Consultation of the Horizon 2020 Advisory Groups: Providing advice on potential priorities for Research and Innovation in the Work Programme 2016-2017

Response of the
Horizon 2020 Advisory Group on European Research Infrastructures including e-Infrastructures

Table of Contents

15/07/2014

Introduction

The European Research Infrastructures (RI) are at the centre of ERA and Horizon 2020. **Their prime mission is to serve and advance scientific excellence***. In doing so they may also have a pivotal role in relation to the aims of European industrial leadership and in tackling grand societal challenges. Further, it must be acknowledged that some RI developed to meet demanding science needs may not be compatible with short term innovation goals.

The research and innovation capabilities of each region/nation determine its position in the globalised knowledge economy. Access to top level RI is crucial for realizing its full potential in this respect. To this end, RI go far beyond concentrations of scientific instruments and data sets. They may also form intellectual and cultural poles at the centre of innovation ecosystems. In this capacity, they have a significant role also at regional/ national levels, as enablers of Research and Innovation Smart Specialisation Strategies (RIS3).

In the current times of economic and societal crises, the prioritisation of RI is critical. In this respect, the Advisory Group identified that the design, construction and operation of pan-European RI face four major challenges:

1. Expanding their role and impact in the innovation chain (**Annex 1**);
2. Ensuring their sustainability and optimal life-cycle (**Annex 2**);
3. Maximizing data exploitation (**Annex 3**) ;
4. Assessing evaluation procedures (**Annex 4**).

Working groups were set up for each one of these major challenges to prepare position papers which are provided as annexes. The working groups further identified three cross-cutting challenges for RI:

- Openness and global character aiming at leadership in research and innovation;
- Skills development, education and training opportunities;
- Maximizing socio-economic benefits related to regional, national and cohesion strategies

In this context, the Research Infrastructures Advisory Group deliberated the questions posed by the European Commission and, following in-depth discussion and investigation, reached consensus on the critical points to address these questions as follows.

Question 1: What are the biggest challenges which require immediate action under the next Work Programme? Which related innovation aspects could be targeted?

Challenge 1: Expanding the role and impact of RI in the innovation chain

- **Exploit the opportunities of RI to push the limits of technological research and development and promote ambitious innovation**
  Seek RI proposals based on scientific excellence needs and exploit opportunities of fostering the development of components, instruments, services and knowledge that push the limits of existing technology challenging industry (including SMEs) to meet new development and operation requirements.

- **Expand users and multistakeholders of RI**
  Promote knowledge about RI and enhance the effective engagement in RI of scientists involved in research projects and research organizations, of industry (including SMEs) by contract research or through employees, and of professors, teachers, students, public administrations, policy makers, citizens. Promote the clustering of complementary RI for promoting multi/cross-disciplinary collaboration, sharing of best practices and cross-fertilization. Communicate the mutual benefit for all involved. Facilitate transnational access to and use of RI and promote usage. Ensure access and use procedures are clearly communicated and diffused.

Challenge 2: Ensuring sustainable funding and optimal life-cycle of RI

- **Ensure long-term sustainable funding and adoption of best user/stakeholder engagement and management practices**
  Foster a long-term view of RI lifecycle and carefully crafted organizational concepts for pursuing sustainable high-end knowledge and technology evolution and innovation. Long-term sustainable funding of research and e-Infrastructures (networks, data and computation) is essential for trust and acceptance between users and infrastructure providers. Long-term funding needs to be based on the involvement of all stakeholders, including the European Commission, as in GÉANT; self-sustainability of RI depends on their nature and some may require partial continued support. Investment and operational costs should be made explicit as well as the associated sources of funds. Flexible business models, taking into account all investment and operational cost items and the various sources of cost coverage, are essential (i) to keep RI sustainable in the long run, (ii) to cope with societal, political and dynamic technological changes, and (iii) to initiate Public Private Partnerships (PPP).

  EU project support tools and funding schemes should be used with full flexibility to allow support (Integrating Activities should be open to application by infrastructures with pertinent partners). Funding should be sustainable also after the establishment of the RI. Also, RI should be supported to bridge the funding gaps between planning, construction and operational phases (avoiding “near death experiences” when the funding of one phase ends without new funding for the next phase, risking to lose the people involved). This demands careful planning of the RI from the very beginning (including construction and operation).

Challenge 3: Maximizing data exploitation

- **Foster the exploitation of Data and Knowledge Infrastructures**
  Future RI are expected to contribute to an open ecosystem of services where data-driven science will blossom in tandem with socio-economic innovation by promoting the sharing and the interoperability of data, scientific results, analytical processes and services in single sited and in distributed and virtual RI. Appropriate harmonized big data management to enable cost-effective and reproducible science and technologies for acquisition, curation and analysis of data, together with methods to validate findings and communicate discoveries with visualization and story-telling must be ensured. Safety and quality of storage and access to data and the IT infrastructure must be addressed. Novel methods are needed to discover and exploit hidden features and relations in huge research data collections by non-conventional, revolutionary techniques. The (re-)use of research data and data infrastructures poses novel ethical, privacy, security, copyright and IPR challenges, and requires the development of appropriate legal, social and technical frameworks. Furthermore, setting up some RI and/or the associated services may require standardization and the assurance of interoperability, even though some others may have an emergent character in their own. Gender should also be considered as a factor in maximising data exploitation, without prior assumption that women and men will automatically equally benefit and with due acknowledgement of the full range of social and economic impacts, whether positive or negative.

  Bottom-up initiatives, such as Research Data Alliance (RDA), are important stakeholders of RI and need to be nurtured due to the essential role they may play in engaging communities and developing open standards well-conceived for different data types and supported by the involved research community.

Challenge 4: Assessing Evaluation procedures of RI

- **Strengthen accountability and evaluation practices**
  Identify all national evaluation bodies/agencies relevant to RI and promote joint work of national and European bodies/agencies in developing harmonised ex-post evaluation and monitoring mechanisms and conducting evaluation exercises together. Long-term commitments require transparency, quality criteria, evaluation procedures and opt-out options. Improve ex-ante evaluation processes based on rigorous metrics and carried out by experienced experts and scientists who are willing to include, where applicable, development potential and societal impact (with attention to characteristics such as gender, age, ethnicity) in their assessments. Such rigorous criteria should be applied for new RI. Strengthen the role of ESFRI under the above conditions.

Cross-cutting Challenges:

- **Enhance openness and the global character of RI to enable Europe to play a leading role in science and innovation**
  Aim at European global leadership on ground-breaking Science. Promote RI that can function as interaction and joint work platforms for bringing together a complementary diversity of disciplines, skills, concerns, approaches, applications, exploiting the potential of open RI and open data at global scale thriving in wide international cooperation. Ensure agreements on cross-borders RI and multiple agents' participation, which also address ownership and sharing issues. Facilitate

transnational access and use of RI, and coverage of the associated costs. Make sure that formal access and use procedures to big RI provide guarantees of equal access and use. Unleash the potential for the enormous quantity of data produced, gathered, managed, analysed, and transformed by European research projects to be made available as a good for society, citizens, businesses and public bodies and services; foster the development of innovative technologies and standards on search, retrieval and discovery technologies, data and text mining, data citation and the linking of data with experiments and publications, analytics, semantics, and exploit opportunities for developing an innovative marketplace for data and services.

- **Foster the development of skilled specialists and users in connection with RI and exploiting education and training opportunities at and with RI**
  Link RI with education and training of researchers, technicians, data scientists and analysts, data curators and archivists, information specialists, students, users, managers. Build up synergies between RI, research projects, and education and training funding instruments. Look at the connection of RI to education and training opportunities as a contribution to enlarge impact, develop a skilful workforce, promote long-term sustainability and the emergence of new ideas and applications. Promote new appointment opportunities in the academic system for staff scientists, engineers, data specialists, with career progression paths and suitable performance metrics.

- **Exploit opportunities of maximum socio-economic benefit of RI in connection with regional and national strategies and funds**
  Develop effective links between European research and regional development funding instruments and establish management processes to assure execution. RI are important enablers of “smart specialisation” strategies. Establish mechanisms for aligning national and European strategies, RI roadmaps and funding programs. Look for new RI location opportunities that promise best socio-economic regional impacts. Launching Regional Partner Facilities (RPF) is a step towards this end.

Question 2: What are the key assumptions underpinning and driving the future developments and expectations for research infrastructures and e-infrastructures, in particular in relation to the challenges identified above (for example, regarding research & innovation, demand side and consumer behaviour, policy needs, or the concerns and expectations of citizens and civil society) ?

- **Europe faces fierce international competition in RI and international cooperation is essential**
  Competitive regions (e.g. Japan, US, China) invest enormous amounts in RI for maintaining their research at a competitive, if not leading, edge. Europe must invest substantial funding to ensure a leading position. Even if EU has a leadership in some key technologies (accelerator technologies and others), it needs to maintain such a position. However, the extensive cross-border and cross-cultural European experience in RI usage has considerable potential to offer examples in leading RI to other regions of the world. There is a real opportunity for EU to transfer technological and organizational knowledge to emerging countries investing in researchers that do not always have the appropriate technical or organizational

knowledge. International cooperation in RI is important for effectiveness and leadership, especially when there are mutual benefits.

- **Science is increasingly data-intensive, multidisciplinary, collaborative and global**
  Research communities increasingly acknowledge the key importance of data for research and innovation. "Big Data" has to be harnessed into efficient, interoperable and global Research Data Infrastructures. Data accessibility and long-term persistence is assumed to improve over time, but requires proper data organizations and a landscape of trusted and certified repositories that not only store data but also participate and contribute in the continuous enrichment of data sets and services.

- **Digitization will go on in all research disciplines**
  The volume of digital research data will continue to increase massively. Data management, sharing and analyzing needs international cooperation, as does dealing with big societal challenges in general.

- **Collaboration of RI with Universities and Research Organisations within vibrant research and innovation ecosystems may help to limit "brain drain" and encourage "brain circulation"**
  Globally leading RI developed to fulfil the needs of research excellence and sustained by vibrant research and innovation communities and organisations will provide important focuses of talent attraction. New means of strengthening communities around RI, including ERIC type of organisation are needed.

- **The operation of RI within research and innovation ecosystems involving Universities, Research Organisations, companies (including SMEs) and funding and support agencies/mechanisms will enable knowledge creation and facilitate translation of research results into wealth**
  The role of research as the driving edge of innovation will continue. More risk-taking is needed in promising new technologies by European industry and experience gained in leading RI will have an effective role to play. When properly operated and well knitted to research communities and other stakeholders RI may have a especial role to overcome the so-called European paradox regarding the gap between excellence in research and comparatively modest wealth creation results.

- **Some future or existing RI need new high level technologies that must be developed in partnership with industry as generic technologies**
  Such new generic technologies may create an attractive market, beyond the RI market. As an example, new forms of PPP should be developed and introduced, by applying fair and mutually advantageous business models.

- **Europe cannot afford the support of all RI expected by the scientific community**
  There is a need to prioritize RI and to develop a rationale for initial go/no-go decisions based on proper ex-ante evaluation and possible continue/opt-out decisions based on ex-post evaluation. This careful approach is to be combined with global prioritisation and world-wide distribution/integration of remotely accessible high value RI facilities.

Question 3: What is the output that could be foreseen, what could the impact be, what would success look like, and what are the opportunities for international linkages?

- **A landscape of first-class sustainable RI and services open to all researchers and to industry, and other interested groups such as policy makers and the public**
  An open cross-border and trans/cross-disciplinary market place for RI services with engaged stakeholders and organisations can have high impact on the generation of new jobs and the potential to drive a whole new economy impacting on discovery acceleration and innovation and opening new avenues for international competitiveness.

- **Reduced fragmentation and costs**
  A balanced and sustained landscape of infrastructures with symbiotic trends will contribute to reduce fragmentation costs and redundancy, allowing more efficient work environments in which researchers can better focus on what they are trained to do and enabling them to efficiently achieve more relevant results. The support of clustering activities is important to reduce fragmentation in the landscape and to build stronger European nodes for international collaboration.

- **Examples of indicators of success are:**
  - **Attraction of talented young researchers and prominent scientists**, from Europe and outside Europe, to and around RI;
  - **Size and quality of education and training** programmes at and with RI;
  - **Effective international linkages** of RI;
  - **Discovery of completely new fields of research – and the way research is done** at or based on RI;
  - **Transdisciplinarity in addressing complex global problems**;
  - **Contributions to push the limits of technology** through the development of novel materials, instruments and services;
  - **Regional economic growth** due to transformational effects of local RI;
  - **Sharing of regional, national, industry, education** funding for Horizon2020 RI projects;
  - **Alignment of strategies and evaluation practices** for RI at national and European levels.

- **International linkages can be enabled by leading edge RI attractive to worldwide researchers and other stakeholders and by setting standards and increasing interoperability of data infrastructures**
  Interoperable and open innovative infrastructures, supported by virtual research environments, will allow for better collaboration and intercultural communication in Europe and worldwide. Common standards and increased interoperability would allow for wider data sharing and enriched research data resources.

- **Improved quality of RI management**
  Education and training of specialists in RI management (organisational, technical, financial, etc.) and the strengthening of accountability and evaluation practices will contribute to better RI management.

Question 4: Which are the bottlenecks in addressing these areas, and what are the inherent risks and uncertainties, and how could these be addressed?

- **Lack of Human Resources**
  Constructing and operating RI can only be done if a critical mass of researchers in different countries and key scientists are explicitly committed to the RI, if they are strongly engaged in building the case for the need of the RI and in its design, and if highly qualified experts are available. For researchers, infrastructure building is often not attractive as a career option and further infrastructure experts (such as, special equipment operators, experts in modern analytical techniques, data scientists, information specialists, technologists, communicators) are scarce. These bottlenecks need to be overcome.

- **Risk of non-robust priority setting**
  In the political environment of priority setting for RI funding there is a high risk of favouring popular short term interests over higher value but longer-term or not so popular investments. Long-term considerations are essential for RI sustainable funding and valuable prioritisation.

- **Differences in administrative requirements, decision making processes and their timelines and regulatory frameworks among Member States**
  The difficulties of overcoming the obstacles associated with differences of bureaucratic, regulatory and decision making processes among Member States frequently impair decisions on RI. It is necessary to align strategies and processes by regular joint work with national and regional organizations.

- **Funding constraints**
  Sustainable funding of construction and operation of RI is expensive. Innovative and flexible business models need to be invented to overcome funding bottlenecks. New types of services, such as those that can be developed on the basis of data collections can demonstrate to stakeholders, the public and politicians the value of RI for society and the economy, including regional development. Sustainability and evolution of RI are subjected to cycles of funding needs throughout the RI overall life; funders must know about these cycles and be ready to reserve the necessary funds. The risks associated with consistent political commitment and continuity of funding are maximised in times of crisis.
  Moreover, there is a lack of funding for generic technologies and research ecosystem as the current model of funding is oriented to research projects or RI, but maintaining long-term competitiveness on technologies and people for next generations of RI is being neglected.

- **Ineffective governance**
  Errors in the setting up of the governance structure can be fatal for RI effectiveness and sustainability, and are difficult to correct after the RI is created. Defining efficient and effective governance structures in a cross-border, cross-disciplinary, multi-stakeholder setting is difficult. Actively involving leading researchers and putting them in the driving seat of infrastructure building is not easy to achieve, but is much needed to ensure goal orientation and to build trust. Even when governance structure is sound, difficulties in ensuring good governance practices along the RI life frequently result in problems to RI effectiveness and sustainability.

- **Techno-legal barriers in research data RI**
  The use of research data and data infrastructures poses new challenges of ethics, privacy, security, copyright and IPR, and requires appropriate legal, social and technical frameworks to overcome these challenges. Europe needs a data framework that lets citizens feel empowered and safe, and where responsible innovation can enable an enhanced society. For research data RI these are fundamental, underlying requirements that need to be addressed to build trust in data-driven research and innovation.

Question 5: Which gaps (science and technology, policy) and potential game changers need to be taken into account?

- **Consultation processes to identify demand for RI and build trust**
  The demand for RI should be identified by widely participated open transparent processes involving consultations of a critical mass of key researchers in different countries, addressing specific relevant issues of goals, characteristics, design, governance, management, services and other aspects.
- **Political and societal vicissitudes**
  Political and societal dynamics which vary widely between the Member States are most important game changers. New governments set different priorities which may change the funding base within a short time. This is a danger for the sustainability of infrastructure construction and maintenance. Bad perspectives have the effect that the best researchers or experts will leave early in the piece with an immediate impact on the quality of services. Only a situation of long-term commitments on the one hand and mechanisms that allow a flexible reaction (business model, payment models, partnerships, etc.) can overcome these uncertainties.
- **Harmonised legislation**
  Having obtained agreement on the ERIC instrument was an excellent step forward towards establishing European cross-border infrastructures as sustainable legal constructions. Still, the Member States need to collaborate more so as to harmonise legislation at various levels including the rules of how to work with sensitive data. Differences in legislation often create huge barriers to developing efficient and cost-effective solutions and in creating the open market place for data and services which are needed to foster a new economy.
- **Global harmonisation of core aspects of RI**
  Global harmonization in the infrastructures area will lead for example to easy data sharing and re-use at a global level, since research is global. Initiatives and organizations such as OECD, CODATA, RDA and WDS at a global level are important to define widely agreed principles that guide infrastructure work. Bottom-up driven initiatives are also required towards overcoming the many practical hurdles that hamper establishing the open market place.
- **Universal principles for direct access to research data**
  A current gap is the absence of abstract universal principles for direct access to and use of research data, based on PID and data address registry, and the mapping of this addressing scheme on the network infrastructure, a consistent/standardised PID system and data citation formats, and a European or international authentication system for RI users.

Question 6: In which areas is the strongest potential to leverage the EU knowledge base for innovation and, in particular, ensure the participation of industry and SMEs?

- **RI construction and new high-end equipment design and manufacturing pushing the limits of technology**
  RI frequently pose technological challenges that require continuously expanding the limits of technology in both construction and component design and manufacturing. These challenges can be great innovation opportunities for industry (including SMEs) to build up expertise and ensure sustainable and competitive advantage in leading-edge technologies which frequently have spill-overs to related products and services to be commercialized in a wider market.

- **Education and training opportunities at RI**
  RI provide specific opportunities for education and training that may have a large impact in innovation capacity and develop a highly skilful workforce for industry and training in specific techniques that may be of special interest to industry (including SME) innovation.

- **Regional impact of RI**
  Setting up Regional Partner Facilities to overcome the uneven distribution of pan-European RI and also strengthening pan-European RI through fully exploiting scientific talent and capacity existing across European regions provide opportunities to strengthen the innovation base of regions, and will have the highest value-added in regions were the innovation potential can increase the most. In particular, "location-blind" policies as in Horizon 2020 and "location-based" policies for cohesion of European regions should be linked further to enhance socio-economic benefits in different Member States regions without compromising the RI mission of scientific excellence.

- **Data providing RI**
  Data providing RI provide special opportunities for innovation:
  (1) Open access to research data that allows companies (including SMEs) to build added-value and services to be commercialized;
  (2) Partnerships of RI with companies (including SMEs) to support and foster the development of new data based services;
  (3) Development of Data Management solutions of wide use (including by SMEs) that can be offered at the marketplace;
  (4) Development of new data analytics with commercial value (including by SMEs);
  (5) Emergence of new business activities and markets (including for SMEs) based on the analytics of Big Data, including in areas of possible high societal impact such as healthcare for the aged, energy supply, climate change and several sorts of "smart" applications in transportation, cities, energy distribution, consumption and production, developing new business opportunities based on completely new pattern detection driven fields.

- **Making use of "Big Data"**
  Meaningful use of "Big Data" requires cooperation with industry both as a provider (e.g. computer, telecommunication) and as part of the user/research community. PPP could be one option for business models in such cases.

- **Testing and development of innovations in RI facilities**
  The access of industry (in particular SMEs) to RI facilities for developing products, processes, services that require special equipment available at RI can also be an important driver of innovation based on RI.

- **Open innovation initiatives at RI facilities**
  RI may provide appropriate settings for some initiatives adopting the open innovation model, involving Big Industry and other actors of the science and innovation system.

