# **GRL**2020 **Europe**

## Paving the Way for a Collaborative Global Research Environment

# Outcomes of GRL2020 Europe

Participants at GRL2020 Europe, the 2nd Global Research Libraries Workshop

27th -28th March 2008 Tirrenia, Italy
www.grl2020.net

# Summary

This report details the main outcomes of GRL2020 Europe, the 2nd Global Research Library 2020 (GRL2020) Workshop, 27-28 March 2008, Pisa, Italy. An overview on the main conclusions of the 1st Workshop (October 2007, Washington, USA) is provided. The Report goes on to define the main aims of the 2nd Workshop, exploring the trends and visions for research libraries with a case study from Washington University, as well as a section dedicated to the importance of global collaboration and data preservation.

Five domain-specific case studies are presented with particular reference to best practices fostered by the communities, and a use case is defined. A dedicated section identifying ten key drivers for GRL2020 is included. The final section explains the main features of GRL2020 and the bottom-up, down-approach adopted with the aim of making GRL2020 a reality.



| GRL2020 Europe Programme Committee | | |
|---|---|---|
| Peter Buneman | University of Edinburgh | UK |
| Donatella Castelli | CNR-ISTI | IT |
| Lee Dirks | Microsoft Research | US |
| Fabrizio Gagliardi | Microsoft Research | US |
| Jessie Hey | University of Southampton | UK |
| Yannis Ioannidis | University of Athens | GR |
| Lizabeth Wilson | University of Washington | US |

## Acknowledgements

GRL2020 wishes to thank Microsoft Research for their support to this publication, to the GRL2020 Programme Committee and to the authors for their valuable insight and contributing work in achieving this informative report.

## Disclaimer

The contents of the post-event report have been created by representatives from members of the GRL2020 Programme Committee and with the editorial support from Trust-IT Services Ltd. Although content of this report is public, we request that any copying or paraphrasing from the writings or works indicated in the present document should clearly acknowledge the report reference source, its authors, organisation names and dates as indicated here. All rights are retained by the original authors and submitters.

# Table of Contents

# 1. A Future Vision

## Mark Parsons, US National Snow & Ice Data Centre

## An Imaginary Journey through the Arctic Observing Network Portal

The global digital library of the future will contribute greatly to improving the working environment of scientists. Exactly how this could affect the daily life of researchers is examined here.

Suppose one day you read an interesting article speculating on the contribution of processes in submarine canyons to the global carbon cycle and decide to explore arctic datasets. Entering the AON data portal, you first encounter icons for terrestrial, atmospheric, oceanic, and human dimensions that contain a summary of data holdings under each discipline. You then have the option to browse datasets by discipline or by theme. Using data exploration tools, you search for canyon processes and determine what relevant meteorological, geophysical, and oceanic datasets are archived, and their availability in space and time.

Although you do not realize it, the information accessed comes from four different data centres in two different countries. For observations that are interesting but unfamiliar, you find links to descriptions of the instrumentation, the methods, and the data processing steps. You also find links to browse images of the datasets and, after inspection of these, you decide flow levels of the X River bear closer investigation, as the X River appears to be associated with the Y Canyon, and both the oceanic and terrestrial environments are well instrumented.

Plotting the time series using the online data display tools, you observe that three years ago, in June, the gauges reported an abrupt drop in water level after a gradual rise through late spring. The screen also shows an icon that looks like the silhouette of a parka. Curious, you click on the icon, and a text box pops up describing a large ice dam that gave way about the time of the abrupt water level drop, with a notation from the Inuit hunter who reported the event. Now you open the relational database interface in the AON portal and frame a query requesting turbidity measurements within 100 km of the mouth of the X River during the timeframe of the ice dam collapse. Within seconds, you have links to data streams and generate another series of plots. These show an increase in turbidity within the Y Canyon two days after the ice dam collapsed. You suspect that you have identified a flow event carrying sediment into the deep Arctic Ocean. Wondering how general these events are, you search for abrupt drops in tide gauge measurements coupled to local increases turbidity measurements for other arctic river systems and find three more candidate events.

It is almost the end of the day and you download your time-series plots and email them to your colleagues twelve time zones away for their review tomorrow. You save your AON session using the password protection you have installed so that you can access the data again tomorrow without having to redo the data searches. Before wrapping up, you post a request to the event detection service, providing the combined tide gauge turbidity criteria as the trigger. Finally, you post a request to the observation scheduling list, starting the process to request time on the docked autonomous underwater vehicle near the mouth of the X River to be triggered

on detection of an event. It has been a productive day.

## 2.Research Libraries - A Changing Landscape

Digital technologies are changing the way science and research are carried out, enabling new methods, experiments and scale. Fast communication networks are connecting scientists, universities and research labs allowing them to work together on new, often global, challenges. Researchers are developing collaboration technologies that go far beyond the capabilities of the Web, in order to use remote computing resources, integrate, federate and analyse information from many disparate, distributed, data resources, together with access to and control of remote experimental equipment. The capability to access, move, manipulate and mine data is the central requirement of these new collaborative science applications.

Many of the societal challenges ahead, such as climate change, world-wide health threats, and international economic issues require a wealth of knowledge and global endeavours. Global partnerships would not only reinforce Europe as a leader in the advancement of digital infrastructures but would also inextricably link research, education and innovation to create a Triangle of Knowledge

Beyond this, some innovative libraries are playing a pioneering role in the movement towards digital libraries (DL) and schools of library science are increasingly schools of information or information science. As Lucy Nowell, National Science Foundation (NSF, USA) points out, 'that fact that the documents of interest are digital, in no way diminishes the need for trusted repositories for them, nor the need for knowledgeable guides, as the information "superhighway" can take users through very bad neighbourhoods'. For centuries, librarians have provided guidance about how to locate material of interest, as well as an understanding of which documents can be deemed trustworthy and which should be challenged. Librarians therefore have a key role to play in the current and future landscape, particularly in terms of domain-specific expertise and acquiring new skill sets in the face of the digital age.

# 3. A Vision for Global Research Libraries – Thought Leadership

The changing landscape of research, discovery and libraries is bound up with a shift towards increasingly multi-institutional and cross-sector approaches, which in turn brings new opportunities and roadblocks for further development. A select group of experts have spearheaded efforts to develop GRL2020 - a vision a for a global research library that is open, interoperable and that fosters collaboration on research challenges.

These experts formed the Programme Committee of GRL2020 US, Washington, US, 30 September – 2 October; and GRL2020 Europe, Pisa, Italy, 27-28 March, 2008, which have helped connect innovators from across the globe. At these workshops, major actors from university libraries, developers, domain specialists, funding agencies, and national and international initiatives have evaluated the state-of-the-art, case and use studies, with the aim of defining the main drivers for GRL2020 and the challenges it would help tackle.

*Public value and user needs are central to the GRL2020 mission. The driving principles behind GRL2020 are developing effective collaborative research environments that tackle issues such as resource curation and sharing and that are founded on an interoperable and sustainable e-Infrastructure. Tapping into the expertise of the global DL community, discipline specialists and funding agencies is key to making this vision a reality.*

Donatella Castelli, CNR-ISTI

## 3.1. Building Blocks – GRL2020 US

Redmond, Washington, USA provided the backdrop for the 1st GRL2020 Workshop, which was held 30 September-2 October 2007 and hosted by the University of Washington. The Workshop brought together experts from Australia, Canada, China, Europe, India, Japan and the USA to define the global research library of the future and discuss its main roles and functions. Additionally, experts evaluated the main implications for users and providers, and the underlying global regulatory and policy framework.

Tony Hey, Corporate Vice President for Technical Computing at Microsoft, underscored that research libraries should foster new approaches aimed at bringing value-add to research and discovery on a global basis. By pursuing this goal, research libraries will be better placed to play a central role in an increasingly multi-institutional and cross-sector environment. Thought leadership from Tony Hey helped focus discussions and build a consensus on core value propositions for GRL2020.

| GRL2020 US Programme Committee | | |
|---|---|---|
| Lee Dirks | Microsoft Research | US |
| Ann Ferguson | University of Washington Libraries | US |
| Tony Hey | Microsoft Research | US |
| Neil Rambo | University of Washington Libraries | US |
| Lizabeth Wilson | University of Washington | US |

## GRL2020 US - Main Outcomes

The workshop helped define core values in terms of long-term data curation and preservation.

