Data challenges at e-IRG

Norbert Meyer, Maciej Brzeźniak
PSNC

Big Data & Open Data, Brussels, May 7, 2014

e-IRG
e-Infrastructure Reflection Group

e-IRG - Purpose

The **e-Infrastructure Reflection Group** (e-IRG) was founded to define and recommend best practices for the pan-European e-Infrastructure(s)

Inspiring Future e-Infrastructures in Europe and Beyond

The e-IRG mission is to pave the way towards a general-purpose European e-Infrastructure.

The vision for the future is an **open and innovating** e-Infrastructure that enables flexible cooperation and optimal use by international user communities of all electronically available resources.

e-IRG Roadmap

Advancing the European e-Infrastructures

Integrating the Scientific and Public Data Level

Intensifying the Cooperation with Stakeholders

A vision towards a European e-Infrastructure Commons 2020

e-IRG
e-Infrastructure Reflection Group

- The vision for 2020 is that Europe needs a single ‘e-Infrastructure Commons’ for knowledge, innovation and science

- A living ecosystem which is open, accessible and continuously adapts the changing requirements of research

- The European e-Infrastructure components and services will have to developed into an ecosystem of **networking, high throughput/high performance computing and data resources supporting virtual research communities worldwide**

A vision towards a European e-Infrastructure Commons 2020

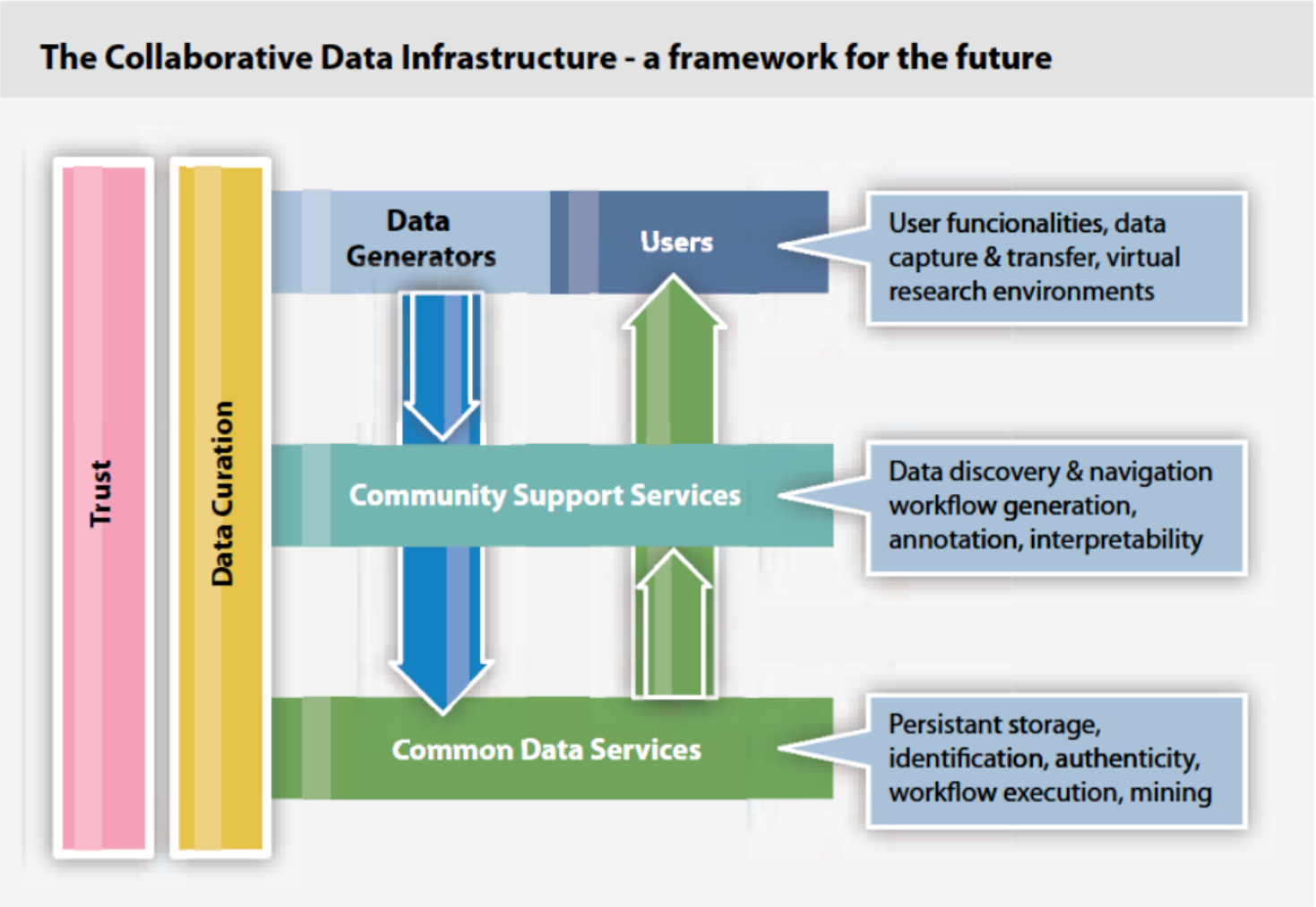
e-IRG
e-Infrastructure Reflection Group

The leading edge users of data and their service providers will have met the challenge to develop an international framework for how different companies, institutes, universities, governments and individuals would interact with their data systems