So, there are numerous ways in which RI can leverage innovation capacity, but most of them are not related to the participation of industry in RI project partnerships. Accordingly, the participation of industry in RI project partnerships or management should not be a criterion to assess innovation impact in ex-ante evaluations of RI nor to judge on it in ex-post evaluations. Instead, the criteria should be directly related to the several ways above identified in which RI can leverage innovation capacity.

Question 7: What is the best balance between bottom-up activities and support to roadmaps or to the societal challenges? What is the best way to address cross-cutting activities, such as social sciences and humanities, responsible research and innovation including gender aspects, and climate and sustainable development? Which types of interdisciplinary activities should be supported?

- **The best balance**
  Due to the natural competition for achieving the best results, researchers would like to have maximal control over the resources needed. This bottom-up control will lead in most cases to the innovative solutions society wants to see. This may be in contradiction with the need to build infrastructures that serve a wider community and which thus need to be put under shared control including top-down influences to achieve the broadly usable solutions necessary. Top-down decisions in principle tend to hamper fruitful innovation; pure bottom-up processes can lead to singular and costly re-inventing the wheel solutions, creating huge maintenance cost challenges. A close interaction between bottom-up and top-down support is desirable to achieve the purpose or objectives (e.g. roadmap or societal benefits) and avoid gaps in the long-term development of these activities. Solid and balanced governance structures need to be established. Excellent personnel from different stakeholder groups need to be engaged continually at all levels in order to achieve the essential balance. A rigorous regular evaluation of all aspects needs to be carried out by undisputable experts, based on transparent criteria and metrics. Opt-out options and contingency plans to continuously adjust the balance, given changed circumstances, should be incorporated.

  Bottom-up calls for either new infrastructures or for integrating infrastructure capacities on certain topical research issues should take into consideration the potentialities of existing infrastructures in order to avoid an ever growing portfolio of new marginally different infrastructures with substantial overlaps with existing ones, taking into account the landscape analyses done by ESFRI.

A rational approach should be ensured to identify thematic topical areas for Integrating Activities (semi-bottom-up with identified topical areas), either addressing types of infrastructures or thematically oriented infrastructures closely coordinated with EU thematic research support.

- **Bottom-up cross-cutting activities in data RI**
  RDA is identified as a good bottom-up process for fostering cross-cutting activities regarding research data infrastructures which has the potential of being highly effective with minimum cost (and therefore deserves support) for ensuring the engagement of researchers, building consensus based schemes and open standards for research data infrastructures, and developing specifications and standards for particular types of data that require the contribution of specialists.

- **Interdisciplinary activities**
  Activities involving a diversity of agents, regarding scientific areas, thematic interests, professional roles, public or private activities, goals and cultures, frequently increase the potential for knowledge creation and innovation in complex science subjects. It is commendable to foster interdisciplinary activities between physical sciences, environment, Life Sciences, Social Sciences and Humanities to address global issues and grand societal challenges, but room should also be left for focused high potential research efforts.

- **Gender issues**
  Top down guidance in EU policy, including funding criteria, is necessary to encourage scientists to explore fully any gender aspects in RI data analysis in order to reap the benefits for the EU, regional and national governments, industry and third sector organisations and in guiding human resource development strategies.

  The design and location of RI should consider gender equality and assessments should be made on whether the gender dimensions of RI projects are adequately addressed.

Conclusions

In conclusion, the group makes the following recommendations for concrete actions to address the challenges, gaps, bottlenecks and key priorities identified in this paper and in the appended four position papers.

The recommendations are listed here (not as a ranked list) and will be developed in further detail in preparation for the development of the Work Programme later this year.

The ***action recommendations*** of the Advisory Group are:

- **Access:** develop mechanisms to enable or enforce trans-national access for established and emerging RI, and support access for remote communities to RI and services.
- **Funding:** enable and leverage a blend of multiple funding sources (H2020, regional, national, cohesion, PPP, other) for diversified long-term funding for RI.
- **Engagement:** Develop clustering between RI, and KETs, Smart Specialisation, and JPIs and Marie Curie initiatives for best synergies, and promote awareness of the ERA framework and exploit best user/stakeholder engagement and application practices in RI utilisation.
- **Training:** support training for RI staff and users by:
    - Developing training programmes for RI staff, including legal training for RI managers, data management training for engineers and data scientists, ethics training and gender awareness, and support mobility for exchange of skills;
    - Developing training and user support for researchers to use international RI;
    - Recognising and fostering the new profession of 'data scientist', including education programmes in schools, undergraduate and postgraduate programmes.
- **Exploit existing strengths:**
    - Re-use existing (in preference to re-inventing) best RI and data management methodologies, tools and technologies, and support and participate in international organisations such as RDA (Research Data Alliance) for best practice in data sharing;
    - Develop an Open Market Place for Services, via a user-driven registry or distributed catalogue of services that are offered by RI service providers;
    - Support the further development of methods for discovering scientific knowledge, coming from machine learning, data mining, and intelligent data analysis, covering the entire life-cycle of big data analytics and applied in multidisciplinary domains to face the grand scientific and societal challenges;
    - Ensure that the investment in human resources is guided by principles of equality of opportunity and outcome;
    - Support Data Fabric projects to enable reproducible science, where automatic processes turn legacy data into re-usable data.
- **Trust and Accountability:**
    - Strengthen accountability and both ex-ante and ex-post evaluation practices;
    - Increase emphasis on interoperability and trustability of RI facilities and services;

- Support the definition of a clear and effective ***techno-legal*** framework that allows responsible scientific research to take place whilst taking account of the interest of the data subjects.

- At all levels, engage with strategic planning and road-mapping to support RI **contributing to innovation and to grand societal challenges**, and promote their visibility in this role.
- Support RI of **horizontal nature**, that underpin multiple disciplines and both basic and applied research.

Group membership

Name	Institution/ Position	Country	Contact
Costas Fotakis	Foundation for Research and Technology-Hellas (FORTH), President **Chair**	Greece	fotakis@iesl.forth.gr
Luis Magalhaes	Instituto Superior Tecnico, Technical University of Lisbon, Professor **Vice-Chair**	Portugal	lmagal@math.ist.utl.pt
Sandra Collins	Digital Repository of Ireland, Director **Rapporteur**	Ireland	s.collins@ria.ie
Anton Anton	Technical University of Civil Engineering Timisoara, Professor	Romania	anton@utcb.ro
Lajos Balint	National Information Infrastructure Development Institute (NIIFI), Director of International Relations	Hungary	lajos.balint@niif.hu
Corinne Borel	CEA Physical Sciences division, Director of External Relations and Partnerships	France	corinne.borel@cea.fr
Sabine Brünger-Weilandt	FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, President and CEO	Germany	sabine.bruenger-weilandt@fiz-karlsruhe.de
Janusz Bujnicki	International Institute of Molecular and Cell Biology, Warsaw, Professor	Poland	iamb@genesilico.pl
Fosca Giannotti	ISTI CNR, Director of Research	Italy	Fosca.giannotti@isti.cnr.it
Lucie Guibault	Institute for Information Law, University of Amsterdam, Associate Professor	The Netherlands	L.Guibault@uva.nl
Geraldine Healy	Centre for Research in Equality and Diversity, Queen Mary University London, Director	United Kingdom	g.m.healy@qmul.ac.uk
Frederic Hemmer	CERN, IT Department Head	Switzerland	Frederic.Hemmer@cern.ch

Rosie Hicks	Australian National Fabrication Facility Ltd, Chief Executive Officer	Australia	rosie.hicks@anff.org.au
Sverker Holmgren	Nordic eScience Globalisation Initiative (NeGI), Program Director	Norway	sverker.holmgren@nordforsk.org
Blanca Miranda Serrano	Andalousian Public Health System Biobank, General Coordinator	Spain	blanca.miranda@juntadeandalucia.es
Maria Teresa Ponce de Leao	National Laboratory of Energy and Geology (LNEG), President	Portugal	mleao@fe.up.pt
Stefano Ragazzi	National Laboratory of Gran Sasso of INFN, Director	Italy	stefano.ragazzi@lngs.infn.it
Orly Reiner	Weizmann Institute Department of Molecular Genetics, Professor	Israel	orly.reiner@weizmann.ac.il
Sanna Sorvari	Climate Research Unit, Finnish Meteorological Institute, Research Manager	Finland	sanna.sorvari@fmi.fi
Dominic Tildesley	European Centre of Atomic and Molecular Computations (CECAM), Director	Switzerland	dominic.tildesley@epfl.ch
Dany Vandromme	National Institute of Applied Science (INSA), Professor	France	dvandrom@me.com
Beatrix Vierkorn-Rudolph	Federal Ministry of Education and Research, Director	Germany	Beatrix.vierkorn-rudolph@bmbf.bund.de
Peter Wittenburg	Max Planck Institute for Psycholinguistics, Senior Advisor	The Netherlands	peter.wittenburg@mpi.nl
Colin Wright	SANReN (South Africa National Research and Education Network), Chief Strategist & Manager	South Africa	colin.wright@wits.ac.za

Annex 1: Innovation and Cooperation

Table of Contents

1. Executive Summary

Research Infrastructures (RI) are the backbone of the research and innovation system. They provide the necessary resources (e. g. equipment, material, data etc.) to perform research and development. They could be single-sited or distributed. At large facilities like CERN they provide access to top researchers from all over the world at one place but this could be virtual access to data bases. Virtual access could also be provided to collections of tissues, biological samples or ancient scripts. Virtual access could also be provided in many cases by Distributed RI.

RI are an intimate part of the research system and follow therefore the same processes or actions which generate innovation as for research projects. But there is a difference between “normal” research projects and RI, namely, the latter provide access to data, archives, collections, beamlines or telescopes e.g. meaning they provide services for the whole community.

The position paper on “Innovation and Cooperation” focus on different types of cooperation, which may lead to innovation: industrial cooperation or cross-disciplinary cooperation and regional cooperation which is dealt with under the heading “socio-economic aspects”. The distribution of excellent RI in Europe is still imbalanced between the

different regions. Since RI are often a stimulator of regional "High-Tech Clusters", it is important that RI are more evenly distributed. Therefore measures are necessary to stimulate the cooperation between large RI and smaller regional RI for expanding the socio-economic impact of RI and exploiting scientific capacity across Europe. Increasing potential synergies with cohesion policies and support of RI through structural funds is crucial in this respect.

Task from the European Commission

The European Commission has set up various Expert Advisory Groups to accompany the successive batches of the funding under Horizon 2020 in the time frame of 2014 – 2020.

RI play an ever increasing role in the research funded by the European Commission since they are

- means to better organize the research all over Europe,
- important to make research more cost-efficient and more competitive on an international scale,
- an important part of the European Research Area (ERA),
- indispensable for the economic growth and welfare in Europe.

For the 2014-2020 period, the foreseen funding of EU will be around 2 500 Millions €, out of the 24 441 millions € allocated to H2020 Priority 1: Excellent Science.

To advise the Commission in implementing this huge amount of funding, a high level expert advisory group has been formed. The group will provide the Commission with recommendations for various relevant topics for RI. During the first plenary meeting on 23rd October 2013 and in discussions afterwards, four topics were selected deserving immediate contribution from dedicated subgroups:

1. Innovation and cooperation
2. Sustainability
3. Data
4. Evaluation

The four subgroups are expected to provide each a clear position paper with possible recommendations, so that these recommendations could be taken into account by the EC for the preparation of the next periods of the work programme of H2020 (i.e. 2016-2017 and 2018-2020)

2. Introduction

RI play an ever increasing role in research and innovation. They are a key factor in addressing Global Challenges such as securing energy supplies, global warming, tightening supplies of water and food and securing quality of life for an ageing population.

Progress in these areas depends on excellent research and innovation capabilities which require access to the best available RI. RI bring together researchers, funding agencies,

politicians and industry to tackle important, cross-disciplinary scientific and technical issues for our continued prosperity and quality of life.

In many cases RI are nowadays too demanding and too expensive to be built and operated by a single Member State or Associated Country alone. Therefore efforts have been brought together especially in the European Strategy Forum on Research Infrastructures (ESFRI) to tackle this issue. The work of ESFRI complements the already existing RI on a national or regional level by setting up more complex RI, single sited ones as well as distributed RI.

ESFRI has started to establish closer links with industry to contribute to the economic and social impact of the ERA and to promote knowledge and innovation partnerships. But innovation means not necessarily industrial cooperation. Innovation accrues often when research connects different disciplines. RI like the European Spallation Source (ESS Neutron), CERN or ESRF (European Synchrotron Radiation Facility) bring together researchers from a lot of different disciplines from all over the world, enabling therefore intrinsically innovation.

The development and construction of components and materials for RI act in many cases as a driver for industrial innovation. It is therefore important that industry is involved in the design and construction of new RI and is informed about upcoming procurements at an early stage. Industrial liaison offices have proven to be an adequate instrument for this purpose.

Many RI, like the ones in materials area and the life sciences have important industrial users. Tightening the links with industry will support technology transfer and ensure the fast delivery of scientific results to industry. Finally socio-economic aspects of RI at a regional level should be taken into consideration.

3. Innovation through Cooperation

3-1. Innovation through cooperation with industry

Industrial cooperation with RI can be characterized in different ways: 1) industry as provider of new designs and components 2) industry as user of RI 3) or both. The first category where industrial companies are the providers of state of the art components is the standard procurement situation. The second category describes R&D activities where industry and RI work together (or collaborate) in a tight manner in cooperative or contract research projects: researchers together with industrials engineers or scientists develop advanced technologies – often a win-win situation. The outcomes are highly demanding technological developments (or more simple “innovation”).

In the third category the RI provide to industrial companies access either as a test facility for innovative developments and products or for early stage basic research to conduct cooperative projects but also for training and exchange programs. A lot of examples for the first case could be found in the life sciences where bio-banks, screening platforms or clinical study centers directly serve pharmaceutical companies to test new products whereas “classical basic science” RI do not attract so far much industrial users.

Gateways for cooperation of RI with industry are numerous and diverse. Industry must be considered in a wider sense, not just as hardware provider but also as provider of software and services, especially when talking about e-infrastructures. In the following, the term "industry" refers to the private sector in general. This especially applies in relation to innovation, as many innovations of today's world are built on ideas, concepts, services and new business models rather than advanced objects and products.

3-1-1. Design and construction

The first two categories are quite conventional and well defined: Industrial partners collaborate with RI to provide either state of the art components (first category) or design and build new components for RI (second category). There are already a large number of examples, like supra-conducting elements for accelerators or tokamaks, adaptive optics for astronomy, sequencing automats for genomic research, high performance computing elements, etc. For many years, these were perceived as the main gateways between private and public sectors for RI, with obvious impact on technology transfer and skill management. Public sector is also requested to train young people (researchers, engineers, etc.) to make them suitable for industry needs. Such a win-win virtuous circle is also benefiting from adapted procedures for public procurement (pre-competitive procurement), aiming at supporting PPPs (Public Private Partnerships) for advanced technologies, prior to the market offers.

All these schemes are well known and widely documented. However, they do not match the way society adapts to the digital world. A major illustration is for instance the SSH (Social Sciences and Humanities) sector.

While RI framework is more and more inclusive for the SSH field, there is no real involvement of the private sector to the design, build and operation of RI. This is due to the fact that the private sector is mostly interested in accessing the content of RI made essentially of public data, rather than contributing to the set-up phase.

Life science RI are somewhere in between the conventional hard science tools (physics, astronomy, condensed matter, e.g.) and new soft sciences (genomics, proteomics, metabolomics, clinical trials, translational medicine, epidemiology, etc.). There is close collaboration with the private sector in some domains (imaging, sequencing, structural biology, e.g.) to design and build modern instruments, while in many other aspects, the private sector is primarily interested in using data or well documented material (biobanks for instance).

3-1-2. Access and usage

The third category for cooperation between (publically funded) RI and private sector (industry and services) is for the access and use of RI. ESFRI is involved in the drafting of an "European charter of access to Research Infrastructures" which will be discussed on European level with all relevant stakeholders in the next months. So far the text concerning industrial use states: "Promoting the **cooperation with industrial users** and allowing them access to RI can be enabled through both quality-based and market-based access regimes.

Bridging the commercial and Research Infrastructure worlds by dedicated initiatives combined with expert support will help to close the gap between scientific excellence and knowledge and technology transfer with industry, the main drivers for innovation."

Some RI are more adapted and suited for this, such as analytical facilities (synchrotrons, neutron sources, FEL) and also high performance or distributed computing facilities (HPC, Grids, Clouds) whereas RI in biological and medical sciences still have to develop appropriate business models and access rules (including IPR, data privacy e.g.).

Usually, the costs of analytical facilities are much too high for an industry to construct and operate its own facility. It is preferable to get access to RI, for a share of them or on a pay-per-use mode. Buying a share of a facility (i.e. at the construction and operation levels) is not current, with the exception may be for aeronautics, where wind tunnels are used by industry and public research. For synchrotron or neutron facilities, the core facility is always funded from public sources, while it may happen that instrumented beamlines may be owned by a private partner. However, this model is rather exceptional. In most cases, industry willing to use RI, prefers to use the pay-per-use model. But at the end, this has only a limited impact on the RI business model. It prevents industry to have a long term strategy in terms of usage of the RI and leaves the full responsibility of the operation and upgrade of the respective RI to the public side. Furthermore, when RI are essentially publically funded and operated, competition rules prevent to allocate more than ~10% of the available resources to private users. Another limitation observed for the pay-per-use model is that industry is reluctant to pay anything. Reasons are usually too high costs and research confidentiality concerns. Facilities like synchrotron or neutron sources or HPC resources have even difficulties to reach the 10% level of private customer funding. Remedy to this relies on the openness of data. If a private user wants to use public facilities, he may benefit the same conditions as public research organisations, provided the results are open to the public. If the results are not made public, then the private user may still access and use the facility but then he has to cover the costs of the usage.

3-1-3. Summary of existing mechanisms

At the design/construction level, the most appropriate mechanism in place today is the pre-competitive public procurement. It fosters collaboration between private and public sectors, and may eventually transform into a PPP for construction and operation of RI, but significant examples still need to be identified.

At the usage level, private investment into a share of a RI is a way that should be explored more (for instance with analytical facilities or HPC resources). There have been examples in the past (HLRS in Stuttgart), but the model did not spread out significantly so far. It has to be analysed further in which areas this may work or not. RI which serve more basic research will need other means for attracting industrial users (e.g. ATTRACT) than RI which serve more applied research (as in the biomedical or materials analytical area). There may be technological areas where new RI meet more the demand of industrial research (e.g. "Australian National Fabrication Facility"). This example is worthwhile to be studied in more detail.

The status of research data (open or not) is considered as a compromise for using publically-funded RI. It allows the private sector to be treated just as public laboratories when the research output is made available, while the use for protected research is still possible, but with an appropriate cost recovery mechanism, to prevent any unfair competition and breaching of the market rules.

For e-infrastructures, the above models apply as well. However, there is a strong pressure from the private sector to provide resources, which raise strong questions about the business model of public services (like clouds), while endangering the long term strategy about scientific data sustainability (maintenance, curation, etc.).

3-2. Innovation through cross-disciplinary cooperation

As mentioned earlier, RI like ESS Neutron, CERN or ESRF could bring together researchers physically at one location to cooperate with each other which is an important means to foster joint developments but it might be more important in the next years to bring together researchers from different disciplines. One good example could be to integrate the digital humanities in other scientific disciplines, something which could give an unforeseen push in many developments and will help to meet the Grand Challenges.
The ESFRI project SHARE-ERIC is one of these examples which are going to cross borders by bringing together their social science data archives with biological samples which will be collected in biomedical RI (e. g. Biobanks). Combining these data for statistical analysis is challenging for several reasons as health-related data are strictly protected by privacy regulations. But using e-Infrastructures and tools for data encoding it is possible to conduct multivariate statistical analysis at various levels of higher aggregation. In this specific example you will be able to combine life histories concerning socio-economic environment, behavioural, environmental and occupational lifestyle factors of several thousands of European citizens in more than 9 countries to biomedical markers for measuring health outcome in later life.

