WORKSHOP ON BIG DATA AND OPEN DATA

7th – 8th May, 2014
Royal Museums of Art and History
Brussels

Executive Summary

ERF-AISBL in collaboration with **Elettra - Sincrotrone Trieste** has organized a first workshop addressing the ***big data*** and ***open data*** issues with special focus on the common problems that all data producing large scale research facilities are facing and will face in the years to come, and on the search of possible solutions. The purpose of the workshop was to provide a fertile ground for the information exchange between RIs and for joint action on specific proposals to be submitted to the Horizon 2020 programme. The speakers have been selected as high profile experts in this field. The workshop attracted about 100 highly concerned scientists and facility managers. All the sessions were plenary. The workshop was extremely effective in stimulating active discussion both during the presentations and as a result of the networking opportunities offered. The goal of this document is collect the results, suggestions, propositions and analysis of pending issues that were formulated during the workshop and form the direct outcome of the initiative being perceived as useful both for facility managers and policy makers.

The slides presented during the workshop and other multimedia material is publicly available and can be downloaded from the conference website:
http://www.elettra.eu/Conferences/2014/BDOD/

In order to facilitate the reader the summary has been structured as a set of bullet points.

Scientific Committee
Anke Kaysser-Pyzalla
Donatella Castelli
Fabio Pasian
Giorgio Rossi (Chair)
Jesus Marco de Lucas
John V. Wood
Juan Bicarregui
Kimmo Koski
Nicolas Menard
Norbert Meyer
Sanzio Bassini
Volker Guelzow

Organising Committee
Fabio Mazzolini
Giorgio Rossi
Ornela De Giacomo
Roberto Pugliese (Chair)

Workshop Programme

Wednesday, 07 May, 2014

Opening Session	***Chair*: Carlo Rizzuto**
13:00 - 14:00	Registration
14:00 - 14:30	Welcome **Giorgio Rossi, Carlo Rizzuto, Wolfgang Sandner**
14:30 - 15:00	Carlos Morais Pires **Data e-Infrastructures Horizon2020**
Global initiatives in this domain	***Chair*: Kimmo Koski**
15:00 - 15:20	Leif Laaksonen **Research Data Alliance**
15:20- 15:40	Peter Wittenburg, Damien Lecarpentier, Kimmo Koski **EUDAT: Shaping the Future of Europe's Collaborative Data Infrastructure**
15:40 - 16:00	Nobert Meyer, Maciej Brzezniak **e-IRG data challenges**
16:00 - 16:15	Coffee break
Big Scientific Data in related domains	***Chair*: Juan Bicarregui**
16:15 - 16:35	Francoise Genova **Big data in the astronomical community**
16:35 - 16:55	Alf Game **ELIXIR - Big data in Bioinformatics**
16:55 - 17:15	Jamie Shiers **Long Term data preservation for HEP (CERN)**
17:15 - 17:35	Massimo Cocco **Big Data in solid Earth sciences Observatories**
17:35 - 17:55	Brian Wee **Big, Complex Environmental and Biodiversity Data**
17:55 - 18:15	Stefano Cozzini **Data and Metadata for Nanoscience**
16:00 - 16:15	**Summary and discussion**
20:00 - 22:30	**Social Dinner**

Thursday, 08 May, 2014

National and International initiatives in big data domain	***Chair*: Jesus Marco de Lucas**
08:30 - 08:50	Niinimäki Sami **Big and Open Scientific Data in Finland**
08:50 - 09:10	Sanzio Bassini **Italian Scientific Big Data Initiative**
09:10 - 09:30	Norbert Meyer, Maciej Brzeźniak, Maciej Stroiński **Polish National Data Storage**
09:30 - 09:50	Sergi Girona **PRACE Big Data challenge**
09:50 - 10:10	Tiziana Ferrari **EGI services for high throughput big data processing in a secure federated environment**
10:10 - 10:30	Juan Bicarregui **PaNdataODI: an Open Data Infrastructure for Photon and Neutron based research**
10:30 - 10:45	Coffee break
Large Scale Facilities and ERF members initiatives in this domain	***Chair*: Giorgio Rossi**
10:45 - 11:05	Roberto Pugliese **Elettra Big Data and Open Data Challenges**
11:05 - 11:25	Jean-Francois Perrin **ILL Open and Big Data challenges**
11:25 - 11:45	Thomas Stibor **Storing and Analyzing Efficiently Big Data at GSI/FAIR**
11:45 - 12:05	Volker Guelzow **Big Data management for large experiments at DESY**
12:05 - 12:25	Rudolf Dimper **BIG Data at ESRF, big problems, big opportunities**
12:25 - 12:45	Giovanni Lamanna **EU-T0, Data Research and Innovation Hub**
12:45 - 13:00	Bill Pulford **Diamond Big Data and Open Data Challenges**
13:00 - 13:15	**Summary and discussion**
14:00 - 16:00	**Facilities available for specific project proposal meetings**

