Knowledge Exchange

The evolving landscape of Federated Research Data Infrastructures

Final report on the situation in the six Knowledge Exchange partner countries

Report dated November 2017

DFG CSC Jisc cnrs SURF Deff Denmark's Electronic Research Library

Title: The evolving landscape of
Federated Research Data Infrastructures

Authored by: Stéphane Goldstein
visit: www.informall.org.uk

DOI: 10.5281/zenodo.1064730

Report dated: November 2017

Contents

Introduction: Background and methodology

1. This report, commissioned from Knowledge Exchange (KE), is an overview and synthesis of the evolving landscape of Federated Research Data Infrastructures (FRDIs) in the six KE partner countries: Denmark, Finland, France, Germany, the Netherlands and the United Kingdom. The fieldwork and study underlying the report were undertaken by InformAll CIC during the first half of 2017, on the basis of interviews with experts from a range of organisations that run federated infrastructures in the respective countries.

2. The report reaches nine broad conclusions, below. It first considers the issues and background to the study, then the methodology deployed, before synthesising and analysing the information gathered during the fieldwork. This information is structured to address the following: what is understood by 'federated'; drivers for FRDIs (push factors and meeting demand); operation of the infrastructures; data-related practices; training and skills; challenges and obstacles; outcomes and impact; and implications for the European Open Science Cloud (EOSC).

Summary of conclusions

3. The report's nine broad conclusions are summarised here, and set out more fully at the end of the document (**para 120**):

Conclusion 1
Federated infrastructures can apply to research disciplines across the spectrum, including physical sciences, life/health sciences, environmental sciences, social sciences and the humanities; they may also be multidisciplinary. Disciplines carry their own characteristics, for instance sensitivity and confidentiality of research data, which impact on the nature of the federations.

Conclusion 2
Essentially, a federated infrastructure is one where a range of distributed services - focused on the actual demands of research - are coordinated by an overarching level, with the aim wherever possible of providing seamless access to research data and tools.

Conclusion 3
Two broad sets of factors drive the emergence and development of federated infrastructures: push factors, which might be also be characterised as top-down, including political and public interest drivers; and demand from users, reflecting a bottom-up approach and research cultures.

Conclusion 4
Infrastructures are often characterised by long-term financial uncertainty, reflecting short-term project funding schemes.

Conclusion 5
The involvement of users is also a crucial imperative, and infrastructures are careful to nurture their relationships with numerous partners within the academic sector and beyond.

Conclusion 6
Infrastructures are characterised by a wide range of practices and services, which vary according to the nature of each initiative and evolve dynamically in the light of researchers' needs.

Conclusion 7
A major challenge for the development of federated infrastructures is the complexity and fragmented nature of the research data environments in which they evolve.

Conclusion 8

Many infrastructures have processes in place to evaluate impact, either through measuring usage in a quantifiable (but in practice, often a limited) way, or through formal review mechanisms overseen by governance bodies. However, impact evaluation is often limited or sketchy and represents a challenge for infrastructures.

Conclusion 9

The emergence of EOSC is generally welcomed, particularly since it is seen as reflecting the same rationale as national infrastructures, albeit at a pan-European scale – with the beneficial scaling up that this could imply.

The issue

4. The management of research data (mostly, but not exclusively, in digital form) has increasingly become an intrinsic part of the academic research endeavour, across all disciplines: not just in the pure sciences and social sciences, but in the arts and humanities too. A diminishing number of researchers do not use, create or share data, whatever forms this might take. As data-centric research methodologies have become more prevalent it has become necessary to develop infrastructures to help ensure that data is validated, accessible, shareable, reusable and effectively curated; and crucially, to make sure that it conforms to the principles of open research data[1]. These infrastructures have developed both within institutions and in the wider national and international context. Within institutions, systems and procedures have been put in place to cover data management planning, ingest, cataloguing, storage and preservation; current research information systems and repositories have been built. Beyond individual institutions, national, international and disciplinary data centres play an important role – some of them long-established; aggregation and identifier services and research data discovery services provide important entry points. And data publications are emerging as media in their own right.

5. This infrastructure provides necessary tools to help the research community navigate its way through the data jungle – but it is fragmented, and often lacking in transparency and sustainability. It involves different sets of technologies, players and competencies, reinforcing the confusing nature of the mix. The European Commission recognised this problem in 2010 in its ‘Riding the wave’ report[2], with its call for the development of an international framework for a collaborative data infrastructure, allowing for the different players – institutions, companies, public bodies and individuals – to interact more easily with the data infrastructure. When in 2011 KE looked to formulate a programme to implement the conclusions from ‘Riding the wave’, it also recognised that “the [data] infrastructure is diverse, fragmented, in flux and organised differently across various disciplines in different countries”[3].

Footnotes

[1] Various organisations have set out principles of open research data, but a particularly good example of a comprehensive framework of principles is the UK’s Concordat on Open Research Data, set out jointly in 2016 by the Higher Education Funding Council for England (HEFCE), Research Councils UK, Universities UK and the Wellcome Trust – rcuk.ac.uk/documents/documents/concordatonopenresearchdata-pdf. See also para 75, which addresses the related issue of the FAIR Data Principles. Another example is the principles agreed on in 2010 by all major German Science Organisations (Allianz), the Alliance of Science Organisations’ Principles for the Handling of Research Data; allianzinitiative.de/en/core-activities/research-data/principles.html

[2] European Commission, High Level Expert Group on Scientific Data, 2010, ‘Riding the wave: how Europe can gain from the rising tide of scientific data’ – https://ec.europa.eu/digital-single-market/en/news/digital-agenda-unlock-full-value-scientific-data-high-level-group-presents-report

[3] Knowledge Exchange, 2011, ‘A surfboard for riding the wave: towards a four country action programme on research data’ – https://repository.jisc.ac.uk/6200/1/KE_Surfboard_Riding_the_Wave_Screen.pdf

6. These analyses provide the rationale for a federated approach to this infrastructure, so that both data generators and data users can easily make use of data services and support services. The report delves into how federalisation – the process of creating federal structures – is understood and defined by the different organisations that have taken part in the study. Jisc, for example, encapsulates the challenge neatly when it describes its vision as "visible data, invisible infrastructure"[4]. To contribute to the implementation of this vision, Jisc itself is currently working on the creation of a pilot shared service to allow researchers and institutions to meet their policy requirements for the deposit and curation of research data[5].

7. In its 2011 report, KE referred to the different components of the research data infrastructure as an ecology, which might be summarised as the relationship between the elements of the infrastructure. Drawing from this, FRDIs could be characterised as an ecosystem, a complex network of interconnected and interdependent systems. Indeed, the recent first report from the EC High Level Expert Group on the EOSC talked about the need to develop a data infrastructure commons, that is, an ecosystem of infrastructures[6]. The current project will shed light on the nature and scope of this ecosystem.

Scope and methodology

8. The project was founded on detailed interviews with experts involved in the management and/or leadership of a range of federated data infrastructures in all six KE partner countries. These individuals were carefully identified and selected, drawing from KE's network of contacts across Europe. Sixteen people were interviewed in all: three from organisations in Denmark, France, Germany and the Netherlands, and two each from Finland and the UK. The organisations represented not only a spread of KE partner countries, but also of disciplinary areas; most were discipline specific, and four were multidisciplinary. Although the interviews tended to focus on national infrastructures, many of the discussions also addressed the international context and corresponding European or global initiatives, with which the national infrastructures are often aligned.

9. The interviews were semi-structured, lasted in average between an hour and an hour and a half, and were undertaken in two stages: a pilot phase, covering five organisations, and a second phase with the remaining eleven organisations. The pilot interviews were articulated around 24 detailed questions. Following a review undertaken at the end of the pilot, in February, the number of questions for the second phase of the project was reduced to nine. Both sets of questions can be found at **Annex A**. A summary of the relevant organisations, by country, is at **Table 1**.

10. With a single exception[7], interviews were audio-recorded. They were not transcribed verbatim; instead, notes were drawn up from the recordings. These notes in turn formed the basis of coding, which picked out the themes and issues that are described and analysed in this report.

Footnotes

[4] Jisc, 2016, 'Towards an integrated UK national research data infrastructure, presentation to a Jisc seminar at the Science & Innovation 2016 Conference – **slideshare.net/JiscRDM/towards-an-integrated-uk-national-research-data-infrastructure**

[5] Jisc research data shared service – **jisc.ac.uk/rd/projects/research-data-shared-service**

[6] European Commission, High Level Expert Group on the European Open Science Cloud, 2016, 'Realising the European Open Science Cloud' – **https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf**

[7] In the case of DNA Bank Network, in Germany, it was not possible to secure an interview in a timely fashion; instead, and exceptionally, the individual representing the organisation provided written answers to the nine questions that were submitted to her.

Table 1 – Organisations that were the subject of the study

Country	Organisation	Disciplinary area	Study interview phase
Denmark	Computerome	Life sciences	Phase 2
Denmark	DeIC – Danish e-Infrastructure Cooperation	Multidisciplinary	Pilot
Denmark	KOR – Danish Advisory Board on Register Based Research	Multidisciplinary	Phase 2
Finland	FIN-CLARIN	Language studies	Phase 2
Finland	Envibase initiative from SYKE – Finnish Environment Institute	Environmental sciences	Pilot
France	CDS – Strasbourg Astronomical Data Center	Astronomy	Phase 2
France	CDSP/DIME-SHS – Socio-political Data Services	Social sciences	Phase 2
France	HumaNum – Digital Humanities Infrastructure	Humanities & social sciences	Phase 2
Germany	GeRDI – Generic Research Data Infrastructure	Multidisciplinary	Phase 2
Germany	GFBio – German Federation for Biological Data	Biological sciences	Pilot
Germany	DNA Bank Network/GGBN – Global Genome Biodiversity Network	Biodiversity	Phase 2
Netherlands	Nikhef – National Institute for Subatomic Physics	High-energy physics	Pilot
Netherlands	NLBIF – Netherlands Biodiversity Information Facility	Biodiversity	Phase 2
Netherlands	SURFsara	Multidisciplinary	Phase 2
UK	ADRN – Administrative Data Research Network	Social sciences	Pilot
UK	Farr Institute of Health Informatics Research	Health sciences	Phase 2

Description and analysis

Contexts

11. The report does not purport to be a comprehensive study of all federated research data infrastructures across Europe. Rather, it is a snapshot, framed to be as representative as possible of strategies, approaches and practices in the six KE partner countries. As outlined in **Table 1**, a wide range of disciplines are covered, and this in itself represents the first lesson of the investigation: federated infrastructures, however they are defined – the possible definitions of 'federated' are considered below – can apply to all or most research disciplines. Of the 16 organisations covered two are in the physical sciences, three in the life/health sciences, three in the environmental sciences, two in the social sciences, two in the humanities and four are multidisciplinary. Federalisation is therefore not restricted to particular disciplines, although different research cultures are liable to have an impact on the way that respective infrastructures evolve.