4. Means to increase the cooperation with industry

To increase cooperation between RI and industry a change of culture is needed within RI. Researcher and industry staff still speak different languages and have different expectations. Cooperation with industry is not (yet) in the focus of RI researchers.

Beyond the different initiatives coming from RI (with Industrial Liaison Officers for example), to foster and accelerate collaboration with industry, there is a need of increasing critical mass of opportunities, to attract industry both as supplier and as RI user. To promote this further, outreach programmes are necessary and “industry days” should be organised. To promote the ideas of cooperation with industry, the socio-economic aspects of such cooperation should be studied in more detail. Another issue is the development and the maintenance of competences of technicians and scientists to support and co-develop with industries new products for new markets.

Usually, the decision for constructing a research infrastructure (RI) is science-driven, i.e. a scientific community defines the needs for a RI and European labs working in the domain get together to design and build this RI, bringing in their specific competences. In the setting up of this RI, innovative techniques to reach the scientific goals are developed, in many cases together with industrial partners. Due to sometimes long construction times for large-scale RI, it is difficult to maintain during the construction time the acquired competences, with the result that sustainability cannot not be (fully) secured.

It is recommendable that the cooperation between different scientific communities and disciplines is significantly increased to be able to use the most advanced technologies and the best available know-how which might be developed in adjacent disciplines. We all know that in many cases real innovation will be generated at the borders of disciplines. And, in such a way, unnecessary duplication of R&D efforts could be avoided. Often, the potential for applications to the society is not fully exploited. Yet, in the ERIDWatch survey, 62% of companies (large groups and SMEs) declared that other markets segments have benefited from technologies first developed for RI. Also, industry is often reluctant to invest in an intermittent market or is not aware of the opportunities. In addition, while a RI is being designed and then built, the focus is put on short-term priorities and it is generally very difficult to find financial resources to continue investing in new and highly innovative technologies within the involved scientific community.

Clustering of RI belonging to a broad scientific domain can address some of the recommendations addressed above: it can help developing fruitful collaboration between RI, sharing R&D of common interest, enlarging the market for industry.

Beyond the necessity of a closer cooperation between industry and research organisations to disseminate the already available know-how there is a need to develop new models for financing large scale RI for the development of ground-breaking technologies and innovative integration schemes, which is mandatory if Europe wants to stay at the forefront of world research in the future.

Such a new model of incorporation of industry and research organisations should be rather technology driven and also product driven. One possibility could be clusters of technology centres, within a smart specialization approach, which would help to develop synergies and complementarities avoiding duplication of work, and will help optimising the sharing of European resources. The recognition by the European Union of the importance of such clusters or platforms will help securing national funding, and could benefit from regional funds, assuring a stronger socio-economic impact.

Such clusters or platforms (some already exist) could foster innovation and transfer to SMEs and industry and enable a faster transfer of innovation to the market. By creating both a large and viable market and a critical mass of industrial partners, it would ensure that EU industry would be ready to respond to invitations to tender in Europe but also in other parts of the world, in particular in Asia, where a lot of large-scale RI are likely to be built in the future.

The established links with universities would ensure the dissemination of the acquired knowledge and the training of top-level engineers, who are desperately needed in the European laboratories as well as in industry.

In many regions, clustering of industry especially SMEs in research and development areas play an important role as incubators for knowledge and technology transfer. An example of such a cluster is the so-called "Forschungscampus" (research campus)[1] in Germany where a close cooperation between industry and research organisations/universities is created. As an example "M^{2}OLIE – Mannheim Molecular Intervention Environment" should be mentioned, where at the University Medical Center of Mannheim, various medical engineering and biotechnology companies, a Fraunhofer Institute and the Universities Mannheim and Heidelberg work together to develop new models for diagnosis and therapy of cancer at the molecular level.

5. RI and Socio- economic aspects

The prime mission of European RI is to serve scientific excellence. In doing so, they may also have a pivotal role in relation to the other two priorities of HORIZON 2020, namely, industrial leadership and societal challenges. In this context, the future RI programme should be more open and integrated, focusing on three major issues:

1) Promotion of industrial involvement and innovation at both the suppliers and users ends of RI.
2) Human capacity building by exploiting the excellent training and educational opportunities offered by RI.
3) Maximizing the socio-economic benefits RI may have, thus facilitating regional and national cohesion policies in Europe.

The last issue (3) is raising a major challenge:

How can socio-economic benefits in different Member States be enhanced, without compromising the mission of scientific excellence?

Knowledge-based economy occurs according to logic of concentration. The socio-economic benefits of the development and operation of RI have been recognized in several EU documents and policy papers. These benefits may be tangible, having direct and measurable effects, such as procurement, the creation of new jobs and spin out companies, as well as public income generation from taxation. It is worth noting that for new RI of the ESFRI Roadmap, the RIFI analysis[2]) has shown that the tangible benefits are to a very large extent localized (e.g. construction work and job creation). There are also intangible benefits, such as the strengthening of scientific and entrepreneurial culture, which may accelerate regional competitiveness and facilitate national cohesion.

[1] www.vdivde-it.de/forschungscampus
[2] (Research Infrastructures: Foresight and Impact. http://rifi.gateway.bg)

Furthermore, there is evidence that the presence of RI can become a pole of attraction for talented young scientists as well as prominent researchers in a region and contribute to the reversal of "brain-drain" and enhancement of "brain-circulation". Overall, RI may mobilize significant cultural, social and educational resources in local communities, including outreach activities.

Sometimes it might be also necessary to provide a small funding for industry for the collaboration with RI to reduce the risk for the industrial partner. If the industrial partner wants to use the result of the research later on, he has to pay in the end the research effort of the RI (example which works at STFC, UK)

Example of the Aquitaine region [3]

> The setting up the Laser Megajoule project at Le Barp has developed a research cluster in optics and photonics which is a magnet for collaboration and R&D projects, bringing together research institutions and industry. The investment of 123 Millions € over 15 years by the Aquitaine Regional Council has created more than 200 new jobs and 2 start-up companies have been set up.
>
> The success of the Aquitaine region was based on an initial technology push from RI and ensuring a long-term critical mass in terms of funding and expertise and a political long lasting support underpinned by a clear investment strategy.

Experience has shown that scientific research, even when it is not directly translated into services and products, influences indirectly and over long term the economy and social/cultural evolution of a country/region by enriching its human potential. The scope for developing top-level science relies on the attraction and education of well-trained scientists and skilled personnel. Indications are that industrial research managers view this as the principal contribution to industry rather than the specific knowledge generated by research.

For the above reasons, RI are well placed to play an enabling role for regional development in the context of Research and Innovation Smart Specialization Strategies (RIS3). They may either be hubs of distributed RI of the ESFRI Roadmap, or operate as regional RI associated with a major RI located elsewhere. In other cases, they may constitute "open" independent entities of territorial interest and impact, stimulating interactions across neighbouring regions and countries. These RI can be strong components of RIS3 provided that their role is well understood and promoted. To this effect, the form of support they may provide to local business communities and their contribution to the overall regional and national economic capabilities have to be clear within the RIS3 strategy.

To accelerate the impact of both tangible and intangible socio-economic benefits by embedding RI in RIS3, the potential synergies between the funding mechanisms of Horizon 2020 and Structural Funds must be further explored and the rules applied must become clearer and simpler. A serious reduction in bureaucracy on EU level but also on MS/AC level

[3] ERF workshop "The Socio-Economic Relevance of Research Infrastructures", Hamburg 31 May/1 June 2012

is mandatory in this respect! It is recommended that further measures and initiatives towards this direction are taken.

The formation of regional RI hubs, which provide good science, technology, talent and entrepreneurial challenges are important for having strong regional impact and simultaneously a positive contribution to pan-European RI.

Here could be an opportunity to set up either “Regional Partner Facilities” as defined by ESFRI and recognized by the Competitiveness Council in 2009 but also to stimulate set up national nodes of distributed pan-European RI.

6. Recommendations

1. Carry out an analysis of potential technical areas of interest for industrial research at RI.
2. Promote existing mechanisms for industrial cooperation in the design and construction of RI, such as the precompetitive procurement; e.g. early involvement of industry in the preparation of calls for tender
3. Promote a culture for cooperation with industry in RI. This also includes some changes in the tendering procedures so that companies are not suspended from future calls after they have been involved in the design of the respective components.
4. Increase usage of e-infrastructures to facilitate the access of industrial researchers to RI (harmonisation in terms of quality, standards, data protection needed)
5. Enhance clustering actions of RI to promote cross-disciplinarity effects on innovation and private sector collaboration
6. Establishing strong links between RI and technology platforms and encouraging industrial participation in RI.
7. Facilitate and enhance further the use of structural funds
8. Develop cross-disciplinary cooperation (e.g. involvement of digital humanities in other scientific areas)
9. Enhance synergies with training and educational programs for engineers, young scientists and technicians to work in RI.
10. Promote the awareness of socio-economic aspects of RI

7. Conclusions

RI have to serve very different R&D communities either concerning their scientific disciplines or alongside the innovation chain from basic to applied science.

Innovation could be promoted not only by direct interactions with industry but also by cross-disciplinary cooperation between different areas of research.

One of the main problems concerning the usage of RI by industrial companies rest with the different interests of researchers and companies (confidentiality, intellectual property, publications, patents e.g.). Therefore a change of culture is necessary within the RI as well as in industry.

RI may be effective tools for enhancing scientific and technological excellence in Europe, while simultaneously countering societal, cultural and economic challenges at regional level, in a manner that effectively promotes the goals of European cohesion and integration.

Subgroup membership

Name	Institution/ Position	Country	Contact
Beatrix Vierkorn-Rudolph	Federal Ministry of Education and Research, Director **Subgroup Chair**	Germany	Beatrix.vierkorn-rudolph@bmbf.bund.de
Anton Anton	Technical University of Civil Engineering Timisoara, Professor	Romania	anton@utcb.ro
Lajos Balint	National Information Infrastructure Development Institute (NIIFI), Director of International Relations	Hungary	lajos.balint@niif.hu
Corinne Borel	CEA Physical Sciences division, Director of External Relations and Partnerships	France	corinne.borel@cea.fr
Costas Fotakis	Foundation for Research and Technology-Hellas (FORTH), President	Greece	fotakis@iesl.forth.gr
Rosie Hicks	Australian National Fabrication Facility Ltd, Chief Executive Officer	Australia	rosie.hicks@anff.org.au
Blanca Miranda Serrano	Andalousian Public Health System Biobank, General Coordinator	Spain	blanca.miranda@juntadeandalucia.es
Maria Teresa Ponce de Leao	National Laboratory of Energy and Geology (LNEG), President	Portugal	mleao@fe.up.pt
Orly Reiner	Weizmann Institute Department of Molecular Genetics, Professor	Israel	orly.reiner@weizmann.ac.il
Dominic Tildesley	European Centre of Atomic and Molecular Computations (CECAM), Director	Switzerland	dominic.tildesley@epfl.ch
Dany Vandromme	National Institute of Applied Science (INSA), Professor	France	dvandrom@me.com

Annex 2: Sustainability

Table of Contents

1. Introduction

Modern Infrastructures

Research Infrastructures (RI)[1] "are a key instrument in bringing together a wide diversity of stakeholders to look for solutions to many of the problems facing society today. RI offer unique research services to users from different countries, attract young people to science, and help to shape scientific communities. New knowledge and, by implication innovation, will only emerge from RI which are of high-quality and accessible. Moreover, RI help to create a new research environment in which all researchers - whether working in the context of their home institutions or in national or

[1] In this report the term Research Infrastructures often includes the aging term e-Infrastructures realizing that in fact most RIs also need to tackle cross-disciplinary issues. In some cases e-Infrastructures are mentioned explicitly. e-Infrastructures are per definition distributed and virtual.

multinational scientific initiatives - have shared access to unique or distributed scientific facilities (including data, instruments, computing and communications), regardless of their type and location in the world. RI are therefore at the centre of the knowledge triangle of research, education and innovation, producing knowledge through research, diffusing it through education, and applying it through innovation"[2]. It was recently appreciated that RI also need to take on a fundamental role in preserving our scientific information and knowledge.

Of course researchers would like to have all facilities they need to follow their research interests in their own environment and under their own control, but given the scientific challenges and the need to include many different resources it is not realistic to expect to arrive at new results in this way. However, modern infrastructures, Google for example, can be seen as such a resource under virtually own control. Huge amounts of digitized material are available and can be analyzed by specific data mining services. However these offers based mainly on textual web-content do not include the huge amount of data objects created and consumed within science daily in an independent and heterogeneous manner.

Types of Infrastructures

The term 'research infrastructures' refers to facilities, resources and related services used by the scientific community at large to conduct top-level research in their respective fields or across fields. Basically we can distinguish 3 types of RI:

- '**single-sited**' (a single resource at a single location),
- '**distributed**' (a network of distributed resources), or
- '**virtual**' (the service is provided electronically).

These types are in general organized by discipline or domain. However, some of the problems[3] associated with distributed and virtual RI are common to the others and require for example, so-called e-Infrastructures that offer services across all or at least some of the communities in the areas of networking, computing or data management. Irrespective of the type of RI, high costs are involved in their construction, operation and further development and expensive experts are required in all phases.

Road-mapping, building and operating research infrastructure requires the engagement of many stakeholders from different countries in the different phases as ESFRI has shown. Overcoming the many sociological, organizational and cultural hurdles involves building mutual understanding and trust, both of which imply long lead times and considerable costs. In addition, researchers who wish to remain globally competitive and achieve the best results will only invest their time, if they can rely on smoothly functioning and robust services which are available for longer time periods and to which they can adopt quickly. In addition, modern RI ought to open their services to the public and where possible involve citizen scientists to meet the changed expectations of societies in a data driven world. The need to offer data specific services on infrastructures adds a level of complexity which needs to be funded by a variety of stakeholders.

[2] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=what

[3] Even large telescopes of today are associated with major data and processing challenges starting from the sensing device until delivery and analysis.

Trust Dimension

Establishing and maintaining trust in the mechanisms of RI requires the active involvement of researchers, they need to be in the driving seat to guarantee the necessary innovation. On the other hand, these RI are in general complex and operated by specialists who need to establish their own dynamics. This field of tension between researchers' interest in innovation and the intrinsic dynamics of infrastructure requires continuous massaging to guarantee continuing success of the infrastructure in terms of their services over time. It bears the risk of failure and thus requires a pro-active strategy for assessing the quality of the services, a proper certification of the processes, the level of user trust that has been established, the state of global competitiveness and the efficiency and cost effectiveness of the infrastructure. Since RI will be evaluated with respect to how well they accomplish their core activities on allocated research budgets there must be mechanisms to stop or re-purpose such infrastructures if the assessments do not show satisfactory results.

ESFRI Process

RI are very well known in history and some even call the library of Alexandria, collecting the knowledge of ancient world, one of the first. Single sited infrastructures, particularly in physics, were established where the costs were so high that they could only be shared by countries. The massive take up of the Internet and the huge potential of distributed structures necessitates rethinking the design and the mechanisms of RI. The ESFRI process, started by the EC and the member states, faced this challenge in an admirable way, since it put these new requirements on the agenda of all relevant stakeholders. It resulted in roadmaps and concrete actions: 48 new research infrastructure initiatives in many research domains were started based on European and national funding. With the ERIC (European Research Infrastructure Consortium) concept, a possible legal form for such European-based RI has been worked out and an increasing number of initiatives seem to be able to turn their initiatives into legal entities by ensuring funding commitment from member states[4]. However, this does not necessarily meet the sustainability needs of researchers as the most relevant users, since funding for new types of distributed and virtual infrastructures is often only ensured for a few years. Funders frequently adopt a different stance in funding large physical facilities from new virtual infrastructures — everyone accepts for example that experimenting with a new type of fusion reactor costs many years of preparation and construction but in the case of the new type of virtual infrastructures it is assumed that the required integration and interoperation can be achieved within a few years. This assumption is wrong, although these RI need to adopt mechanisms that allow them to commence offering services at an early stage. Also usage costs are often not budgeted. Funders are hesitant to make long-term commitments, and in so doing endanger the success of these infrastructures[5].

Elements of Sustainability

With respect to a definition of the term "sustainability" Wikipedia examples: "In ecology, **sustainability** is how biological systems endure and remain diverse and productive. Long-lived and healthy wetlands and forests are examples of sustainable biological systems. More recent accounts have broadened the idea of sustainability to include social wellbeing, resilience and adaptation

[4] The ERIC legal concept is not the only one known for research infrastructures. GEANT/DANTE for example are using a different legal form.

[5] It took years for states to understand the potential of railway systems and to see it as their role to help building nation-wide infrastructures to easily move people and goods.

across four domains: ecology, economics, politics and culture. In economics-centered accounts, sustainability requires the reconciliation across the "three pillars" of sustainability: economic demands, environmental resilience, and social equity."

When applied to RI we should include, for example,

a) the capability to survive under changing cultural, political and economic circumstances by offering attractive opportunities that fit the demands of societal stakeholders;
b) the capability to turn research content and workflows substantially in a direction that enables tackling the small and big research challenges of the time by making use of state of the art technology and in particular the new possibilities offered by the Internet;
c) the capacity to work along innovation cycles, i.e. at certain moments in time smaller or larger measures are required to maintain a position that allows doing competitive research. While for large physical installations this often means replacing sensor technology by a new generation, in distributed information infrastructures this means replacing the existing layer of connectivity[6] by a new one that yields a new level of integration and interoperability including semantic technologies.
d) a combination of providing stable services based on proven procedures and technologies and the will to invest in innovation where the cycles are determined by a number of factors. Whatever these cycles are, researchers need to be assured that infrastructures will be supported at least for 10 years.

With some exceptions in particular for large single-sited and distributed physical facilities, long-term sustainability is not yet ensured which will hamper take-up by researchers.

Recently a discussion was started stating that it is the task of infrastructures to also guarantee the long-term visibility and accessibility of data which are part of our record of science[7]. For some experimental data it is certainly true, for example, that they are being replaced by measurements from new sensors with much higher spatial and temporal resolution. Often such data does not need to be stored for longer than a period of roughly 10 years to follow good scientific practice. Much data however, has a scientific and societal value requiring permanent storage. European initiatives such as OpenAIRE and EUDAT have been setup to meet these long-term requirements at European level. At national and even research organization levels strategic discussions are taking place to set up permanently funded centers that can preserve data over long periods and give access to it. In the case of EUDAT[8], European and national strategies come together since the involved centers are to a large extent also pillars in national or organizational strategies. Therefore long-term sustainability is at this moment ensured by national or organizational policy decisions, but not by European policy decisions.

Commercial Offers

An increasing number of commercial services are being offered particularly in the domains of data, information and knowledge, since industry realizes their enormous value which will grow further in

[6] With connectivity in this context many layers ranging from APIs to semantic interoperability are meant.
[7] High Level Expert Group Report "Riding the Wave"
[8] Similar holds for other regions such as DataONE in the US.

the future. This does not seem to be a viable solution in the long run despite attractive service offers, as industry:

1. needs to address a mass market and most scientific infrastructure building and data provisioning does not address a mass market;
2. is interested in offering specific added value services that can be maintained easily while research is interested in maximal flexibility;
3. is interested in establishing dependencies and dominance which will hamper innovation whilst research is dependent on innovation[9];
4. will offer continuity only as long as this brings sufficient profit and as long it can survive in competition;
5. so far as information goes is largely in the hands of US companies and a competitive European industry is not in view[10];
6. has a current business model which is likely to be changed to costs per service.

Taking all these factors into account it is obvious that researchers cannot rely on commercial applications only. They will make use of commercial or industrial offers as long as they are of interest and affordable. However, close collaboration with industry has shown its high impact on driving innovation, i.e. a bi-directional fertilization is highly attractive.

In the following chapter we will elaborate on the reasons why RI should be sustainable. In chapter 3 we elaborate on some dimensions that influence sustainability and in chapter 4 propose recommendations.

References

- ESFRI Roadmaps 2008

2. Reasons for Sustainability

In the previous chapter we described what we understand under “sustainability” and indicated that it is part of the trust building which is essential for take-up by researchers. In this chapter we will describe other reasons that make RI and their sustainable funding a necessity.

