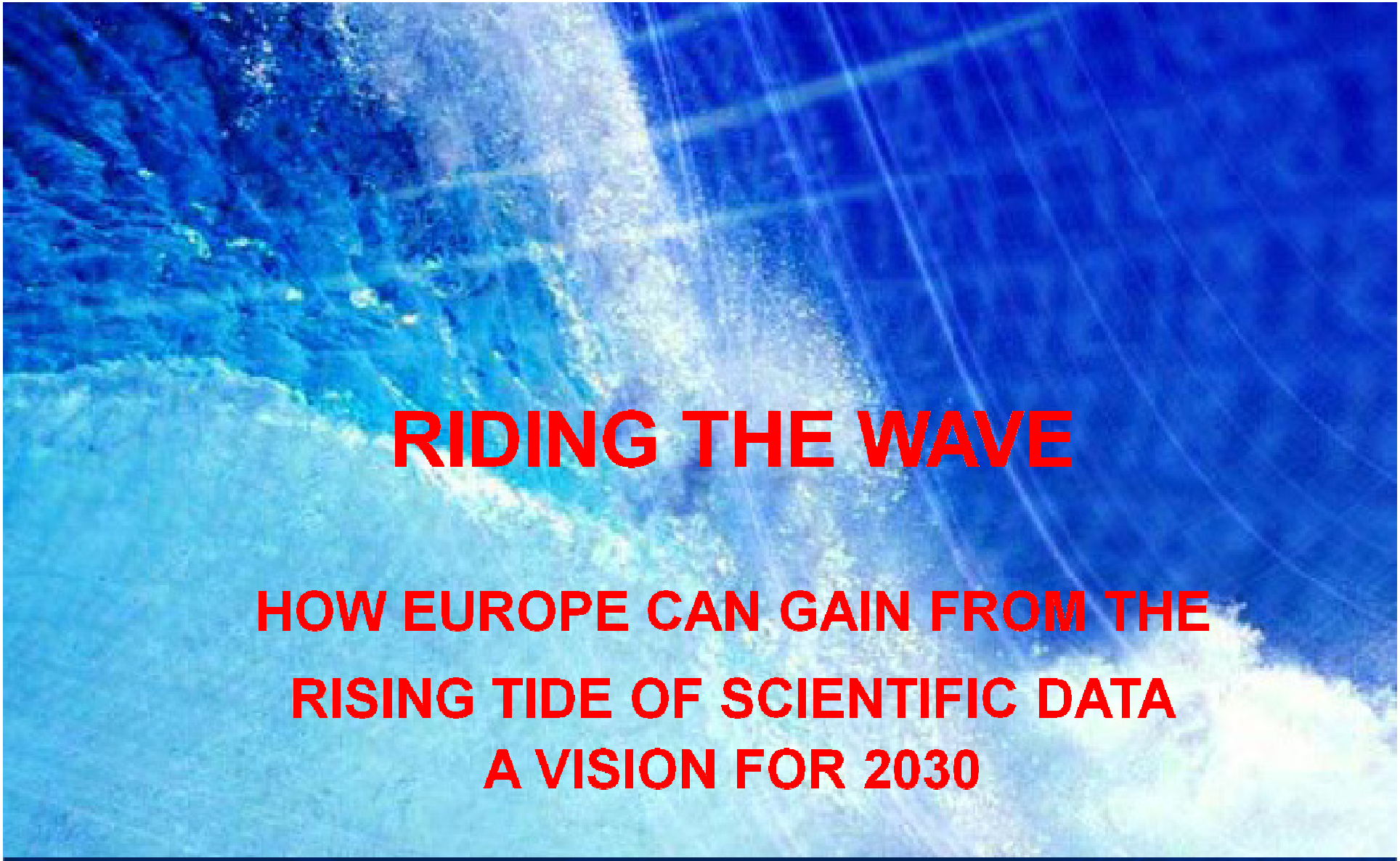


RIDING THE WAVE

HOW EUROPE CAN GAIN FROM THE RISING TIDE OF SCIENTIFIC DATA
A VISION FOR 2030

Report of the High Level Expert Group on Scientific Data

Kostas Glinos, Head for e-Infrastructures, European Commission

Outline

- **Context**
- Vision
- Integration & initial wish list
- Benefits
- Obstacles

Digital Agenda for Europe
the policy context

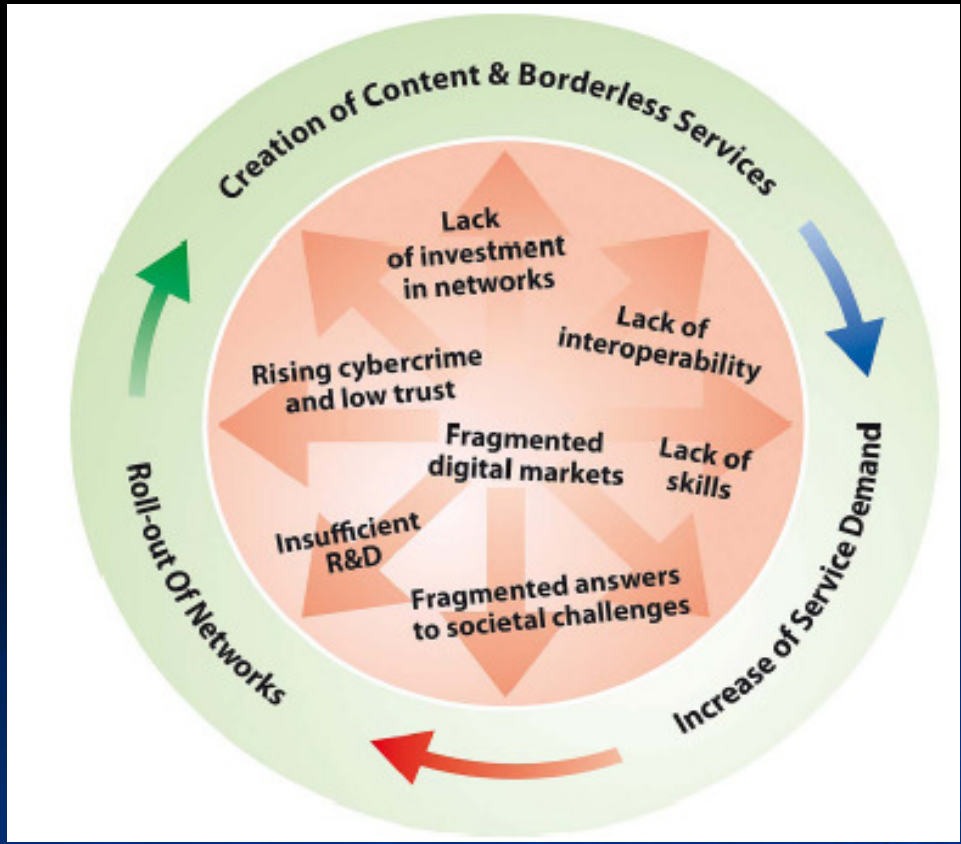
"The Digital Agenda for Europe outlines policies and actions to maximise the benefit of the digital revolution for all. Supporting research and innovation is a key priority of the Agenda, essential if we want to establish a flourishing digital economy."