Collaborative Data Infrastructure

A vision towards a European e-Infrastructure Commons 2020

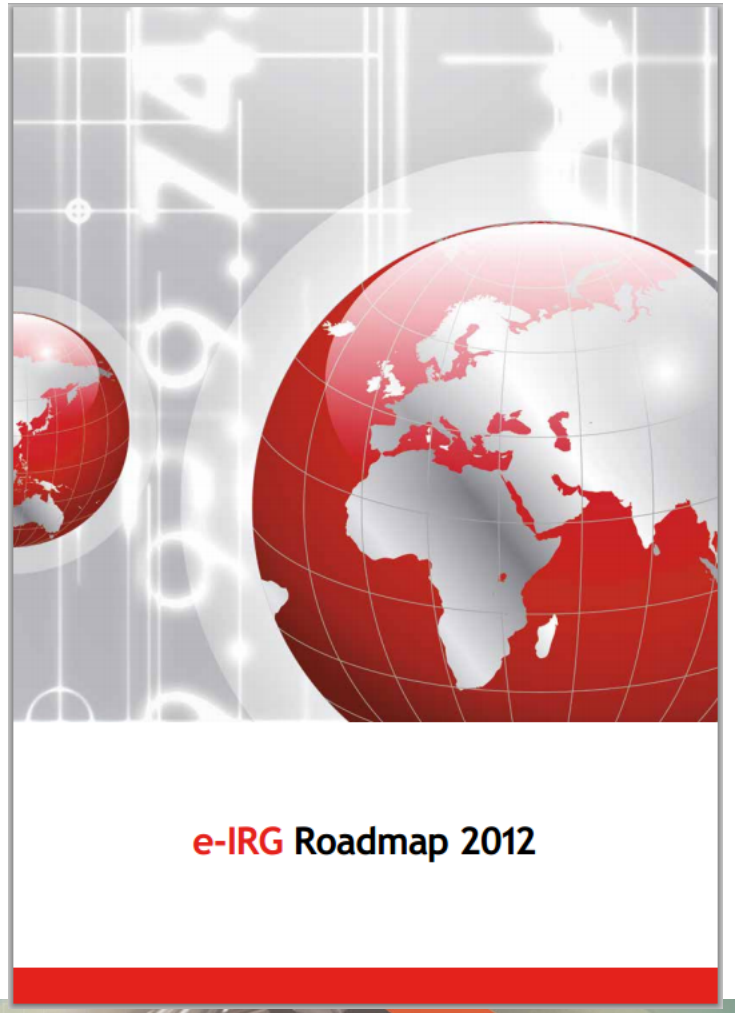
e-IRG
e-Infrastructure Reflection Group



e-IRG Roadmap

More about the plans in the Roadmap

http://www.e-irg.eu



Identified stakeholders

- End user
- Data owner
- Infrastructure provider
- Computer science researchers
- Service provider
- Policy maker

Importance of requirements

Large datasets

Restricted access

Metadata structure

Federated AAI

Accounting

Data provenance

Integration

Interoperation

High trust and security

Reliability

Access

Advanced search functionality

Data preservation

Interpretation of data

Unified access

High quality

Simplified access

Searching information

Important stakeholders

Topics covered by the Blue Paper on DM

1. **Grand Challenges and ESFRI Requirements**

2. **Data e-Infrastructure**
3. **Reliability and Replications**

4. **Metadata**
5. **Unified Access and Interoperability**
6. **Security**

Co-operation with big RI

e-IRG and ESFRI

Research infrastructures (RIs), such as the initiatives on the ESFRI roadmap, produce and are dependent on rapidly increasing amounts of data.

For research and society to take full benefit of the major investments in RIs the data needs to be **made openly and easily available for researchers**, over wide spans of fields, in sustainable settings.

RECOMMENDATIONS

Blue Paper

White Paper

for policy makers

Policy makers are recommended to take action to ensure that

- Roles (e.g. end users, data owners, infrastructure providers, service providers, and researchers on data management) are identified and , when appropriate, partitioned between different actors to ensure effective and cost-efficient solutions, fulfilling the needs of the end users and data owners.
- Governance and mandates for different actors are clarified and their way of interacting is sufficiently formalised. Actors specializing on different tasks ensures that synergies can be exploited, leading to cost-efficient implementation of services. Clear responsibilities and formalised relations ensure that the relevant quality of services can be maintained. Funding paths are defined and sustainability for all parts of the e-Infrastructure is secured.
- Costs for different services and procedures are made transparent and that different options for implementing them are investigated.