12. A common feature of the 16 infrastructures is that they are facilitators for a range of services that help researchers get the most out of distributed data sources. Their broad aim is therefore to serve as portals that enable researchers, as easily and intuitively as possible, to work with datasets that might otherwise be difficult to access; in this vein, SURFsara was described in its interview as a knowledge broker – perhaps a useful way of describing FRDIs in general. The infrastructures may variously assist not just with access, but also with authentication, authorisation, identification (AAI), linkage, pooling, sharing, interoperability, security, standards, addressing legal/ethical issues, openness/publication, storage and more broadly with research data management (RDM). They provide a collaborative platform, an opportunity for researchers to work together and, where appropriate, to benefit from advice, support and sometimes training. In some cases, the range of services provided is quite broad. For instance, GeRDI explicitly seeks to support researchers throughout the successive stages of the research data lifecycle. In other instances, the service provided might be more focused, as in the case of data.deic.dk[8], the cloud-based storage infrastructure provided by DeIC. A second lesson is therefore that although FRDIs might all broadly be described as research data facilitators, they variously provide a range of distinct, albeit related services. There is no single model of service provision.

13. All 16 infrastructures play a national role in their respective countries, but several of them are also part of broad international networks and may serve as the national nodes of global initiatives. This is particularly the case in the following:

- FIN-CLARIN is the Finnish arm of CLARIN, a Europe-wide initiative on digital language resources
- CDS, in France, is integrated within the International Virtual Observatory Alliance (IVOA)
- the German DNA Bank Network is fully affiliated with the Global Genome Biodiversity Network, (GGBN)
- NLBIF, in the Netherlands, is part of the Global Biodiversity Information Facility (GBIF)
- Nikhef, also in the Netherlands, is affiliated to the Worldwide Large Hadron Collider Computing Grid (WLCG)

FRDIs, even when their focus is national, are more often than not closely aligned with international partners.

Footnotes

[8] https://www.deic.dk/en/data_deic_dk

What is understood by 'federated'?

14. There is no single, overarching definition of 'federated' that emerges from the study. The federalised nature of infrastructures is envisaged in different ways, often even within each of the six countries. Indeed, in the case of KOR, in Denmark, the interviewee felt that his organisation could not even be described as federated.

15. Some of the suggested definitions are very generic: for instance, 'federated' could mean independent but co-operating entities [FIN-CLARIN], or non-centralised organisations, made up of parts plus a coordination hub [ADRN]. But there are more specific perceptions as well, with an emphasis on providing an organisational or coordinatory layer which enables ease of access through a single point of entry, or ease of data management in general, with common standards and the complexity of underlying structures largely hidden; the coordinatory layer is the only component of the infrastructure visible to users [Computerome, DeIC, CDS, GeRDI, Nikhef]. The definition and application of common data standards, whether national or international, could also form part of the coordinatory layer [CDS]. For SYKE, federalisation implies deploying the means of allowing different stakeholders to make effective use of resources and to have free use of data from different sources, with a strong emphasis on openness. HumaNum sees federalisation in terms of the geographic dispersal of individuals, through a network of humanities and social sciences (HSS) data experts, gathered in sub-disciplinary consortia, and scattered across French regions in a variety of research institutions. For SURFsara, it is partly about data sharing, with individual institutions gaining access to resources in other institutions in a secure and trusted way. These definitions imply a degree of decentralisation or localised empowerment, and/or means for researchers easily to make use of resources that are heterogeneous and/or widely distributed, either nationally or internationally. DNA Bank Network suggested that decentralisation enhances engagement with users. In these terms, a federal approach might be seen as means of allowing researchers to find easy points of entry into complex, constantly-shifting infrastructure ecosystems; to put it in other terms, to cut through complexity.

16. Although no interviewee made explicit reference to the concept of subsidiarity, this was implicit in the view, expressed by DNA Bank Network, that efficiency dictates a division of labour between what is done at the central and decentralised levels: the central function provides a sharing platform, but it is the local level, where researchers actually work, that best ensures the quality of the data. At an even more fundamental level federalisation may be viewed, in Nikhef's eloquent terms, as an essential property of science, a necessary means of facilitating global research. Science is necessarily about collaboration, and federalisation offers scientists a ready-made framework so that they don't need to invent different collaborative tools for specific activities and purposes.

17. In some cases, 'federated' is perceived in its more conventional geographical sense. Thus ADRN's structure is federated across the devolved nations of the UK (England, Scotland, Wales, Northern Ireland), each hosting an autonomous centre; similar arrangements apply to the Farr Institute, although these do not cover Northern Ireland. DNA Bank Network also pointed to the geographical organisation of science and education, with the respective responsibilities of the national federal level and the regional entities, which themselves set their policies and priorities on issues such as research data storage. In the view of both ADRN and DNA Bank Network, federalisation is thus closely tied with political imperatives.

18. Finally, federalisation is also associated with the respective responsibilities of different governance structures or stakeholders, either nationally or internationally [NLBIF, SURFsara].

19. Clearly, there are differences in what is understood by federalisation. And sometimes, as reported by GeRDI, the concept is not well understood at all, either by researchers or among research data centres. But there is a common thread whereby, broadly, a federal approach is founded on the relationship between a coordinatory level and distributed or devolved, autonomous hubs. Ease of access to, or use of, distributed resources also features prominently in some of the definitions.

Drivers: introduction

20. Two sets of drivers influence the emergence and development of federated infrastructures. On one side, there are push factors, determined top-down by stakeholders such as funders, infrastructure providers, government and other public agencies, and also by the general social and political context in given countries. On the other side, drivers reflect demand from users - that is, from the research community itself, in a bottom-up manner. The study revealed that both sets of factors are important determinants for almost all of the investigated FRDIs.

Drivers: push factors

Social and political context

21. The most overarching push factors are those that reflect the societal and political context, and the cultural environment in which particular policies or initiatives can take hold, tapping into broad societal needs. The Computerome interviewee described the FRDI agenda as being driven by "real politics". From GeRDI's perspective, the drive towards federated infrastructures is seen in terms of the social responsibility of science and scientists. Science is not just self-referential: excellent, peer-reviewed science is no longer in itself sufficient. It is increasingly important for science to democratise itself by demonstrating its societal benefits and impact. This puts healthy new pressures on scientists, and obliges them to set out new narratives to satisfy public curiosity, explain to society what they are doing and promote public understanding of science. There is thus an imperative to address societal challenges, which implies cross-disciplinary working [GeRDI]. In Finland, the aims of the Open Science and Research Initiative (**para 26**) include ensuring that the possibilities of open science are widely utilised in society; promoting the trustworthiness of science and research; and increasing the societal and social impressiveness of research and science. Political will, legislative amendments and international development (such as EU policies) strongly support this trend and place new kinds of demands on the management of data [SYKE].

22. At a more disciplinary level, the politically driven agenda of climate change, with its socio-economic ramifications, has acted as a driver to integration, nationally and internationally. As part of this agenda, biodiversity has thus become a major current topic [GFBio]. In the biomedical field, one federalising driver is the development of personalised medicine, an area of research characterised by the use and merging of different data types, including registry data from different sources. This necessitates many technical developments to access data in parallel, and also imposes significant legal and organisational challenges. In Denmark, the political agenda behind personalised medicine is a strong driving force [KOR].

23. Finally, a different sort of political consideration has influenced the development of ADRN and the Farr Institute in the UK. There, the geographically federated nature of these two initiatives is determined largely by the existence of devolved governments in

England, Scotland, Wales and Northern Ireland[9], each of which has a degree of autonomy when setting its research priorities. It is noteworthy that in the three smaller countries of the UK (ie Scotland, Wales and Northern Ireland), progress has been faster owing to smaller populations, existing countrywide data infrastructures and policies to build on, and existing relationships with devolved government departments as data controllers.

National public policy

24. Governments and public agencies, often reflecting the societal and political factors outlined above, provide another set of drivers for federated infrastructures. These drivers emanate from national policies on research data, including open data, with multi-stakeholder engagement on the part of leading public agencies and other key national players. The interviews suggest that the situations in Denmark, Finland and the UK are especially marked by such overarching policy initiatives.

25. Two years ago, Denmark launched a national strategy on RDM, which includes policies for handling data at the national level. Federated approaches are being developed as part of this, and there are pilot projects in different sectors. data.deik.dk is the most notable of these, with a focus on cloud data storage [DeIC].

26. In Finland, much of the impetus has come from a drive to encourage open data. The National Research Data Initiative (TTA), which ran from 2011 to 2013, was a broad-based cooperative network for the development of research data services and the promotion of open knowledge and interoperability. As a result of the initiative, a centralised research data infrastructure using research data enterprise architecture and metadata models was developed. This was succeeded in 2014 by the still current Open Science and Research Initiative (ATT), which is based on cooperation between several key players, including government ministries, universities, research institutions, research funders (such as the Academy of Finland and Tekes – the Finnish Funding Agency for Innovation), Finnish Social Science Data Archive (FSD), the National Library of Finland and the Federation of Finnish Learned Societies [SYKE].

Footnotes

[9] There is a caveat in that, unlike ADRN, the Farr Institute does not run a centre in Northern Ireland (however, it has two centres in England, based in London and Manchester, as well as in Scotland and Wales).

27. Although open data in Finland is a powerful underlying influence on the development of the national research data environment, there is no specific national policy or strategy for federated data infrastructures. However, a strong policy is in place to create interoperable national services, with national guidelines that provide information management guidelines for public administration.

28. In France, the research infrastructure roadmap, updated in 2016, includes elements relating to the management of research data, to which all infrastructures wishing to subscribe to the roadmap should adhere [CDS].

29. In the UK, the national context is provided by the government's Open Data White Paper[10], published in 2012, which stressed the importance of ensuring effective access to national data collections for research and policy development. The Digital Economy Act, which became law in April 2017, includes clauses on data sharing for research purposes. Separate legislation and regulations relate more specifically to the use of health data (the Care Act and the Caldicott review of data security, consent and opt-outs in the health and social care system). The government's financial allocation to the science and research budget, in 2012, comprised funding earmarked for innovative technologies, including big data – a fraction of which went to finance the setting up of ADRN. But, as explained below, (**para 77**), public agencies can act as brakes as well as drivers: the risk-aversion demonstrated by the UK's National Health Service (NHS) towards the sharing of sensitive, patient-derived data can make the work of federated infrastructures more difficult [ADRN, Farr Institute].

30. Finally, at the international level, the European Commission's endorsement of ERICs (European Research Infrastructure Consortiums)[11] may have also constituted a political push factor [FIN-CLARIN].

Research infrastructure

31. In the hierarchy of push factors, the next level down relates to drivers from research funders, universities and infrastructure providers, which to an extent take their cue from the frameworks and guidance set by national public policy.