### Core Values for the GRL2020

*Innovation and knowledge creation rely on the sustained availability of information and serve as drivers for discovery. The creation of public value is central to the mission of global research libraries. Selection, sharing, and sustainability are long-standing components of library missions, and remain so as library assets shift from paper to digital formats. Long-term curation of content is critical, and requires focused effort in the development of systems and standards to support them in the long digital future ahead.*

A series of current impediments need to be addressed if the vision of GRL2020 into reality.

| Current Impediments |
|---|
| Funding for research and learning is fragmented and suffers from steep disparities globally |
| Intellectual Property and copyright constraints increase friction in the information supply chain |
| Complexity of the stakeholder environment impairs interoperability and information flow |
| Cross-sector tensions and proprietary perspectives dilute resources and leadership |
| Infrastructure deficiencies, especially in developing countries, limit the scope and effectiveness of recognized solutions |
| Economic and technological sustainability are problems at all levels |
| Skills appropriate to the 21st century information world are scarce in a 20th century workforce |
| Disparate political, economic and cultural environments often impede collaboration. |

## 3.2. GRL2020 Europe– Focus and Main Aims

Co-hosted by Microsoft Research and the Italian National Research Council - Institute of Information Science and Technologies (CNR-ISTI), GRL2020 Europe built on recommendations and conclusions from the 1st workshop, GRL2020 US, seeking also to identify the grand challenges and objectives for Global Research Libraries of the future.

GRL Europe was designed to transcend political boundaries and demand new collaborative user-friendly frameworks that are sustainable in the long-term. Lee Dirks (Microsoft Research) defined the main aims of the Workshop focusing on:

- **Linking infrastructure:** established and new infrastructures, including skilled personnel systems, standards, and policy frameworks must be tightly-linked and easily interoperable through efforts such as leveraging community protocols and the development of suites of standards.
- **Ensuring sustainability:** effective international funding partnerships must be established through close interaction with key funding agencies, foundations, and at relevant private sector organisations and corporations in a concerted effort to establish common goals and work toward eventual sustainability and independence.
- **Changing people:** a core aspect of re-architecting research libraries will involve advancing professional training in the field. Skills required include archiving, data curation, library science, information technology, and computer science. At present, experts are often hired from other fields. A key goal is to define how to foster a new mentality and approach in the allied information professions.
- **Building community:** joint outreach and advocacy efforts around explaining the GRL2020 vision for how the global research library should evolve and how it must continue to scale to support of global information problems are required.
- **Working together:** ultimately, the workshop seeks to identify a project, forum, research programs, or a proof-of-concept idea with the aim of launching this effort in a concrete manner.

## GRL2020 Europe - Main Outcomes

The two-day workshop was attended by 45 of the world's most highly regarded experts in the area of research libraries and scientific data archives. Valuable perspectives from funding agencies were offered through presentations from Lucy Nowell, US National Science Foundation (NSF) and Carlos Morais-Pires, European Commission (EC). The Workshop also featured focused talks on a variety of domains. These included insights into the state of the art and requirements of the Biodiversity research community (Thomas Garnett, Smithsonian Institute), the High Energy Physics community (Jens Vigen, CERN), the Earth System Science community (Mark Parsons, National Snow and Ice Data Centre); eScience and eResearch in relation to the emergence of the economy of knowledge (David Prosser, Scholarly Publishing and Academic Resources Coalition Europe); a perspective into the rise of Digital repositories in India and challenges facing developing countries (ARD Prasad, Indian Statistical Institute);

the importance of libraries and information retrieval for the Food and Agricultural Organisation (FAO) of the United Nations ( Johannes Keitzer, FAO).

Roundtable discussions focused on user, technology and organisation perspectives led by Christine Borgman (University of California), Reagan Moore (San Diego Supercomputer Center) and Wolfram Horstmann (Bielefeld University Library), respectively. Valuable time was dedicated to in-depth discussions on evolving an agenda for GRL2020 and evaluating the research challenges that GRL2020 will help tackle in terms of international partnerships and from both technical and non-technical perspectives.

*I was impressed with the quality and depth of discussions, both group and individual-to-individual, at the workshop, but I was most struck by the clear realisation that global collaboration is essential to achieve the tremendous opportunities presented by digital scholarship, research and science. Such global collaboration requires the formulation and the shaping of a vision that reflects multi-discipline, multi-organisational, multi-national partnership activities that help move from projects to programmes related to data. GRL2020 helped advance better understanding of the elements of this vision.*

Peter Young, National Agricultural Library, USA

# 4. The Changing Face of the Research Library – Trends & Visions

As research becomes increasingly global, both in its aspirations and in reality, so must libraries. The rapid dissemination of findings, the creation of new tools and platforms for information manipulation, and open access to research data have rendered the more traditional institution-based library approaches to providing access to information inadequate and outmoded. Easy-to-use search engines providing access to a vast array of content have changed daily information seeking behaviour and expectations.

Lizabeth Wilson (University of Washington) is among the innovators and pioneers helping to reshape libraries into flexible learning spaces designed to meet the variety of user needs, such as collaborative and individual study, hi-tech, hi-touch instruction. The shape and form of the emerging library is associated with a series of challenges. Key issues include:

- The expectations of faculties and students, now and in the future.
- The role of global research libraries in the future in supporting a transforming university mission in a technology enabled world.
- GRL2020 investments and focus in the face of limited resources, conflicting priorities, proliferating user groups and often competing clientele.

In the future, faculties, students, and researchers will be able to access and use the information they require when and where they want it around the world, and in the format most appropriate to their needs. User needs will be woven into the fabric of the search for knowledge, and into the flow of discovery and research.

The scholarly communication system will be accessible and affordable irrespective of geographical boundaries and institutional affiliation. Physical and virtual libraries will improve research productivity and facilitate the generation of knowledge.

> User needs woven into the fabric of the search for knowledge

David Prosser Scholarly Publishing and Academic Resources Coalition Europe (SPARC Europe) highlighted five new roles of the library:

1. **Maximising dissemination of authors' work**
2. **Promoting of the institution through the promotion of research performed within the institution**
3. **Increasing participation in research publication: organising peer-review, alerting services, searching tools, etc**
4. **Providing virtual research environments that take advantage of Web 2.0 tools to fulfil the eScience needs of researchers and political masters and funders.**
5. **Taking responsibility for the long-term preservation of an institution's intellectual output (theses, data, publications, etc.).**

According to Johannes Keizer, Food & Agriculture Organisation (FAO), the future of GRL implies a change in the set of skills required of a librarian:

- The ideal librarian of the future will have two qualifications, a subject qualification and a qualification in knowledge organisation & information science.
- She/he will work much more with humans in facilitating knowledge exchange in an organisation.
- She/he will be much more technology conscious and will be able to use web technology.
- He/she will not only have an exquisite knowledge about existing information sources on the web, but also about the different tools and techniques to access them.

## 4.1. Collaboration across national boundaries

Libraries around the world will be even more interdependent and intertwined than ever before—not only with each other, but with stakeholders, information providers, knowledge creators, and users. A new cross-sector and global orientation is paramount.

Collaboration across national boundaries and institutions is the defining characteristic of global research libraries in the future. Pressing issues include scholarly communication, digital libraries, data repositories and information literacy, all of which require the contributions of many.

***Collaboration is imperative. Collaboration is needed to re-frame scholarly publishing and dissemination. Collaboration is fundamental to a digital library that has purpose and value. Collaboration will help engender a knowledge rich global community.***

It is imperative that collaboration opportunities need to be grasped by the digital library community. Indeed, formal discussion on library-data centre collaboration, which will potentially lead to refereed publications in the growing body of informatics literature, are currently being proposed. Other opportunities for collaboration include increased metadata sharing between libraries and data centres through OAI-PMH and related protocols in organized consortia, and the development of formal data management curricula and education programs.

However, collaboration is fraught with a number of roadblocks. While collaboration often does not come naturally, it is difficult to mandate. Budget structures, administrative lines, geographic boundaries, and reward systems can all create barriers to collaboration. It is crucial to learn how to cross borders and that innovative organisations pay attention to supporting the skills and providing the latitude needed for effective collaboration.

A truly global research library needs to encompass developing countries which are currently less equipped to deal with such issues. Despite the availability of highly-qualified IT personnel & a thriving economy, India's engagement with the GRL is impeded by a number technical barriers, such as

the lack of widespread internet access in universities; high access costs; and frequent power cuts.

Cross-sector & global orientation as crucial assets for tackling pressing needs

A number of challenges for GRL are pertinent to developing countries such as India, and a number of roadblocks need overcoming in order to ensure that the wider community can benefit from the new opportunities. ARD Prasad, Indian Statistical Institute, highlighted the challenges facing developing countries:

Political issues - researchers need motivating to support open access, both as a community and through government policy support, so as to make open access a truly effective movement.

