The need for expanding spatial data capabilities: Challenges and opportunities

Author: Dr Ines Hessler, ACENET

Introduction

Spatial data are prevalent in most research areas -- in the medical and social sciences, in earth science disciplines like oceanography, geology, and climate sciences, and in more unexpected areas such as national defence, economics, and psychology. Some of these are traditional producers and consumers of spatial data, while others are emerging and growing their use of this kind of data.

Due to the increased availability and accessibility of advanced sensors, instruments and information technologies, the pace of generating “big” and complex data has exploded in the last decade. This is significantly increasing the strain on available storage and data processing resources, and highlights the importance of developing efficient and continuous workflows and information exchange pathways.

Canada is well positioned to build a competitive advantage by developing capabilities to enhance appropriate spatial data lifecycle support, addressing aspects of the production, collection, analysis, curation and storage of such data streams. This, in combination with strengthening efforts to upskill talent in the exploitation and effective management of complex spatial data, will enable the development of new decision-making and knowledge-exchange platforms, thus providing unprecedented opportunities for science, industry, innovation and the national economy.

What makes spatial data ‘special’?

Spatial data (also known as ‘geospatial data’) are data that identify the location of a geographic feature (e.g. river, trench) and boundaries that separate different regions on Earth. The data are georeferenced, often gridded, stored as coordinates, are frequently associated with one or multiple additional dimensions (e.g. depth, elevation, time), and can be mapped. Spatial data are only complete when their three components - location, attributes, and time - are combined. Their attributes can be varied and plenty depending on the type of data (e.g. raster, vector, structured, unstructured), their generation (e.g. measurement or modelled), and the non-spatial contextual information (e.g. coordinate system, type of feature).

So what sets spatial data apart from other kinds of data?

Context is key: The definition and description of attribute information and other metadata is crucial to add context to the data. The list of attributes and other supporting metadata can be lengthy. It may

include details about the data's spatial resolution, methodology used for acquisition, information about the input data used in its generation, the time frame covered, authorship and ownership information, and many more. While context is also important for other types of data, the complexity of spatial data requires a significant amount of additional information to be available so that others can interpret the data unambiguously and/or reuse them.

Bigger and bigger spatial data: Spatial data play a significant role in the era of big data, as collection and production methods become more and more technologically advanced. Satellites and unmanned aerial vehicles (commonly known as drones) for instance, are airborne, dynamic, geographically distributed and location-aware sensors that collect data rapidly and continuously, challenging researchers and professionals to handle and process these massive data streams efficiently. The data's wide application in areas of navigation, telecommunications, agriculture and urban planning, to only name a few, contributes to the industry's tremendous growth with an increasing number of private organizations investing in and even launching satellites themselves.

Variety of sources: Besides being prevalent across numerous disciplines (e.g. agriculture, climate and ocean sciences, health sciences), it is common in the spatial research community to require data from various sources to undertake the desired analyses. The climate modelling community is a great example of researchers combining numerous data sets (e.g. weather observations, atmospheric chemistry, ocean and solar data, land surface and ice coverage information) into one model to represent the processes and interactions that drive Earth's climate as realistically as possible in order to make predictions about the past and future state of our climate.

Challenges and Opportunities

ACENET undertook consultations with a cross-section of representatives from the Atlantic Canadian spatial data community, including the Ocean Tracking Network, the Canadian Integrated Ocean Observing System's (CIOOS) Atlantic Regional Association, researchers from Dalhousie University, Saint Mary's University, the University of Prince Edward Island, and Memorial University of Newfoundland. During these consultations we explored the numerous digital research infrastructure challenges the community is facing.

Challenge 1: Spatial data gets BIG quickly

As previously indicated, spatial data are often BIG, making it challenging to extract meaningful information in a timely manner and frequently requiring the application of High-Performance Computing (HPC). These huge data volumes require reliable storage space that can handle petabyte-scale volumes of data. In addition, spatial data come in more than one flavour - structured (e.g. gridded data produced by climate circulation models) or unstructured (e.g. videos captured by drones), with both benefiting from different storage solutions. It is also advisable for storage to accommodate data in their native format to avoid unnecessary delays when attempting to convert vast amounts of data.

To provide a better understanding of the scale, the Coupled Model Intercomparison Project - the poster child of the World Climate Research Programme that is currently in its sixth iteration - is an international climate model effort that produced an estimated 15 to 30 petabytes (PBs) of binary data. Another spatial community key player, the European Union's Copernicus programme that makes Sentinel satellite data publicly accessible through a hub portal, grew to approximately 160 PBs in 2019.

Opportunity 1: Offer access to large-scale data storage

Based on our consultations, the spatial data community requires vast amounts of file and object storage (manages data as distinct units, non-hierarchical structure) to store research results and support their work. Currently, the Compute Canada Federation's (CCF) HPC clusters offer almost 100 PBs of storage, including active, project, nearline and archival. Considering the large data volumes produced in this community, the currently available storage may not sufficiently support the BIG data producers who would benefit from scalable resources.

Challenge 2: Accessibility of input data

The spatial modelling community requires access to various input datasets to force their models. For example, in order to successfully run an atmosphere-climate model for the Canadian Arctic, regional temperatures, precipitation, albedo, ice cover, wind and potentially many more data sets, depending on the complexity of the model, are required to obtain meaningful results. As these data are often obtained from various sources and can be of a significant size, the simple assemblage of relevant input data may be time consuming and requires sufficient temporary storage.

