OPINION ARTICLE

Check for updates

Related to other papers in this special issue	4 (p40); 12 (p122); 19 (p192); 26 (p257)
Addressing FAIR principles	F, A, I, R

State of FAIRness in ESFRI Projects

Peter Wittenburg[1†], Franciska de Jong[2], Dieter van Uytvanck[2], Massimo Cocco[3], Keith Jeffery[4], Michael Lautenschlager[5], Hannes Thiemann[5], Margareta Hellström[6], Ari Asmi[7] & Petr Holub[8]

[1]Max Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany

[2]CLARIN ERIC, CLARIN ERIC Drift 10, Utrecht 3512 BS, The Netherlands

[3]EPOS ERIC, Via di Vigna Murata 605, Rome 00143, Italy

[4]Keith G Jeffery Consultants, Faringdon, UK

[5]DKRZ Ringgold standard institution, Bundesstr. 45a, Hamburg, Hamburg 20146, Germany

[6]Department of Physical Geography and Ecosystem, Ringgold standard institution, Lund University, Scienc Sölvegatan 12, Lund 22362, Sweden

[7]University of Helsinki Ringgold Standard Institution – Institution of Atmospheric and Earth System Sciences, P.O.Box 64, Helsinki 00014, Finland

[8]BBMRI-ERIC Neue Stiftingtalstrasse 2/B/6, Graz 8010, Austria

Keywords: Infrastructure; FAIR Metrics; GO FAIR Matrix

Citation: P. Wittenburg, F. de Jong, D. van Uytvanck, M. Cocco, K. Jeffery, M. Lautenschlager, ... & P. Holub. State of FAIRness in ESFRI projects. Data Intelligence 2(2020), 230–237. doi: 10.1162/dint_a_00045

Downloaded from http://direct.mit.edu/dint/article-pdf/2/1-2/230/1893423/dint_a_00045.pdf by guest on 15 July 2024

ABSTRACT

Since 2009 initiatives that were selected for the roadmap of the European Strategy Forum on Research Infrastructures started working to build research infrastructures for a wide range of research disciplines. An important result of the strategic discussions was that distributed infrastructure scenarios were now seen as "complex research facilities" in addition to, for example traditional centralised infrastructures such as CERN. In this paper we look at five typical examples of such distributed infrastructures where many researchers working in different centres are contributing data, tools/services and knowledge and where the major task of the research infrastructure initiative is to create a virtually integrated suite of resources allowing researchers to carry out state-of-the-art research. Careful analysis shows that most of these research infrastructures worked

† Corresponding author: Peter Wittenburg (E-mail: peter.wittenburg@mpcdf.mpg.de, ORCID: 0000-0003-3538-0106).

© 2019 Chinese Academy of Sciences Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license

on the Findability, Accessibility, Interoperability and Reusability dimensions before the term "FAIR" was actually coined. The definition of the FAIR principles and their wide acceptance can be seen as a confirmation of what these initiatives were doing and it gives new impulse to close still existing gaps. These initiatives also seem to be ready to take up the next steps which will emerge from the definition of FAIR maturity indicators. Experts from these infrastructures should bring in their 10-years' experience in this definition process.

1. INTRODUCTION

The European Strategy Forum on Research Infrastructures (ESFRI) [1] was established in 2002 as a response to the increased relevance of state-of-the-art infrastructures for modern research. As a policy level organisation, ESFRI's mission is to support a coherent and strategy-led approach to policy-making on research infrastructures in Europe. ESFRI distinguishes between physical infrastructures located at one place (common in natural sciences, for example CERN [2]), distributed physical infrastructures (such as the new antenna systems being set up for example by SKA [3]) and virtual distributed infrastructures. In the latter, many scientists located at different places work together to produce integrated collections of digital objects (data, software, etc.) intended for data-intensive science. In this paper we will focus on those infrastructures which are widely distributed in nature which increasingly often can be found in many scientific disciplines. The ESFRI process has produced roadmaps from 2006 on and the first projects in various research areas[1] started in 2009 with their work. Currently there are ESFRI projects in almost all areas of research, some have the status of a formal legal entity funded in a sustainable manner by the EC and the European member states.

