea surface temperatures--"climate data records" is the term that we use today, and forming a hub to serve those data out to other people.

I released my website, called the Pathfinder Cafe, in late 1993 or early 1994. It was perhaps one of the first thousand websites in the world. This was all really early days--a total wild west. It was very exciting because we were not only helping to create this dataset, but we were able to serve out these data all over the world. This experience really opened my eyes to how exciting it was to not only do your own research, but to also help other people do theirs. I was totally hooked. We had early days of web services. We enabled people through the website to launch Matlab scripts to extract SST data from what were big satellite datasets--something that you would never do today because it allowed other people to use your machine. It was very exciting--all innovation, all the time. We had the ability to serve others and help them do their work at the same time.

After completing your doctoral degree, did you have other positions before getting to NOAA (National Oceanic and Atmospheric Administration)?

I did a 2-year postdoc at NASA (National Aeronautics and Space Administration) Goddard. Then I did a 2-year rotating position at the US Naval academy where I taught satellite oceanography, while continuing my work with sea surface temperatures, climate data records, and understanding climate change on the oceans.

In 2001, I became a visiting scientist at NOAA. That was at the National Oceanographic Data Center (NODC), which is now today called NCEI - National Centers for Environmental Information. My job was to continue my work on climate data records and sea surface temperature. One of the primary missions of the center was to share data. Our job was to aggregate data from all over the world, build value-added products, and then share all of it out.

Is the mission of NCEI pretty much the same today as it was then?

Yes, although our organizational structure has changed. Four years ago there were three different data centers and they were all separate ESIP members. Then we merged into one, and became NCEI. But the mission of bringing data in and archiving it for the long term, and then serving it out to the world is a fundamental core part of our mission. From that, we built value-added products to help others do their work.

You've got a career trajectory of about 30 years, and things have changed quite a bit over that time. Can you describe how the political, socioeconomic, and technology context has changed through this time period and how its influenced your career path?

The common thread throughout all of this is that there has long been a tradition among like-minded people around collaboration and data exchange, but among others it wasn't the norm. Back in those earlier days, NODC had to run around begging for data. We would take data however people wanted to give it to us. Of course, we had our standards and aspirational goals, but more or less we were willing to do whatever we needed to do to make that data useful to others. So our archives are full of this non-standardized data. Some of it had really minimal metadata descriptions. Then we would do the extra work to feed it into our value-added products, which is what people were often after. That was a pretty constant baseline for the first 10 years or so of my career in the federal government.

The real game changer was when the federal Open Data Policy was released by the Obama Administration in 2013. Anyone that collected data using federal funding had to make their data available to everyone. It isn't fully enforced, but in federal agencies like NOAA, it shifted the paradigm. No longer did we have to go around begging for data. Now everyone was coming to us with their data. It was a complete role reversal. It put a lot of pressure on us because we had to respond with good standards, best practices, and tools. Within NOAA, every program that generates data now has the responsibility to get it into the archive, so we need to be there to help support that.

Does that mean that NOAA now has requirements for the data?

It puts us in a better position to ask for data to be prepared in a certain way, but we see ourselves as more the helping hand. We don't turn down data if it doesn't fit a certain set of requirements. We provide standards and best practices, and ask that people try to follow them. Our mentality is that we will continue to develops tools to help people to do that. We don't expect the National Marine Fisheries Service, National Ocean Service, or National Weather Service to start changing what they do, because they are collecting data to meet their own missions. The data center is there to help their data sharing process, so they can make sure that their data are reusable and they get the most return of investment of their data.

Is NOAA also motivated to do that work because the data contributes to its value-added products?

Within NOAA, we are the only long-term data archive, but we are not the only one that produces value-added products. A lot of other people across and outside the agency do that. But that is a strong motivator, because NOAA does have a heavy science element.

We bring in tons and tons of data, but not all of it gets fed into value-added products. Where we do have the resources and expertise, taking that archived data and feeding it into our value-added products helps make sure that the archived data is in really good shape for someone to use it in the future. It is sort of a cyclical feedback cycle approach when it is done really well. So in the process of using the data today, while the generator of the data is still alive, you can ask them questions about it.

Do data centers compete with each other for data?