Financial issues - libraries are frequently the most affected by spending cuts. Open content and collaboration between libraries could provide a solution, particularly for less funded libraries.

Legal issues - copyright laws may restrict the development of digital libraries unless transfer copyright laws are addressed nationally and globally.

Technological issues - Open Standards would ensure DLs are long-lasting and that information is reusable. Grid technology is an essential issue to be addressed. A barrier could be the costs of migration issues.

Major Barriers (geographical, linguistic, political prejudice) - While universities are being connected via the internet, progress has been slower than needed. Much more effort is needed to address language barriers in developing countries. Overcoming the largest barrier, political prejudice, is paramount if DLs are to become truly global.

## 4.2. Collaboration between Libraries and Data Centres

Modern science's increasing use of ICT means that the amount of data being produced by researchers is growing every day. The sheer size of digital data resulting from experiments or captured from sensor arrays is beyond belief, as we move from petascale to exascale data and computation. Data with geospatial and temporal attributes, such as that from seismic or environmental sensors, is irreplaceable. Data from clinical trials and other costly studies for healthcare is almost impossible to replicate.

One of the challenges facing domain-specialists and library communities is data preservation and defining the validity of data for long-term preservation. Librarians and archivists have an important role to play in helping us know what to preserve and how to protect it, while also providing open access. Despite the fact that preservation may be the best understood problem by researchers, it is also the most intractable. Although what needs to be done is clear, how this is achievable causes problems.

According to Mark Parsons (US National Snow & Ice Center), the Open Archival Information System Reference Model (OAIS) provides a good baseline of what is required for the preservation of data. However, detailed implementation is only beginning to be defined for the Earth sciences and it is clear that this process is both complex and costly. This may be the most critical and fruitful area for library involvement. Not only do libraries have centuries of experience in data preservation and curation, but they also have the necessary respect and support of institutions and society to provide long-term sustainability. A core challenge is to develop a sustainable business model whereby the entire scientific community and society at large can contribute to the sustained preservation of unique and critical data in an era of rapid technological, social, and environmental change.

The vision for the future is the development of a Data Utility which is:

- Simple
- Predictable
- Reliable
- Extensible
- Accessible (usable)
- Durable

It is essential that Libraries and data centres work together to achieve this and that they become more actively involved with the sharing of metadata. The vision for Libraries should guide thought leadership issues regarding: interface design and issues of predictability to ensure a user-friendliness; interoperability: the use of tried and true technology and practices; transfer mechanisms; communication protocols; usability; software design; cost models; periodic overhauls of systems recognized simply as a cost of doing business; data preservation; distributed vs. centralized data management; and the development of a utility model of a hybrid of distributed systems tied together by a more central authority such as the phone system.

With the establishment of initiatives such as IPY, the Electronic Geophysical Year (eGY); the Global Earth Observing System of Systems (GEOSS); the International Council of Sciences goal to take a leadership role in data management for the sciences (ICSU 2004); a proposed Union Commission for Data and Information in the geosciences; and the GRL2020 workshops, it is clear that the time is ripe for greater interdisciplinary collaboration and coordinated Earth system informatics.

***In particular, GRL2020 could work to build greater collaboration and information sharing, even convergence, between libraries and data centres.***

# 5. Case Studies and Best Practices

The case studies below illustrate the efforts currently ongoing in diverse domains to tackle specific challenges and advance the state-of-the-art in repositories and better serve the respective communities. The Annex includes a table of related, current EC-funded initiatives, many of which were highlighted at the 2nd GRL2020 Workshop.

## 5.1. Biodiversity: Biodiversity Knowledge Ecology and the Biodiversity Heritage Library

### Thomas Garnett, Smithsonian Institution Libraries

The field of biodiversity is intimately linked to a broad spectrum of research and data types, ranging from taxonomic identification to satellite imagery. Biodiversity is an example of knowledge ecology, that is, knowledge generated by analysing information from diverse sources for a more effective evaluation. The growing divide between born-digital, analogue and textual data raises key questions concerning conceptualisation, management, and software support. Comprehensive research in several disciplines like biodiversity requires considerable overheads. A number of activities are underway to address these issues. Initiatives include digitalising the legacy literature of biodiversity through OCR and providing taxonomically structured search tools making it openly available to readers and harvestable by intelligent software. Additionally, the development of interdisciplinary and international partnerships aimed at effectively managing large-scale research resources for the long-term are taking place. A change in the paradigm for modes of scientific communication, that is the taking back of the scientific journal, is also in progression.

**Drivers**: Finding reliable information in the face of the overabundance remains challenging. Search systems such as Google provide access to large volumes of information, commonly without consideration for the trustworthiness of data. Furthermore, the searches are biased by purchased priority listing and the preferences of others, rather than objective assessment of content & value.

One of the specific areas connected with biodiversity is understanding the wider effects of urbanization taking place in specific world regions, such as increasing deforestation in Latin America). From a broader perspective, data preservation in the domain of biodiversity is a pressing need for taxonomic literature as it has a long citation life & fast decay rate.

**Outcomes**: Data analysis leveraging expertise of diverse organisations & advanced technology resources (sensors/satellite), providing key information on the effects of changing conditions in the environment, in order to incorporate data from different sources and generate enhanced models of human behaviour on the net carbon dioxide balance. One example is the remote sensing data sets such as those from NOAA ESRL Carbon Cycle Greenhouse Gases Group provide information with very high sensitivity on shifting atmospheric gas levels in response to changing conditions. Local, remote and satellite imagery contribute key information

on changing silt loads in major rivers, while the residue making its way to oceans is detected by local chemical oceanographers and remote underwater sampling machines (ROVs), sending information to the world oceanographic centres, the information is assembled in models that also incorporate reflectance data from satellite imagery systems. Data analysis and sampling also takes place at the National Geophysical Data Center; the Center for Sponsored Coastal Research; and Oak Ridge National Laboratories.

**Best Practices:** The Biodiversity Heritage Library (BHL) is a consortium of ten natural history, botanical, and research institute libraries that collectively hold a substantial amount of the world's published knowledge on biological diversity. Data curation, preservation and management are among the issues being addressed. The BHL promotes data preservation with key goals centred on digitalising core sources (complete works); open access (information can be re-purposed, re-used & re-formatted; and congruence (coherent with a dynamic knowledge ecology).

The BHL value-add lies in its bibliographic accuracy for all materials so that all data can be re-purposed and re-used as needed, while ensuring congruence of original printed materials to digital versions. The value-add of BHL offers an advantage over Google, which does not meet the precision requirements of the user community.

Additionally, the Encyclopaedia of Life has developed a taxon names-based cyber-infrastructure, in order to effect indexing and organizing of any data object from biodiversity databases, data sets, on-line documents, and historical repositories, such as the BHL.

Users are empowered to integrate models with molecular bio-markers to have cross-walks between molecular surveys and geological cores by using taxonomically fuzzy search systems to extract structured data sets that bear on target and related coccolithophores that allow for the existing models to be refined and tested.

**Themes**: Open Access; data preservation; data analysis.

## 5.2. Earth Sciences: The Special Sensor Microwave/Imager (SSM/I): A Case of unanticipated User Communities

### Mark Parsons, National Snow and Ice Data Center

Defining best practices
underpinning an effective network

The original purpose of the special Sensor Microwave/Imager (SSM/I) was to support operational weather forecasting for the US Air Force and Navy. However, today the data is used for a broad array of global land, ocean and atmospheric monitoring applications. Analysis of remote sensing data is now invaluable to global climate change studies as it provides researchers with the longest records of sea ice concentration and northern hemisphere snow cover.

**Drivers**: The increased demand for detailed information regarding sea ice following recognition of its impact on climate, ecology and society.

**Outcomes**: The study of arctic caribou calving patterns has been aided by the exploitation of SSM/I data. Correlating SSM/I derived snow water data with caribou calving data has enabled to discover more about these arctic mammals. The use of the applications is well beyond the scope envisioned by the original designers of the sensor.

The evaluation of studies conducted has helped identify a number of pain points that need resolving. One conclusion is that the definition for data release with respect to aspects of time, and data use with respect to data type (applied or research) is vital. It is therefore important that thought processes and planning go beyond the technical information, documents, people and practices.

Three International Polar Year workshops have helped define some of the practices required to develop such a socio-technical system. The related themes of building trust and understanding quality were persistent in these workshops and guide the practices that underpin an effective network. One workshop explored how researchers search for and understand data outside their expertise.

