Information Services & Use 29 (2009) 103–121
DOI 10.3233/ISU-2009-0602
IOS Press

International Council for Scientific and Technical Information (ICSTI) Annual Conference – Managing Data for Science

June 9–10th, 2009, Ottawa, Canada

David J. Brown
Director, SCR Publishing Ltd.

This 2009 ICSTI annual conference was essentially about ‘Data’ – the primary, often raw, quantified, initial outputs from research projects which are held in digital form in a variety of media formats and repositories throughout the world. In some cases the raw data is stored on disc in a researcher’s personal filing system never again to see the light of day. In other cases subject-based datasets have become huge accumulations which test the ability of the science community to handle and curate them. The social and economic loss to Science in general in not having a consistent international policy to manage data and datasets is huge, and is growing as more and more scientific areas become data-centric in their approach.

In recent years it has therefore become evident that data has become a critical asset to researchers, and how access to such data is being made available and curated has become a dominant theme in scientific policy discussions. From humble beginnings as being seen as ‘supplementary information’ or as digital files which are used for one particular research experiment and then left fallow, recent years have witnessed a ‘data deluge’, a fire hose of data, which in many subject areas is challenging the refereed article as the primary source for new information. Data, along with text, does not consume itself in the ideas and innovations which are sparked off, but instead become an endless fuel for future creativity if managed effectively. And that is the challenge – to manage such an important research resource effectively for the benefit of the economy, society, Science and research in general.

As such ICSTI is to be complimented for highlighting the Data issue and bringing together some leading international experts on the various aspects of data and dataset management at its annual Summer conference held in Ottawa, Canada, from 9th to 10th June 2009. Whilst there were many librarians in the 160 or so attendees, it is unfortunate that the publishing community was notable for its absence – even though data is becoming (in some sectors has already become) a vital element of research communication activity. In ignoring data, in sanctioning it to be handled by other stakeholders, the publishing community risks losing a major stake and role in the future of a significant part of the multimedia scientific/technical publication portfolio.

The importance of the data issue, not just for the publishing industry but for the wider information community and end users at large, was addressed by a number of leading international authorities during this ICSTI conference. The local host and organiser of the conference was NRC-CISTI (National Research Council – Canada Institute for Scientific and Technical Information). Despite fears that the prevailing flu pandemic might take hold, and global economic recession may conspire to create difficulties for the organisers and reduce attendance levels, this was not the case and Ms. Pam Bjornson, conference chair and director general of NRC-CISTI, and her team, are to be congratulated for ensuring a smooth, efficient and stimulating conference.

NRC-CISTI, besides organising this successful conference, was also able to announce the launch of a major new initiative – a 'Gateway to Scientific Data' in Canada. This initiative enables the Canadian research community to have access to Canadian scientific datasets and important data repositories. It was a timely launch given the topic for this particular ICSTI conference being held at the Library and Archives of Canada in Ottawa – "Managing Data for Science".

The first keynote speaker at the conference was **Lee Dirks**, Director of Education and Scholarly Communications at Microsoft's External Research. Mr. Dirks gave a formidable tour de force of the topic 'eResearch, Semantic Computing and the Cloud'. In doing so he touched on the digital tidal wave which has affected data, and how data was moving upstream in the research process to the extent that it is being integrated into existing tools and workflow processes. Data in no longer confined to being a workbench or laboratory issue – it has become an integral part of the whole scientific research life cycle.

Leading on from this Lee Dirks saw that data management was an enabler for semantic computing and some of the services which are being derived. He also described some of the 'clouds' which were being created to harness the power of digital computing and data creation on massive scales, as well as investigating the role of software in all this. In his opinion this will increasingly become an accepted model for scientific research in future.

Referring to a recent issue of *Communications of the ACM* entitled 'Surviving the Data Deluge', Lee Dirks added that data management should not be seen solely as a technological issue. There is also a significant sociological perspective with change being necessary in the way people interact with the emerging massive datasets. The evolution of multicore parallelism and the power of the client and Cloud offering access anywhere at anytime are challenges and opportunities which society has yet to fully come to terms with. For example, what will people do with 100 times more computing power, and how will they cope with more scientific data being generated in the next five years than in the whole history of mankind? Issues such as effective data collection, data processing and archiving are challenging sociological as much as technical concepts. The speaker referred to a blog from Joe Hellerstein from UC Berkeley that said "...we're not even to the Industrial Revolution of Data yet". We are starting to see the rise of automatic data generation 'factories' such as software logs, UPC scanners, RFID, GPS transceivers, video and audio feeds, etc. What opportunities will data-centric web services and 'cloud computing' open up? At present people appear to be having to do more work on managing data and less on usage of the data.

Taking the life cycle of information supporting research – from data collection, research and analysis, through authorship, then publication and dissemination, and finally storing, archiving and preserving the results, Lee Dirks saw the need for not only much more collaboration but also the need for a new types of collaboration to make the best of the data deluge. There is also the requirement to find more effective searching and finding – discoverability – of relevant data and data sources. We are still in the early days of data use and management.

Lee Dirks claimed it is advisable to integrate data management into the wider scientific research process. There is, according to the speaker, the need to move from the traditional static summaries of work towards dealing with rich and dynamic information processes such as laboratory research which has not been the preserve of information specialists in the past. To provide reproducible research – to be able to use original or new methodology and tools and apply them to the raw data. To enable the evolution of the content within the document to become part of the web paradigm. Again, we are witnessing just the tip of the iceberg of data's potential value within Science.

In effect the emerging procedures are to enable new types of services to be conducted on and developed from the data. To bring data within every part of the Research Lifecycle. We are seeing that new organisations are recognising the possibilities for this and usurping traditional publishing roles in providing easy and effective data access. Lee Dirks went on to postulate on some of the new types of reporting on research which are driven in part by the increasing pace of Science. Reproducible research, better collaboration, offering interactive data, producing dynamic documents, creating 'mash-ups' – all are features and consequences of the new data-centric era and the stakeholders who are emerging to exploit them.

Data is now easily shareable. The Sloan Digital Sky Server (http://cas.sdss.org/dr5/en/) contains some 3 terabytes of free public data provided by 13 institutions with 500 attributes for each of the 300 million 'objects'. In effect it is a prototype virtual e-Science laboratory. In astronomy, some 930,000 distinct users access the SkyServer (compared with the 10,000 officially recognised 'professional astronomers' worldwide). Over the past six years there has been 350 million web hits. This demonstrates how open access to data can extend the reach of Science into new traditionally 'disenfranchised' areas, the amateur scientists and the general public. In the 'GalaxyZoo.org' web site there are some 27 million visual galaxy classifications, many provided by the general public. 100,000 people participate in open access blogs. This is one example of a data-driven 'citizen science' made possible by access (in this instance) to data which is free.

