e-IRG Workshop 22-23 May 2013, Dublin

Around 90 participants attended the e-IRG workshop organized in Dublin on 22-23 May 2013 during the Irish EU Presidency. There were two main themes during the workshop. The first one was the European eco-system of services on data, and the second the coordination of e-Infrastructures which is a central element of the e-IRG strategy.

Keynote: Legal Interoperability, Paul Uhlir, National Academy of Sciences, USA

Paul Uhlir started with a definition of Legal Interoperability for data as follows: "the legal rights, terms, and conditions of databases from two or more sources are compatible and the data may be combined by any user without compromising the legal rights of any of the data sources used". His talk focused on Intellectual property (IP) under public law focusing on copyright and database protection law. The main points were the following:

- Legal interoperability is important for data re-users (e.g., public researchers) rather than for end users who are just consumers.
- Public law status quo is uncertain and can be very restrictive for public research users of databases. Data users may even ignore the law.
- Public domain status provides greatest interoperability and freedom for users, but no control or protection for the producer/original rights holder.

Keynote: Ways to improve Global Data Sharing and Re-use, Alberto Michelini**,** Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Alberto Michelini focused on global data and re-use perspective from the side of a seismologist, also involved in the EPOS ESFRI project, and from the side of EUDAT developing horizontal services for other projects. The main points were summarised as follows:

- Individual communities have their own thematic services developed throughout many years and, in general, they are happy with them.
- Data sharing has enormous potential but the feeling is that there is not yet enough expertise on the resulting advantages, e.g., on the science that can be done by "mixing"/correlating different information.
- EUDAT is developing primarily "core services" common to all the e-Infrastructures.
- Building an e-Infrastructure is very demanding given the diversification of the communities in terms of different levels of data organization development/maturity
- We must capitalize on the existing developments in order to avoid to loose pieces (communities) along the way. The true actors are the communities.
- The communities have their own running services.

An Eco-System of Services on Data - Introduction and selected service, Peter Wittenburg, Max Planck Society, Netherlands

Peter Wittenburg stated that after several years of EU funded research on Research Infrastructures and e-Infrastructures, projects should rely on services from other projects or in case these services are not appropriate or non-existent they should build new ones. So a repository of searchable services is required with clear statements on terms of usage and sustainability. And the next question would be who has sufficient authority to work on such a repository. Zenodo http://bit.ly/12UQhcK by OpenAIRE and CERN sharing research data across Europe was mentioned as good service example.

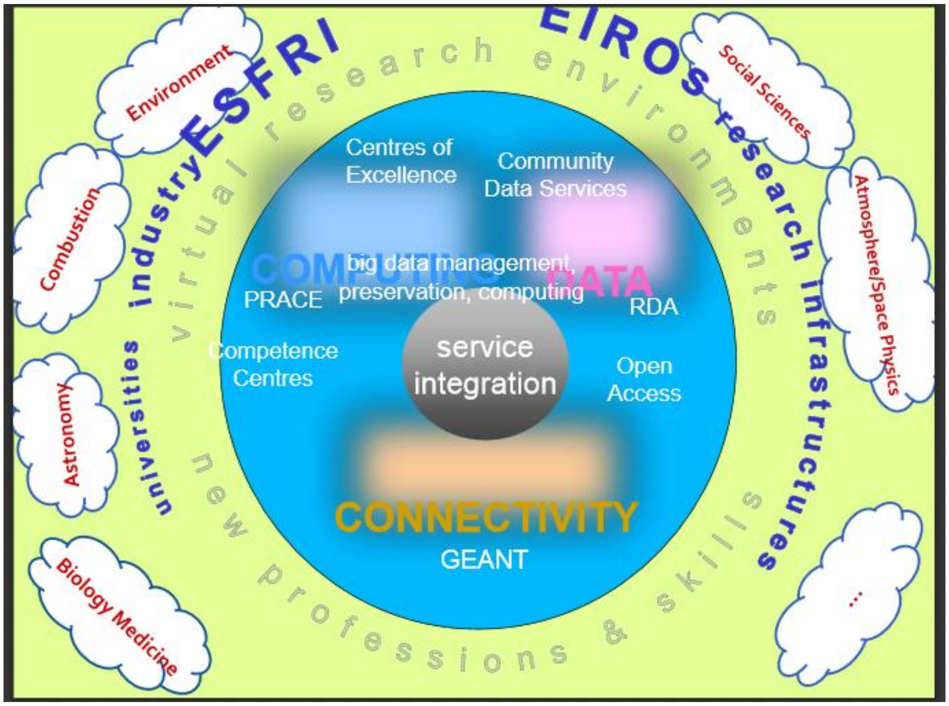
IULA Web services for text data, Núria Bel, IULA-TRL, Universitat Pompeu Fabra, Spain