From the viewpoint of different stakeholders, building and maintaining RI only makes sense if there is continuity for a sufficient period of time. Dependent on the types of RI, we indicate some differences:

- ‘**single-sited**’ RI are mostly based on big investments in a specific physical facility supporting a set of experiments or observations. Despite these facilities becoming obsolete after a period of time, the availability of specifically skilled experts and the existence of specific industry in the neighborhood suggest continuation;

[9] With respect to innovative services in the web some industry is currently in a leading position, but this may and will change again.

[10] Helix Nebula is a good initiative, but it can only compete at storage level. But the competitive gain can only be achieved at the services level.

- '**virtual**' RI are forming at the other extreme in that huge efforts have been invested into setting up the virtual "connectivity" layer (connecting, integrating and achieving interoperability). Also, in this case, specially trained experts need to be available to maintain the RI and industry branches emerge that both rely on and add to this expertise.

The following reasons for sustainability of RI with slightly different foci dependent on the type can be found in the literature:

- Cross-national RI allow the big challenges of our times which require cross-disciplinarity and make data and services available to all interested. Not only do they have a high scientific value, but also a huge societal impact.
- These RI connect experts and integrate efforts across borders and have a high potential to push innovation by virtue of this cross-disciplinarity, its cross-country interaction and when a fruitful partnership with industry can be established.
- Due to their cross-disciplinarity and cross-country setup they have the potential to prevent duplication of efforts and thus promise to increase cost effectiveness. Consequent to their contribution to structuring the landscape they result in the further strengthening of coordination.
- In circumstances where the need is for researchers to integrate data and services from different disciplines and even sub-disciplines it is apparent that infrastructures have the potential to allow researchers to focus on their research work again and leave the typical integration and interoperability work to trained experts. However, trust in long-term availability and robustness is essential for researchers to invest time in learning how to make best use of them.
- At those locations where RI have their main centers (be it single sited, distributed or virtual[11]) we will see a societal impact in so far as new enterprises will emerge and thus new jobs will be created.
- For all types of RI discontinuity would result in a loss of difficult to build up expertise, in a tremendous degradation of the innovative potential of a region or at various places and thus a loss on the investments that have been made.

A recent science workshop on data co-organized by RDA/Europe and the Max Planck Society, which attracted 17 leading scientists from Europe, clearly revealed that researchers would like to have robust and highly available infrastructures that can facilitate the integration and interoperability of their work. The meeting concluded that a number of principles must be met to achieve acceptance of infrastructures, including: openness, long-term availability and availability of expertise and guidance.

[11] From almost all virtual research infrastructure initiatives we have seen that they rely on a network of strong centers (expertise, facilities, and services).

3. Dimensions of Sustainability

3.1 Political and Social Dimension

As has been stated, during the past decades scientific progress has become increasingly reliant on large scale collaborative efforts involving shared facilities and resources and variable numbers of researchers and participating organizations, often geographically distributed among the different countries. Initially these were confined to physics and astronomy, nowadays however they extend to other scientific branches, social sciences and humanities. This leads to **big RI being in the position of transcending national borders and economic blocks**. Therefore **their funding is, or can be, influenced by cross – national harmonization and political agreements.**

However, it becomes clear that social and subsequently political debate and interest around the different problems is not at the same level everywhere. For instance, while climate science debate is reaching increasing vehemence in the western world, it is rarely debated elsewhere.

Not only does the **scientific mission** need to be agreed. The international consensus is also important to determine the **location of the facilities and Infrastructures** and the configuration and dynamics of the scientific community organized around them.

Probably it could be stated that **Big Research Infrastructure expenses** that involve the decision of different governments from different countries for a long time, **would deserve significant political agreement at our European Institutional level**. Long-term investment in such scientific development requires detailed understanding of the potential impact sustained on solid based data of utility, results, scientific interest, transparent managing models etc. A complete analysis of the **social impact of these structures will be required:** what will be the impact, where will this impact materialize, what form is it likely to take, who is likely to benefit. In two words: the potential to contribute to scientific progress but also and more widely to impact in society and economy. What has been called "**the big picture of the big science**."

In any sphere of social life it is not easy to assess how much influence particular people, ideas, products or organizations do have on others. We are forced to look for indicators or ways of measuring this influence. We cannot define research impact only in terms of the auditable influence of research on other actors or organizations. It also concerns social outcomes (business, government, civil society, etc.) either positive or negative.

This kind of analysis should be organized around different elements:

- ✓ **The scientific impact itself**. The core scientific research. How does research influence agendas in adjacent fields of science? What can be the motivations of Governments to fund big Infrastructures and what factors influence those motivation.
- ✓ **Human resource development**: teaching, training, local economic development, direct and indirect jobs.
- ✓ Potential for **Innovation**: Technical R&D assistance for infrastructure building and management. Markets for engineering supply.
- ✓ **Industry collaboration**: joint research and development projects in collaboration with external stakeholders. **Spin-off and Spin-out companies creation**.

- ✓ **Service provision**: access to unique facilities goods and data (bits and atoms)
- ✓ **Delivery of impact**. Intellectual capital. Social knowledge and popularity of the research. Contribution of the research towards accumulation of human capital in member countries. Track the impact of the research on macroeconomic indicators, both, at local level and across member countries (GDP, HDI)
- ✓ Is the research actually **addressing societal challenges** and concerns?
- ✓ Due to their own characteristics we may very often find **monopolistic structures** among the big research programs. A pure monopolist becomes a single seller. It can come from organic growth or through alliances or mergers (mainly the case of the big RI), via horizontal or vertical integration. Probably the challenge is not to prevent these monopolistic structures but to **prevent the presence of barriers to entry** for those Academies or Institutions (not involved from the beginning) that can be interested, and may add value to the project. It is not a matter of promoting competence since the interim philosophy is to fund collaborative unique structures and projects. On the other side of the coin, the **true cost of the monopoly power may be the loss of productive efficiency** that can lead to a **waste of scarce resources**. The quality of the processing and services need to be carefully analyzed and the waste of public funding adequately prevented.

3.2 Governance and Organizational Dimension

What and *how* are often the first questions that the scientists and science community need to tackle when initiating and planning a new research infrastructure. The question related to *what* (namely on the content) is obviously the main topic and forms the core of the RI operations. Thus the content needs to be addressed first but the content matters are immediately followed by a question of *how* (namely the organization, governance and resources). As the first ideas of the organization are emerging in the beginning and they are tackled mainly by the scientists, the first organizational plans of the RI are based on the scientific orientation, not on previous experience of setting-up European or international governance structures. Only rarely in the initial phase does the planning team have in-house expertise on planning organizational and governance structures at the level of multi-national/European or international RI. The expertise in setting-up the governance is either increased over the course of the planning, construction and operational phases (internal learning process) or it will be brought into the RI project in the later phases whilst the first decisions on the organization of the RI and the governance have already been made.

A good and feasible governance structure is the key to be a successful and sustainable research infrastructure. It is much more difficult to change a poor governance structure later than for example a poor management of the research infrastructure. A good governance structure requires high commitment and orientation from both the science community but also from the political stakeholder (funder) level. So, both bottom-up and top-down processes need to be managed. As mentioned above, the planning of the governance is first done mainly by the scientists and in the later phases the stakeholders will commit to the development of the governance structures. It is often the case that the stakeholders are much more experienced in creating and operating governance structures than the involved science communities. However, the set-up team for governance needs to be equipped to deal with a multitude of political perspectives, seeking consensus solutions that work and have processes on how to solve rising conflicts.

Over the course of the ESFRI process there have been many cases where the scientists have successfully learned to handle the organizational and governance matters and involve political stakeholder level. But there are also examples where RI projects have failed because the planning/construction team has had difficulties in building credible governance together with the finances for the research infrastructure. It is rarely the content which has led to a failure in implementing the RI plans.

It is not easy to fill-in all the requirements and political positions of the stakeholders and the science community into a balanced, efficient and effective governance structure. There is a risk to create fat and complex governance structures in relation to the operations just because there is a need for seeking political consensus solutions. Normally this is reflected in very complex power definitions (e.g. voting rights) and mixing the role of the bodies, decision-making and day-to-day management issues. This risk can be decreased by building trust among the partners and by reserving enough time for setting-up the governance structure. Therefore the work of establishing a sustainable governance structure requires time and a patient attitude. It is also crucial to clearly identify the roles and the responsibilities of the different bodies in the governance structure and acknowledge that different phases of the life-cycle of research infrastructure may need different governance structures.

It is also a challenge to maintain a high-level quality and state-of the art services for the user communities over time. Therefore, it is important to recognize that the drivers of the research infrastructure – also on governance issues - are the science and user communities. Especially in the case of distributed RI which have many national nodes/actions, it is crucial to keep the national level science communities involved and motivated also in the operational phase. This task requires resources and good coordination and should not be overlooked because it directly impacts on the sustainability of the research infrastructure.

Different types of RI require different governance solutions (centralized vs decentralized) and when seeking the best governance model for the research infrastructure there are plenty of examples in the research infrastructure landscape from which it is possible to learn and seek best practices and models. The main issue is ensuring that the governance and organizational structure best serve the scope and functions of RI. Current RI have chosen many different legal and governance models and it would be very helpful to have a European wide survey on analyzing what are in practice the strengths and weaknesses over the life-cycle for different types of research infrastructure in relation to legal and governance models. At least, in principle, ERIC model allows an establishment of flexible governance structures but time will show how sustainable the governance structures created around the ERIC legal model are. In addition, knowledge transfer and staff exchanges between the RI would enhance sharing of the best practices and expertise. Commission could facilitate these processes by offering support to RI to enhance the collaborative actions dealing with governance and management related issues.

Setting-up a legal entity with workable governance structure is not an easy or a cheap task and when a science community has a plan for a new research infrastructure it would be wise to analyze the research infrastructure landscape very carefully and seek out whether there are already operational RI that might have close enough science focus and that would be interested in a close partnership.

Because it might be so that the planned research infrastructure could be merged into an existing research infrastructure as a new operational infrastructure component without creating a totally new legal entity with the heavy governance structure. This kind of development would also reduce the risk of potential fragmentation in the RI landscape, which is a real danger if all the selected ESFRI research infrastructure projects are establishing their own legal entities and governance structures. Commission can play an important role in initiating and facilitating the European level collaboration between the science communities, promoting wider partnerships and integration processes, such as creating funding instruments supporting the upgrades of the existing RI with new infrastructure components.

RI sustainability needs a continuous quality assessment and the simple measure for the quality is the usage of the RI services by the user communities. In addition to monitoring the performance of the RI operations and the services, also the governance structures and management should be a subject to periodical evaluations. The methods of evaluating the governance and the management structures should be transparent and coherent among those RI that the periodical prioritization assessments over the RI landscape assess. These kinds of evaluations are important also when the exit strategies of the RI are planned. ESFRI and Commission are in a central role for ensuring the balanced and transparent evaluations among the RI. They also have an important role for communicating the results of the assessments to the national stakeholders who are responsible for national research infrastructure developments and roadmaps.

3.3 Financial and Business Dimension

Financial aspects and business models with respect to *sustainability* of RI (RI) are briefly investigated below. This important aspect of sustainability is emphasised here because of the overall not-for-profit character of RI activities. In most cases, there is no need for a fast return on investments, and there is no need for an effective profit-making model. In the case of scientific advancement however, it is not straight-forward to calculate any form of explicit return-on-investment. However, in a wider indirect sense, via economic exploitation of the research results achieved from utilising the available RI, the RI investments can in fact produce a return in a short period of time. On the other hand, there is a need for careful handling of the financial aspects (costs and cost coverage) and also a need for sophisticated business models towards arriving at an assessment of the financial stability of RI development and operation. The achieved financial stability is a basic pre-condition of sustainability.

Indeed, financial aspects play an increasing role in the sustainability of RI. In contrast to the early periods when conscious and organised development and utilisation of such infrastructures started, nowadays the financial element amongst the sustainability dimensions turns out to be a major determinant rather than a negligible component. Moreover, the case gets to be increasingly demanding in the event that availability of traditional financial resources to cover the costs of developing and operating the RI are inadequate. In this latter case the introduction and due application of complex business models for developing and operating the infrastructures is needed. Such business models should take into consideration:

- all technical and organisational aspects,
- all stakeholders (owners, developers, operators, users, supporters, etc.),
- short-medium-long range processes and procedures,
- input and output channels of financial resources,
- flexibility and adaptivity of handling those financial resources and processes by taking into account the complete value chain, and
- the resources necessary for the planning, decision making, controlling, and evaluation of components of human activities devoted to those financial matters and the business model based operational issues.

The evolution of this economic approach has emerged from the fact that central (government etc.) financing is no longer able to completely cover the RI development needs and especially the costs of operating the RI. Although there is considerable on-going effort on behalf of the EC and the Member States within the EU, and also in a number of overseas countries, to duly stimulate and support building, developing and operating infrastructures in general and RI in particular, there is seemingly insufficient funding and it is likely to further decrease as a percentage contribution towards covering the overall costs.

Indeed, there are remarkable processes going on:

- complexity and size of the infrastructures increase,
- procurement, development, and also operational costs are fast escalating,
- geographically and organisationally distributed but simultaneously operationally integrated combinations of RI components are playing a gradually increasing role, whilst wide, remote use of the costly, and in many cases unique, infrastructures and RI services gradually substitute for the traditional, individual building, operating, and exploiting of infrastructures and services.

Obviously the traditional funding practice has to be revisited and new business models are to be developed, tested, and, in cases where appropriate, duly applied.

Here an important question arises: What are the features of the traditional approach and what different peculiarities characterise the supposedly new models?

- Historically, in the early periods, individual research organisations made in-house decisions about investing in instrumentation, i.e. equipment and tools needed for doing research, especially scientific experimentations. Since wise investments could result in re-usable instrumentation (by several research teams and for several research topics, etc.), building local RI soon started investing consciously into such re-usable instruments. In-house financial aspects did arise, sometimes these were resolved by involving outside funding but practically never by generating income by offering the RI components to outside users. Accessing the infrastructure has obviously been free for the researchers belonging to the related research organisations (owners of the procured instruments).

- The next period can be characterised by the parallel process of rapidly increasing infrastructure costs (both building and operating RI) and wide recognition of the potential to be gained by co-operating both in the building and operating of those expensive infrastructures. Common exploitation of the infrastructural resources and opportunities by co-operating (or even competing) R&D organisations turned out to be the only possible option of successfully building, operating, and accessing such high value infrastructures and related services. Most of the services were offered mutually by the involved research organisations to each other without requesting any charge or fee for using them, and this simple exchange of services proved to be a viable, well working solution at that time.
- Another milestone arrived with the emergence of the ERA (European Research Area) initiative. Joint efforts at all levels on the European scale to build and operate most complex, unique, sometimes distributed but always integrated RI and enabling remote access for all and any interested R&D role-players has turned out to be a major process by exploiting FP5, FP6, and FP7 project funds allocated specifically to such specific RI (with the expert advice from the ESFRI in the background) and also national funds available on MS level. Of course such a transformation of the financing for building-developing and operating RI has happened primarily (and almost exclusively) with high complexity, high value, widely exploitable RI and consequently most of the lower complexity, lower value RI components remained under the traditional financing frameworks.
- However, the last few years have demonstrated that the increasing complexity and elevated costs of building and operating especially those extreme RI no longer allows a financially sustainable opportunity. Allotted EU and national government funds are not adequate to cover the increasing costs and therefore new models, based, at least partly, on a “pay for use” approach are to be introduced where the costs of building and developing the infrastructure are, at least partially, covered by the potential and early practical users, while accessing and using the completed infrastructure and its services are charged so that the fee paid by the users (researchers) contributes to covering the operational costs. Special aspects should be taken into account in progressing towards financing of jointly accessible infrastructural resources when, on one hand, global scientific co-operation is strengthening, and on the other hand, industrial partners in PPP frameworks enter the picture.

Five important comments should be made here.

1. First, there is of course no common, uniform business model for all infrastructures. A set of such models can perhaps be developed and introduced so that in the simplest case the traditional financing is maintained while in more complex cases some of the models can well be fitted to the specific features and characteristics of the infrastructure and its operation.
2. Second, it is of great importance to better learn to split operational costs from innovation costs, since their underlying funding scheme is completely different.
3. Third, paying for access to research infrastructure has never been taken into consideration in the past when calculating aggregate project costs, whether estimated or predicted. Therefore the need for assessing such costs when preparing project proposals and later, during the execution of the projects, is something new, which of necessity will have to be understood, accepted, and also learned, experienced, and finally, mastered by the researchers, research teams, or research organisations and also research funding bodies.

4. Fourth, following the transition into financial and business modelling based operations, it is recognised that there will be a period of the VRC (Virtual Research Communities) based symbiosis of the involved research organisations. A symbiosis extending not only to joint R&D activities but also, among others, to joint investments and joint coverage of the operational RI costs by the VRC constituents/members. It is still not known if such a development in integrating resources for achieving common goals will result in a "pay for using the resources" or a different principle and practice. Nevertheless, healthy finances will remain a primary pre-condition of RI sustainability – which, will practically probably also coincide with the sustainability of the related VRC itself.
5. Fifths, compatibility and interoperability as well as related standardisation, are indispensable pre-conditions for proper operation of distributed-combined-integrated RI (or RI components). Since joint financing (investment and cost coverage) assumes and, in successful cases, offers seamless integration of the involved infrastructure components, attaining such a compatibility and interoperability is a prerequisite. Fortuitously, often e-Infrastructures offer functionality for transferring information between the RI building blocks and services, as well as between the users and the RI accessed by them, therefore the compatibility and interoperability requirement is in most cases taken care of by the e-Infrastructure. Here is a task for RDA to help in harmonizing the principles and procedures in the coming years.

An interesting mixture of the above options can be foreseen in the near future as flexibility and adaptivity become extremely important factors of building and operating RI – a challenge both for infrastructure owners and for infrastructure users, whether they be fully or partly separate, or fully or partly integrated.

Moreover, flexibility and adaptivity will be needed in continuously developing the applied business models themselves, partly because of the regular improvement of the methods and their use, and partly because of the changing conditions and circumstances in building, developing, and operating the RI.

The next few paragraphs list some specific aspects of how appropriate financial approaches and business models can be developed and introduced towards achieving the sustainability of RI.

The following cases should be distinguished and handled separately:

- Small individual infrastructure components vs large distributed infrastructures (and versions in between)
- Infrastructures for research vs infrastructures for research-development-innovation
- RI in general vs e-RI offering cross-disciplinary services and achieving economy of scale factors.
- Different contributors (funders of building-developing-operating the infrastructure) and different users (accessing free or charged) services
- Specific (public vs private) partners in PPP collaborations
- Regular (traditional) procurement vs PCP (Pre-Commercial Procurement) in building the infrastructures

Building, developing, and operating a RI by using appropriate business models will enable a good balance between available financial resources and emerging costs even in the case of limited (or zero) non-repayable central funding – simply by using a well-established, wise, careful cost strategy and a friendly, transparent, fair charging policy {*need examples*}.

Best practices are not yet available but surely will follow soon.

However, early recommendations (for builders, developers, operators, users, and funders) can already be derived:

- develop various business models matched to characteristic RI types,
- define RI classes (with associated business models) and allocate concrete RI to specific classes (wrt. type of financing in building, development, and operation),
- follow life-cycle of RI by taking into account specific financing aspects and specific business models (depending on class allocation),
- separate clearly operation/maintenance costs from innovation costs
- provide user access* to RI in accordance with the allocation (the selected RI class),
- refine allocation (perhaps also the business model) if needed,
- establish set of best practices,
- urge research project owners to consider charges for accessing RI services in the execution of their projects.

A final comment on sustainability (the keyword of our study here): sustainability of a healthy RI can be strengthened by a good business model, but even healthy RI can lose sustainability if an improper business model is applied, while of course sustainability of ill RI cannot be established even by applying the best business models.

_____

* Cf. EC proposal on a ”European Charter for Access to Research Infrastructures”, 21 Feb 2014

3.4 Legal and Ethical Dimension (some experts to be asked)

Modern RI even when single sited are federated in some way:

- users can remotely access experimental facilities
- users receive and process the results of experiments (data) in distributed infrastructures
- users build collections of data from different sources spread across Europe and even beyond
- etc.