Opportunity 2: National mirrors of selected reference datasets

Providing one system that can house both subsets of large-scale reference datasets and the computational resources required for analysis would foster more efficient workflows. The selection of suitable datasets could be based on their value to a broad range of users. In addition, the new data generated from user analyses could be made available to other research groups, thereby promoting reusability and reproducibility of research outputs.

Challenge 3: Only subsets of data are required

In climate science, it is common to simulate a number of climate parameters on a global scale, covering time spans that may involve decades. These data are frequently reused in impact analyses concerned with shorter time frames, a limited set of parameters and/or are regionally focused. Instead of downloading entire datasets that can be several hundred TBs in size, it's more practical to limit access to the relevant subset of the data. Applications such as ERDDAP (Environmental Research Division's Data Access Program) - a data server that provides a simple, consistent way to download subsets of gridded and/or tabular scientific datasets in common file formats - are widely used in the community. It is not uncommon for independent research groups to each own and operate such servers. This duplication is inefficient, wastes valuable resources, restricts the greater community's access to data, and negates the data's reusability.

Opportunity 3: Add subsetting server when expanding or adding new infrastructure

Subsetting servers like ERDDAP offer a data-retrieval architecture that has demonstrated high value for accessing various types of large-scale spatial data ranging from modelled, in-situ measured, to remotely sensed, via the web. Deploying ERDDAP or similar servers on national digital infrastructures would simplify the access and retrieval of relevant data subsets, hence making the flow of data more efficient. The expertise necessary to implement and deploy such servers is widely available, both nationally (e.g. Ocean Networks Canada) and internationally (e.g. US National Oceanic and Atmospheric Administration), and has proven to make access to scientific data easier and more consistent.

Challenge 4: Beginning-to-end data lifecycle support

To reduce the amount of poorly managed data, the inefficient use of resources and the devaluation or loss of data assets, the availability of relevant infrastructure components that allow for the development of workflows, the set-up of quality check-points, and consistent procedures and policies are key aspects in guiding data through all stages of their lifecycle. No infrastructure currently exists in CCF to sufficiently support the spatial data community in managing their data throughout the lifecycle, why individual research groups and projects develop in-house measures to fill gaps, which oftentimes perish at the end of the funding cycle. This overall challenge may not be unique to the spatial data community, but certain aspects of it are, such as choosing and supporting community-relevant metadata standards and data access protocols.

Opportunity 4:

Providing researchers with a one-stop-shop for spatial data lifecycle support that includes resources to create, process, analyze, manage and make them discoverable and reusable would be the ideal solution, thus offering substantial potential to develop seamless data workflows. However, even small steps taken toward providing more comprehensive data lifecycle support, would set the right tone and pave the way for future service improvements. These could include adding a data management plan requirement to existing CCF resource allocation processes, or encouraging data producers to make their data accessible to a wider audience where appropriate.

Aside from the technical aspects and related support requirements, continuous communication and education is critical to ensure the community increases its data management literacy and is aware of the benefits of FAIR (Findable, Accessible, Interoperable, Reusable; https://www.force11.org/group/fairgroup/fairprinciples) data.

How did they tackle the above challenges in another country

The National Computational Infrastructure (NCI) in Australia is operating the country's largest HPC, cloud and storage facility that has its services and architecture geared towards satisfying the needs of the spatial data community. It brings researchers from the Australian government, academia and industry

together to address questions of global importance. The NCI operates as an unincorporated collaborative venture of a number of leading national research organisations, including The Australian National University, the Commonwealth Scientific and Industrial Research Organisation (CSIRO), the Australian Bureau of Meteorology, and Geoscience Australia, all of which make significant contributions to world-leading research efforts based on spatial data. The NCI is supporting over 5000 users from various disciplines with many of them relying on *Gadi*, their latest supercomputer.

The infrastructure and service components that make this system particularly valuable for spatial data researchers include, but are not limited to, data collection management services, provision and development of portals and cloud-based research environments, availability of decades of climate data, availability of a THREDDS server for dataset subsetting and a GeoNetwork instance to describe and find community-relevant data, as well as a state-of-the-art supercomputer with more than 100 PBs of rapid-access data storage.

The NCI is addressing common community challenges and makes valuable contributions to spatial data related research activities, which have implications beyond the scientific realm and "directly contribute to the safety and well-being of the nation". Even though the NCI system is strongly geared towards accommodating the needs of the spatial data community, it is still considered a general purpose system that is accessible to users from other disciplines, fostering interdisciplinary research activities.

Take away

Research and innovation activities, particularly in Atlantic Canada, have a strong focus on the ocean and climate sciences, which are known to be large-scale producers and users of spatial data. It is home to a number of initiatives of national and international relevance that further highlights the importance of spatial data to the region. These include the Ocean Frontier Institute, the Ocean Supercluster, and the Canadian Centre for Climate Change and Adaptation, which combined received several hundred million dollars in federal funding to drive Canada's ocean, blue economy and climate research agenda. In addition, the Atlantic provinces are home to multiple public sector organizations such as Fisheries and Oceans Canada, Defence Research and Development Canada, and Environment and Climate Change Canada, with relevant research activities that are often closely tied to universities and other research organizations.

Canadian governments already recognized the importance of the spatial sector through initiatives such as the Ocean Supercluster, the Ocean Frontier Institute and CIOOS. Recognizing the challenges and exploring the opportunities we've outlined to increase digital research infrastructure support to these communities is a logical next step in enabling Canada to expand its sphere of influence in topics of global importance including climate change and blue economy, and advance its international economic and scientific competitiveness. We consider NDRIO to be an obvious facilitator, providing the platform for collaboration, open communication and access to digital infrastructures, services and support.