Statements / Questions / Key Issues – T.R. Connor

Within Biology data is often shared, but rarely reused. Why is this? It probably relates to several reasons

1. Accessibility of data
    a. For large biological projects with multiple data types, and multiple files, downloading/accessing data can be extremely time consuming
    b. Data searchability is often poor
    c. Data is often unlinked (eg metadata is in a table in a PDF in a paper, sequence data in the European Nucleotide Archive)
2. Infrastructure limitations
    a. Often biological IT infrastructure has been built piecemeal and so is not well designed for the task
    b. Often Biologists don't understand the systems that they procure, and don't focus on key components
    c. Biologists have, compared to other Data Intensive fields, a generally lower skill level in computing than would be ideal (most non-bioinformaticians can't use UNIX, and can't install software on a UNIX machine)
    d. Biological infrastructure is generally local (single groups often buy/own single servers for example)
3. System incompatibility / inability to share software
    a. Most papers within biology will include work done on local systems with bespoke environments; it is not easy to replicate these to reproduce the results
    b. Software is often shared as a git repository, with no incentive for the author to maintain the code, or ensure it is portable
4. For many groups there may not be an incentive to share data as widely as possible
    a. Some researchers have strong feelings of data ownership
    b. Mechanisms for citing datasets are currently still developing (therefore there is a potential impact problem)
    c. Researchers may want to exploit their data set for other research questions, which they don't want other people to be able to attempt to answer

This is compounded by the fact that in my experience some Biologists do not want to share data, or believe sharing the minimum amount of data is acceptable. There are a growing number who believe that data should be free, and available to all, and that we should try to make this possible by developing systems and approaches to support this – however, this view is not universal, and there is a cynical view that even if there are resources available, people won't use it.

What lessons can we in Biology learn from other fields in this area? How can we build cross-discipline collaborations? How can we develop best practice? And how can we build best practice across Data Intensive Science?

Is it possible to develop views about data sharing and reuse within our students? Providing them with the tools to share data, in order to bring about a generational change in the way in which data sharing is viewed?

I suspect that what we need is to begin coalescing a young, active, engaged group of Data Intensive Biologists, who can develop the systems and approaches to share data in a way that is relevant to the practice of modern data intensive biological research. I think this is an area where the RDA can help, on several levels. Firstly, to aggregate researchers across continents. Secondly to connect researchers across disciplines, to spread and develop best practice across data intensive science. Thirdly, to facilitate the creation of resources to enable data and method sharing. Fourthly, when systems/approaches become available, to help champion these, and to help researchers to work with other stakeholders such as publishers to establish standard approaches for sharing data, based on what we know about data production, data richness and data quantity today.

How can we more effectively work with stakeholders including government, publishers, funders, charities and researchers to facilitate and enable research data sharing?

While data types, and what they describe may be different, fundamentally data is all the same; we use the same technologies to store it, and the same hardware to analyse it. There are a lot of people who like to believe that their data is somehow special; however, in reality there are very large areas of commonality between fields, often without those fields realising. How can we start to get people to see the similarities, rather than the differences between data and how we analyse and store it?

With an increasing number of large infrastructure projects developing large scale research storage and compute, can we move towards a situation where data is universally available; an 'eduroam' for data storage and software? Can we develop systems that enable researchers to share, and access data wherever they are?

And can we develop universal researcher identifiers, so that research data is tagged to individual(s) as they move institutions?

The cloud and virtualisation technologies offer enormous promise to enhance the portability of software and enable more effective sharing of datasets. Imagine a situation where research is performed on a VM, and at the point of publication a researcher snapshots that VM, creates a DOI to it, and includes this in their paper. This would then enable any other researcher to literally pick up where the paper ends. On this basis, how can we make use of cloud technologies to enhance data and software reuse and sharing?

Are attempts to build on existing tools/systems (many of which were designed in the 1980's or earlier) detrimental to the long term needs for data reuse and making data available, and do we need, instead, to attempt to develop wholly new systems/environments/software tools to enable data sharing? This might

be a risky venture, but could support for a number of well targeted small scale pilot projects that are independent of current systems, all with the potential to scale up, provide us with the tools that we need going forward.

Are the systems for data sharing we have currently suitable for their task? How would Google do data sharing? And what does this tell us about where current systems are good, or poor.