2. EXAMPLE INFRASTRUCTURES

In this paper we use the CLARIN, ICOS, EPOS, IS-ENES and BBMRI infrastructures as examples and draw some generic conclusions. These research infrastructures (RI) were all started with the goal to create an integrated and harmonised domain of digital objects that could be easily exchanged, integrated and reused to enable new kinds of research questions. Researchers formulated an increasing interest to work with remote facilities and make use of distributed databases for example.

2.1 CLARIN

The CLARIN ERIC [4] (Common Language Resources and Technology Infrastructure) research infrastructure, which has become a legal entity in the meantime, was started with the intentions to overcome the fragmentation in the domain of language resources and to make data and tools much more findable,

[1] ESFRI domains include Energy, Environment, Health & Food, Physical Sciences & Engineering, Social & Cultural Innovation and Data, Computing and Digital Research Infrastructures.

Downloaded from http://direct.mit.edu/dint/article-pdf/2/1-2/230/1893423/dint_a_00045.pdf by guest on 15 July 2024

accessible and reusable and to make steps to increase interoperability. The following major dimensions are tackled by CLARIN:

- harmonising the domain of metadata descriptions by developing a component based system (CMDI) supported by easy-to-use tools
- promoting the sharing of metadata, harvesting all globally available metadata about language resources and creating a search portal (VLO) based on fast indexes and semantic mappings
- motivate researchers to increase the quality of their resources and to make them available via trustworthy repositories/centres that are assessed by CoreTrustSeal [5]
- developing distributed workflow frameworks allowing everyone to carry out analytics on textual data based on goal-driven tool orchestrations
- working on advanced concepts such as switchboard schemes to increase the interoperability of data types and tools
- clarifying ethical and licensing aspects for language resources

From the beginning CLARIN worked on increasing Findability, Accessibility, Interoperability and Reusability, although the FAIR principles [6] were not known at the start time. Widely agreed FAIR maturity indicator tools will therefore be applied by CLARIN when they will become available.

2.2 BBMRI

Biobanking and BioMolecular resources Research Infrastructure (BBMRI-ERIC) [7] is a pan-European research infrastructure with 20 national nodes to overcome fragmentation in the domain of biobanks which store all types of human biological samples, such as blood, tissue, cells or DNA, data on the research participants (consenting patients/donors) and data associated with the samples, as well as other biomolecular resources that can be used in health research. The intention is to bring together all the main players from the biobanking field – researchers, biobankers, industry, and patients – to boost biomedical research.

Since its inception BBMRI-ERIC focuses on the following main topics, which are largely related to the FAIR principles:

- *Findability:* basic findability is provided by BBMRI-ERIC Directory [8], which contains aggregated descriptors of collections of data and biological material stored in the biobanks. Technology preview of BBMRI-ERIC Locator [9] allows for obtaining estimates of available cases and biological samples, based on a federated search mechanism.
- *Accessibility:* Basic accessibility information is already available in the BBMRI-ERIC Directory and further negotiation of access permissions is supported.
- *Interoperability:* In order to improve interoperability of sharing biological material and associated data, BBMRI-ERIC is working on community standards such as MIABIS 2.0 Core [10] or MIABIS Sample/Donor Data Model [11]. BBMRI-ERIC has also established an Interoperability Forum [12], which aims to provide a vendor-neutral platform to standardize APIs and data models related to the biobanking domain.

Downloaded from http://direct.mit.edu/dint/article-pdf/2/1-2/230/1893423/dint_a_00045.pdf by guest on 15 July 2024

- *Reproducibility and reusability:* Having biomedical research facing dramatic reproducibility challenges for more than a decade [13-20], BBMRI-ERIC promotes quality management in biobanks and leads development of provenance information management standard in ISO TC276 (PWI23494-1).
- *Privacy protection:* Dealing with sensitive human data and human biological material, BBMRI-ERIC is developing internal policies for optimum use of privacy enhancing technologies, in order to retain maximum value of the data made available for the research while minimizing risks for the research participants donating their data for research.

