|  |  |
| --- | --- |
| **World Meteorological Organization**  **EXECUTIVE COUNCIL TASK TEAM ON DATA POLICY AND EMERGING ISSUES**  **First meeting** Geneva, 17 to 18 May 2016 | **EC-TT-DPEI-1/Doc. 4.1(2)** |
| Submitted by: CBS  10.V.2016 |

**AGENDA ITEM 4: Emerging Data Issues**

**AGENDA ITEM 4.1: Input from presidents of technical commissions and presidents of regional associations**

**SUMMARY OF ISSUES**

**WMO Centric Input from CBS**

At PTC 2016, the President of WMO asked the Presidents of Technical Commissions to prepare essays on the topic of “big data” to inform the work of EC-TT-DPEI and the subsequent work by CBS.

EC-TT-DPEI has considered submissions from the TCs including CBS that are centric to their domains. This document tries to present a WMO centric view of BD issues focusing on the Information Value Chain as a tool for helping to structure the WMO approach. This document also presents some ways Big Data issues could be categorized. It is hoped this will be helpful for the EC-TT-DPEI as a segway to considerations of its future work and the review to be undertaken per agenda item 4.4.

The Table of Contents is available only electronically as a Document Map[[1]](#footnote-1)\*.

**4.1(2).1 Perspective**

C:\Users\Fred.Branski\Documents\orgs\WMO\WMO TCs\WMO CBS\CBS Leadership\CBS-MG\WMO CBS MG-16 Dec15-16\CBS-MG-16_AllDocs\CBS-MG 16 Big Data Essays\BD Essay v2 pic_Page_1.tiffThe issues as discussed in this document presume the reader has also read ET-TT/DPEI INF04-1(3) or has an understanding of the meaning of the term of art “Big Data” (BD) as well as the concept of an Information Value Chain (IVC) noting that information refers to both data and products. Also, despite the use of the word “data” in the term “Big Data”, the concepts are inclusive of products as well. As was explained, “Big Information” may have been a more accurate term but we will move forward with the common usage of “Big Data” and consider it to be inclusive and defined rather broadly.

Likewise, within the scope of this consideration, the definition of stewardship is also broad. I repeat here that definition. Essentially, **stewardship** covers all the matters pursuant to the management and fitness of data and information elements as well as their relevant metadata. This includes policies, regulations, guidelines, practices and processes, and administration to ensure compliance with those items. Stewardship begins with generation of data and information and continues through its manipulation, usage, storage, sharing, archiving and evaluation. It applies to existing data and information as well as planning for new data and information or new uses of data and information. Individuals or organizations with stewardship responsibilities may have end to end roles or pick up and drop various roles at any point or points along the data and information lifecycle.

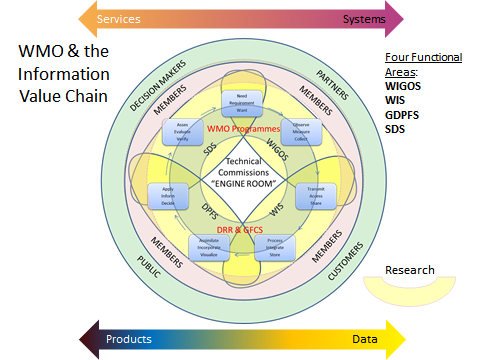
It is again valuable to consider how other IT analysts define BD. Gartner defines BD as “… **high-volume, high-velocity and high-variety information assets** that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” The bolding is mine. The 3 Vs, volume, velocity and variety are the key descriptors of BD. The second part of the Gartner definition “…demand cost-effective, innovative forms of information processing for enhanced insight and decision making” is important because it moves us down the information value chain toward knowledge and using information to enable services.

**4.1(2).2 WMO and the Information Value Chain**

In WMO, information in its various forms is arguably essential to all its Watches, Programmes, and activities; likewise with WMO Members. In order address how information and BD affects these functions and hence devise a management methodology, it would be good to have a framework around which to examine issues and impacts and then to manage them. Actually, WMO already has a very good framework with which to do so. This framework was originally envisioned and then created for the World Weather Watch (WWW). It has evolved over time and continues to do so. That framework as it exists today includes four main components, the WMO Integrated Global Observing System (WIGOS), The WMO Information System (WIS), the Global Data Processing and Forecasting (GDPFS) and the Strategy for Service Delivery (SDS). As agreed and directed by WMO Congress, these four areas serve all programmes of the WMO.

These four areas along with research can be used to examine all of the Watches, Programmes, and activities of WMO or its Members. Likewise all of the TCs manage components and activities that contribute to these five considerations; four functional areas and research which can be applied to any of the areas.

As with research, the activities of the TCs also contribute to two major crosscutting activities of WMO and its Members, Disaster Risk Reduction and the Global Framework for Climate Services. These extend across the entire information value chain.