From this description we can conclude that modern RI lead to complex legal and ethical issues since different legal systems and cultures are involved. Overcoming legal and ethical hurdles for easy and secure access to distributed facilities is very time consuming and sometimes, as in the case of sensitive medical data for example, hardly possible. Legal and ethical systems emerged as a result of long trust building processes, thus changing them will requires time. In particular through ESFRI Europe has started to build pan-European infrastructures, it is urgently required to start a phase of legal and ethical harmonization to make infrastructures work more cost-effectively.

The ethical dimension includes many aspects up to ethically correct scientific behaviour (breaches of research integrity) which is at the core of trust building in an era which is very much determined by anonymous relationships between the actors involved in all phases of research or research data usage.

In addition we recognize that there is a general awareness that data, information and knowledge has a high scientific and economic value. As a consequence, various stakeholders started trading such entities, i.e. closing it from free and open access. In some, disciplines such as the humanities and social sciences, this creates hurdles that seem to become bigger.

Federated methods for distributed authentication and authorization are still in their infancy and do not work in a pan-European scenario although secure mechanisms are important to protect against intrusions of all sorts. Since these methods require actions on the technological and in particular on organizational/political levels we need a concerted action involving different stakeholders to come to a secure environment as a basis for building trust.

As a reflection on this complicated situation a few trends can be recognized:

- The Internet and in particular the World Wide Web is a growing space of open information where in particular the young generation is establishing its own culture often ignoring rights.
- A world-wide open access movement is changing the attitude towards making access to publications and data free of barriers. Of course it is accepted that there are good reasons to protect some of the data.
- Governments more often come up with new rules stating that publically funded research must end up in publically accessible results.
- Industries such as Google are big enough to define their own legal terms and to widely ignore restrictions. Since they are operating worldwide almost everyone has access to their services.

Another aspect of the legal dimension is the legal framework that has been setup to establish RI as legal entities at the European level. The ERICs are owned by the member states, which ensure their anchoring in the national roadmaps which is important at a political level. The first such initiatives have followed the lengthy procedure and have been established as ERICs. Yet it is too early to speak about experiences.

In appendix A we will include some statements about legal issues which are important in this context and which have been provided by legal advisors working in RI.

3.5 Technical Dimension

CDI Framework

Modern RI and e-Infrastructures are an intrinsic part of an eco-system of infrastructures which the EC's High Level group on Scientific Data, when restricting the focus to data, called a Collaborative Data Infrastructure. Whatever the type of infrastructure to be established, it ought to be designed in such a way that it can become part of this modern distributed landscape. This landscape will be such

that it has users in various roles, located at a wide variety of locations, connecting with a variety of devices including mobile, comprising various discipline specific infrastructures with centers at various places and cross-disciplinary e-Infrastructures which are in general also established as distributed and virtual networks. Any design must take account of the specifications and requirements put forward by this eco-system of infrastructures else it will fail. In EUDAT it was found that two types of participation can be imagined: (1) “Using an infrastructure” in general means to establish an easy interface with the RI having restricted functionality but without having to adapt the users own infrastructure, and (2) “Joining an infrastructure” means to adapt the user’s way of operating to make use of all advanced functionality. Finally, although joining might be the optimal way, yet there is usually insufficient agreement on basic issues and principles and consequently “joining” is a method for the future when RDA will have produced the required results.

Functions and Services

RI need to provide a whole range of functions and services. These can be aligned variously along the continuum between creation and consumption of data and a variety of activities around this lifecycle process. This comes close to a tier concept which we can identify in some form in many disciplines.

In some form infrastructures are involved in the process of creating data by running big experimental facilities, maintaining sensors, executing simulations, running campaigns to collect observations or by engaging in massive crowd sourcing. Infrastructures, independent of the type, need to take care of the generated data, perform appropriate preprocessing according to the scientific needs, make data accessible in useful ways and push data towards institutions that will further manage it. Issues such as quality control, process documentation etc. are important for subsequent processing.

Other institutions will take care of functions which can typically be associated with the term lifecycle management and which typically will be carried out by some flavor of certified repositories[12]. Data needs to be managed, curated and stored permanently which includes the assigning of PIDs, metadata and organizing it properly. PIDs are typically issued by special, certified centers and metadata is typically harvested and aggregated to form interesting and searchable portals offered by different types of service providers. PID assignment and the availability of metadata are at the basis of data publishing and data referencing either for processing or for citing purposes. Curation, which means keeping data accessible and interpretable, can have many forms and may require special campaigns, i.e. curation costs can hardly be estimated.

Giving access to data is a function of the repository; however access can lead to a wide variety of ways of processing data. Proper data management needs to be policy guided taking account of explicit statements about processes to be carried out on the data for various purposes. Data analytics, derivation and enrichment can have many different forms - often including the combination of data sets within trusted federations. Important is that data processing requires explicit knowledge about its structure, its semantics and its provenance all of which should be accessible via the metadata. To manage the increasing amounts of data and its complexity there is a

[12] Certification is here indicated as a fact. In reality we are far away from having certified repositories although first assessment processes have been defined by DSA, DIN and RAC-ISO.

trend to make use of agreed and documented workflows. In some disciplines we can see specializations in so far as that some institutes are focusing more on the management aspects while others more on the analysis part. Since data and processing are two sides of the same coin infrastructures must keep both aspects in view. Both activities require specialised experts and continuity is necessary for efficient and cost effective service provision.

Maintaining software that is used for management and processing is costly; therefore a distinction needs to be made between infrastructure software which needs to be robust and special scientific software which needs to be subject to continual innovation. Sustainability of systems implementing an infrastructure needs to be added as a dimension to assessing the sustainability of the research process. A modular architecture, which takes care of standards, is embedding into distributed landscapes, which allows combining robustness of basic functions and innovation for cutting-edge tasks cannot be designed from scratch, but must evolve depending on the trends and technologies. Frequently, research imposes access restrictions for various purposes, thus security is an important aspect. Distributed systems are much more difficult to handle with respect to security, hence it is essential that today's RI which are inherently distributed at some layer need to set aside funds for appropriate experts.

Since RI often work at the cutting edge of technology or the introduction of new technology and methods, it is necessary to train experts to fulfill all requirements. Also, domain experts and scientists need to be trained so as to be able to adopt new methods and appropriately adapt their systems. Infrastructures need to give support and help to users since the RI are in reality complex systems that will have failing components and further are not immediately understandable in all their functionalities.

Implications for Sustainability

RI of all types are complex technical systems that are based on interacting components in distributed landscapes where robustness of basic components needs to be combined with continuous innovation for components close to science and where secure mechanisms are absolutely essential. The technical systems of RI are in themselves living entities in all phases due to technology developments, the need to replace inefficient components and interfaces by adding new ones and incorporating new functional wishes. Trained specialists are required to guide these processes during all phases.

Disruptions in funding streams will cause system errors within a short time and fluctuation in the team of experts will lead to inefficiencies, malfunctioning and loss of trust of users. Sustainability of RI is a necessity from a technological point of view.

3.6 Human Resource Dimension

Training, education and human capital development are key to exploiting RI to their full potential.

Users benefit from training and user guidelines pertinent to RI to inform their data preparation, their use plans, and their findings; trained users will be more effective and consume fewer resources.

Higher Education courses targeting data science, data management, data analytics, digital archiving, and information science should be prioritised with suitable accreditation, in order to train the cadre of data scientists required for both RI and also the data industry. As well as dedicated courses, hands-on skills workshops and online educational resources for researchers who find themselves using significant quantities of data can enable improved data management and use. Use and re-use of RI should be incentivised via metrics that feed into career progression. Similar to the 'h-index' there need to be standardised, commonly used metrics to capture the use and citation of RI.

Data metrics are currently under development, and a standard approach to data publication, citation and attribution would speed up this process, with the adoption of standard metrics into funding assessment processes incentivising their use. Data scientists need to be rewarded and incentivised with a recognised career progression path, so as to build a sustained skills base to ensure that RI can be sustained.

The relationship between RI and Research Organisations (Universities, Colleges, and Research Institutes) should be reviewed to maximise human synergies, through exchange of personnel and shared training and skills exchange. Research Infrastructure staff should receive the same supports and benefits as RPO staff.

Gender balance should be addressed for Research Infrastructure human capital. Research Infrastructure staff should work towards gender balance, and promote gender role models where appropriate. In implementing a transparent access charter for RI, care must be given to ensure gender balance appropriate to the range of access requests received.

4. Summary Statements

In this chapter we present our statements which we extracted from the previous document and from the discussions about it.

4.1 Biggest Challenges

Long-term Sustainability

Long-term sustainability of research and e-Infrastructures (networks, data and computation) is mandatory in order to establish a relation of trust and acceptance between users and infrastructure providers. The current hesitance of funders to give long-term commitments in particular for virtual RI is counterproductive since it inhibits researchers' take up. Long-term commitments require quality criteria, evaluation procedures and opt-out options. Long-term funding needs to be based on the involvement of all stakeholders, including the European Commission, as has been shown successfully for GEANT. It is not realistic to assume full self-sustainability from a Research Infrastructure; governments will need to keep infrastructures at high profile. Flexible business models are essential to cope with societal, political and technological changes.

Complexity Requirements

Modern RI are distributed and virtual since data and algorithms are widely sharable. Such virtual infrastructures are inherently complex, requiring culture and language bridge building, stakeholder agreement, agreements on sharing and interoperability which normally go beyond the individual infrastructures, excellent people to a) construct and operate the multilayered systems that implement the infrastructure; b) take care of the required security mechanisms and c) help users in innovative ways to find solutions and do problem solving. Only having long-term strategies in place will allow the generation of justifiable roadmaps and perspectives to fully engage developers to tackle complexity.

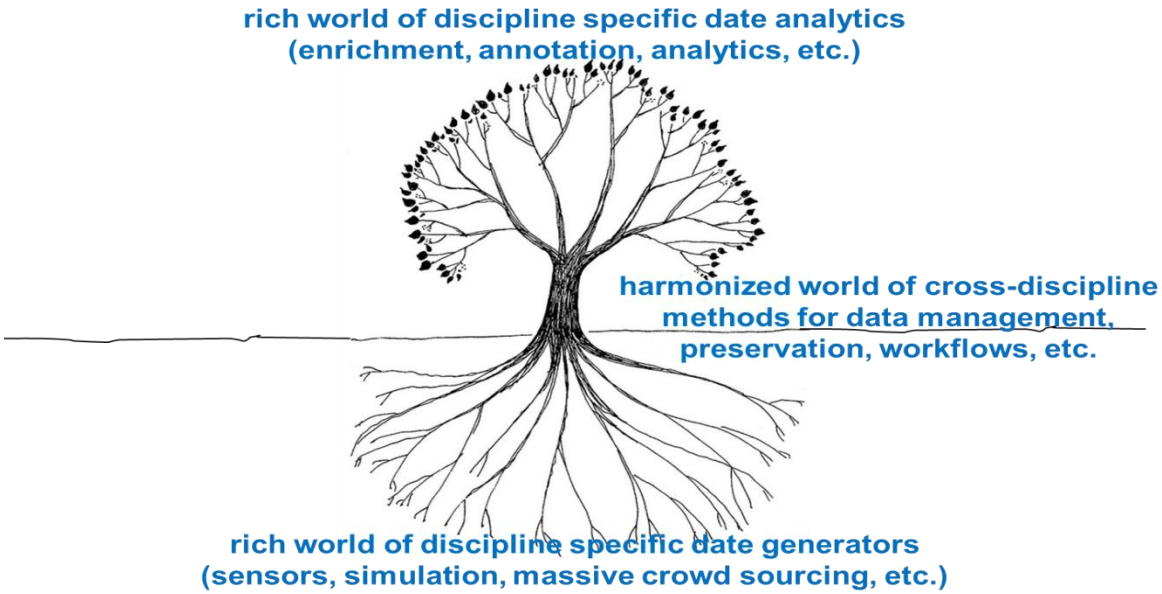
User Acceptance and Innovation

In particular although the new virtual infrastructures and the changing e-Infrastructures now better understand that user orientation and cost-effectiveness are primary goals they still need to find acceptance by users. Short-term and innovation expectations from policy stakeholders seem to be in opposition to the needs of researchers who wish to have robust and functioning services based on proven technologies. Finding a good balance is extremely difficult given the impatience of the various stakeholders. Innovation and service provisioning need to be clearly split since they are based on different funding principles; mixing them will lead to unclear cost structures and responsibilities. Acceptance building is a process which takes much time.

Evaluation and Decision Taking

One of the biggest challenges is to arrive at a balanced evaluation process based on rigorous metrics which is carried out by objective, constructive thinking experts and scientists who are willing to include the development potential into their considerations. Wise decisions based on infrastructure roadmaps, funding possibilities, developmental potential and societal impact are urgently needed to evolve infrastructures to professional services, and yet be cost-effectiveness in a changing eco-system of infrastructures. Opt-out options must be possible as well as requiring drastic adaptations

due to changing requirements (see the diagram below as one example). Further, establishing a balanced evaluation process and adjusting it regularly only makes sense when there is a long-term perspective.



This diagram[13] indicates that the middle layer of data organizations and data management evolves to a harmonized island of sharing and interoperability based on an increasing amount of global agreements. This does not mean that we will have for example a reduction of data formats, but that we will have a landscape where we are able to deal more efficiently with the heterogeneity based on meta-standards. This diagram also stresses the need to reshape the eco-system of RI, since currently almost any infrastructure is building its own data management components. RDA is the motor of harmonization in this middle layer.

4.2 Key Assumptions

What are the **key assumptions** underpinning and driving the future developments and expectations for RI and e-infrastructures, in particular in relation to the challenge(s) identified above (for example, regarding research & innovation, demand side and consumer behaviour, policy needs, or the concerns and expectations of citizens and civil society)?

International Competition

Competitive regions (e.g. Japan[14], China, US) invest enormous amounts of funding to establish RI with the intention of maintaining their data driven research at a competitive, if not leading edge standard. Dependent on the respective socio-political-economic situation different strategies are applied. In particular long-term sustainability is being taken care of via different mechanisms. To properly address this challenge Europe should must invest substantial funding to remain in a leading position.

Trusted Repositories and Registries

We assume that data accessibility and persistence will improve over time. This can only be achieved, however, when we can rely on proper data organizations and on a landscape of trusted and certified

[13] This diagram is translated from ideas presented by Juan Bicarregui.
[14] Using the example of the Exascale supercomputer, Satoshi Matsuoka said at ICRI2014 that true forward-looking innovation and research is possible because funders invest speculatively without guarantee of return, while industry investment tends to be more short-term/ROI focused.

repositories that not only store data but also participate in the continuous enrichment and processing of the stored data in the data fabric sense. A world-wide registry of such repositories needs to be set up to assist in user orientation and to enable new data services by anyone interested. Strategies are needed to preserve ephemeral data sources such as the web in trusted repositories.

Industry Take-Up

The role of research as the driving edge of innovation will continue and RI will consolidate as described above based on wise decision making. A consolidation towards re-occurring basic components will motivate industry to participate, use available services and offer new advanced services. Europe with its weak IT industry needs to become a place where more risk-taking based on experiences gained in the leading infrastructure work occurs. A change in thinking is required—this can only be achieved over a longer period of time.

Better Classification and Understanding

Europe, due to its extensive cross-border and cross-culture infrastructure work, has the potential to offer examples to other regions of the world. We need to move towards excellent classifications of infrastructures and their components, not only to become cost-effective, but also to adapt solutions to other circumstances. Carrying out such classifications requires long-term investigations.

4.3 Output & Impact

What is the **output** that could be foreseen, what could the **impact** be, what would success look like, and what are the opportunities for **international linkages?**

Landscape of first-class Services

The first and most important output will be a landscape of first class services open to all researchers and other interested groups such as policy makers, industry and the public. This however requires an open cross-border and cross-disciplinary market place for services where everyone can comment in open fora about the services and where open interaction will indicate which infrastructures are doing well and which services in particular are yet need to be developed. Such a landscape will dramatically foster research since traditional borders will be overcome, new combinations of data and services will enable new fields and insights and new groups will be able to participate. A market of new types of services will be opened with a consequent high impact on the generation of new jobs, which may well be regional at those locations where infrastructures are centered. Further this open market place of services needs to be stable before researchers and others will invest their time; however it has the potential to drive a whole new economy, particularly for the young. It will help to make the currently underexploited European infrastructures visible and usable worldwide. The necessary basis for this is a sustainable landscape of infrastructures.

Political Relevance

Another output of infrastructures is that they promote a stepwise transcending of national borders and economic blocks in Europe which is of high political relevance. The need for consensus finding is an act of political collaboration which will strengthen Europe. This output should not be underestimated in an era that will suffer from increasing instabilities.

Fragmentation and Costs

Another impact of building a balanced and sustained landscape of infrastructures with symbiotic trends is a dramatic reduction in fragmentation and redundancy, which in turn leads to more efficient work environments in which researchers can focus on what they are trained to do. Having a new generation of Virtual Research Environments at their disposal which gives them flexible access to a rich service offer, so enabling these researchers to achieve more relevant results in the same time frame leading to the conclusion that overall costs will not be increased.

4.4 Bottlenecks

Which are the **bottlenecks** in addressing these areas, and what are the inherent **risks** and uncertainties, and how could these be addressed?

Economic Constraints

Infrastructure building is expensive, costs time and needs to be funded by re-purposing parts of the budget. Currently it seems to be increasingly difficult for governments to convince all stakeholders to rank infrastructures on priority lists and to fund all infrastructure initiatives that have been initiated. Innovative and flexible business models need to be invented, including PPP, to overcome these expected funding bottlenecks. New types of services, such as on data collections, need to be offered to convince the public and politicians and hence open access to additional revenues. The fact that RI contribute to regional development and to social and economic welfare needs to be stressed. After having agreed on a start phase and assuming that the infrastructure is fully functional, financially difficult moments in time occur when renewal of major parts will be required and asking for new large investments for innovation which most probably will cannot be funded without governmental support. Sustainability requires that funders know about these cycles and are ready to reserve funds.

Priorities

Funders need to decide on rankings within the ecosystem of infrastructures and to set priorities between the scientific domains. Such decisions carry an enormous risk of turning out to be inappropriate in future and to possibly favour popular short term interests[15]. Yet it is not obvious how to ab initio organize a prioritization that includes long-term considerations: a process which is essential for sustainability. A permanent council of scientists, infrastructure providers and funders needs to be established which has the capability to regularly monitor the development of the infrastructure landscape and to give high level and balanced advice on optimisations and priorities.

Governance

Defining efficient and effective governance structures in a cross-border and cross-disciplinary setting as is as is the case in Europe is not simple. Lean and professional governance structures are the basis for a functioning RI and need to be the outcome of careful discussions between all stakeholders. Complex and inefficient governance structures are often the result of distrust, competition and egocentrism. Political balance, which has high cultural and political value, does not lead to efficient

[15] Big challenges such as health and climate stability are being discussed widely ignoring the relevance of cultural stability and the role of humanities for societal stability.

and effective structures per se. Actively involving leading researchers and putting them into the driving seat of infrastructure building is not easy to achieve, but is much needed to achieve goal orientation and to build trust. Errors in governance can be fatal for sustainability and are difficult to correct subsequently.

Human Resources

Constructing and operating RI can only be done if researchers are engaged and if highly qualified experts are available. For researchers, infrastructure building is often not attractive as a career option and in further the infrastructure experts (communicators, technologists) are scarce which creates a severe bottleneck in building top level infrastructures. Urgent actions are required to ensure stability of infrastructures and thus increase the sustainability potential. Community building, such as by programs like RAMIRI, is a very important building block.

4.5 Gaps and potential Game Changes

Which **gaps** (science and technology, policy) and **potential game changers** need to be taken into account?

Political & Societal Vicissitudes

In a multicultural setting as in Europe, political and societal dynamics which vary widely between the member states are the most important game changers. New governments set different priorities which may change the funding base within a short time. This is a danger for the sustainability of infrastructure construction and maintenance. Bad perspectives have the effect that the best researchers or experts will leave early in the piece with an immediate impact on the quality of services. Only a situation of long-term commitments on the one hand and mechanisms that allow a flexible reaction (business model, payment models, partnerships, etc.) can overcome these uncertainties.

