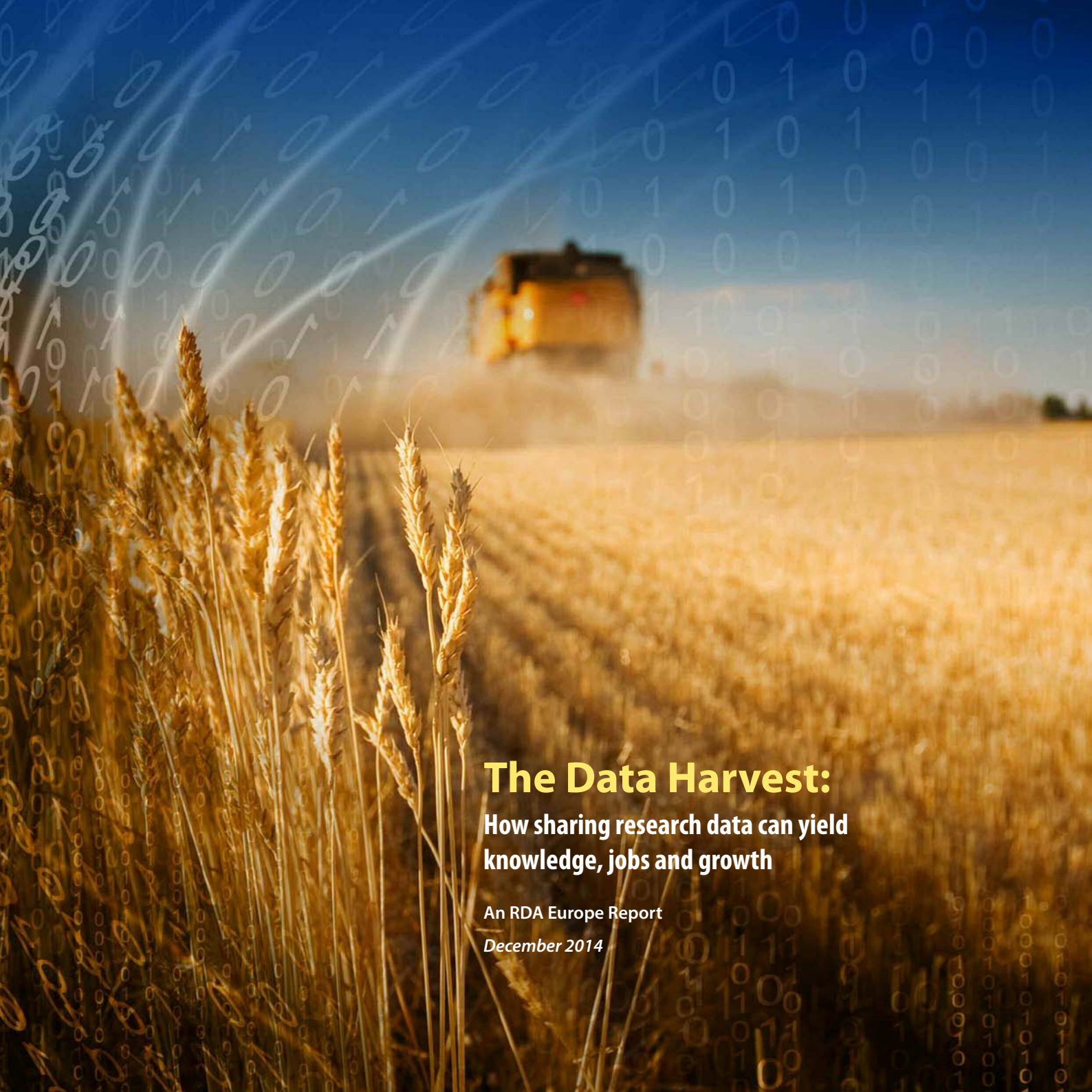


# The Data Harvest:

## How sharing research data can yield knowledge, jobs and growth

An RDA Europe Report

*December 2014*

# The Data Harvest

## How sharing research data can yield knowledge, jobs and growth

**A Special Report by RDA Europe**


**The Data Harvest Report team**

Francoise Genova, RDA TAB & Strasbourg Astronomical data centre CDS, France
Hilary Hanahoe, RDA Secretariat & Trust-IT Services Ltd., United Kingdom
Leif Laaksonen, RDA Europe Co-ordinator & CSC – IT Center for Science, Finland
Carlos Morais-Pires, European Commission, Directorate General CONNECT, Belgium
Peter Wittenburg, RDA TAB & Senior Adviser, Garching Computing Center of the Max Planck Society, Germany
John Wood (chair) RDA Europe Chair & Secretary General Association of Commonwealth Universities, United Kingdom



**Contributors**

Editors: Richard L. Hudson and Nuala Moran, Science|Business
Design: Chris Jones, Design4Science Ltd
Illustration: Fletcher Ward Design
Photographs: European Commission, ESA/ATG medialab, EPFL RTEmagic, SSilver/Bigstock, The Human Brain Project





Printed by RDA Europe

A LETTER FROM THE CHAIR

# Planning the data harvest


**"Riding the Wave," Report to the European Commission by the High Level Group on Scientific Data, October 2010.**

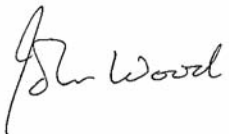
Nearly five years ago, a group of us – specialists in data science, research policy and R&D management – met at the request of the European Commission to assess the impact of an epochal development: The rising tide of scientific data.[1] From scanners, telescopes, accelerators, sequencers and other instruments around the world, a torrent of new data was flowing. Added to that: all the analysis, messaging, imaging and simulation that constitute the working drafts of 21st century physical sciences, plus the untold billions of words, images, and sounds of the humanities.

Since then, it has become clear to many that this isn't just about the volume of scientific data; rather, it reflects a fundamental change in the way science is conducted, who does it, who pays for it and who benefits from it. And most importantly, the rising capacity to share all this data – electronically, efficiently, across borders and disciplines – magnifies the impact. For the first time in human history, we can arrange to have exactly the right minds working on the right problems with the right data.

But now a new challenge becomes clear: an economic one. Just as the World Wide Web, with all its associated technologies and communications standards, evolved from a scientific network to an economic powerhouse, so we believe the storing, sharing and re-use of scientific data on a massive scale will stimulate great new sources of wealth. It turns data into a type of infrastructure, transforming the enterprise of science so anyone, anywhere, anytime can use and re-use data. It will mean new products and services, new companies and jobs. New trade flows will develop, and the competitiveness of nations will again be in play. So far, Europe is in good shape: Its scientific base is strong, and it has been investing heavily in the past decade in the infrastructure and policies to lead. But it cannot count on success.

Thus, this report is to sound a warning: Europe must act now to secure its standing in future data markets. In these pages, we outline the benefits and challenges, and offer recommendations. We speak as leaders of the Research Data Alliance, an EU-funded organisation that, with counterparts in the US and Australia, is working to speed and smooth the path for scientific data sharing around the world.

As a new Commission and Parliament begin work, we urge they pay heed. The seeds have been sown. Now is the time to plan the harvest.



**John Wood** *Chair, Research Data Alliance-Europe*
*Co-Chair, RDA Foundation (Global)*

## The Research Data Alliance

In 2013, the EU, US and Australian governments launched one of the most significant ventures in the quest for global data-sharing: The Research Data Alliance. Its aim: to promote the international cooperation and infrastructure that scientific data sharing will require. It now has over 2,350 members from 96 countries.

Among the problems it is tackling:

- What kind of infrastructure is needed to handle this data-rich science?
- How do you find the right data in the right lab that you need quickly?
- How do you manage permission, privacy and proper access to the data?
- What new software tools are needed to analyse all this data?
- How can we improve the use of computer simulation in science?
- How do you ensure the scientific data don't get lost or corrupted?

RDA's vision is of researchers and innovators openly sharing data across technologies, disciplines, and countries to address these and other grand challenges of society. RDA's mission is to build the social and technical bridges that enable data sharing, accomplished through the creation, adoption and use of the social, organisational, and technical infrastructure needed to reduce barriers to data sharing and exchange.

Scientists & researchers join forces with technical experts in focused Working Groups and exploratory Interest Groups. Membership is free and open to all on www.rd-alliance.org.

## EXECUTIVE SUMMARY

History goes in cycles, and a new one is now beginning. Over the past 25 years, we have seen the Internet grow from a technical tool to a global economic force on which millions of jobs depend. A key character in that drama has been the scientific community; in fact, it was at a European physics lab, CERN, that the World Wide Web was invented – and it was the global scientific community that first recognised its potential, and pushed its development. Now, the story is about to repeat: A new digital technology is coming. At its core is the scientific community. And, while we don't know how the story will end yet, we do know it will be important.