Neelie Kroes,

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

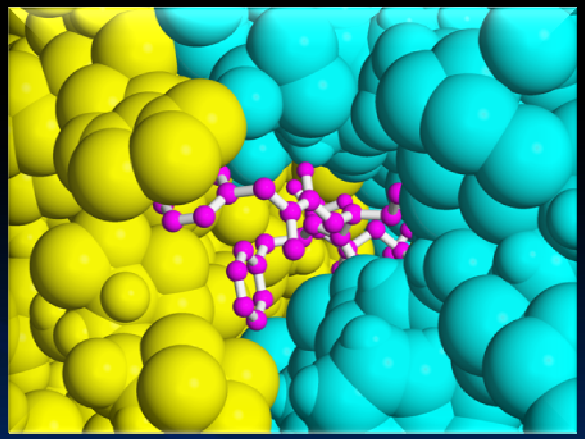
Digital Agenda for Europe
the policy context

DAE is one of the flagships of **“Europe 2020: a strategy for smart, sustainable and inclusive growth”**



Science and ICT

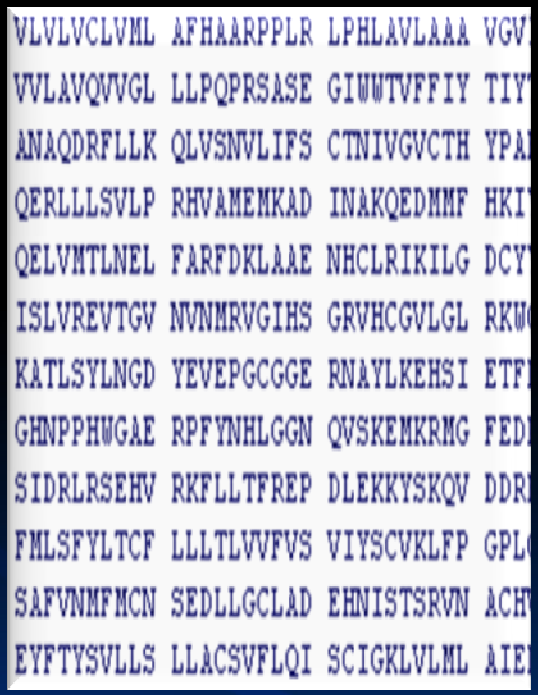
- High-speed communications and advance computation give rise to the era of e-Science.



During the 2006 pandemics alarm, Asian and European laboratories analysed drug components against avian flu using thousands of computers distributed in network grid during 4 weeks!

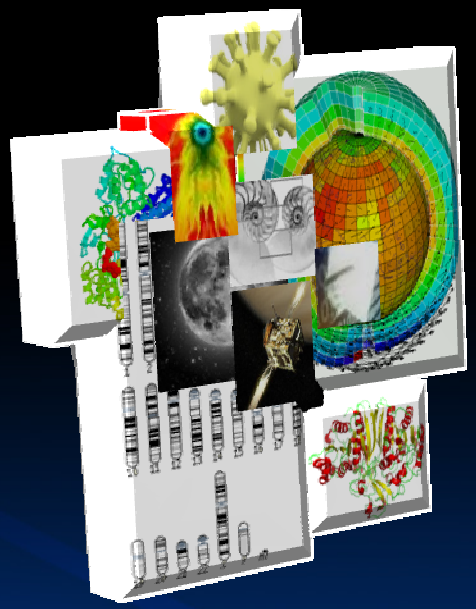
This work would have taken 100 years on a single computer!

Global collaboratories



- With a proper scientific e-Infrastructure, researchers in different domains can collaborate on the same data set, finding new insights.
- They can share the data across the globe, protecting its integrity and checking its provenance.
- They can use, re-use and combine data, increasing productivity.

Global collaboratories



- They can engage in whole new forms of scientific inquiry and treat information at a scale we are only beginning to see.
- … and help us solving today's Grand Challenges such as climate change and energy supply.

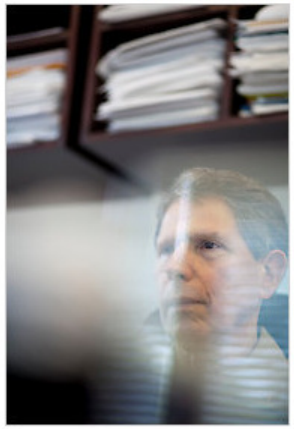
Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

In 2003, a group of scientists and executives from the National Institutes of Health, the Food and Drug Administration, the drug and medical-imaging industries, universities and nonprofit groups joined in a project that experts say had no precedent: a collaborative effort to find the biological markers that show the progression of Alzheimer's disease in the human brain.

RECOMMEND
TWITTER
COMMENTS (155)
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Enlarge This Image



Michael Temchine for The New York Times

Neil Buckholtz, chief of the Dementias of Aging Branch at the National Institute of Aging, in the National Institutes of Health.

Now, the effort is bearing fruit with a wealth of recent scientific papers on the early diagnosis of Alzheimer's using methods like PET scans and tests of spinal fluid. More than 100 studies are under way to test drugs that might slow or stop the disease.

And the collaboration is already serving as a model for similar efforts against Parkinson's disease. A $40 million project to look for biomarkers for Parkinson's, sponsored by the Michael J. Fox Foundation, plans to enroll 600 study subjects in the United States and Europe.

The work on Alzheimer's "is the precedent," said Holly Barkhymer, a spokeswoman for the foundation. "We're really excited."

