



# **Riding the wave**

**How Europe can gain from the rising tide of scientific data**

Final report of the High Level Expert Group on Scientific Data  
A submission to the European Commission

October 2010

# **Riding the wave**

## **How Europe can gain from the rising tide of scientific data**

**Final report of the High level Expert Group on Scientific Data**

**A submission to the European Commission**

**October 2010**

© European Union, 2010

Reproduction is authorised provided the source is acknowledged.

The views expressed in this report are those of the authors and do not necessarily reflect the official European Commission's view on the subject.

Printed by Osmotica.it

The members of the HLG would like to thank the project GRDI2020 for supporting the meetings logistics and the arrangements for the publication of the final report. GRDI2020 is a coordination action project funded by the European Seventh Framework Programme for Research and Development (FP7) under Grant Agreement RI-246682.

# Unlocking the full value of scientific data



**“Information and Communication Technologies (ICT) are the most recent transformational factors in science.”**

**T**he Digital Agenda for Europe outlines policies and actions to maximise the benefit of the digital revolution for all. Supporting research and innovation is a key priority of the Agenda, essential if we want to establish a flourishing digital economy by 2020.

Scientific research is supported by its infrastructures: technical tools and instruments and socio-economic systems for organising and sharing knowledge. These have been in constant change for many centuries reflecting advances in technology and change in political systems. Key inventions like the microscope or the telescope resulted in huge

scientific progress by allowing the validation or rejection of theories; and the invention of book printing in the 15th century and the organisation of knowledge in research libraries allowed unprecedented access to knowledge.

Information and Communication Technologies (ICT) are the most recent transformational factors in science. They enable close and almost instantaneous collaboration between scientists all over the world and they provide access to unprecedented volumes of scientific information that can in turn be processed on powerful computational platforms. Many younger scientific disciplines would not even

exist without access to these technologies. Today ICT-based infrastructures (e-infrastructures) have become an essential foundation of all research and innovation.

This is reflected in the European Commission and EU Member States investing in different domains of e-infrastructures. Together we have been working on connecting researchers, scholars, educators and students through high speed research networks like GÉANT, providing access to shared grid and cloud computing facilities, and developing supercomputing capacity for very demanding applications through the European partnership PRACE. To complement these developments, Europe is putting the seeds for the emergence of a robust platform for access and preservation of scientific information.

All these are and will remain important elements underpinning European research and innovation policies. However, with robust infrastructure for data transmission and data processing in place, we can now start to think about the next step: data itself. My vision is a scientific community that does not waste

resources on recreating data that have already been produced, in particular if public money has helped to collect those data in the first place. Scientists should be able to concentrate on the best ways to make use of data. Data become an infrastructure that scientists can use on their way to new frontiers.

Making this a reality is a more difficult task than it may seem. To collect, curate, preserve and make available ever-increasing amounts of scientific data, new types of infrastructures will be needed. The potential benefits are enormous but the same is true for the costs. We therefore need to lay the right foundations and the sooner we start the better. This report of the High-Level Group on Scientific Data will be an invaluable input for formulating our research and research-infrastructure policies. I invite every citizen and every organisation involved in scientific research to take note of this report and to use it as a reference point when discussing the priorities of EU research investments.

**Neelie Kroes**

*Vice-President of the European Commission,  
responsible for the Digital Agenda*



FROM THE CHAIR

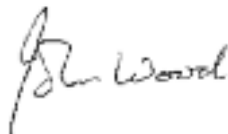
## On the challenges ahead



present the report of the High Level Group on the future of scientific data. The importance of facing up to the challenges before us is crucial if European research is to remain at the leading edge globally.

The resulting actions that we propose will affect all areas of research, not just big science. This range has been reflected in the group as we have considered the impact on, for example, the humanities, publishing, and bio-diversity in addition to large international science facilities. Indeed, getting it right will affect the way research is done in the future and will be instrumental in ensuring that the challenges before us are solved in a holistic way rather than allowing individual disciplines to dig entrenched positions. Just how students will be trained in the future, or how the profession of “data scientist” will be developed, are among the questions the resolution of which is still evolving and will present intellectual challenges for both privately and publicly supported research. Critical to everything is how trust can not only be fostered but ensured so that the “Fifth Freedom of Knowledge” is pursued with vigour for the good of all society.

In addition to the High Level Group coming from a diversity of backgrounds, the liveliness of the discussions and the working atmosphere have been a delight and I thank the members for their excellent contributions. Also my thanks to the Commission staff who have entered into the debate with an exemplary degree of open-mindedness. Finally I would like to acknowledge the assistance of the various people who came to the group to share their thoughts and experience with us from around the world, to rapporteur David Giarretta who brought the discussions together into a coherent structure and action plan, and to Richard Hudson who miraculously took our stream of consciousness ideas and turned them into a coherent report.

A handwritten signature in black ink, appearing to read 'John Wood', written in a cursive style.

**John Wood**  
*Chair*



A fundamental characteristic of our age is the rising tide of data – global, diverse, valuable and complex. In the realm of science, this is both an opportunity and a challenge. This report, prepared for the European Commission's Directorate-General for Information Society and Media, identifies the benefits and costs of accelerating the development of a fully functional e-infrastructure for scientific data – a system already emerging piecemeal and spontaneously across the globe, but now in need of a far-seeing, global framework. The outcome will be a vital scientific asset: flexible, reliable, efficient, cross-disciplinary and cross-border.

The benefits are broad. With a proper scientific e-infrastructure, researchers in different domains can collaborate on the same data set, finding new insights. They can share a data set easily across the globe, but also protect its integrity and ownership. They can use, re-use and combine data, increasing productivity. They can more easily solve today's Grand Challenges, such as climate change and energy supply. Indeed, they can engage in whole new forms of scientific inquiry, made possible by the unimaginable power of the e-infrastructure to find correlations, draw inferences and trade ideas and information at a scale we are only beginning to see. For society as a whole, this is beneficial. It empowers amateurs to contribute more easily to the scientific process, politicians to govern more effectively with solid evidence, and the European and global economy to expand.

But there are many challenges. How can we organise such a fiendishly complicated global effort, without hindering its flexibility and openness? How do we incentivise researchers, companies, and individuals to contribute their own data to the e-infrastructure – while still trusting that they can protect their privacy or ownership? How can we manage to preserve all this data, despite changing technologies and needs? How to convey the context and provenance of the data? How to pay for it all?

Our vision is a scientific e-infrastructure that supports seamless access, use, re-use, and trust of data. In a sense, the physical and technical infrastructure becomes invisible and the data themselves become the infrastructure – a valuable asset, on which science, technology, the economy and society can advance. Our vision is that, by 2030:

- All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.
- Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.
- Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories work to international standards, to ensure they are trustworthy.

- Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of publicly generated data.
- The innovative power of industry and enterprise is harnessed by clear and efficient arrangements for exchange of data between private and public sectors, allowing appropriate returns to both.
- The public has access to and can make creative use of the huge amount of data available; it can also contribute to the data store and enrich it. All can be adequately educated and prepared to benefit from this abundance of information.
- Policy makers are able to make decisions based on solid evidence, and can monitor the impacts of these decisions. Government becomes more trustworthy.
- Global governance promotes international trust and interoperability.

There is a clear role for government in all this; and we offer a short-list of action by various EU institutions – building on work already begun across the EU in recent years, and complementing efforts in the US, Japan and elsewhere in the world.

### **1. Develop an international framework for a Collaborative Data Infrastructure**

The emerging infrastructure for scientific data must be flexible but reliable, secure yet open, local and global, affordable yet high-performance. There is no

one technology that can achieve it all. So we need a broad, conceptual framework for how different companies, institutes, universities, governments and individuals would interact with the system. We call this framework a Collaborative Data Infrastructure, and we urge the European Commission to accelerate efforts – in Europe and around the globe – to make it real.

