# Where does all the data come from?

## Introduction

Data without context is of little or no value. It matters where data has come from and that provenance is something the data must carry with it. Of course digital information is being generated in large quantities each day and depending on the source the information or data come a variety of issues from appropriate semantics to describe it, to integrity and completeness of the data, and all data comes with a cost for keeping it. We tend to think we can and perhaps should keep all digital research outputs to allow future reuse. Some have a vision of a future where data can be searched, found, and federated with new data…. But of course data will be lost and perhaps a more likely future will be that of digital archaeologist piecing together the historic past.

Data born digital comes in many sizes and shapes and from a vast range of sources – in our daily lives we create digital information about ourselves and our lives by shopping with credit cards – [put a stastitic about visa here] – using on line systems – Microsoft’s cloud operational support get a trillion lines of enquiries every ? [forgotten – will check notes] – creating videos [stats about youtube].

Science research creates data from observations, experimentation and simulation and of course research outputs include publications in digital form. In the following chapter we will consider a set of examples and raise some of the issues that we face with respect to these forms. We leave it for later chapters to help the reader understand what the resolution might be to these issues.

## Some Examples

Data in our lives

* Social – tweets, music, email, [ref dirty data], digital books
* Transportation
* Healthcare (images)

Observations

## One angle on Astronomy

At the present time hundreds of astronomers, computer scientists and technology engineers are designing the next generation radio telescope - the Square Kilometre Array[[1]](#footnote-1). It is anticipated that construction of the first phase of the telescope will begin in 2016 with the full telescope completed and in operation by 2022. The SKA will likely be located in either Australia or South Africa, in a desert so as to have little or no interference, but is a collaborative effort of over 50 groups in 19 countries.

The present design [1,2,3] has a combination of aperture arrays in the core and up to 3000 phased array feeds on dishes and a collecting area of approximately one square kilometre with receptors extending out to a distance of 3000km from the centre of the telescope (figure 1). It will allow a sensitivity of more than 50 times that of existing telescopes, and 10,000 times the survey speed and will provide data to answer fundamental science questions on gravitation and magnetism, galaxy formation and even the question of life on other planets. The design of the SKA is developing through studies based on the science requirements, Pathfinder telescopes that provide experience of design options, and technology capability considerations.



*Figure 1 Possible configuration of SKA receptors and artist’s impression of SKA core(from [21])*

The SKA provides a fabulous information technology challenge with a typical data rate from each dish antenna on the order of 100Gbs-1 aggregating to over 100Tbs-1 [4] and need for Exaflop computation [5] for post-processing. The infrastructure required to support the various science cases will need to range from real-time capability to transport and analyse the data at these high-data rates and the capacity to store and “publish” the data for later analysis and interpretation by the global astrophysics community. The computational systems will likely range from specifically designed FPGA-like units to exascale computing and Cloud-like data centres. The communications infrastructure will range from intra-chip and inter-chip with optical fibre to the correlator and on to a high-performance computer, to trans-oceanographic with the latter having data rates of at least 100Gbs-1 over the general network providers. The SKA will succeed or not depending on both the physical implementation of the telescope design and the software infrastructure that will enable it. The software infrastructure required to realise this information technology challenge is itself has been identified as > 2000 person year task [6] but even this may not take full account of the complexity of the task.

The analysis of the raw data will require exascale computing capability and without this the data will be of no value.

 Cornwell et al [7] estimate the computational requirements in the context of the Top500 and show them to be beyond the scale of projected performance in the timescale required (figure 2).

Alongside the challenges of the computational infrastructure are the related challenges of powering the infrastructure. This includes the power required for the core antennas (~30MW) and the remote stations ~0.5MW) and the various computational components including high-performance computing (~40MW) data transmission (~?MW) and Cloud (or equivalent) provision (~?MW). Energy provision is a major design factor in the delivery of the telescope – plans to mitigate the energy constraints include renewable energy sources (sun) providing the station power and it is easy to see that the SKA could possibly be the largest Green IT project ever to be considered.

* Earth observations
* Sea (your example)
* Smart Energy Grids

Experimentation

* Bio
* LHC
* Diamond

Simulation

* Weather
* Climate
* Bio
* Exascale

Data from other Data

Issues of tracking provenance –

# Digital libraries

## Data in and out of context

* Open data
* Data Vaults
* Semantics, ontologies etc,
* Energy requirements of data
* Data Architectures (different ways data appears – sensor streams, giant instruments, nodes of a supercomputer …)
* And how the Cloud fits in

# How Green is your data?

# Counting Joules

As indicated above the ecosystem of computational resources required to enable the SKA provides a plethora of challenges and opportunities where energy efficiency is concerned.

Moving data takes energy whether the data is moving from L2 cache on chip or within a transatlantic Data Cloud. Effective management of that data communication is a major component of any optimal energy model. There has been a great deal of research on wireless network communications and sensor networks, where devices are most often low-energy devices with battery constraints. Indeed a lot of research has been done in general on low-energy devices including computational algorithms from which we might learn. The existing communications network across the globe relies on hundreds of thousands of switches and routers and these, unlike wireless or mobile infrastructure, do not power down when idle. The network is a complex system of different technologies and it is difficult to predict the energy consumption under different circumstances so sending data 10 times as fast might use interfaces that use 100 times the amount energy or in some cases (e.g. newer optical ones) 1000 times less []. A good overview of network energy costs for realistic configurations can be found in []. There are a number of research efforts considering the issues around Green network communications including INTERNET [] and others can be seen at recent conferences [,]. Of course as energy-aware systems are developed there are still issues of what they optimise – usage or cost? Qureshi et al [] illustrate effective ways in which energy-aware data centres can optimise cost by moving computation to nearby states where electricity costs are less.

Optimizing energy usage in large-scale data centres and Clouds is almost a science in itself. McKinsey and the UpTime Insitute [] indicate that the energy used by data centres in the US is becoming a significant percentage and is likely to overtake airlines in terms of carbon emissions. The report states that the average data centre uses as much power as 25,000 households, but that estimate is probably somewhat out of date as in the last few years Cloud provisioners have built very –large scale data centres across the US []. On a somewhat smaller scale at the Oxford Supercomputing Centre (OSC) we have developed software that intelligently powers down components of the systems at times of under utilisation. The indications are that this will provide significant savings.

The McKinsey report identifies a number of issues around the effectiveness of data centres including siloed organisations and limited transparency that match very well with the findings of the HPC/NA for that community. McKinsey make the recommendation that metrics be defined that are not only measuring the facility but are linked to the applications using it and the processes integral to it. A consortium called the Green Grid is now in place with the aim to develop standards and best practice for data centres. The recommendation regarding metrics is one that we, as a community, should also take on board as we develop exascale technology.

## Conclusions

There’s a lot of it and it hard to deal with.

1. www.skatelescope.org [↑](#footnote-ref-1)