The key to the Alzheimer's project was an agreement as ambitious as its goal: not just to raise money, not just to do research on a vast scale, but also to share all the data, making every single finding public immediately, available

Multimedia

Backstory With the Times's Gina Kolata

What's Popular Now

1938 in 2010

Obama to Call for $50 Billion Public Works Plan

Well
Tara Parker-Pope on Health

Vegetarian Recipes for Barbecue Season
September 3, 2010

Sunday Shopping Linked With Less Happiness
September 3, 2010

Creating a Safer Kitchen
September 3, 2010

Testing the Bonds of Doctor and Patient
September 2, 2010

Do Fluorescent Lights Trigger Migraines?
September 2, 2010

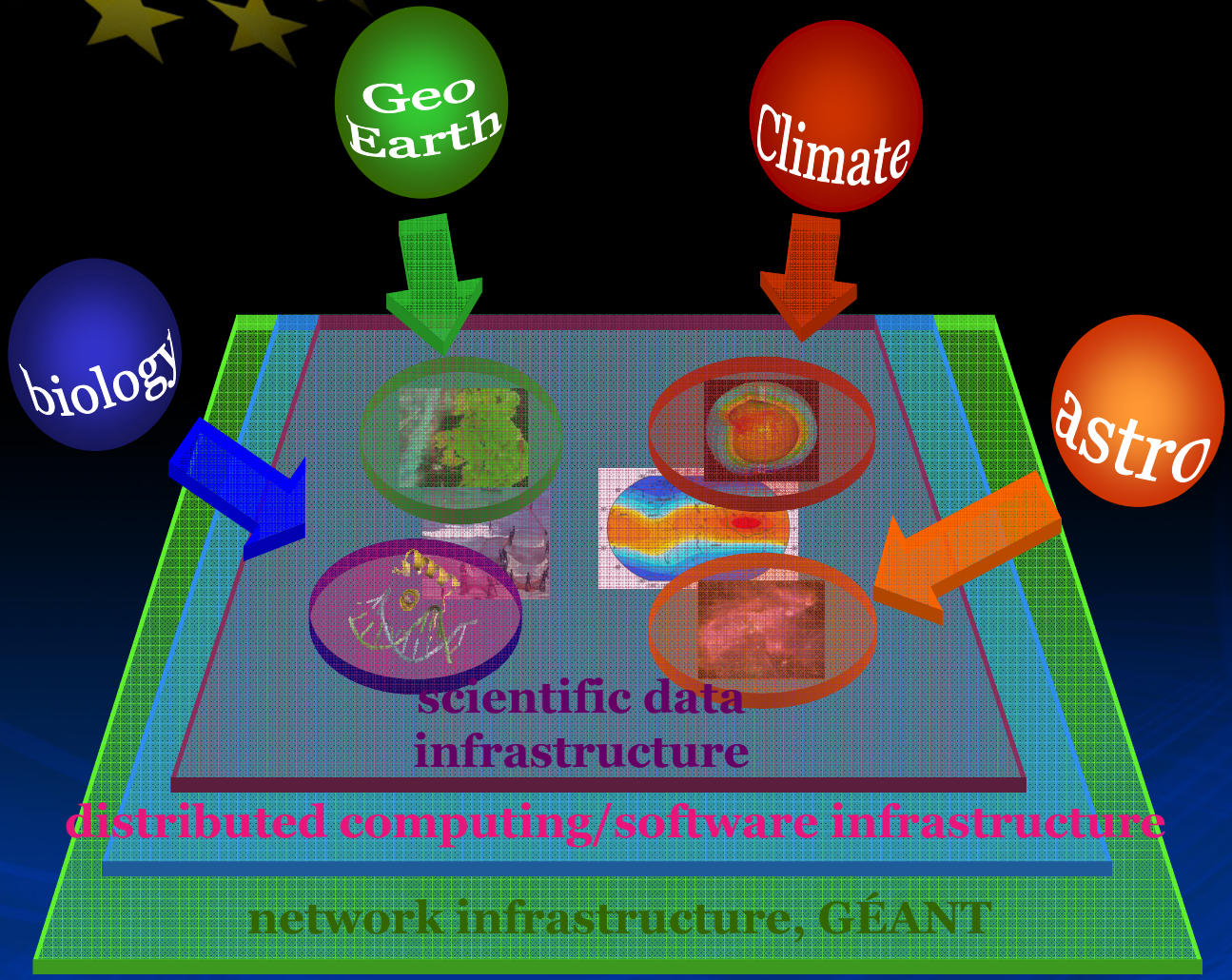
Get the Opinion Today E-Mail

Sign up for the highlights of the day in Opinion, sent weekday afternoons.

Sign Up

See Sample | Privacy Policy

Scientific Data Infrastructure

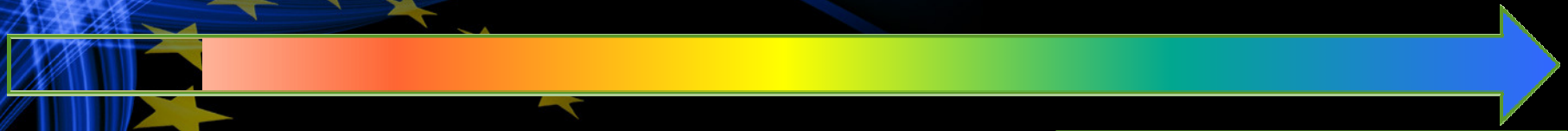


Rising tide of data…

"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."

Report of the High-Level Group on Scientific Data, October 2010

"Riding the Wave: how Europe can gain from the raising tide of scientific data"



1990

- Web not yet begun
- XML not yet begin
- Internet speeds kbps in universities and offices
- 300,000 internet hosts
- Data volume ??
- XXX researchers
- Few computer programming languages
- Transition from text to 2D image visualisation