People do have choices, particularly if they are outside of NOAA and federally funded. NSF (National Science Foundation) is a good example. Some NSF programs, like the Ocean Sciences division, suggest sending data to NCEI because we have a long standing relationship with NSF. In fact, NSF was one of the initial sponsors for the NODC. We have a long history of bringing in data from other agencies and international partners.

I don't feel like it's exactly competitive. It's more collaborative. An example is that NSF funds domain repositories, like BCO-DMO (Biological and Chemical Oceanography Data Management Office) or CCHDO (CLIVAR & Carbon Hydrographic Data Office). They are much smaller than NCEI, while having a closer relationship to a specific observing community. People from the biological and chemical oceanography community tend to send their data to BCO-DMO, which is more of a short-term archive with specialized interfaces. This is great from the NCEI perspective, because BCO-DMO then sends data to us for long-term preservation. So this is a highly effective and collaborative process with different repositories fulfilling different roles. We bring our economies of scale and provide a generalized set of services, because NCEI must serve a much broader community. In contrast, other groups have specialized biological access tools that are targeted to NSF researchers.

NCEI has base systems for finding and accessing data of all kinds, but in certain pockets where we are funded by NOAA or have specialized in-house experience, we may also provide community-specific portals. For example, NOAA has an ocean acidification program and they have funded NCEI to create a data discovery mechanism to support data from that program. Another example is the Coral Reef Conservation Program, which benefits from the data management expertise at NCEI, while providing specialized capabilities (domain knowledge) for NCEI. This partnership works really well.

We think of these types of partnerships as higher tiers of data stewardship. The base of these tiers is the fundamental long-term preservation and basic access--discipline-agnostic archiving and access mechanisms. The next tier up has more enhanced access services. Above that, there is more automated quality control, and then above that we could build some value-added products. So when there are more financial resources or a partnership in place, we can provide even more specialized data stewardship.

You mentioned that the Open Data Policy was one major milestone. Are there any other milestones that have really shifted the field?

NODC was created in 1960. In the early days, NODC was only interested in specific datasets. As people were sending their data packages other parameters might be there, and those extra parameters just disappeared. NODC would just pull out the temperature and salinity data, for example, and put those in the specific databases. There was also poor communication about this process. This created a lot of anxiety, because a researcher would contact NODC later and ask for their data back, and it would be missing key data.

So, in the 80s or 90s, NODC decided to shift its paradigm by keeping the data as the provider gave it to them. This decision played into the approach of using tiers of data stewardship, and underlies how NCEI approaches their data stewardship today. Fundamentally, if the data center were to ever lose its funding, the last thing we would turn off would be archiving the original data package that is sent by a provider. This change happened because of an NODC director's specific mandate, which aimed to reduce the anxiety and lack of trust from the community. Across the whole archiving landscape, that was a really key moment for us.

Of course, another major event was when the world wide web came along, although this happened before I was at NODC.

Where do you think Earth Science data / informatics is going in the future? What do you envision?

We are on the cusp of a whole new world--another one of these major paradigm shifts. In order to solve complex problems, you need a diverse set of players to help solve the problem. That

diversity can take a lot of different forms: gender, age, etc. It is well known that diversity helps solve problems better and faster--the “diversity bonus”.

In the data world, the same thing could be true. We have a tremendous amount of diversity, but we don’t know how to leverage it because it’s all in different formats and non-standardized--it’s hard for an individual that is tackling a problem to tap into that diversity bonus data. Along comes the cloud, which offers new opportunities. NOAA is aggressively looking at how to get all of their data into the cloud--there are various pilot projects and experiments going on, looking for new business models. For example, the Big Data Project is a partnership between NOAA and the big commercial cloud providers, looking for a win-win-win situation for the government, cloud vendors, and third party users of the data. So all of the data could go into the cloud, and the cloud can be provided for free or at a reduced cost because the cloud providers can benefit through the paid use of the data by third parties, leading toward new innovation, spurring the economy, and so forth.