Legislation

Having obtained agreement on the ERIC instrument was an excellent step forward towards establishing European cross-border infrastructures as sustainable legal constructions. Still, the member states need to collaborate more so as to harmonize legislation at various levels including the rules of how to work with sensitive data. Differences in legislation often create huge barriers to developing efficient and cost-effective solutions and in creating the open market place for data and services which are needed to foster a new economy.

Global Harmonization

Global harmonization in the infrastructures space will lead for example to easy data sharing and re-use at a global level, since research is global. Initiatives and organizations such as OECD, CODATA and WDS at a global level are important to define widely agreed principles that guide infrastructure work. Bottom-up driven initiatives, such as RDA, are also required towards overcoming the many practical hurdles that hamper establishing the open market place. RDA is a very young initiative and needs support. Infrastructures need to be motivated to actively take part in the bottom-up processes to establish RDA as the natural authority where agreements with wide impact can be made.

4.6 Innovation

In which areas is the strongest potential to leverage the EU knowledge base for **innovation** and, in particular, ensure the **participation of industry and SMEs**?

Innovation Potential

Infrastructures have a high innovation potential in different dimensions: (1) in their construction phase many barriers requiring novel solutions need to be overcome (political, sociological, cultural, and technological); (2) if set up in properly infrastructures attract creative scientists and technologists who invent these new solutions and keep driving towards new applications and research enabled by these new solutions; and (3) they have regional impact by involving interested industry (SMEs), the public, schools etc. Sustainability will play a key role in establishing the trust which is key to achieving engagement.

Industry Engagement

Industry and especially SMEs, being unable to think over a long range, are interested in stability, growth and strengthening. Although the simplest measure of their success in these respects is income and profit, they are probably interested in co-operating with the related research community if that co-operation promises to add to their stability, growth and strengthening. Components and application scenarios (on the infrastructure side) and products/services (on the industry side) should be thoroughly investigated, evaluated, and compared, in order to find those areas where common interest and joint effort will create an attractive, coercive force bringing research and industry closer to each other, and initiating a kind of symbiosis in this sense (where PPP is but one possible option). One such domain is being indicated by the tree diagram (see above) which indicates cross-disciplinary markets at two levels: (1) Data Management solutions needed by everyone and (2) the emergence of collected metadata and data domains that form a gigantic resource that will evolve into completely new business activities and markets in particularly suited for SME and startups.

Societal Challenges

Infrastructures are established with high expectations to enable tackling the main grand societal challenges by creating an easy to access infrastructure and by combining the knowledge, data and tools that are necessary. It can be expected those areas with strong leverage potential such as healthcare for ageing, energy supply and climate change will push innovation and motivate industry/SMEs to invest so allowing the running of Big Data applications, i.e. develop completely new pattern detection driven fields.

4.7 Balance Bottom-Up vs. Top-Down Support

What is the best balance between bottom-up activities and support to roadmaps or to the societal challenges?

Balance

Due to the natural competition for achieving the best results first, researchers would like to have maximal control over the resources needed. This bottom-up control will lead in most cases to the innovative solutions society wants to see. This is in contradiction with the need to build infrastructures that serve a wider community and which thus need to be put under shared control

including top-down influences to achieve the broadly usable solutions necessary. Top-down decisions in principle tend to hamper fruitful innovation; pure bottom-up processes can lead to singular and costly re-inventing of the wheel solutions. Solid and balanced governance structures need to be established, excellent personnel from different stakeholder groups need to be engaged continually at all levels in order to achieve the essential balance. A thorough regular evaluation of all aspects by a neutral board of experts needs to be carried out based on transparent criteria and metrics. Opt-out options and contingency plans to continuously adjust the balance given changed circumstances should be incorporated.

Cost Effectiveness

Decisions towards cost-effective solutions regarding the complex service landscape in the emerging ecosystem of infrastructures requires a strong top-down component, since decisions are required that will stop certain activities that may have relevance for only a small group of researchers. Finding solutions is a non-linear process that requires communication skills across countries and disciplines. Social sciences and humanities often have the experts that can moderate such discussions.

Common Principles

Also, putting in place some common principles such as respecting financial constraints, spending effort on training young people, taking care of gender balance, respecting ethical boundaries, etc. requires a strong top-down influence that looks beyond the short-term results that can be achieved.

Appendix A - Legal Considerations

Principles

While sharing and openness, guiding principles of academic ethics, are also cornerstones of RI, legal rules applicable to services and materials shared within RI grant exclusive rights (monopoly) to authors, producers, publishers or research subjects. The conflict between these two approaches: openness on one hand and exclusivity on the other is a source of many obstacles to the sustainability of RI.
The most important legal frameworks relevant to RI are copyright, *sui generis* database right and personal data protection. In addition to that, the framework regulating the liability of service providers also has an impact on the functioning of RI.

Copyright law grants to authors of original works an exclusive right to copy their works, and to communicate them to the public. Even though raw data (for example statistical data or data coming from experiments) should be regarded as facts and therefore not qualify for copyright protection, other materials (texts, images) are copyrightable. In the European Union, copyright expires seventy years after the death of the author; therefore, in most cases research materials are subject to copyright protection. Moreover, copyright also protects original collections (even if their constitutive parts are unoriginal, e.g. collections of raw data (datasets)) and software. Thus, such material can only be shared within RI on a basis of a statutory copyright exception, or on a basis of a permission granted by the right holder (a license).

Directive 2001/29/EC allows Member States to introduce exceptions for non-commercial copyright research (art. 5.3 (a)). Unfortunately, while most (if not all) Member States did introduce such exceptions, most of them are in fact extremely narrow (s. 52a UrhG, art. L122-5.3°(e) CPI - compared to a broad exception in s. 29(1) CDPA) and their unclear wording makes them of no practical use for researchers.

In the recent decade - marked by the development of participative Web and public licensing schemes such as Creative Commons - the role of copyright licenses has increased. In practice, the vast majority of scientific resources are accessed and re-used via a licence rather than via a statutory exception. On one hand, a big part of these resources (e.g. Wikipedia entries) is licensed under more or less restrictive public licenses (Creative Commons, GNU GPL etc.); on the other hand 'bespoke' licensing agreements are negotiated directly with right holders (authors, publishers, developers). None of these solutions are perfect. 'Bespoke' licenses are costly and their negotiation process time-consuming; as the positions of researchers in the process is comparably weak, the final licenses are usually very restrictive. Note that in most EU jurisdictions contractual clauses may override statutory copyright exceptions, which means that licenses may even deprive researchers of the benefits - albeit limited - of statutory research exceptions. Public licenses, while being a great tool, may be victims of their own success: indeed, rightholders without proper legal assistance may choose inappropriate public licenses (e.g. a license with a « no derivatives » requirement for datasets, making it impossible to build on the data).

Finally, a plethora of existing contractual solutions makes interoperability between licenses an important, if not unsurmountable, obstacle on the road towards sustainable RI. In fact, materials licensed under two incompatible licenses cannot be « combined » together (without violating the conditions of one of the licenses), which greatly reduces the benefits of sharing them within RI. For example, Creative Commons licenses are generally incompatible with free software licenses; bespoke licenses are in practice rarely compatible with any public license; incompatibilities are possible even within one licensing scheme (e.g. « share-alike » and « no derivatives » requirements are generally regarded as incompatible).

Apart from the economic rights, EU copyright law protects also moral rights of authors, the most important of which is the right of attribution of authorship. Moral rights in many Member States are perpetual, and cannot be transferred or waived. While the attribution of authorship is a guiding principle of academic ethics, it may be an obstacle to some re-uses of scientific materials shared within RI. A big dataset (e.g. a language corpus) may be combined of thousands, if not millions, of texts, and the obligation to attribute every single one of them may lead to a phenomenon known as 'attribution stacking'. Some researchers, being aware of this, consciously decide to waive their moral rights via a waiver such as CCzero. In practice, this instrument, due to the fact that in most jurisdictions moral rights cannot be waived, raises serious questions regarding its legal validity and enforceability. The use of waivers may therefore have unintended consequences.

***Sui generis* database right** grants to database producers an exclusive right to prevent extraction and re-utilization of a substantial part of their databases. This right is independent of copyright (i.e. from originality of a database and its constitutive parts) and is instead focused on the investment in the production of the database. The right may therefore allow database producers to 'block' access to material that is unoriginal (e.g. phone numbers) or in the public domain (e.g. medieval poetry). It may therefore have a huge impact on the sustainability of RI. While the Directive allows Member States to provide for limitations on the exclusive rights of database producers (art. 9), in practice databases are accessed on a contractual basis (via, for instance, terms of service or a similar document). As in the case of copyright exceptions, contracts may override exceptions to the sui generis database right, making operations such as text and data mining impossible without violating the contract (or at least without signing a specific agreement concerning text and data mining, presumably for a (high) fee).

For a long time, there were no public licenses allowing licensing the *sui generis* database right. The situation has changed with the introduction - and, hopefully, generalization - of Open Data Commons and Creative Commons 4.0.

Personal data are defined in a very broad way as any data relating to an identified or identifiable person. In principle, personal data cannot be processed without consent of the data subject (i.e. the person to whom the data relates). The existing framework, due to the broad definition of both 'personal data' and 'processing', is an important obstacle to the creation of scientific material shared within IR. 'Personal data' may not only include 'typically' personal information such as one's name, age or address but also sensitive data, such as information about one's health, ethnic origin, political opinions or sex life, or even x-Ray photos which, with the development of technology, may serve to identify a person. Personal data protection is of course of crucial importance for RI in the field of

medicine, but also in human science (e.g. language material may also contain personal data). Obtaining the data subject's consent is not always possible and it slows down the development of science; anonymisation, on the other hand, may also be a very difficult and time-consuming task, given the progress in identification technology. Some Member States may provide (usually in limited cases) for research exceptions to personal data protection; the new Regulation will most likely contain such exceptions (albeit rather limited, especially in case of medical research).

The key problem is the « fragmentation » of **statutory exceptions** and their interpretation by courts in Member States, combined with lack of certainty regarding **applicable law** (particularly in copyright cases). In practice, researchers who relied on a broad research exception in one jurisdiction to create their material and who then shared it within a RI may be sued for violation of foreign law which does not provide for equally broad research exceptions (vide: French or German law regularly applied to Google by French or German courts on the basis of their services 'targeting' the French or German public).

In addition, the sustainability of RI is impossible without Internet Service Providers, in particular **hosting providers**. According to art. 14 of Directive 2001/31/EC, hosting providers are not liable for the content that they host, as long as they do not know it or remove it immediately after being informed of its illicit nature. In practice, hosting providers in RI often review the content that they host in order to make sure that it is interoperable with the whole RI, relevant and of sufficient quality. By doing this, however, they may involuntarily deprive themselves of the benefits of the liability limitation presented above; they may therefore be held liable e.g. for copyright infringement or unlawful processing of personal data.

Moreover, the newly revised **Public Sector Information Directive** may increase the availability of research data (at least those from publicly funded projects), which would facilitate the creation of RI. It is too early, however, to evaluate its impact. Finally, another aspect of the legal dimension is the **legal framework** that has been set up to establish RI as legal entities at European level. The ERICs are owned by the member states which ensures anchoring them in national roadmaps and thus importantly at political level. The first initiatives have followed the lengthy procedure and have now been established as ERICs. Yet it is too early to speak about experiences.

Subgroup membership

Name	Institution/ Position	Country	Contact
Peter Wittenburg	Max Planck Institute for Psycholinguistics, Senior Advisor **Subgroup Chair**	The Netherlands	peter.wittenburg@mpi.nl
Anton Anton	Technical University of Civil Engineering Timișoara, Professor	Romania	anton@utcb.ro
Lajos Balint	National Information Infrastructure Development Institute (NIIFI), Director of International Relations	Hungary	lajos.balint@niif.hu
Sabine Brünger-Weilandt	FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, President and CEO	Germany	sabine.bruenger-weilandt@fiz-karlsruhe.de
Janusz Bujnicki	International Institute of Molecular and Cell Biology, Warsaw, Professor	Poland	iamb@genesilico.pl
Sandra Collins	Digital Repository of Ireland, Director	Ireland	s.collins@ria.ie
Blanca Miranda Serrano	Andalousian Public Health System Biobank, General Coordinator	Spain	blanca.miranda@juntadeandalucia.es
Sanna Sorvari	Climate Research Unit, Finnish Meteorological Institute, Research Manager	Finland	sanna.sorvari@fmi.fi
Colin Wright	SANReN (South Africa National Research and Education Network), Chief Strategist & Manager	South Africa	colin.wright@wits.ac.za

Annex 3: Big Data (Data Deluge & Future Research Data Infrastructures)

Table of Contents

Executive Summary

Modern science (including humanities), is data-intensive, multidisciplinary, collaborative and global. Research environments are becoming more and more virtual. The conduct of modern science therefore runs in parallel with the development of the phenomenon known as 'Big Data'. In science, Big Data needs to be harnessed into efficient, interoperable and global Research Data Infrastructures (RDIs), based on standards and standardization. To this end, future RDIs must strive to reach three goals:

- *Data-intensiveness*: Relying on the appropriate technologies in order to keep up and suggest new avenues of investigation to researchers. This can be achieved by increasing efficiency of data management, curation, search, sharing, and transfer, as well as by managing the complexity of the analytical process. Moreover the involvement of society in the Challenges of H2020 mean that new RDIs must be developed to allow flourishing of public research here, providing the appropriate organizational and legal frameworks.
- *Interdisciplinarity and collaborativeness*: The RDIs must provide the mechanisms and incentives for sharing data and results of experiments (different level of interoperability and semantic enrichment) as well as to realize experiments by

combining resources (data and methods and results) belonging to different communities. This call for tools facilitating the governance of complex analytical process and for sophisticate search and retrieval tools supporting resource discovery.

- *Globality*: RDIs have the role of knowledge accelerator and must support different ways of contributing to the literature where access is crucial. Search technology and text mining play an important role, as well as linking scientific literature to data that participated to experiments.

The position paper envisages making the following recommendations:

1. Establishment of a Pan European Network of Research Data Infrastructure. Moving from a collection of RDIs to an ecosystem of RDIs, leveraging Big Data and the collective knowledge generated by citizens, emphasizing openness and ease of access, especially for young generations.
2. Adopt an appropriate normative scenario for personal data in research and novel form of conduct code for personal & social data
3. Foster the development of skilled specialist with a strong educational effort on a new generation of data scientist

Introduction

Technology is ubiquitous and very much part of public and private organizations and individuals. People and things, content and value chains are becoming increasingly interconnected. Smartphones, buildings, cities, vehicles and other environments and devices are filled with digital sensors, all of them creating evermore data. New high-throughput scientific instruments, telescopes, satellites, accelerators, supercomputers, sensor networks, and running simulations are generating massive amounts of data. Often referred to as a data deluge, massive datasets are revolutionizing the way research is carried out, which results in the emergence of a new fourth paradigm of science based on data-intensive computing (Hey, Tansley, & Tolle, 2009).

Big Data also has the potential to become the main enabler of reality mining by driving nearly every aspect of our modern society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences. The new availability of huge amounts of data, along with advanced tools of exploratory data analysis, data mining/machine learning, and data visualization, has produced an important change in the scientific methodologies as well as the way of cross-disciplinary scientific work, including the humanities

This paper describes the main issues around Big Data as they will play out in the coming years in the area of scientific research. It focuses on the technical challenges (A), on data management (B), on responsible access to data (C) and finally, on the future of Big Data and e-Infrastructure (D).

A. Technological challenges

In order to be able to exploit huge volumes of datasets, the research data infrastructure (RDI) needs to harness the accumulating data and knowledge produced by the communities of research, optimizing the data movement across scientific disciplines, enabling large increases in multi- and inter- disciplinary science while reducing duplication of effort and resources and integrating research data with published literature. In synthesis, modern science is: data intensive, multidisciplinary, open, global and participatory, and in order to empower the next generation of research(ers), RDIs will have to tackle the following technological challenges:

- Large-scale in data volume (Big Data)
- Wide-scale in data contextual diversity (Little data)
- Scientific data should be discoverable, understandable and assessable
- Scientific data and publications have to permeable to knowledge boundaries: syntactic, semantic and pragmatic boundaries (risk: semantic distortion)
- Scientific data as an essential element of the scientific communication
- Scientific communication should encourage modularity: it should allow for non-linear reading.

There is no proper definition of the term “Big Data”, which is also true for several of the terms such as eScience and cyber-infrastructure that have recently been invented, when describing some of the new challenges for science. We need to accept this as a positive since it leaves us space for the dynamics we obviously need to characterize the challenges and solutions. This trend can be characterized by a number of properties such as:

- Volume and complexity of data is increasing
- Velocity with which data is being created and characterized is changing
- Variety of data in all respects and the challenges of combining variety
- Veracity related to aspects such as trust in dealing with data, i.e. statistical significance or reliability in terms of re-use for further purposes.

An aspect frequently associated with Big Data is the potentiality of the Data Mining paradigm in finding patterns in large collections of different data sources with the help of mathematical methods. A major theme correlated with Big Data is to turn data streams into insights often without making too many assumptions. Therefore, algorithms such as machine learning or evolutionary algorithms are particularly relevant in finding patterns and extracting insights.

Despite measures to protect data where necessary, Open Data and easily accessible data will play an important role in creating the large collections across disciplines that are embraced by the term “Big Data”. Thus the term “Big Data” also stresses the close relationship with the aspects of managing and processing the data as well as their long-term preservation and availability.

A.1 Data Types

All types of data are of relevance to Big Data analytics. There is the old distinction between structured and unstructured data, where "structure" pre-supposes a schema in XML or in a database structure description that helps to interpret data. Other data types are unstructured textual data aggregated from the web which can widely be interpreted without additional structure and context descriptions. Another useful distinction is regular versus non-regular data: regular data is typically generated by sensors or software simulations. The amounts of data can be large as a consequence of the increasing time and spatial resolution of the underlying generation processes; however their regular structure, described by a very simple schema, allows easy interpretation. Certain processing will require detailed knowledge of context descriptions such as about filter characteristics. Non- regular data, as created by all types of pre-processing and derivations, mirror the richness of science, i.e. they are very heterogeneous and their correct interpretation requires contextual and provenance description encapsulated in metadata. In general, non-regular data is associated with the long tail of data bearing mostly condensed information, but nevertheless their amounts are huge and they will also be the object of Big Data analytics.

A special type of data is dynamic data (streams) in so far as this data is used for analytics, even while the data is changing. This is true, of course, in the case of aggregation of web information. But it is also true in case of sensor data where different time slots are filled in with delays of different sizes or where data is created by massive crowd sourcing, through the unpredictable response of the public. Two of the special challenges with this type of data are how to manage them efficiently and how to refer to the snapshots over time when the changes are occurring at unpredictable moments.

New types of massive data that are used for typical Big Data applications are for example: SMS messages, blogs etc. where each message is extremely short, but where the millions of users create Big Data. These are the data originating from the digital breadcrumbs of human activities, sensed as a by-product of ICT systems being used every day: desires, opinions and sentiments leave their traces in web pages and blogs; in the social media in which we participate; in the query logs of the search engines we use; in the tweets we send and receive; social connections leave their traces in the network of phone or email contacts, in the friendship links of social networking sites; movements leave their traces in the records of our mobile phone calls and in the GPS tracks of our on-board navigation system. In other words, these Big data offer new opportunities to observe and measure how our society intimately works, even though they have been sensed for other purposes: this makes the process of "making sense" particularly challenging.

A.2 Data storage and access

Big Data is certainly too large and complex to store and analyze by using traditional IT approaches such as file or database systems. These traditional approaches are gradually being replaced by virtualization technologies such as clouds, that offer just a simple hash tag to access objects, and a strong parallelization engine based on *map-reduce* kind of schemes to scale up data access. Thus storage and access technologies need to be

replaced to meet the developing requirements. But it is not just the storage and access aspects that need to be changed, we also need to structure our data space in so far as trusted repositories are concerned to store and manage the data. This needs to be done together with the metadata information that enables interpreting it and that offers the possibility of permanent access based on persistent identifiers.