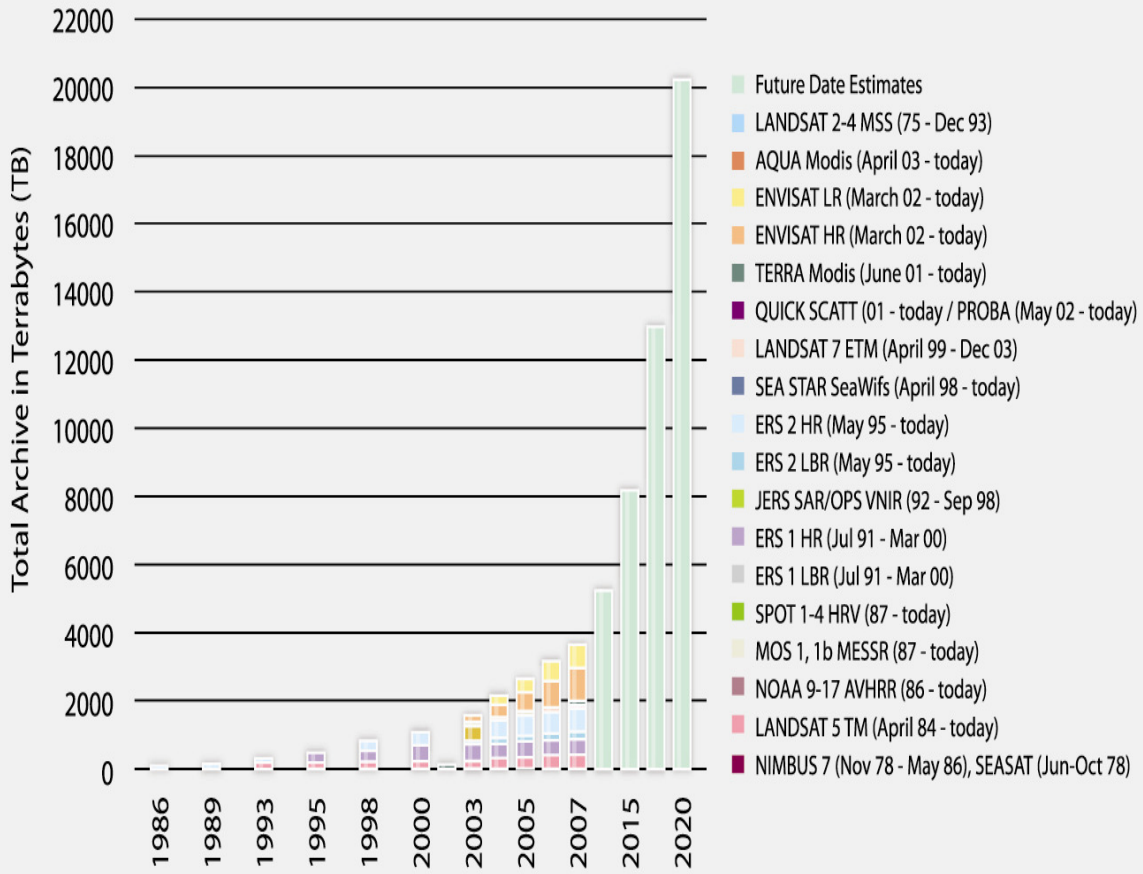
2010

- Web 2.0 started
- XML widespread
- Internet speeds Mbps widespread
- 600,000,000 internet hosts
- 5.10^{18} bytes of data
- Millions of researchers
- Many new paradigms for programming languages
- 3-D and Virtual reality visualisation

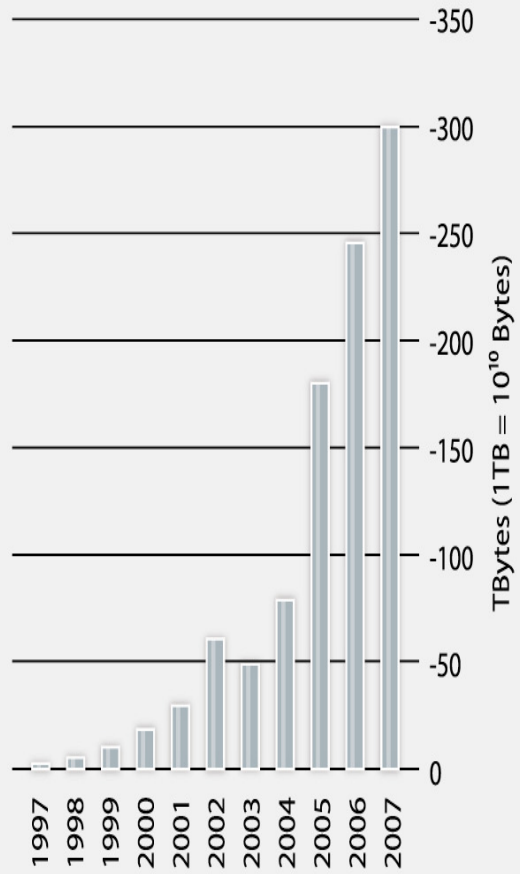
2030

- Semantic Web
- XML forgotten
- Internet speeds Pbps widespread
- 2,000,000,000,000 hosts
- 5.10^{24} bytes of data
- Billions of citizen researchers
- Natural language programming for computers
- Virtual worlds

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



Yearly Data Creation on NICE



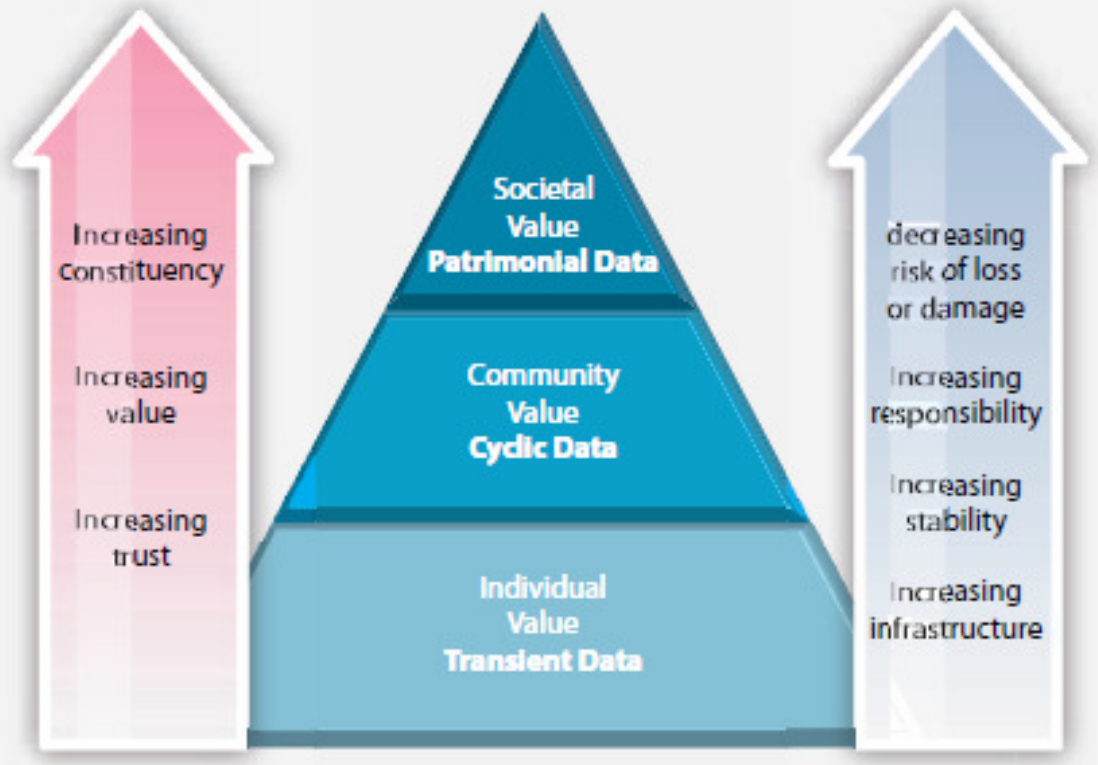
The data pyramid - a hierarchy of rising value and permanence