Nevertheless, the speaker pointed out that there are some concerns with data sharing. There are issues of data integration and interoperability, particularly of datasets in different but related subject domains. There are technical issues of consistency in annotations. Agreement on formats also needs to be made. Provenance also has to be resolved, as does the issue of privacy and security. There are some services which have either challenged some of these constraints or swept them aside. Swivel has arisen as a cross data searching platform. IBM 'Many Eyes' has also pushed on the frontiers, as has Google's 'Gapminder' and Metaweb's 'Freebase'. CSA's 'Illustra' has also adopted a novel approach.

As a result some of the old commercial concepts are being challenged. Some enlightened organisations are providing software using open source, enabling a whole range of applications (APIs) to be developed around them. IBM and Redhat were cited as examples of open source. In the library world we are seeing institutional repositories becoming sources for free access to grey literature, supplementary information, theses and, of course, raw data, as well as the traditional research article in a pre-published form. It has led to there being various flavours of repository software being introduced to help local institutions capture and make available new datasets and new information formats. Added to which there are enhanced operability standards emerging, though still having some way to go before full interoperability is achieved.

Some specific examples of the new data resource repositories which are emerging include the US government's 'data.gov'. This is an expression of the Obama drive for transparency and 'openness', leading to trawling around for data not only from within the US Executive but also from other US federal agencies. A data catalogue has been developed which includes access to the data in two ways – through the data catalogue and using various other external access tools. Then there is the Department of Energy (OSTI) led project to create a more textual 'WorldWideScience.org' building of the more national 'science.gov' in the USA. This was commented upon in a subsequent paper given by Richard Boulderstone from the British Library who currently chairs the WWS international consortium.

The semantic web, a logical home for much of the data-centric activity, will not happen on its own accord. It has to be enabled. It has to grow from the grass roots. In this respect there is a distinction to be drawn between semantic-based technology and the semantic web. There are examples of semantic-based technology such as machine learning, neural networks, ontologies and inference software. The semantic web itself is just one of many tools at our disposal. Their combination to leverage on the collective intelligence of the community can be seen when experts in the field openly share their knowledge and experience through such applications as Connotea, BMC's Faculty of 1,000 (even though these are mainly manually-based at present) and specialised social networking sites using Web 2.0 approaches.

Another major step forward can be found in Lee Dirk's other main theme in his presentation. He cited examples of how 'cloud computing' is being brought in to analyse, process and visualise data. It offers the creation of a world where all data is linked and interconnected through machine-interpretable information. All the data will be stored, processed and analysed 'in the cloud'. The cloud is a linked, network of mainframes around the globe which can share the burden of massive data analyses. The advantage of using 'cloud computing' is that there is no need to build up a single big infrastructure – it already exists – it just needs to be brought together in a uniform way. A number of organisations are involved in creating the cloud, organisations such as Amazon, HP, Google, etc.

Lee Dirks suggested there are three types of cloud computing – there is the utility computing (which is the infrastructure). This is offered on a pay-as-you-go basis enabling large dataset users to process and analyse data as and when they need to. They do not need to create a physical infrastructure to process such data. Some of these are based on Amazon's S3 and EC2 offerings. Then, secondly, there are platforms which provide a service making use of the infrastructure (such as Google's AppEngine and Salesforce's force.com). And finally there are end user applications which make use of the infrastructure and platforms – such as Google, Amazon, Twitter, Flickr and other major Web 2.0 applications. Lee Dirks felt that it would soon be the case that the research sector would follow the commercial sector in adopting the cloud and semantic concepts. Even within the commercial sector the cloud landscape is still evolving. It is being facilitated by such tools as Flickr, SmugMug for photos, YouTube, SciVee for video, Slideshare for presentations, Google Docs for word processing. These are perhaps the tip of an iceberg, with the full scope of such social networking to appear over the next few years.

In recognition of the importance of data analysis within society the National Science Foundation has created the NSF DataNet infrastructure. Two major awards have already been made, others will be due soon. DataNet involves data preservation in a whole new way – changing the culture specifically around preservation. The two main beneficiaries of the DataNet awards so far include John Hopkins and the University of New Mexico as lead organisations, supported by many of the country's top research centres. They are five year projects with the possibility of an extension for a further five years. It is an attempt to change the culture of work undertaken in a data rich environment.

A particular example of how 'cloud computing' is developing in the area of preservation and provenance can be seen in the DuraCloud project. D-Space and Fedora have recently merged to create the DuraSpace organisation. DuraCloud is the preservation aspects of this in the cloud. The infrastructure was primarily funded by the Mellon Foundation. DuraCloud then goes to Amazon, Google, HP, Yahoo and Microsoft – all cloud service providers – to get the necessary computing power. The key thing about the cloud computing is that it is not one single source of computing power – it is the use of multiple repositories and resources. Throughout it is essential to separate the technical issues form the business issues, and to ensure that quality remains paramount in any service provision given.

John Milibanks (director of Science Commons) has commented on these cyber infoproviders – he suggests it is more than just computers – it is not just the machines which make the cyberinfrastructure work. Software alone is not the answer either. It is people, policy, legal frameworks. Connecting with people and getting them involved in the process is crucial. Having the right legal and policy regimes is essential. We need to fill the gap which is emerging in the management of the data resources and its applications or else the opportunities for the efficient use and reuse of data will pass us by. We need a change in the sociology of scientific research and its use of information.

After the coffee break **Fran Berman**, Director of the San Diego Supercomputing Centre in California, discussed 'Mobilising the Deluge of Data'. Fran Berman is also co-chair of the important Blue Ribbon Task Force on Sustainable Digital Preservation and Access which has a staff of over 300 looking at aspects of cyberinfrastructure and sustainable data preservation in the US. Fran Berman also mentioned that she is on the point of moving to Rensselaer Polytechnic Institute in New York.

Science has become a key agenda item for the new US federal administration. According to President Barak Obama "Science is more essential for our prosperity, our security, our health, our environment and our quality of life than it has ever been before".

There are many opportunities arising from the 'data deluge' which has been going on in the background but with this comes many challenges in creating useful information services. Dr. Berman used as an example the application of data to understand the impact of large-scale earthquakes in the southern San Andreas Fault area in California. They use computer models to predict seismic activity. They create grids or blocks using a super computer to collate the relevant data for each block. The aim is collect evidence to enable new building codes to be developed and to manage effective responses to a major earthquake.

There are many significant social questions which can be answered using large datasets. Some of these issues were commented upon in an IDC White Paper entitled "The Expanding Digital Network: A Forecast of Worldwide Information Growth through 2010" (March 2007). There are some datasets which are more important than others – those with political relevance, for example, have high social priority, as do environmental datasets dealing with issues such as the ozone layer and global warming in general. The fact that high resolution data and images from NASA have been lost is a major concern. It raises issues of stewardship of highly valued data – for example, what data should be saved, how does one save it, and how does one maintain an important data collection? The consequences of moving into a data-rich society need to be analysed from a variety of socio/economic/politico aspects.