32. In the Netherlands, the main driving forces behind federated infrastructures are research institutes, ICT providers and funders. Thus Dutch funders are starting to include federated RDM requirements in their calls. The Dutch National Research Infrastructure Roadmap (which is related to the ESFRI roadmap) has issued a call that, for the first time, includes a requirement for research infrastructures to provide research ICTs. SURF is now in dialogue with the main Dutch funder, NWO, to include similar requirements in most relevant calls [SURFsara]. We have seen above how, in Finland, the Open Science and Research Initiative brings together universities and funders, along with other public agencies.

33. The Dutch e-infrastructure environment is characterised by a complex web of relationships between key players, with SURF (and more particularly SURFsara) providing some of the coordinatory 'glue'. The Netherlands, with its traditional collaborative research culture, has favoured bottom-up approaches – and perhaps inevitably, this means multiple lines of accountability, no single overlay, no definite hierarchy, no firm overarching governance and a federative environment made up of interlocking structures that create a fragile equilibrium. One of the manifestations of this informality is that although Nikhef was set up to support the high energy physics community, it has evolved into a service geared to researchers across a much wider range of disciplines. Nevertheless, even in the case of the Netherlands, there are some imperatives that are determined through national policy. Thus the resources pledged to the

Worldwide Large Hadron Collider Computing Grid (WLCG), come from the Dutch national e-infrastructure, rather than domain-specific initiatives; and the generic (ie not focused on a specific discipline) nature of the infrastructure conditions the contractual terms of the Dutch relationship with WLCG.

34. There are many research data centres in Germany, and they tend to be disciplinary. But in the light of moves towards EOSC and the impetus provided by the German Council for Information Infrastructures (RfII)[12], there is now an expert-derived consensus about the need to develop an ecosystem of data infrastructures, to overcome data silos and ensure improved interconnections between existing infrastructures. Two initiatives were therefore established in Germany at the end of 2016, covering different parts of the scientific community, and both bringing together key stakeholders in the research infrastructure:

› The Helmholtz Data Federation (HDF), a consortium of six partner high-performance computing centres plus a community with a focus on big data which, because of its scale, is not easily transferable over the internet[13]

› The GeRDI project itself, targeted at the long tail of smaller-scale science, bringing together five partners from the worlds of high-performance computing, librarianship, academia and data networking

These are the two building blocks forming part of the German contribution to the emerging federated infrastructure environment in Europe [GeRDI].

35. In the UK, the Administrative Data Taskforce, with representation from the main relevant public funding agencies (Economic and Social Research Council (ESRC), Medical Research Council (MRC) and Wellcome Trust), worked between 2010 and 2012 to set out the challenges and opportunities for accessing administrative data in research. This work provided a foundation for the creation of ADRN as a federated entity. In addition, the Data Forum, sponsored by ESRC, published its five year plan[14] which set out a vision of how high quality empirical research in social sciences should rely on high quality data stemming from a wide variety of sources, and on associated infrastructures.

36. A final driver under this heading is simply the financial savings attributed to federalisation. In astronomy, characterised by large, expensive infrastructures such as telescopes, it is more cost-effective to develop mechanisms to reuse data than always to gather it afresh from observations. The incentives here are about saving public money and optimising investment in the infrastructures, in addition to enabling additional science by reuse of data [CDS].

Footnotes

[10] https://data.gov.uk/sites/default/files/Open_data_White_Paper.pdf

[11] https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=eric

[12] http://www.rfii.de/en/home. In 2016, RfII published its recommendation on ‘Performance through Diversity’, which recommended the establishment of a Nationale Forschungsdateninfrastruktur (National Research Data Infrastructure, or NFDI), to serve as the backbone for research data management in Germany – http://rfii.de/?wpdmdl=2075

[13] http://fz-juelich.de/SharedDocs/Meldungen/IAS/JSC/EN/2017/2017-03-hdf.html?nn=373302

[14] ‘UK Strategy for Data Resources for Social and Economic Research, 2013-2018’ – http://esrc.ac.uk/files/research/uk-strategy-for-data-resources-for-social-and-economic-research/

Other players

37. It is worth noting the possible role of other players in driving for federated infrastructures. Relevant cross-disciplinary initiatives include the Research Data Alliance, the International Council for Science's (ICSU's) World Data System (WDS) and the Committee on Data for Science and Technology (CODATA). In biomedicine, KOR suggested that drivers include advances in clinical data for health research, influenced by the requirements of large Danish corporations – for instance, the global pharmaceutical companies Lundbeck and Novo Nordisk – which have a commercial interest in acquiring different types of data from sources such as the Program for Clinical Research Infrastructure (PROCRIN)[15]. Publishers are another relevant set of commercial players and they may also act as drivers: in molecular biology, they require the deposit of molecular sequence data in gene banks; in astronomy, data attached to publications are also made available by CDS, in collaboration with academic journals.

Weight of different drivers

38. Nikhef posed the question about whether the difference in driving forces, and/or national set-ups, is a determinant of success for federated infrastructures. Any given driver is not necessarily a conclusive success factor, but it certainly helps if an organised community is willing to invest in a joint, federated infrastructure. But at the same time, this is not invariably necessary because some global initiatives that do not have contributory mechanisms or are less strongly coupled have also succeeded, even where computing was not a primary concern or inherent to given communities. For NLBIF, operating in the area of biodiversity, there are two sets of further players with distinct, albeit slightly overlapping, needs: non-governmental organisations (NGOs), which are important data collectors, and museums, which are opening up their collections through digitisation and through the depositing of their data with GBIF. Museums in the Netherlands are a major driver; they run their own scientific infrastructures and look to be important partners within European platforms.

Drivers: demand – meeting user needs

39. Drivers also reflect the needs of users – members of the research community, not just in individual countries but globally. The CDS interviewee stressed the fundamental importance of representing users' views and of allowing researchers to play an active role in data centres. As with push factors, most of the interviews reported that demand from users, in one form or other, plays a role in the development of federated infrastructures. However, whereas push factors can often be identified through fairly specific developments or initiatives led by well-defined players, demand is less easy to define; it is more diffuse and less tangible, and therefore more difficult to quantify.

Relationship between push and demand

40. In the first instance, there is a crossover between push and demand, where infrastructures show how they acquire an understanding of user needs. Interviewees provided examples from their respective countries:

- In Denmark, the rationale behind the federated infrastructure across Abacus and Computerome is a capacity to identify and meet the needs of users who approach the services, and to orient them towards either of the two services. This model works satisfactorily for them, as it enables the infrastructure providers to accommodate researchers on the basis of their different needs [Computerome]

- DeIC is designed and developed by the national e-infrastructure organisation, based on accumulated knowledge of researcher needs from different sectors. It is a new initiative, but initial user feedback is good. Pilot schemes are currently being implemented with active engagement from a number of research groups [DeIC]

- Also in Denmark, KOR feels that it responds to demand from researchers, and to a growing research culture that reflects the understanding that access to data matters and is important for research. KOR is in constant dialogue with researchers about what sort of access is needed, and what it should focus on. There is therefore an inbuilt, organisational desire to respond to the wishes of the community [KOR]

- GeRDI's first work package was to develop use cases, showing how an FRDI could work; the communities are currently being consulted via questionnaires to ascertain their needs and demands. GeRDI is also working on a demo system (which should be ready in the second project year), showing the basic FRDI functionalities as part of a drive to explain a concept that is not always well understood

- A reflection on needs is important for HumaNum too: its experts, embedded in sub-disciplinary consortia (**para 14**), provide advice and guidance, and importantly, work with their hosts on jointly elaborated solutions

41. Importantly, infrastructures need constantly to adapt and to reflect the evolving nature of research. In Germany, there is a requirement for relevant, research-intensive science organisations such as Helmholtz to run dynamic services that are aligned with researchers' needs [GFBio]. Such dynamism is significant: academic e-infrastructures are characterised by innovation, which is associated with instability. In that sense, they are not designed to provide professional services to ISO standards. Academic services must continually adapt to researchers' needs – so even if the infrastructure is a little shaky and occasionally fails, adaptability to needs, and the benefits of having access to leading edge services, are more important [Nikhef].

Cultural factors

42. Demand is affected by the different research cultures in the various communities, with some (for instance, high energy physics and astronomy) being very collaborative with regards to RDM. In high energy physics, arguably one of the most successful aspects of the international infrastructure has been the federated aspects of bringing users together; it is the trust infrastructure that has been most appreciated by members of that community. Global data sharing, without limitation to certain communities, is deeply rooted in the culture of astronomy and its multipolar organisation. Other disciplines, such as the life sciences, are in the process of developing their collaborative methodologies. So, the better organised research communities are either demanding federative approaches, or starting to demand them [SURFsara, Nikhef, CDS].

43. Specific data practices may also reflect research cultures. Astronomy is a science that has long depended on a capacity to draw and interpret data from distributed sources, such as instruments and telescopes. Given the heterogeneity of such sources, it follows that interoperability lies at the heart of astronomy research. There is therefore a culture-driven imperative to favour infrastructures that allow for interoperability and for the development of the International Virtual Observatory Alliance (IVOA) as the host of interoperability standard definition [CDS].

Footnotes

[15] http://kea.au.dk/en/en/PROCRIN.html

44. HumaNum highlighted the variable nature of demand from within the HSS community. Those researchers who are used to working collaboratively are more likely to be open to making use of research data methodologies; for such groups, there is a need and demand for HumaNum services. But for cultural reasons, including resistance to e-research, other researchers feel less comfortable with the sort of tools on offer; for those categories, the level of demand is correspondingly lower. HumaNum sees part of its role as nurturing and effecting the cultural acclimatisation that would help researchers understand the benefits of good research data practice across the entire HSS spectrum.

45. The Nikhef interviewee elaborated on how long-term shifts in research culture can impact on needs and demand. In this instance, the evolution of the international infrastructure over a period of several decades has been conditioned by the culture and practices of the high energy physics community as it develops towards maturity. This is a discipline in which, over a long period of time, researchers have come to realise the importance of computing as an integral part of experimental design and research support, and this has allowed for a formal structure within which computing capability can develop – a key underpinning factor for the evolution of infrastructures that reflect user needs. According to the Nikhef interviewee – and from the perspective of high energy physics – such integration of computing and research is less the case for other disciplines such as astronomy, where the approach is more fragmented and less coordinated (although conversely, according to the CDS interviewee, access to data is more open and integrated in astronomy than in particle physics – **paras 42** and **43**). It is thus important that there should be a convergence between the people building the instruments / data centres and those doing the analysis.

46. In a similar sense, a culture characterised by increasingly international working may also be a driver. There might be limited federalisation at the national level but a much greater degree of it internationally, with researchers therefore getting accustomed to the benefits of federated approaches. International practices can therefore encourage a trend towards supporting federalisation at the national level [KOR]. In astronomy, international dialogue and working practices are long-established features of the research culture. From its inception in the 1970s, CDS was given an explicitly international remit, and half of its scientific board are non-French scientists. An international division of labour is important, because that might be the only way of ensuring a critical mass of expertise in given areas.