### **2. Earmark additional funds for scientific e-infrastructure**

Development of e-infrastructure for scientific data will cost money, obviously – and as there is a significant element of public good in this, so there must be a significant degree of public support. One obvious source is found in the EU's Structural Funds – a portion of the budget mostly used to build roads, industrial parks and other key infrastructure, targeted at those regions of Europe most in need. Already, a portion of this budget is earmarked for research and innovation, including digital infrastructure. We call upon the European Council to expand the funding possibilities.

### **3. Develop and use new ways to measure data value, and reward those who contribute it**

If we are to encourage broader use, and re-use, of scientific data we need more, better ways to measure its impact and quality. We urge the

European Commission to lead the study of how to create meaningful metrics, in collaboration with the 'power users' in industry and academia, and in cooperation with international bodies.

#### **4. Train a new generation of data scientists, and broaden public understanding**

We urge that the European Commission promote, and the member-states adopt, new policies to foster the development of advanced-degree programmes at our major universities for the emerging field of data scientist. We also urge the member-states to include data management and governance considerations in the curricula of their secondary schools, as part of the IT familiarisation programmes that are becoming common in European education.

#### **5. Create incentives for green technologies in the data infrastructure**

Computers use energy; and as the tide of scientific data rises further the energy consumption risks rising in tandem. We urge the European institutions, as they review plans for CO2 management and energy efficiency, to consider the impact of e-infrastructure and prepare policies now that will ensure we have the necessary resources to perform science.

#### **6. Establish a high-level, inter-ministerial group on a global level to plan for data infrastructure**

It makes no sense for one country or region to act alone. We urge the European Commission to identify a group of international representatives who could meet regularly to discuss the global governance of scientific e-infrastructure. It should also host the first such meeting.





# I. Riding the wave

**W**e all experience it: a rising tide of information, sweeping across our professions, our families, our globe. We create it, transmit it, store it, receive it, consume it – and then, often, reprocess it to start the cycle all over again. It gives us power unprecedented in human history to understand and control our world. But, equally, it challenges our institutions, upsets our work habits and imposes unpredictable stresses upon our lives and societies.

Science is both producer and consumer of this data– and we urgently call on our political leaders to grasp the opportunities it creates. Success can create economic growth and a fairer, happier society. Failure will undermine Europe’s competitiveness and endanger social progress. Knowledge is power; Europe must manage the digital assets its researchers generate.

Science has a pivotal role in this phenomenon, and this report focuses on the infrastructure needed to manage scientific data. Our purpose is to provide a vision and action plan.

Why the focus on scientific data?

For starters, science is a cause of this data wave. Scientific discovery led to the microprocessors, optical fibres and storage media with which we create, move and store the data. And the continuing process of scientific discovery – in all disciplines from astronomy to economics – is generating a growing share of that new data. In one day, a high-throughput DNA-sequencing machine can read about 26 billion characters of the human genetic code. That translates into 9 terabytes – or 9 trillion data units – in the course of one year; alongside it is a wealth of related information that can be 20 times more voluminous. The total data flow: more than 20 new US Libraries of Congress each and every year. That is from one specialised instrument, in one scientific sub-discipline; enlarge that picture across all of science, across the world, and you start to see the dimension of the opportunity and challenge presented.

Most importantly, however, our focus is on scientific data because, when the information is so abundant, the very nature of research starts to change. A

feedback loop between researchers and research results changes the pace and direction of discovery. The “virtual lab” is already real, with the ability to undertake experiments on large instruments in other continents remotely in real time. Researchers with widely different backgrounds - from the humanities and social sciences to the physical, biological and engineering sciences – can collaborate on the same set of data from different perspectives. Indeed, we begin to see what some<sup>1</sup> have called a “fourth paradigm” of science – beyond observation, theory and simulation, and into a new realm of exploration driven by mining new insights from vast, diverse data sets. For the first time, large-scale and complex “whole body” solutions become possible for some of society’s Grand Challenges of energy and water supply, global warming, and healthcare.

Just how will we train people to work in this environment? What tools will we need to move, store, preserve and mine these data? How to share them? How to understand them, if you are in a different scientific discipline than that in which they were created? As a researcher, how will you know the data you access on another continent are accurate, uncorrupted and unbiased? What if those data include personal details – individual health records, financial information or Internet habits? These are just a few of the profound policy questions posed by this new age of data-intensive science.

Nowhere in the world are these questions adequately addressed. But we believe Europe has a special responsibility to lead, rather than to react, in this domain. The European Research Area – despite its oft-noted difficulties – remains today one of the top three scientific powers of the world, and if measured by the number of published scientific papers alone, it out-produces the US and Japan; it thus contributes more than its fair share to the scientific data tide. But that also means it has unique skills to address the challenges, through the strength of its best research institutions, the diversity of its technical talent, and the unique ability of its researchers to collaborate across borders, industries and disciplines.

Throughout human history, the interrelation between science and the technology for recording it has been deep and productive. In the ninth century, the spread of paper underpinned the Golden Age of Islamic science, as Greek

and Roman works of science were translated and then superseded. From the 15th century onward, the printing press permitted scholarship to travel far and wide – so a Copernicus could more easily influence a Galileo. In just the past 60 years we have seen information and communications technologies applied to such diverse fields as reaching the moon, harnessing nuclear energy and beginning to control cancer.

In this report we are not trying to second-guess the future; it will certainly be different from anything we can imagine now. But what we can do is to push for the difficult policy questions to be addressed, so that important options are not closed off and the science done today will be available to researchers tomorrow. We point to a pathway that is 'technology-neutral' – based on concepts broad enough to embrace whatever new forms of information and communications technologies we develop over the next generation. This requires developing principles for interoperability (technical, semantic, legal, and ethical), verification and reliability – at local, regional and global scale. It requires new incentives for sharing and protecting data of different types, whether that data is precious and guarded or abundant and open. And it requires a framework to review all these principles at regular intervals.

The European Union has an important, coordinating role in achieving this vision – through its Digital Agenda, its Framework Programme and the policies embodied in its European Research Area initiatives. Equally, there is the opportunity for the EU institutions to lead in creating a common, world-wide vision. The EU Competitiveness Council of late 2009<sup>2</sup> called on the European Commission to address the issue of e-infrastructure for science, and this High Level Group is part of that effort. As we publish this report, the product of six months of collective thought and research, we now call on the EU institutions to move beyond study and into action.

We are on the verge of a great new leap in scientific capability, fuelled by data. We have a vision of how Europe could benefit rather than suffer, lead rather than follow. But we urge speed. We must learn to ride the data wave.

**Keep constantly in mind in how many things you yourself have witnessed changes already.**

**The universe is change, life is understanding.**

*Marcus Aurelius,  
121-180*



## II. Welcome to the data world

**“We humans have built a creativity machine. It’s the sum of three things: a few hundred million computers, a communication system connecting those computers, and some millions of human beings using those computers and communications”**

Vernor Vinge<sup>3</sup>

**W**e live in the Information Age; and nowhere is that name more apt than in science and technology. Technical information in all forms, whether statistics, images, formulae or know-how in the broadest sense of the term, has already transformed our view of the world – and much more is yet to come. A few examples, to sketch out the possibilities ahead:

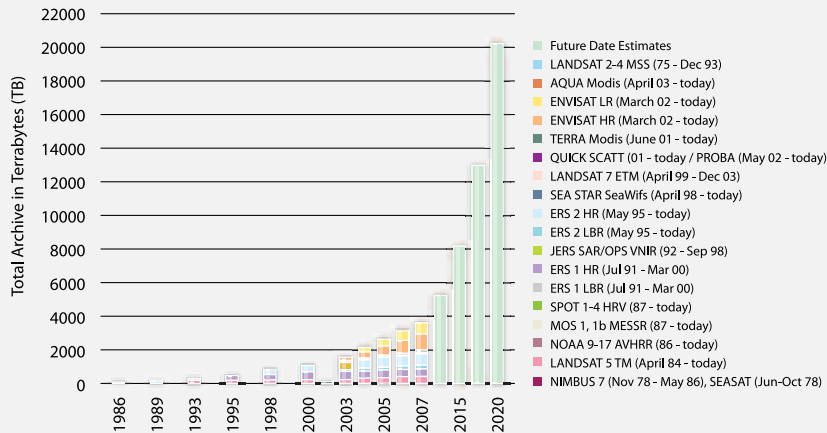
- Currently, about 2.5 petabytes – more than a million, billion data units – are stored away each year for mammogrammes in the US alone.<sup>4</sup> World-wide, some estimate, medical images of all kinds will soon amount to 30% of all data storage.<sup>5</sup> These could be a goldmine of data for epidemiological and drug research, if made accessible in appropriately anonymised form to researchers.
- ‘Smart meters’ for electricity consumption, now being installed in many EU countries, produce the equivalent of one CD-ROM of data for each household every year. Scale that up to 100 million households, and you have a vast repository of data for economic and behavioural analysis of people’s energy consumption.<sup>6</sup>
- Astronomy is a well-recognised ‘power user’ of data – but we are barely at the start of this trend. From 2020, the Square Kilometre Array, a new international radio telescope on the drawing board, could generate 1 petabyte of data every 20 seconds – a fire-hose of numbers requiring unimaginable processing power.<sup>7</sup> Yet that data will push the limits of the observable universe out by billions of galaxies, perhaps back to the first moments after the Big Bang.

- This century opened with the first “reading” of the human genome. By August 2009, digital records on more than 250 billion DNA bases, from various species, were stored in the US government’s public GenBank database<sup>8</sup> and an entirely new discipline of science had emerged: systems biology. This uses computers to simulate, at the sub-molecular level, exactly how DNA, proteins and the other chemical components of life interact – and in time, it will transform the practice of health sciences. “Organisms function in an integrated manner...but biologists have historically studied (them) part by part,” said Nobel Laureate David Baltimore. Systems biology “is a critical science of the future that seeks to understand the integration of the pieces to form biological systems.”<sup>9</sup>

As these examples suggest, the increase in scientific data isn’t simply a question of more information, more storage disks and more optical pipes to move it all – though that is certainly part of it. It is more profound than that: it changes the way we do our science, and opens entirely new fields of research.

And these new fields require, from the start, an international effort. One current project, 1000 Genomes<sup>10</sup>, is comparing the complete DNA sequences of more than 1,000 individuals from around the world to define what makes us different from one another – an inquiry with at least as many humanistic as scientific overtones. Geographical information systems, popularised in Google Maps, are changing the way we study economic, agricultural and demographic trends world-wide. And the global Internet offers an extraordinary new tool for behavioural research. Epidemiologists have studied the frequency with which people search online for keywords such as ‘flu’, as a way to monitor disease spread. Other researchers, trying to understand how people would react to pandemic alerts, have looked at the way online gamers in ‘World of Warcraft’ congregate around the digital equivalent of disaster zones, as a clue to new disease-control strategies.<sup>11</sup>

**Evolution of ESA's EO Data Archives between 1986-2007 and future estimates (up to 2020)**

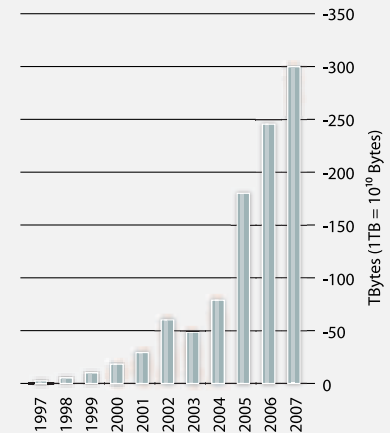


Left: After a long, slow rise since 1986, the volume of earth-observation data from the European Space Agency's satellites passed three petabytes in 2007 – three million, billion bytes. The projection for 2020: A seven-fold rise.

But it is not just what is studied, but also who studies it and how, that is affected by the data tide. For the first time in two centuries of growing professionalisation in science, new ICT tools allow the increasing involvement of amateurs; there are simply too many observations required for the professionals to do it all on their own. An example is GalaxyZoo, a Web portal through which amateurs help astronomers explore the Universe.<sup>12</sup> Biodiversity monitoring depends to a large extent on the power of observations by tens of thousands of volunteers who send their notes on species in defined areas to a central data repository.<sup>13</sup> For instance, the Swedish Species Gateway provides a navigable visual interface linked to geographical information systems.<sup>14</sup> These examples represent an important social and political trend. Empowering informed 'citizen-scientists' will also empower science.

Where will this lead? Consider five diverse scenarios that we believe to be entirely possible in coming decades.

**Yearly Data Creation on NICE**



Right: In Grenoble, the European Synchrotron Radiation Facility is a super-microscope studying anything from the propagation of cracks in steel to the surface proteins on the influenza virus. In the decade to 2007, its annual data output rose more than a hundred-fold. And it is just one of about 50 synchrotrons world-wide.

## SCENARIOS OF THE FUTURE



### Scenario I: Science and data management

Marie is working in a genomics project linking twelve large labs over four continents. The task of data management is intense. A group from China is preparing to feed its data into the consortium's processing pipeline. At the same time, the project's accounting system shows two research efforts elsewhere are behind schedule due to a microscope producing sub-standard data; the data flow from that machine is automatically blocked from the system. Some of Marie's graduate students, meanwhile, are trying to verify recently published results from a competing group of researchers - but their own lab equipment is different, so the work-plan needs to be modified. Next issue: A meeting with lawyers over a dispute between two of the research consortium's members. There is an argument over who controls what data before and after processing of the group's research results. Until it is settled, the data bank is holding all the files in escrow.



### Scenario II: Science and the citizen

Carlos likes bugs. He watches them the way a bird-lover tracks Canada geese – and he feeds his observations into a system for professionals to analyse. Out walking in a field one day, he spots something interesting – an insect species in an unexpected place. He queries the remote database for any relevant information that could explain it, and then checks in with his fellow amateurs. Their hypothesis: The insect may have changed its food preference to a different plant species. But why? Carlos posts his observation; and in the coming days other enthusiasts are watching for a similar anomaly. The system automatically analyses their observations, asks them more questions, and checks for incorrect information. It also looks for correlations with other databases – weather patterns, soil conditions, maps of flora distribution.

This wealth of observation allows a professional entomologist to test his hypothesis on 'preferred attractors in chaotic ecological processes.' Carlos is acknowledged in the resulting open-access publication.



### Scenario III: Science and the data set

Anneli has a grant under which she is allowed free access to 10 years of measurements by the global cell-phone sensor network, stored in cross-continental archives. This network uses the miniature sensors standard in cell phones to monitor local temperature, air quality, wind speed, light intensity, noise levels and other parameters – and links it to GPS information. The information is all kept in regional archives, with open interfaces so researchers can query them uniformly. With her team, Anneli wants to investigate correlations between the environment and the spread of illness – and for the disease information, she is looking at anonymised, geo-tagged messages sent by people mentioning the disease. She intends to clean the resulting data set and make it publicly available via her university's institutional repository. From there, it could become the scientific equivalent of a Top-40 song – played by others around the world. Her chances for tenure rise.



### Scenario IV: Science and the student

Roger is working on an international PhD. It's a relatively new programme, in which a student applies to become a member of an international team working on a big problem that affects all people. His group is comparing many forms of non-verbal communications between cultures. It has several hundred members and his university tutor is one of the nodal points contributing expertise in 'synergistic communication between biological components.' Others in the network are using archaeological evidence to study communications between ancient Mesopotamian and Hellenic cultures; some are studying computer-computer interactions between different systems; yet more are studying communications in refugee camps. Each node contributes to the whole. Results are communicated as they happen, and there are daily, virtual-presence planning sessions. Roger had to sign a contract not to misuse data or contribute anything that is not for the common good – such as externally sourced information that he has not thoroughly checked for provenance.