Additionally, the ability to communicate with data experts in order to assess the quality of data in question is viewed as a critical piece of an interdisciplinary data discovery system. The importance of being sensitive to the intricacies and thought methods used by different user communities is vital, particularly in consideration of the often mutually exclusive approaches. A good case in point is the geographic information systems community, which sees the world as a collection of features with geographical footprints on the surface of the earth, discrete objects often described by a set of 2-dimensional shape and geometry characteristics. The view of the world to fluid-earth scientists such as climate modellers; on the other hand, is as a set of observations or measurements that are described by parameters such as temperature and velocity, that vary as continuous functions in 4-dimensional space-time. Behaviours of parameters are governed by a set of equations.

**Best Practices**: The development of new user communities. Fostering the need to build trust & to understand the importance of quality as a guide to practices underpinning an effective network. Promoting the ability to communicate with data experts to assess the quality of data as an essential feature in an interdisciplinary data discovery approach, and taking into account different approaches to data analysis.

**Themes**: Creation of new user community; collaboration with user communities; data preservation; data analysis for climate change.

# 5.3. High Energy Physics: PARSE.insight; INSPIRE and SCOAP3

## Jens Vigen, European Organisation for Nuclear Research (CERN)

The HEP community consists of 20,000 active scientists distributed across a number of large international laboratories and many universities. Experiments carried out in this field generally have a long life circle compared with other disciplines, with approximately 5,000 articles published every year, 90% of which are based on theoretical arguments.

The access of HEP information by the community relies heavily on community services such as subject repositories and lab-supported databases. A recent poll of the HEP community found that 90% of respondents used these methods for accessing information with only 9% use internet search engines such as Google.

> Defining the infrastructure for sustainable permanent access and use of digitally encoded information

**Drivers**: Like Astronomy and Climate Science, the highly complex nature of HEP data makes it difficult to be made Open Access. The highly complex nature of HEP data generated by a large community coupled with a complex procedure for final results. Enormous amounts of data generated.

**Expected outcomes**: PARSE.Insight (Access, Preservation and Curation) is an FP7 European funded project launched in 2008 to work with providers and users of scientific information and repositories to deliver insight into the issues of sustainable permanent access and provide cross-fertilisation of ideas and requirements between providers and users, the research community as a whole and national/European funding agencies. The project aims to help define the infrastructure needed to preserve and use the digitally encoded information on which the HEP society increasingly depends, but which is so fragile. These digital resources contain the intellectual inheritance for future generations and which need to be exploited to the fullest, right now, across domains.

The outcome of PARSE.Insight will be vital for the development of HEP repositories. Further streamlining of HEP repositories is also important with the curation and preservation of articles being archived in INSPIRE.

INSPIRE is the project name of a new High Energy Physics (HEP) information system which will integrate present databases and repositories to host the entire corpus of the HEP literature and become the reference HEP scientific information platform worldwide. It will empower scientists with new tools to discover and access the results most relevant to their research; enable novel text- and data-mining applications; deploy new metrics to assess the impact of articles and authors. In addition, it will introduce the Web2.0 paradigm of user-enriched content in the domain of sciences with community-based approaches to the peer-review process.

Peer review and Open Access publishing will also flourish with the development of SCOAP3, the Consortium for Open Access Publishing in Particle Physics which is working towards Open

Access publishing in High Energy Physics. In this model, HEP funding agencies and libraries, which today purchase journal subscriptions to implicitly support the peer-review service, federate to explicitly cover its cost, while publishers make the electronic versions of their journals free to read. Authors are not directly charged to publish their articles OA. The collection of preprints through arXiv also aids the process of streamlining.

**Best Practices**: Dealing with large amounts of often complex data, streamlining repositories, peer review of content. Defining the infrastructure needed to preserve and use the digitally encoded information on which the HEP community increasingly depends. Ensuring digital resources are exploited to the fullest, now and in the future.

**Themes**: Curation, preservation, open access

## 5.4. Nanotechnology: nanoHUB: Integrating Open Access and Web 2.0 Functionality into eScience and eResearch

### David Prosser, SPARC Europe

> Defining the infrastructure for sustainable permanent access and use of digitally encoded information

According to the Liber Development Plan (2007-2010), a primary role of the research library is to offer its user community the most efficient means of access and preserving the globally accumulated scholarly knowledge in their field of interest. Research outputs can be integrated into e-Science/e-Research. This can be done by taking content from repositories; adding Web 2.0 functionality; increasing the desire for collaborative working; creating resources that serve the community in new ways by providing not just content, but a complete research environment; making institutional repositories part of the infrastructure that allows e-Science to take place (across all disciplinary and geographic boundaries). An excellent example of the integration of Open Access into e-Science/e-Research is given by nanoHUB.

**Drivers**: integrating research outputs into eScience/eResearch by taking repository content and adding web 2.0 functionality.

**Outcomes**: the development of nanoHUB, a web-based resource for research, education & collaboration in nanotechnology with a large user community (25,000 nanotechnologists around the world). NanoHUB hosts over 1100 resources added by users including courses; on-line presentations; learning modules; podcasts; animations; and teaching materials. Simulation tools are also offered which are accessible from users' web browsers, meaning that users can learn about nanotechnology and simulate devices. The nanoHUB also provides a collaboration environment via workspaces, online meeting and user groups.

The nanoHUB was created by the NSF-funded Network for Computational Nanotechnology (NCN). The NCN is a network of 6 US universities with a vision to pioneer the develop-

ment of nanotechnology from science to manufacturing through innovative theory, exploratory simulation, and novel cyberinfrastructure. NCN students, staff, and faculty are developing the nanoHUB science gateway while making use of it in their own research and education. Collaborators and partners across the world have joined the NCN in this effort.

**Best Practices**: Open access through web 2.0 functionality to satisfy the need for a collaborative working environment; creating resources that serve the community in new ways by providing a complete research environment; making institutional repositories part of the infrastructure that allows e-Science to take place across all disciplinary and geographic boundaries.

**Themes**: Open access, web 2.0 functionality

## 5.5. Funding Agency: UK Repository Programme, Joint Information Systems Committee (JISC)

### Malcolm Read, JISC

Research papers and open access were the first drivers for repositories based on the assumption that open access improves research. The challenge that has now emerged and that needs addressing is data. By making data available, we can improve research and learning through verification, replication and understanding. Funding serves to create repositories, populate them and connect them with the aim of creating unified national repositories.

A culture of assessment responding to the need for an explicit focus on the user.

Repositories UK has developed a framework for technical interoperability, and policies for management. The aim is to set technical interoperability based on shared standards, and policy frameworks based on shared practice for the evolving UK federated network of digital repositories. Issues addressed include data ownership policies and IP; policies on open access; curation & preservation; technical interoperability.

A number of data-related challenges are faced by the programme, such as size, number, deposit rate, and nature of the use of data. The domain-specific expertise of librarians plays a key role in building discipline links across repositories to tackle rights and ethical limitations.

**Drivers**: improving the long-term availability & access to digital content through a network of repositories providing capacities for teachers, learners, & researchers to use & share content though a series of related projects, more details examples of which are provided in the annex. Fostering collaboration between research councils, national libraries, university research departments, computing departments, learning technologists and administration departments.

**Outcomes**: The JISC programme offers universities and their institutional repositories network an opportunity to establish themselves in a pivotal role in contributing to developing a digital repositories network providing better access to and management of intellectual outputs, in-

creased capability within the sector to manage these assets for education and research, and an infrastructure supporting the sector into the future.

**Best Practices** are a driving principle for JISC, which supports technical interoperability & policy frameworks. The drive towards best practices is reflected in the sharing & re-use of digital assets in co-operation with frameworks, such as DRIVER. Fostering collaboration between research councils, national libraries, university research departments, computing departments, learning technologists and administration departments.

**Themes**: policy; sustainability (financial/funding issues); interoperability; motivation & managerial issues.

## 5.6. University of Washington: Transforming the University Library to Meet User Needs

### Lisbeth Wilson, University of Washington

The future of libraries will be determined in large part by how the community collectively responds to the networked world, and anytime, anyplace expectations and realities. Universities will be measured by how well they disseminate knowledge and will need to find ways to share intellectual effort, in order to advance discovery and educate students for the fast-changing, largely unpredictable future. While the mission of the library is expected to remain constant, that is, meeting community information needs by gathering, organising, preserving, creating and disseminating knowledge, the strategies to achieve these goals will continue to change. This case study focuses exclusively on the Libraries at Washington University.

**Drivers**: Washington University's main objective is to respond to the changing role of the university library, serve as a user-centric organisation, make the best possible use of resources and market services to match user priorities by fostering a culture of assessment. To this end, the university has developed triennial surveys designed to assess user needs. Users are actively listened to with action taken on the views that emerged. Surveys, usability testing, environmental scanning, LibQUAL, focus groups, and learning outcomes are all part of the Washington toolkit.