Fran Berman gave some statistics on the amount of digital data currently being churned out. There are 50,000 proteins in the Protein Data Bank Structures, equivalent to 25 terabytes. Stored data from the ENZO cosmological simulations amounts to 500 terabytes. Google Earth has some 71 terabytes. Even the Library of Congress, seen primarily as a repository of printed publications, manages nearly 300 terabytes of digital data, 230 of which were 'born digital'. In 2007 the amount of digital information available in the world finally exceeded the amount of available storage (some 264 exabytes). The 'date resourcing' gap is widening in each successive year. Some of this data is more important to access than others and some need preserving more than others.

With regard to what data should be saved, there are three levels to consider. Saving data of interest to society at large (such as census data, presidential emails, etc.) are crucial. Then there is the issue of saving data of relevance to the research community (such as protein data bank, national virtual observatory, etc.), and finally, saving data of relevance to the individual (such as personal photos, medical records, etc.). Each poses its own unique policy challenges.

How does one save it? – There are a whole variety of preservation technologies, which are themselves in a state of transition. Besides the new technologies there are new data standards, best practices, new protocols, etc. For example, with regard to best practices there are issues concerning replication, metadata development and planning ahead. There is a need to plan ahead for database selection – which one's to keep, for portal creation. Also, where the dataset would fit in. Whether it can be used for such applications as data mining, mash-ups and other semantic web purposes. How the dataset will be curated and archived is also an issue.

According to Fran Berman a good cyberinfrastructure must be reliable. A file may have a shelf-life of 1 year; a tape for 5 years; a system for 5 years; and an archive for 50–100 years. Each such format needs to be taken into account. There has to be sustainable economic models which support digital data access and preservation. There is a need for ongoing funding and governance to make sure there is a sustainable future. The funding support can be through a subscription service, through advertising, through an institutional subsidy or on a pay-as-you go basis. This means that we may need to be creative, to open our minds on how the commercial underpinnings can be put in place. Creating partnerships is one way.

Ms. Berman provided a list of Best Practices required for Digital Preservation. These include:

- Replication of multiple copies some off-site.
- Linking the data with quality metadata to facilitate easy access.
- Planning ahead to allow for transition of the data to new and emerging media.
- Include costs of creating the data in the information infrastructure.
- Building resource security.
- Being aware of appropriate regulations, policies and penalties.

An attempt to tackle some of these problems is being done through a Blue Ribbon Task Force in the USA. Economic sustainability is being seen as a targeted issue. For more information on this activity see the Blue Ribbon Task Force web site at http://www.brtf.sdsc.edu. Some of the key figures on this task force include Fran Berman, Chris Geer, Cliff Lynch, Chris Rusbridge, Paul Ayris from the UK (UCL) and Lee Dirks from Microsoft.

Overall there is a need for a Master Plan for Data. The master plan should include the framework for what is valuable in data, and cover issues such as stewardship and costs involved, assessing data needs, mapping the data to these needs, and a system for measuring the success of data provision. Fran Berman indicated that there are three levels of stewardship of data – a 'gold' standard, a 'silver' standard and a 'brass' standard for the different levels of data stewardship. Having said that, there is no need to wait for the Master Plan to get things started. As new research paradigms emerge so these create an even greater need for data mobilisation and an effective ultimate Data Master Plan.

Richard Boulderstone then stood up to the plate. Richard Boulderstone is director of e-Strategy and Information Systems at the British Library. He currently leads the British Library's efforts to create a large-scale digital object management system that will become the primary repository for the UK's legal deposit collection of electronic resources. He described how data was being managed within the British Library. His basic message was that the British Library is adopting a customised and focused approach to data management.

Richard Boulderstone gave a historical overview of the British Library (BL) to set the scene, to reflect on the extent of the challenges, and to provide the context for the British Library's activities in data management. The BL was formed in 1972 and moved into its new headquarters in St. Pancras over a decade ago, St. Pancras being one of the two main operational centres (the other at Boston Spa in Yorkshire in the north of England). The BL has a wide-ranging cultural mission as the nation's central research library. One of the BL's more recent aims is to provide a support service for Science beyond that of just providing an international document delivery service from its extensive collection of journals and serials subscriptions. The means for achieving this is through an understanding of what the end user, the researcher, really wants – getting to know the users. Part of this programme includes an irregular

'Talk Science' open forum at the British Library during which a leading expert gives their opinion on a particular scientific research theme (with John Millibanks from Science Commons, referred to by Lee Dirks in his presentation, being the next scheduled expert). Users' involvement in SecondLife is also being tested. Undertaking questionnaire surveys, commissioning case studies, and investigating behavioural research focusing on specific research disciplines are some of the other mechanisms used.

One research discipline which has been given particular attention by the British Library is biomedicine. In January 2007 the British Library took on a significant role in the development of UK PubMed-Central (UKPMC). Supported by a consortium of funding agencies, led by The Wellcome Foundation, UKPMC involves an operational partnership between the British Library, the European Bioinformatics Institute (EBI) and MIMAS from University of Manchester. Building on the database of some 1.4 million articles included in the PubMed Central service, the intention is to make it easy for UK based biomedical researchers, funded by UK funding agencies, to include their final research reports within the UKPMC repository. More recently, four additional work packages have been agreed on which brings in features required by the UK and European biomedical markets, less required in the heavily centralised funding situation in the United States. A wider range of information services will be made available through UKPMC including some advanced work on text mining. Meanwhile the demand being made on UKPMC from users is currently not as much as had been anticipated, and whilst hypertext links can be a useful service according to Dr. David Lipman, the original sponsor of the PubMed Central concept, extensive use of hypertext linking can overwhelm the end user. In such innovative areas it may pay to be cautious.

The British Library also involved in more generic scientific domains. This includes Richard Boulderstone's chairmanship of the WorldWideScience.org project, and the inclusion of the British Library's electronic table of contents within the service to provide wide appeal and access to published research articles. The addition of China as a further major national information supplier to the WWS service demonstrates the geographical spread now being achieved. A formal signing ceremony to include China as a participating member in WWS was included later in the conference.

Another exciting project which has considerable input from end users is Research Information Centre (RIC). This is a partnership between the BL and Microsoft. It is an attempt to bring together through one single interface access to information which covers all aspects of the research cycle or workflow. Initially it is aimed at a biomedical community, particularly working with National Institutes of Health Research (NIHR) in the UK to iron out bugs and to help develop new user-demanded features.

Another new project which the BL has initiated in response to market demand is the digital library system. Using mirror sites currently in place in London (St. Pancras), Boston Spa and Aberystwyth (national library of Wales) a support system for digital archives of digital information is being put in place. In due course Edinburgh may also join the network, as may Cambridge and Oxford universities.