### Scenario V: Science and data-sharing incentives

Hans, rooting in the basement one day, finds an old laptop with a video of Grandpa on a boat. He is a young man in the video, wearing a diving costume. In the background is a marvellous beach. The video goes on to show underwater scenes with bizarre fish and colourful coral. The video is entertaining – but where was it made? Hans can get the answer in a few minutes. He goes online to a centralised mapping service, to which he uploads parts of the video. The service has smart pattern-matching algorithms, using huge reference collections. Soon, different mapping probabilities for the video fragments are returned, pointing out the most probable area where the video could have been made: The Maldives, before global warming drowned them. This is a bit of personal trivia for Hans, but a new data set for science. So there is a price for the service: Hans must let his video fragments stay in the central database, enriching it further and making it even more useful – for professional scientists, too.

Are these five scenarios fantastical? Not at all. There are already hundreds of projects, in the EU and elsewhere, that are precursors to the features described in these scenarios.

For example:

- The European Space Agency, recognising the importance of its satellite data for climate-change research, has launched a Long-Term Data Preservation programme that merges all earth observation data from across Europe<sup>15</sup>. At the same time, the EU's GENESI-DR project is creating a grid-based computing system for accessing and processing the huge amount of earth observation data which will become available. Both use fundamental results from the CASPAR EU project on how to preserve digitally encoded information.
- Humanities researchers are creating CLARIN, a system to establish an integrated and interoperable research infrastructure of language resources and tools. In doing so, they are already tackling proper data management as a key dimension of the system for the scholarly community.
- Astronomers around the world are creating the International Virtual Observatory to allow researchers everywhere to access and use data from hundreds of astronomical data sources. Also to be included are results of computer modelling and simulation – for it is not just raw observations that are the business of modern astronomy, but also the models built from them.
- As part of Framework Programme 7, the European Commission and EU member-states are investing in a broad range of e-infrastructure projects. The GEANT research-data network, for instance, connects over 40 million users, 8,000 institutions and 40 countries.<sup>16</sup> Other projects provide access to cooperative grid-computing platforms, develop supercomputing capacity, and lay the groundwork for the access and preservation of scientific information.

So, if this be dreaming, it is done with eyes wide open. But there remain many challenges to address, as well.

**The future of e-infrastructure for scientific data is bright - and already, extensive work is underway to make it a reality.**



### III. Facing up to the challenges

Certainly, creating a scientific world based on e-infrastructures will not be easy. For starters, it is technically difficult. The scale and complexity of this global scientific asset – with all its sensors, instruments, workstations and networks – are truly massive. There are many planning pitfalls, common to all large infrastructure projects. There can be ‘choke points’: technical or industrial problems that, if unrecognised, stop the show. People can get locked into sub-optimal technologies; think of the QWERTY/AZERTY computer keyboard, with its inefficient but now immutable layout. Gateways, originally created to join disparate systems, can later become barriers to progress in themselves. Short-term funding decisions can undermine the system’s longer-term development. What works best for a local user could hamper global functions.

The list of pitfalls is long. Success requires careful, coordinated and agile planning, on a global as well as EU level. E-infrastructure for science is one area where fragmentation of effort is more than inconvenient or inefficient; it is inimical. But the technical issues are only the beginning of the challenges to be overcome. Consider:

- How will we preserve the data? As we all have seen, the media in which we store information change constantly – from magnetic coils, to tape, to disk, to USB key, to ‘cloud’ storage, and so on in an endless chain of invention and obsolescence. What will be the point of storing all this scientific data if, a century from now, it has degraded, been corrupted, or is simply too difficult for anyone but a well-equipped expert to use?
- How will we protect the integrity of the data? Even today, it is easy for a determined individual to alter or corrupt digital data (think of the constant controversy over Wikipedia entries.) As the data tide rises higher, how will we detect unauthorised alterations? Should every researcher, and indeed every citizen, have access to the data repositories? Should there be different levels of access allowed?
- How will we convey the context and provenance of the data? Given the emerging trend to make all publicly funded research data publicly available, just how will users from a wide range of backgrounds understand and query

the data they are accessing, and recognise the special circumstances under which it was collected? Already, in medical research, potentially fatal errors can arise by researchers inadvertently misinterpreting the drug-trial data collected by others; so-called 'meta-analysis,' to manage such complexities, is far from a certain science.

- How will we pay for all this? What new funding and business models will we need, so that everyone – researchers, enterprises, citizens – have adequate incentive to contribute to the data infrastructure? What kinds of data, under what circumstances, should be free?
- How will we protect the privacy of individuals linked to the data? We have already seen how easy it is for supposedly safeguarded data – whether tax files or health records – to be lost or misused. On one hand, access to this data is vital to researchers studying the economy or public health. On the other hand, carelessness in handling the data compromises our safety and security. How will we resolve this paradox?

Many of these issues involve trust. Data-intensive science operates at a distance and in a distributed way, often among people who have never met, never spoken, and, sometimes, never communicated directly in any form whatsoever. They must share results, opinions and data as if they were in the same room. But in truth, they have no real way of knowing for sure if, on the other end of the line, they will find man or machine, collaborator or competitor, reliable partner or con-artist, careful archivist or data slob. And those problems concern merely the scientific community; what about when we add a wider population? Many fields require the public to cooperate in supplying data (wittingly or not). How will we judge the reliability and authenticity of data that moves from a personal archive into a common scientific repository? If science is to advance, all these questions of trust must be answered by the infrastructure, itself.

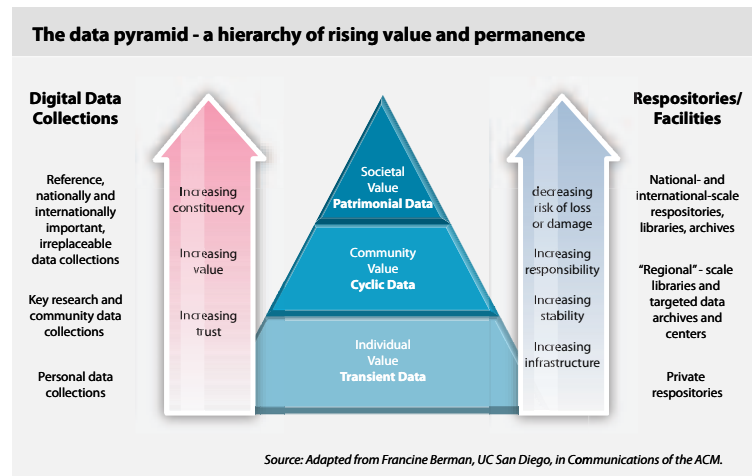
In dealing with many of these issues, we believe a few broad principles arise:

#### **Data as infrastructure**

Our stock of intangible knowledge, expanding at today's hyper-speeds, needs to be thought of as a new kind of asset in itself, that serves all. As such, it requires

professional analysis and engineering. Its contents are heterogeneous – different data formats, value and uses. There is tremendous value in having the data made seamlessly available, to use, reuse and recombine to support the creation of new knowledge. And the data must be available to whomever, whenever and wherever needed, yet still be protected if necessary by a range of constraints including by-attribution licenses, commercial license, time embargos, or institutional affiliation.

A data pyramid (below) suggests the complex data ecology. At the bottom of the pyramid lie the most abundant, transient forms of data – billions of personal data files across the planet, on private disks and storage services, of obvious value only to the few who create or use them. At the top of the pyramid is patrimonial data – high-value, irreplaceable data of importance to an entire nation or society, redundantly stored in national or international trusted archives. In the middle is cyclic data – a mid-range of data created and used in a specific task, community or region. The new data infrastructure must cope with all these data classes.



### Interoperability

Diversity is a dominant feature of scientific information – diversity of data formats and types, but also of the people and communities that generate and use the data. Even within the same scientific community, there are different points of view, different ways of analysing, sharing and handling data. There is also diversity in how the data are stored, categorised and mapped. There is diversity in who can access what kinds of data, and how – from tightly protected military satellite images to freely accessible Google Earth views. And as science advances, diversity is bound to increase.