The story is about sharing scientific data on a truly massive scale. The sheer volume of data spilling from telescopes, gene sequencers or environmental monitors is vast. So too is the torrent from such diverse disciplines as sociology, economics or linguistics. We often feel as if we are drowning in words, numbers, sounds and images – and we are. But when data volumes rise so high, something strange and marvellous happens: the nature of science changes. Problems that were previously not even recognised suddenly become tractable. Researchers who never met, at different institutions and in divergent fields, find themselves working on related topics. Work that previously plodded along from one experiment or hypothesis to another can accelerate. And what's the vital catalyst for all this? The ability to share the data – in huge volumes, over vast distances, across disciplines and institutions. And then to analyse, re-interpret, re-use and re-think it.

This is the future of science: a global data commons, a virtual science library spanning the globe.

We are, today, starting to move. In Europe, a host of projects – national, EU, regional – is now pioneering how the system will work. Developers are working on systems to share and exploit satellite data to measure the thermal efficiency of cities and buildings to preserve the climate, or track tigers in the wild to preserve biodiversity. Scientists are sharing brain scans and genomics databases to find new medicines. In the policy world, the EU and several member-states have been successfully promoting Open Access – first for research publications, and now for data.

Why should we care? Because, just as the World Wide Web has transformed our lives and economies, so this new data wave will matter eventually to every one of us, scientist or not. In the first instance, developing the tools, systems and businesses required for this will create jobs, revenues and economic growth; the cost – growing over time to something on the order of 5 per cent of research budgets – is large but, if the market incentives are set correctly, will be shared between the private and public sector. Already, economists have shown how scientific investments of a narrower scope have yielded great returns: For instance, in the US, one study estimated the $13 billion in government spending on the Human Genome project and its successors has yielded a total economic benefit of about $1 trillion. A British study of its public

economic and social research database found that for every £1 invested by the government, an economic return of £5.40 resulted. Even bigger numbers have been circulating about the impact of Big Data, a related trend. However it is measured, the economic and social benefits will be large.

That means Europe's leaders, including its new slate of European Commissioners and Parliamentarians, must act – or go down in history as the politicians who missed the Next Big Thing. We, European members of the Research Data Alliance, an international effort to stimulate and coordinate work on data sharing, propose the following actions:

1. **DO require a data plan, and show it is being implemented.** We want a system to let researchers around the globe gather, store and manage, share, re-use, re-interpret and act upon each others' data. For that, every EU member-state should have a plan to develop the tools, infrastructure, skills and funding to take part - and the EU should update its own plans to coordinate the European effort. Internationally, every country wanting to join coordinating bodies like RDA should also have a plan implemented.

2. **DO promote data literacy across society, from researcher to citizen.** Embracing these new possibilities requires training and cultural education – inside and outside universities. Data science must be promoted as an important field in its own right. Use and evaluation of data must be embedded in all curricula, from primary school to post-doctoral programme. EU R&D programmes should incorporate data training and skills. And public workers, who control scientifically vital databases on populations and environment, need training.

3. **DO develop incentives and grants for data sharing (and don't forget Horizon 2020).** Few people will act without incentives – whether direct grants from EU programmes, or indirect market incentives to private investors. For Horizon 2020, the upcoming Work Programme for 2016-17 should reflect the growing importance of data sharing – in funding for experiments, business models, communities and analysis. Incentives will be needed for industry, in public-private partnerships or direct government procurement of innovative infrastructure. Clarity is needed on who owns a scientific data set, so a balance can be struck between public access and private gain. And within universities, a cultural change is needed so that good data management is seen as important in tenure and other rewards.

4. **DO develop tools and policies to build trust and data-sharing.** Perhaps the biggest challenge in sharing data is trust: How do you create a system robust enough for scientists to trust that, if they share, their data won't be

lost, garbled, stolen or misused? The problem is partly technical: Much work is needed to develop the underlying infrastructure, identifiers, meta-data, systems and networks – and for that, again, public funding in Europe and international coordination by RDA will be needed. But in the end, it is the culture of science that we are talking about, and that will take a generation to change.

5. **DO support international collaboration.** The biggest benefits will come from cross-fertilisation with other disciplines, regions, cultures and economic systems. Our organisation, the Research Data Alliance, with its 96-country membership, exemplifies the kind of global coordination that will be needed. But Europe must speak with one voice as the work advances, and that means the European institutions must lead. Long-term thinking and support will be needed to work globally.

6. **DON'T regulate what we don't yet understand.** Sharing scientific data on this scale is new; we don't know yet what opportunities will arise, or what problems will dog us. Until then, we urge forbearance from those who would wish to regulate too hastily. Issues such as privacy and ethics should be handled in consultation with the wider data and scientific community.

7. **DON'T stop what has begun well.** Much effort, expense and brainpower, across the EU, has been invested in making data sharing a reality. It will be a temptation, with a new Commission and Parliament in Brussels, to change course, re-order priorities and move funding lines around. Don't.

# 1. INTRODUCTION

**"We never, ever in the history of mankind have had access to so much information so quickly and so easily."**

Internet pioneer Vinton Cerf[2]

From small seeds, great trees can grow. In the 1970s and 1980s, computer scientists gradually transformed what had started as a US military project into a network for scientists to communicate. Then, in 1989, this fledgling "Internet" took a leap from public-sector project to economic powerhouse: A computer scientist at CERN, the high-energy physics lab in Switzerland, developed a way for non-specialists to use the network without knowing computer code. Then came browsers, search engines and e-commerce. Today, the World Wide Web connects some 2 billion people. It contributes more GDP globally than energy or agriculture.[3] And, on a social level, it can make new friends or topple old governments. But, policy makers take note: Scientists kick-started its growth, by sharing information.

Now a new seed is starting to germinate. And again, it depends on the scientific community – specifically, its need to share and exploit the vast new data sets of 21st century science. This is more than swapping files or following Web links. It is building the global, virtual digital libraries of the future, preserving all the outputs of research so they can be re-used by anyone, anywhere, anytime. For that, data must be structured, stored and catalogued in a way that makes it accessible across disciplines and countries. New software and hardware tools will be needed. New institutions will arise, and old ones change. New skills will develop – and greater funding, ambitious projects, and wise policies will be needed to manage the privacy, security, trade and copyright issues that arise. What we are planting now is the seed corn for new knowledge.

Since the early days of the Internet, the volume, quality and accessibility of

data have grown so quickly that it is pointing to something entirely different: A new kind of science, with profound implications for all, across all disciplines from physics and biology to economics and literature.

First, consider the sheer volume of data involved. It starts with the huge amounts of information now generated by telescopes, DNA sequencers, weather sensors and other instruments across the globe. Add to these the billions of pages of literature, history and other humanities databases that are being digitised and made searchable online. Now add data from public services like hospitals, land registries and environmental monitoring stations. In total, the world is generating 1.7 million billion bytes of data every minute. That is enough to fill 360,000 DVDs.[4] It is a limitless sea of data owned by or available to researchers – something on a scale never before seen in science.

But it isn't just the existence of this data that counts: It's the ability to share and re-use the data, across disciplines and institutions and countries. We are moving rapidly towards the day when an epidemiologist in New York can instantly access anonymised patient data from Shanghai, and work with drug researchers in Basel to find new medicines – indeed, we have already started to see it in the increasingly routinised global response to new influenza strains, and the urgent, internationalised research effort on Ebola. Soon, an historian in Johannesburg will be able to look into archives in Nairobi to analyse social trends. An ecologist in Melbourne could study wildfire distribution in parts of the US and suggest new ways to fight them at home in Australia.

The outcome: The right minds get the right data at the right time. From this will come new medicines, products, and solutions for a complex world. Scientists will change the way they work and think. Data will be stored, managed, annotated and curated in a virtual digital library spanning the globe and disciplines. It will allow others to test and reproduce the findings, to re-use and combine the data. In this data-centric world, new hypotheses can be imagined and tested. New disciplines, such as climate research or personalised medicine, can flourish. New alliances, ever-shifting to suit the latest problems, can be formed and re-formed around the globe. In growing numbers, ordinary citizens can join in – adding their observations, opinions, and policy preferences.

The vision is breath-taking: From an abstruse scientific tool grows a technology that can turn the entire globe into one vast network of thinking, sharing, competing, collaborating and generally 'doing' together. Science –

physical and humanistic – becomes the baseline, not just for knowledge, but for prosperity and peace. Too idealistic? Look at what science has already achieved. Who could have predicted, 50 years ago, that starvation would cease to be the dominant worry in most of the developing world, or that in just the past 25 years the global mortality rate for under-5s would nearly halve? Who could have foreseen that a diagnosis of breast cancer would no longer sound like a certain death sentence, or that our ability to learn and think would be so enriched by computers?

So this is not dreaming. Europe has already invested heavily in the foundations of this world science library – as have the US, Australia, Canada and many other research-intensive economies. The initial investment is scientific, but the ultimate return is economic and social. It will take much work to harvest these benefits. If we are to succeed, we must act now.

## What more is needed?

We have the Internet, the World Wide Web, and billions of humans – scientific or not – interconnected already. So what's new or special about the kind of global, virtual science library we envision?