Digital Data Collections

Reference, nationally and internationally important, irreplaceable data collections

Key research and community data collections

Personal data collections



Respositories/ Facilities

National- and international-scale respositories, libraries, archives

"Regional" - scale libraries and targeted data archives and centers

Private respositories

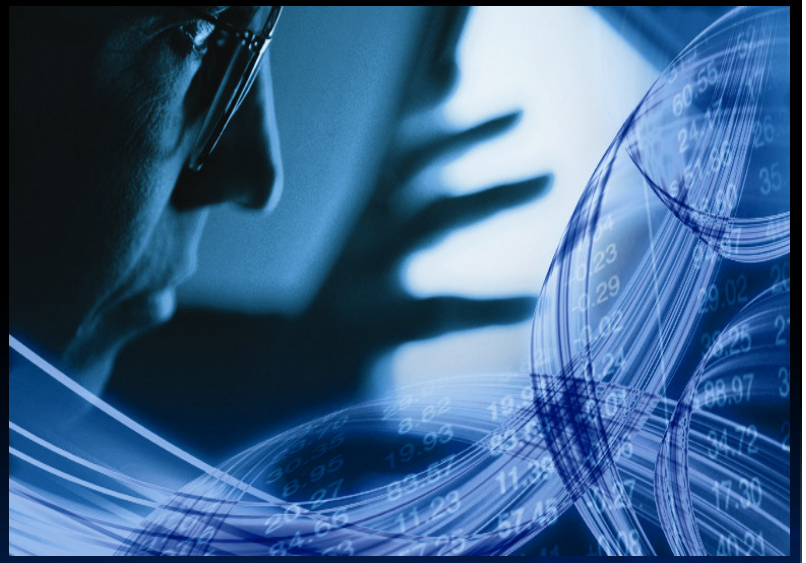
Source: Adapted from Francine Berman, UC San Diego, in Communications of the ACM.

Outline

- Context
- **Vision**
- Integration & initial wish list
- Benefits
- Obstacles

Vision 2030
high-level experts group on Scientific Data

"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."



High-Level Group on Scientific Data

"Riding the Wave: how Europe can gain from the raising tide of scientific data"

Vision 2030

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

- 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.
- Legal issues are worked out so that they encourage, and not impede, global data sharing.
- The scientific community is supported to provide its data and metadata for re-use.
- Every funded science project includes a fixed budget percentage for compulsory conservation and distribution of data, spent depending of the project context.

IMPACT IF ACHIEVED

- Data form an infrastructure, and are an asset for future science and the economy.

Vision 2030

(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 the data can be trusted.

- 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.
- Propose a directive demanding that data descriptions and provenance are associated with public (and other) data.
- Create a directive to set up a unified authentication and authorisation system.
- Set Grand Challenges to aggregate domains.
- Provide “forums” to define strategies at disciplinary and cross-disciplinary levels for metadata definition.

IMPACT IF ACHIEVED

- 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.

Vision 2030

(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 work to international standards, to ensure they are trustworthy.

- Propose reliable metrics to assess the quality and impact of datasets.All agencies should recognise high quality data publication in career advancement.
- Create instruments so long-term (rolling) EU and national funding is available for the maintenance and curation of significant datasets.
- Help create and support international audit and certification processes.
- Link funding of repositories at EU and national level to their evaluation.
- Create the discipline of data scientist, to ensure curation and quality in all aspects of the system.

IMPACT IF ACHIEVED

- Data-rich society with information that can be used for new and unexpected purposes.
- Trustworthy information is useable now and for future generations.

Vision 2030

(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.

- ❑ EU and national agencies mandate that data management plans be created.

IMPACT IF ACHIEVED

- ❑ Funders have a strategic view of the value of data produced.

Vision 2030

(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 for both.

- Use the power of EU-wide procurement to stimulate more commercial offerings and partnerships.
- Create better collaborative models and incentives for the private sector to invest and work with science for the benefit of all.
- Create improved mobility and exchange opportunities.

IMPACT IF ACHIEVED

- Commercial expertise is harnessed to the public benefit in a healthy economy.

Vision 2030

(6) The public has access 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.

- Create non-specialist as well as specialist data access, visualisation, mining and research environments.
- Create annotation services to collect views and derived results.
- Create data recommender systems.
- Embed data science in all training and academic qualifications.
- Integrate into gaming and social networks

IMPACT IF ACHIEVED

- 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.

Vision 2030

(7) Policy makers can make decisions based on solid evidence, and can monitor the impacts of these decisions. Government becomes more trustworthy.

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

IMPACT IF ACHIEVED

- Policy decisions are evidence-based to bridge the gap between society and decision-making, and increase public confidence in political decisions.

Vision 2030

(8) Global governance promotes international trust and interoperability.

