

The Data Challenge

October 2007

Until a couple of decades ago researchers were used to collecting their formal information to support their research work through a diet of published books, journals and original study, with informal communication providing additional ideas and information at conferences, meetings and through one to one contact.

The arrival of the digital age has changed that, and the change is escalating. Driven by Moore's Law which forecasts that there is a doubling of computer power every 18 months; by Gilder's Law which indicates that the available bandwidth in the US triples every year, and Metcalfe's Law claiming a network which grows in proportion to the square of the number of people using it – the so-called 'Fax effect' – all these have combined to create a new digital environment.

The combined effect has been spectacular. A study by Stanford University entitled 'How much information?' showed that by 2003 there were 5 exabytes of information being created each year, and that 'paper-based' communication represented a mere 0.01% of the total. Digital information held on hard-discs was the main medium. Even this 5 exabytes is but the tip of the digital iceberg when more recent social developments are taken into account – every time one drives through a congestion area; every time one passes a CCTV; every flight taken – 'digital footprints' are left everywhere as part of daily life and being recorded onto datasets.

In terms of the scholarly research community this digital phenomenon has created a new method of working. No longer is the published, refereed literature the only or even prime source for research information. Access to the raw data which underpinned the research effort undertaken by others has emerged as a critical tool. Building on the shoulders of giants has taken a new twist as these giants have taken on the form of robotic accumulations of hard data deposited in subject-based and increasingly (potentially) institutional repositories.

The main subject areas where huge subject-based data repositories currently exist include astronomy, bioinformatics, environmental sciences, weather forecasting, NASA, particle physics, medicine and health, social sciences. More are about to come on stream as global collaboratories at places such as CERN and Los Alamos begin to churn out massive amounts of new data.

Researchers undertaking projects in these areas where such datasets have become part of the firmament now can or have to alter their traditional information-seeking behaviour pattern. They can access these stores of data, often for free, and get first hand and quick access to other research team's results. In addition, in some cases they can input their own data into the data network enabling immediate feedback and response. In a few subject areas the whole communication procedure has switched from a text-based system

to online data analysis and manipulation, with print-based publications there purely as occasional statements of the progression of the Record of Science and used for fund application purposes.

Data has become the new Intel as one leading light in the industry has claimed. What does this mean for traditional publishing and librarianship? How will the traditional publication system cope with the threats and opportunities? It poses an immense challenge.

It may be understandable, but questionable nonetheless, that the scholarly publishing sector has not hitherto participated as a central player in this new digital information system. In most cases it has decided not to get involved leaving it to others to take on the challenge. It is true that the business case in support of active participation in the provision, management and particularly curation of large datasets – in an era when open access and open systems are proliferating – does not sit well with meeting financial targets set by shareholders or institutional budgets. But one must question, given the scale of the current data activity in crucial areas such as bioinformatics and physics/astronomy, whether the traditional publishing industry could become marginalised by excluding itself from dataset trends and activity.

Particularly as data and datasets are now becoming part of something even bigger. National and international organisations are making huge infrastructural commitments to creating e-Science and e-Research by investing in Grids and data networks. The NSF report on Cyberinfrastructure in the US last year “Cyberinfrastructure Vision for 21st Century Discovery”, NSF, March 2007 (<http://www.nsf.gov/pubs/2007/nsf0728/index.jsp>), and recent investigations on data and infrastructure by the OSI in the UK (<http://www.nsf.gov/pubs/2007/nsf0728/index.jsp>) are examples of policy setting initiatives to upgrade the research system to cope with the ‘data deluge’ (Dr Tony Hey, Microsoft). As one speaker at an STM annual meeting in Frankfurt commented a couple of years ago, (in astronomy) we are facing a ‘firehose of data’. Nations and government agencies are competing to ensure that their research infrastructure can cope. Equally, the focus is on making such data exchange ‘free’, with the OECD leading an initiative to ensure that there is global consistency on such free access, has added spice to the concoction.

And even further in the future, if advocates such as Sir Tim Berners-Lee (CERN and Southampton University) have their vision upheld, the semantic web would see datasets as an intrinsic part of this future intelligent Web mechanism. Though the full ramifications of the semantic web are several years away, there are parts of its weaponry already being applied in some of the recent information applications. Whilst some are critical about whether the semantic web in its current conception is achievable, there are elements of it which are in progress. In the meantime we have the Web 2.0 initiatives and the whole social networking and social collaborative schools which feed off the free data resources being put together by the scientific communities themselves.

Given this burgeoning activity, the traditional publishing community does need to address what role it can play in serving the information needs of the digitised research community.

Can the same competences developed for ensuring relevance in the Oldenburgian conception be applied to the new dataset driven information system?

Ensuring credibility, one of the four Oldenburg pillars, is an unanswered question facing the dataset providers. Each subject or discipline, using its own cultural approach combined with modern expediency is approaching the issue differently. Some have addressed standards and are making sure interoperability is achieved. Others have not. Is there some work to be achieved in setting standards and linking between datasets and other forms of media including research articles? One of the biggest challenges is to bring the small, isolated individual datasets into the public domain – it has been estimated that such small author-created datasets amount in aggregate to two to three times the total amount of data currently within the large curated datasets at discipline level. There is a challenge here.

Much depends on the quality of the metadata applied to individual records within the dataset to promote access, and so far this has been a thankless and often under-performed task of dataset creators. Is this a role the library community could perform? There are suggestions that some large research libraries are moving towards metadata provision for material which will go into their local Institutional Repository – why not perform the same task for the subject-based repositories? Certainly libraries, through their cataloguing and indexing skill set, would make them qualified to have a role in this area.

There are some serious challenges which the existing stakeholders – not just publishers and librarian - face if they are to successfully cross the valley of death and transition from a manuscript-based system to a digitally integrated and screen-based service. Whether it is the investment capabilities of the larger commercial and learned society publishers which will ensure their transition, or whether it is the smaller flexible information providers who work with specialised parts of the community to develop new targeted services mixing text with data and other media – the jury is still out on who will succeed. But a vision of the future which includes the challenge created by the dataset offerings is one which all parties need to address.