On the crucial topic of 'datasets', Richard Boulderstone volunteered that the BL does not have a master plan in this area yet. Studies and user feedback is being sought to determine how users currently find such information and where it located. The BL does not believe it should take on the function of archiving, storing and preserving such material on its own. However it is working within the new consortium set up which allocates digital object identifier (DOI) numbers to datasets (see Jan Brase's presentation later in the conference proceedings). The question still remains what additional services researchers want without duplicating what other centres are doing. Issues such as metadata harvesting, linking, creating catalogue access, ingest and preservation are still to be resolved.

The final speaker in the first morning of the conference was **Dr. Christopher (Chris) Greer**, director of the National Coordination Office at the Federal Networking and Information Technology R&D

Programmes in Virginia, USA. He is also co-chair of the interagency working group on digital data of the NSTC committee on Science. The title of his presentation was 'Science in five dimensions: Digital data for Cyberscholarship'. It focused on the respective roles of government, academia and industry in providing digital preservation and access to scientific data.

In setting the scene Dr. Greer referred to the exploding digital universe, as already commented on by Dr. Fran Berman. In the opinion of John Gantz, chief research officer at IDC, the digital universe will grow from 160 exabytes in 2006 to 1,600 exabytes in 2011 – a ten-fold increase in five years. This is dramatic! Already, in 2007, the amount of digital information surpassed the available storage capacity (see earlier), and in five years time information will be twice the available storage capacity. However, unstructured information accounts for 90% of the digital universe. According to IDC, 'preservation intense' information will grow nine fold over the next five years.

This 'expanding digital universe' is creating new ways of conducting scientific research. One example cited by Dr. Greer was the data generated by the Large Synoptic Survey Telescope (LSST) which will produce – in one night – 30 terabytes of data. As the speaker pointed out, 'the widespread availability of digital content creates opportunities for new forms of research and scholarship – this is called 'cyber-scholarship'. Even the smallest bit of information can produce a giant leap forward in innovation and creativity. This involves repurposing digital data as much as new primary data creation.

The Interagency Working Group, of which Dr. Greer is part, is a unit of the US Committee on Science. This in turn reports to the National Science and Technology Council which itself is a subset of the Office of Science and Technology Policy responsible to the Executive Office of the President at the White House. As we heard from Fran Berman, President Barak Obama has pledged his support for Science as part of his administration's agenda.

The National Science and Technology Council has been charged with providing a national framework for facilitating access to information and data, and implementing the agreed recommendations for R&D, education, science, technology and engineering. The remit covers a wide range of digital formats (raw data, audio, video, algorithms, etc.). The main conclusions are included in the report 'Harnessing the Power of Digital Data for Science and Society', published in January 2009.

The guiding principles were that communities of practitioners need to be brought into the process, and that not all data needs to be preserved. However, not yet resolved is who makes the decision on what is kept and who is brought in.

The report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council recommended that a unit should be charged with digital preservation and access at a federal level. This unit should also be a single point of contact for global collaboration. The Federal Agency policy needs to provide guidelines for those who create the data. Each federally-funded research project requires a project plan to be produced for the preservation of the data created during each project. Such issues as IPR (intellectual property rights) and formats for interoperability should be included.

According to Chris Greer librarians have a critical role to play in the preservation of the digital asset. There are a number of ways this can be done. Implementation phases can be introduced. Templates are being developed for agency fund allocation. Examples of data management plans – best practices – are also being produced. General procedures set at federal agency level are also being made available at state, local and even research level. First principles which are being adopted should enable archivists of digital data to operate more effectively in future.

The first presentation in the afternoon came from **Dr. James (Jim) Mullins**, professor and Dean of Libraries at Purdue University. Purdue University has a sound scientific record and a strong international orientation. Founded 1869 by a gift from John Purdue it has established premier programmes in engineering, agriculture, hospitality and tourism, business, computer science and communications. It has over 40,000 students in 2008/2009 with the third largest international student enrollment in US – over 6,000 for that year. In Autumn 2009 12% of first year undergraduate class will be international students and 60% of the graduate school will be from overseas.

Professor Mullins took up the cudgel of defining what role there is for librarians in a data-centric world. The title of his presentation was ‘New Roles for Librarians: the Application of Library Science to Scientific/Technical Research’. This was in essence a case study of Purdue University’s approach to dealing with the cyberinformation environment.

However, Professor Mullins first set the context for data curation, and by reference to the National Science Board’s publication “Long-lived digital data collections: Enabling research and education in the 21st century”, sought to establish that data curation was important to data authors, data managers and data scientists alike.

As far as the role of the library is concerned he described HUBzero™. This is the technology which supports the creation of dynamic web sites that connects a community in specific areas of scientific research and education. These are nanoHUBs. nanoHUB projects are essentially Web 2.0 services for scientists in targeted research areas. Purdue university libraries have played a leading part in the development of these hubs.

Purdue university has a high rating within US research centres – it appears sixth in Google’s top 200 universities and this is largely due to the impact and appeal which its development of nanoHUBS is having. Purdue still retains the IPR of these eight or nine research areas where the nanoHUBS have been created but the intention is make them open source as soon as possible.

It was determined early on that research scientists needed the libraries to help resolve their information problems, particularly in the current era of rapid technology change in information delivery. Librarians needed to trigger the scientists’ interests however. This involvement is essential to enable librarians to have a role in the future information world through developing such useful services as nanoHUBS.

In addition, Professor Mullins pointed out that in many research areas scientists are working in teams, sometimes remote from each other. In 2008 a sponsored research project was undertaken by D2C2 or the Distributed Data Curation Center (with grants of some $618,000 in first year alone mainly from the National Science Foundation). A Working Group was set up which asked scientists in certain areas how they made use of the information they wanted and acquired. This developed information solutions in areas such as pharmacology, cancer research, etc. The speaker gave a number of examples of the studies undertaken and indicated the range of research grants received by D2C2 in their work of curating and investigating dataset applications.

There are therefore a number of opportunities open to librarians to become more proactive in charting a new future for research information creation, dissemination and curation, and these depend on close interaction with the Faculty to understand their real and emerging needs for information.

The next speaker was **Dr. Liz Lyon**, director of UKOLN, a library research unit based at the University of Bath in the UK. She currently also serves as a member of the NSF Advisory Committee for Cyberinfrastructure. She focused on the growing requirement for coordinated data curation, digital preservation and data management within the scholarly knowledge cycle. In particular she felt there is a need to adopt the concept of ‘Team Science’ to reflect the multi-partite involvement in sustaining future information services. In fact the title of her paper was ‘Libraries and Team Science’.

Her premise was that multi-author collaboration, across several institutions, produced better science than research done by a singleton. Science is a social activity, increasingly evident as such services as Tweet, Blog, Share, Mash and Tag proliferate and become part of the social milieu. Trusted parties working collaboratively on research projects become crucial, and new tools are required to assist in such multisite collaboration. An example of a 'team science' approach can be seen in the area of e-Crystals research – where chemists, computer scientists and informatics specialists all act in unison in a project on crystal research. It involves several departments at Southampton University, DCC and UKOLN.