We want any researcher, anywhere, at any time and in any discipline to be able to reach across the globe and gather data – to store it, test it, re-use it, analyse it, and find new insights in it that even the original authors hadn't imagined. Repeating: we want the right minds, with the right data, at the right time. That's a tall order that requires change in:

- The way science works and scientists think
- How scientific institutions operate and interact
- How scientists are trained and employed
- The way data networks operate, data tools function, and the basic vocabulary of scientific data is parsed
- The way we handle problems of data privacy, security, ownership and ethics
- The international order for funding, coordination and regulation of scientific data.

# 2. WHERE WE ARE TODAY

The vision is grand: To create, across the globe, the means for researchers to work together with ever-growing data flows – instantly, efficiently and creatively, crossing all boundaries of discipline, geography, institution or policy. In short: To make the whole world a single, living lab. This laboratory of Open Science will have all the functionality required to allow data to be analysed, validated, integrated and re-used. For this to happen, new tools and services must be layered on top of the physical computer networks, servers and storage systems.

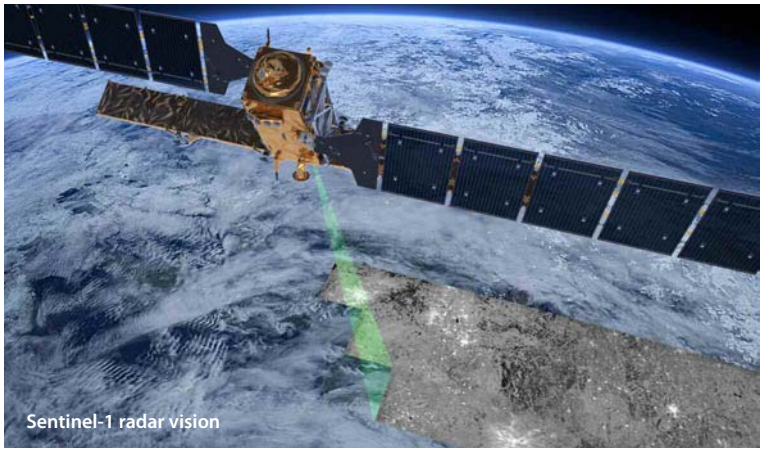
Of course, a revolution like this won't happen all at once. It comes in steps, gradually gaining force until, one day, its importance becomes obvious to all. Here are some dispatches from the front lines of this data revolution:

**The data space: COPERNICUS**

**Space is no longer** the exclusive realm of rich governments; it is an industrial sector of growing importance – and a good example of the chain of value that data sharing can create, from government to small company to individual citizen.

From April 2014, the first of a suite of Copernicus[5] earth observation satellites launched by the European Space Agency started to send images back to earth. Each orbit of the earth will generate data streams of several terabytes, delivering unprecedented temporal and spatial resolution and data continuity. All of this information will be freely available – to public authorities, to scientific and commercial users, and to the general public. And research[6] commissioned by the European Space Agency says the system could generate a financial benefit of some €30 billion and 50,000 new jobs by 2030.

That includes jobs at small companies. For instance, Latvia-based ThermoCERT uses the data as the basis of tools for monitoring the thermal efficiency of buildings and urban areas. Another example is CAMEA, a project developed by a Hungarian landscape planner using Copernicus images to certify that crops have been grown in an ecologically sensitive way. And then, the satellite data can also feed citizen-science work – like Tiger Nation[7], a project to track the 1,700-odd Bengal tigers still alive in the wild in India. Photographs taken by


Sentinel-1 radar vision

some of the 2.2 million tourists who visit Indian game reserves each year are uploaded to a central database, where image recognition software identifies individual tigers from their stripes, with the associated satellite location data allowing them to be tracked.

**The humanities Web: CLARIN**

**Data are not just numbers**: words, sounds, images and other cultural artifacts are also part of the information torrent. Clarin[8], a €104 million EU project, is knitting together the archives of universities, libraries and other public institutions across Europe so all this data can be easily found, searched and used by researchers – whatever their native language.

The uses of the system already can be astonishing. Where does your family come from? If you're Dutch, the Clarin system can let you trace how family members from generations past moved around the country – all through interlinking public records. Or, perhaps you want a Polish summary of the content of *Le Monde* on a particular date in 2008? How about a search through Estonian or Russian literature by keywords – that doesn't get confused by complex case-endings or different archival systems? How about a word-map showing the relationship between all the cognates of 'bread' in Danish and Finnish?

The ambition of Clarin is enormous: To permit every scholar in the humanities and social sciences in every EU and associated country to

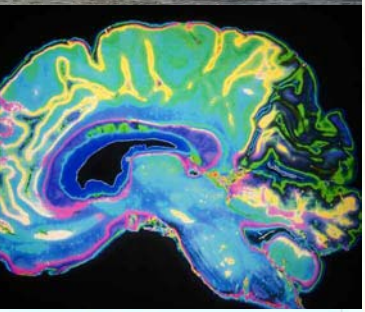
have direct, single sign-on access to every digital data collection containing language-based material owned, or made available by, public bodies. These will be annotated and tagged to make searching easy; and researchers will be able to build their own virtual collections of material from different sources in different countries. There are also language tools to annotate, explore, exploit, enhance, analyse, manipulate and visualise data. At the same time, Clarin is making it possible to feed results of research projects back into data collections so that other researchers can use them. The aim: data and results are preserved in a sustainable way, with persistent identifiers.

**Brain injury: A SHARED DATABASE**

**Traumatic brain injury** is the most complex ailment in our most complex organ. There are 10 million cases worldwide each year; and it is the leading cause of disability in individuals under the age of 45. Estimates[9] put the annual cost of caring for people with brain injury in the US at $60 billion a year; across Europe the cost is at least €33 billion.

The Collaboration in European Neurotrauma Effectiveness Research in Traumatic Brain Injury is a €30 million EU project sharing information among 60 hospitals and 38 science infrastructures. It will collect data on 20,000 to 30,000 patients with traumatic brain injury. There is a range of possible treatments, and this project aims to help doctors decide which to use in which circumstances.

## Other European data sharing projects



- **LifeWatch** provides the e-Science infrastructure underpinning research into biodiversity and ecosystems across Europe. It helps researchers study the spread of invasive species that can out-compete native species, threatening extinction. It helps monitor Europe's wetlands, and provides a single point to collect information on its migratory birds. The resulting data can be cross-referenced with other sources, relating to weather and climate, for example. Constructing consistent geographical datasets within LifeWatch from across Europe makes it easier to assess the effect of climate change or agricultural practices on biodiversity. (www.lifewatch.eu)

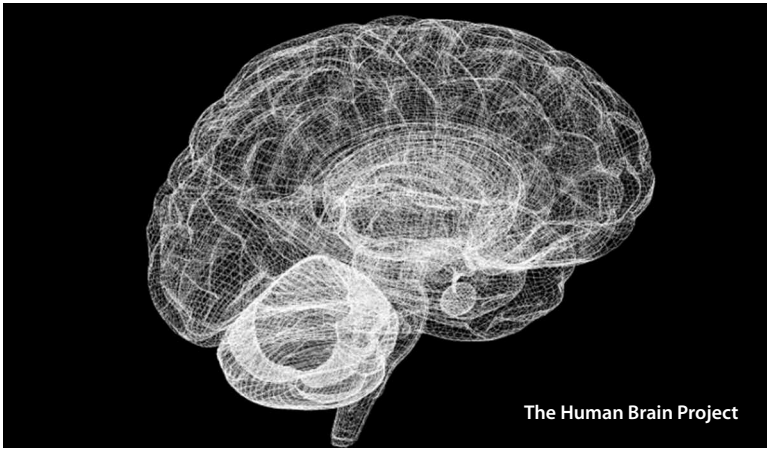
- **Pharmacog**: a €27.7 million pan European partnership involving 15 academic institutions, 12 global pharmaceutical companies and five SMEs, to help pick out compounds with the potential to treat Alzheimer's disease at the very early stages of drug discovery, by improving preclinical research and animal models. There have been many expensive late-stage failures of drugs in development; the aim is to ensure that only compounds with well-validated biology advance into clinical development. (www.imi.europa.eu/content/pharma-cog)

- **Walter – Wadden Sea Long-term Ecosystem Research**: another example of an initiative by institutions to pool data and plug gaps in monitoring networks, to give a comprehensive ecosystem-level view of this important Dutch inland sea. (www.walterproject.nl).

- **The Human Brain Project**: this €1 billion EU initiative involves 135 partners in 26 countries working across many disciplines, from high performance computing to medical informatics, to model the human brain. One of the biggest hurdles to be overcome is that existing data from brain research has not been collected and stored in a standardised, systematic way. (www.humanbrainproject.eu)

- **Elixir**: A pan-European project uniting Europe's leading life science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research. (www.elixir-europe.org/)

- **European Medical Information Framework (EMIF)**: A project to provide a common architecture for sharing data, with seven countries contributing 48 million patient records. An example of how this might be applied is an Alzheimer's disease project, EMIF-AD, which will mine these records to look for links between genes, biomarkers and outcomes in cases of the neurodegenerative disorder. (http://www.emif.eu/)




Similar projects are in progress elsewhere. The findings from the European project will feed into the International Initiative for Traumatic Brain Injury Research, involving the EU, Canada and US. This represents a global effort to coordinate and harmonise clinical research activities across the full spectrum of brain injuries, with a long-term goal of improving outcomes and lessening the global burden by 2020.

