

# The Virtual Observatory: what it is and where it came from

Andrew Lawrence

Institute for Astronomy, University of Edinburgh,  
Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK  
email: al@roe.ac.uk

The Virtual Observatory (VO) is driven partly by new science goals, partly by external technological changes, and partly by the flood of data coming our way. More astronomy is done on line through organised ‘science ready’ archives, and with processed data (source catalogues etc) from large survey projects. Users assume on-line availability of anything useful. The normal methods to date are to download files and analyse them at home, but very likely in the future data will be analysed *in situ* and analysis software will be as standardised as data reduction software is now. Many science goals require combining data from multiple archives (e.g. crossmatching different wavelengths), and often demand the processing of huge amounts of data, e.g. rare object searches, computing correlation functions, etc. All users want to be ‘power users’.

Two technical bottlenecks make this hard - firstly CPU-disk I/O, which has grown much slower than Moore’s law, and secondly end-user network bandwidth (the infamous ‘last mile problem’). These bottlenecks mean that jobs like searching and analysing very large databases are best performed as *services* provided by the data centre, or by collaborating data centres, joined by fat pipes. Such ‘on tap’ services will need to be standardised, professional and reliable. New internet software technologies (especially XML, SOAP, and Grid software) mean that computers can exchange data and/or share processing power in a transparent way, thus these kind of standardised services are plausible. The final driver for the VO is data expansion. Many archives now grow at TB/yr rates, several will be growing at 100TB/yr over the next few years, and soon we will have facilities like ALMA and SKA which will deliver PB/yr. Dealing with these rates is one of logistics. Modern facilities need organised data management, and methods for serving science ready data to users. From the users point of view, the problem is not the number of bits, but the number of archives. If they all speak different languages, we are sunk. The two key requirements are therefore *archive interoperability* and *well funded data centres*.

The aim of the VO is that using astronomical data from all over the world should feel just as transparent as using the Web does today. Like the Web, the VO is not a monolithic system – it is a way of life. It is an agreed set of standards, a community of interoperable data collections, and – perhaps not so obviously – a collection of interoperable software modules. If pieces of software have the same screw-threads as it were, we do not have design and build some grand system – it will grow organically. The VO is not an enormous warehouse; it is not a hierarchy, like the LHC Grid; it is not a peer-to-peer system, like Napster. Rather, it is a small set of *service centres* and large population of *end users*. Its not so different from internet shopping.

To achieve this vision we need global standards; well funded data centres; working data services; infrastructure software; VO aware client tools; and VO aware data mining services. Considerable progress has been made on standards - data access protocols, service metadata, table column semantics, software interfaces - by the formation of the *International Virtual Observatory Alliance (IVOA)*.