Achieving an interoperable data infrastructure in the midst of such heterogeneity is a significant challenge. None of the potential benefits of the scientific data wave will be harnessed unless – given the proper access rights – it is easy and cheap to rummage through relevant data files anywhere in the world, in any field. An epidemiologist in Geneva studying the latest flu virus will benefit greatly from being able to tap easily into DNA databases in London of 1918 Spanish Flu victims – and the epidemiologist's work should be accessible to a public health official in Hong Kong, a systems biologist in San Diego and a medical historian in Boston. That's all possible today, but with great effort, skill, cost and time. A leap forward in interoperability will change that.

### Incentives

How can we get researchers – or individuals – to contribute to the global data set? Only if the data infrastructure becomes representative of the work of all researchers will it be useful; and for that, a great many scientists and citizens will have to decide it is worth their while to share their data, within the constraints they set. To start with, this will require that they trust the system to preserve, protect and manage access to their data; an incentive can be the hope of gain from others' data, without fear of losing their own data. But for more valuable information, more direct incentives will be needed – from career advancement, to reputation to cash. Devising the right incentives will force changes in how our universities are governed and companies organised. This is social engineering, not to be undertaken haphazardly.

### Financial models

All of this costs money – so who pays, and how? To a considerable extent, scientific e-infrastructure represents a public good. It is vital that governments and taxpayers step in to provide the critical funding in those instances. Our data future will look bleak if the public sector under-invests. Of course, there is private

## Scientific e-infrastructure – a wish list

**The ideal data infrastructure for science will have a long list of technical characteristics. Here are some suggestions.**

- Open deposit, allowing user-community centres to store data easily
- Bit-stream preservation, ensuring that data authenticity will be guaranteed for a specified number of years
- Format and content migration, executing CPU-intensive transformations on large data sets at the command of the communities
- Persistent identification, allowing data centres to register a huge amount of markers to track the origins and characteristics of the information
- Metadata support to allow effective management, use and understanding
- Maintaining proper access rights as the basis of all trust
- A variety of access and curation services that will vary between scientific disciplines and over time
- Execution services that allow a large group of researchers to operate on the stored data
- High reliability, so researchers can count on its availability
- Regular quality assessment to ensure adherence to all agreements
- Distributed and collaborative authentication, authorisation and accounting
- A high degree of interoperability at format and semantic level

*Adapted from the PARADE White Paper at <http://www.csc.fi/english/pages/parade/>*

gain as well. When a government laboratory contributes its raw research data to the global e-infrastructure, it is certainly saving private users the expense of running those experiments on their own. Equally, when a private company contributes its own files to the system, it also helps the public researchers. It is important to devise funding mechanisms that enable all to contribute as well as to benefit, through an increased return on investment.

These issues can be resolved. We have experience of past changes in how we store, share and manage valuable assets. As the technology of food and transport evolved, society moved from self-supporting farmers to town markets, and from markets to a range of supermarkets and specialty shops. In finance, we moved from private hoards to communal banks to international markets. The same path from individual control to international exchange must be trodden by data – indeed, it is already happening.

**It is important to devise funding mechanisms that enable all to contribute as well as to benefit, through an increased return on investment.**

## Scientific e-infrastructure – some challenges to overcome

---

### Collection

How can we make sure that data are collected together with the information necessary to re-use them?

### Trust

How can we make informed judgements about whether certain data are authentic and can be trusted?

How can we judge which repositories we can trust? How can appropriate access and use of resources be granted or controlled?

### Usability

How can we move to a situation where non-specialists can overcome the high barriers to their being able to start sensible work on unfamiliar data, perhaps using intelligent automated tools for an initial investigation?

### Interoperability

How can we implement interoperability within disciplines and move to an overarching multi-disciplinary way of understanding and using data?

How can we find unfamiliar but relevant data resources beyond simple keyword searches, but involving a deeper probing into the data?

How can automated tools find the information needed to tackle unfamiliar data?

### Diversity

How do we overcome the problems of diversity – heterogeneity of data, but also of backgrounds and data-sharing cultures in the scientific community?

How do we deal with the diversity of data repositories and access rules – within or between disciplines, and within or across national borders?

### Security

How can we guarantee data integrity?

How can we avoid data poisoning by individuals or groups intending to bias them in their interest?

How can we react in the case of security breaches to limit their impact?



## Scientific e-infrastructure – some challenges to overcome *continued*

---

### **Education and training**

How can the citizen make these benefits available for sensible investigations, and how can they be safeguarded from fakes?

How can scientific e-infrastructure foster and increase popular interest and trust in science?

How can we foster the training of more data scientists and data librarians, as important professions in their own right?

### **Data publication and access**

How can data producers be rewarded for publishing data?

How can we know who has deposited what data and who is re-using them – or who has the right to access data which are restricted in some way?

How do we deal with the various ‘filters’ that different disciplines use when choosing and describing data? What about differences in these attitudes within disciplines, or from one time to another?

### **Commercial exploitation**

How can the infrastructure benefit from commercial developments in data management?

How can the revenue-generating expertise of the commercial world be brought into play for the long-term sustainability of these resources?

### **New social paradigms**

How can we learn from the wisdom of crowds about what and whom to trust, while avoiding being misled by concerted campaigns of deceit?

### **Preservation and Sustainability**

How can we be sure that the important information we collect will be usable and understandable in the future; in particular how can we fund our information resources in the long term?

How can we share the costs and efforts required for sustainability?

How can we decide what to preserve?



## IV. A vision for 2030

The creation of scientific e-infrastructure is a means, not an end. It is a means to new science, new solutions and new progress in society. We cannot predict what the world will be like in 2030, but we can state some broad principles of what it should be like if scientific e-infrastructure is by then the major contributor to society, the economy and science that we expect it to be. All of these principles – our vision – point in the direction of an infrastructure that supports seamless access, use, reuse, and trust of data. It suggests a future in which the data infrastructure becomes invisible, and the data themselves have become infrastructure – a valuable asset, on which science, technology, the economy and society can advance. We will know we are well on our way to realising this vision when we see the following milestones achieved:

### **1. All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.**

This may sound obvious – but it is by no means so.<sup>17</sup> Today we see the relative priorities of society in constant flux on the Internet and other electronic media. In a world of limited resources, how urgent is a packet of scientific data compared to home videos? How much is it worth to create reliable back-up and storage systems for what may seem today like transient chat messages, but could tomorrow become vital behavioural or epidemiological data? Thus, the first task is simply to get the message out that scientific e-infrastructure is important to society.

**Expected impact:** The intellectual capital of Europe is used to generate economic and scientific advances now, and that capital is safely preserved for further exploitation by future generations.

**Risk of Inaction:** Resources for funding take a back seat to more pressing concerns, and data decays through neglect. When critical data – whether about climate, new medicines or historic monuments – are needed later on, it will be too late.

## **2. Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.**

**Expected impact:** Researchers can today access online sources, but it is a small fraction of all data produced. In future, the breadth and depth of data available to them will grow dramatically, whether their discipline is demographics, ocean chemistry, high-energy physics or astronomy. Scientists' efficiency and productivity will rise because they know they can access, use, reuse and trust the data they find. Inspiration or serendipity can lead to unexpected results. Cross-fertilisation of ideas and disciplines will produce novel solutions, and promote greater understanding of complex problems.

**Risk of Inaction:** As the volume and diversity of scientific data increase, and as research becomes more multi-disciplinary, researchers struggle to understand and correlate data – especially if from another field. They may not find the data at all. Or if they find it, they are not sure it is what it claims to be. As a result, researchers become increasingly isolated, narrow specialists; wide-ranging, serendipitous results become more difficult.

