Who needs Sudoku when you have a puzzle like Data Management?

Dave Spiegl, MSC, Toronto, ON, Canada; and P. Wong, R. Campbell, O. Jacobsen, and B. Fehr

1 - Introduction

The Meteorological Service of Canada (MSC) collects a vast array of irreplaceable data. Handling of data has always been a core strength of the MSC, but the evolution of technology and the ever-increasing availability and volume of new data warranted significant changes to the way MSC manages these data. With an emphasis on ensuring data quality, MSC is tackling the way it handles operational data and metadata with a multi-year, life-cycle managed data infrastructure initiative called the DMF or Data Management Framework. The DMF rationalizes the flow of monitoring data throughout the MSC infrastructure to ensure that internal and external clients as well as MSC partners have reliable and relevant access to meteorological, hydrometric and air quality data and metadata for their immediate (real-time) and future (archived) requirements.

The MSC DMF project has been underway for approximately 2 ½ years and will continue for several more years as a project, eventually transitioning into a MSC program area responsible for the long term data management functions of the organization.

The DMF is expected to be deployed as a distributed system in several cities corresponding with the major MSC operational centers within Canada.

2 - Improving Data Management

The MSC, like most other organizations in the world, are finding themselves struggling with an ever growing volume of data. These increases come as a result of many factors including advances in science and technological improvements making what was once difficult to measure or predict now operationally possible.

These increases in data volumes would normally be considered to be a positive thing as with more data comes more possibilities for newer or improved products or services. But these increases in data volumes also create additional challenges, and if not managed properly can create additional risks to an organization. The days in which an organization can count on a limited number of 'data experts' to control the access and interpretation of specific data sets or types are long gone, due to the shear volume of data and the ease of data exchange via the internet or other high-speed data networks that are now readily available. Combining broader access with more and larger data sources along with less domain knowledge can be a dangerous combination - one which if not managed properly could even lead to negative results in cases where incorrect data, or data that has not been properly interpreted, is used in the generation of critical products or services.

The need to not only store data, but to manage this ever growing asset has become

clear in the minds of MSC management and was a large driver behind the initiation of this MSC DMF project a couple of years ago. Improving the management of data, to a large degree involves the MSC doing a better job at managing, and associating, the related metadata that describes the data being managed. With more, and better metadata available, users depend less on the expert domain knowledge they once relied upon allowing MSC to more easily handle the larger data volumes it expects in the future.

3 - Data Quality

An important aspect of managing data is in understanding and making known to all who will use it, the quality attributes of the data. For some applications it is unacceptable to use anything but data of an agreed upon quality standard, but for many users it is both acceptable and in fact desirable (versus no data at all) to be able to access lesser quality data – so long as the level of quality of the data is well known and understood.

There is an opportunity within the MSC DMF project to improve, in some cases, the overall quality of specific MSC data values through either improved or more standardized quality control algorithms applied in an automated fashion. As well, the knowledge that data quality indicators will be more prevalent and available to data users will in and amongst itself result in an overall data quality improvement.

4 - Improving Data Access

Another key deliverable of the DMF project has to do with improving how users or applications access data. Although most operational MSC data today is available

'somehow' to users there are a lack of standards regarding not only how mechanically data is accessed but in many cases which data can and cannot be made available to users and when. Gaining access to data should not be dependant upon who you know within the organization or on the technical skills of an individual or group to figure out how to interact with various data stores. A limited number of standard and supported methods for gaining access to data will be delivered by the DMF project, and a common and approved data access policy will be implemented both from a policy and technical perspective that will ensure consistent and appropriate data access to users, partners, and clients.

Discovery and access methods will be developed based upon widely accepted and recognized standards. This will both reduce the amount of development effort and will improve interoperability with others outside of MSC.

A consistent answer will be provided to users independent of where within MSC they initiate their query.

5 - Efficiency

Although no specific targets have been set there are nonetheless management expectations that the DMF project will deliver to the MSC overall efficiency improvements in the longer term. These efficiency improvements are expected to come through the increased standardization in data access; through streamlining and standardizing data management functions horizontally across different data types; and by reducing duplicate processing. Operational efficiencies, once realized, can be applied to further other MSC initiatives

that today are not possible due to an existing shortage of resources.