BBMRI-ERIC community has also proposed a specific extension called FAIR-Health [21], primarily focusing on reproducibility and on privacy protection policies. In specific cases, BBMRI-ERIC promotes also accessibility and utilization of biobanks by collecting large collections of data; this is demonstrated by a colorectal cancer cohort (CRC-Cohort) of 10,380 data sets from 25 biobanks from across Europe to foster cancer research accessible via a unified metadata set. The CRC-Cohort is being integrated into the tools described above.

BBMRI-ERIC offers services related to implementation of quality management in biobanks and guidance on ethical, legal, and societal issues that biobankers and researchers may encounter, in particular due to the EU General Data Protection Regulation.

2.3 EPOS

European Plate Observing System (EPOS) [22] facilitates integrated use of data, data products, and facilities from distributed research infrastructures for solid Earth Science in Europe and brings together different stakeholders to develop new concepts and tools for accurate, durable, and sustainable answers to societal questions relevant to the environment and human welfare. It is integrating the diverse and advanced European Research Infrastructures for solid Earth Science, and is building on new e-science opportunities to monitor and understand the dynamic and complex solid-Earth System. EPOS is tackling the following major tasks:

- helping Earth scientists and others to develop a more holistic understanding about the underlying processes of Earth's dynamics by providing an integrated view on observational data, data products, extracted knowledge and solutions;
- aggregating information of about 400 elements (data, data product, software and services) within the EPOS federation combining, for example, satellite and in-situ earth observations to model surface deformations and tectonic processes causing earthquakes;
- offering legal solutions securing a common and shared data policy for open access and a transparent use of data, and guaranteeing mutual respect of the intellectual property rights;
- promoting open standards and developing new standards where necessary in collaboration with other European and global initiatives in earth science to tackle data sharing and interoperability;
- building a virtual research environment (the Integrated Core Services) providing discovery, access, workspace, visualisation and processing services representing a practical solution to data interoperability and a feasible integration of services shared with scientific communities.

Downloaded from http://direct.mit.edu/dint/article-pdf/2/1-2/230/1893423/dint_a_00045.pdf by guest on 15 July 2024

The highly fragmented landscape consisting of national and international research infrastructures covering a variety of scientific domains requires integration to establish the EPOS research infrastructure and the adoption of shared practices to improve findability, accessibility, interoperability and re-use for the benefits of earth science. In addition to these FAIR dimensions, finding solutions to improve data and metadata quality and ensure long-term accessibility are important as well.

2.4 ICOS

Integrated Carbon Observation System (ICOS ERIC) [23] is a pan-European research infrastructure with a mission to provide standardised, long term, high precision and high quality observations on the carbon cycle and greenhouse gas budgets, and their perturbations. ICOS first entered the ESFRI Roadmap in 2006 and became an ERIC legal entity in 2015.

The ICOS observing network consists of over 130 observation stations, each related to one or more of the three domains: Atmosphere, Ecosystem and Ocean. The collected data is processed and quality controlled at Thematic Centres (one for each domain), before being openly distributed via the ICOS Carbon Portal data centre.

All ICOS data are meant to be easy to find, available for open access, fully traceable, complete with all relevant metadata, and interoperable with other (environmental) data and services. Indeed, ICOS has been committed to making its data and services FAIR even from before the term was coined, as outlined in the ICOS Carbon Portal concept paper from 2012.

The ICOS Carbon Portal service list includes

- data ingestion & storage, including the minting of persistent identifiers;
- staging data from the repository to HTC resources;
- easy-access cataloguing on top of an ontology-based metadata database (RDF triple store accessible via a SPARQL endpoint);
- provisioning of dynamic landing pages for any digital object described in the metadata store, including data sets, observation stations, data type specifications and concepts;
- single-sign on authorization, authentication and identification (AAI) for ICOS services;
- a virtual research environment (VRE) platform for user-initiated data processing (based on Jupyter Notebook running on virtual machine instances);
- data discovery, including searching, visualising and downloading of ICOS-related data products including usage tracking.

As far as possible, the Carbon Portal bases all its data management and computing services on Open Source technology.