## **3. Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories is guided by international standards, to ensure they are trustworthy.**

**Expected impact:** Researchers are rewarded, by enhanced professional reputation at the very least, for making their data available to others. Confidence that their data cannot be corrupted or lost reassures them to share even more. Data sharing, with appropriate access control, is the rule, not the exception. Data are peer-reviewed by the community of researchers re-using and re-validating them. The outcome: A data-rich society with information that can be used for new and unexpected purposes.

**Risk of Inaction:** Information stays hidden. The researcher who created it in the hope it can yield more publications or patents in the future holds on to it. Other researchers who need that information are unable to get at it, or waste time re-creating it. The outcome: A world of fragmented data sources – in fact, a world much like today.

#### **4. Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of publicly generated data.**

**Expected impact:** Research productivity rises, through easy access and re-use of data. Funders take a strategic view of the value of data – and plan investments logically and consistently. R&D activity grows globally. New and unexpected solutions emerge to our major societal challenges.

**Risk of Inaction:** The public sector unnecessarily spends money on producing data over and over again, because they are lost or cannot be found. Data that are of the greatest value to the public (of a “public goods” nature) are a special loss. Researchers overlook important insights, because they cannot access or understand potentially vital data from others around the world. Opportunities for progress and prosperity are missed. Investment slows.

#### **5. The innovative power of industry and enterprise is harnessed by clear and efficient arrangements for exchange of data between private and public sectors, allowing appropriate returns to both.**

**Expected impact:** Data generated for one purpose are re-used for others, and the pace of innovation – social and technological – rises. Commercial research capability is strengthened by public research, and broad expertise is harnessed to the benefit of all. Mobility and cross-fertilisation between the commercial and academic sectors increase, amplifying the impact of innovation and new discoveries. New companies, jobs and fortunes result. European industry is more competitive.

**Risk of Inaction:** Suspicion and adversarial attitudes develop between private and public sectors. A vicious circle sets in of ivory-tower academics and under-investing industrialists. Europe’s competitiveness suffers.

**6. The public has access to and can make creative use of the huge amount of data available; it can also contribute to the data store and enrich it. Citizens can be adequately educated and prepared to benefit from this abundance of information.**

**Expected impact:** Citizens can share and contribute to the scientific process. They understand the benefits and risks of new technologies better, and more rational political decisions emerge. The young are inspired by an ambition for new discoveries, and join the ranks of scientists and engineers in far-greater numbers.

**Risk of Inaction:** Citizens feel increasingly distrustful of and isolated from science, and resistant to technology. They are easily misled by pseudo-science and political demagoguery. The supply of engineers and scientists is inadequate to society's needs.

**7. Policy makers are able to make decisions based on solid evidence, and can monitor the impacts of these decisions. Government becomes more trustworthy.**

**Expected impact:** Policy decisions improve, and public confidence in the entire political process rises. It is possible to correct policy mistakes, whether economic or social, in real time. People gain confidence in government, and political participation rises.

**Risk of Inaction:** Ill-informed political decisions lead to bad results, and our economic, environmental and social problems mount. Citizens lose confidence in their leaders. An impenetrable wall of data separates the governors from the governed.

**8. Global governance promotes international trust and interoperability.**

**Expected impact:** Citizens have access to the world's store of information without unnecessary boundaries. A framework for global interoperability maintains a common, public space for scientific data. This instils trust and ensures that the best minds can make use of information no matter where they are. World trade grows, and societies prosper.

**Risk of Inaction:** The divide between the information-rich and the information-poor grows. Some of the best minds are isolated, and new ideas go un-exploited. The world is a poorer place.

## Who benefits from scientific e-infrastructure

Beneficiaries	Benefits
<b>Citizens</b>	<p>Appreciate the results and benefits arising from research and feel more confident in how their tax money is spent</p> <p>Find their own answers to important questions, based on real evidence</p> <p>Pass on knowledge and experience to others, and make a contribution to the knowledge society beyond their immediate circle and life-spans</p>
<b>Funders and Policy Makers</b>	<p>Make evidence-based decisions</p> <p>Eliminate unnecessary duplication of work</p> <p>Get greater return on investment</p>
<b>Researchers</b>	<p>Have all data and tools easily available, increasing productivity</p> <p>Cross disciplinary boundaries, gaining new insights and producing new solutions</p> <p>'Stand on the shoulders of giants'</p>
<b>Enterprise and Industry</b>	<p>Use the best available information for R&amp;D, increasing productivity</p> <p>Create new knowledge, markets and job opportunities</p> <p>Provide a strong industrial and economic base for European prosperity</p> <p>Increase opportunities for mobility and knowledge exchange</p>



## V. A call to action

The scientific and social benefits of our vision are numerous. But there are many other practical reasons to act. ICT is one of the main engines of economic growth. It is to our age what paved highways, national railroads and inter-continental telegraphs were to earlier generations. Yet in Europe, the industry underpinning this vital economic activity has had many difficulties. And, as the European Competitiveness Council has noted, “the ICT impact on productivity growth is lower in the EU than in major trading partners.”<sup>18</sup> A concerted European effort to build e-infrastructures for science will stimulate market demand for ICT. It will pull the best from ICT researchers, engineers and industrialists, spurring growth and jobs. And it will pave the digital highways that European science will need.

There is a clear role for government in all this. We urge our leaders to take into consideration the following:

- A good framework for the governance of data will be a source of strength in the most knowledge-intensive industries, fostering the growth of companies, goods, and services with the highest value-added. Those regions of the world that lead this policy debate, and develop the technologies and industry to support it, will gain competitive advantage.
- Scientific e-infrastructure is essential if we are to address the Grand Challenges of today. Understanding climate change, finding alternative energy sources, and preserving the health of an ageing population are all fiendishly complex, cross-disciplinary problems that require high-performance data storage, smart analytics, transmission and mining to solve.
- Social cohesion will depend in part on how fairly and openly knowledge and information flow within our region and between the public and private sector. If information is power in the knowledge economy, governments must ensure that the benefits are appropriately distributed. Governments must work effectively through public-private partnerships to develop e-infrastructure.
- International collaboration is essential; there is no such thing as a purely local or national network anymore. We must collaborate in global architectures and governance for e-infrastructure, and we must share costs and

technologies for archiving, networking and managing data across the globe.

With this preamble, we offer a short-list of action by various EU institutions. Of course, we recognise there has already been much work done in the field. The Commission has funded several projects to develop distributed computing environments, databases for discipline-specific content, and libraries for new types of online communications. There has been much debate – from the Commission, the Council and the Parliament – about the need to speed development of scientific e-infrastructure. And we note that many other public bodies have begun considering these matters: For instance, the group reporting to the US Office of Science and Technology Policy recently published its own agenda and recommendations for ensuring long-term access to digital information.<sup>19</sup> But more, urgent, concrete action is needed from all parties, we believe. First steps include:

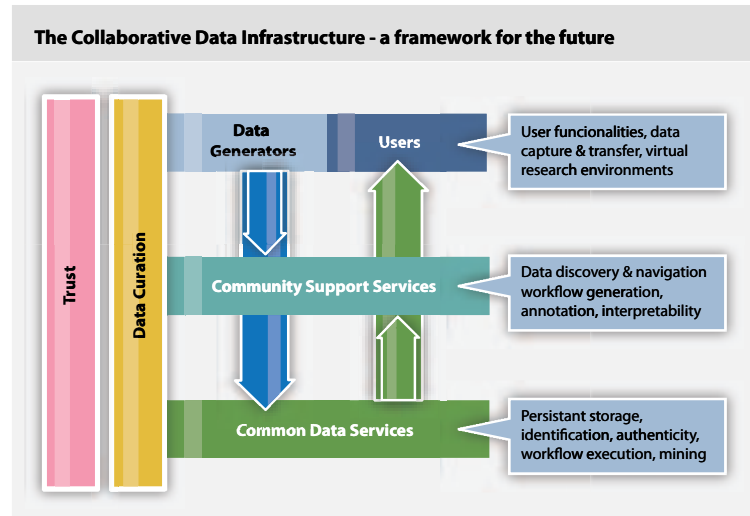
### **1. Develop an international framework for a Collaborative Data Infrastructure**

The emerging infrastructure for scientific data must be flexible but reliable, secure yet open, local and global, affordable yet high-performance. Obviously, this is a tall order – and there is no one technology that we know today or can imagine tomorrow to achieve it all. Thus, what is needed is a broad, conceptual framework for how different companies, institutes, universities, governments and individuals would interact with the system – what types of data, privileges, authentication or performance metrics should be planned. This framework would ensure the trustworthiness of data, provide for its curation, and permit an easy interchange among the generators and users of data. For the sake of illustration, we outline below the broadest building blocks of such a framework.