Driving best practices – the building blocks towards a national digital repositories network & the sharing of digital assets

**Outcomes**: The surveys clearly indicate that the remote access of on-line information is the most preferred approach. Self-reliance within the library space is a key requirement. IT and on-line information resources are perceived as an enabler in terms of learning and research. Desk-top delivery of full-text resources is ranked as a top priority by faculties, under-graduates and post-graduates. Furthermore, Physical facilities are used extensively by students who account for more than 90% of the 4.4 million annual visits. Undergraduates use the library as their primary space for research, course work, individual study, socialization, group work and presentations.

The library is considered as the largest and most important learning space at the university, open 24 hours a day.

Focus Groups are dedicated to understanding specific information needs, such as the community of Bio-scientists. Interaction with the community has helped define the following recommendations:

- More electronic access is required
- The library is seen primarily as an e-journal provider
- Most subjects use the virtual library rather than the physical library
- Article databases are greatly underused
- There is a great need for personal information management
- Bioscience researchers are multi-disciplinary and multi-institutional collaborators working with people within UW, across the nation, and around the globe

**Best Practices:** The resulting Washington Toolkit enables a culture of assessment through the involvement of all stakeholders in the decision-making process. A culture of assessment is an environment in which decisions are based on facts, research and analysis, and where services are planned and delivered in ways that maximise positive outcomes and impacts for library clientele.

The culture of assessment is a fundamental part of the process of change and the creation of the research library for the 21st century as it responds to the need for an explicit focus on the user and a continual evaluation of the landscape. Strategic actions include listening to users and understanding where a difference can be made in connecting people with knowledge. User groups and their respective needs are defined through assessment tools, focus groups and learning outcomes.

**Theme**: End users.

# 6. The Importance of Identifying & Defining Use Cases

## Mark A. Parsons, US National Snow and Ice Data Center

Use cases are required by the global research library community to fully define the requirements of a future data system. Use cases should illustrate potential roles for libraries and data services, not only in the construction of robust search and exploration systems, but also by providing relevant expertise (curation) to better enable cross-disciplinary data discovery and use.

It is also paramount that the term 'use case' is clearly defined. One example of a use case is a data seeker needing to identify and assess a diverse set of data to examine the relationship between sea ice retreat and Arctic coastal changes (exposure to storminess, wave erosion, and coastal retreat) and assess the vulnerability of coastal settlements to projected reductions in sea ice. The figure below summarises the results of that use case developed at a workshop conducted by Mark Parsons two years ago. This use case only examines a small set of services that a data service should provide for someone seeking data.

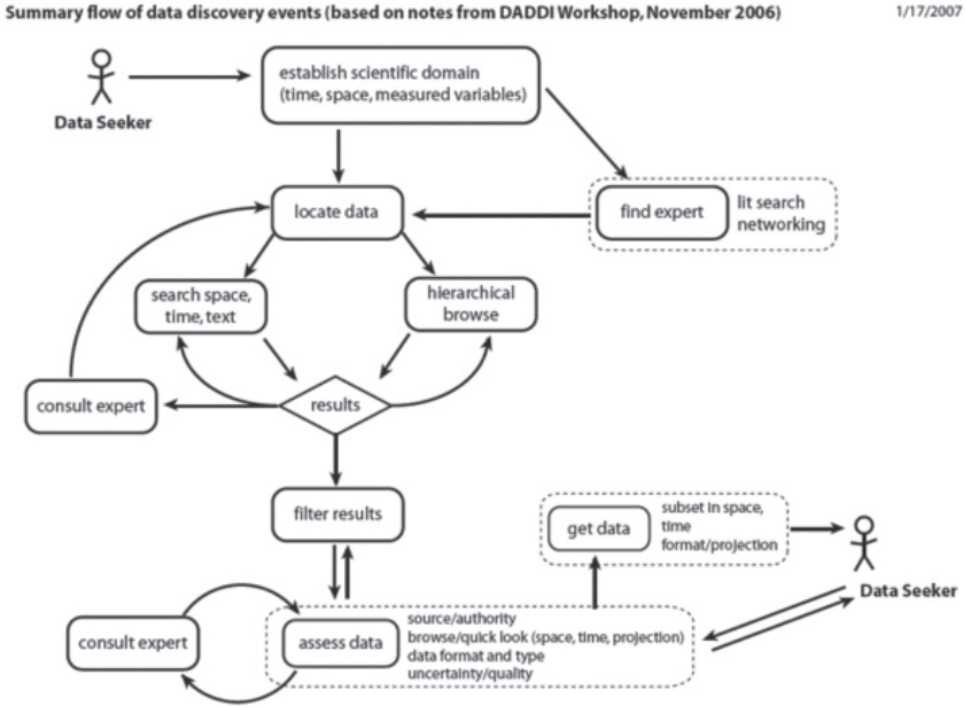


*Figure 1. Use case diagram created at an interdisciplinary workshop to describe how a user might search for data related to Arctic coastal processes. The full (draft) use case model is available at http://wiki.esipfed.org/index.php/SolutionsUseCase_CoastsOcean_Arctic_1a*

In another example, the National Research Council (NRC, 2006; p. 63) describes a more sophisticated future data portal serving a comprehensive Arctic observing network. Both these examples, while interdisciplinary, deal only with Arctic system science.

***GRL2020 needs to extend to all disciplines, including the humanities. Perhaps by examining a broad array of use cases we can determine some common themes that reach across all of them. What are the core informatics research problems we need to address regardless of whether we are managing information on astrophysics or James Joyce? Those core problems or themes could provide the basis of a research agenda to be developed at the next GRL2020 conference.***

# 7. Ten Drivers for GRL2020

As Thomas Garnett (BHL) points out in his Position Paper, there are great economies of scale by sharing the curation and preservation of research data, while allowing for a certain level of discipline control. The economies result from a number of issues, such as spreading costs over a wide base. Such sharing will require an organisation that can be virtual or incorporated. Special skills and new career paths will be needed; institutional, scientific and national support, which may include funding, willingness to contribute content, guidance for user-based services; research community buy-ins are also needed. David Prosser (SPARC) comments that potential repositories are cheaper as well as being more accessible for users. In addition, editorial value-add of repository content will be as high as that currently offered by publishers.

A number of key drivers for GRL2020 emerged from the position papers, dedicated tracks and interactive discussion featured in the 2nd GRL2020 Workshop. The themes explored are inextricably bound up with the development of a collaborative research environment, service provision and the user communities of the future. The table below provides an analysis of the key drivers related to the vision for GRL2020, which should be user-centric, based on open access, ensure high-quality data preservation & trust, and be interoperable.

| | |
|---|---|
| **User Focus** | **Responding to an ever-changing landscape:** libraries must cater for a new generation of users not only demanding information in digital format, but also becoming the future creators of content.<br><br>**Prioritising the future direction of libraries:** responding to the needs of future users should be the driving force. Libraries should become flexible and collaborative spaces.<br><br>**Fostering a culture of assessment:** evaluation should underpin an environment of continual change and encompass an explicit focus on user groups and a continuous assessment of the landscape.<br><br>**Shaping the new roles of users:** researchers need to be educated on aspects of data management, data creation, issues of provenance and interface use. In the future, actors may play multiple roles, whereby the citizen scientist is curator and the library (physical and virtual), the publisher and the repository.<br><br>Scientists, scholars, depositors and contributors will add value by tagging or making new discoveries. Seekers will aim to discover and retrieve information from the DL; in their role as "intelligent agents", they may also become contributors.<br><br>**Enabling users**: processes for creating and depositing data need to be very simple for users. Complexity must be hidden from users, for example, by writing data deposition into the software, so it is automatically produced. While the goal is to provide basic functions for depositing and discovering information and associated meta-information to be simple and transparent, users will need to be trained and educated in this process. |