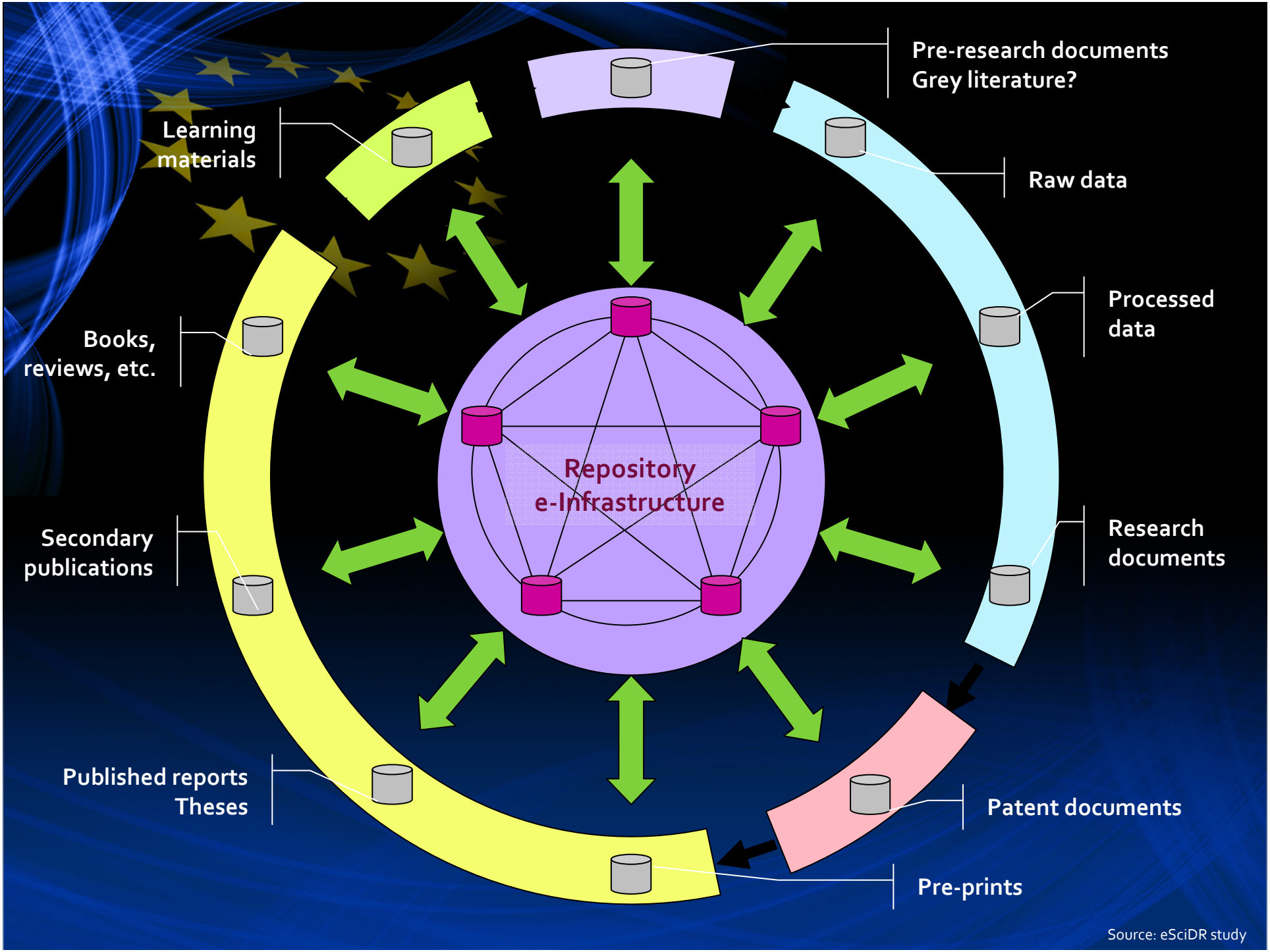
- Member states should publish their strategy, and resources, for implementation, by 2015.
- Create a European framework for certification for those coming up to an appropriate level of interoperability.
- Create a “scientific Davos” meeting to bring commercial and scientific domains together.