Ris and e-Infrastructure providers

- Ensure that data formats are standardised and contain sufficient information on the data (metadata) to enable global usage within the discipline, across disciplines, and in new research settings that could possibly not be envisaged at the time of creation of the data.
- Build e-infrastructure solutions consisting of multiple layers, successively adding more specialised higher-level services using standardised interfaces. Here, different layers can be provided by different actors.
- Define and successively move towards a **common second-level data storage layer** where cross-related requirements between different RIs are identified and utilised to enhance cost-efficiency and quality.

Data e-Infrastructure - recommendations

- Define **business cases and requirements** for ESFRI projects related to data access and data infrastructure
- **Define cross related requirements** generalized for all ESFRI communities
- Provide a current state of the data infrastructure within ESFRI
- Ask the service providers to provide a **sustainability policy** and to state what they can offer ESFRI projects
- Need a **redundant data infrastructure** where strong data centres form a backbone for data access and preservation
- Define a role of each ESFRI project. For the near future, we have to assume a high degree of **specialisation**, where the roles of service/application provider, infrastructure provider and data owner/producer will be separated. Only really big organisations will still be able to consolidate all of these roles..

Data e-Infrastructure - recommendations

- Define **business cases and requirements** for ESFRI projects related to data access and data infrastructure
- **Define cross related requirements** generalized for all ESFRI communities
- Provide a current state of the data infrastructure within ESFRI
- Ask the service providers to provide a **sustainability policy** and to state what they can offer ESFRI projects
- Need a **redundant data infrastructure** where strong data centres form a backbone for data access and preservation
- Define a role of each ESFRI project. For the near future, we have to assume a high degree of **specialisation**, where the roles of service/application provider, infrastructure provider and data owner/producer will be separated. Only really big organisations will still be able to consolidate all of these roles..

Metadata

Communities and interoperability

- Establish **metadata service managers** and give them a greater role in supporting 'newcomers' with their specific requirements to join, and setting up cross disciplinary metadata search
- Give recommendations for distributing responsibilities wrt.
  - metadata quality
  - providing guidance for interoperability (across disciplines)
  - providing metadata services (across disciplines)

Best practices

- Enable easy and **standardised metadata**
- Establish **federated data catalogues**
- Pay attention to aspects of granularity. The need for describing sets of resources (data-sets) at different levels of granularity calls for different metadata schemas.

Security

- Investigate on all ESFRI project the security requirements
- Check if the community is ready to use a **federated authentication process** and the cost of the transition phase
  - **Federated authentication process:** The user should be able to handle the authentication process with a similar user experience as with the most common web applications and leverage existing institutional or national electronic identities (such as the European Citizen Card) to access institutional or community attribute servers to gain access to distributed data and services.
- **Implement data encryption**: data-centric, file-level encryption that is portable across all computing platforms and operating systems should be available to users as a way to increase data protection, confidentiality and integrity in transit and at rest.
- **Influence the EU Data Protection directive under revision** which should address the aspects which are important especially from the data owner and end user point ov view.

Open questions from the public consultation process

1. **business model:** who should pay for what? you only recommend to look at business cases. in such a document you would need to discuss this point at least that investment is needed if data is to be kept at all.

2. **Standards:** who enforces them? what is the motivation to use them? who pays the people holding the standards and assuring the quality?

3. **common interfaces**: who defines these? the market does not do it, although some efforts are there to standardize on interfaces, data is always ignored because there is money to be made with nonstandard interfaces. who pays the people enforcing and managing the common interface?

Open questions from the public consultation process (cont.)

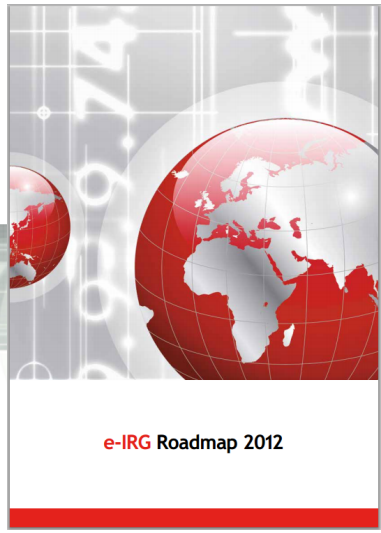
4. **trust network:** who are the really trusted providers of authentification and authorization information? can it really be decentralized? or should there be some kind of passport office? this is very important for medical data, personal data, and all other data that needs to be protected.

Good topics for further investigations (?)

Thank you!

Reports available on

www.e-irg.eu/publications



e-IRG "Blue Paper" on

Data Management