Metadata information itself has to be based on worldwide standards and formats. Big Data will need to rely on automatic procedures that can be repeated frequently with slight modifications. In order to be successful and efficient this requires a "data fabric" like approach. Skilled data management thus becomes an issue and here the Research Data Alliance should be mentioned as a facilitating organization and structure to offer ways to improve our ways of dealing stepwise with data.

In many cases, it will make sense to have computational capacities close to the data store since Big Data necessitates processing large amounts of data. Infrastructures such as EUDAT, where the data centers also have the power to offer machine cycles in a distributed fashion have the potential to meet the requirements of the future. However, in a Science 2.0 framework difficult rights solutions must be implemented to simultaneously allow for the required flexibility and nevertheless maintaining a certain security level.

A.3 Big Data Analytics

Although analysis intuitions behind big data are pretty much the same as in small data, having bigger data consequently requires new methods and tools for solving new problems, or solving the old problems in a much better way. In Agrawal et al. (2012), authors identify five different processing phases in the use of Big Data a) Data Acquisition and Recording; b) Information Extraction and Cleaning; c) Data Integration, Aggregation, and Representation; d) Query Processing, Data Modeling, and Analysis; and e) Interpretation.
The following challenges may underlie many, and sometimes all, of these phases:

- *Lack of Semantics*. Sensed data are low-level and semantically poor because they expose the raw details of the measurements allowed by the ICT infrastructure that generates them. The big size of data does not always overcome semantic deficiency when modeling complex phenomena. This calls for semantic enrichment methods based on Machine Learning and Data mining aimed at capturing sense hidden in data. Scientific communication across disciplinary boundaries needs semantic enhancements in order to maintain the interpretative context and make the text/data intelligible to a broad audience composed of specialists in different scientific disciplines.

- *Heterogeneity and incompleteness* that calls for novel data collection, fusion and aggregation method able to combine vastly diverse data coming from different sources but speaking about the same phenomena. This goes well beyond database integration technology towards alignment, statistical matching and learning, entity resolution, uncertainty and fuzzy reasoning. In general it requires a novel dimension for data fusion.

- *Statistical significance*: the very nature of big data challenges the traditional notions of significance, quality and likelihood, because data expose part of the ground truth of a phenomenon sensed and observed at irregular times and with procedures that are alien to analytical goals. This calls for novel semi-supervised quality and significance checks, where the plausibility of data analysis w.r.t. what is known becomes a basis to project the veracity of new findings? Also, participatory crowd sourcing will become a standard way to leverage the intelligence of the crowd to ascertain the veracity of data.

- *Timeliness:* this refers to the cases where the analysis outcome becomes irrelevant unless delivered in a specific fragment of time. The challenge lies in designing methods where partial results can be computed in advance, so that only a small amount of incremental computation is needed when new data arrives. Indexing structures are typically the instrument employed to facilitate this goal when the searching criteria are specified. However, in the context of Big Data, new types of criteria may be specified or may even change dynamically, calling for new index structures. Designing such structures becomes particularly challenging when the data volume is growing rapidly and the queries have tight response time limits.

A.4 Governing the complexity of the collaborative scientific discovery

Research teams must be allowed to share data and results and to collaborate and learn over time and across geographic, organizational, and disciplinary boundaries. Moreover, data coming from different sources need to be integrated in order to obtain a richer yet coherent picture of the phenomena under study. Big data and the increasing availability of data tools and services and the intensive interactions between globally distributed research teams demand far more mediation in order to allow heterogeneous parties to communicate and interoperate. The analysis of big data is an interdisciplinary research area, which requires that experts in different fields cooperate to harvest the potential of big data. A comprehensive big data network architecture must be established to help scientists and engineers in various fields access different kinds of data and fully utilize their expertise, so enabling the requisite cooperation to complete the analytical objectives.

The need to combine data from different sources is one of the most prominent characteristics of Big Data. This amounts to a multimodal type of analysis where the individual components can largely differ, not only with respect to their time and spatial resolution, but also with respect to their nature. In some fields of study, algorithms need to understand all the differences and need to rely on the availability of schemas, contextual information, semantic categorizations, etc. Another issue is how to integrate data sets assuming that they are generated and stored at different locations.

Data sharing: The modern Science paradigm depends largely on the ability to reconcile information from multiple sources and to make geographically and institutionally separated research teams interoperable. Several levels of interoperability must be considered, for example:

I. Data Exchangeability (exchange of meaningful information)
   - Syntactic exchangeability
   - Semantic exchangeability
II. Functional Compatibility (Compositionality/Replaceability)
III. Logical consistency
IV. Policy Compatibility (legal compatibility)
V. Data Usability

Data Reduction: In the case of big data even during the creation process the challenges for data reduction will increase not only per institution but also across institutions since more of them will be confronted with big data. Even for the humanities, sensing massive crowd sourcing data will need the application of smart preprocessing to manage the data volumes. Real-time event analytics and outlier detection are applications that focus on reducing the incoming data. Parallelization techniques will be required as well to keep the costs low. Importantly: It is still an open question on who will be able/ has the right to decide on preservation or cancellation of data?

Transferring Big Data: this a further challenge where we clearly see that current network bandwidths do not scale with the amounts of data. Here it is worth to mention newer networking NREN paradigms such as: BoD (Bandwidth on Demand) and data specific DMZs (data de-militarised zones). Currently NREN move towards wide-ranging 100 Gbps and even higher backbones. A major bottleneck remains the disparity between NREN and local institutions infrastructure, networking equipment and expertise. In addition to extending the network capacities new methods such as distributed analytics need to be applied where analytic processes are started at different locations working on queries that have been translated to the specific data types to be expected.

Mastering the interdisciplinary discovery process: Scientific workflow is a key component in a research data infrastructure as it orchestrates e-science services so that they co-operate to efficiently implement a scientific application. A workflow is a precise description of a scientific procedure – a multi-step process to coordinate multiple tasks acting somewhat like a sophisticated script. Services supporting the scientific workflows liberate scientists from the drudgery of routine data processing so they can concentrate on scientific discovery. Such processes shoulder the burden of routine tasks, they represent the computational protocols needed to undertake data-centric science and they open up the use of processes and data resources to a much wider group of scientists and scientific application developers.

New proposals in literature (Ceri et al., 2013) advocate the need for a new approach to data analysis in support of *computational inter-disciplinarily.* These approaches are based on mega-modeling, which is a holistic data and model management system for the acquisition, composition, integration, management, querying and mining of data and models, capable of mastering the co-evolution of data and models and of supporting the creation of what-if analyses, predictive analytics and scenario explorations. Mega-modeling provides a comprehensive theory and technology of data driven model construction, model search, model fitness evaluation, model composition, model reuse and model evolution.

Linking scientific literature & data space: another challenge stems from research teams relying on a large number of diverse and interrelated datasets but having no way of

managing their scientific data spaces in a principled fashion. Linking data will allow the sharing of scientific data on a global scale and interconnect data between different scientific sources. Linking data refers to the capability of publishing data on a data space in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets and can in turn be linked to from external data sets.

Modern science requires integrated support of the whole research data life cycle. Scientists need to publish their raw data sets, experimental details, analytical methods and visualizations, in addition to traditional scholarly publications. As both datasets and articles are made available in electronic form, the distinction between them will disappear. It is the task of the linking technology to support the next step, namely their integration. A generalization of the linking data concept leads to the creation of Linked Scientific Data Spaces (of disciplinary or interdisciplinary scope). OpenAIRE[1] and Eudat[2] are going in this direction, but also interesting proposals are coming out of other FP7 projects, for example Qlective[3] that introduces the concept of scientific data space as a knowledge accelerator, i.e. a self-organizing socially intelligent information system.

B. Data Preservation and archiving

The term Big Data seems to indicate that the only challenge is the sheer size of information, but one must also include the capture, storage, long-term preservation, curation, search, analyzing, retrieval, sharing, transfer, analysis and visualization of data. Data management refers to every step necessary to make sure that the data can be accessed and used over time in a sustainable way. Managing data goes beyond the technical challenge; it is very much an organizational one and refers to data governance; data origination; data integration; data quality, and information management. Managing research data can be challenging and requires attention to be paid to data curation, a function which the Data Curation Centre (DCC), in the United Kingdom, describes as “maintaining, preserving and adding value to digital research data … [Data] curation enhances the long-term value of existing data by making it available for further high quality research.”

As G. Little (2012) points out, ‘making research data available and preserving it means that, aside from allowing students and faculty to conduct their research, published experiments and tests can be recreated to verify and confirm findings and results; existing data on one topic, including data gathered by national, state, or local statistical agencies can be re-used now and in the future by researchers and policy makers in other areas’.

In order to elaborate norms for data management as well as useful tools for this purpose, it may be useful to make an inventory of the ongoing management practices and initiatives across Europe and elsewhere to filter out the best practices among them. Data Management entails the proper preservation and archiving of scientific data. Not only storage and access aspects need to be adapted, but the data space needs to be structured insofar as trusted repositories need to store and manage the data as well

[1] https://www.openaire.eu/
[2] www.eudat.eu/
[3] www.qlectives.eu

as the metadata information that enables interpretation and offers the possibility of permanent access based on persistent identifiers. Metadata information has to be based on worldwide accepted standards, and the metadata themselves have to refer to content on the one hand and to technology on the other hand. Skilled data management thus becomes an issue and here efforts under the Research Data Alliance (RDA) should be mentioned as mechanisms to improve our ways of dealing stepwise with data.

The future will adjudicate how much further software development methods will have to adapt so as to move away from traditional engineering models and to better meet the new requirements. Certainly we need to move much closer to automatic workflows that can be executed in a variety of ways and can flexibly react to changed settings and wishes. Of course, these workflow scripts must not simply be documented but rather document the steps of the workflows by for example automatically adding detailed provenance records to the metadata after each processing step. A common problem for automatic workflows is how to deal with errors introduced at a certain step. Since manual interventions will increasingly be problematic, error treatment will be a serious challenge.

B.1 Policies, mandates, plans for data management

RDIs will function in a seamless manner provided that their implementation and enforcement follow clear and standardized guidelines. Horizon2020 projects must already produce a Data Management Plan (DMP) (Guidelines on Data Management in Horizon 2020), which describes the data management life cycle for all data sets that will be collected, processed or generated by the research project. It is a document outlining how research data will be handled during a research project and even after the project is completed, describing what data will be collected, processed or generated and following what methodology and standards, whether and how this data will be shared and/or made open, and how it will be curated and preserved. The DMP is not a fixed document; it evolves and gains more precision and substance during the lifespan of the project.

Strengthening data management and release practices is imperative. To ensure that research data are managed and maintained throughout their life cycle, institutions must adopt effective data asset portfolio management approaches. To this end, data management policies and practices will need to be developed for funding proposals and as grant conditions.

C. Responsible access to data

Using Big Data represents a significant economic opportunity for Europe. Prolific use of Big Data would add tens of billions of Euros in value to the EU's aggregate GDP. This would result chiefly from higher productivity amongst researchers and from the effects ('externalities') of increased levels of research. At present, the use of Big Data and mining tools by researchers in Europe appears to be lower and probably significantly lower, than is the case in the United States and some other countries in the Americas and Asia. This probably reflects, among other factors, disadvantages created by the

European legal framework with regard to the use of Big Data. Providing researchers with on-going, reliable access to high quality content for research is said to involve a significant investment in validation, correction and refinements to content, plus investment in systems to hold that content in a secure manner.

C.1 Intellectual Property Rights Issues

Big Data offers opportunities that are certain to become more important as researchers acquire the skills and the technology to address and investigate datasets of increasing size, complexity and diversity in all media: text, numbers, images, audio files and in any other form. Big Data discussions provoke complex IPR issues compounded specifically by:

- The inherent copyright and/or database rights which might exist in original texts
- The levels of adaptation and processing required to create the derived data
- The intended use of the outcomes.

Legal research has shown that whilst scientific articles, monographs and reports almost always attract copyright protection, research data itself is not protected by copyright law and seldom by other legal norms (Guibault and Wiebe, 2013). In Europe databases and their structures are protected under a *sui generis* database right[4] provided that sufficient investment has been undertaken in establishing the database. This right is unique in the world and is conferred only on European residents or European based entities. Whether scientific databases are protected by this specific right is a question of fact which needs to be established on a case-by-case basis. If protected, the use of the database is subject in Europe to the prior authorization from the rights owner. (Massive) extraction of data for the purpose of analysis is not covered by specific exception of the Database Directive. Moreover the limitation allowing scientific use of databases is optional and, as a result, has not been fully harmonized across the Member States.
The European legislator must re-evaluate the EU's legal framework with regard to copyright and database protection, in order to support the international competitiveness of Europe's research base. There is a serious risk that Europe's relative competitive position as a research site for the exploitation of 'Big Data' will deteriorate further, if steps are not taken to address the issues discussed in this report. The results of this might well include a loss of talent and a loss of investment to more favorable research locations.

C.2 Licensing Issues

On 17 July 2012 the European Commission published its Communication to the European Parliament and the Council entitled "Towards better access to scientific information: Boosting the benefits of public investments in research". In light of this Communication, researchers using Big Data should not face restrictions. This position is now reflected in the EU's Horizon 2020 strategic research framework. In the model grant agreement for Horizon 2020 the Commission states that the beneficiaries must:

[4] Directive 96/9/EC of the European Parliament and Council of 11 March 1996 on the legal protection of databases, OJ L 77, 27.3.1996, p. 20–28.

- deposit in a research data repository and take measures to make it possible for third parties to access, mine, exploit, reproduce and disseminate — free of charge for any user — the following:
  I. the data, including associated metadata, needed to validate the results presented in scientific publications as soon as possible;
  II. other data, including associated metadata, as specified and within the deadlines laid down in the data management plan (see Annex I);

OA principles entail more than just granting access to research data free of charge[5]. The core of OA principles demands that research data be available for any type of re-use by any user. As intellectual property rights (IPRs) may attach to the scientific output, it is important to license the data in a clear manner so that users immediately know what they can and cannot do with the data. Especially important for Big Data, licensing research output under OA conditions assumes that the person or entity that applies the license owns the IP rights on such scientific output. Ownership issues may be at stake in cases of public/private partnerships. Efforts should also be made to encourage rights owners to use compatible licenses to avoid imposing unnecessary transaction costs on researchers who would need to assess whether the combined use of different databases is allowed under the different licensing terms. Efficient and transparent RDIs will demand that databases be correctly labeled to allow optimal re-use possibilities.

C.3 Privacy

Discussions on privacy issues and the role of data mining, profiling and data warehousing date back to the nineties. However, as an ever-larger amount of data is being digitized, shared across organizational boundaries and re-used for secondary purposes, privacy and data protection have become even more pressing policy issues (Mckinsey Global Institute, 2011). The proliferation of ubiquitous computing ('Internet of Things', ambient intelligence…) in combination with the growing possibilities for the linking and analysis of data creates the additional challenge that even data which would, taken alone, not raise privacy concerns, may expose wide-ranging impressions of the person concerned, including very sensitive personal data (Cas, 2011). Sets of correlated data that could be considered insignificant or even trivial can provide intimate knowledge about, for example, lifestyle or health risk, if data mining is applied[6].

[5] Legally binding definitions of 'open access' and 'access' in this context do not exist, but authoritative definitions of open access can be found in key political declarations on this subject. These definitions describe open access as including not only basic elements such as the right to read, download and print, but also the right to copy, distribute, search, link, crawl, and mine.

[6] M. Hildebrandt, 'Profiling and the identity of the European citizen.' in M. Hildebrandt and S. Gutwirth (eds.), *Profiling the European Citizen: Cross-Disciplinary Perspectives,* Dordrecht: Springer, 2008, p.304. The aggregation and analysis of digital clinical data from medical records, for instance, may reveal information that help payers and regulators to improve clinical decision making, but may also hold risks for patient privacy

Current EU rules on data protection provide a high level of cross-sectorial protection for the privacy of individuals, imposing strict limits on the collection and use of personal data[7]. Generally speaking, researchers who in the context of their projects wish to process personal data must comply with the rules on data protection. European research project consortia involved in the mining of information on social networking sites have highlighted the difficulties experienced in seeking the consent of the data subjects. The requirement for obtaining user consent (and the administrative burden surrounding it), as well as difficulties relating to the allocation of responsibilities and the principal prohibition of the processing of certain categories of (so-called "sensitive") personal data, hinder the conduct of research and the development of innovative and competing tools involving user data. In other words, researchers have a responsibility to deal with data ethically, even when data are publicly accessible, especially when it is impossible to obtain consent by each person represented in the data as is often the case with Big Data. (van den Hoven et al, 2012) This means "both accountability to the field of research, and accountability to the research subjects" (Boyd et al., 2011).

But the current regulatory framework pertaining to data protection is clearly not geared towards the reality of Big Data, and the need of data science research to share and integrate data for statistical and scientific purpose. Future RDIs need to be able to rely on a clear and effective legal framework that allows scientific research to take place whilst taking account of the interest of the data subjects. Moreover, future RDIs should provide the ecosystem where new grounded practices are experienced and consolidated under special provision provided under special regulation at European level.

C.4 Scientific integrity

Data citation refers to the practice of providing a reference to data in the same way as researchers routinely provide a bibliographic reference to printed resources. This practice corresponds roughly to the obligation under certain open content licences, such as the Creative Commons Licences or the Open Data Licence administered by the Open Knowledge Foundation, to give attribution to the author of the work or the maker of the database. Unfortunately, no universal standards exist for citing quantitative data. Neither Creative Commons nor the Open Knowledge Foundation has issued guidelines in this regard. In fact, the norms relating to the granting of attribution or the citation of quantitative data are often dependent on the customs in force in each discipline of science. Clear guidelines regarding the citation of research data need to be established to ensure that scientists can take advantage of future RDIs in accordance with the norms of scientific integrity in force in their own discipline.

[7] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995, p. 31–50; and see: Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), 25.1.2012, COM(2012) 11 final, 2012/0011 (COD).

D. Future of Big Data

The Big Challenge for EU will be how to multiply the value of the many projects and transforming Big Data into a consumable good for the coming society, individual citizens and households, businesses and public bodies and services. RDIs will play an important role in this. We are in the Era of data-intensive science. The future RDIs have to be equipped with the appropriate Big Data technologies in order to keep up with new constraints:

- efficiency of data management (noSQL paradigms and cloud computing play important role here) and curation, search, sharing, transfer;
- managing the complexity of the analytical process is a key issue (much research and innovations in analysis and visualization of data are still ongoing, Big Data Visual Analytics is crucial to facilitate the data exploration).

There is a completely new contribution to science from Big data produced by ICT services: relationship with society. This is orthogonal to many disciplines, and the use of such data will increase with the Big Societal Challenges of H2020. The new RDIs panorama has to be ready to allow flourishing of public research here, providing the appropriate organizational and legal frameworks

1. technological big issues here is complexity of the "making sense" process -- semantics and veracity that call for advances in analytics
2. participation/crowdsourcing for validation are crucial here
3. privacy enhancing technologies and conduct code are crucial here
4. special regulatory context for research using personal data (same as biological human tissues)

Modern science is interdisciplinary and collaborative. The RDIs have to provide the mechanisms

1. for sharing data and results of experiments (different level of interoperability and semantic enrichment)
2. to realize experiments by combining resources (data and methods and results) belonging to different communities. This call for tools facilitating the governance of complex analytical process in a workflow style or mega-modeling. This call also for sophisticated search that supports resource discovery.

Modern science is global (share the findings to the scientific community). The RDIs of global science has the role of knowledge accelerator and must support different ways of contributing to the literature:

1. reachability is crucial (99% of published papers have never been accessed). Here search and retrieval technology and text mining play an important role.
2. linking to data that participated to experiments (OpenAir , EuDat are first step in this directions, or Living Science from Qlective.eu

Europe will be ready in (2016-2020) to create a Pan European Research & Innovation Infrastructure, a sort of Planetary Nervous System in the form of a Network of Infrastructures, which provide Big Data from different sources (including open data and participatory sensing) together with analytical skills and technologies, in order to boost innovation of businesses and public administrations, creativity and self-awareness of

citizens, research of multi-disciplinary scientists, wisdom of policy makers and managers, education of students. The Pan European Network of RDI will have the very ambitious goal of providing the ground not only for research, but also for innovation and creativity at industrial and societal level. It will be a sort of wind tunnel for new discoveries, services, products and societal innovation.