These are just a few examples. But Europe is not alone in seeing the importance of scientific data sharing. The Australian government hopes to save property and lives, by sharing data on wild fires – a particularly costly problem there. Likewise, it is joining other governments on a project to share data on the southern seas – to track fisheries, environmental conditions and weather patterns on a vast scale. In the US, the National Science Foundation is funding Earthcube[10], a network for data sharing across the wide range of disciplines relevant to geoscience. The aim is to create a unified structure integrating different scientific inputs to create a single view of the earth system - from the sun to the centre of the earth.

In sum, it is no longer in doubt that data-sharing, on a vast scale, among millions of scientists, will happen; it already is. And this will increase productivity, benefit industry, support trade, inform and legitimise policy - and save lives. As a result, over the past five years, the pace and breadth of collaboration on scientific data have increased, opening up to more analysis, validation, checking and re-use. There are still many obstacles, and many challenges. Thankfully, we are starting to see more and more policy action around the world.

## The Open Access movement

The most significant policy change underpinning these projects has been the now-unstoppable momentum of Open Access – the notion that publicly-financed research results and data should be freely available to all scientists. From the beginnings of the movement nearly a decade ago, it encountered stiff opposition from private scientific publishers. But already by 2011 about half the scientific papers published that year were freely available, the Commission says.[11] That is, at the normally sedate pace of academic policy change, an overnight revolution. The European Commission re-committed itself to the cause in 2012, saying: "Information already paid for by the public should not be paid for again each time it is accessed or used and …it should benefit European companies and citizens to the full."[12]

But making journal papers available through Open Access is only a staging post. To extract the full value, advocates say, all Open Access research papers should be available through a single source. This is the objective of OpenAIRE, (Open Access Infrastructure for Research in Europe, at www.openaire.eu) an EU project set up to support the implementation of open access in Europe. It has supported open access to almost 8.5 million publications from 461 data sources. Next, as part of its Horizon 2020 programme, the EU has launched the Open Research Data Pilot, which considers commercial interests, privacy and security aspects of open access and seeks to maximise access to, and re-use of, data generated by the projects. Another initiative, EUDAT – European Data Infrastructure (www.eudat.eu) – is creating a pan-European infrastructure for e-science to glue together other ambitious projects. And several member-state projects are also underway – such as a Dutch effort (www.u2connect.eu) to provide access to scientific information from the repositories of all Dutch universities and a number of research institutes.

This significant progress is not to say that all the problems have been dealt with. Technical issues remain; and there are important matters of governance and oversight that require attention in opening up data for sharing. This is particularly so in the case of personal medical information. In the UK, the medical charities Wellcome Trust and Cancer Research UK have joined with two funding bodies in setting up an expert group to advise on the emerging scientific, legal and ethical issues associated with data access for human genetics research and cohort studies.

Open Access initiatives have been underway elsewhere. In 2004, science and technology ministers of 30 countries called on the OECD to develop

guidelines for the sharing of, and access to, digital research data.[13] A notable point came in 2013, when US President Barack Obama launched the $200 million Big Data Research and Development Initiative, which directs federal agencies that get more than $100 million a year in research funding to make results available within one year of publication. And, in a related move, the US has created a counterpart to Europe's OpenAire programme – to make sharing and preserving data easier for universities and research institutes.

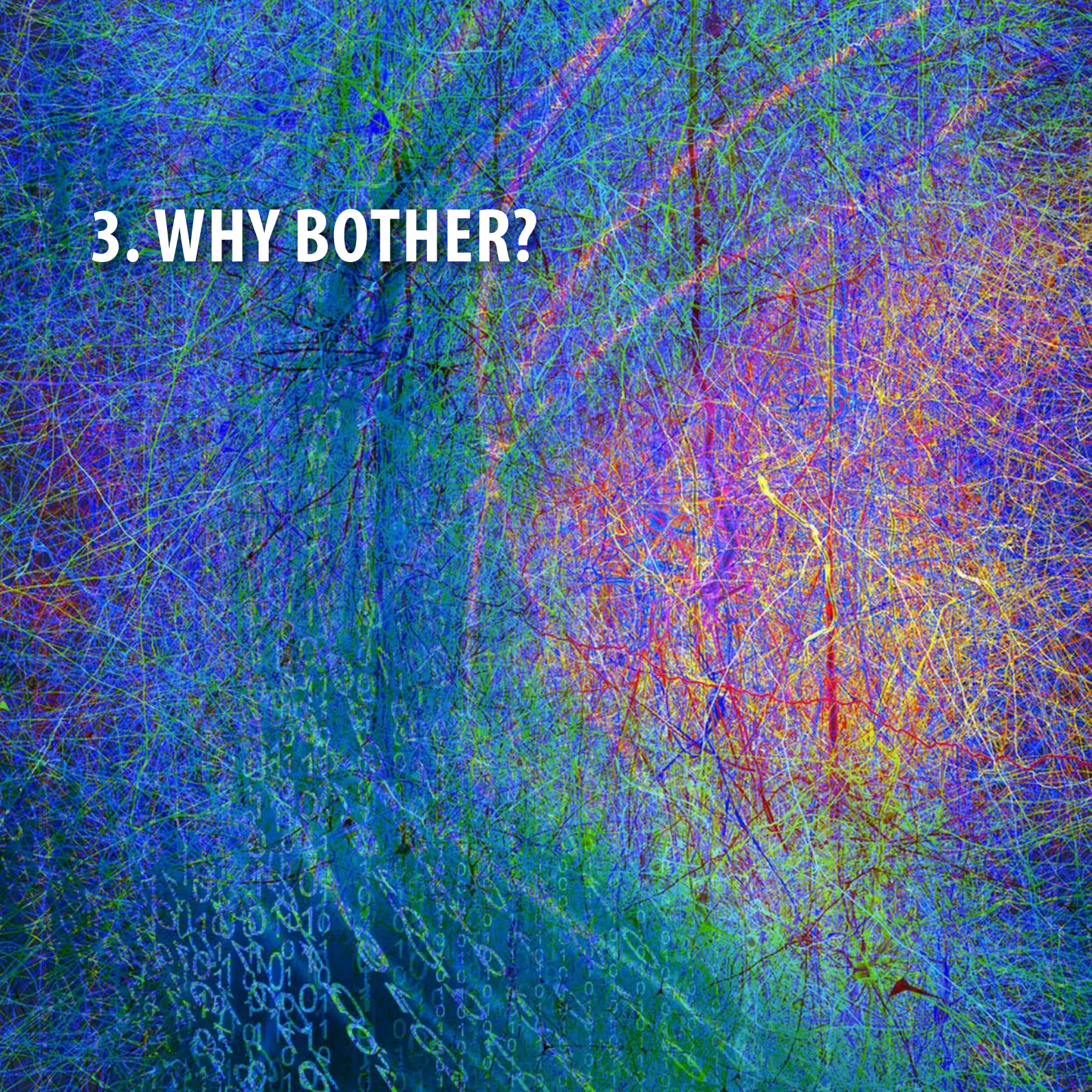
This is detailed, technical work, but the economic impact is never far from politicians' minds. In announcing the Big Data initiative, John Holdren, Director of the US Office of Science and Technology Policy, said:[14] "These policies will accelerate scientific breakthroughs and innovation, promote entrepreneurship, and enhance economic growth and job creation."

## The broader context

The policy work goes well beyond Open Access, however. There's infrastructure: The EU has spent several billion euros over the past decade, in various programmes, on machines and networks to generate and transport high volumes of data. Its policy initiatives have been advancing – from, in February 2007, a communication[15] on access, dissemination and preservation of the scientific information flowing across this physical infrastructure. Now under its new research and innovation programme, Horizon 2020, there will be further investment in infrastructure - for example to strengthen the GÉANT research network, and develop a 5G mobile network with the capacity to transmit "big data". It is readying a multi-million euro public-private partnership on Big Data towards the end of 2014. It plans to invite proposals for Big Data "lighthouse" initiatives to enhance daily life, advance Europe's competitiveness, and improve public services.

In 2012, the EU joined the US and Australian governments in launching the Research Data Alliance (www.rd-alliance.org) to promote international cooperation in the field. And then, in 2013, the G8 countries adopted an "Open Data Charter" to act by the end of 2015 on a set of five basic principles – for instance, that public data should be open to all "by default" rather than only in special cases.[16]

"It's pretty clear that in the 21st Century, data drives everything, from the health sciences to climate change," said Francine Berman, chair of the US RDA branch.[17] "But there's only so far you can go in solving problems using your own data and your own team. Today, you need to reach across boundaries."