| | |
|---|---|
| **Open Access** | **Commitment to maximum public access:** there must be open and harvestable repositories running on a community-operated infrastructure by an international consortium of librarians and computer scientists.<br><br>**Compliance with policy body endorsements:** open access is paramount to the Lisbon Agenda aim of making the EU the most competitive and dynamic knowledge-driven economy by 2010. Policy bodies, such as the National Institutes for Health, USA and the Organisation for Economic Co-operation and Development Committee for Scientific and Technological Policy, foster open access (Source: David Prosser, SPARC).<br><br>**Tackling data & IPR challenges:** there is a pressing need to address data and particularly issues regarding Intellectual Property Rights (IPR), such as ownership of data and copyright laws. Collaboration with librarians can help identify which resources should be open access. A comprehensive survey and metrics of open access repositories would be valuable. David Prosser (SPARC) highlighted that "all research data from publicly-funded research is to be publicly accessible following a review of the current Intellectual Property Rights (IPR) regulations". |
| **Preservation** | **Ensuring preservation** of valuable data & long-term curation: have become growing priorities with effective management and replication of digital holdings required to minimise the risk of data loss.<br><br>**Bringing value-add:** libraries can offer value-add to machine-readable information, creating new forms of cataloguing.<br><br>**Developing new tools:** tools for tackling provenance are vital as data is constantly improved and corrected. User friendly intelligent tools for metadata creation; complex document management; rule-based systems for automating administrative functions and assessment criteria evaluation; annotation; personalisation & customisation. There is also a need for new tools for improved semantic discovery, registries with semantic vocabularies; and scholarship communication services.<br><br>**Selecting data for preservation:** this is a challenging area for both policy-makers and researchers. Expertise on data suitable for preservation lies within the community. It is therefore important that preservation processes are developed in concert with institutional communities in order to maintain user community involvement. |

| | |
|---|---|
| **Trust and Quality** | **Fostering trust:** digital materials should be trustworthy and reliable with objective searches based on an objective assessment of content and value.<br><br>**Ensuring quality:** the completeness, consistency and authoritative source should be proven for each digital reference collection with links to data supporting them. Accurate provenance information is vital for this and built-in provenance tools for authors are required, which in turn will improve the quality and accuracy of search engines. These requirements should be enforced by granting agencies. Solutions for self-validating and checking data quality is an example of the value-add of DLs, setting it apart from the web. Socialisation processes are also important for this in order to qualify data appropriate for conservation. |
| **Data & Scalability** | **Tackling the deluge of data:** DLs are facing a deluge of data with library and information professionals required to handle vast quantities of information, including digital data, complex and compound information objects like 3D models and compound information objects. Automated procedures need to be put in place for curation, administration, data management and validation.<br><br>**Improving Data Management:** there is a need to take into account not only the huge volumes of information, but also the different scenarios where content can be used and re-used together with the different categorisation schemas possibly applicable to them.<br><br>**Tracking provenance:** as scientific data tends to be re-used, changes to the source are often not propagated and hence the data format or semantics can change, making tracking impossible. Methods for tracking provenance in such a distributed stateless environment are vital. |

| | |
|---|---|
| **Inter-operability** | **Driving interoperability:** for e-Science to be successful, interoperability is vital as isolated repositories bring little value to the research community. The establishment of a universal register would assist the interoperability of all scientific data regardless of original format.<br><br>**Building a seamless infrastructure:** ways to integrate and create a layered infrastructure design for interoperability of national, institutional and subject repositories across platforms, applications and data formats are needed, in order to create a seamless infrastructure layer of content and a global service network. Transition from bit stream manipulation to structured information manipulation is required. Interoperable context-based semantics are also required.<br><br>**Fostering Open Standards:** driving forward open standards is one way to ensure that DL content is reusable and future proof. However, migration issues could prove to be costly. Mapping of ontologies, schemas and services as well as a definition of methodologies to be used for the exchange of standards is necessary for interoperability. In order for interoperability to be achieved, libraries and communities should be motivated to collaborate with standards organisations. |
| **Sustain-ability** | **Cost- Benefits analysis:** whether libraries will rely on public funding or become self-sustainable is an open question. However, key issues that need addressing centre on the current & future costs involved, meeting capital & recurrent costs, and the need for a cost-benefit analysis. A full quantitative and qualitative analysis including a Cost efficiency analysis and Cost-benefit analysis of the DL communities would be a valuable asset for evaluating the real potential for sustainability.<br><br>**Continuous assessment & collaboration:** if investment in R&D is to continue and the connections between access to knowledge, technology transfer and wealth increases, accountability and assessment of universities, libraries and research groups will grow. Collaboration within and between institutions and commercial service providers are two possible solutions for attaining self-sustainability. |

| | |
|---|---|
| **Services** | **Importing & Exporting content:** libraries will compete on value-added services rather than on the size of collections in attracting users. Increasingly the two principle and simultaneous services on offer will be importing content from outside the institution to local researchers, and the 'exporting' of local research results to the wider academic community and beyond.<br><br>**Creating a global service network:** if each library focuses on a specific service (Cornell – ArXive, NIH – ePrints, OAISTER – uMich) which is then shared globally and customised to meet the needs of multiple user groups, a global service network may emerge. Such collaboration has a long-standing tradition with libraries.<br><br>**Facilitating concerted efforts to create digital content:** the e-Infrastructure would be both stable and interoperable globally through joint and co-ordinated efforts to create digital content. Once assembled, this content should be easily accessible with features, such as multilingual searching; data weighting & results ranking, as well as scalable browsing.<br><br>**Services requiring development:** user friendly tools for depositing are required. The development of access controls is a current impediment. User friendly intelligent tools for metadata creation; complex document management; rule-based systems for automating administrative functions and assessment criteria evaluation; annotation; personalisation & customisation; improved semantic discovery, registries with semantic vocabularies; and scholarship communication services. |
| **Libraries** | **Leveraging the expertise of the librarian community:** issues such as data ownership, preservation & curation, open access, quality & management are all within the expertise of the librarian community, which makes collaboration a vital asset. Experience in the UK has showed that both IT and librarian skills are necessary, with librarians playing a pivotal role in the management of repositories. Current pointers indicate that that in the future, librarians will have two qualifications: a subject qualification and a qualification in knowledge organisation/ information science. Skills and subject-based expertise input will aid the development of interoperable context-based semantics which are dependent on knowledge of context vocabulary. Another possible approach is outsourcing repository management based on successful developments in the UK, which merit further investigation. |

## Unique Positioning

**Libraries as catalysts for facilitating knowledge-sharing:** The central position of libraries in the world of research should be exploited by DLs, which should disseminate information, facilitate knowledge sharing between institutions, establish cross disciplinary connections and identify synergies. This is especially important as increasingly differing and polarising user communities develop. Cultural barriers between scientific disciplines, between data managers and researchers, between libraries and data centres need to be bridged with the development of an infrastructure

# 8. Defining the Features of GRL2020

*The Global Research Library will be an interoperable network of services, resources, and expertise designed to facilitate the process of research and the selecting, sharing, and sustaining of the outputs of research.*

Fabrizio Gagliardi, Microsoft Research

There is a general consensus on the important concepts of data curation and access 'utility' – a core infrastructure of science that is simple, predictable, reliable, extensible, accessible and durable. However, beneath the basic simplicity of this concept there lies deep complexity, structure, planning and professionalism. Creating such a level of infrastructure requires great collaboration around standards, maintenance, and professional development and certification. Cultural barriers between scientific disciplines, between data managers, researchers, librarians and data centres must be bridged. The time is ripe for greater interdisciplinary collaboration, which GRL2020 could facilitate and foster.

A truly global research library needs to encompass developing countries that are currently less equipped to deal with such issues. Despite highly-qualified human resources, the engagement of developing countries is impeded by a number of technical, economic, government and cultural barriers. Overcoming major barriers, such as geographical, linguistic and political prejudice, is paramount if Digital Libraries are to become truly global.

It is widely viewed that the Global Research Library of the future should be:

- Multi-ethnic, multi-cultural and multi-lingual.
- A collaborative and global environment, which emphasises the ethical issues surrounding data.
- Purposefully inclusive, attending to different cultures.

One of the Grand Challenges facing stakeholders is to establish a cross-disciplinary approach, where data plays a key role and to develop a research agenda which demonstrates that GRL2020 is crucial for solving key issues. Such an approach would broaden the subsets of prime beneficiaries and leverage the expertise of developers and domain specialists that exists on an international level. This approach would also make it possible to harness the digital infrastructure to the requirements of researchers and user groups and test the technology against these needs. Furthermore, a case study focusing on a key issue of global relevance would resonate with potential funding agencies, and ultimately the wider community.

A top-down, bottom-up approach is therefore considered to be the most effective way of establishing a collaborative partnership aimed at developing a global research environment.

# 9. Driving Forward GRL2020

## 9.1. Demonstrating the Value-add of GRL2020 - Bottom-up Approach

Cross-domain approaches provide tangible examples of where data is required to solve important problems, as well as real requirements to measure the value-add of collaboration. Examples highlighted by experts include Bio-diversity; Climate Change and global Public Health. Overarching objectives would focus on establishing a culture of data contribution that currently exists in only a few fields and bringing together data from multiple sources and nations to enable solutions.