Besides her role as director of UKOLN which undertakes basic research on fundamental library and informatics issues, Dr. Lyon is also a director of the Digital Conservation Consortium (DCC). In this capacity Dr. Lyon described some of the research projects which are being funded as part of the curation life cycle model of research. These include:

(1) *Leadership*, involving senior managers (vice chancellors and above) in universities providing awareness and advocacy in support of new information services.

(2) *Policy*:

(a) There are a number of national and international reports from DCC and JISC which provide comparative information on curation activities.

(b) At a local level, Neil Beagrie produced a study on 'Digital Preservation Policies' in October 2008. This provided some additional high level pointers for mapping UK institutional strategies.

(3) *Planning*:

There is a draft 'Data Management Plan Content Checklist' produced for the Digital Curation Centre by teams at Edinburgh and Glasgow universities. This is still going through a consultation phase and as such remains work-in-progress. It covers such issues as Legal Rights and Ethical Issues, as well as Resourcing, Deposit and Long-term Preservation.

(4) *Audits*:

This looks at the amount of legacy data available within each institution. DCC is promoting the idea that audits be undertaken of this data. A 'Data Audit Framework' was developed in areas such as geosciences, archaeology and the humanities during the period May–July 2008.

(5) *Engagement*:

This is where greater proactivity with the faculty is being recommended, ensuring that past R&D projects are brought into the mix, case studies are included, etc. They have been done in areas such atmospheric data, neuro-imagery, tele-health and architecture.

(6) *Repositories*:

(a) This includes 'Policy making for Research Data in Repositories – A Guide' sponsored by JISC and compiled by the Data Information Specialists Committee – UK in May 2009. This focuses on Metadata provision and Ingest, Access and Preservation of data.

(b) Also, 'Preservation Planning for Crystallographic Data' which is a JISC e-Crystals Federation report has been compiled. It looks at quality assurance issues in particular.

(7) *Sustainability*:

Metadata provision is a key element to ensure sustainability and access to data records. Another report from JISC, DCC and the e-Crystals programme looked at 'Preservation Metadata for Crystallography Data'. This tackled issues such as creating a PREMIS data dictionary, and delving into OAIS. OAIS involves joint collaboration between partners supporting the exchange of information on open access material.

(8) *Access and reuse*:

‘Community Criterion for Interoperability’ is being investigated, which includes bringing together standards and procedures such as CIF, InChi, DOI, the various embargo and rights policies, etc. The aim being to instil trust through understanding of what is possible and legal and achieving mutual benefits.

(9) *Training and skills*:

Of critical importance is the provision of effective training and courses specifically around the topic of ‘curation’ and the curation life cycle model for research and information needs. Several such courses are currently being aimed at end users, librarians and repositories. An online version of these courses is also in preparation.

(10) *Community building*:

There is a diverse physical group involved in data curation, each having and operating within their own ‘silos’ of information and skill sets. The fragmentation issue is being addressed at the various International Digital Curation conferences, one of which will be held in London in early December 2009.

Looking towards the future Liz Lyon addressed the issue of skills required for the process of data management. The challenges were to some extent covered in a report produced by Key Perspectives for JISC in July 2008 on the topic of ‘Skills, Roles and Career Structure of Data Scientists and Curators’. From this it appears that there are not enough trained and professional data librarians available. The report distinguished between a number of different facets of data management as a career – from data managers, data creators, data librarians to data scientists. Different skills requirements were matched against each of these target groups. From a *CILIP Update* in June 2008 it appears that there are ‘only 5 dedicated data curators’ in the UK and these are largely ‘accidental’. This leads to the conclusion that there should be a greater commitment given to teaching data management skills within library/information schools. The issue is as much about people as it is about the technical features of data curation.

For more information about the digital curation centre’s activities, see: http://www.dcc.ac.uk/.

The next speaker was **Jan Brase** from the German National Library of Science and Technology (TIB) in Hannover, Germany. This centre has focused on issues relating to data management since 2004, with support coming from the German Research Foundation (DFG). In 2005 TIB became a registration agency for DOI (digital object identifiers). It has the responsibility for allocating DOI’s for datasets under an agreement reached with the International DOI Foundation (IDF). Jan Brase is coordinator for the DOI Registration Agency for research datasets at TIB. During the past 4–5 years the TIB and its trusted partner organisations have registered some 600,000 datasets with DOIs.

Jan Brase described the progression of information from the data, through the analysis to information and from there to knowledge and wisdom. However the data which supports and is behind this edifice is often lost in the passage of time. What is needed are stronger, more robust and sustainable data centres, and the ability to find descriptions about the datasets so that access to the data can be made through library catalogues and other search mechanisms. It requires a process of establishing persistent identifiers, something DOI offers. The datasets can then become citeable which gives them a new level of respectability and acknowledgement.

The citeable dataset should have the same look and feel as the article to which it refers. If this can be achieved some of the benefits include:

- High visibility of the dataset.
- The results can be verified.
- The original data can be reused (which can become a reward mechanism for the scientist). It confers reputation on the authors.
- It avoids wasteful duplication of research effort.
- It provides motivation and stimulus for new research (to update earlier results).

The Brussels Declaration in 2007 by the STM publishers made it clear that publishers accept the principle that access to such datasets should be free of charge. However there is a cost involved in data creation management within the scientists/researchers community itself (which can vary from €5,000 to 5 million), at data centres (approximately 1% of the data creation costs, or from €50 to 500) and the inclusion of the data metadata within the library catalogue (which was estimated at €0.1–1). However these costs do not include costs of storage and curation (the middle layer, currently managed by data centres).

The problem is that in many research disciplines the infrastructure for data management is missing. There is also inadequate funding available for dataset curation. There are policies missing for consistent data submission, to enforce submission. The carrot is to include the data within the citeability process – so that the individual researcher or team can be rewarded through recognition of the work done to create the new data.

For the library, the record of the dataset would look just like that of any other artefact – it would appear similar to a book including a DOI persistent identifier. The DOI would direct the user to the appropriate data-centre where the dataset resides and is curated.

A number of experiments are being undertaken by TIB to demonstrate the versatility of the dataset identification process. The first is with Elsevier and its Earth Science journals on ScienceDirect. A link is included within the journal article which takes the user directly to the appropriate data centre where the supplementary data is to be found. The other experiment is with Thieme Verlag where the project is to provide links to the source data from chemical research experiments. The chemical research data included as supplements to articles in two chemical journals will be registered with DOI's by TIB. The datasets will be stored at Fachinformationszentrum Karlsruhe (FIZ Karlsruhe). And access will be free of charge.