# 3. WHY BOTHER?

Getting scientists to share and build upon each others' data has always been a complex business. Nearly two centuries ago, when Charles Darwin returned from his voyage to the Galapagos, he brought back a vast trove of data: 770 pages of diary, 1,383 pages of geological notes, 1,529 specimens preserved in spirits and another 3,907 dried.[18] Some of it he kept to himself, as he began work on what would become *On the Origin of Species*. But some of it, he shared with others - expressly to speed the tedious job of cataloguing and analysing everything. Since then the history of evolutionary biology has been a process of building communities around new data, public and private.

Today, researchers have no less hesitance about sharing data; there is just so much more of it available than in Darwin's day. By collaborating, scientists can extend their intellectual reach, tackle more complex problems, and step out of the confines of their disciplines. But the benefits of sharing go well beyond the immediate scientific impact: Increasingly, it means money – new products and jobs, trade advantage, economic growth. In the final analysis, it's the economic implications that lift it up the policy agenda.

## Benefit 1: Creating jobs, spurring growth

It seems obvious: Data have more value shared and used, than hidden and unused. But quantifying that value, in a form that would satisfy an economist, is a task that is only now getting underway. Some recent studies are suggestive:

- **The multiplier effect***:* In 2012, the British government commissioned some research[19] on the economic value of one of its important scientific collections, the Economic and Social Data Service. This curated database and archive, with 23,000 users, exists to promote wider use and sharing of research and teaching in the social sciences. The study's conclusion: the ease of access and of finding the right data means users make efficiency gains of more than £100 million per annum. For every £1 spent on the Economic and Social Data Service, there is a value to the economy of £5.40.

- **The DNA dynamo**. A separate study counted the economic value of the US Human Genome Project, to sequence the human genetic code; significantly, the project included an international agreement to put sequence data in the public domain. The study[20], by the Battelle Institute in 2011, showed that the $3.8 billion that the US government invested in the project from 1988 to 2003 yielded $796 billion in economic output – in new medicines, equipment, services, jobs and more. An updated study, published in 2013, added in the impact of the subsequent public investment in genomics research of $9.1 billion after the genome was sequenced. The estimated total return: $1 trillion.

- **Big Science**: EU member states spend €10 billion a year running capital-intensive, shared Big Science research facilities – from synchrotrons to telescopes. Of course, they generate knowledge; and sharing data is part of how they do it (remember: The World Wide Web began at the CERN particle accelerator). But they also promote innovation through the transfer of knowledge to their industrial supplier companies, by providing knowledge inputs that are distinct from those coming from any other networks, according to a recent study[21] by Erkko Autio, an Imperial College London professor. The unique frontier-pushing infrastructure demanded by many Big Science projects can lead to major opportunities for the high tech firms which supply the facility. They provide a platform for global research networks. And they provide research and training services that would otherwise not be available to firms, according to Autio.

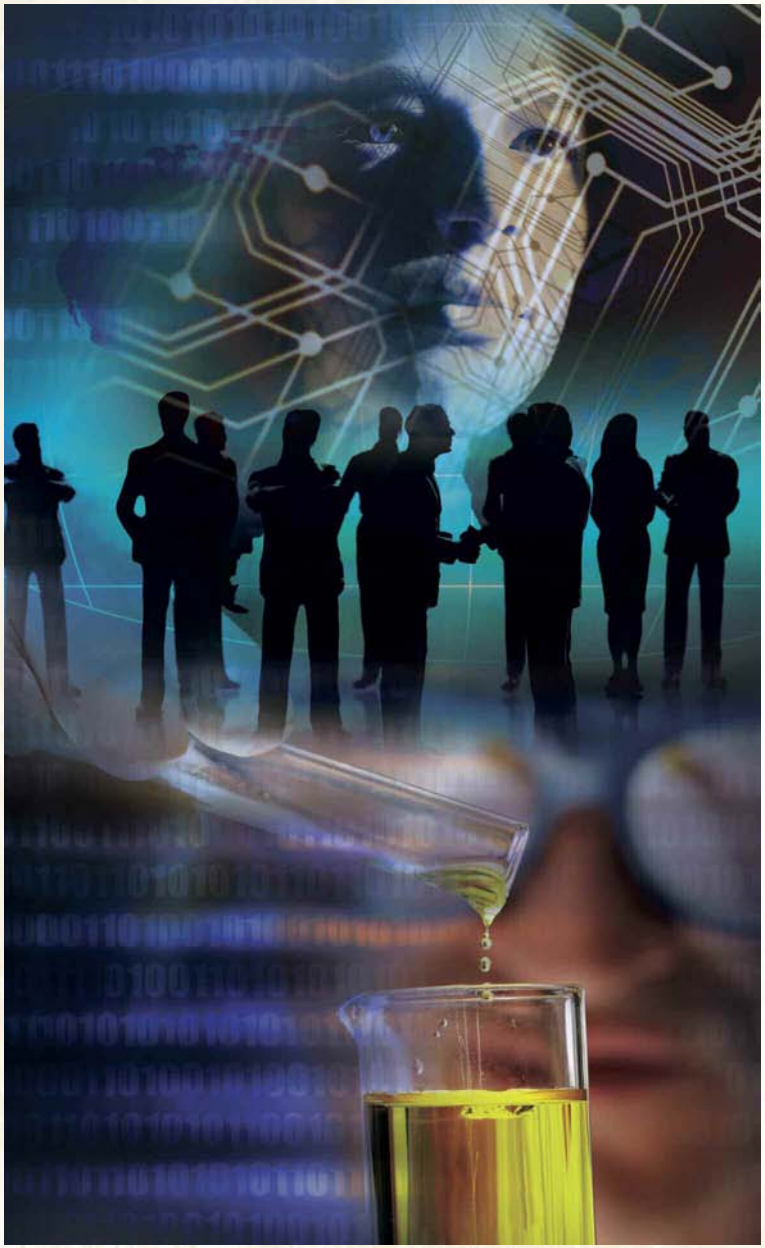
- **Big data**: For both private and public sectors, better use of 'Big Data' matters. For instance, in economics research, a special problem is getting accurate employment numbers – and, because government depends on them, bad numbers mean bad policy. A study by the Research Institute of the Finnish Economy, published in August 2014, concluded that the analysis of Google search queries can improve the accuracy of unemployment statistics.[22] More broadly, a report for Germany's Federal Ministry of Economics and Energy in March 2014 concluded that Big Data will foster big changes in markets and in companies. "These disruptive changes can result in substantial opportunities and competitive advantages for businesses in Germany," say the authors.[23].

## The benefits of open data

**The Citizen**: All people will benefit from the products and services that are developed around open data and sharing – directly or indirectly. More accountable, efficient and effective businesses and government result. Most importantly, citizens are empowered, have the information they need to make decisions in all spheres of life; they are engaged.

**The Entrepreneur**: Open data is a source of inspiration for entrepreneurs and provides the raw material for new products and services. Examples include the opening up of weather data, leading to the private sector provision of information services, making global positioning data available to a mass market in satellite navigation systems, and making Human Genome data freely available to the genomics sector. No one organisation has the money or the expertise to extract the full value from its data. Opening it up to entrepreneurial imagination will foster unthought-of innovation.

**The Scientist**: Freely exchanging data will transform the nature of what it means to be researchers. It will make their work easier and faster, as more data and tools are put within reach. It will open new research avenues, crossing old boundaries of discipline, institution or country. It will create new career opportunities, and get more researchers crossing borders. And, through greater engagement with fellow-citizens, it will enhance their status and relevance in society at large.



## Re-using public data. By the numbers

What is it worth for the government to share its data? Plenty, according to a recently published review of several studies. Here are a few numbers:

- **€28 billion** in 2008, rising to around €32 billion in 2010: That is the size of the direct market for re-use of public sector information in the EU.
- **€140 billion** a year: That is the probable total value of that information, counting both direct and indirect uses of the data across the economy.
- **€40 billion** a year: That is how much the economic value could grow, if policies were changed to make it easier for people and companies to get at the information.
- **€2 billion** a year: That is the potential savings in preparing environmental impact statements if the relevant public sector information were more easily available.
- **€1.4 billion** a year: The economic benefit if, through more rapid and comprehensive access to public information, average citizens could just save two hours a year in their routines.

*Source: "Review of recent studies on public sector information re-use and related market developments." Graham Vickery, Information Economics, Paris, August 2011.*

- **Business creation**: Making public sector data available directly stimulates the formation of companies that use it in innovative ways. For instance, ASEDIE, a Spanish information industry association, reported in 2013[24] that 444 companies were using public sector information. They had a combined turnover of € 900 million and 9,971 employees. These businesses identified quality, accuracy and accessibility as the key requirements for making public sector information amenable to commercialisation. Almost 37 per cent use public sector information from other countries in the EU, and urge that standardising formats, so that all data can be handled in the same way, would promote business creation.