Downloaded from http://direct.mit.edu/dint/article-pdf/2/1-2/230/1893423/dint_a_00045.pdf by guest on 15 July 2024

2.5 IS-ENES

IS-ENES is the infrastructure project of the European Network for Earth System Modelling community (ENES) [24] with the aim of developing a common climate and Earth System Modeling (ESM) and data research infrastructure in Europe. IS-ENES started in 2009 as part of the ESFRI roadmap and is continued by its broad activities in supporting the 6th IPCC Assessment Report [25] and promoting the scientific goals of the global Earth System Modelling community. IS-ENES main tasks were to further integrate the European climate modelling community, to ease the development of full ESMs, to foster the execution and exploitation of high-end simulations, and to support the dissemination of model results and the interaction with the climate change impact community.

To achieve the goals, the global climate modeling community is working on the data standard Climate Modeling Intercomparison Project (CMIP), the 6th version of which is currently being completed. Centres such as the World Data Climate Centre (WDCC) [26] are the strong pillars for this community and have the task to carry out proper data management, long-term archiving and data publishing. Centers such as WDCC need to meet and demonstrate high quality standards, which is the reason for participating in the CoreTrustSeal quality assessment procedures.

Through CMIP6 and associated policy rules high quality standards are achieved. Sustainable findability and accessibility is guaranteed by associating all data with PIDs. For all hierarchical levels, starting at the level of individual data objects up to collections and published data sets, extensive metadata is created, which can be exported in different formats or accessed via OAI-PMH, for example. Accessibility and Interoperability is achieved by supporting open standards for metadata and data (netCDF-CF) as much as possible and by ensuring that all schema and vocabulary definitions are accessible. Reusability is essential for the ENES community and thus clear license terms are defined for all metadata (CC0) and data objects (CC-BY 4.0). Harmonised provenance recording needs to be improved across the various centers engaged in ENES and beyond. Therefore, IS-ENES can claim to have been widely FAIR compliant before the principles were published.

3. CONCLUSIONS

In this paper we investigate various strategies and approaches taken by ESFRI initiatives towards implementation of the FAIR principles. For this purpose, we have taken CLARIN, EPOS, BBMRI, ICOS and IS-ENES as examples, assuming them to be representative for the many others that are confronted with a highly distributed and fragmented domain of resources. All these initiatives started in 2009 with the clear mission to improve the conditions of carrying out cutting edge data intensive science in the respective fields that will lead to deeper scientific insights. We see reoccurring patterns in these research infrastructures such as increasing quality and trust in data and in trustworthiness of the care takers (repositories), solving the issues of persistency of data services and stability of references, creating an integrated domain of metadata facilitating the creation of a joint index and catalogue, addressing the challenge of making semantic encodings more explicit and determining community wide standards and best practices.

Downloaded from http://direct.mit.edu/dint/article-pdf/2/1-2/230/1893423/dint_a_00045.pdf by guest on 15 July 2024

Without having the "FAIR" principles in mind at the start, improving findability, accessibility, interoperability and re-use were amongst the key missions of all these distributed research infrastructures. The recent emergence and wide adoption of the FAIR principles have validated their pioneering efforts. Additionally, other dimensions such as creating a culture of data sharing, improving data management/stewardship, looking for persistent solutions based on continuous funding streams, improving the readiness level of services and investing in improving the skills of all actors have been key for the ESFRI initiatives. Many of the highlighted ESFRI initiatives have actively participated in Research Data Alliance (RDA) [27] working and interest groups, which they regard as a platform to exchange ideas and to work on agreements within the global communities as well as across them.

We believe that the ESFRI initiatives we studied are all ready to start the work to integrate the application of FAIR maturity indicator tests to their data and service curation workflows. However, given their long experience with developing current data management practices, there is also an expectation from their side to be involved in defining these indicators and establishing the associated assessment procedures. Massimo Cocco from EPOS recently expressed the priorities this way: *FAIR is a realistic goal, but we need practices, not more principles!* This view is seemingly widely shared by the ESFRIs initiatives: the FAIR principles are important, but taking all necessary steps will cost considerable resources and efforts, and convincing all actors to change their practices will take time. ESFRI initiatives are participating in the RDA FAIR Data Maturity Model Working Group proposing concrete mechanisms to conform with FAIR principles before establishing metrics for assessment.