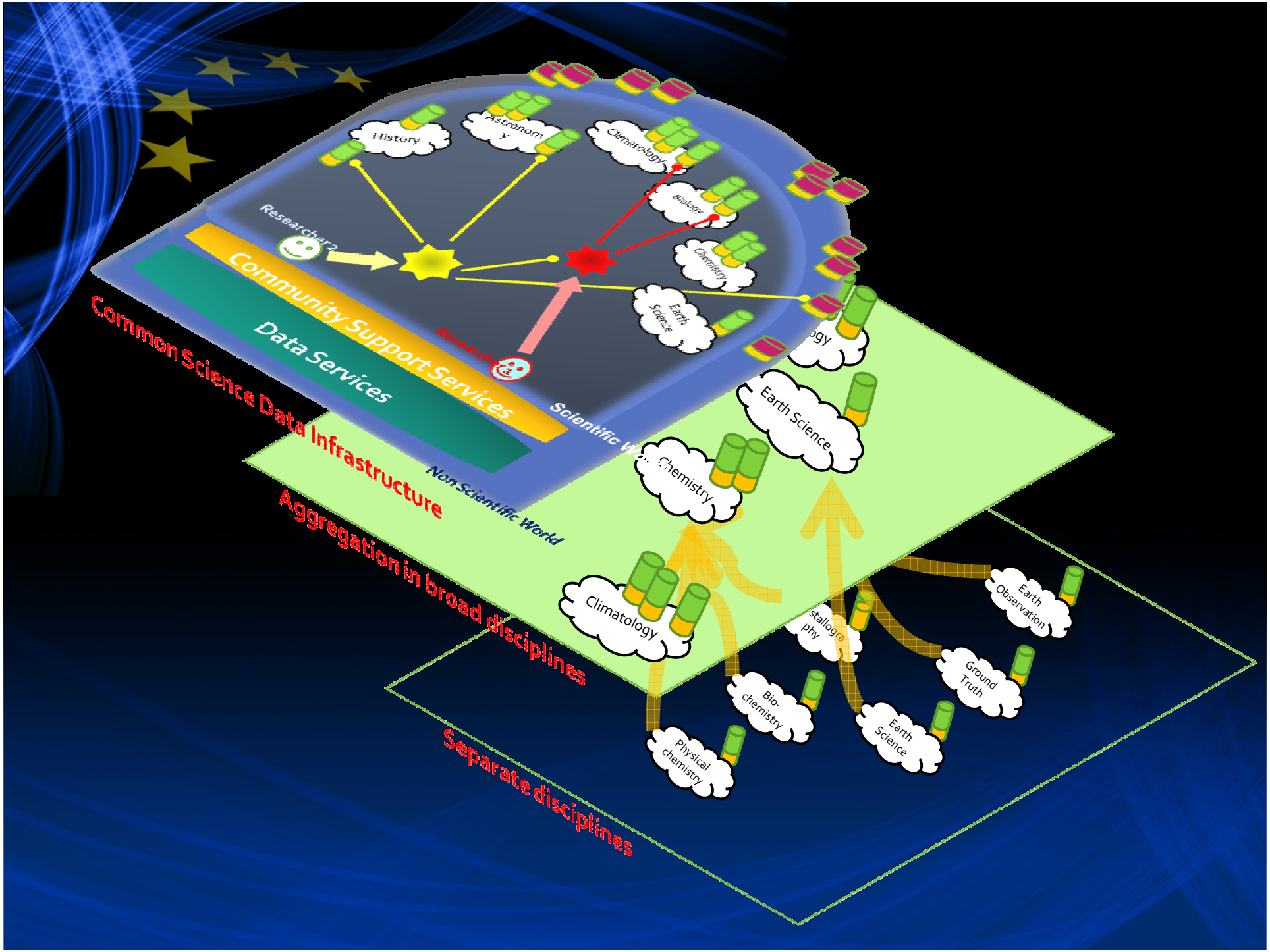
IMPACT IF ACHIEVED

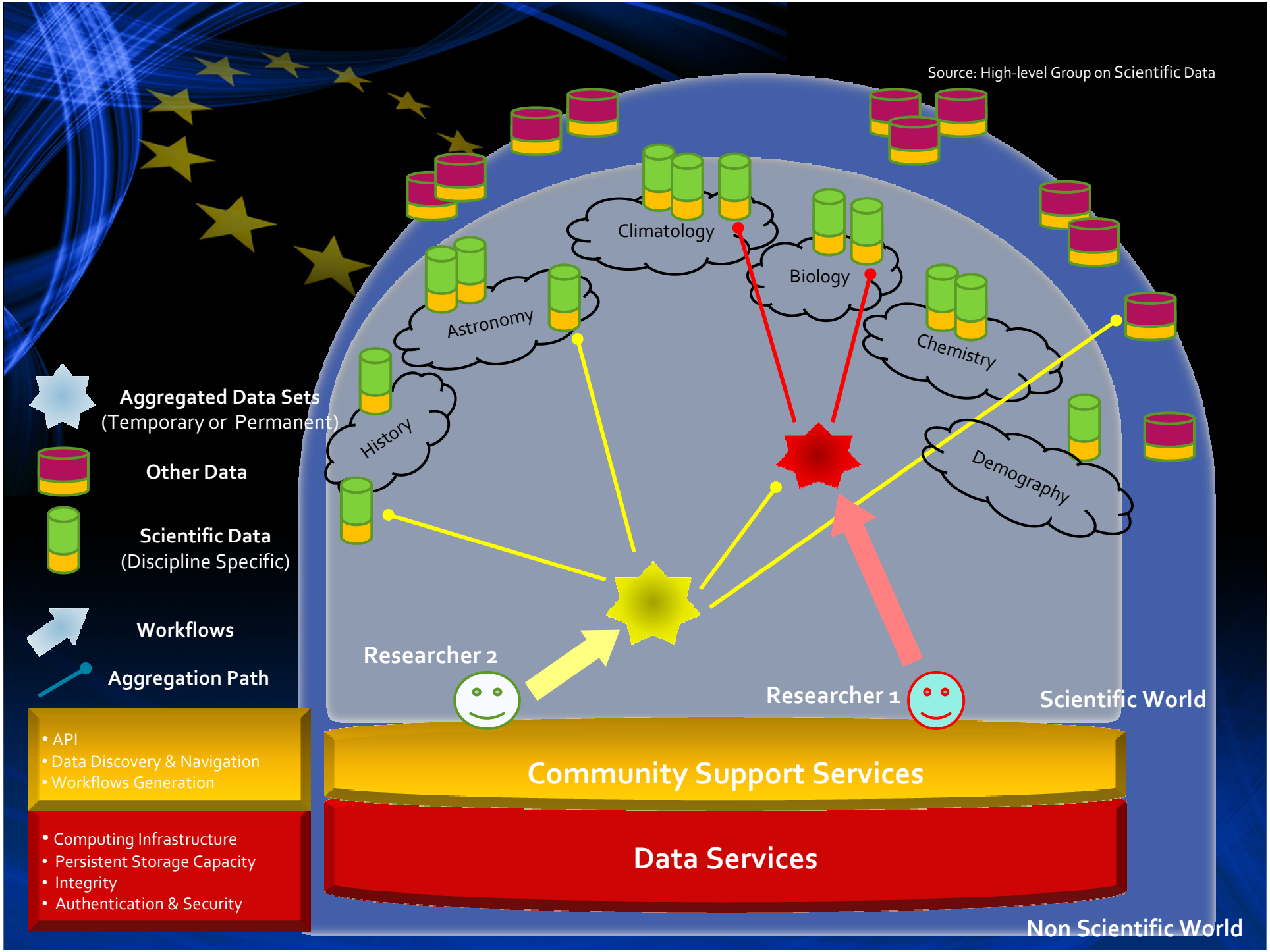
- We avoid fragmentation of data and resources.