So here we are sitting on this vast, diverse collection of data about oceans, atmosphere, sea floor, space weather, etc. But it’s almost impossible for anyone to really leverage this diversity. If all of these modern tools (e.g., machine learning, deep learning, natural language processing) can be brought to bear on these diverse data, then we have some hope of being able to use it (to its fullest extent). So part of that is on NOAA. You can’t just throw the data into the cloud and expect all of this to happen--we need to be very systematic about it. We need to figure out the right way to transform it so that it is machine ready.

The cloud is really where we are going, and it’s not in the distant future. It is happening now. NASA is working on it now, and other agencies and other data repositories. Some people call it the data lake or the data platform. If we can get that data diversity from all of their applications and different angles, we will see a tremendous explosion in innovation.

It is so exciting to think about a data diversity bonus, as well as how that might affect and be affected by the people diversity bonus, particularly in developing countries.

It has been a long tradition in the sciences to exchange data. We participate in communities like the ICSU (International Council for Science) World Data systems and the Committee on Earth Observing Satellites (CEOS), Group on Earth Observations, World Meteorological Organization--all international groups. They all have their own mechanisms for how to bring data in and bring it back out. So this is something that NOAA and the federal government at-large tries to do systematically. We have data.gov, as well as the GEOSS Common Infrastructure and CEOSS International Directory Network and “CWIC” system (CEOSS WGISS Integrated Catalog) for finding all of the data collections and data granules. We are trying to be a player in all of these. So as we are bringing in and aggregating all of the datasets, we are also trying to make sure that it is going out through all of these different vectors. Schema.org is another example. All of our collections are exposed here, so the search engines can come in and

peruse this information. We have been using schema.org for a number of years, but only recently is it getting more active, so it is exciting to see that.

We can't take for granted that if we build it, they will come. You can't have catalogues on the web and expect people to find it. People may have heard of the National Weather Service, but they certainly won't know that that's part of NOAA, as well as all of the other stuff that we do. So it's on us to get our data out there. Get it into the cloud and available to international partners. Because if we're not proactive about it, it's a big world, and people won't necessarily find it.

There are so many databases out there, it's almost impossible to know where to begin.

That brings in the concept of trusted repositories because in this day and age, trust is more important than it's ever been. There are cases where there has been data vandalism. So finding ways to ensure authenticity, so that people know that the data came from NOAA and NOAA is a trusted partner. Maybe 50 years ago it wasn't so critical, but today big money is being made off of environmental data. It's a huge business. Whether it's the insurance industry or Walmart looking at weather patterns so it knows when to stage sweatpants in Chicago.

When the economy becomes more intertwined with environmental data, money is involved and that creates drivers for less than ethical behavior in regards to the data. We've got to find ways to make sure that our data are authentic, so people can confirm the authenticity of the data. There is a recent example of a hurricane forecast--NOAA publishes these cones of certainty. Someone took one of those, modified it, and then tweeted it out. This was then re-tweeted out thousands of times before the National Weather Service was like, No, no, no, that's not real. It's on us to find ways to make sure that people can verify that the data come from a trusted source and not just from some guy creating trouble.

Is there a system in place for verifying data sources (trusted data sources)?

Most of what gets discussed are technical solutions. We already pay attention to making sure that the delivery of data is not corrupted in transit--that things are encrypted and sent over secure networks. More recently, people are thinking about the use of block chains, which is a way to verify the authenticity of something through provenance. How to apply that to environmental data? That is still new. I don't know if anyone has set up any pilots, but it is something that has been talked about at least a few times here at this meeting (2018 ESIP summer meeting). I wouldn't be surprised if we had to pursue something like that in the near future. But that will be exciting too.

It's all about building and maintaining trust of the users of our data. Its like human relationships. It just takes one mistake to destroy human trust, while it takes years to build and maintain trust, especially after a loss of trust. There was a session today (at the 2018 ESIP summer meeting) about the core trust seal and trusted digital repositories certification. These are the beginnings

of this sort of thing--how to verify that where you are getting your data from is a solid, reputable dealer of the data. There is a lot more to come from this angle.

[*Disclaimer: The opinions expressed in this interview are those of the author alone and do not necessarily reflect official NOAA, Department of Commerce, US government policy, or any other organizations listed. This interview also represents an "oral history" (a recollection of history), so its value is in the personal perspectives and insights of the interviewee, rather than specific dates, years, and titles for reference.*]