The experiences gained from the International Polar Year projects, for example, illustrate that Climate Change involves fields such as astronomy, bio-diversity, and chemistry and entails the creation of data sets. The creation of a climate change library would address a number of key issues, such as how human resources and data are brought together in a value-added way; how interoperability and sociological barriers could be overcome; exploring the possibility of building use case studies around data sets across a number of disciplines, which poses a number of challenges for fields that traditionally do not consider such an approach as part of their core work, and understanding how different groups of people can be connected and catalysed.

The table below summarises the key requirements, both technical and non-technical, which would pave the way for a collaborative research environment.

| | |
|---|---|
| **Key Technical Features** | Curation & data quality; data processing synthesis; information retrieval for modelling and simulation; life cycle of scientific processes & management tools; preservation; provenance management; interoperability, visualisation and workflows. |
| **Synergies with the User Community** | Engagement with domain-specialists & users is paramount to define requirements and test the technology against those needs. Communicating and sharing with the subject community serves to attain valuable feedback on data, new tools, and the ability to access past evidence. |
| **Core Messaging for user groups & the wider community** | Sensitivity to the outside world is important to communicate the new processes, with a user-friendly approach to the underlying concepts. Core messaging should focus on the value-added services, while hiding the complexity from users, showcasing best practices and the benefits gained by the wider community. |

Specific challenges centre on the location of data, the relevance of the data, the combination of and distinction between audiences for data sets, and the provision of the type of data for researchers (geographically or scientifically tagged; domain specific). The processes estab-

lished should involve domain scientists, in order to leverage their specific knowledge on target communities and use of data. Challenges surrounding research libraries regard the possibility of re-formatting data, dealing with semantic issues and human-to-human curation.

The digital infrastructure should illustrate efficient, authoritative access to data, making the case for data contribution to solve key problems. Data must make a compelling argument and generate real knowledge that is of value. It is vital that communities clarify what information infrastructure is required, in order to map and connect the data. The goal is to ensure preservation of high-quality data requiring different workflows and tools. While there have been efforts in the workflow community to generate provenance information, it is not clear where the data will be preserved. Different workflows should be employed to meet the requirements that enable end products to be created. An issue to be addressed is therefore the implications for libraries to make data available with the required tools and how workflows can be preserved.

The ultimate goal of the infrastructure is to enable the solving of problems more quickly, speeding up the time from research to publication and availability to people through the ability to locate large amounts of data spread across global communities, and to re-use and re-purpose data to transform it into knowledge that serves both the research community and ultimately the wider community.

## 9.2. The Role of Policy-Makers & Funding Agencies – Top-down Approach

The area of Scientific Data is a priority in the 7th European Framework Programme of Research and Development. In particular, the Capacities Programme frames it in the e-Infrastructures domain dealing with Digital Repositories, Computing Grids and High-speed networks serving Scientific and Education communities

The paradigm shift that characterises the vision of GRL2020 is an important opportunity. Europe seeks to bridge research, education and innovation in connection with the Lisbon Agenda, which is aimed at making the EU 'the most competitive and dynamic knowledge-driven economy by 2010. At present, Europe is engaged in projecting a European Research Area (ERA), that is, a space for research with enhanced access to scientific information. Progressed tools and methods, fast networks and grids form the basis of this paradigm shift, which will generate data to be transformed into knowledge. Member states have a strong interest in an efficient scientific information system that maximises the socio-economic impact of investments in research and technological development. The interest in Global Research Libraries stems from the role they can play as infrastructures for research and education to face the challenge of modernising higher education systems.

From a US perspective, Global Research Libraries open a window of opportunity, not only to tackle key issues such as data preservation and sharing but, primarily to engage in complex global research challenges that require the expertise of many disciplines across national boundaries. The NSF's Office of Cyberinfrastructure (OCI) has developed a vision to combine expertise in library and archival sciences, computer, computational and information sciences,

cyber-infarstructures and engineering through the DataNet Partner program. OCI's vision for DataNet Partners recognises the importance of international connections and collaborations. Libraries and librarians have a crucial role to play in preserving and supporting access to and use of data. The challenge is to ensure the effective transformation of the research library to meet emerging requirements, in order to play a part in the global collaborative research environment.

# 10. Conclusions & Recommendations

The GRL Community of experts and domain specialists has defined a vision to develop a research environment enabling collaborative working across disciplines and geographical boundaries. Technological requirements have been identified with emphasis placed on open access, interoperability, and digital curation , while underscoring the importance of interoperability. The future sustainability of GRLs remains unclear and calls for a cost-benefit analysis to be developed over time. Experts have additionally evaluated a number of key cultural, ethical, and political issues which need addressing through targeted support actions.

The GRL2020 community is now poised to evolve a research agenda and chart a course that visualises next steps. This Report aims to serve as a foundation for the evolved agenda and future steps. The domain-specific case studies showcase some of the achievements to date. An extended portfolio of case studies, together with a set of use cases, would help foster best practices and fully define the requirements of a future data infrastructure. The GRL2020 website with an in-built wiki will serve as a forum that supports and further enhances these achievements.

Top technology recommendations centre on a more in-depth evaluation of state-of-art systems, such as production systems; ease of use; the ability to hide complexity from users and the ability to implement control policies. The establishment of Working Groups is considered to be valuable, helping to define the role of shared test-beds, the development of semantics and contributions  of existing repositories within the GRL2020 scenario.

In conclusion, an evaluation of the real opportunities to advance international, inter-disciplinary and cross sector collaborations can pave the way towards developing GRL2020

# 11. Annex I - Participants at the 2nd GRL2020 Workshop, 27-28 March 2008, Pisa, Italy

| Name | Surname | Organisation / Affiliation | Country |
|---|---|---|---|
| **Programme Committee** | | | |
| Peter | Buneman | University of Edinburgh | UK |
| Donatella | Castelli | CNR-ISTI | IT |
| Lee | Dirks | Microsoft Research | US |
| Fabrizio | Gagliardi | Microsoft Research | US |
| Jessie | Hey | University of Southampton | UK |
| Yannis | Ioannidis | University of Athens | GR |
| Lizabeth | Wilson | University of Washington | US |
| **Participants** | | | |
| Nicoleta | Albu | Politehnica University of Bucharest | RO |
| Pam | Bjornson | Canada Institute for Scientific and Technical Information | CA |
| Cristine | Borgman | University of California | US |
| Harry | Bruce | University of Washington | US |
| Graham | Cameron | European Bioinformatics Institute (EBI) | UK |
| Kurt | De Belder | Leiden University Library | NL |
| Nicholas | Ferguson | Trust-IT Services | UK |
| Luigi | Fusco | European Space Agency | IT |
| Thomas | Garnett | Smithsonian Institute | US |
| Stefan | Gradmann | Humboldt Universität zu Berlin | DE |
| Wolfram | Horstmann | Bielefeld University Library | DE |
| Johannes | Keizer | Food & Agriculture Organisation | IT |
| Simon C. | Lin | Academia Sinica | TW |
| Paola | Mancini | Scuola Normale Superiore di Pisa | IT |
| Anton | Mangstl | Food & Agriculture Organisation | DE |
| Natasa | Milic-Frayling | Microsoft Research | UK |
| Reagan | Moore | San Diego Supercomputer Center | US |
| Carlos | Morais-Pires | European Commission (EC) | BE |
| Silvana | Muscella | Trust-IT Services | UK |
| Lucy | Nowell | National Science Foundation (NSF) | US |

| | | | |
|---|---|---|---|
| Stephanie | Parker | Trust-IT Services | UK |
| Mark A. | Parsons | National Snow and Ice Data Center | US |
| Enrica | Porcari | Consultative Group on International Agricultural Reseaerch (CGIAR) | IT |
| A.R.D. | Prasad | Indian Statistical Institute | IN |
| David | Prosser | Scholarly Publishing and Academic Resources Coalition Europe (SPARC) | UK |
| Randy | Ramusack | Microsoft Research | US |
| Malcolm | Read | Joint Information Systems Committee (JISC) | UK |
| Eloy | Rodrigues | Universidade do Minho - Serviços de Documentação | PT |
| Federico | Ruggieri | National Institute of Nuclear Physics | IT |
| Alexander | Sotnikov | Joint Supercomputer Center | RU |
| Eldon | Ulrich | University of Wisconsin | US |
| Edward | Van den Berghe | Ocean Biogeographic Information System | US |
| Dany | Vandromme | The Pan-European Gigabit Research and Education Network | FR |
| Jens | Vigens | European Organisation for Nuclear Research | CH |
| Alex D. | Wade | Microsoft Research | US |
| Leigh | Watson Healy | Outsell | US |
| Peter | Young | National Agricultural Library | US |

# 12. Annex II - Overview of Break-out Sessions

## 12.1. User Perspectives, Christine Borgman, University of California

*Digital libraries will be embedded in the day-to-day environment.*
*Building collaborative platforms: interoperability & standards should be common goals.*

### Users: multi-roles

The same person may have multiple roles. Scientists, scholars, depositors & contributors will add value by tagging or making new discoveries. Seekers are individuals or agents wanting to discover & retrieve from the DL. Intelligent agents will play an increasingly important role and may also be contributors. The notion of the user shaping the vision is associated with some users being seekers, others contributors.