One study[25], by McKinsey consultants, tried to summarise the impact of open, shared data – not just in science but across the economy. It suggested that seven sectors – education, transportation, consumer products, electricity, oil and gas, health care, and consumer finance – could generate more than $3 trillion a year in additional value as a result of open data. Already, it is giving rise to hundreds of entrepreneurial businesses, and helping established companies to segment markets, define new products and services, and improve the efficiency and effectiveness of their operations. Open data can break down information gaps across industries, allowing companies to share benchmarks and spread best practices that raise productivity. The study concluded: "Although the open-data phenomenon is in its early days, we see a clear potential to unlock significant economic value by applying advanced analytics to both open and proprietary knowledge."

## Benefit 2: Boosting research productivity and creativity

Sharing and re-using data change the way science is done, and who does it; that has unexpected consequences. Already, through online tools, researchers are getting thousands of people to report sightings of wild life, observe galaxies, help track infectious diseases. Scientists themselves are becoming more open, sharing early findings, working in large-scale international collaborations, making their background data accessible.

For instance, our knowledge of climate change – and certainly its political impact – would be much smaller today had it not been possible, in just the past decade, for disparate scientists in different disciplines to compare notes, share their data, and create complex computer models collaboratively. The very idea of an international, multi-disciplinary peer review group, such as the

International Panel on Climate Change, springs from scientific data sharing on a massive scale. The famous search for the Higgs Boson, at the CERN particle accelerator, was enabled by thousands of researchers and engineers sharing and analysing data together. As these examples suggest, ready access to vast data sets changes the scientific method, itself, with deep epistemological implications. Today, science is no longer a simple matter of testing discrete hypotheses; constant interaction with vast data sets suggests an ever-changing set of new hypotheses, each assessed with varying degrees of confidence.

Some refer to this broad set of changes as Science 2.0. Indeed, the European Commission recently launched a public consultation on its meaning, impact, and policy implications. But in the real world today, we see its effect in graphic terms. In 2014, in the six weeks after scientists at Harvard-Smithsonian Center for Astrophysics published direct evidence for cosmic inflation - the Big Bang theory as it is popularly known - more than 200 papers appeared by other scientists who drew on the Harvard data to reach new insights. Such collaborations are becoming the norm as data volumes mount across the globe – at a rate of 30 per cent a year, according to the European Commission.[26]

## Benefit 3: Helping people, engaging citizens

When floods struck the south of England in January and February 2014, an experiment in data-sharing – under crisis conditions – began. The results are a model of how citizens can benefit directly from data sharing.

In one of the most badly affected areas, the Somerset Levels, over 600 houses and 17,000 acres of agricultural land were inundated. Villages were cut off. The army was called in to assist the over-stretched emergency services. High volume pumps were brought in from the Netherlands. As part of its response, the UK government instigated a one-day event, Flood Hack, at which more than 200 developers volunteered to make apps for communities hit by the floods. The results included "Don't Panic," a system that allows people with and without Web access to get help, ranging from the delivery of sandbags to information. The data are recorded for future analysis and real-time response planning. Another was Flood Feeder, an aggregation tool that visually presents a feed of flood and related data. The initiative points the way to better flood management across Europe based on real, constantly updated environmental data.  A report[27] in the journal *Nature Climate Change* estimated

## The Era of Data-Driven Science

In July 2014, the European Commission launched a public consultation on what it calls Science 2.0. Herewith, the official view:

"Science 2.0 defines systemic changes that are currently taking place in the way the science and research system functions. It is characterised by an open, collaborative networked way of doing research, that has been referred to as Facebook for scientists. While the feedstock is big data, it requires many people to make inputs.

"Science 2.0 is enabled by digital technologies and driven by the globalisation and growth of the scientific community, providing the means to address the Grand Challenges of our times. Science 2.0 has impacts on the entire research cycle, from the inception of research to its publication, and on the way this cycle is organised. It also affects the evaluation of the quality and impact of research."

*Source: ec.europa.eu/research/consultations/science-2.0/consultations_en.htm*

economic losses from flood damage are set to rise from €4.9 billion a year to €23.5 billion by 2050. Making better use of environmental data is an obvious way to help manage the damage.

Similar hacks have happened elsewhere in Europe. In May 2014 the "Homer" project organised a hackathon to unlock the value of public sector information in the countries of the Mediterranean.[28] The winning app, Geostep, was developed by a team from Montenegro, and provides the smartphone answer to a printed guidebook. It suggests attractions to visit based on a user's location, and then provides further information on arrival at a site.

Citizen-science projects are spreading across Europe, based on shared data. One EU-funded project, Everyaware, is building on the fact that low-cost sensing technologies now allow citizens to collect environmental data, and blend these with their personal, subjective perceptions. For example, all smartphones have microphones that can record noise pollution; many also have thermometers for recording temperatures. Social networking tools can help collect and distribute data from thousands of smartphones, to be analysed, interpreted and visualised. Everyaware aims to pull these elements together in a single technology platform for conducting environmental surveys and analysing the results.

Can this strengthen democracy? Hopefully: Citizens not only gain greater insight into what is being done in their name, but they can also look at the data themselves and suggest policy improvements. Politicians, take note: It will change your business forever.

## Meeting the challenges

As with any phase-shift in our thinking, data sharing poses several knotty problems. Edward Snowden's revelations have politicised the whole field of data protection and privacy. So how to secure the benefits of sharing research data, without undermining the rights of the original researchers and funders, or compromising the privacy of citizens? As a society, we don't yet have an answer. But bit by bit, there's some progress:

- **Medical records**: Within the research world, medical records pose the biggest policy challenge. Nobody wants to see their health history end up on Facebook or in their employee files; yet everybody wants researchers to find new medicines. The European Commission recently proposed a

new Data Protection Framework to protect privacy while enabling research – and the heads of Europe's leading medical research organisations at first applauded.[29] But amendments advanced by the European Parliament in 2014, the researchers say, undermine the proposal and risk making vital research unworkable. [30]

- **Online privacy**: Beyond medical research, many social sciences are affected by how personal data are treated online. The courts have started to move: In May, the European Court of Justice enshrined a "right to be forgotten" online – forcing dramatic changes in the way Google and other search engines operate.[31] The impact on research is as-yet unclear. But EU Justice Commissioner Viviane Reding called the decision "a clear victory for the protection of personal data of Europeans."[32]

- **Intellectual property rights**: There is a fine line between the moral imperative of allowing taxpayers free access to research they have funded, and the practical need to provide incentives for industry to invest and commercialise research results. The Commission wants all information generated in Horizon 2020 R&D projects to be freely available; and its €3 billion Open Data Pilot will study how to do this.

## The cost of not sharing

It's hard to prove a negative; but most researchers and economists believe there is a cost of *not* sharing, as well. This would be measured in products that don't appear, jobs that aren't created, medicines that don't save lives, and efficiencies in research that don't happen. There is a social cost too, in slower progress in understanding disease, in addressing environmental and energy challenges. Further, many complex problems can only be addressed by combining information from multiple data collections. In an earthquake, knowing what areas are most at risk requires mixing sensor data, population density data, information about the built environment and transport and communications infrastructure – as well as inputs from earthquake simulations.

It is evident that there is a risk to scientific research if the data tools are not up to the job, and data are mis-used or misunderstood. It would become impossible to reproduce results, retarding the pace of research and leaving a trail of scientific errors. And don't forget the unquantifiable cost of *not*

sharing *not* conclusive scientific results. Perhaps the best example is in the handling of 'negative' clinical trial results. When a drug fails to show the hoped-for effect, others should be told, to avoid duplicate studies and to inform future research.

If that's still too theoretical for you, there are well-documented cases of the cost of not sharing. In 2006 the World Health Organisation complained[33] that China was withholding information about bird flu, making it hard to track how the virus was mutating and spreading. The H5N1 virus first infected humans in 1997 during a poultry outbreak in Hong Kong. Since its widespread re-emergence in 2003 and 2004, the avian virus has spread from Asia to Europe and Africa and has become entrenched in poultry in some countries, resulting in millions of poultry infections, several hundred human cases, and many human deaths. Since that initial diplomatic fuss in 2006, however, there has been some progress in sharing international data about the disease. Today, WHO is coordinating the monitoring of bird flu globally, and in March 2013 received from China the first reports of a new subtype of the virus, H7N9, infecting three humans, two residents of the city of Shanghai and one resident of Anhui province. The Chinese authorities continue to update WHO, most recently providing details of four further cases in June 2014.[34] So there can be progress.

# 4. THE WAY AHEAD

History can repeat itself. Today, it is the job of European policy leaders to prevent that happening – at least in the area of scientific data.

The danger lies in the history of the Internet and World Wide Web. That technology, as outlined earlier, began growing with the scientific community – initially in the US, but soon in Europe, as well. But it was American companies that first spotted the commercial potential and launched user-friendly browsers and money-making services. Likewise, it was the US government that wrote the rules of the road for what was by then being called "the Information Superhighway." European politicians didn't really catch on until 1993, when the first EU R&D project on the Web was funded. Result: For most of the past 25 years, it has been American companies that have reaped most commercial benefit from the Web. Europe still has no real answer to the likes of Google, Linked-In, Facebook or Amazon; and while challengers are now rising, they come from Asia, not Europe. That, as an economist would say, is a great example of "path dependence."