**The Commission has funded several projects to develop distributed computing environments, databases for discipline-specific content, and libraries for new types of online communications.**



This figure suggests, in the broadest possible terms, how different actors, data types and services should interrelate in a global e-infrastructure for science. Data generators and users gather, capture, transfer and process data - often, across the globe, in virtual research environments. They draw upon support services in their specific scientific communities - tools to help them find remote data, work with it, annotate it or interpret it. The support services, specific to each scientific domain and provided by institutes or companies, draw on a broad set of common data services that cut across the global system; these include systems to store and identify data, authenticate it, execute tasks, and mine it for unexpected insights. At every layer in the system, there are appropriate provisions to curate data - and to ensure its trustworthiness.



This Collaborative Data Infrastructure is a map to be filled in by thousands of different actors across the globe and over many years. But we call upon the European Commission to accelerate efforts to make this map. And it should consider requiring that all relevant EU research projects should, when it comes to considering their data management, fit into such a framework.

## 2. Earmark additional funds for scientific e-infrastructure

This is expensive. And as e-infrastructure for scientific data has a public dimension, so it should also have appropriate public funding. There are several possible funding sources – including some ideally suited for major infrastructure projects of this sort. The EU's Structural Funds are already used to build new schools, roads, industrial parks and other key infrastructure, targeted at those regions of Europe most in need. Already, a portion of these Structural Funds are earmarked for research and innovation. This need, for data generation and maintenance, cuts across that part of the budget – and all EU programmes, innovation-related or not. We call upon the European Council to increase the amount spent specifically on e-infrastructure for scientific data.

### **3. Develop and use new ways to measure data value, and reward those who contribute it**

Who contributes the most or best to the data commons? Who uses the most? What is the most valuable kind of data – and to whom? How efficiently is the data infrastructure being used and maintained? These are all measurement questions. At present, we have lots of different ways of answering them – but we need better, more universal metrics. If we had them, funding agencies would know what they are getting for their money – who is using it wisely. Researchers would know the most efficient pathways to get whatever information they are seeking. Companies would be able to charge more easily for their services. We urge the European Commission to lead the study of how to create meaningful metrics, in collaboration with the ‘power users’ in industry and academia, and in cooperation with international bodies.

### **4. Train a new generation of data scientists, and broaden public understanding**

Achieving all this requires a change of culture – a new way of thinking about when you share information, how you describe or annotate it for re-use, when you hide it or protect it, when you charge for it or give it away. It requires new knowledge about how researchers use and re-use information, in different disciplines and countries. We urge that the European Commission promote, and the member-states adopt, new policies to foster the development of advanced-degree programmes at our major universities for this emerging field of data science. We also urge the member-states to include data management and governance considerations in the curricula of their secondary schools, as part of the IT familiarisation programmes that are becoming common in European education.

### **5. Create incentives for green technologies in the data infrastructure**

Computers burn energy – vast quantities of it. Data centres absorb about 2% of world electricity production. Computer assembly also consumes precious minerals, lots of fresh water and adds to CO<sub>2</sub> production. Clearly, as hardware components multiply into the trillions, environmental constraints will tighten. So

the ICT industry must be incentivised to change its production and distribution methods, to go greener. But the issue goes beyond hardware. When a researcher makes a copy of a data set, he or she consumes resources – virtual though the action may seem. Indeed, basic information theory tells us, whenever we bring order to information we are adding to its energy. This fact must be understood, and factored into our broader environmental policies. We urge the European institutions, as they review plans for CO2 management and energy efficiency, to consider the impact of e-infrastructure and prepare policies now that will ensure we have the necessary resources to perform science.

## **6. Establish a high-level, inter-ministerial group on a global level to plan for data infrastructure**

As stated previously, it makes no sense for one country or region to act alone. Interoperability requires that there be reciprocal agreements between governments – the digital equivalent of trade treaties. There must also be agreement that all countries contribute, according to their usage and needs, to the global effort; free riders can endanger the system. We urge the European Commission to identify a group of international representatives who could meet regularly to discuss the global governance of scientific e-infrastructure. It should also host the first such meeting.

There are many other actions we believe essential to the development of e-infrastructure for science; we detail more in the Annex, and provide a list of potential ‘show-stoppers’ that will need attention. We believe that we all benefit from a far-seeing, collaborative and open approach to science and the e-infrastructure to support it. We urge action now.

**We believe that we all benefit from a far-seeing, collaborative and open approach to science and the e-infrastructure to support it. We urge action now.**

## The 2030 Vision – and the recommendations

Vision	Summary Recommendations	Impact if achieved
<p><b>All stakeholders, from scientists to national authorities to the general public, are aware of the critical importance of conserving and sharing reliable data produced during the scientific process.</b></p>	<p>All member states ought to publish their policies and implementation plans on the conservation and sharing of scientific data, aiming at a coordinated European approach.</p> <p>Legal issues are worked out so that they encourage, and not impede, global data sharing.</p> <p>The scientific community is supported to provide its data and metadata for re-use.</p> <p>Every funded science project includes a fixed budget percentage for compulsory conservation and distribution of data, spent depending on the project context.</p>	<p>Data form an infrastructure, and are an asset for future science and the economy.</p>
<p><b>Researchers and practitioners from any discipline are able to find, access and process the data they need. They can be confident in their ability to use and understand data, and they can evaluate the degree to which that data can be trusted.</b></p>	<p>Create a robust, reliable, flexible, green, evolvable data framework with appropriate governance and long-term funding schemes to key services such as Persistent Identification and registries of metadata.</p> <p>Propose a directive demanding that data descriptions and provenance are associated with public (and other) data.</p> <p>Create a directive to set up a unified authentication and authorisation system.</p> <p>Set Grand Challenges to aggregate domains.</p> <p>Provide “forums” to define strategies at disciplinary and cross-disciplinary levels for metadata definition.</p> <p>Work closely with real users and build according to their requirements.</p>	<p>Dramatic progress in the efficiency of the scientific process, and rapid advances in our understanding of our complex world, enabling the best brains to thrive wherever they are.</p>
<p><b>Producers of data benefit from opening it to broad access, and prefer to deposit their data with confidence in reliable repositories. A framework of repositories is guided by international standards, to ensure they are trustworthy.</b></p>	<p>Propose reliable metrics to assess the quality and impact of datasets. All agencies should recognise high quality data publication in career advancement.</p> <p>Create instruments so long-term (rolling) EU and national funding is available for the maintenance and curation of significant datasets.</p> <p>Help create and support international audit and certification processes.</p> <p>Link funding of repositories at EU and national level to their evaluation.</p> <p>Create the discipline of data scientist, to ensure curation and quality in all aspects of the system.</p>	<p>Data-rich society with information that can be used for new and unexpected purposes.</p> <p>Trustworthy information is useable now and for future generations.</p>