### Solutions, Challenges & Impediments

A key issue is identifying the requirements of each type of user.

| Solutions | Challenges & Impediments |
|---|---|
| Generalise | Privacy (tracking provenance and intellectual property) |
| Personalise | Fears of misuses |
| Adapt | Sustainability |
| Re-combinable/interoperable | Motivation |
| Easy tools for depositing: more work done on search; much needed here in terms of interface | Access controls |
| Branding & badging | Semantic interoperability |
| Standards development, participation | Right-sizing |
| Tracking provenance & IPR (Intellectual Property Rights) | Building collaborative platforms and tools |
| Feedback & Reporting back | Lack of resources in public schools & libraries |
| Provide solutions to self validate or check quality | Issue of competition, especially between schools |

| | |
|---|---|
| Review mechanism for data | Services that give better quality are important |
| Transparency of location in distributed environments | |

## 12.2. Technology, Reagan Moore, San Diego Supercompter Center

**6 Key Challenges for the GRL2020 Vision:** *Size of collections, Multi-linguality, Socialization (data life cycle), Interoperability, Registries, Data formats*

### Overview

This Group explored whether Library metaphor/context is sufficient, as well as the differences between hard science data sets, which are large scale, and humanities data sets, which mainly comprise unstructured data. Other issues include long-term preservation & how shared collections could be created from distributed data.

The Group assumed we are dealing with digital data as opposed to digitalization. Reasonable expectations include new data formats, as well as complex & compound information objects, such as 3D models and compound information objects.

| Six Challenges for the GRL2020 Vision | |
|---|---|
| Size of collection | Exabyte (petabyte collections exist today). Science collections with trillions of records. |
| Multi-linguality | Both for interpreting record content & discovering relevant records. Concern with OCR limitations for interpreting languages. Influences support for semantics |

| | |
|---|---|
| Socialisation (data life cycle) | Collaboration support mechanisms. Mechanisms to build multiple levels of consensus on collections.<br>Quality control, associated with properties of objects; one way in which DLs are distinct from the web.<br>This is a socialization process:<br>• Socialisation issues for what qualifies as a reference collection<br>• Bringing data in requires multiple levels of consensus across multiple groups<br>• Policy development per collection<br>• Relates to quality control |
| Interoperability | Across platforms, applications & data formats; modular infrastructure design for extensibility; layered infrastructure design for interoperability.<br>Transition from bit stream manipulation to structured information manipulation. Context based semantics – a contentious issue – structure on semantics, meaning associated with terms depends on the context of use; discovery & browsing require knowledge of context for vocabulary – often implicit rather than explicit, depending on community that assembled the data. |
| Registries | Semantic vocabularies. Information repositories- global namespace for data. |

| | |
|---|---|
| Data Formats | Structure characterisation:<br>• Minimal entity description for compound objects<br>• Structure syntax standards – DFDL, multivalent, UVC)<br>• Relationship characterization<br>• Relationships implicitly assigned to structure by application<br>• Fedora, FoxML – relationships among named entities<br>Behaviour characterisation: sets of operations that can be performed on the relationships<br>• iRODS micro-services<br>• multivalent behaviour<br>• Fedora behavior |

## Emerging Technologies

1. Generalisation of context-based semantics
- Context includes semantic, spatial, temporal & functional relationships
- Faceted ontologies
- SKOS - Simple Knowledge Organization System
  - Good within a single domain
  - Needs extension to multiple domains
  - Needs inclusion of authority files

2. Goal is the elimination of the artificial barrier between humanities and hard science data sets
   - Rule-based data management systems
   - Differentiate between communities on basis of management policies

## Digital Library Infrastructure: Possible Approaches & Issues

1. Services
   - Desire for improved semantic discovery
   - Desire for registries
   - Desire for simple user interface
   - Desire for scholarship communication
2. Collection management
   - Rule-based systems for automating administrative functions, assessment criteria evaluation
   - Desire for standard functional primitives for interoperability
3. Desire for:
   - Robust systems
   - Reliability
   - Production quality
   - Controlled sharing (authentication / authorization)
4. No consensus on state of current systems: pointing toward need for further communication on:
   - Production systems
   - Ease of use
   - Ability to hide complexity from users
   - Ability to implement control policies

## Q&A Session – Points Highlighted

- Difficulty of communication across languages, cultures & professions.
- Challenge in addressing different communities – communication & links with players from different "groups", e.g. scientists/scholars, that is, hard science versus humanities. There is a real risk of using divisive language and of increasing the gap between different communities, which would not be conducive for interdisciplinary approaches.
- Elimination of the barriers between science & cultural heritage data, as well as between science & humanities.
- Leverage of the power of language technologies & socially adept technologies.
- Need to enlarge the community & the discussion, how to move towards agreed semantics.
- Need for global agreement.
- Need to be community specific – registries created with common methodologies.
- The establishment of Working Groups would be a valuable overall contribution.
- There is a need for shared test-beds.

# 12.3. Organisers Perspectives

## Wolfram Horstmann, Bielefeld University Library

*The future of libraries is in the hands of the kids, not of the libraries. Scientific insight cannot be replaced by high throughput.*

## Emerging Groups

Conventional groups comprise:

Users

- Institutions
- Libraries
- Funding agencies & tax-payers.

The emerging, new groups are made up of:

- Virtual organizations
- Project-driven initiatives
- Globalised subjects

## Polarising Factors

- Subject-based versus Institutional.
- Disciplines have different speeds for development: Disciplinary differentiation shapes customers faster than institutional service providers can handle.
- Conservative versus progressive: Document repositories may take 20 years to be established as an accepted service. Globally accepted data services can be created in just a couple of months (e.g. BioMed, Astro, GEO).
- Text versus data: conventional library media will become less important.

## Libraries in 2020

There is a common consensus that libraries cannot stay as they are as users are tending to increasingly sideline their services. So what are the most likely new roles for libraries, assuming that they will have a role to play?

New roles

- Interfacing between publication & data – serving as curators
- Acting as cross-walks between subject-based services
- Focusing on 1 specific service?

## Libraries & data – a disputable liaison?

The library scenario

- GRLs can be the Interface between publication and data
  - Apart from subject based curation
  - Apart from pure IT (storage)
- Librarians can become Curators
  - Adding value to (machine readable) information
  - Creating new forms of cataloguing (tagging?)
- The data centre scenario
  - Data centres provide literature and data

## Libraries & Subjects – a new dream-team?

The more specific subjects become, the stronger is the need to bridge gaps will be. Libraries have all the prerequisites to deal with this:

- Orientation towards users
- Bridging between subject-based applications

Librarians can be facilitators and advisors

- Add value to (machine readable) information
- Creation of new forms of cataloging and tagging

## Libraries & Services – a new service puzzle

No one library is able to serve

- Differentiating user communities
- Differentiating subjects
- Differentiating data types

What if each library focused on 1 service?

- Collaboration has a long standing tradition in libraries, e.g., Cornell (hosting ArXive), NIH (hosting PubMedCentral), Soton [Southampton] (ePrints), OAISTER (uMich), BASE (Bielefeld)

## Conclusions

The Institutional Library still is the central unit of organisation - but it is embedded in global service network.

## Challenges & Impediments

- Positions & Careers
- What are we doing?
- Mandates, "Institutional Ego"
- Inertia

## Q&A Session – Points Highlighte

- Data will become the publication & where credit will be given
- Knowledge management infrastructures will be institutional and global
- Different levels of services and sources thereof exist
- Researchers lose out when they don't rely on libraries
- Many alternatives along with and not of libraries exist
- Useful methodology should be used to achieve clarity – take an extreme position and then attempt to articulate why it isn't true.
- Users don't have to physically come to the library to be using library services

Young people will bring totally new approaches, different from ours.

- Change from accumulation logic to selection, aggregation, appraisal logics. Selective aggregation is an important function
- Differentiation of the library as collections from the skills of those who run them. The skill set is a critical factor and distinct from the buildings and collections