Workshop Summary

- ERF is organized in Chapters to allow its Associates to interact on specific issues. One Chapter will be set-up on big and open a data in order to capitalize and pursue the work started with this workshop.
- The viewpoint of the EC on Data e-Infrastructures in Horizon 2020 was presented. Europe is riding the research data wave. Data clearly needs logic machines. Scientific Data heterogeneity is a major issue. EC supports specific projects and global initiative such as the Research Data Alliance (RDA). EC recommends open access and preservation of scientific information (see Open-AIRE and the requirement of a Data Management Plan).
- The RDA is building the social and technical bridges that enable open sharing of data. It aims at accelerating and facilitating research data sharing and exchange through working groups and interest groups.
- EUDAT common data services for science based on the universally accepted Collaborative Data Infrastructure have be presented. EUDAT is now ready to service science. EUDAT is developing a sustainable funding model based on services and contracts.
- e-IRG and its data challenges have been introduced as well as the cooperation with ESFRI facilities. Open questions coming from a public consultation process are related to business model (who should pay for what?), standards (who enforces them? what is the motivation to use them? who pays the people holding the standards and assuring the quality?), common interfaces (who defines these? who pays the people enforcing and managing the common interface?) and trust networks (who are the really trusted providers of authentication and authorization information? can it really be decentralized? or should there be some kind of passport office?).
- A set of success stories related to big data in the astronomy community have been presented: there is a strong tradition of international collaboration; a common data format has been adopted since the 70s (FITS); data are open (in general after a rather short proprietary period). All the available infrastructures are driven by community needs (on-line observation archives, on-line services). Key for success is seamless access to data and interoperable tools. Another important role is played by common sense and rules of thumb like "when it works it works!" which means that new infrastructure are better if applied to new challenges.
- The ELIXIR big data and open data infrastructure for bioinformatics has been presented, and its longer term strategy outlined.
- The long-term Data Preservation strategies for High Energy Physics have been presented. CERN is currently storing about 100PB. These data has to be preserved, re-used and shared, now and in the long-term. This is now both affordable as well as technically possible but requires substantial economic effort. CERN relies on public money. It has close relationship with the funding agencies, who are tightly involved in the approval of the scientific programme and the details of how resources are allocated. There is a good understanding of the costs, a proven funding model and a multi-decade outlook. In CERN

experience it is important to avoid overloading the core infrastructures with too many functionalities and this matches well the perceived philosophy in the "infrastructure" versus (complementary) "Virtual Research Environments" calls currently open through H2020.

- The Big Data issues in solid Earth sciences observatories have been presented. The planned Research Infrastructure to face and solve these issues is EPOS. To achieve the best results, it needed continuous orchestration between scientific communities and ITs (e.g., scalability, AAI). The communities are maturing and the ITs are progressively assuming the relevant issues and envisaging solutions. Interactions with industry in Earth sciences require effective strategies and particular attention (ethical issues, use and re-use of scientific data).
- Big, complex environmental and biodiversity data issues have been presented as well as the NEON US project. The main challenges in this field are mainly related to the complex interactions between processes in nature, which are inherently difficult to model and measure. These translate to challenges in provenance management, identifier management, and semantics.
- Data and Metadata issues in the field of Nanoscience have been presented as well as the NFFA project that includes a first data repository for nanoscience including growth and sample nanofabrication protocols as well as characterization data with large scale radiation sources,
- A set of contributions from both national and international initiatives has been presented. Finland has a high number of researchers. By 2017, Finland is aiming to become a leading country in the openness of science and research. Big data is recognized as a very important new technology and a national strategy is in the making. Main problems facing big data are lack of experts, complexity of data and legislative issues.
- The Italian Scientific Data Initiative involving CINECA was presented. CINECA is a research infrastructure supported by National and European (coordinated funding). Scientific community cost contribution is limited to marginal and operational costs. Standard service provision best practices are commonly in use by CINECA, which is fully ISO certified. Pay for service model (applicable for CINECA being a private organization) does imply costs for VAT and couldn't be widely applicable. Open tender for service procurement may be very difficult to apply because of the difficult to define suitable terms of reference documentation for in progress scientific activities.
- The Polish National Data Storage service has been presented. It is based on the 'Common Data Services' model. Services are provided to individuals, SMEs, Universities.
- PRACE's impressive achievements have been presented. PRACE has recently launched an expression of interest on Big Data and received 33 proposals from 8 scientific domains. PRACE is developing a proposal for the handling and access to large datasets arising from the use of PRACE Tier-0 resources. Access is and will be provided purely based on the scientific excellence of the proposals.
- European Grid Infrastructure Services of big data processing in a secure federated environment were presented. The long term vision is for One

European HTC and cloud infrastructure for RIs technically integrated with EUDAT and PRACE and world-wide, complemented by commodity services from commercial providers and for a distributed network of competence centres.