Getting the right balance

47. What about achieving a balance between meeting needs, but not interfering overbearingly in researchers' daily practices? Infrastructures need to steer a path between providing too strong a steer for researchers and presenting them with data without supporting them. There should be an imperative to make it easy for researchers to use services, but in a way that's not too prescriptive. Although infrastructures should improve the efficiency and quality of research, their purpose is not to answer researchers' research questions. Their role lies somewhere in the middle: neither completely detached, nor completely involved [FIN-CLARIN].

48. A related question is the extent of the service that federated infrastructures should aim to provide, because there are limits to their capacity to meet researchers' expectations. As noted in the description of possible meanings of federalisation (**paragraph 14**), federated infrastructures are often understood to be about ease of access to research data, with minimum fuss. There is a debate in Denmark about extending the scope and breadth

of DeIC as a national service to meet researcher needs, and determining how many discrete activities to develop as part of the service. However, it would be problematic if each researcher were to expect his/her own personalised service within a federated structure [DeIC]. In Germany, GFBio's mandate is determined by the service profiles that institutional research partners must agree with it. Memorandums of Understanding (MOUs) have been drawn up to set out needs and expectations. But not all of the services offered are useful for a federated infrastructure (for instance, they may be important for individual projects, but not on a broader scale), so streamlining may be required; prioritisation and filtering out may be needed to define the scope of a core service and to establish a corresponding business model. This represents a significant current challenge, more so than technical issues [GFBio].

49. A further balance to achieve is between the views of stakeholders who have a short-term interest in given activities (such as those involved in a time-limited project) and those with a longer-term stake and a greater interest in sustainability. For GFBio, the proportion of partners who take such a longer-term view may be as high as 80%, which demonstrates a clear understanding of, and a strong commitment to, the aims of the infrastructure – a good driver for the whole initiative [GFBio]. The relationship between short- and long-term imperatives is also explored below (**para 51**).

Operation of the infrastructures

50. This part of the report addresses the operations of FRDIs, particularly their financing, governance and the way that they engage with users and stakeholders.

Financing and funding

51. To a significant extent, for funding purposes, the various federated infrastructures are treated as projects – that is, the funding streams are for limited periods, often associated with set-up costs, with no guarantee of longer-term support. For example:

- Current funding for FIN-CLARIN comes from the Academy of Finland. But the academy's mandate covers only the development of new services, not the support of existing services. In principle, the Ministry of Culture and Education could also provide funding, but unlike the academy, it is not currently equipped to evaluate FIN-CLARIN; an evaluation mechanism would need to be set up for this funding to be possible

- For GeRDI, funding is for a three-year period. Beyond that, further funding will be requested, as overall objectives are too ambitious to be achieved over three years. The second phase is intended to focus on the rolling out of the technology, rather than the basic development aspects

- For DNA Bank Network, current financial support is secured through a business model that supports an international secretary (GGBI); however, further network extension and development demands long-term technical support

- For ADRN, the funding provision is more ambitious, and covers a five-year period

52. However, this project-based approach isn't universal. Thus data.deic.dk was conceived originally as a traditional development project, but recently restructured to transform it into an operational organisation.

53. It follows that financial sustainability is an issue for FRDIs, and for the stewardship of research data in general, as testified by the OECD/Global Science Forum (GSF) Project on Sustainable Business Models for Data Repositories[16]. Short-termism, characterised by a succession of small projects, is inefficient [DeIC]. The question of sustainability should be addressed from the outset of initiatives – and this is important not just for national infrastructures, but for international ones too, including EOSC [GeRDI]. In Germany, DFG is looking to develop funding models to fund general infrastructure, beyond individual research projects. Over the next few years, data infrastructures will have to account for a larger proportion of research budgets. The federal and Länder governments aim to provide these budgets for a German National Research Data Infrastructure (NFDI) as proposed by the RfII (**para 34**). Ultimately, sustainability may be dependent on multiple funding sources [GFBio]. One possible way forward is to seek longer-term funding through user subscriptions, and this is being envisaged by DeIC and SYKE, for its Envibase initiative. Federalisation need not necessarily lead to increased costs; SYKE's view is that a federated infrastructure could even lead to cost savings, mirroring the view expressed in the context of astronomy (**para 36**).

Governance

54. Most interviewees did not elaborate on their infrastructures' governance arrangements. For those that did, there was an emphasis on multiple levels of accountability. This is perhaps a reflection of the complexity of federated infrastructures. For instance, GFBio's governance functions through a variety of complementary structures including a steering committee, a strategic advisory board and a general assembly; this is essentially a democratic structure, but there are some top-down approaches too, where broad-based discussion (such as what happens in the general assembly) is complemented by expert views from the advisory board. In the UK, ADRN runs a similarly multi-headed governance structure. There is internal governance within each of its four centres, in England, Scotland, Wales and Northern Ireland. These four centres, plus the Administrative Data Service, are overseen by an ADRN board, a directors' group that offers strategic direction, and an operations group, made up of project and user services managers.

55. Governance arrangements may also be envisaged at the international level. In high energy physics, the Worldwide Large Hadron Collider Computing Grid (WLCG) has a top-level governing body and a management board, which allows the community to take a collective decision on developing WLCG. There is also a wide board, with representation from every participating country, which delegates some of its powers to an executive board. The management board also oversees the executive directors – both of which have real powers, based on an MOU framework that extends globally [Nikhef].

Involving users

56. The governance arrangements described above are designed to seek input from users. They complement the means deployed by infrastructures to understand or evaluate user needs, as described above (**para 40**). In Denmark, the KOR secretariat actively facilitates a two-way dialogue between researchers and registry owners. An example is the recurring requests from researchers to adjust registry data access such that it is easier to combine with other types of data. The dialogue also addresses the utilisation of access data, including AAI infrastructure issues and legal issues. KOR also helps to define use cases, in situations where researchers wish to add types of data other than registry data; it facilitates a dialogue with registry owners about how best to go about that.

Partnerships and stakeholder involvement

57. Typically, federated infrastructures work in close partnership with universities, research institutes, data centres, other data infrastructures, research funders and government departments as well as with individual researchers and research groups. Such partnerships may manifest themselves through the formal consortial arrangements which characterise some of the infrastructures, for instance:

› FIN-CLARIN is a consortium made up of 11 partners, mostly universities, but also including a research institute and an IT centre

› We have already noted how, in Finland, the Open Science and Research Initiative (ATT) brings together a wide range of partners from different sectors (**para 26**)

› Each of the four ADRN centres is made up of a partnership between universities and other players such as government (in Scotland and Wales) and statistical agencies

› In the Netherlands, SURFsara is part of a consortium, Research Data Netherlands, which also brings together DANS, an (inter)national data archiving service, and 4TU.ResearchData, the data sharing and reuse collaboration run by the four Dutch technical universities

58. HumaNum described the particular workings of its relationship with relevant research funders. These now systematically direct research teams to HumaNum, so as to help funding applicants address research data challenges through the use of HumaNum tools and services. There is thus an understanding among funders about the imperatives of obtaining advice and support on research data issues in order for researchers to formulate persuasive research proposals.

Footnotes

[16] http://codata.org/working-groups/oecd-gsf-sustainable-business-models

59. Some infrastructures are involved in partnerships that reflect their particular disciplinary settings. The registry initiatives that have a bearing on health research, KOR and the Farr Institute, have close relationships with health agencies in their respective countries: the Danish Health Data Agency (Sundhedsdatastyrelsen) and the UK's National Health Service (NHS). KOR also works with the Danish Statistical Bureau, with regards to the socio-economic data that is collected under the auspices of the Economy Ministry. As noted above, statistical agencies are also among the partners of ADRN centres: the Office for National Statistics (ONS), National Records of Scotland and the Northern Ireland Statistics and Research Agency. In astronomy, CDS is involved in IVOA and works in worldwide partnership with agencies in charge of telescopes, with other providers of added-value data services and with academic journals. In fields of environmental sciences and biodiversity, museums are important partners. NLBIF has partnerships with the National Museum of Natural History in Leiden, and with a number of local museums; SYKE's partners include LUOMUS, the Finnish Museum of Natural History. NLBIF also works in partnership with NGOs which, as noted above (**para 38**), are important data collectors in their own right – many such NGOs work on data from single plant or animal species, but some also run platforms where relevant data gets aggregated.

60. Closely associated with the development of partnerships is the imperative to engage actively with stakeholders, and to deploy mechanisms that enable a dialogue to take place with them. The report has already touched on these issues with regards to users, under the headings of meeting user needs, governance and involving users.

61. The first step is often to bring stakeholders together as part of the formulation and early development of the infrastructure. As noted above (**para 25**), Denmark launched a national strategy on RDM a couple of years ago. The development of this has provided an opportunity for a range of key players – national archives, national library, university libraries and institutional chief information officers – to come together. Such multi-stakeholder dialogue has been slow and sometimes difficult, but it has allowed all relevant parties to contribute to a joint process and to broaden perspectives. On the basis of this effort, DeiC is now in a position to provide coherent solutions that will save time and effort for researchers. It can therefore be said that the engagement strategy itself was federated – an important lesson applicable to others [DeIC]. Strategies for stakeholder engagement might involve a formal process for identifying them, and GeRDI is currently defining precisely who the stakeholders are – essentially, infrastructure providers and user communities. But given the complexity of federated infrastructures, identifying stakeholders (and getting them interested) can be challenging, as suggested below (**para 94**).

62. Some infrastructures have formal mechanisms in place to provide channels of communication for engagement and to underpin outreach. Examples include:

- GeRDI's first work package was to develop use cases, demonstrating how a FRDI could work; the research communities are currently being consulted, via questionnaires, to ascertain their needs and demands. This is associated with the development of a demo system (which should be ready in the second project year), explaining basic FRDI functionalities. This is part of the outreach narrative mentioned below. Close cooperation between research communities and project partners (infrastructure providers) is guaranteed by GeRDI community managers, who enable an effective dialogue

- For its part, GFBio has recently instituted an outreach work package, which will be intensified over the next two years, with visits to institutions, participation in informal discussions and making presentations. Outreach material is being streamlined and expert staff members are being trained to ensure consistency of messages. There are also approaches to scientists who are known to be working on proposals, to get a feel for emerging projects. Also, in an effort to achieve some sort of common language, GFBio has started to draw up MOUs between itself and its project partners, to chart respective responsibilities and longer-term prospects; to agree the service profiles of partners, and to get long term commitments for supplied services

- In the UK, each ADRN centre has a communications team, whose target communities are academic researchers and government analysts. And ADRN as a whole has a communications plan and strategy in place

- In Finland, FIN-CLARIN seeks to be relevant to the needs of researchers by informing them about its services, notably through its website and a programme of site visits to universities, which are deemed to be fruitful

63. Engagement with the wider public can be an important and valuable aspect of FRDI outreach. The Farr Institute interviewee emphasised the important of such interactions when dealing with sensitive, patient-centred data. Public engagement is mandatory (as is the case for any project of this sort in the UK), through Patient and Public Interaction (PPI) strategies. The Farr Institute runs a PPI Working Group, with representatives from all four centres, which works closely with patient groups both nationally and locally. The Farr Institute Health e-Research Centre (HeRC)[17] in Manchester runs a Citizens' Juries initiative, where groups of 12 people provide their views on particular projects. HeRC also runs its #datasaveslives campaign that gets messages out to the public and to health professionals about the benefits of research use of health data. There are often patient representatives on Farr Institute governance structures, emphasising the importance of understanding people's views about the use of their data. Once such a dialogue takes place, participants are very open about making their data available.