6 - Approach

The MSC is considered a mission-critical government service and as such must continue to operate 7x24 with minimal service interruptions. As such the DMF project has taken a more cautious incremental approach to the development of its deliverables, and is facilitating in its implementation a significant parallel acceptance period before transitioning these deliverables into operation. Smaller incremental releases not only permit smaller and more frequent successes to be shown, but also permit the organization to make course corrections if necessary in the event that the organization later determines that it had incorrectly described what its requirements were. Although this approach takes longer to implement than a single 'big bang' implementation there is considerably less risk to the organization in taking this approach and it does permit users of the DMF to become more comfortable with the deliverables before being required to migrate operationally over to the DMF. It is also easier to coordinate smaller DMF deployments with other non-DMF initiatives such as the NinJo forecaster workstation deployment within MSC.

The DMF project actively tries to engage as many organizational areas of the MSC as possible so that a broad range of team experience can be assembled, increasing the probability that more user requirements will be met and increasing the probability that broader buy-on to the DMF deliverables will be achieved organizationally. Although this adds additional complexity to managing the project the benefits far outweigh any difficulties that may arise. The DMF team

has incorporated several techniques to help it operate more efficiently in a distributed team environment. The DMF team is presently spread out across 8 cities within Canada.

7 - Challenges

The challenges associated with this large DMF project undertaking are many. Some challenges are based on organization or governance, as historically different areas of MSC were each responsible for their own portion of the corporate data asset and have historically managed this asset well but to differing and varying degrees of rigour or standards.

Other challenges exist as a result of the large scope of the DMF project and the data infrastructure the project is looking to improve. The DMF project scope involves improving both the management of and access to meteorological, hydrometric, and air quality data and metadata, covering both current and forecast data (used in forecast or client services type applications) as well as historical data (used by climatology, science, etc). Many parts of MSC are potentially impacted by this initiative and thus there is a great need to 'get it right'. A higher level of rigour is needed to ensure this project is properly implemented which adds significantly to the financial resources required and puts pressures on other parts of the MSC program.

The technical challenges associated with the project, especially in areas where data access requirements are time sensitive can as well be difficult to manage. Although storage, network, and compute cost/performance ratios continue to go down year after year the demands on these services due to increasing data volumes continue to rise. The rate at which these technologies are

changing, often necessitating significant reworking of applications, is also on the rise making it difficult for organizations to keep up. There continues to be significant competition to attract or retain the necessary technical staff needed to implement large technical initiatives.

8 - Standards-based

A large part of the DMF project will be focusing on improving access to many different kinds of data, and since access is not only internal within MSC but also entails data exchange between MSC and its many partners and clients it is important that the project implement data exchange mechanisms based upon well recognized industry and/or domain standards. This will not only facilitate easier and faster deployment based upon being able to incorporate already implemented work, but will also in the longer term also increase the amount of and ease of data exchange with other clients and partners who have also chosen to adopt a standards-based approach to development.

9 - Collaboration

The data management and data access challenges that the MSC is facing are certainly not unique to the MSC. As such it is obvious that we are not the only organization that is currently wrestling with this problem and as such there must clearly be opportunities out there within the meteorological or related communities for MSC to collaborate with to the mutual benefit of all involved.

Every organization will certainly have some requirements that will be unique to how they operate, but even so it is inevitable that there will exist a sizeable core of common

functions that will be needed by all and that if planned correctly could be implemented in a coordinated and collaborative approach.

Collaboration requires both willingness and awareness. It is through venues such as this AMS conference that MSC is attempting to make others aware of the work it is currently undertaking and to learn what others are doing in this area. Although MSC can often be unaware of what some other countries outside of North America are developing we are nonetheless fully prepared where appropriate to collaborate with others on similar initiatives such as with the present case of Canada's active participation in the NinJo consortium. With advances in both communication and computer technologies over the past many years these international collaborations are becoming easier to participate in. The willingness to collaborate is there.

10 - Conclusion

Although the challenges at times appear daunting the MSC is committed to making improvements in its data management and data access through this Data Management Framework project. The expected benefits to MSC include not only properly managing these important assets, but we hope to see improvements in data quality in many areas and hope to also see efficiencies resulting from implementing more common and standardized approaches.

MSC is happy to share what it has learned over the past 2 ½ years of this project with others, and is very interested in learning from others their related experiences. MSC is open to collaborating with others who are also committed to improving their data management activities.