- The PaNdataODI an open data infrastructure for photon and neutron based research has been presented as well as the plans for the exploitation of the results achieved so far.
- Elettra Big Data and Open Data Challenges have been presented. Elettra and FERMI@Elettra light sources and the CERIC-ERIC data production capability have been estimated. Relevant aspects of Big and Open scientific data have been underlined as well as the Elettra Virtual Laboratory and future plans towards a sustainable solution to the Elettra Big Data Challenges. Elettra is planning to integrate its facility to the CINECA scientific storage resources for long-term data preservation. It is important anyhow to be aware of the added value of storing data. Not all the data has to be stored and suitable data format design and compression can greatly reduce the costs.
- The Big Data and Open Data challenges of ILL have been presented. Datasets are huge and it is no more possible for RI users to move data at home. There is the need of remote data analysis. Users & Facility (data producers) rewarding (methods & metrics) is still largely insufficient.
- The approach to storing and analysing efficiently big data produced at GSI/FAIR has been presented as well as defining the trigger of data selection by software.
- The approach to Big Data management for large experiments at DESY has been presented.
- BIG data means big problems but also big opportunities. This is the vision of ESRF. The Photon and Neutron Data Analysis As A Service (PaNDAAS) proposal has been presented.
- The EU-T0 is a hub of knowledge and expertise that optimises the investment of the research funding agencies in proven e-infrastructure. EU-T0 leverages the existing successful federated data and computing infrastructures in Europe, already supporting WLCG, EGI, some ESFRI and worldwide research projects, to build a federated virtual European Tier 0 data-management and computing centre for multi-disciplinary applications. EU-T0 commit in technological innovation developing approaches to manage extremely large and heterogeneous data sets scaling to Exabytes.
- Diamond Light Source Big Data and Open Data challenges have been presented.
- Research Infrastructures as well as European e-Infrastructures both face a similar issue of sustaining the costs of operating their services in the long term, and of ensuring consolidation and expansion of their infrastructures in terms of capacity and capabilities.
- The EC played in the past a key role in ensuring a coordination and leadership in provisioning of a core networking infrastructure for the benefit of the National Research Network that expand it and complement it nationally. The motivations of these are economy of scale in procuring services, innovation, and sharing costs of operations. Thanks to this today European collaborations can rely on a

sustainable network infrastructure, without having to face the issue of procuring these facilities themselves.

- We are today at a stage where ICT procurement is fragmented and left to the responsibility of the RIs.
- Therefore RIs and e-Infrastructures will benefit from a European coordination in capacity and service provisioning. This has the potential of allowing RIs to be part of a single procurement framework, reduce service costs, share resources and operations costs of their ICT infrastructures.
- On the other hand, this will foster sustainability of e-Infrastructures, by relying on a solid customer base.
- A stronger coordination action between e-Infrastructures and RIs to analyse problems and opportunities in service provisioning is needed.

Conclusions

As e-IRG and ESFRI move towards an integrated analysis of the new research infrastructure projects, including upgrades of existing research infrastructures, it has become urgent that EC programmes, like H2020, address the support of data-producers and data-managing infrastructures in an integrated way.

The pressure to develop adequate IT solutions in the Research Infrastructure projects could be better served through a further integration of calls between the DG-RTD and DG-CONNECT in the field of Research Infrastructures.

The Workshop therefore strongly suggests integrated calls are most appropriate for RIs, possibly from WP2016-2017.

E-Infrastructures and common services, including data services, with a direct interface with RIs, and transversal (e.g., all the RIs in a given field of science or sharing a similar formatting of data or technology of data analysis) should be proposed and supported with the aim of shaping the European / International Research Area interfacing and networking with the IT solutions that individual RIs develop optimized for their special features.

Sustainability of transversal e-infrastructure cannot be afforded by the data producing RIs that, on the other hand, must refine their efforts to be compliant with the best-accepted data formats and analysis protocols emerging in the field.
Storing and retrieving scientific value is the goal and this requires a large effort in defining metadata, formats, appropriate semantic data mining tools, and advanced data analysis and computation facilities that should operate at the proper scale (local, cloud) for each scientific community or application. This requires a continuous dialogue between data producers, RI users, IT developers and legal/ethical issue administrators. The Big and Open Data issue is of very high potential impact, but it may absorb excessive resources (money, energy, human

resources) if pursued without a continuously refined strategy requiring the interplay of all data stakeholders.

All the actors in the field must assume a collaborative role and optimize the overall effort. We all should find the way to stimulate the collaboration between the different projects and facilities.