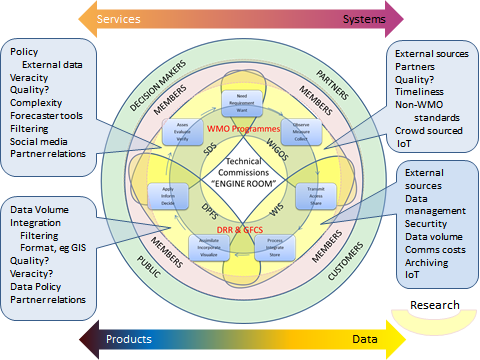
The figure above is meant to diagram these relationships. It demonstrates the various relationships discussed above. It shows the relationship of the four functional areas to the IVC. provides the crosscutting areas are shown by the concentric donuts beginning with the WMO Programmes, DRR and GFCS in the center. They are surrounded by research which supports them as well as the various blocks of the IVC. Members are in the donut surrounding these various elements and Members are in turn surrounded by their users or rather those whom they provide services to. This is meant to indicate that it is the Members whom WMO and its components and activities serve and the Members who then in turn provide services externally.

With this framework in mind then we can examine for each of the functional areas how information is acquired used, processed and passed on gaining value as it goes. We can also use this knowledge to examine and bring focus to the issues, impacts, contributions and challenges that BD brings into the equation. We can then also formulate strategies to address BD.

Without some structure to the effort, the net result would be disorganized, likely a bit chaotic, difficult to understand and almost impossible to address.

**4.1(2).3 BD so what are the issues?**

The figure below takes the WMO IVC and shows some of the BD issues connected to the functional areas. This is not meant to be all inclusive, but rather a demonstration of some issues and how they can be mapped. It should be noted that some issues may apply to more than one functional area, however its likely them it will express itself differently and may well even arise from different information sources.



The TCs and Programmes can also map where their responsibilities and contributions lie. Then it’s straightforward to link those responsibilities to the identified BD issues.

**4.1(2).4 Characterization of Big Data Issues**

Mapping of BD issues and considerations to the functional areas cross referencing them with applicable TCs, Programmes or crosscuts will enable construction of matrices that can help Members understand how the issues may affect them as well as provide a mechanism to track strategies for dealing with them.

However there is another way of characterizing BD issues that has value. It is clear that some issues or at least their impacts and challenges are primarily in the area of technology. These are often the case for those BD issues related to volume or velocity.

Other issues have clear policy considerations and these often are associated with the issue of BD variety. This could include privacy and intellectual property concerns as well as the more traditional data sharing or cost recovery issues.

Lastly, there is the characterization of BD issues that are related to how they apply to services or more specifically to users. This user connection is very important and affects how NMHSs will service user needs. Users vary in their sophistication and ability to deal with information. They likely have similar issues with volume, velocity and variety.

So, it would also be useful to characterize BD issues as to whether they are policy related, technology related or user-interface related. Some may lie in more than one of these characterizations.

**4.1(2).5 BD and the Private Sector**

There are multiple issues with the private sector and information management within the WMO and it Members. Some of these involve changes in sourcing, provision and sharing of information that has historically been done by Members. Although these are not necessarily volume or velocity issues, it may be argued that it is a variety issue. As such, this is an area that will need consideration.

However, some information today is primarily provided by the private sector. Also more and more, traditional information is also being sourced through the private sector. Lastly, often new information types and sources especially those either outside of the mainstream hydromet realm or from social media sources are often only available either through the private sector or with their assistance. In many cases, NMHSs may not have the skills, infrastructure, or agility to be capable these new BD sources.

The private sector also has a role to play in the area of processing BD in ways NMHSs can’t. And likewise the private sector has a role with the dissemination or delivery of information through avenues not often done well by NMHSs.

**4.1(2).6 Closing Considerations’**

This document is presented to stimulate thinking about how BD issues can be organized and addressed. Discussions in the meeting will feed into the follow on review and work the Task Team. There are many considerations within the documents, papers and presentations at this meeting. There are also additional sources of BD information that the entities here have within their disposal. I’m sure you have already been addressing some of these concerns. CBS has certainly been considering BD related issue as it evolves those systems it is responsible for. It will be good to undertake this review to help WMO bring a coordinated approach to dealing with BD.

This last section is simply a list of some aspects to be considered. It is not nor intended to be complete.

IDSS drives the need for not just observational data but data associated with factors and domains closely tied to users’ needs and sometimes alien to NMHSs. It also changes what and how we share our data and deliver our services.

As computational costs of models increase with more timesteps, higher resolution, etc processing demands dramatically increase. Can we increase sharing of pre-processed analysis sets? Can we move more of the modelling and post processing towards users’ devices, and let them extract whatever level of detail they need?

Traditional 'big data' analytics will become more important – to help us understand and respond to our users; to help us and them mine more information and value from each other’s data.