Today, we are about to take another step-change in our ability to communicate and share information. As explained earlier, this is beginning as a tool for researchers: Creating the infrastructure, standards, rules and institutions that will enable the biggest, hyper-speed growth of online data usage since the early days of the Web. We can already foresee billions in new equipment and services will be purchased – from somebody – as more and more scientists join globally. What we can't yet foresee is all the new commercial services and products that some bright sparks, somewhere in the world, will start to build out from it. But, if history is any guide, we can predict it will be big – bringing lasting economic gain to whichever part of the world moves first. And with that, of course, will come as-yet unpredictable changes in society, politics and our daily life. Remember the Twitter effect, during the Arab Spring of 2011?

The weight of history, then, is pressing on European leaders today: Make this a story that Europe writes, not just reads.

That will require major investment in data infrastructure, the creation of new user communities and institutions, writing new policies to encourage growth, and much more. And, because other parts of the world are already

starting to move as well, it will require an unprecedented degree of global cooperation and coordination. Herewith, our Do's and Don'ts for Europe's leaders.

## 1. DO require a data plan, and show it is being implemented

**We want a system to let researchers around the globe gather, store, share, re-use, re-interpret and act upon each others' data – a global digital commons for science.** That effort has to start at home, with each member-state of the European Union figuring out its part in the story, and how it will be fulfilled. Moreover, the European Commission must update its own plans to incorporate these 28 individual efforts, and integrate them into the global endeavour. Of course, several member-states have already formulated plans and moved far down the road of action; noteworthy are the efforts of Britain, Germany, Denmark, the Netherlands and Finland – and, together, some of their computing institutes now collaborate in a regional effort called Knowledge Exchange (www.knowledge-exchange.info). But other countries have yet to act, and may need help from their neighbours and the EU to do so.

These plans must be comprehensive. It isn't just about which institutes will get which funding lines – though that certainly matters if the system is to grow or be sustainable. There are harder issues, embedded in the culture and institutions of each member-state. How will a country's universities manage academic career tracks, to encourage researchers to share? What curricula will be needed to educate young researchers – or, indeed, the entire population – about the hows and whys of sharing and re-using data? At what age should the training begin? How will research institutions sustain their investments, so the data aren't here today and gone tomorrow? Global data-sharing will wreak great change in the methods, institutions and careers of the scientific world. That isn't something undertaken lightly, and should involve all relevant government departments and research institutes in each country, so a coherent European approach evolves.

And on an international level, we believe formulation and implementation of an appropriate data plan should also be the entry ticket for any country wanting to play. As the RDA grows and adds new countries, all the members should at least be working from the same book, if not on the same page. A plan is the first vital step towards getting politicians in a country to focus.

## 2. DO promote data literacy across society, from researcher to citizen

**The transition from standalone personal computers in the 1980s to 4G smartphones, music streaming and cloud computing in our age was long and, for many people, challenging**. How much harder will this next step, to global data sharing, be for the current and next generation of researchers? How quickly will grant agencies adapt? How hard will ordinary people, with a growing interest in citizen-science, find it? A massive programme of human engineering will be needed – in professional training, general education and cultural attitudes. The EU can help lead this effort, through programmes like Horizon 2020 for research or Erasmus-Plus for education. The member-states, again, will need plans and timetables. And at the front line of this effort: individual universities, school systems and professional bodies. Among the tasks:

- **A first-class science**: Data sharing provides the foundation for a new branch of science. It must be acknowledged as such, and ranked alongside other major disciplines. This means fundamental changes in the incentives, career paths and academic status of what has, hitherto, been a fairly low-profile discipline. At the same time, data science needs more professional bodies, and internationally transferrable qualifications. This is homework for university administrators, education ministries and learned societies.

- **Data education**: Training in the use, evaluation and responsible management of data needs to be embedded in curricula, across all subjects, from primary school to university. We have seen this before: For instance, the teaching of intellectual property moved from the law schools to the engineering faculties at many technical universities; and, now, it can be found in the curricula for many masters and doctoral students across Europe. In the same way, elementary training about data science and

handling must be viewed as a standard part of every young researcher's education. Again, this is on the to-do list for universities and education ministries.

- **Training within EU projects**: The majority of EU-funded research projects rest on collaboration. A data management plan should be required for every EU project proposal, spelling out how data will be used and shared and how people will be trained to do it. This is an urgent action for Horizon 2020.

- **Government and public sector training**: Some of the biggest, and scientifically most interesting, data sets are in public hands – healthcare records, income statistics, environmental monitoring. As the appeal of Big Data grows, there's mounting recognition that the public sector needs to rethink when and how it grants access to its data riches. To help this thinking, national ministries and agencies could usefully add data training to their own employee requirements.

## 3. DO develop incentives and grants for data sharing (and don't forget Horizon 2020)

**This new data revolution won't come easily; it means changing the way many people, institutions and companies think about what they do and how they collaborate**. It will also be very expensive. Our informal estimate is that the infrastructure and operation of a truly effective data-sharing system could cost on the order of 5 per cent of total research budgets. For the Commission, which spends over €10 billion a year through its Horizon 2020 programme, that would amount to half a billion euros. Scaled up globally, it seems massive. But so, too, the economic return. Thus, we would expect funding to come from both private and public sectors – and for that to happen, proper incentives will be need, some financial, some institutional. For instance:

- **Horizon 2020**. The EU's flagship R&D programme, worth €79 billion over seven years, is a prime actor. Curating data and making it available will add costs for research communities, and this needs to be reflected in grants and other funding. But we will need many experiments and pilots to see what technologies or business models work best in this new domain. Funding will

be needed to build the community of data scientists; to support tool development; to kick-start major public-private infrastructure projects; or simply to give a bright idea a small chance to grow. The new Commission can start with the 2016-17 Work Programme for how it will spend the Horizon 2020 money. But we also urge member-state and charitable funders to accommodate this need, as well, when providing research grants.

- **Incentives for industry**: Commercially, we're still in the early days of this revolution. What products might be needed? What services? What business models could help pay for it all? Of course, industry can be trusted on its own to find the gold mines eventually; but a few test drillings, in the form of joint industry/university projects in specific applications, can help point the way. The Commission should consider this in Horizon 2020. And at the same time, it and other public procurers should use the power of the purse. Together, their procurements represent about 16 per cent of the EU economy; using a bit of that money to procure innovative data-sharing systems would provide powerful incentive to industry.

- **Intellectual Property Rights**: Who owns a scientific data set? At present, you generally cannot patent what's in a database, trademark it or copyright it. True, you can try to protect it as a commercial asset – like a building or a trade secret; but that's governed by contract law, country by country, with enforcement costly and time-consuming. This uncertainty is bad for private investors. They won't put their own money into data-sharing unless there's some simple, clear way of earning it back with profit. What's needed: Solutions. The EU could start by funding some studies, to float a few new ideas for when and how to protect data. Ultimately, this will require international agreement (see below.)

- **Recognition for data generators and custodians**: The system of academic preferment and grant giving needs to recognise scientists for generating datasets, in the same way as they are today recognised for published papers. That's something Europe's universities, education ministries, publishers and funders should tackle.

## 4. DO develop tools and policies to build trust and data-sharing

**For one scientist to share data with another, there must be trust. Forget for the moment all the technical details: Sharing is a human behaviour, and one that only happens under the right conditions**. People have to trust that, by sharing data with others, they will gain more than they lose. That gain could be something quite selfish – help getting work done, new career contacts, or simply fame. It could be altruistic: advancing knowledge, helping humanity. Whatever, the reward must exceed the risk – which can seem substantial. What if your data get lost or garbled, hacked or deleted? What if they are hijacked by a rival scientist claiming credit for your work? What if they are misused or misinterpreted? Imagine: You're a psychologist studying teenage smoking, and one day you find some of your data selectively mined for pro-industry conclusions in a meta-study funded by Big Tobacco. Shouldn't you have some kind of 'moral right' to protest, as exists under copyright law in many countries?

The good news, of course, is that there is already some progress being made with these problems. On the ethical questions, several research and government organisations around the world have taken a first obvious step: Write a code of conduct, so at least there's no ambiguity about what's considered correct and incorrect behaviour. For instance, the European Commission in 2005 published its first European Charter for Researchers, and in 2010 the European Science Foundation and the Federation of All European Academies published their "European Code of Conduct for Research Integrity." But there remains much more to do – and much of it will be buried inside the infrastructure, as technical features. These include ways of ensuring that the data remain accessible, can't be tampered with, can be easily annotated and accessed, and can be hallmarked so we know where, how and by whom data are generated.

Some of these issues are under study in the various working groups of the Research Data Alliance – but the solutions will need to be adopted by public procurers, research institutions and the researchers themselves. We urge the Commission to facilitate that process, through its funding programmes and its Digital Agenda policies.