The idea will be to move on from a collection of RDIs to an ecosystem of RDIs, leveraging Big Data and the collective knowledge generated by citizens, emphasizing openness and ease of access, especially for young generations.

This vision was in the e-IRG 2012 roadmap as well as in the report from the Reflection group on e-IRG for 2013. They both foresaw, for the evolution of the current e-Infrastructure towards Horizon 2020, a common data infrastructure integrating a set of coherent data services exposed to users by means of an interoperable set of underlying e-Infrastructures. This Collaborative Data Infrastructure is a key element towards enabling user communities to get on with the business of science, while the generic data services they need are provided by various actors.

References

1. Agrawal, D., et al, White Paper: Challenges and Opportunities with Big Data, Feb 2012. (url: http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf)
2. Barabási Albert-László, Chaoming Song & Dashun Wang Affiliations Corresponding author Nature 491, 40 (01 November 2012)
3. Barabási, Song, Wang, Handful of papers dominates citations Correspondence in Nature 491, 40 (01 November 2012)
4. Boyd D., Crawford K., Symposium on the Dynamics of the Internet and Society, September 2011 (available at: http://ssrn.com/abstract=1926431) (2011)
5. Cas J., 'Ubiquitous Computing, Privacy and Data Protection: Options and Limitations to Reconcile the Unprecedented Contradictions', in S. Gutwirth et al. (eds.), *Computers, Privacy and Data Protection: an Element of Choice*, Springer, 2011, p.152
6. Ceri, S. et al.: Towards mega-modeling: a walk through data analysis experiences. SIGMOD Record 42(3): 19-27 (2013)
7. e-IRG Reflection Group White Paper 2013, www.e-irg.eu/
8. G8+5 White Paper Draft v2.0: 5 Principles for an Open Data Infrastructure, March 15, 2013
9. Giannotti F. et al, A planetary nervous system for social mining and collective awareness. Eur. Phys. J. Special Topics 214, 49–75 (2012)
10. Graham M., Shelton T.: Geography and the Future of Big Data, Big Data and the future of Geography in Discourses in Human Geography, Dialogues in Human Geography 3(3) 255–261, (2013)
11. Guibault L. and Wiebe A. (eds.), *Safe to be open - Study on the protection of research data and recommendations for access and usage*, Göttingen University Press, Göttingen, 2013, p. 33-34.
12. Guibault L., 'Licensing Research Data Under Open Access Conditions under European Law' in Beldiman (ed.), *Information and Knowledge: 21st Century Challenges in Intellectual Property and Knowledge Governance*, Cheltenham, Edward Elgar, 2013, pp. 63-92
13. Guidelines on Data Management in Horizon2020, 16 December 2013, available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h202
14. Hester J.R. , Closing the data gap: Creating an open data environment, *Radiation Physics and Chemistry*, Volume 95, February 2014, Pages 59-60.
15. Hey T., Tansley S., Tolle K., The Fourth Paradigm: Data-Intensive Scientific Discovery, WA: Microsoft Research (2009), http://research.microsoft.com/en- us/collaboration/fourthparadigm/

16. Krishnan K. , **'**Chapter 12 - Information Management and Life Cycle for Big Data', in *Data Warehousing in the Age of Big Data*, 2013, Pages 241-250.
17. Little Geoffrey, Managing the Data Deluge, *The Journal of Academic Librarianship*, Volume 38, Issue 5, September 2012, Pages 263-264.
18. McKinsey Global Institute (2011). Big data: The next frontier for innovation, competition, and productivity, at p.107; http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.
19. McKnight W., 'Chapter Seven - Master Data Management: One Chapter Here, but Ramifications Everywhere' *Information Management*, 2014, Pages 67-77.
20. NEM (Networked & Electronic Media): Big and Open Data Position Paper http://nem-initiative.org/wp-content/uploads/2013/11/NEM-PP-016.pdf
21. Thanos C. A vision for global research data infrastructures. In, Data Science Journal, vol. 12 pp. 71 - 90.
22. UN Global Pulse, White Paper: Big Data for Development – Challenges and Opportunities, May 2012, http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf
23. van den Hoven J., Helbing D., Pedreschi D., Domingo-Ferrer J., Giannotti F. andChristen M. FuturICT – The road towards ethical ICT. Eur. Phys. J. Special Topics 214, 153-181 (2012).

Appendix 1: Useful Terminology

Data ownership refers to both the possession of and responsibility for information. Ownership implies power as well as control. The control of information includes not just the ability to access, create, modify, package, derive benefit from, sell or remove data, but also the right to assign these access privileges to others. According to Scofield (1998), the term 'ownership' could be replaced with 'stewardship', "because it implies a broader responsibility where the user must consider the consequences of making changes over 'his' data".

Data stewardship is the management and oversight of an organization's data assets to provide business users with high quality data that are easily accessible in a consistent manner. It implies a broader responsibility than "data ownership" where the user must consider the consequences of making changes over 'his' data". It implies also the identification of strategies related to the convenience of maintaining the data or re-executing the simulation.

Data Privacy refers to the existence of possible issue related to privacy or confidentiality in relation to the data used by the community.

Data handling is the process of ensuring that research data is stored, archived or disposed of in a safe and secure manner during and after the conclusion of a research project. This includes the development of policies and procedures to manage data handled electronically as well as through non-electronic means.

Data integrity is important in ensuring the integrity of research data since it addresses concerns related to confidentially, security, and preservation/retention of research data. Proper planning for data handling can also result in efficient and economical storage, retrieval, and disposal of data. In the case of data handled electronically, data integrity is a primary concern to ensure that recorded data is not altered, erased, lost or accessed by unauthorized users.

Data Integration. In data management, data Integration is the problem of combining data residing at different sources, and providing the user with a unified view of these data. Another instance of data integration concerns the management of data acquired with different sensor types, where an emerging need is how to integrate different media (still images, video, 3D models, text, audio) to build a single, integrated representation of the phenomena under study.

Data Linking. A Digital e-Infrastructure may not act as a data integration system when semantic integration is not possible or practical; in this case, the infrastructure follows a co- existence approach, providing linking data. Linking data refers to the capability of publishing data on a data space in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from external data sets.

Data preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. This broad definition of data preservation refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change.

Long-term preservation can be defined as the ability to provide continued access to digital materials, or at least to the information contained in them, indefinitely.

Data curation refers to the active management of data through its life cycle of interest and usefulness to a designated community. Data curation activities enable data discovery and retrieval, maintain its quality, add value, and provide for re-use over time. As such, it includes all processes in the organization that involve data management.

That is, pre-ingest initiatives; ingest functions; archival storage and preservation; and disseminating and providing access to data for its designated community.

Data sharing refers to the policies related to the access to data.

Traceability is the ability to verify the history, location, or application of an item by means of documented recorded identification. End-to-end control and traceability of access to the data, controls of access history and actions performed by users.

Data protection refers to the set of techniques such as file locking and record locking, database shadowing, disk mirroring, to ensure the availability and integrity of the data.

Data Security means protecting data from destructive forces and the unwanted actions of unauthorized users.

A procurement policy is simply the set of rules and regulations that are put in place to govern the process of acquiring goods and services needed by an organization to function efficiently. In this case it refers to rules or constrains in the acquisition of data equipment or infrastructures.

Data Storage is a data conservation policy which defines dataset quality and quantity constraints for an experiment to be reproduced with that dataset.

Legal issues refers to any regional, national or EC laws regulating the access to data, the sharing of data, the preservation, security and privacy policies.

Commitment constraints refer to any commitment or constraint derived from European, national, or community projects or initiatives, pre-existent, that has to be respected or fulfilled and has or may have an impact on data management matters at large.

Subgroup membership

Name	Institution/ Position	Country	Contact
Fosca Giannotti	ISTI CNR, Director of Research **Subgroup co-chair**	Italy	Fosca.giannotti@isti.cnr.it
Lucie Guibault	Institute for Information Law, University of Amsterdam, Associate Professor **Subgroup co-chair**	The Netherlands	L.Guibault@uva.nl
Janusz Bujnicki	International Institute of Molecular and Cell Biology, Warsaw, Professor	Poland	iamb@genesilico.pl
Sabine Brünger-Weilandt	FIZ Karlsruhe - Leibniz Institute for Information Infrastructure, Karlsruhe, President and CEO	Germany	sabine.bruenger-weilandt@fiz-karlsruhe.de
Frederic Hemmer	CERN, IT Department Head	Switzerland	Frederic.Hemmer@cern.ch
Blanca Miranda Serrano	Andalousian Public Health System Biobank, General Coordinator	Spain	blanca.miranda@juntadeandalucia.es
Maria Teresa Ponce de Leao	National Laboratory of Energy and Geology (LNEG), President	Portugal	mleao@fe.up.pt
Sanna Sorvari	Climate Research Unit, Finnish Meteorological Institute, Research Manager	Finland	sanna.sorvari@fmi.fi
Peter Wittenburg	Max Planck Institute for Psycholinguistics, Senior Advisor	The Netherlands	peter.wittenburg@mpi.nl
Colin Wright	SANReN (South Africa National Research and Education Network), Chief Strategist & Manager	South Africa	colin.wright@wits.ac.za

Annex 4: Evaluation

Table of Contents

1. Introduction

To accompany the launching of next research funding cycle of the European Union (H2020) spanning from 2014 to 2020, EC has set up a series of Expert Advisory Groups (EAG) to advise her on the implementation and the strategy of the various components of the H2020 programme, as successor of the former Framework Programmes for Research and Technology (FP5, FP6, FP7).

Among these components, Research Infrastructures (RI) play an increasingly important role since early 2000's as a building block of the European Research Area, in order i) to better organize research at the European scale, ii) to make it much more cost efficient, iii) to enhance excellence in the global competition and iv) to boost their impact on member states economies.

For the 2014-2020 period, the foreseen funding of EU will be around 2 500 M€, out of the 24 441 M€ allocated to H2020 Priority 1: Excellent Science. To advise the Commission in implementing such massive funding, a Research Infrastructure EAG group has been formed. The group will provide EC with recommendations for various relevant topics for RI. During its first plenary meeting on 23rd October 2013, the group selected four topics deserving immediate contribution from dedicated subgroups:

1. RI and innovation
2. Cooperation and harmonisation: industrial engagement and socioeconomic aspects
3. Sustainability issues
4. Data

At the time of the membership definition of the subgroups, it was agreed to merge #1 and #2 (innovation and cooperation with socio-economic world) and to add an extra subgroup about ex-post evaluation.

The final subgroup configuration is therefore as:

1. Innovation and cooperation
2. Sustainability
3. Data
4. Evaluation

The four subgroups were expected to provide each a focused position paper with recommendations before the end of 2014, in such a way that outcome can be taken into account by the EC for the preparation of the next periods of the work programme of H2020 (i.e. 2016-2017 and 2018-2020)

2. Context

RI are a key element for structuring the European Research Area, because i) they contribute to improve the efficiency of large scale public funding beyond the capacity of a single member state, ii) they guarantee the excellence of research performed with them, iii) they foster collaboration between European researchers (including in terms of mobility), sharing common instruments and iv) they support the position of European research in the competitive global scene.

The actual landscape of European RI is framed by the action of the European Strategy Forum for Research Infrastructures (ESFRI), complemented by existing or new infrastructures built and funded at the national level, by the EU member states. This framework is supported by the ESFRI roadmap and by national roadmaps published by the large majority of EU member states.

The ESFRI roadmap is an on-going process. First published in 2006, with 35 projects, it was updated in 2008 bringing the number of RI of pan-European relevance to 44. The latest update focusing on projects dealing with energy, food and biology was published in December 2010. Having identified 48 projects of new RI (or major upgrade on existing ones) so far, ESFRI is now focusing on their implementation for the next few years. The next update of the roadmap will be carried out in 2015.

The ESFRI RI have been extensively discussed and evaluated before being included in the roadmap. To perform this evaluation, which is also closely linked to the decision of countries to participate or not, to the investment and to the operating costs of these pan-European entities, a deep analysis was necessary. ESFRI has set up a dedicated working group about evaluation, which did focused on the ex-ante evaluation of the projects.

This dedicated working group was set up by ESFRI in order to clarify the evaluation needs and associated methodologies. The "Evaluation Report[1]" provides a clear synthesis of the evaluation process for RI. However, it is related strictly to the ex-ante phase, as it aimed to provide rules, advises

[1] http://ec.europa.eu/research/infrastructures/pdf/esfri_evaluation_report_2011.pdf#view=fit&pagemode=none

and guidelines prior to the decision of creating a new RI and eventually its inclusion in the ESFRI roadmap, together with the possible prioritisation process.

The ex-ante evaluation has therefore been widely described and covered by the ESFRI working group. The present paper is rather focused on the ex-post phase of evaluation, as there is a need for monitoring the RI during their active life, both to justify their funding, monitor the science excellence and eventually prepare the decision to stop the RI, including the decommissioning phase.

In summary, the ESFRI evaluation report states that :

- ***ex-ante* evaluation is necessary for a robust decision-making process leading to the setting up of a new infrastructure, and for major upgrades or reorientation of existing RI.**
- ***ex-post* evaluation is mostly based on facts and results. It is used to demonstrate the quality of the research output and achievements, to account for the resources invested and to monitor value for money and cost effectiveness, including appropriate management of the RI.**

3. Role and benefit of Evaluation

Objectives of evaluation of RI are quite generic and can be grouped into 3 sets:

1. Scientific and technological excellence and impact
2. Socio-economic impact and competitiveness
3. Governance and financial management

These objectives are common with the EC evaluation framework for FP7 projects, and should be applicable also for evaluations at the national level.

Furthermore, applying these framework objectives, it provides also link with other position papers expected from H2020-RI-EAG:

1. Sustainability: The key aspect for the success of RI is their sustainability. This is obviously depending on the funding, for which the justification will heavily rely on a positive evaluation (objective #1)
2. Innovation and cooperation: Outcome of evaluation will also provide information and arguments for these aspects (objective #2)

Position paper about “big data” cannot be linked to specific objectives of evaluation, as it emphasizes the transverse features of the e-infrastructures. It brings at the forefront their role, whatever the scientific disciplines are. However, e-infrastructures, which are likely to be concerned by “big data” issues, should be also submitted to evaluation.

There is no doubt that evaluation is needed for RI. Most European countries have already a longstanding tradition in evaluating scientific activities, like it is already implemented for research organisations, laboratories and higher education institutions[2]. ESFRI has already established a very

[2] See EQAR (www.eqar.eu) and EQNA (www.eqna.eu)

strong prototype example with the development of the pan-European RI roadmap. At the national level, there are many more RI, depending on national authorities, typically Ministry of Research or Research Council. A recent survey made by the EC produced an exhaustive inventory of all RI of pan-European interest[3]. Most of these infrastructures are subject to monitoring and evaluation by their own funding authorities. Many of them, but not all, are evaluated according to a common framework, through their participation of EC funded projects (FP7). However, there is no shared framework of RI evaluation between all EU member states.

4. Evaluation process

Ex-post evaluation at the national level has been scarcely treated in the ESFRI report. To assess this, we can summarize the situation for those countries, which have documented the report:

- Finland, Italy do not mention ex-post evaluation at all
- Spain, France, Hungary, Romania, Sweden and UK distinguish ex-post evaluation, with ad-hoc process, usually under the responsibility of the funding organisation (ministry, Research Council, etc.).

We could expect that since this 2011 report, the situation has significantly improved and that more information is available from the UE28, nevertheless there are still significant room for improvement to harmonise the evaluation process.

The evaluation for pan-European or global RI is not a trivial action as it is complicated by the diversity of membership and interests. For large RI organisations (say CERN, ESO, EBI/EMBL, or other EIROFORUM members), there is no doubt about the quality of scientific activities, the socio-economic impact or the management quality. These organisations have de facto, their own evaluation system always based on international expert committees. The need for evaluation is not disputed, as the process is recognised and accepted by all member countries. Decision for one country to withdraw from such large international organisation is not based on scientific quality, but rather on political context or eventually for budgetary difficult situation.

Typical topics, on which ex-post evaluation does matter, but final decision is likely to be taken at the political level: are: Neutron sources in Europe, particle collider beyond LHC, radio-astronomy – SKA, large telescopes – eELT, Free electron lasers – XFEL, etc.

However, the recent creation of ERIC, the new legal instrument for pan-European RI is going to extend significantly the RI landscape, especially in domains where the scientific communities are more loosely organised than high energy physics or astronomy. As these are set up under the auspices of the EC, there is certainly a stronger need for an harmonised system of ex-post evaluation to ease the monitoring of these new entities, up to their termination, whenever this may happen, including the decommissioning phase.

[3] http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=mapri

Having in mind a truly European Research Area, in which a large variety of RI exists, common standards, criteria and indicators have to be defined to be used by the members of the ERA with general validity, i.e. fully applicable also when evaluating RI owned by a single member state or institution.

Within ESFRI, there has been a full recognition that the ex-post evaluation framework has a significant overlap with the ex-ante evaluation, in terms of criteria. However the indicators needed to document these criteria, are still to agree on between all EU member states and associated countries. Development of a common set of indicators is required.

5. Evaluation outcome

The outcome of ex-post evaluation allows quality control of research and technical activities and checks compliance with national research policy. Some countries expect also answers about how well strategic long-term goals of their research policy are followed, and how it is coordinated with the European strategy (for instance for neutron sources, synchrotron radiation or neutrino detection) and possibly with RI roadmaps

An important aspect of the evaluation relates also to the compliance to European strategy related to the open access (infrastructures and research data), for which the EC has demonstrated a strong will to improve the benefit to the scientific communities (through the various funding instruments of FP7, and now within H2020).

6. Recommendations

1. Identify all independent evaluation organisations at national level and achieve a joint/common activity at the European level for sharing policies and best practices, and possibly joint actions between these organisations.
2. Establish a common ex-post evaluation framework (criteria) for RI, both belonging to ESFRI or national roadmaps. These frameworks, inspired of the ex-ante evaluation process of ESFRI, should be recognised and endorsed by the various evaluation agencies at the national level.
3. Establish a common set of indicators, agreed by all national evaluation agencies, making possible joint work at the pan-European level, without requiring the set-up of a pan-European evaluation organisation.
4. Require from MS a permanently updated roadmap of national RI of pan-European interest.

7. Conclusion

Based on previous work of ESFRI evaluation working group, an approach for ex-post evaluation of RI is recommended to harmonise national and pan-European activities and policies. Ex-ante evaluation has already been widely worked out within ESFRI and at the national level, as it is a required phase for decision makers before committing for heavy investments on a long-term basis. Beyond the start of RI, it is needed to proceed also with ex-post evaluation to guarantee scientific excellence, socio-economic impact and quality management. Ex-post evaluation is already required to align national and pan-European or global strategies, and eventually provides the various funding authorities with rationale to decide termination and decommissioning of RI, when their usefulness is no longer demonstrated.

Subgroup membership

Name	Institution/ Position	Country	Contact
Dany Vandromme	National Institute of Applied Science (INSA), Professor **Subgroup Chair**	France	dvandrom@me.com
Lajos Balint	National Information Infrastructure Development Institute (NIIFI), Director of International Relations	Hungary	lajos.balint@niif.hu
Geraldine Healy	Centre for Research in Equality and Diversity, Queen Mary University London, Director	United Kingdom	g.m.healy@qmul.ac.uk
Rosie Hicks	Australian National Fabrication Facility, Chief Executive Officer	Australia	rosie.hicks@anff.org.au
Beatrix Vierkorn-Rudolph	Federal Ministry of Education and Research, Director	Germany	Beatrix.vierkorn-rudolph@bmbf.bund.de
Colin Wright	SANReN (South Africa National Research and Education Network), Chief Strategist & Manager	South Africa	colin.wright@wits.ac.za