64. Citizen science is another manifestation of engagement with the public. In the context of SYKE and Envibase, citizens' observations are used in environmental monitoring. They are also relevant in astronomy, through online data gathering tools such as Galaxy Zoo[18]. Envibase, through the importance that it attributes to openness, stresses a distribution of work that involves cooperation with citizens and companies.

Footnotes

[17] https://www.herc.ac.uk
[18] https://www.galaxyzoo.org

Data-related practices

65. The study uncovered a wide range of features and data-related practices associated with research data and the operations of federated infrastructures. These factors are obvious components of any research data environment, and therefore feature naturally in federated environments. In most if not all cases, federalisation enhances capacity to address or improve such practices – although in some cases, it also points to problems or blockages. The interviews highlighted some (but far from all) individual practices in the case of each FRDI.

Research data management

66. Research Data Management is perhaps the most holistic of the identified data-related practices, because it relates to the overall relationship that researchers have with data. But curiously, given its ubiquity as an overarching research data concept, it featured explicitly in only a small number of cases. The most expansive references to RDM were provided by GeRDI and HumaNum:

- GeRDI's service provision is framed around the research data lifecycle elaborated by the UK Data Archive; GeRDI thus provides functionalities associated successively with searching for and finding research data, processing it, analysing it and deriving new data from the analysis

- HumaNum promotes a generic workflow for all HSS, based on its own take on the data lifecycle, and articulated around data storage, management, sharing, analysing, reusability/interoperability, linking to scholarly publications and achieving visibility, for instance through search engine optimisation. HumaNum deploys a range of tools for many of those functions, often designed jointly with user communities within sub-disciplinary consortia. It recognises the value of raising awareness about the importance of good RDM, and researchers often appreciate the benefits that this confers

Among other infrastructures, GFBio wishes to be part of international drives to ensure that RDM is priced into research and recognised as an integral part of scholarship. SURFsara is piloting the use of the iRODS software suite[19] to establish proof of concepts for RDM.

Access

67. As outlined above (**para 14**), one of the key characteristics of FRDIs is to provide an organisational layer that enables ease of access to research data. The capacity to gain intuitive, transparent access not just to data but also to tools, is a recurring feature for most of the infrastructures covered in the study. The ease with which data may be accessed is especially important: FRDIs can provide a single point of access from which researchers can define the characteristics of the data that they are looking for, irrespective of where it is located, without needing to know where it is located and without the inconvenience of contacting individual data centres [GeRDI, Nikhef]. In the long term, the level of easy access to sophisticated data will be such that, in some disciplines such as archaeology and biology, it will be less necessary to go out into the field to test hypotheses and/or to monitor phenomena (although other fields, such as astronomy, continue to rely heavily on observational data, for instance obtained from telescopes); instead, it could be done primarily through virtual, global, shared spaces – data contained in a variety of clouds which researchers can tap into once the initial data is stored. The growing incidence of federated infrastructures will facilitate this [GFBio].

68. Where data is of a sensitive and/or confidential nature, providing facilities, support and advice for accessing resources in a safe and controlled way is of paramount importance; this is especially the case for the registry data whose access is mediated by KOR, ADRN and the Farr Institute. Thus registries may employ support officers, who help researchers gain access to relevant data and provide both administrative and technical support, and who also ensure that access complies with relevant data protection legislation [KOR].

AAI – authentication, authorisation, identification

69. Identification and authentication of authors are facilitated by initiatives such as ORCID, which can be integrated in FRDIs [GFBio]. Mechanisms such as the AARC Blueprint Architecture[20] facilitate the transition to a model where researchers gain access to resources based on their well-known credentials [Nikhef]. Verification of user identity, and a unified approach to the processing of requests and the granting of permissions, can be an important prelude to gaining access to federated resources, and some interviewees [Computerome, KOR, the Farr Institute] referred to two-factor authentication; in the case of the Farr Institute, the sensitive nature of the relevant data necessitates tight controls prior to disclosure. Issues associated with AAI featured prominently in the interview with FIN-CLARIN. The interviewee pointed out that, while in theory federated login allows access to distributed resources internationally, in practice legal constraints and regulatory/policy requirements, which vary from country to country (and sometimes even region to region, within countries such as Germany, where legal frameworks vary between each of the Länder), can act as real bureaucratic barriers. Such situations complicate AAI processes, and can sometimes lead to considerable delays before researchers are granted access to particular resources. The forthcoming General Data Protection Regulation (GDPR)[21] could change this: a rigorous pan-European standard could foster harmonisation and deter individual countries from setting differential AAI standards.

70. There is a trend to push for mandatory user identification, but open systems can also work; the possibility should be left open not to require or implement AAI. When CDS removed the requirement for user identification for one of its services, the service usage increased very significantly [CDS].

Usability and interoperability

71. Interoperability and usability are important factors for federated infrastructures. Usability also incorporates factors such as efficiency. Efficiency in this sense allows easy access to very large numbers of different data sources, which are not harmonised and not machine-readable, each with their own formats, metadata structures and content standards. The challenge is brokering for integrated pools of data from different domains to be exploited without significant manpower effort. Interoperability may also apply, for instance to curatorial ontologies. GFBio deploys a terminology server that harvests ontology data from many ontologies worldwide, allowing for a common language and harmonised naming conventions.

Footnotes

[19] iRODS is an open source system which sits as a management overlay for data storage – https://irods.org/

[20] https://aarc-project.eu/architecture/

[21] The GDPR was approved by the European Union in April 2016, and comes into force in across EU member states in May 2018 – http://eugdpr.org/

Standards

72. Different data sources are characterised by different standards and ensuring their compatibility, or converging them, can be a major endeavour – although it is likely to be an organisational rather than a technical or computing challenge. As related by FIN-CLARIN, the CLARIN initiative addresses this through qualifying for the Data Seal of Approval[22], which is granted at the level of the individual repository, as well as through certification of its centres, valid for three years, after which re-certification is required. In this scenario, it doesn't matter that end-users don't necessarily understand the exact nature of certification; what's important is that acceptable service is provided, and that the different CLARIN centres across Europe can be reassured that each one adheres to recognisable standards, so that there is a common language for cooperation. In France, CDS also uses the Data Seal of Approval, although for slightly different reasons: in this instance, the rationale is to obtain an external evaluation that CDS is a trustworthy repository and as a check for its own procedures. Finally, the Data Seal of Approval is also used as a basis for certification by WDS, as part of procedures for accrediting its own members[23].

73. Appropriate standards are crucial in the context of complex regulatory frameworks relating to sensitive data, where there are obligations to ensure data protection and confidentiality. For instance, the Farr Institute's activities are subject not only to ethical approval (see below) but also must meet NHS Information Governance Toolkit[24] certification and, in most cases, ISO27001 accreditation for information security management systems[25]. For its part, ADRN offers an accreditation service to researchers ('Sure Training'), which needs to be renewed every couple of years. These are rigorous requirements.

74. Standards might apply internationally too. In the case of astronomy, a discipline characterised by a long-established culture of international collaboration, technical standards are agreed through the relevant global body, in this instance IVOA [CDS].

75. The FAIR Data Principles (making data Findable, Accessible, Interoperable, Re-usable)[26] provide a recognised gauge for demonstrating adherence to good research data practice. Demonstrable adherence to FAIR can help to ensure that data infrastructures reflect the requirements of different research disciplines. This might be especially challenging for institutional data repositories, which are not always attuned to disciplinary needs and which may also be content to adhere to the lesser requirements suggested by Dublin Core [CDS]. To help address this challenge, one of the pillars of the GO FAIR initiative (**para 114**) aims to instigate cultural change to make the FAIR principles a working standard in science. Besides FAIR, there are other initiatives describing fitness of use of data. These include the Group on Earth Observations (GEO) Data Sharing Principles[27] and the framework on data quality standards developed by the Federation of Earth Science Information Partners (ESIP)[28]. In addition, the WDS/Research Data Alliance (RDA) Working Group on Assessment of Data Fitness for Use[29] is also looking into developing solutions around assessment of data quality criteria [GFBio].

Security

76. As with AAI and standards, data security features prominently where data is of a sensitive nature; again, this is important for registry data, such as that serviced by KOR, ADRN and the Farr Institute. It is therefore important to secure the agreement of data owners – registry or non-registry data – and to reassure them about the security of the data and

the safeguarding of their clients, patients and of the population in general [KOR]. In the case of ADRN, the network's overarching aim is to enable safe access to linked, de-identified administrative data for public benefit. On that basis, ADRN has set up a state-of-the-art, safe physical and digital infrastructure, based on the 'five safes': safe projects, safe people, safe settings, safe outputs and safe data. The Farr Institute has worked with Jisc to develop the concept of 'Safe Share', which provides solutions for encrypting data as it travels from point to point through Janet, the UK's academic high speed network, and for decrypting once it reaches designated 'safe havens'.

Ethical and legal issues

77. Ethical and legal issues are closely associated with data protection, and have already been touched upon under the headings of AAI, standards and security. There is an imperative to ensure data protection and, in the case of health research, to preserve patient confidentiality. We have already seen, in particular, that meeting regulatory requirements can act as a significant and potentially frustrating brake on accessing and sharing data. Key stakeholders – public agencies, including health service managers – may be very risk-averse about providing access to the sensitive data that it collects and curates. It can sometimes take years or even decades, and much high level user engagement and negotiation with relevant agencies, to link research data and demonstrate that it can be managed securely [Farr Institute]. Moreover, data breaches can lead to significant penalties for research organisations.

78. The constant struggle to make data available in the face of data protection imperatives and ethical obligations reflects the difficulty of reconciling demand for access from the research community with the need to address the caution of public agencies. Technological approaches such as the provision of data safe havens are part of the solution, but even these solutions are not always agile, because of the information governance constraints. This can be frustrating for researchers who just need to get on and do their research.