Although we may be focused on handling ever increasing amounts of data, we may find that the issue is not using all that data but how do we find the right data or the right amount of data that makes a significant difference. In a sense, how to we shrink BD back to small data or more appropriately the right data?

The value comes not from scale but from the ability to extract useful information and make decisions on the smallest, or is that smartest, data sets.

The value of data lies in it being fitness-for-purpose. How well does it meet new service needs? Does it support new monitoring, policy and research goals?

How will we use crowd-sourced data – fit for what purpose? Verification, phenomena or situation awareness, rapid response? Has the wind changed direction, where is the fire front, how big is the hail, how deep is the snow etc.

The concepts of mesonets taken to a new level, high density, urban environments, the Internet of Things – a whole new realm of opportunities.

How will BD be exchanged, processed, disseminated and archived?

What are the most effective dissemination infrastructures when BD just keeps getting bigger?

Where will the WIS and its dissemination and exchange infrastructure need to be in 10 years' time?

Do we need to increase high-level processing on the data before exchange.

Increased used of data portals for data discovery. Improved metadata to make datasets more discoverable within WIS or by popular search engines (e.g., Google …..)

Do we rely on existing IT tools and infrastructure or do we rely on global players in IT and the cloud – private sector (Google, Amazon, Microsoft …..)

Interoperability is key. Furthering compliance with standards and governance to enable interoperability: OGC, INSPIRE, ISO 19xxx series, NetCDF-CF, WIS, GEOSS …

Web services, APIs for data retrieval and manipulation. Processing done close-to-the-data.

Leverage Cloud Computing: (NIST) A model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

Software As A Service (SAAS): Users can use the applications of the service provider via the network. Examples: Google Apps, SalesForce

Platform As A Service (PAAS): Users can deploy their applications on the cloud infrastructure within the technical framework offered. Examples: Google App Engine,

Infrastructure As A Service (IAAS): Clients can rent computing, storage or network capacity according to their need. They control their part of the system. Examples: Amazon EC2

Cloud computing has matured and is increasingly used. The development of commercial cloud services has made significant progress. US companies lead the market with large commercial clouds (Amazon, Google, Microsoft, Oracle, IBM, Vmware, etc.)

Google and global IT players are approaching Earth Observation data holders to offer their services to develop business based on information from different datasets (EO data with other non-EO data).

With Big Data from multiple sources, it is very difficult to move the data to the processing. Instead move the analysis, processing and visualisation to the data.

Pressure is increasing to reduce the cost of running IT infrastructure.

Legacy Model: Move the data to users’ computers

New model: Move the users’ computers to the data

Web services provide a standard means of interoperating between different applications, running on a variety of platforms and/or frameworks

The Open Geospatial Consortium (OGC) have defined a number of specifications for accessing geospatial information ... maps, features, grids etc. The OGC is a non-profit international organization founded in 1994. It is based on consensus from governments, private Industry, Academia and NGOs. Some standards are fast tracked in ISO. The aim is to ensure interoperability for geospatial data and services.

Are NMHSs nimble enough to adopt and apply new approaches to handling BD before they become outdated? It’s hard for government agencies with strict procurement processes; maybe this is something that encourages us to look at private sector partnerships

Finding the right balance between focusing on data, systems, science and service – these are our 'nexus of forces', making sure we maintain an alignment and balance across them. That has driven the big revolution in forecasting – as satellite data exploded and supercomputers got bigger, our NWP/science got smarter and our services got better. The challenge in the future is too keep that balance – or at least, to keep the tension between these 4 things at the right level to push the others along and upwards together. Too much data and it is wasted time and money; if the science cannot keep up, we won't extract the value from the data; etc

Security – everything is at risk these days. As they say, all systems will be infiltrated. The challenge is to know when and where it has happened and recover quickly. Understanding what you have to protect, where the value is, what is reproducible, what is critical to services is key. It's not just about 'big' data but valued data.

Verification – data that can tell you that you are delivering value, delivering accurate and trusted forecasts, warnings, etc. You need the proof to demonstrate that the services you deliver are right, to build confidence to ensure that they are trusted and applied, in order to deliver real value to users (economic, safety etc).

**BD may the meteor of volcano that drives an NMHS extinction event or it may be the holy grail that gives us new life – it all depends on how we prepare and deal with it.**

**Our ability to observe and process are outpacing our ability to analyse and distribute.**

**This is made more complex and difficult because we are moving from a paradigm where we have had control of the data because we generated it to one where we are increasingly need and use externally sourced data.**

1. \* On a PC, in MS Word 2010 go to “**View**” and tick the “**Navigation Pane**” checkbox in the “**Show**” section. In MS Word 2007 or 2003, go to “**View**” > “**Document Map**”. On a Mac, go to “**View**” > “**Navigation Pane**” and select “**Document Map**” in the drop-down list on the left. [↑](#footnote-ref-1)