However, much attention within ICSTI itself is currently being focused on the joint initiative by a number of national centres to create a worldwide registration agency known as Datacite. It had its beginnings in a Memorandum of Understanding signed at the previous ICSTI (Winter) meeting held on March 2nd 2009 in Paris at which representatives from a number of national libraries in Europe agreed to march along the same DOI road in providing persistent identifiers to datasets. The participants include INIST (France), Technical University in Delft (The Netherlands), Denmark, TIB (Germany), the British Library (UK) and ETH Zurich (Switzerland). This international collaboration is now been opened up beyond Europe with interest in joining being expressed by CISTI (in Canada) amongst others. In future DOI data registration will no longer just be a TIB and German affair (though the path-breaking work done by TIB was much appreciated) – henceforth it has international reach through the Datacite consortium.

Dr. Ellsworth LeDrew, Professor in the Department of Geography and Environmental from the University of Waterloo in Canada, then described a particular example of a dataset challenge. It involved analysis of long data series and its preservation and making it available, particularly related to activities associated with the International Polar Year (IPY). There are some 200 projects with 50,000 research scientists involved. There are a whole series of honeycombed projects with the need to provide integration.

The premise adopted by the speaker was “Hypotheses come and go, but data remains the same”. However there are large gaps in the data required to support the International Polar Year work. Data which was collected before 1882 is largely no longer available. During the 1930s and early 1940s most of the data was lost in world conflicts. Much data from the period 1957–1958 has also gone. To ensure that the most efficient use is being made of what data is available an International Data Management Centre has been set up, and the Canadian Data Management Committee and Policy unit feeds into this. The aim is to move towards an ‘information commons’ – a portal which will allow access to all data even before it is formally published.

The question of ‘whose data is it?’ was posed – informatics is one of the pillars of Science and as such it is critical to ensure that data be managed effectively. As far as IPY is concerned it is a question not only of managing the data on scientific information, but also social information, health and physical issues. Another critical aspect is getting metadata registered – to enforce this future funding will no longer be provided to an individual researcher unless adequate metadata is also provided with the research results and datasets The data needs to be acknowledged. The data goes into a Polar Data Catalogue for Canadian Scientists in ArcticNet.

There is also an issue of privacy over the data records which has to be addressed. For example, in Canada the Inuit want to know what is happening to the environment within which they live. However, this can cause some problems – for example, making data on the movement of caribou (through collars attached to the animals) results in easy kills for the hunters which is not the aim of the exercise, and as such this data – and others like it – need to be protected. But in general the intention is to make the metadata generally available. Other ongoing issues include making sure the data is interoperable, that equitable access is achieved, that the digital divide should not be made wider, and the role of data centres be considered.

Another major data centre with a huge data management challenge is CERN in Switzerland. According to **Dr. Tim Smith** from CERN’s IT Department, Big Science being achieved through Big Collaboration which is leading to Big Repository Services. The data which is being managed at CERN is primarily derived from the Large Hedron Centre Ring which has a 27 kilometer circumference, running 100 meters underground not only in Switzerland but also in neighbouring France. It is known as a ‘Cathedral of Science’. Some 15 petabytes of data each year will be created and will be transmitted around the globe. This data will be delivered at three levels – Tier One is where all 15 petabytes will be collected with CERN itself. There are then tiers 1 through to 10, ten centres worldwide each taking 10% of the data created. The third level is where 140 computer centres worldwide are involved in specific data analysis. Not only of the data itself, but also the calibration, e-logs and simulation. Raw data is reduced to reusable data which the libraries are able to work on. This creates new skills and traditions for librarianship. However scientists need to recognise that such engagement by the library community is necessary. Why should libraries be involved? – a new skill set may be required.

There are a whole series of information services which are available to the scientists – including postprints, theses, conference proceedings, reusable data, raw data, photos, videos, animation, etc. Scientists want all these sources connected in some relevant way. They want a ‘one stop shop’ for their information access. In this respect Libraries have a significant role to play. They can build the architecture and develop the structure for managing such multimedia.

One specific example of the importance of libraries is in the development of Institutional Repositories (IRs). CERN has long been a leader in the development of IRs, to the extent of developing open source software (Invenio) to create INSPIRES, a next generation institutional repository for high energy physicists. It builds on the original SPIRES database, well recognised and accepted by the international physics community. It has been so successful that it has attracted interest from outside CERN, with some 30,000 unique visitors accessing the system – which far exceeds the known number of particle physicists.

Keywords are distributed the same way and are curated. The CERN digital library becomes a data handling warehouse which consists of 8 file servers, 2 Web servers and 4 streaming media servers. It allows for Web 2.0 collaboration and commentary. It creates a personalised digital library offering citation analyses. Full text can be mined. Tag Clouds are created from the keywords.

All this means that scientific communication in high energy physics is very rapid. Traditionally the scholarly publishing system took 5–6 months for scientific information to be refereed and 'published'. With the accent now being given to preprint services, to 'open' IR services, the improvement in speed has been profound. In essence there is data everywhere. There is published data (such as in tables and figures); there is supporting data also in tables and figures; there is highly complex data (multi-dimensional); there are data analysis objects; there are data analysis programmes; and there is raw data including access and analysis tools.

As was explained, INSPIRE was formed from the merger of the data which was part of SPIRES (SLAC) database using Invenio technology. INSPIRE therefore represents the whole corpus of particle physics information. Besides the data and text stored internally within INSPIRE it provides for external connection to ArXiv, NASA, the Astronomical Data System (ADS), etc. At this point users start asking for more and more services. They are now able, for example, to track all the records from a given author (such as in NAMES) broken down by type of media citation.

This was brought about through intense engagement with the research scientists. The aim is to provide them with what they really, really want and to develop new information services accordingly.

The next speaker was **Paula Hurtubise** from Carleton University in Canada. She is also Section Chief at Statistics Canada and the manager of the Ontario Council of University Libraries' (OCUL) project. The focus of her presentation was a project called Ontario Data Documentation, Extraction Service and Infrastructure Initiative (http://search2.odesi.ca). Or, as was claimed, odesi is a voyage in data discovery. Odesi is the product of a partnership between university libraries, business and government in Ontario. It is a tool designed to enhance and stimulate social scientific inquiry by enabling researchers to share data and documentation, leading to new and unanticipated discoveries. Odesi is an intuitive web-based statistical data exploration and extraction tool hosted on Ontario's Scholars Portal. It is part of 'Ontario's e-Science in Action' programme. Through distributed access, it also addresses the significant disparity in the availability of data resources across Ontario's academic institutions.

The aim of odesi was to create an intuitive data portal for researchers, teachers and students that would 'inspire, develop and support research excellence'. Traditionally such data resided in silos with separate access and search procedures. Odesi was also intended to remove the need for specialists in this process.