79. A related issue, although one that relates potentially to all research data, is its legal status, as interpreted by data centre owners or national authorities, who may have requirements that data cannot leave their premises or country/region of origin [KOR, Farr Institute]. There is therefore a need for a dialogue with data centres, to address their fears that FRDIs could make them lose control of their data, lead to the undermining of IPR on data, or modifications of licensing terms. It is important to explain clearly that FRDIs are essentially mediators between data centres and the community, that IPR remains firmly vested with the data centres and, more broadly, to set out exactly where the respective responsibilities of FRDIs and data centres lie. One often unresolved question is what happens when new data is created from existing datasets that have different licensing terms, or different degrees of openness [GeRDI, NLBIF].

Footnotes

[22] datasealofapproval.org/en/
[23] icsu-wds.org/services/certification
[24] igt.hscic.gov.uk/
[25] https://en.wikipedia.org/wiki/ISO/IEC_27001:2005
[26] For an explanation of the rationale and applicability of the FAIR principles, see Wilkinson M.D. et al, 'The FAIR Guiding Principles for scientific data management and stewardship', Scientific Data, Article number: 160018 (2016) – nature.com/articles/sdata201618?utm_source=feedburner&utm_medium=feed&utm_campaign=Feed%3A+sdata%2Frss%2Fcurrent+(Scientific+Data)
[27] earthobservations.org/dswg.php
[28] http://wiki.esipfed.org/index.php/Information_Quality
[29] rd-alliance.org/groups/assessment-data-fitness-use

80. A similar difficulty, as flagged by CDSP, is the lack of awareness about IPR on the part of researchers themselves; they often fail to realise that they do not own the data that they create, and that ownership is vested with their employing institutions. Moreover, researchers who work with qualitative data, as in social sciences, often have a possessive attitude towards it, sometimes resulting from what they perceive as a special relationship with study participants.

81. Open cloud solutions might help where, as suggested above, there is a requirement to keep data in its place of origin. Crucially, this would work by taking the analytics to the data, running comparative analyses in different locations and amalgamating summary statistics and outputs – rather than shipping large, raw datasets with very stringent security requirements and geographical constraints [Farr Institute].

Sharing and linking

82. Along with providing access, allowing for sharing of data and tools is another major feature of federated infrastructures. Sharing can usefully take place across disciplinary boundaries. For instance, the methodologies developed for Computerome, which among other things handles personal medical data, can also be deployed in the digital humanities and the social sciences. HumaNum has found that resources developed for one particular HSS community can be adapted for other communities, and looks to drive such cross-boundary dissemination. There is even occasional demand for HumaNum tools from researchers beyond the fields of HSS. Federalisation can thus lead to cross-disciplinary fertilisation. And in addition, researchers might also use FRDI collaborative platforms to share experiences – this could be valuable for researchers who work in overlapping fields [Nikhef].

83. Sharing and linking can be valuable, too, for stakeholders from different backgrounds. The cross-disciplinary opportunities afforded by FRDIs, and the linkage or combination of different datasets, can lead to completely new, unexpected insights or research questions which would not otherwise have been possible [GeRDI]. The linking of different kinds of registry data, and also the linking of registry data with other sorts of very large datasets that are of long-term value in different areas of research, involve a growing degree of complexity and organisation which in themselves are drivers for higher degrees of coordination and federalisation [KOR]. In health research, there is a strong case for data sharing and linkage when investigating outliers or rare diseases, whose characteristics can only be properly investigated if data is compared and analysed on a global scale. For this reason, patients affected by such conditions (and their families) are often very strong advocates of data sharing [Farr Institute].

84. Sharing may also sometimes be approached with caution; as reported by DeIC, there is a desire from the research community in Denmark for a specifically national cloud service, keeping data within Danish jurisdiction, allowing for a more controlled data sharing cloud-based solution. Sharing can also be problematic in contexts where data is of a sensitive nature, notably where it is derived from citizens – typically in the case of registry data. In these instances, governments and public agencies, such as the NHS, may be reluctant to share the data that they generate [ADRN, Farr Institute]. The UK hopes to address such wariness through legislation, with a provision on data sharing for research purposes included in the recent Digital Economy Act (see also **para 29**); however, this will exclude health and social care data [ADRN].

85. The GFBio interviewee reflected on whether federalisation had fostered a sharing culture within the biodiversity community. He felt that, in practice, this was not the case because sharing pathways have not yet been incorporated into researchers' daily practices. Such pathways can only emerge over the next five to ten years through cultural shifts, enabled through workable RDM arrangements, proper funding, training and guidance.

Openness and publication

86. Data openness is deemed important for some of the interviewed infrastructures. In Finland, there is a strong political imperative across government to encourage open science for societal ends. This is driven by the Open Science and Research Initiative (**para 26**), bringing together diverse players. The premise of openness provides a context for federated infrastructures, so that the principles of free and open data enable the distribution of work and cooperation between different operators. Data is stored either where it is produced or where storage is most affordable, while open interfaces make the information accessible to everyone [SYKE]. As well as national factors, the predisposition of certain research cultures to openness may also influence FRDIs; in disciplines such as biological sciences and astronomy, the international environment encourages openness of research results and data outputs across the world – albeit on the assumption that partners in other countries are equally committed to the open exchange of knowledge [GFBio]. In astronomy, data is shared internationally. Data from telescopic observations is in general open after a proprietary period (often one year), during which it is reserved to the original observers, which facilitates acceptance of openness by the community. In the field of biodiversity, steps are being taken to develop GBIF into a data publication platform for researchers, and to encourage biodiversity researchers to use the facility. This is part of a general move to encourage data openness [NLBIF].

87. As with sharing, some players approach openness more cautiously. University researchers may feel less motivated than public agencies about open data; they may reason with their own individual interests in mind and so may not appreciate the benefits of open data. Here too, cultural shift is needed for attitudes to change [NLBIF].

88. Data publication is closely associated with drives to openness and good RDM practice, and offers the opportunity of bringing publishers into the frame as another key stakeholder group. There are efforts to achieve a common strategy in areas such as data publication, in collaboration with academic journals and commercial publishers (**para 37**). The development of data publication standards, through concepts such as rigorous peer review, data citability and visibility, and the deployment of DOIs, is part of this effort. GFBio is working with RDA to develop a quality/usability label for research data (**para 75**).

Training and skills

89. Training in good data practices represents a key part of data infrastructures, even if not a physical part. The GeRDI interviewee made a pertinent point on the relationship between training and fostering cultural change: one of the conclusions from the initial EOSC High Level Expert Group was that, with regards to effecting good research data practices, cultural change among researchers represents 80% of the challenge whereas technological change represents only 20%. Researchers still often tend to neglect data curation, because their main priority remains over-focused on using their data as a means of getting good research publications. As suggested above (**paras 42 44**), there could be some way to go before cultural shift takes place, and this is required for research data to be recognised as a valuable research output in its own right. It is important for skills to be nurtured in areas such as the proper understanding and application of metadata standards. For this reason, GeRDI is developing a training plan for those communities who wish to become part of the GeRDI system in future; its intention is eventually to run a fully-fledged training programme that would enable data centres to become part of its federated infrastructure.

90. Other FRDIs recognise the place of training, and the nurturing of skills, in their service offering. Within ADRN, each of the four constituent centres is funded to provide an offering that includes training and capacity building. The Farr Institute runs a working group on capacity building, geared to the training of the next generation of health informatics researchers. This covers the coordination of its master's degree courses and its leadership programme, which identifies early career researcher talent. In the Netherlands, training relating to data architecture and to research data management planning is organised under the auspices of the Research Data Netherlands consortium (**para 57**).

In astronomy, training activities for scientists are run through a project which coordinates European activities around the Virtual Observatory. The project organises hands-on schools and provides templates of tutorials that are reused in national contexts. In France, training activities for data providers are run by ASOV, the French arm of IVOA, which organises discussions on good practice and the use of data tools; these sessions also attract participants from related disciplines other than astronomy [CDS].

91. Not all federated infrastructures have a training offer: with biodiversity, although GBIF has provided training in the past, most of its member organisations now have experts in place to assist with the deployment of its data publication toolkit, so it is presumed that there is little call for training [NLBIF]. HumaNum does not have the resources for providing formal training; it feels that appropriate training can often be better addressed upstream by research institutions themselves. For CDSP, training shortcomings cause concerns. In social sciences quantitative analysis it is important to foster a data culture, which can only be done through teaching appropriate methodologies. But the problem in France, and more widely in Europe, is a shortage of relevant teachers. The CDSP interviewee pointed to the UK, where the Nuffield Foundation's Q-Step programme[30] is a serious attempt to address this through the funding of teaching programmes in a range of universities – which has beneficially acted as a magnet for students from across Europe.

Challenges and obstacles

92. Unsurprisingly, the development and implementation of federated infrastructures is beset by many challenges and obstacles. As we noted earlier, the concept of federated infrastructures is not always well understood; there is still a big explanatory task, for instance to dispel the misconception that infrastructures are simply a search engine and to explain them in terms of their relation to the research data lifecycle. Many of the difficulties facing FRDIs have been alluded to in earlier sections of the report, and they are reviewed here more systematically.

Complexity

93. Federated infrastructures reflect highly complex, fragmented research data environments, and this was reported by several interviewees; according to ADRN, even major research funders may find it difficult to coordinate activities in the face of such complexity. It can be a struggle to cope with different legal frameworks, administrative systems, funding regimes, regulations and policy environments. Indeed, as we have suggested in the section on definitions of 'federated', an important rationale for FRDIs is to cut through complexity to enable users to access resources as seamlessly as possible. But however much these complexities are concealed from researchers and other stakeholders they do not disappear – and addressing them remains one of the significant challenges confronting federated infrastructures. And, as highlighted by DeIC, there may be competing visions about how best to cope with complexity and fragmentation – and a corresponding need to develop a cohesive approach.

94. In complex environments, a first difficulty is to ascertain the range of different user needs and requirements and, on top of that, to get users and other stakeholders interested [DeIC]. Finding ways of engaging with end-users, ie with individual researchers, can also be problematic [SURFsara]. There are related problems associated in finding a common language that different stakeholders can relate to. In Germany, there is a dual challenge of generating a dialogue with players in IT departments, who often are not exposed to research, and also developing an understanding between different domain-specific experts and IT departments. Similar challenges might apply with regards to players within other professional cultures, such as curators and taxonomists [GFBio]. In Denmark, a vast number of organisations own registry data and it is important to engage all these different players in a unified dialogue. There is much technical variation characterising how the registries are founded – for instance, different access models; this underlines the importance of a rational approach to the processing of requests and applications from researchers, as outlined in the section on AAI (**para 69**).

95. A further dimension of complexity was highlighted by Nikhef. Infrastructures can become overly complex if too many factors are built into them (the Nikhef interviewee illustrated this with the example of attempts to build open grid services architecture). Complexity of this sort can lead to the building of unstable systems and reduces the likelihood of something useful emerging from a given infrastructure, which is why it may take ten years or more for it to settle down with appropriate applications and a capacity to process large amounts of voluminous data.