## 5. DO support international collaboration

**It's obvious: This is a global effort. Collaboration within the EU is important, but the biggest benefits will come from cross-fertilisation with other regions, cultures and economic systems.** The Research Data Alliance, with its international membership, exemplifies the kind of global coordination that will be needed. But there is much more to do.

- **EU in the lead**: Within Europe, the EU institutions have been most active in coordinating policy for data sharing, and have started to play a leading role internationally. It's important now that, as this field develops, Europe speak with a single voice on data sharing. This must be coordinated across disciplinary and sector silos.

- **Global harmonisation**: The essence of data sharing is to provide equal access. To ensure this happens, there must be global harmonisation of standards and rules.

- **Long-term support**: Data sharing is not a one-off project; it will become part of the fabric of research. As such there must be long-term thinking, across borders, about how this will affect trade, security, economies and our social fabric.

- **Intergovernmental coordination**: Data sharing is poised to become a permanent feature of how science is practised. This means there needs to be a permanent structure – a World Data Organisation – to maintain coordination and supervision. The RDA is an important step in this direction.

## 6. DON'T regulate what we don't yet understand

**One of the greatest risks in government is for politicians to act too soon.** A global data library sounds like a big deal – and it is. But we don't yet know which technologies will work best, which policies will safeguard our privacy and security, and who will benefit most and how. There will be a great temptation – especially in the post-Snowden era – to regulate.

We urge forbearance – at least until the world's scientific community has more experience with the costs and benefits. Issues such as privacy and ethics should be handled in consultation with the data and scientific communities, as well as with society at large.

## 7. DON'T stop what has begun well

**A final warning, for what some might call the blindingly obvious**. A great deal of effort, expense and brainpower, across the EU, has been invested already in making data sharing a reality. It will be a temptation, with a new Commission and Parliament in Brussels, to change course, re-order priorities and move funding lines around. Don't.

## A final word…

Numbers matter. A crowd behaves differently than a few individuals; an avalanche has different dynamics than a snowball; and today's rising volume of scientific data brings a change in quality, not just quantity. It changes the way science is performed, the problems it can tackle, and the speed and nature of the solutions it finds. It generates new products, services and jobs. It challenges old policies – and, in time, will change the way policy itself is made, in all spheres. This requires a bold response. We are starting to see it in major capitals around the world. We are also starting to see a collaborative approach, with the formation of the Research Data Alliance to help coordinate policies. But, equally, the challenges are great: It will take money to build the infrastructure, support for clever pioneers and entrepreneurs, and creative solutions to the policy problems with data privacy and security.

These are all issues which we urge the new European Parliament and Commission to take up, urgently. The strength of Europe's science base is at stake, surely. But more broadly, so is the Union's competitiveness globally – its ability to create jobs, attract investment, and maintain the social fabric Europe values. Exactly how this will all evolve, we don't yet know; but we do know it will be important. As the British Internet pioneer Tim Berners-Lee put it: "The Web as I envisaged it, we have not seen it yet. The future is still so much bigger than the past."

## REFERENCES


[1] cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf
[2] phys.org/news159644537.html#jCp
[3] www.mckinsey.com/insights/high_tech_telecoms_internet/internet_matters
[4] European Commission. "Commission urges governments to embrace potential of big data." IP/14/769. 02/07/2014.
[5] www.copernicus.eu/
[6] Http://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Free_access_to_Copernicus_Sentinel_satellite_data
[7] www.tigernation.org/
[8] www.clarin.eu
[9] www.center-tbi.eu/project/background
[10] www.earthcube.org
[11] European Commission: "Open Access to Research Publications reaching 'tipping point.'"IP13/786. 21/08/2013.
[12] European Commission. "Scientific data: open access to research results will boost Europe's innovation capacity." IP/12/790: 17/07/2012.
[13] http://www.oecd.org/science/sci-tech/38500813.pdf
[14] http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf
[15] European Commission. COM (2007) 56 final of 14.2.2007.
[16] "G8 Open Data Charter": https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex
[17] Davidson, Osha Gray. "Data drives everything (but the bridges need a lot of work.) Earthzine: 11 June 2014. http://www.earthzine.org/2014/06/11/data-drives-everything-but-the-bridges-need-a-lot-of-work/
[18] Desmond, Adrian and James Moore. "Darwin." WW Norton & Co., London: 1991.
[19] Houghton, John. "Economic Impact Evaluation of the Economic and Social Data Service." Centre for Strategic Economic Studies, University of Victoria, March 2012. www.esrc.ac.uk/_images/ESDS_Economic_Impact_Evaluation_tcm8-22229.pdf
[20] battelle.org/media/press-releases/$3.8b-investment-in-human-genome-project-drove-$796b-in-economic-impact-creating-310-000-jobs-and-launching-the-genomic-revolution
[21] www.gov.uk/government/uploads/system/uploads/attachment_data/file/288481/bis-14-618-innovation-from-big-science-enhancing-big-science-impact-agenda.pdf
[22] "Big Data Google Searches Predict Unemployment in Finland" http://www.etla.fi/en/publications/33195/
[23] Markl,Volcker, Thomas Hoeren and Helmut Krcmar. "Innovationspotenzialanalyse für die neuen Technologien für das Verwalten und Analysieren von großen Datenmengen." www.dima.tu-berlin.de/fileadmin/fg131/Publikation/BDM_Studie/bigdatamanagement-short-EN-finalv1.pdf
[24] www.asedie.es/images/asedie%20informe%20del%20sector%20infomediario.pdf
[25] www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information
[26] European Commission. "Have your say on the future of science: public consultation on Science 2.0." IP/14/761. 03/07/2014.
[27] Nature Climate Change 4, 264–268 (2014) doi:10.1038/nclimate2124
[28] www.hack4med.homerproject.eu/info/
[29] www.wellcome.ac.uk/stellent/groups/corporatesite/@policy_communications/documents/web_document/WTP055585.pdf
[30] League of European Research Universities. "The EP´s position on the General Data Protection Regulation threatens EU Research!" 6 October 2014. http://www.leru.org/index.php/public/news/the-eps-position-on-the-general-data-protection-regulation-threatens-eu-research/
[31] curia.europa.eu/jcms/upload/docs/application/pdf/2014-05/cp140070en.pdf
[32] www.facebook.com/permalink.php?story_fbid=304206613078842&id=291423897690447
[33] www.newscientist.com/article/dn10439-who-blasts-china-for-withholding-bird-flu-samples.html#.U8-wMiwo_IU?
[34] www.who.int/mediacentre/factsheets/avian_influenza/en/

**RDA Europe acknowledges the many contributions from:**
Fragkiskos Archontakis, European Commission, Directorate General Research & Innovation, Belgium
Paul Ayris, Director of UCL Library Services and UCL Copyright Officer; President of LIBER; Chair of the LERU community of Chief Information Officers, The Netherlands
Juan Bicarregui, Head of Data Division, STFC, United Kingdom
Rachel Bruce, Innovation Director for digital infrastructure, JISC, United Kingdom
Donatella Castelli, Senior Researcher,"Istituto di Scienza e Tecnologie della Informazione, "A. Faedo" National Research Council, Italy
Patrick Cocquet, RDA Council & CEO Cap Digital, France
Sandra Collins, Chair of the ALLEA E-Humanities Working Group & Director of the Digital Repository of Ireland (DRI), Ireland
Fabrizio Gagliardi, Barcelona Supercomputing Center, Spain
Wolfram Horstmann, Director of the Göttingen State and University Library, Germany
Peter Linton, Senior Advisor - Burson-Marsteller, Belgium
Norbert Lossau, Vice-President, Georg-August-Universität Göttingen, Germany
Marja Makorow, Academy of Finland's Vice President for Research
Monica Tarazona Rua, European Commission, Directorate General Research & Innovation, Belgium
Hans Pfeiffenberger, Head of IT Infrastructure, Alfred Wegener Institut for Polar and Marine Research (AWI), Germany
Raphael Ritz, Head of Data Division, Garching Computing Center of the Max Planck Society, Germany
Herman Stehouwer, Garching Computing Center of the Max Planck Society, Germany
Jens Vigen, Head of the CERN Scientific Information Service, Switzerland
Doris Wedlich, RDA Council & Biology, Chemistry, and Process Engineering Division Head, Karlsruhe Institute of Technology, Germany

**In October 2010, the High Level Group on Scientific Data presented the "Riding the Wave," report to the European Commission outlining a series of policy recommendations on how Europe could gain from the rising tide of scientific data. Over 4 years later, a team of European experts have generated this new report to outline how Europe must act now to secure its standing in future data markets. It offers recommendations to European policy makers while outlining the benefits and challenges. The seeds have been sown. Now is the time to plan the harvest.**

RDA Europe, the European plug-in to the global Research Data Alliance, is funded by the European Commission under the 7th Framework Programme (FP7-INFRASTRUCTURES-2012-1 - ID 312424)