The sources of data imported into the system were various. Primarily they came from Statistics Canada, from CORA Nestar (which included government information) and the ICPSR Archive. A number of factors then came together in 2007 which included the OCUL and Ontario Buys (an agency which looks for efficiency in purchasing material) to offer enhancements to the education and research processes within Ontario. The Scholars Portal became the delivery mechanism, DDL standards were included as was appropriate commercial software. The resulting odesi system was seen as a mechanism for encouraging community input and participation in the overall service aims. It recognised the existence of many boutique data centres.

Eleven universities became participating partners. Software was included which enables tables, charts, and graphs to be created (Nesstar). The Scholars Portals was set up to enable diverse browsing and search features to be included. It is Ontario's attempt to mobilise public data for the good of its citizens.

Professor Peter Fox was the first speaker on the next day. His position is at the 'Tetherless World Research Constellation Chair, Climate Variability and Solar-Terrestial Physics at Rensselaer Polytechnic Institute'. He addressed some of the definitions which set out the structure of the new data-intensive informatics industry. The title of his presentation was "X-informatics, data science and the full life cycle of data information and knowledge in Earth and Space sciences". He personally is active in the area of semantic e-Science, as well as virtual observatories and organisations. Also in something he referred to as 'X-information'. These semantic e-Science issues were particularly focused on earth and space sciences.

Professor Fox gave an overview of the whole data spectrum. He commented on how data is obtained by multiple means (from instruments and models), using various protocols, in differing vocabularies, using assumptions and often inconsistent metadata. There is no easy and smooth path to creating data interoperability.

It is all made even more difficult in what the speaker referred to as 'semantic heterogeneity', the production of large scale data, the complex data types, the many legacy systems, etc. But despite these challenges the need to come to grips with Data Science is important as Data becomes the fourth pillar on which Science is built. At present there is a gap between Science and the underlying infrastructure and technology which is available to support it.

The speaker introduced the term 'cyberinfrastructure' as the new research environment that supports advanced data acquisition, data storage, data management, data mining, data visualisation and other computing and information processing services over the Internet. He also described 'informatics' as "the structure, behaviour and interactions of both natural and artificial systems that store, process and communicate data and information" (Wikipedia). Modern informatics dates back to 1957/1958 when the two disciplines of information science and computing split up and it is only now that informatics is beginning to get a thorough grounding as the two traditional disciplines coalesce again. In essence there has been a move from IT, through cyberinfrastructure, to cyber informatics, and then core informatics to science informatics.

Professor Fox then gave a description of some of the data issues facing specific date-rich disciplines. He described how atmospheric sciences was coping with informatics (see http://www.vsto.org). Also how climate searching was being impacted by the new informatics approach. Solar physics was also a beneficiary. He felt that the arrival of aspects of the semantic web was beginning to provide benefits to the research process, not least in the area of enabling a much broader audience to access and gain advantage from the application of informatics in these areas.

There is the need for those active in the cyberinfrastructure and informatics fields to work with social scientists and scientists much more closely than has been done in the past. But here is the difficulty. There are very few trained in this specialised area of informatics in science. Courses are required in such areas as semantic e-Science, on Data Science, on X-Informatics and on interdisciplinary digital rights. For effective development in this area the participants need to be part of the science community. With regard to semantic web methodology there are too many unknowns in taking technology forward and applying it. However it should be achieved through an understanding of user needs, and through cooperation between the library and the computer sciences.

The next speaker was **Dr. Katy Börner**. Dr. Börner is the Victor H. Yngve Associate Professor of Information Science in the School of Library and Information Science at Indiana University. The presentation was an illustration how visualisation techniques could help science policy makers and funders understand the impact which science funding was having on the economy. In particular she described how computational scientometrics could inform Science Policy.

Dr. Börner started her presentation with a general overview of scientometric workflow, with computational scientometrics being defined as the study of science by scientific means. The output from such an approach are maps of knowledge domains. The source of data to produce such pictures of how scientific subjects relate one to another is from the ISI database of 1 million research papers in 2002. The results of looking at the clustering or interrelationship between these subjects areas were that 670 clusters of journals were identified.

From this Dr. Börner was able to look at the funding patterns within the US Department of Energy (USDoE) – it highlighted the clustering of funding in the physics and chemistry areas. This may be obvious, but the visualisation of the relationship shows the priority areas as well as the spillover areas for DoE funding. The same was shown for the National Science Foundation funding with computer technology, chemistry/physics and geosciences taking precedence. The speaker also showed that fund allocation pattern of the NIH which in this case stressed the biosciences, notably biochemistry and some chemistry, as the key points.

Armed with such factual evidence, Dr. Börner's team undertook 34 detailed interviews, each lasting some 40 min on average, with 34 policy makers. Four of these were outside the USA. A number of questions were asked of these policy makers including 'what databases do they use?', 'are they aware of the structure of science?', 'what monitoring do they do?', 'what impact measures are used?', etc. From these interviews three main themes of what policy makers wanted emerged. These were:

- Information on science structure and its dynamics.
- Impact issues.
- Feedback cycles.

A practical example of the work which has emerged from this Science of Science approach is the development of a Cyberinfrastructure portal. Included within this science of science portal are some 23 million records, with strong visualisation support tools. It also includes a network workbench tool. The scholarly database had, as of May 2009, some 170 registered users. The workbench tool enables the co-authorship of Medline authors, for example, to be visualised. The same was done for patents citation network. Some of this work comes from the activities being undertaken by the MESUR programme at Los Alamos, being pioneered by Bollen and van de Sompel et al.

A particular example of how such scientometric technique quantifies the output from an industry can be seen from the work of the Council for Chemical Research in the US. Using such scientometric tools, the CRC showed how the US economy generated $8 billion in taxes, of which $1 billion was spent by federal agencies on R&D for chemical research. Added to this was the $5 billion spent by the chemical industry itself to produce a $10 billion operating income. This helped to produce a $40 billion growth in the US gross national product and an increase in employment of 600,000. It also resulted in $8 billion in taxes, and so the cycle continues. The timeline for such chemical research impact on commercialization is twenty years.

Dr. Börner was therefore able to demonstrate that by using a suite of cyberinfrastructure applications and through discussing how the sharing, analysis and visualisation of published knowledge this can all help to inform science policy at a national level.

Brian McMahon from the International Union of Crystallography (IuCr) in Chester, England, then talked about the 'interactive journal'. There are a number of different forms which interactive journals can take. There are instances of online community peer review. There are also examples of post-publication commentary and responses. It can also include dynamic links out to other databases, and also multimedia presentations. Other examples include three-dimensional data visualisation. There are also interactive aspects included in medical imagery. The various permutations are endless and daily becoming more innovative.

However, it does appear that journal publishers are conservative in introducing such innovative procedures. There is an exception in the area of crystallography. There is an e-Crystallographic dataset which has Java applets embedded. The Journal Mol includes visualisation of crystals. They are even able to display and amend publications in PDF through the use of the Adobe Reader. The question is how useful are these tools? Much depends of the quality of the reviewers and the review process itself. It also depends on document size. There may be additional software requirements. There is also the issue of archiving which needs to be resolved effectively. In general it is important for these various services to meet the needs of scientific enquiry and study.