Outline

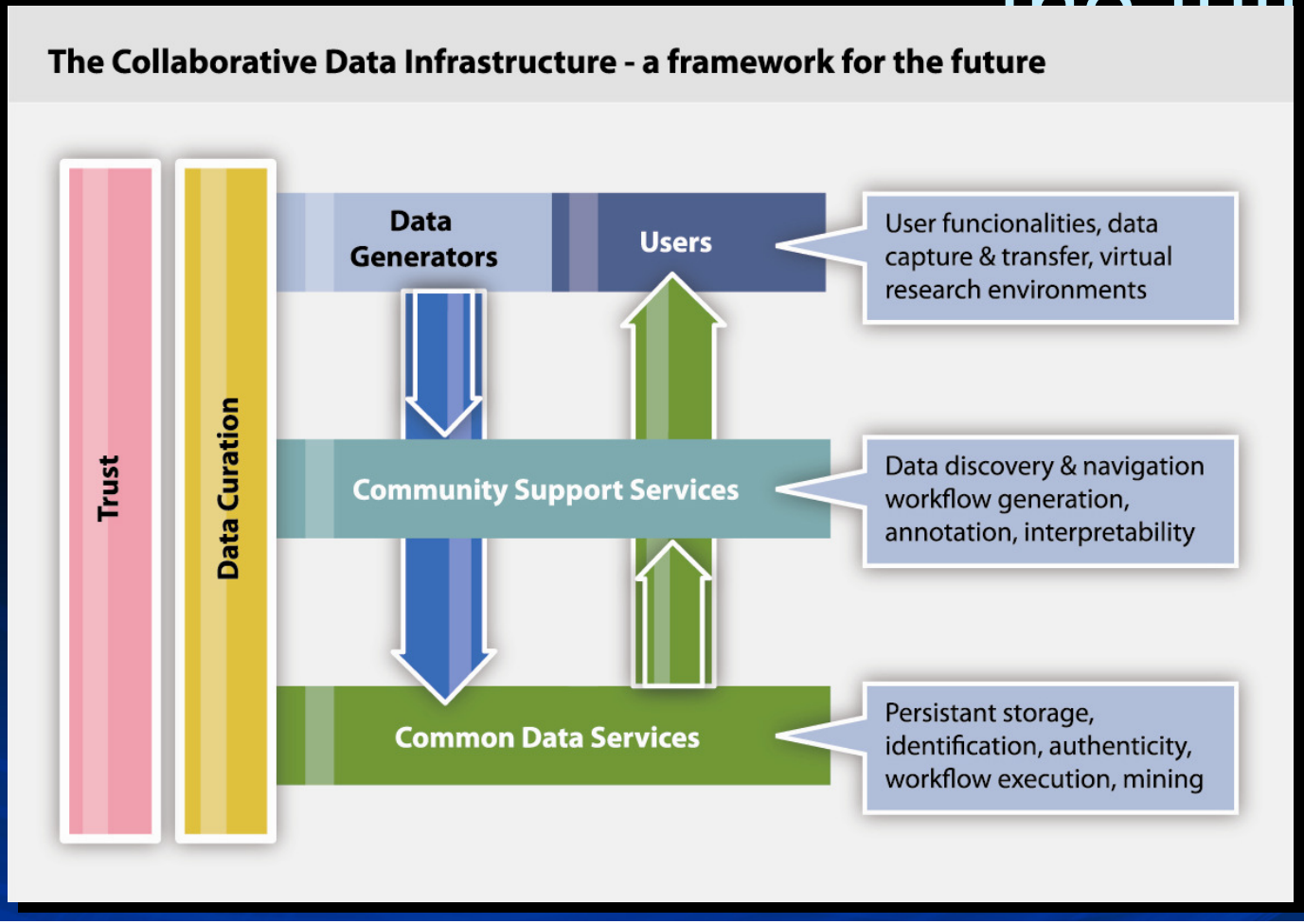
- Context
- Vision
- **Integration & initial wish list**
- Benefits
- Obstacles


Pre-research documents
Grey literature?
Learning materials
Raw data
Processed data
Books, reviews, etc.
Repository e-Infrastructure
Secondary publications
Research documents
Published reports
Theses
Patent documents
Pre-prints
Source: eSciDR study


History
Astronomy
Climatology
Biology
Chemistry
Earth Science
Researcher 2
Community Support Services
Data Services
Researcher 1
Scientific World
Non Scientific World
Common Science Data Infrastructure
Aggregation in broad disciplines
Separate disciplines
Earth Science
Chemistry
Climatology
Earth Observation
Ground Truth
Bio-chemistry
Physical chemistry
Earth Science


Source: High-level Group on Scientific Data
Climatology
Biology
Astronomy
Chemistry
History
Demography
Aggregated Data Sets
(Temporary or Permanent)
Other Data
Scientific Data
(Discipline Specific)
Workflows
Aggregation Path
Researcher 2
Researcher 1
Scientific World
• API
• Data Discovery & Navigation
• Workflows Generation
Community Support Services
• Computing Infrastructure
• Persistent Storage Capacity
• Integrity
• Authentication & Security
Data Services
Non Scientific World

A collaborative Data Infrastructure – a framework for the future



Initial wish list

- 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 date
- 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

Outline

- Context
- Vision
- Integration & initial wish list
- **Benefits**
- Obstacles

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

Outline

- Context
- Vision
- Integration & initial wish list
- Benefits
- **Obstacles**

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

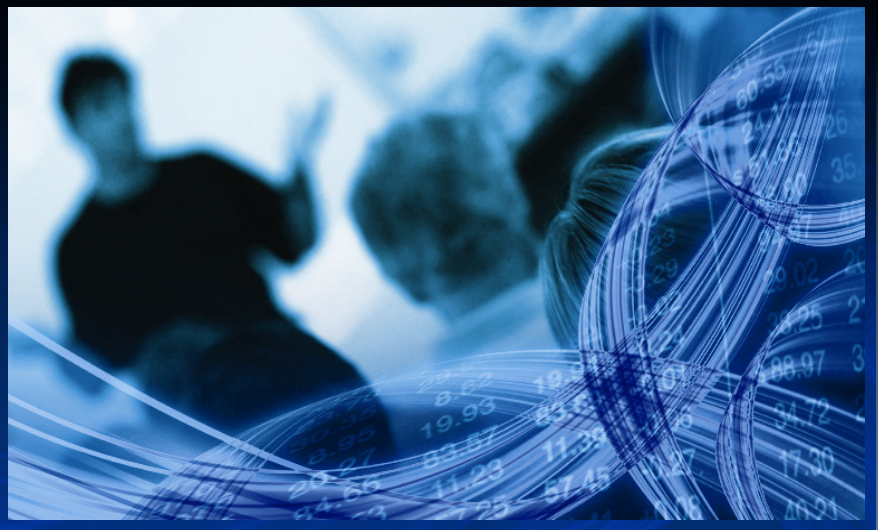
Digital Agenda for Europe

”Making this a reality is a more difficult task…”

Vice-President Neelie Kroes, Commissioner for the Digital Agenda, received the HLG report from the chairman of the group, John Wood, on 6 October 2010.

This report on Scientific Data will be an invaluable input for formulating the European research and research infrastructure policies.

All, citizens and organisations, are invited to take note of this report and use it as background reference when discussing EU priorities.



e-Infrastructures
underpinning a creativity machine…

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

Vernor Vinge

(*Nature, Vol 440, March 2006*)

Members of High Level Expert Group on Scientific Data

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

Thomas Andersson – Prof. of Economics and former President, Jönköping University; Senior Advisor, Science and Innovation, Sultanate of Oman

Achim Bachem - Chairman, Board of Directors, Forschungszentrum Jülich

Christoph Best - European Bioinformatics Institute, Cambridge (UK) and Google UK Ltd, London

Françoise Genova - Director, Strasbourg Astronomical Data Centre; Université de Strasbourg/CNRS

Diego R. Lopez - RedIRIS

Wouter Los - University of Amsterdam; Coordinator of 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, CERN

Peter Wittenburg - Technical Director, Max Planck Institute for Psycholinguistics

Rapporteur: **David Giaretta -** STFC and Alliance for Permanent Access