Nuria Bel introduced the Institute of Linguistics and her group in her university in Barcelona, and stated that they are using web services for enabling researchers to exploit text data. Services include cleaning and anonymisation, quantitative information and statistical analysis, and annotation and enrichment. Sustainability of the services was highlighted as an important element at the end of the presentation and follow-up discussions.

e-Infrastructures and digital ERA, Kostas Glinos, European Commission

Kostas Glinos introduced the digital European Research Area (ERA) as the seamless online space for the circulation of knowledge and technology and stated that the objective of Digital ERA is to make it possible that all researchers in Europe could benefit from e-Infrastructures and Digital Science. After summarising the ERA Communication actions on Digital ERA, he summarised the status of Horizon 2020 programme and then focused in the area of e-Infrastructures.

He stressed that the traditional layered view of e-Infrastructures is not practical anymore as researchers felt that all the layers were separate, and presented a new circular one with emphasis on service integration. The main e-Infrastructure elements are inside the circle and the user communities are around them. Other elements are the VREs (Virtual Research Environments) and the new professions and skills required in this new integrated environment. Finally he presented the timeline for publishing the first e-Infrastructure H2020 calls in December 2013.



In his concluding remarks Kostas Glinos mentioned that the EC is working closely together with member states and stakeholder organisations essential to make progress in e-Infrastructures, and that the roles of e-IRG and ESFRI are important in this environment. H2020 will focus more sharply on user-led service integration, long-term sustainability, big data and policy challenges including openness, innovation and seamless ERA.

Track 1 - Coordination of e-Infrastructures

The main open questions were introduced in the beginning of the session by its chair Lajos Balint; namely the complexity of the topic, whether coordination is enough and whether e-IRG is well-positioned to do such a coordination activity.

HelixNebula Cloud was one of the e-Infrastructures presented in this track by Bob Jones offering a European cloud computing partnership where big sciences (CERN, EMBL and ESA), team up with big industrial cloud businesses. Connections with EGI (testing the HelixNebula BlueBox federated services) and DANTE/NRENs/GEANT were identified, and in the discussion that followed the potential for an e-Infrastructure commons was acknowledged.

Stephen Moffat shared his perspective from IBM Research on the applications and services needed in the next years and the involved IT challenges. Key points mentioned were that technologies are becoming easy to use, openness is a dominating trend and standards for interoperability will gradually eliminate vendor lock-in.

Sandra Collins talked about the Digital Repository in Ireland stressing why Open Access is important; namely because the results of publicly-funded research should remain public, the research findings should be shared with the wider public and that knowledge transfer is enhanced. She then elaborated the approach and policies of the Irish Digital Repository.

Bob Jones in his second speech noted the need for a User Forum of pan-European organisations and projects that operate at an international level including ESFRI cluster projects, EIROforum members, Flagship Projects and ERF (European Association of National Research Facilities). The main topics to be addressed by the Forum would be among others the expectations in e-Infrastructure services, interoperability and sustainability of services, best practices, collaboration and creation of common services, common understanding between service providers and users and user aggregation of needs for industrial or other supply.

In the discussions several points were raised including:

- Sustainability depends on wider impact which is not easy to measure.
- Mechanisms to involve users in designing the services are missing.
- There is a large potential for collaboration and common services need to be found.

Stefano Cozzini introduced the e-Infrastructure requirements for nano-foundries and fine-analysis in the area of Nanoscale Science Research identifying as a basic one the data repository for storing all data from related centres. A prototype was also developed for early testing identifying weak and strong points and that coordination with other e-Infrastructures/ services is essential.

Jon Ison from EMBL introduced the bioinformatics infrastructures ELIXIR and BioMedBridges, and some basic principles that e-Infrastructures need to comply with including collaboration, efficiency, productivity and innovation. Data issues include access, openness, privacy and interoperability. Other issues were also highlighted including software tools, Identifiers and vocabularies.

Sverker Holmgren summarised the White Paper 2013 content, one of its recommendations being a single “e-Infrastructure Commons” for knowledge, science and innovation. Key features should be the high-quality services that are well-managed and seamlessly integrated from a user’s point of view and the dynamicity of service in the evolving future. The need for the Commons was explained, highlighting the insufficient coordination, the legal challenges, the lack of visibility of services, the lack of business models for sustainability, the lack of models for integration with commercial providers and the lack of coherence from many user communities. The proposed approach is the establishment of the Commons through a strategic effort between users, suppliers and other actors. Community building, service provision and innovation are the three core functions of the Commons. In the discussion that followed it was clear that there are business models but nobody likes them and that this needs to be confronted.