Footnotes

[30] nuffieldfoundation.org/q-step

Cultural shift

96. We have seen above (paras 42 to 46) how cultural factors affect demand for federated infrastructures. But such factors can sometimes be a brake too. In the realm of HSS, some communities lag behind other disciplines with regards to willingness to use research data methodologies. In these cases, a major pedagogical challenge for federated infrastructures is to promote cultural acclimatisation, effecting shifts in research culture through persuasion, raising awareness about how new methodologies can benefit individual research projects and helping users to evolve their research data habits in the light of technological developments.

Finance

97. Another major set of challenges relates to securing sources of finance for federated infrastructures. As expressed bluntly by the SURFsara interviewee, everyone realises that RDM is important, but no-one wants to pay for it. In his view, RDM funding suffers from research evaluation systems that remain aligned with conventional outputs such as research publications and numbers of PhD students. Long-term availability of research data, which is an important outcome of the research process, does not properly feature in research evaluation and therefore funders are less likely to provide financial support that helps ensure its availability. Only political pressure – through allocating more budgetary resources and/or adapting funders' key performance indicators – is likely to put RDM, and its associated infrastructure, at the heart of the research evaluation process.

98. Short term or narrowly focused approaches to funding may also be problematic. For instance, in Germany, the willingness of autonomous science organisations such as Helmholtz, Fraunhofer and the Max Planck Society to invest in cross-cutting initiatives is limited; there appears to be reluctance to spend money for services that go beyond individual research facilities. This may limit their capacity to foster federated infrastructures, whose operations can be labour-intensive, with high staffing costs [GFBio]. In Finland, current funding for FIN-CLARIN covers only the development of new services, not the support of existing services – which is problematic, since constant expansion leads to an inevitable need to finance increasing capacity [FIN-CLARIN]. And sometimes, there are only narrow windows for funding opportunities: when the Economic and Social Research Council funded ADRN, it had to commission the project very quickly to take advantage of the available capital funding. The main commissioning challenge was to implement the standard peer review processes for grants of this size within a tight timetable owing to time restrictions in capital funding allocations [ADRN].

Standards

99. The issue of standards has already been addressed above (paras 72 73 74 75).
There is a challenge in brokering for integrated pools of data from different domains. Different standards apply, and in GFBio's experience, putting in place mechanisms to allow for compatibility is a major and difficult endeavour, which is being addressed. Within given domains, there is therefore a need to converge to agreed standards. This is more of an organisational issue than one of computing power [GFBio].

Legal and regulatory issues

100. The legal status of research data can present a significant obstacle to access and sharing. This is particularly true of sensitive and/or confidential data, as exemplified by the material that is kept in registries. As outlined above (**para 79**), many registry owners do not allow data to leave their premises. This seriously hampers researchers' ability to run simulation and modelling, since data distributed between and confined in different sites cannot be processed in parallel. Added to that are the concerns, also expressed by data centres, about undermining their IPR on data (**para 79**). We have seen also how, at the European level especially, AAI and the associated issue of cross-country login can in itself represent a challenge (see paragraph 69). The problem of federated login is that, in theory, systems to allow this are in place but, in practice, legal constraints and regulatory or policy requirements (which vary from country to country) can act as real bureaucratic barriers [FIN-CLARIN].

Other challenges

101. *Defining the scope of the service:* as outlined above (**para 48**) in the context of GFBio's activities, there is a significant challenge in identifying priorities for federated infrastructures, defining the scope of their core service and establishing a corresponding business model.

102. *Storage and curation:* research data in most of the research disciplines is not rationally stored and curated. It is mostly decentralised, on personal computers or institutional or university data centres where it does not adhere to appropriate standards and is not easily accessible. A few research disciplines are characterised by centralised data management (for instance geological sciences, biological sciences, social sciences). Conversely, in astronomy, research data management is dispersed with common access rules. However, awareness of the necessity of long-term data storage and curation has increased recently [DNA Bank Network, CDS].

103. *Staffing:* staffing and employment issues may present problems. FRDIs activities can be dependent on very few people driving the infrastructures forward and often these have very limited back-up support. In bioinformatics, bioinformaticians and curators might be employed only temporarily within the federated infrastructures. Third party funding schemes often do not pay for international collaborations, for instance, to cover bioinformatics support for the establishment of a node in another country. Communication between different partners, especially across the globe with different working hours, is sometimes difficult [DNA Bank Network].

104. *Public attitudes:* with regards to sensitive, patient-related health data, members of the public mistrust the private sector. They are happy for data to be de-identified and accessed for public good by researchers, and look for assurances about this, for instance through lay representation on relevant governing structures. These are concerns that the Farr Institute addresses as part of its stakeholder engagement (**para 63**), but earning the trust of the public represents a challenge too.

Outcomes and impact

105. The report so far has reviewed the drivers behind federated infrastructures, the contexts in which they operate and their characteristics. But a crucial question remains: how effective are they in practice? What impact are they having on the research communities (and indeed on the broader range of stakeholders) that they serve? This is arguably the trickiest of all the questions, and not all interviewees were able to address it. In some cases, impact was described in terms of efficiency and cost-effectiveness or with regards to increased usage. It is clear that there is a correlation between outcomes, impact and user needs. The continued active engagement with users and other stakeholders is therefore likely to be crucial in determining the benefits of FRDIs; and in ascertaining how these benefits are perceived.

106. Some infrastructures look at impact in a quantifiable way. FIN-CLARIN deploys performance indicators, including tools to measure usage of resources; Google Analytics is used to measure usage levels, including applications for restricted resources, number of logins and where these logins occur. It has thus been possible to gauge increased usage, between 2015 and 2016, of language collections. HumaNum has been charting the markedly increased participation at its presentational events (also at presentations of its tools organised by third parties), and the increased demand for its expert advice – symptomatic of the success of its network and outreach capabilities. CDS monitors the large throughput of requests, processed through the Virtual Observatory, for the data it holds. GFBio pointed to the measures deployed by the PANGAEA Data Publisher for Earth & Environmental Science[31]; when this was set up, the intention was to encourage researchers to work on metastudies, which are founded on data from many different sources. Use of PANGAEA has indeed increased, with 20,000 unique users a month and scientists accessing data that they did not produce themselves. The vision is to measure impact by data citations in literature. In this respect, Scholix[32] – a high level interoperability framework for exchanging information about the links between scholarly literature and data – is an important development which can be used by individual data centres as well as FRDIs [GFBio].

107. Other infrastructures described reporting mechanisms in place to evaluate impact. DNA Bank Network measures outcomes and impact through review boards and the evaluation of achievements. At HumaNum, the chair of the scientific board provides the infrastructure's steering committee with an annual report on achievements. The Farr Institute has an obligation to report annually to its consortium of funders on a variety of impact factors, such as publications, student numbers, industry engagement and follow-on funding. KOR aggregates views on the usefulness and success of access to various registry datasets – although it does not evaluate the access environment on an aggregated, national basis. For DNA Bank Network, measuring outcomes and impacts is achieved through background information tracking on number of searches, number of requests from different countries and frequency of requests.

108. In other cases, outcomes and impact were evoked in general terms, without specific information, but broadly pointing out the benefits of the relevant infrastructures. CDS suggested that federalisation, over the years, has changed the way that astronomers conduct research – this is a major impact, albeit one that is difficult to measure. DeIC reported that community engagement with an open source environment has helped to develop capability for RDM and enabled learning on how to

handle research data. ADRN suggested that its federated approach helps to address the socio-economic impact requirements of its funder, the ESRC, through providing local data that can influence local services.

109. Benefits are also perceived in terms of cost effectiveness. An advantage of federalisation is that it allows for economies of scale, critical mass and a division of labour, and it provides access to expertise that might not otherwise have been available [Computerome]. Federalisation helps to ensure a more optimal and cost-effective utilisation of resources, which are made available or shared across infrastructures, and reduce the need for each institution or researcher to acquire such resources [SURFsara]. Value is also added through the way that federated infrastructures allow for working outside silos, thereby avoiding duplication of effort [Computerome, Farr].

110. Nevertheless, it remains the case that the measurement of impact can be challenging. The FIN-CLARIN interviewee stated that, in spite of the deployment of indicators as outlined above, it is difficult to quantify the measurement of success at present; for data resources in language technology, which is a fairly narrow field with a strong national focus, it is not clear what success represents. In other instances, federated initiatives are not yet sufficiently developed for meaningful impact to be ascertained. For DeIC, there is a need for a bigger user base, more deposited data and for greater visibility of the service before being able to reach views about experience. And according to NLBIF, the impact to date of federalisation has been limited, because at the local level of countries and regions the success of other structures is often greater than that of the global player, GBIF.

Implications for EOSC

111. A couple of interviewees hadn't heard of EOSC, and others did not really comment on its prospects. Nevertheless, among those who addressed the question, the general consensus was that EOSC represents a welcome development. Indeed, as the rationale behind it is similar to the justification for the development of other national or international federated infrastructures, the arguments in its favour have been rehearsed in other parts of this report. The view from Nikhef is that EOSC should enable transparent roaming across Europe, with the same service visible everywhere; and it should allow a choice about how to get to that service, to places where researchers feel they can get the best support, wherever that may be. That view is reinforced by Computerome, whose interviewee suggested that EOSC is a significant step for research because it would bring federated infrastructures to everyone and hugely broaden access, within clouds, to data and tools. So, in line with the suggested definitions of 'federated' set out above (**para 14**), EOSC is perceived as a means of easing access to resources and cutting through complexity. But there was also a note of concern, from DNA Bank Network, about the extent to which an additional interdisciplinary infrastructure would add value to what exists already in a European context.

Footnotes

[31] https://pangaea.de/ ; see also Diepenbroek, M. et al (2017). 'Terminology Supported Archiving and Publication of Environmental Science Data in PANGAEA', Journal of Biotechnology, https://doi.org/10.1016/j.jbiotec.2017.07.016

[32] http://scholix.org/home

112. It follows that key to EOSC's future success will be a business/finance model founded on an agreed and effective division of tasks between it and the resource providers. This should allow for the latter to provide their services across Europe, and researchers to access them, in a transparent way. But EOSC will fail if there is no such consensus. Getting the business model right could be more important than governance [SURFsara].

113. Conversely, the CDS interviewee suggested that governance could be the key factor for EOSC, because of the crucial importance of ensuring that the initiative is able to put user needs at its centre. Researcher representation on governance bodies could be a way of achieving this, although it is recognised that this would be difficult, given the large number of research disciplines.