Vision	Summary Recommendations	Impact if achieved
<b>Public funding rises, because funding bodies have confidence that their investments in research are paying back extra dividends to society, through increased use and re-use of publicly generated data.</b>	EU and national agencies mandate that data management plans be created.	Funders have a strategic view of the value of data produced.
<b>The innovative power of industry and enterprise is harnessed by clear and efficient arrangements for exchange of data between private and public sectors, allowing appropriate returns to both.</b>	<p>Use the power of EU-wide procurement to stimulate more commercial offerings and partnerships.</p> <p>Create better collaborative models and incentives for the private sector to invest and work with science for the benefit of all.</p> <p>Create improved mobility and exchange opportunities.</p>	Commercial expertise is harnessed to the public benefit in a healthy economy.
<b>The public has access to and can make creative use of the huge amount of data available to them; it can also contribute to it and enrich it. Citizens can be adequately educated and prepared to benefit from this abundance of information.</b>	<p>Create non-specialist as well as specialist data access, visualisation, mining and research environments.</p> <p>Create annotation services to collect views and derived results.</p> <p>Create data recommender systems.</p> <p>Embed data science in all training and academic qualifications.</p> <p>Integrate into gaming and social networks.</p>	Citizens get a better awareness of and confidence in sciences, and can play an active role in evidence-based decision making and can question statements made in the media.
<b>Policy makers are able to make decisions based on solid evidence, and can monitor the impacts of these decisions. Government becomes more trustworthy.</b>	Propose a directive to ensure that public data is available (with security where applicable).	Policy decisions are evidence-based to bridge the gap between society and decision-making, and increase public confidence in political decisions.
<b>Global governance promotes international trust and interoperability</b>	<p>Member states should publish their strategy, and resources, for implementation, by 2015.</p> <p>Create a European framework for certification for those coming up to an appropriate level of interoperability.</p> <p>Create a “scientific Davos” meeting to bring commercial and scientific domains together.</p>	We avoid fragmentation of data and resources.

## What could jeopardise the vision?

Impediments	What we could do to overcome them
Lack of long term investment in critical components such as persistent identification	Identify new funding mechanisms Identify new sources of funding Identify risks and benefits associated with digitally encoded information
Lack of preparation	Ensure the required research is done in advance
Lack of willingness to co-operate across disciplines/ funders/ nations	Apply subsidiarity principle so we do not step on researchers' toes Take advantage of growing need of integration: within and across disciplines
Lack of published data	Provide ways for data producers to benefit from publishing their data
Lack of trust	Need ways of managing reputations Need ways of auditing and certifying repositories Need quality, impact, and trust metrics for datasets
Not enough data experts	Need to train data scientists and to make researchers aware of the importance of sharing their data
The infrastructure is not used	Work closely with real users and build according to their requirements Make data use interesting – for example integrating into games Use “data recommender” systems i.e. “you may also be interested in...”
Too complex to work	Do not aim for a single top down system Ensure effective governance and maintenance system (c.f. IETF)
Lack of coherent data description allowing re-use of data	Provide “forums” to define strategies at disciplinary and cross-disciplinary levels for metadata definition



# About the High Level Group

The High Level Expert Group on Scientific Data was charged by the European Commission's Directorate-General for Information Society and Media to prepare a "vision 2030" for the evolution of e-infrastructure to scientific data.

Chair: **John Wood**, Secretary General of the Association of Commonwealth Universities

**Thomas Andersson**, Professor of Economics and former President, Jönköping University; Senior Advisor, Science, Technology and Innovation, Sultanate of Oman

**Achim Bachem**, Chairman, Board of Directors, Forschungszentrum Jülich GmbH

**Christoph Best**, European Bioinformatics Institute, Cambridge (UK)/Google UK Ltd, London (from September 2010).

**Françoise Genova**, Director, Strasbourg astronomical Data Centre; Observatoire Astronomique de Strasbourg, Université de Strasbourg/CNRS

**Diego R. Lopez**, RedIRIS

**Wouter Los**, Faculty of Science at the University of Amsterdam; Coordinator of preparatory project LifeWatch biodiversity research infrastructure; Vice Chair Governing Board of GBIF

**Monica Marinucci**, Director, Oracle Public Sector, Education and Research Business Unit

**Laurent Romary**, INRIA and Humboldt University

**Herbert Van de Sompel**, Staff Scientist, Los Alamos National Laboratory

**Jens Vigen**, Head Librarian, European Organization for Nuclear Research, CERN

**Peter Wittenburg**, Technical Director, Max Planck Institute for Psycholinguistics

Rapporteur: **David Giarretta**, STFC and Alliance for Permanent Access

Report Text: **Richard L. Hudson**, Science|Business

The HLG wishes to acknowledge the following individuals for their invaluable contribution to the discussions: Mirko Albani, Peter Doorn, Fabrizio Gagliardi, Daron Green, István Kenesei, Puneet Kishor, Kimmo Koski, Norbert Lossau, Linda Miller, Bernd Panzer-Steindel, Günter Stock, Ilkka Tuomi.

Design: Design4Science Ltd. Illustrations: Fletcher Ward Design

## REFERENCES

- <sup>1</sup> Hey, Tony; Stewart Tansley and Kristin Tolle, Eds. "The Fourth Paradigm: Data-Intensive Scientific Discovery." Microsoft Research. Redmond, Wash: 2009. PDF at <http://research.microsoft.com/enus/collaboration/fourthparadigm/>
  - <sup>2</sup> Council of the European Union. "The future of ICT research, innovation and infrastructures - Adoption of Council Conclusions." 25 November 2009.
  - <sup>3</sup> Vinge, V. "The Creativity Machine". *Nature*, Vol. 440. March 2006.
  - <sup>4</sup> Hey, A.F.G. and A.E.Trefethen, in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, G.C. Fox, A.J.G. Hey, Eds. Wiley, Hoboken, NJ, 2003.
  - <sup>5,6</sup> Beyea, Jan. "The Smart Electricity Grid and Scientific Research," *Science* 328: 979, 21 May 2010.
  - <sup>7</sup> "The Square Kilometre Array: Factsheet for Scientists and Engineers." SKA Program Development Office, April 2010. [http://www.skatelescope.org/PDF/100420\\_SKA\\_Factsheet-Scientists-Engineers.pdf](http://www.skatelescope.org/PDF/100420_SKA_Factsheet-Scientists-Engineers.pdf)
  - <sup>8</sup> National Centre for Biotechnology Information. "What is GenBank?" <http://www.ncbi.nlm.nih.gov/genbank/>
  - <sup>9</sup> Institute for Systems Biology. "Systems biology – the 21<sup>st</sup> century science." <http://www.systemsbiology.org>.
  - <sup>10</sup> The 1000 Genomes Project. <http://www.1000genomes.org/>
  - <sup>11</sup> Lofgren, Eric T. and Nina H. Fefferman. "The untapped potential of virtual game worlds to shed light on real world epidemics." *The Lancet Infectious Diseases*, VII:9 (625 – 629), September 2007.
  - <sup>12</sup> <http://www.galaxyzoo.org/>
  - <sup>13</sup> Irwin, A. "Constructing the Scientific Citizen: Science & Democracy in the Biosciences," *Public Understanding of Science* vol.10, pp.1-18 (2001).
  - <sup>14</sup> <http://www.artportalen.se>
  - <sup>15</sup> <http://earth.esa.int/gscb/tpd>
  - <sup>16</sup> <http://www.geant.net>
  - <sup>17</sup> Survey results from the PARSE.Insight project (<http://www.parse-insight.eu/>) show the lack of awareness of preservation and reluctance to share data.
  - <sup>18</sup> Council of the European Union. *Ibid.*
  - <sup>19</sup> Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information." February 2010. [http://brtf.sds.c.edu/biblio/BRTF\\_Final\\_Report.pdf](http://brtf.sds.c.edu/biblio/BRTF_Final_Report.pdf)
- For the 'Data Pyramid' graphic on page 18, the HLG wishes to acknowledge: Berman, F. 2008. "Got data? A guide to data preservation in the information age." *Communications of the ACM* 51, 12 (Dec. 2008), 50-56. <http://doi.acm.org/10.1145/1409360.1409376>

The High Level Expert Group on Scientific Data was charged by the European Commission's Directorate-General for Information Society and Media to prepare a "vision 2030" for the evolution of e-infrastructure for scientific data. After meetings and consultations from December 2009 through June 2010, the group presents its outlook and recommendations.