Track 2 – Legal Interoperability

Paul Uhlir chaired the session and presented GEO (Group on Earth Observations) Data Sharing Working Group, the CODATA group and the RDA-CODATA Interest Group on Legal Interoperability. It was agreed that the current activities are more working towards landscaping current licenses/laws/solutions than providing solutions for new problems.

Pawel Kamocki and Ville Oksanen presented legal issues from the CLARIN ERIC (Common LAnguage Resources and Technology Infrastructure). They presented the CLARIN data lifecycle and related actors, and it was then stated that the linguistics corpus (a large and structured set of texts) needs content, which is either copyrighted, or in the public domain, or even something in between (in old datasets). "Laundry symbols" for flagging the legal status of datasets were then recommended. It was then declared that CLARIN supports statutory exceptions for research possibly under the existing legal framework; yet more harmonisation is needed.

Andrew Cormack presented the early findings of the e-IRG Task Force on Legal Issues with focus on the commercial use of e-Infrastructures and topics such as data protection, state aid, procurement and software licences. Among the several challenges in the area, it was stated that the current Access Policies (or acceptable usage policies) and software licenses are barriers for the commercial use of e-Infrastructures and that the current data protection law (i.e. personal data protection) is not well compatible with the e-Infrastructure model (i.e. who is the data controller and data processor in e-Infrastructures). In the way forward the following were proposed:

- Offer a standard set of licenses or recommendations for licenses for researchers to adopt.
- Try to clarify how several legal definitions apply to e-Infrastructures.
- Support statutory exceptions for research use.
- Clarify applicability of copyrights for research data.
- Explain the extent of liability of service providers.
- Prepare a cookbook for content owners vs. providers about data ownership.

The key messages included:

- Researchers should maintain the copyright of their work (using an open access model).
- But couple this with default licensing options (otherwise there may be no license or too many licenses).
- Research funders get together and adopt a common access policy to data (for preservation of data that also bring funding obligations).

Other messages comprised the following points:

- Education for researchers on licensing is needed.
- A conference on legal issues with key stakeholders is needed.

- RDA-CODATA Legal Interoperability is willing to host many of these discussions.

Keynote: e-Infrastructures and Customers, Richard Kenway

Richard Kenway provided his views on the relationship between e-Infrastructure service providers and users, mostly given his involvement in the PRACE initiative and that the users should have a clear empowering role. He also presented the related HPC strategy and the associated discussions at the Competitiveness Council that have not concluded successfully. In the discussion that followed there were several comments stating that:

- There is full agreement on empowering users; however the problem is that there are so many users with so many requirements, possibly conflicting.
- Different user communities have different levels of maturity.
- Sampling user requirements may not be the correct way to go; a structured survey is required which has a non-trivial cost.
- Understanding the e-Infrastructure is a pre-requisite for understanding the users.
- Commission bodies work on consensus and if one or two Member States are strongly objecting it is difficult to pass a decision.
- Users' inputs have to be systematically collected; users should have a prominent role in governance.
- Balances should be kept between funding providers and users; some felt that funding should go to the providers and then feedback is received from users and other felt that some direct control of funding to users will make a next step.

Panel on Data Interoperability

Leif Laaksonen chaired the panel and introduced the session by stating that currently a big number of organizations are addressing interoperability issues at various levels such as CODATA, IETF, W3C, RDA, ESFRI RI, etc. The main questions that need to be addressed are among others:

- What is global research data interoperability?
- Who are the main stakeholders to achieving global research data interoperability?
- Who are the main implementers of global research data interoperability?
- How can we create data infrastructures that overcome fragmentation?
- What do we expect from the different organizations, in particular from RDA?
- Do we need a bottom-up or top-down process or do we need both?

Françoise Genova provided a first set of answers:

- Diversity is key for research data and that this diversity needs to be accommodated by data infrastructures and services.
- Main stakeholders are scientists using data, data providers and technologists and these are the main implementers of global data interoperability.

- Generic frameworks for some part of the services building an interoperability layer can help overcome fragmentation and that best practices are very helpful.
- RDA comes in the right moment having a practical approach and a collaborative spirit.
- Both bottom-up and top-down processes are needed in this environment.

Kostas Glinos also recognised the wide diversity of data and agreed that in the same time generic services can be provided. Engagement with communities is needed to enable openness and sharing by supporting the e-Infrastructures that are generic but also thematic (e.g., ELIXIR). Other points mentioned were encouraging the sharing and re-use of data (discoverability, PIDs, establish data producers, users and funders), support the full range of functionality for the ecosystem including data analytics and also bring computing together with data.