114. Because in effect, EOSC would operate on the same principle as national FRDIs – an 'ecosystem of infrastructures', as suggested above (**para 7**), or a system of systems [Nikhef] – the experiences of existing infrastructures are likely to inform its own future development and policies. In a similar vein, EOSC should take into account the legacy of existing infrastructures. Conversely, it is important that these should 'fit' well within the EOSC constellation; and that there should be clarity about conditions of access of existing infrastructures and their users to EOSC. The principle of capitalising on existing infrastructures is also enshrined in the Global Open FAIR (GO FAIR) initiative[33], a proposal for the practical implementation of EOSC, formulated by a group of 'early mover' member states and aimed at the open and practical implementation of the recommendation in the EOSC High Level Expert Group first report (**para 7**).

115. In evolving such a role as an ecosystem, there is a risk that EOSC might focus too much on developing new, pan-European centralised tools (even though in some circumstances there is justification for them), to the detriment of what a European cloud might do most usefully: act as a coordinator and aggregator of existing, national or disciplinary tools in order to make them interoperable. This point was stressed by CDS: EOSC should build on existing, tried and tested disciplinary infrastructures, and provide generic services for communities which do not have these disciplinary frameworks, but not build a rigid mould around generic e-infrastructures.

116. At the same time, EUDAT[34], an infrastructure of integrated data services and resources to support research, might be seen as a pan-European precursor to EOSC. For instance, the IVOA registry of resources is already integrated in EUDAT registry B2FIND[35]. On the assumption that EOSC builds on the achievements of EUDAT, there is a great potential to provide easily shared and co-maintained services for European countries. It could also help to get computation closer to the data. However, this would require a high level of maturity from all parties involved [FIN-CLARIN].

117. Other precedents exist at the regional, sub-European level. Within the group of Nordic countries (Denmark, Finland, Norway, Sweden), the Tryggve project[36] provides a platform for collaboration on sensitive bioinformatics data, allowing for coordination and unified access mechanisms to data whose location remains decentralised, either for reasons of security/ confidentiality (which is important for sensitive data) or of quality assurance. There are precedents too at the disciplinary level: the Belmont Forum[37] is a partnership of funding organisations, international science councils, and regional consortia committed to the advancement of interdisciplinarity and transdisciplinarity in environmental change research. Such approaches, founded on decentralisation of data, researchers and structures, could serve as a model for EOSC [KOR, DNA Bank Network].

118. Finally, because of the scale at which it would operate, EOSC could greatly increase the scope for data and tools to be shared across disciplines, thereby overcoming silos and fostering cross-disciplinary fertilisation [Computerome].

Footnotes

[33] https://dtls.nl/fair-data/go-fair/
[34] https://eudat.eu/
[35] https://eudat.eu/services/b2find
[36] https://wiki.neic.no/wiki/Tryggve
[37] https://belmontforum.org/

Full conclusions

119. As outlined at the beginning, this report draws on the 16 interviews that have taken place over the past few months. It is intended as a synthesis of those discussions but it is not a comprehensive account of everything that was said, nor of the overall research data infrastructure in the relevant countries. Its coverage is necessarily selective, and seeks to highlight common themes, to contrast approaches and to flag up what appeared to be interesting ideas. In these respects, all the interviews provided valuable material, although it should be stressed that they often reflect only the views of the selected interviewees; the report is not always, therefore, a full representation of the 'official' line of the organisations in question. Nevertheless, what it expresses is valuable in that it provides honest perspectives from experts with significant experience of research data infrastructures.

120. Nine broad conclusions may be drawn from this evidence, and these are set out below. They do not reflect every finding, but illustrate the themes and issues that have emerged most strongly. They are also reproduced in summary at the beginning of the report.

Conclusion 1

Federated infrastructures can apply to research disciplines across the spectrum, including physical sciences, life/health sciences, environmental sciences, social sciences and the humanities; they may also be multidisciplinary. Similar infrastructure drivers, characteristics or functions may relate equally to different disciplines, with one notable exception: the sensitive and confidential nature of registry data, with all the data protection, legal and ethical constraints that this implies, tends to apply particularly to health sciences and social sciences.

Conclusion 2

There is no single definition of 'federated' that emerges from the interviews. However, there is a broad common denominator which most interviewees would identify with, and around which a common definition might be articulated: **essentially, a federated infrastructure is one where a range of distributed services are coordinated by an overarching level**. This is very generic, but it may suffice to derive a definition that can be understood and agreed across research disciplines and countries.

Conclusion 3

Two broad sets of factors drive the emergence and development of federated infrastructures: push factors, which might be also be characterised as top-down; and demand from users, reflecting a bottom-up approaches. Both are relevant in most national contexts.

Push factors might be said to correspond to a hierarchy. At the topmost and most overarching level are the factors that reflect broad social and political imperatives, such as current societal concerns – for instance, addressing climate change, the desire for open societies. Next down in this hierarchy are the requirements of national public policy, driven by government and public agencies in areas such as open data. Below this, are drivers from the players in the broad research environment: funders, universities, infrastructure providers and sometimes from other players too, including commercial organisations.

Demand, or bottom-up factors, reflect different research cultures, which may vary considerably, from the highly collaborative (for instance, high energy physics, astronomy) to those communities who may find it more difficult to embrace an e-research culture (exemplified by some humanities).

There is also a crossover between push and demand, whereby national or international players take steps to ascertain the needs and expectation of the research communities that they serve. In a way, this is about top-down organisations trying to foster a bottom-up dialogue.

Conclusion 4

Infrastructures are often characterised by long-term financial uncertainty, with funding typically taking the form of project grants allocated for finite periods – which raises the issue of sustainability. This sort of short-termism is not helpful for the development of FRDIs, and funding models need to evolve to reflect the strategic place of research data in the research process.

Conclusion 5

The involvement of users is also a crucial imperative, and infrastructures are careful to nurture their relationships with numerous partners within the academic sector and beyond. Engagement with stakeholders can form an important part of this, and efforts are often made to foster and maintain a dialogue with users – reflecting a desire to ascertain demand – and other players, including those within broader society.

Conclusion 6

Infrastructures are characterised by a wide range of practices and services, which vary according to the nature of each initiative and evolve dynamically in the light of researchers' needs. There is no single model of service provision, no template that might apply universally to all infrastructures. This is not surprising, since these are determined to an extent by the particular circumstances and cultures that have fostered their emergence and development. Some FRDIs recognise the importance of a holistic approach to their service offer, articulated around the entire research data lifecycle. For most of them, a key characteristic of their activity is to provide the means of allowing easy, intuitive and seamless access to distributed resources, irrespective of the latter's location. Other factors that FRDIs often address include AAI (authentication, authorisation, identification), with all the legal issues that this raises; usability/interoperability; data standards; security; ethical and legal issues, particularly where sensitive data is involved; sharing and linking; and openness. Training programmes and the nurturing of skills feature for some infrastructures, but this is not something that all FRDIs are in a position to support.

Conclusion 7

A major challenge for the development of federated infrastructures is the complexity and fragmented nature of the research data environments in which they evolve. There are difficulties in addressing the different legal frameworks, administrative systems, funding regimes, regulations and policy environments and also in ascertaining the range of different user needs. Other challenges reflect the sort of practices outlined above. They include the slow process of effecting cultural change, identifying reliable sources of finance, ensuring compatibility of standards and addressing the maze of legal and regulatory requirements.

Conclusion 8

Many infrastructures have processes in place to evaluate impact, either through measuring usage in a quantifiable (but in practice, often a limited) way, or through formal review mechanisms overseen by governance bodies. Achieving cost-effectiveness is also seen as an important benefit, achievable for instance through the economies of scale that federalisation can afford. However, quantifying the measurement of success remains difficult, and impact is often perceived in fairly general terms.

Conclusion 9

The emergence of EOSC is generally welcomed, particularly since it is seen as reflecting the same rationale as national infrastructures, albeit at a pan-European scale – with the beneficial scaling up that this could imply. EOSC's future success will depend on the consensual formulation of a well thought-out business and finance model, and also a solid governance structure, which capitalises on a clear division of tasks between it and the different resource providers. EOSC should also ensure that it puts user needs at its centre. EOSC might add most value if it evolves as an aggregator of existing services, rather than as a provider of new, centralised tools

Annex A Interview questions

Pilot phase interviews

National overview

i. What is the general status of the existing infrastructure for research data in your country and how is it organised?
ii. Is there a specific national policy or strategy with respect to development of and support for federated data infrastructures?
iii. What makes an initiative 'federated' in the national context?

Specific national example(s); description

Factual description of the federated infrastructure:

iv. Discipline(s) involved
v. Lead and partners involved
vi. Mandate, governance and (long term) funding
vii. Structure and organisation
viii. Services and their business model
ix. Involvement of target community

Assessment of the federated infrastructure:

x. What were the reasons to decide for a federated data infrastructure?
xi. What or who sparked the idea and pushed the initiative?
xii. What are or were the obstacles to implement the federated infrastructures?
xiii. How are these challenges met?
xiv. Do issues dealing with public-private services play a role in the federated infrastructure? If yes, please identify do's and don'ts
xv. What have been the successes of the federated infrastructure?
xvi. What is the experience with respect to acceptance by the scientific communities and other relevant stakeholders?
xvii. What lessons have been learnt that could be valuable for others?

Questions addressing the overall purpose of identifying:

xviii. What does it mean for researchers that the services are federated?
xix. Are there any positive or negative consequences for researchers, having to deal with federated (rather than single, autonomous) services?
xx. In what way does the federated infrastructure take the researchers' interests into account?
xxi. What are lessons learned, experiences that can help emerging FRDIs, such as EOSC?
xxii. Are you aware of the developments in EOSC and the EOSC pilot projects?
xxiii. Are there any experiences in your own FRDI that you think could be of use for EOSC and the EOSC pilot projects to take into account?
xxiv. What do you think are the biggest challenges, fail and success factors for emerging FRDIs?

Phase 2 interviews

i. To provide us with context, please tell us briefly about the status and organisation of the existing infrastructure for research data in your country.
ii. How would you define 'federated'?
iii. What services and/or facilities do FRDIs provide in your country, and what are their characteristics, for instance structure, organisation, governance, stakeholder engagement, training provision?
iv. What are the factors influencing or driving FRDIs in your country and internationally, from the perspective of public agencies, infrastructure providers, research funders and other stakeholders?
v. How and to what extent do FRDIs in your country reflect demand from and the culture of the research community?
vi. How does the relationship between the factors at Q4 and the demand at Q5 affect the development of FRDIs?
vii. What have been the challenges in developing FRDIs in your country, and how have these challenges been addressed?
viii. What value does federalisation add, and by extension, what is the impact of FRDIs, and what mechanisms are being put in place to measure outcomes and impact?
ix. In the light of your experience of FRDIs in your country, please give us your thoughts on the development of EOSC, and what role this might usefully play in Europe

Knowledge Exchange Office
C/ O Jisc,
One Castlepark,
Tower Hill,
Bristol, BS2 0JA

t: 0203 697 5804
e: office@knowledge-exchange.info