Downloaded from http://direct.mit.edu/dint/article-pdf/2/1-2/230/1893423/dint_a_00045.pdf by guest on 15 July 2024

REFERENCES

[1] European Strategy Forum on Research Infrastructures. Available at: https://www.esfri.eu/.
[2] CERN. Available at: https://home.cern/.
[3] SKA Telescope Project. Available at: https://www.skatelescope.org/.
[4] CLARIN ERIC. Available at: https://www.clarin.eu/.
[5] Core Trust Seal. Available at: https://www.coretrustseal.org/.
[6] FAIR Principles. Available at: https://www.go-fair.org/fair-principles/.
[7] BBMRI ERIC. Available at: http://www.bbmri-eric.eu/.
[8] BBMRI Directory. Available at: https://directory.bbmri-eric.eu/.
[9] BBMRI-ERIC Locator. Available at: https://search.germanbiobanknode.de/.
[10] R. Merino-Martinez, L. Norlin, D.van Enckevort, G. Anton, S. Schuffenhauer, K. Silander, L. Mook, P. Holub, R. Bild, M. Swertz & J.E. Litton. Toward global biobank integration by implementation of the minimum information about biobank data sharing (MIABIS 2.0 Core). Biopreservation and Biobanking 14(4) (2016), 298–306. doi: 10.1089/bio.2015.0070.
[11] MIABIS Sample/Donor Data Model. Available at: https://github.com/MIABIS/miabis/wiki/Data-describing-Sample-Donor.
[12] BBMRI Interoperability Forum. Available at: http://www.bbmri-eric.eu/news-events/bbmri-eric-bbmri-uk-launch-interoperability-forum/.
[13] L.P. Freedman, I. M. Cockburn & T.S. Simcoe. The economics of reproducibility in preclinical research. PLoS Biology 13(2015), e1002165.doi: 10.1371/journal.pbio.1002165.

[14] F. Prinz, T. Schlange & K. Asadullah. Believe it or not: how much can we rely on published data on potential drug targets? Nature reviews Drug discovery 10(2011), 712. doi: 10.1038/nrd3439-c1.
[15] C.G. Begley & L.M. Ellis. Drug development: raise standards for preclinical cancer research. Nature 483(2012), 531–533. doi: 10.1038/483531a.
[16] C.G. Begley. Reproducibility: six red flags for suspect work. Nature 497(2013), 433–434. doi: 10.1038/497433a.
[17] P. AC't Hoen, M.R. Friedländer, J. Almlöf, S. Michael, P. Irina, Y. A. Seyed, Jeroen F J Laros … & L.Tuuli. Reproducibility of high-throughput mRNA and small RNA sequencing across laboratories. Nature biotechnology 31(2013), 1015–1022. doi: 10.1038/nbt.2702.
[18] M. Bissell. Reproducibility: the risks of the replication drive. Nature 503(2013), 333–334. doi:10.1038/503333a.
[19] A. Mobley, S. K. Linder, R, Braeuer & Z.Leonard. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. PLoS One 8(2013), e63221. doi: 10.1371/journal.pone.0063221.
[20] S.J. Morrison. Time to do something about reproducibility. eLife 3(2014), e03981.doi: 10.7554/eLife.03981.
[21] P. Holub, F. Kohlmayer, F. Prasser, M.T. Mayrhofer, I. Schlünder, G.M. Martin ... & D. Strapagiel. Enhancing reuse of data and biological material in medical research: From FAIR to FAIR-health. Biopreservation and biobanking, 16(2) (2018), 97–105. doi: 10.1089/bio.2017.0110.
[22] EPOS Research Infrastructure Project. Available at: https://www.epos-ip.org/.
[23] ICOS Research Infrastructure Project. Available at: https://www.icos-ri.eu/.
[24] ENES Research Infrastructure Project. Available at: https://portal.enes.org/.
[25] 6th IPCC Assessment Report. Available at: https://www.ipcc.ch/activities/sixth-assessment-report.
[26] World Data Climate Centre. Available at: https://www.dkrz.de/up/systems/wdcc.
[27] Research Data Alliance. Available at: https://www.rd-alliance.org/.

Downloaded from http://direct.mit.edu/dint/article-pdf/2/1-2/230/1893423/dint_a_00045.pdf by guest on 15 July 2024