Bob Jones stated that ESFRI Research Infrastructures need to be convinced that the e-Infrastructures will be sustained for many years ahead. Furthermore, besides big sciences there is also the long tail of science that needs to be served and that European industry should also be engaged in providing services for research and science. As implementers of global data interoperability Bob Jones named “pathfinder” initiatives such as GEANT, EGI and PRACE but also EUDAT, OpenAIRE+ and HelixNebula. A common platform with three main areas is the solution to fragmentation, namely a network with authentication, authorisation and persistent digital identifiers, small number of computing facilities for generic cloud and data services and software tools to provide added value services. Bob Jones finally argued in favour of a user driven future e-Infrastructure via empowered stakeholders (the User Forum proposal).

Sandra Collins emphasised global good practices, i.e. practices that apply for all communities, including preservation, citation and PIDs, metrics, training, open metadata and access, etc. She then presented DRI statistics for formats used by institutions (from pdf, to rtf and word and xml) along with similar statistics for metadata (where simple and flat Dublin Core was dominating). She closed with the dilemma of consensus vs. discipline approaches.

Paul Uhlir stated that it is not trivial to identify which is the right way to go on data interoperability, especially with a multitude of approaches (national and EU). He said though that there is a good such mix which is to one’s credit. For physical e-Infrastructure he argued in favour of a minimum set of top-down activities (pointing out that you can’t build a road with crowdsourcing). And then a dialogue between the bottom-up and top down is needed. But again funding always come top-down. Regarding one-size does not fit all he argued about a regime that on one hand is vertically integrated (one dimension-communities) and on the other is the horizontal e-Infrastructure (maybe with multiple layers or components) (other dimension). He closed by stating that ownership in the physical sphere is easy, but more work is focusing on virtual layer, and ownership is a lot more fuzzy and especially data, where there is no copyright of facts. In the public sphere you gain value by including people.

Alberto Michelini stated that most of the questions were answered but summarised his views:

- Global research data interoperability is about providing tools and services to analyse data across communities.
- Main stakeholders depend on the type of services but mainly researchers, educators, citizen scientists and industry.
- Main implementers are the data centre communities and the big data centres.
- Fragmentation is inevitable and a first step is to engage with large communities in adopting some standard "abstract model".
- RDA should be acting as a forum where the "abstract model" is discussed and defined; also RDA should be aware of community developments and should provide guidance.
- Both bottom-up and top-down approaches are needed.

During the discussion the following points were made:

- As long as there is no pressure for user communities to come closer to e-Infrastructures, they will prefer to take their own way and even build their own e-Infrastructures.
- Historical analogies with evolving infrastructures that started from users and then gradually became public.
- The right balance between users and providers was repeated with view on sustainable e-Infrastructures.
- The issue of project-based funding and each project doing its own bureaucracy and not having enough resources to federate with other projects.
- It is much easier to get a new project, than to maintain an existing one.
- References to DataCite for datasets citation (using DOIs and other ids) as a best practice; yet a small fraction of data is published there.

Data Services

Peter Wittenburg re-introduced the registry of services and welcomed ideas for how to go about it. Giuseppe Fiameni and Daan Broeder presented services and best practices from EUDAT and CLARIN in this area mainly on lower level services such as data storage and repositories. Then Jon Ison concluded with a presentation on BioMedBridges.

Peter Wittenburg summarised the highlights of the workshop:

- User involvement.
  - A lot of changes are required including culture.
  - A right balance between providers and users is needed.
  - Empowering users is essential but not without assigning responsibilities and accountabilities to them.
- Top down vs. bottom up.
  - Both are needed. Horizontal vs. vertical (community) view.
- Legal issues
  - Happy to see some directions.
  - Legal harmonization is needed.

    - Need progress in the legal area beyond plain landscaping.
    - Set of default licenses is essential, clarifying IPR issues.
    - Exemption for research is needed. See what the related European Parliament decision will be during summer.
- Basis for persistence (sustainability)
- Service registry (EIROforum cluster suggestion).
- Panel on interoperability:
    - No one solution fits all.
    - We need various projects removing barriers.
    - Interoperability is a global activity.
    - RDA has a role to play (follow IETF).
    - Many stories about data and you have to listen…

Sverker Holmgren wrapped-up the workshop by stating that it was a very fruitful workshop. The White Paper was discussed and once again data was in the centre of discussions. On the e-Infrastructure commons there is also wide acceptance, however, the issue is how to make it happen. Yet, there were a lot of discussions about the "big elephants" such as sustainability and user empowerment. For some time they were ignored, but now it is time to deal with them.

The e-IRG chair finally thanked the audience, all speakers, Peter Wittenburg as Programme Committee chair, and of course John Boland and Aideen Kelly and the rest of the HEAnet team for the excellent organisation and dinner.