The International Union of Crystallography (IuCr) and the National library of Medicine (NLM) are collaborating on the project, and other trusted partners include OSA, Wiley, Elsevier and Springer. The initial results were reported at a workshop in the Winter of 2010. See http://www.iucr.org.

Dr. Johannes (Jan) Velterop is chief executive of a semantic-based company – Knewco – but at the conference he was representing a new consortium which is being created. This is the Concept Web Alliance. It has arisen because there is such great fragmentation in policies being adopted within the overall information industry and there is the need for a forum to bring these disparate parts together.

At the moment we are facing many problems in dealing with data. There is the concept of 'datarrhoea', too much information. Then there is the fact that a lot of data is being lost over time, and finally much additional valuable data is hidden in the Deep Web. Essentially, the industry is suffering from 'information overload' or as Jan Velterop refers to it as H2Know – there are oceans of information. So what has to be done?

Jan Velterop spoke to the topic 'Beyond Open Access: Maximizing the Use of Scientific Knowledge'. There is the requirement to navigate among large amounts of information. In so doing the stimulus is also to create new data. There is the need to make funds available so that information can be reused.

Dr. Velterop went on to discuss the concept of 'Triples'. In semantic terms we think in triples – a concept (a) has links (b) to another concept (c). There is the need to reduce things to simple concepts. So concepts 1 and 3 may be databases. The link, concept 2, is what we do with it, and is part of the knowledge database.

There are a number of strands to what we should be doing with data in future, according to the speaker:

(1) We need to learn how to navigate across large amounts of information, and not just to ingest it.
(2) We need to make additional funds available for data harmonisation, preservation, curation and reducing data loss – not just to produce more and more new data.
(3) Identify every single bit of knowledge for easy and unambiguous find-ability.

This implies we need consensus on how we name things that overlap – an electronic concept identifier (CID). We need to identify every bit of knowledge and to things that relate to it. It all requires a new international platform to support global scientific data cooperation. This is where the Concept Web Alliance comes in. The stated aims of the Concept Web Alliance are to: "To enable an open collaborative environment to jointly address the challenges associated with high volume scholarly and professional data production, storage, interoperability and analyses for knowledge discovery". There are some 20 or so groups already affiliated. See http://conceptweblog.wordpress.com.

The final keynote speaker was **Paul Uhlir**. He is Director of the Board on Research Data and Information at the National Research Council in Washington, DC. His presentation focused on open access and as such differed in approach and content from most of the previous speakers. His main aim was to convince the audience that open access was the way of the future, and that open knowledge environments (OKEs) were essential for the dissemination of public knowledge.

He claimed there is a revolution taking place in scientific information. "Revolution and Evolution in Scientific Information" was the title of his presentation. There are major policy issues arising from the economic, social, technical and institutional issues which are arising. In essence there is huge difference between the print and the global digital network paradigms. There are a number of principles which should be followed, some inherited from the old world and some from new technologies. All lead towards open access according to Paul Uhlir.

These principles are:

(1) There should be maximisation of public good from public fund investment.
(2) Monopolies should be avoided.
(3) Marginal zero costs should be taken advantage of.
(4) Freedom of enquiry and collaboration on research should be emphasised.
(5) Stuff should be made accessible to new discovery tools.
(6) The characteristics of the record should be maintained, notably the quality.

The first five of these principles are part of the open access agenda. However, if such free access is not achieved there are some restrictions which emerge to scientific discourse:

- Research costs are increased.
- Opportunity costs are lost.
- Barriers to innovation are maintained.
- There is a widening in the digital divide, between OECD and third world countries.

'Openness' should be the default, though there may be a few exceptions such as those involving privacy and legitimate confidentiality, etc. Open access and unrestricted reuse of research data and information produced from public funding online is in most cases far superior to proprietary and restricted dissemination, which maximizes value for the disseminating organisations rather than for the content producer and user community.

There is however different business models in achieving openness, between green, gold and grey routes. Even within the green route there are different channels. Paul Uhlir's ambition was to see the emergence of effective Open Knowledge Environments (OKEs) at universities. This can include thematic repositories organised around OA (gold) journals or databases. There should be commonly accepted tools for licensing of content, such as being developed through the Creative Commons, GNU, etc.

Examples of where there has been a community response to the adoption of openness as the basic framework can be found in the genomics area where common international standards are adopted. However, the barriers to achieving full openness in scholarly communication and OKE's in universities should not be underestimated. The speaker gave examples of some of these barriers. These included:

- Implementation and acceptance of new policy and institutional frameworks may be slow given the conservative nature of many academic institutions.
- There are inadequate incentives for individuals, communities, institutions, and governments to take part in the new open business paradigm.
- Long-term financial sustainability of different information models and OKEs are not yet proven.
- It may be necessary to overcome pressures within universities to commercialise their OKE.
- Legal and social issues have to be considered – such as IPR, privacy and national security.

Nevertheless, Paul Uhlir did highlight some significant advances being made to create an open agenda. These included infrastructural developments such as the important advances being made in the open-source software movement (e.g., Linux), distributed grid computing or e-Science (e.g., LHC@Home) and open data centres and archives (e.g., GenBank, EROS Data Center), federated open data networks (e.g., World Data Center Service, Global Biodiversity Information Facility, NASA DAACs). In addition, these contextual changes are also spawning developments at the informatics level including the many new open access journals (approximately 4,000 OA journals), open institutional repositories (IRs) for scholarly works, open repositories for publications in a specific subject area (e.g., the physics arXiv, CogPrints, PubMed Central in US and UK) and free university curricula and lectures online (e.g., the MIT OpenCourseWare). There are also emerging disciplines or applications commons, peer production of information, and integrated open knowledge environments (e.g., INSPIRE, IPY commons, odesi, virtual observatories, wiki encyclopedias).

All in all, a groundswell of support for the open access movement which is as much a revolution as an evolution in the way specialised information is being created and disseminated in the new Millennium.

In assessing the open access movement, Paul Uhlir was able to complete the circle on a conference dedicated to exploring the role of Data in Science. For data management to become a serious part of the information system, to capitalise on the huge global investments made and being made in scientific data creation, the need for free and open access to such data is essential. It enables the data to be mined in all sorts of ways, many currently unimaginable. However such burgeoning growth in a scientific research asset in an open access environment puts pressure on traditional business models employed to provide sustainable information services. There are still many barriers to climb over.

In tackling the Data Lifecycle from a number of perspectives the 2009 Annual ICSTI Conference has done much to remove the scales from the eyes of those data stakeholders present at the meeting. It showed just how important the issue of Data Management in all its aspects has become for those involved in promoting the aims of an efficient scientific information system for the future. There was general consensus at the end of the meeting that the ICSTI Conference was again a great success.